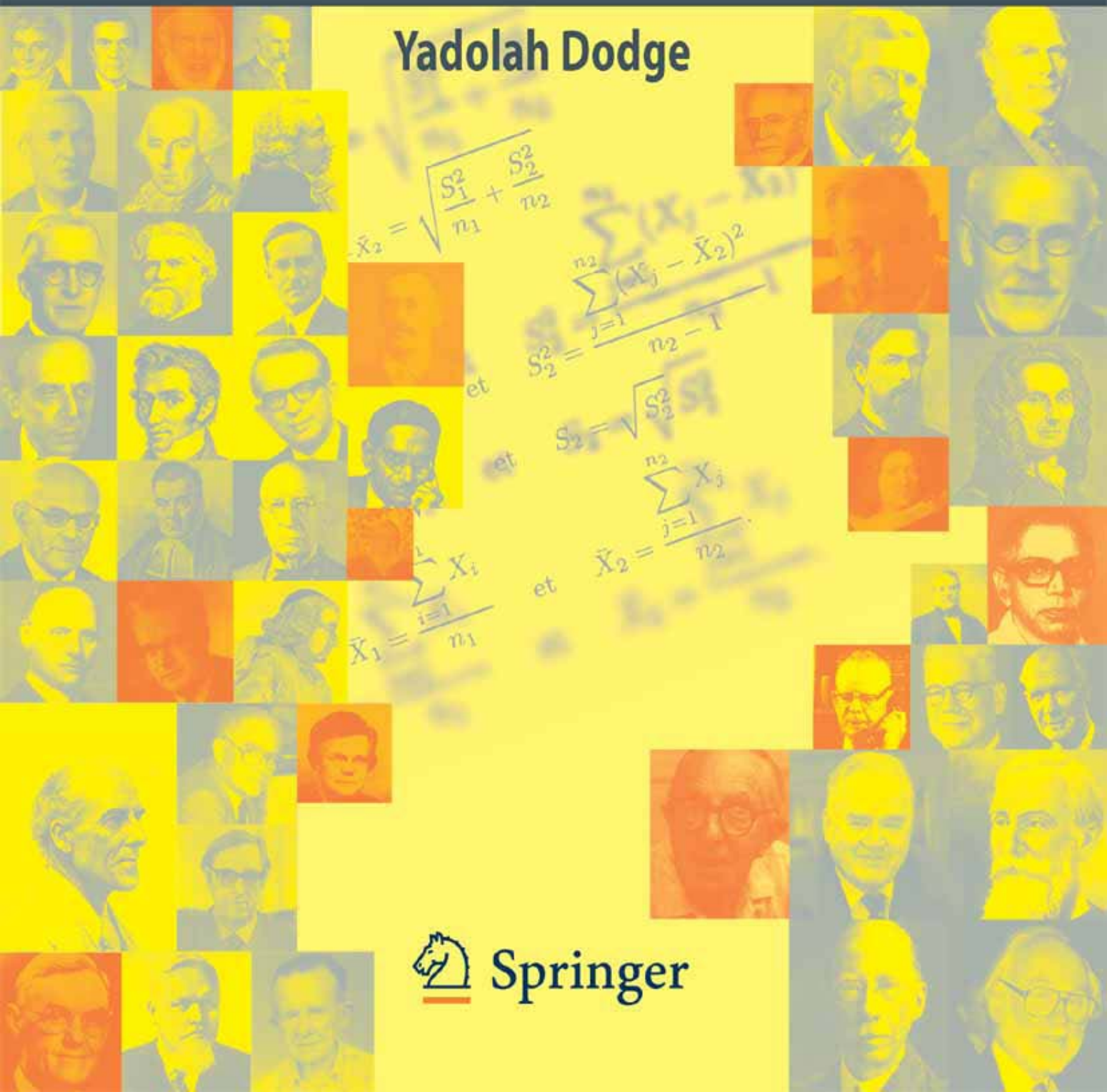
STATISTIQUE

Dictionnaire encyclopédique

Yadolah Dodge



Springer

Statistique

Dictionnaire encyclopédique

Springer

Paris
Berlin
Heidelberg
New York
Hong Kong
Londres
Milan
Tokyo

Yadolah Dodge

Statistique

Dictionnaire encyclopédique

Springer

Yadolah Dodge

Professeur honoraire
Université de Neuchâtel
Suisse
yadolah.dodge@unine.ch

ISBN : 978-2-287-72093-2 Springer Paris Berlin Heidelberg New York

© Springer-Verlag France, Paris, 2007
Imprimé en France

© Dunod, Paris, 1993 pour la première édition

Springer-Verlag France est membre du groupe Springer Science + Business Media

Cet ouvrage est soumis au copyright. Tous droits réservés, notamment la reproduction et la représentation, la traduction, la réimpression, l'exposé, la reproduction des illustrations et des tableaux, la transmission par voie d'enregistrement sonore ou visuel, la reproduction par microfilm ou tout autre moyen ainsi que la conservation des banques de données. La loi française sur le copyright du 9 septembre 1965 dans la version en vigueur n'autorise une reproduction intégrale ou partielle que dans certains cas, et en principe moyennant le paiement de droits. Toute représentation, reproduction, contrefaçon ou conservation dans une banque de données par quelque procédé que ce soit est sanctionnée par la loi pénale sur le copyright.
L'utilisation dans cet ouvrage de désignations, dénominations commerciales, marques de fabrique, etc. même sans spécification ne signifie pas que ces termes soient libres de la législation sur les marques de fabrique et la protection des marques et qu'ils puissent être utilisés par chacun.

SPIN : 12053858

Maquette de couverture : Jean-François Montmarché

Garde-toi bien de la fumée des cœurs blessés
La blessure du cœur se rouvrira sans cesse.
Tant que tu le pourras, n'oppresse pas un cœur
Le soupir d'un seul cœur renversera le monde.

Saadi Shirazi (1184-1271), poète persan

Cet ouvrage est dédié à tous les enfants de par le monde, qui sont les victimes innocentes des atrocités humaines.

Du même auteur

Aux Éditions Springer-Verlag France

Dodge Y (1999) Premiers pas en statistique. Springer-Verlag, Paris.

Dodge Y (2002) Mathématiques de base pour économistes. Springer-Verlag, Paris.

Dodge Y (2005) Optimisation appliquée. Springer-Verlag, Paris.

Autres éditeurs

Arthanari TS and Dodge Y (1981) Mathematical Programming in Statistics. John Wiley and Sons, New York.

Arthanari TS and Dodge Y (1993) Mathematical Programming in Statistics. Classic edition, John Wiley and Sons, New York.

Dodge Y (1985) Analysis of Experiments with Missing Data. John Wiley and Sons, New York.

Dodge Y (1987, 1989, 1996) Mathématiques de base pour économistes. Éditions EDES puis Presses Académiques, Neuchâtel.

Dodge Y (1987) Introduction à la programmation linéaire. Editions EDES, Neuchâtel.

Dodge Y (1987) (Ed.) Statistical Data Analysis based on the L1-norm and Related Methods. North-Holland, Amsterdam.

Dodge Y, Fedorov VV and Wynn HP (1988) (Eds) Optimal Design and Analysis of Experiments. North-Holland, Amsterdam.

Dodge Y (1989) (Ed.) Statistical Data Analysis and Inference. North-Holland, Amsterdam.

Dodge Y, Mehran F et Rousson M (1991) Statistique. Presses Académiques, Neuchâtel.

Dodge Y (1992) (Ed.) L1-Statistical Analysis and Related Methods. North-Holland, Amsterdam.

Dodge Y and Whittaker J (1992) (Eds) Computational Statistics I. Physica Verlag, Heidelberg.

Dodge Y and Whittaker J (1992) (Eds) Computational Statistics II. Physica Verlag, Heidelberg.

Birkes D and Dodge Y (1993) Alternative Methods of Regression. John Wiley and Sons, New York.

Dodge Y (1993) Statistique : Dictionnaire encyclopédique. Dunod, Paris.

Dodge Y (1997) (Ed.) L1-Statistical Procedures and Related Topics. IMS Lecture Notes and Monographs Series, Volume 31, Berkley, U.S.

Dodge Y (1999) Analyse de régression appliquée. En collaboration avec V. Rousson. Dunod, Paris.

Dodge Y and Jureckovà J (2000) Adaptive regression. Springer-Verlag, New York.

Dodge Y (2002) (Ed.) Statistical Data Analysis based on the L1 norm and related methods. Birkhauser, Basel.

Dodge Y (2003) The Oxford Dictionary of Statistical Terms. Oxford University Press, Oxford.

Avant-propos

L'idée de cet ouvrage est née d'un véritable besoin, celui de combler une lacune importante dans la littérature relative à la science statistique. En effet, il n'existait à ce jour aucun dictionnaire de statistique en langue française. Il est vrai qu'il s'agit d'une science relativement jeune, reconnue en tant que telle depuis le début du xx^{e} siècle seulement.

Notre objectif est de donner la définition des concepts les plus couramment utilisés tout en complétant l'information chaque fois que possible, par des historiques, des aspects mathématiques, des domaines et limitations, des exemples, des références et des mots-clés, qui sont donnés avec leur équivalent en anglais (un mot-clé étant une notion qui peut être liée au mot de base en fonction de plusieurs critères). Les mots apparaissant en gras dans le texte indiquent des correspondances qui renvoient à d'autres notions statistiques, permettant d'expliquer et de développer le terme considéré. Il faut toutefois préciser que ce dictionnaire n'a pas la prétention de réunir en un seul volume toutes les notions de statistique existantes à ce jour, cette science étant en constante évolution.

Le premier dictionnaire de cette science à avoir été publié est celui de Sir Maurice Kendall et William Buckland, en 1957. Commandé par l'Institut international de statistique avec l'appui financier de l'UNESCO, il a pour objet la définition des termes statistiques les plus utilisés dans le monde scientifique. En 2003, la sixième édition paraît sous le nom *The Oxford Dictionnary of Statistical Terms*, édité par moi-même.

En 1968, une encyclopédie internationale des statistiques en deux volumes est éditée par William Kruskal et Judith Tanur (rééditée en 1978). Elle comprend 174 articles de différents auteurs classés par ordre alphabétique.

Dès 1982, Samuel Kotz et Norman Johnson, ont publié une encyclopédie des sciences statistiques (*Encyclopedia of Statistical Sciences*) en dix volumes comptabilisant 6 604 pages au total. Or, de l'avis même des éditeurs, ces quelques volumes ne peuvent contenir tous les termes, concepts, notions et méthodes nés des recherches de ce siècle.

En 1998, Peter Armitage et Theodore Colton ont publié une encyclopédie de biostatistique en six volumes (*Encyclopedia of Biostatistics*).

La dernière encyclopédie en date est *The Encyclopedia of Environmetrics* éditée par A.H. El-Shaarawi et W.W. Piergorsch.

Le présent dictionnaire intéressera un vaste public comme celui des sciences économiques et sociales, des sciences humaines ou des sciences de la vie et de la terre. Il est aussi bien destiné aux chercheurs de divers domaines de sciences appliquées qu'aux étudiants envisageant d'approfondir la théorie statistique et ses applications. Certains domaines importants de la statistique moderne n'ont pas été abordés dans cet ouvrage, telles les méthodes statistiques génétiques et la reconnaissance d'image ; mais un dictionnaire exhaustif est impossible à réa-

liser en un seul volume. Nous serions reconnaissants aux lecteurs qui auraient des remarques à formuler, de bien vouloir nous les communiquer afin que nous puissions en tenir compte lors d'une prochaine édition.

Si le lecteur désire approfondir certains sujets, il pourra se reporter à la liste des références à la fin de chaque définition qui, bien que non exhaustive, pourra servir de guide. Il est également conseillé de consulter l'*Encyclopedia of Statistical Sciences*.

Ce manuscrit a été imprimé pour la première fois en 1993 avec le même titre par les éditions Dunod, Paris. Depuis, il a été complété et amélioré.

Pour la première édition, je tiens à remercier tout particulièrement Nicole Rebetez, Sylvie Gonano et Jürg Schmid pour leur étroite collaboration, ainsi que Séverine Pfaff qui a réalisé la saisie de cet ouvrage. Je témoigne ma profonde reconnaissance aux professeurs Gérard Antille, Bernard Fichet, Farhad Mehran et Jean-Pierre Renfer pour avoir apporté les dernières corrections à cet ouvrage, de même qu'à Rachid Maraï pour ses illustrations. Je remercie également mes assistants : Sandrine König, Valentin Rousson, ainsi que les professeurs Ludovic Lebart, Valérie Isham et Joe Whittaker pour leur précieuse collaboration. Enfin, je tiens à remercier le Fonds national suisse de la recherche scientifique (requêtes n° : 20-5648.88 et 20-28989.90) pour sa contribution financière.

Pour cette deuxième édition, je remercie Elisabeth Pasteur, Aline Lykiardopol et Jimmy Brignony pour leur participation à l'élaboration de cet ouvrage. Enfin, sans l'aide précieuse d'Alexandra Fragnière, qui a corrigé et recorrigé avec beaucoup de passion, je n'aurais pas pu présenter cet ouvrage sous sa forme actuelle.

Yadolah Dodge
Professeur de Statistique
Université de Neuchâtel, Suisse
Mars 2004

La statistique : une science omniprésente

INTRODUCTION

La statistique est une discipline qui concerne la quantification des phénomènes et des procédures inférentielles. Elle a trait, en particulier, aux problèmes de mise en œuvre et d'analyse des expériences et échantillons, à l'examen de la nature des erreurs d'observation et des sources de variabilité, et à la représentation simplifiée des grands ensembles de données. La théorie statistique est le cadre qui fournit un certain nombre de procédures que l'on appelle *méthodes statistiques*. Le terme *statistique* est parfois utilisé dans d'autres sens, par exemple, pour se référer non pas à une discipline globale comme décrite ici mais, de manière plus limitée, à un ensemble d'outils statistiques comprenant des formules, des tableaux et des procédures, Dans un sens plus étroit, le terme *statistique* est aussi employé pour désigner un ensemble de données numériques, comme par exemple, les statistiques de chômage en Suisse, et lorsqu'il est au singulier, il peut encore être utilisé pour désigner un paramètre numérique, une estimation calculée à partir d'observations de base. La statistique est une discipline dont les racines remontent loin dans l'histoire. Considérons, par exemple, la notion de moyenne arithmétique qui est une des plus anciennes méthodes utilisée pour combiner des observations afin d'obtenir une valeur approximative unique. Son emploi semble en effet remonter au troisième siècle avant notre ère, époque où les astronomes babyloniens l'utilisaient pour déterminer la position du soleil, de la lune et des planètes. La statistique en tant que discipline est née de l'étude et de la description des états, sous leurs aspects économiques et surtout démographiques. En effet, des formes de recensement se pratiquaient déjà à l'époque des grandes civilisations de l'Antiquité, où l'étendue et la complexité des empires nécessitaient une administration stricte et rigoureuse. Les pratiques dans ce domaine ont bien entendu fortement évolué avec le temps, les méthodes nouvelles se sont développées, et ainsi de la simple collecte des données, la statistique a peu à peu pris un sens nouveau. De plus à la notion de *comptage* ou de *mesure* est venu s'ajouter le problème de la comparaison. Or, la comparaison de mesures nécessite une connaissance de leur exactitude, de leur précision, ainsi qu'une maîtrise des techniques pour formuler l'incertitude rattachée aux valeurs. C'est là que la statistique inférentielle et la théorie des probabilités prennent toute leur importance. Dès le XVII^{e} siècle, de grands savants ont participé au développement de la statistique telle qu'on la connaît

aujourd'hui. Citons, parmi tant d'autres, Jakob Bernoulli au XVII^e siècle ; Pierre Simon de Laplace, Thomas Bayes et Roger Joseph Boskovich au XVIII^e ; Adrien-Marie Legendre, Carl Friedrich Gauss et Siméon-Denis Poisson, au XIX^e ; Karl Pearson, William Sealy Gosset, Ronald Aylmer Fisher, Jerzy Neyman et Egon Sharpe Pearson, Sir David Cox, Jack Kiefer, Bradley Efron, C.R. Rao, Peter Huber et Eric Lehmann au XX^e. Tous ont contribué à faire de la statistique une science en soi. La statistique moderne est beaucoup plus qu'un ensemble de méthodes, d'outils ou de techniques isolées utilisées dans des sciences spécifiques. Il existe une unité des méthodes statistiques, une logique cohérente, une base scientifique, mathématique et probabiliste, solide. La statistique n'est véritablement reconnue comme discipline en tant que telle que depuis le début du XX^e siècle. C'est donc une science relativement jeune, en pleine expansion que ce soit au niveau de la recherche fondamentale ou des applications. Par ailleurs, de nombreuses branches nouvelles, ainsi qu'un grand nombre de spécialisations se sont développées : la théorie de la décision, des séries chronologiques, l'analyse multivariée, l'économétrie, la théorie des jeux, l'inférence non paramétrique, reéchantillonage, envirométrie, biostatistique, statistique génétique, statistique spatiale et simulation (MCMC), modélisation (GLM, régression), modèles graphiques.

RECHERCHE ET STATISTIQUE

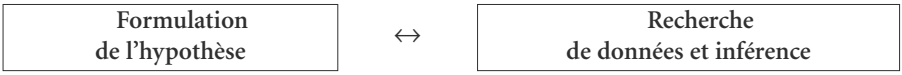
La recherche est une investigation ou une expérimentation critique ayant pour objectif la découverte et l'interprétation correcte de nouveaux faits ou de nouvelles relations entre différents phénomènes. Elle a également pour fonction la vérification de lois, conclusions ou théories acceptées, et à la lumière de faits nouveaux, le développement de nouvelles lois, de nouvelles conclusions et de nouvelles théories.

Comment accédons-nous à la connaissance des phénomènes naturels ? Quels sont les processus mentaux impliqués et comment aborder les investigations ? Comment juger si les données et les résultats obtenus sont exacts ? Ces questions ont en effet troublé la pensée humaine et sont restées longtemps un sujet de discours philosophiques. Cependant, les progrès récents de la logique et de la science statistique nous ont permis de leur apporter des réponses satisfaisantes.

La connaissance d'un phénomène naturel est habituellement définie en des termes et des lois qui permettent de prévoir les évènements de façon précise dans les limites raisonnables du possible. C'est le cas pour la loi de la dynamique de Newton et pour celle de la génétique de Mendel.

Comment ces lois ou théories s'établissent-elles ? Il existe une méthode scientifique. En premier lieu, une loi est formulée en tant qu'hypothèse chargée d'expliquer certains évènements observés, puis, les conséquences de l'hypothèse sont résolues par un raisonnement déductif et vérifiées par d'autres observations provenant d'expériences soigneusement conçues. Si les

données sont en contradiction avec l'hypothèse, cette dernière est écartée et une nouvelle hypothèse est formulée. Dans le cas contraire, elle est provisoirement acceptée et on lui accorde le statut de loi avec des limites spécifiques. Ceci laisse la possibilité de remplacer la loi au cours du temps par une autre loi plus universelle appuyée par un nombre plus important de données. Le cycle logique de la méthode scientifique de la recherche peut être schématisé ainsi :



De ces deux phases, la *recherche de données* est traitée par la statistique, alors que la *formulation de l'hypothèse* naît de la curiosité et de l'observation. Ces deux phases demandent beaucoup d'imagination et des connaissances techniques poussées pour aboutir à un progrès scientifique. Il est à souligner que les lois scientifiques ne progressent ni par le principe de l'autorité, ni par le mythe ou la philosophie moyenâgeuse. **La statistique reste la seule base des connaissances actuelles**.

Grâce à une analyse fine des données destinées à tester des hypothèses fixées, la statistique permet au scientifique d'utiliser au maximum ses capacités créatives et de découvrir des phénomènes nouveaux. Une découverte est, sans doute, le fruit d'une illumination soudaine, Mais ceci arrive seulement à un cerveau capable de trier le bon grain de l'ivraie, de chercher des éléments nouveaux et d'être attentif à l'inattendu pour mettre le tout ensemble, afin de provoquer l'étincelle qui donnera naissance à une découverte. Le travail du statisticien consiste à faire ressortir le maximum d'informations des données pour que les chiffres révèlent ce qu'ils contiennent. Selon son expérience, ses connaissances dans le domaine concerné et sa collaboration avec les experts compétents, le statisticien peut faire un examen plus approfondi des données pour comprendre un phénomène caché.

Cependant, l'acquisition de la connaissance n'est pas toujours liée à des théories ou des lois. Quelle quantité de blé produira-t-on cette saison ? La fumée provoque-t-elle le cancer ? Qui a écrit Hamlet : Shakespeare, Bacon ou Marlowe ? Quelle pourra être la place exacte d'une tumeur dans le cerveau d'un patient ? Comment les langues ont-elles évolué ? Comment mesurer l'effet d'un vaccin contre la polio ? Voici quelques questions ayant été résolues par la statistique et qui ont ainsi contribué au développement des connaissances.

EXEMPLES

Considérons maintenant quelques illustrations d'approches statistiques au service de la recherche de la vérité ou de nouvelles lois.

Diagnostic statistique des maladies

Les médecins font des diagnostics de maladie sur la base de quelques tests et symptômes. Le fait que deux médecins puissent donner des diagnostics différents, à partir de mêmes données et cela même pour des cas simples, montre que les tests sont insuffisants et que chaque médecin utilise sa propre expérience pour établir un diagnostic. Bien sûr, on pourrait effectuer davantage de tests (ce qui augmenterait considérablement les coûts). Mais le problème réel est celui de l'efficacité des tests. Tous les médecins ne sont pas en mesure de comprendre ou de travailler avec une grande masse de données qui, le plus souvent, créent la confusion. Il faut de plus pouvoir relier de petits indices tirés de tests spécifiques pour arriver à la meilleure conclusion. Pour faire face à de telles situations, les statisticiens ont développé ce qu'on appelle l'analyse de variantes multiples par laquelle l'information des mesures individuelles (comme les tests cliniques) peut être combinée de façon efficace pour répondre à une question spécifique (par exemple, quelle est la nature de la maladie ?).

En utilisant les dossiers des patients et l'expérience des meilleurs médecins, on peut construire une fonction discriminante fondée sur divers tests cliniques et ainsi déterminer les tests et valeurs qui peuvent au mieux distinguer les différentes maladies.

Le style littéraire quantifié

La statistique ne se limite pas seulement à la recherche scientifique. À en juger par les articles de la presse scientifiquement « branchée », on est frappé de voir l'accent mis sur les méthodes quantitatives et statistiques, même dans les domaines les plus inattendus comme les arts.

Discutons de quelques explications statistiques dans le domaine de la littérature et de l'étude des langues. L'œuvre de Platon est restée intacte après vingt-deux siècles et ses idées philosophiques tout comme son style, ont été étudiés dans les moindres détails. Malheureusement, personne n'a mentionné ou ne savait l'ordre chronologique correct de ses trente-cinq dialogues, de ses six pièces courtes et de ses treize lettres. Le problème de la chronologie des œuvres de Platon se posait déjà au siècle passé, mais aucun progrès n'apparut. Il y a quelques années, les statisticiens se sont mis à l'œuvre et ont réussi à établir un ordre qui semble assez logique.

La méthode statistique commence par établir pour chaque paire d'œuvres un index de similitude. Dans une étude menée par Lilian I. Boneva (*Mathematics in Archeological and Historical Sciences*, 1971, pp. 174-185), l'index était fondé sur la fréquence dans chaque œuvre des trente-deux descriptions possibles des cinq dernières syllabes d'une phrase.

En se fondant uniquement sur l'hypothèse que les œuvres les plus rapprochées dans le temps se ressemblent d'avantage que celles écrites à une autre période, une méthode a pu être développée et appliquée pour résoudre le problème de leur ordre chronologique.

En étudiant la parenté des langues indo-européennes (le latin, le sanscrit, l'allemand, le persan, le celte, etc.), la linguistique a découvert un ancêtre commun qui a été parlé, on le

pense, il y a quatre mille cinq cent ans. Si un tel ancêtre existe, il y a donc un arbre évolutif des langues qui s'est ramifié à différents moments de l'histoire. Est-il possible de construire cet arbre des langues comme les biologistes l'ont fait avec l'arbre évolutif de la vie ? Il s'agit d'un problème très intéressant et scientifiquement important qui relève de ce que l'on appelle la *glottochronologie*. En utilisant une grande quantité d'informations sur les similitudes entre les langues et un raisonnement fort compliqué, les linguistes ont réussi à établir quelques branches principales de l'arbre, mais les relations exactes entre elles et le moment de l'histoire où elles se sont séparées restent à déterminer. Mais, une approche purement statistique de ce problème, tout en utilisant beaucoup moins d'informations, a pu donner des résultats encourageants.

La première étape d'une telle étude est de comparer une catégorie de mots appartenant à des langues différentes et ayant une utilité commune (l'œil, la main, la mère, etc.). Les mots de même sens appartenant à des langues différentes sont marqués d'un « + » s'ils sont analogues, et d'un « - » dans le cas contraire. La comparaison de deux langues devient ainsi une séquence de « + » et de « – ».

En utilisant uniquement cette information, Morris Swadesh (*Proceeding of the American Philosophical Society*, 1952, 96, pp. 452-463) a proposé une méthode pour estimer l'époque de la séparation de chaque paire de langues. Une fois les dates de séparation de chaque paire de langues connues, il est facile de construire un arbre évolutif

La statistique servant à détecter les erreurs

Puisque l'acceptation d'une théorie dépend de sa vérification sur des données observées, le scientifique peut être tenté de *truquer* les données afin de pouvoir imposer ses idées par des théories bricolées. À part le fait que cette pratique est immorale, elle cause aussi beaucoup de tort au progrès de la science. Sans doute, une théorie fausse, sera-t-elle repérée un jour ou l'autre. Il est toutefois possible qu'elle nuise à la société aussi longtemps qu'elle est acceptée. Un exemple récent est celui de la fraude impliquant Cyril Burt, *le père de la psychologie pédagogique de l'Angleterre*, alors qu'il étudiait le quotient d'intelligence (QI). Sa théorie, selon laquelle les différences d'intelligence étaient héréditaires et ne dépendraient pas de facteurs socio-économiques, dirigea la pensée politique en matière d'éducation sur une mauvaise voie. Or, on a montré que l'augmentation du taux de vitamines B et de fer chez les femmes enceintes des groupes socio-économiques désavantagés, affectait positivement les scores moyens de leurs enfants de 5 à 8 points. Il semble évident que des facteurs autres que l'hérédité interviennent aussi dans les résultats des tests.

De telles falsifications peuvent être détectées par le fait que, comme Haldone l'a mentionné, l'homme est un animal méthodique qui ne peut pas imiter le désordre de la nature. En utilisant cette limitation, les statisticiens ont pu inventer des techniques permettant de détecter des données falsifiées.

Quelques grands scientifiques ont falsifié des données dans leur souci de faire avancer des théories qu'ils croyaient irréfutables et indispensables. R.A. Fischer a examiné les données expérimentales à partir desquelles Mendel découvrit les lois fondamentales de l'hérédité. Dans la plupart de ses expériences, les rapports des phénomènes observés étaient presque trop proches de ceux prévus par Mendel. Si on lance, par exemple, une pièce de monnaie 100 fois, il y a des chances d'obtenir 50 fois face, mais si vous prétendez obtenir ce résultat très souvent, on pourrait vous soupçonner de fraude. Ainsi Fisher a montré que la précision des données de Mendel (en accord avec sa théorie) était frappante à un point tel que si l'on calculait la probabilité de ce qu'il a trouvé expérimentalement, on arriverait à un événement qui se produirait 7 fois sur 100 000. Il commente ainsi cette probabilité rarissime : *malgré le fait que l'on ne peut pas donner d'explication satisfaisante, il très probable que Mendel n'ait pas voulu montrer les résultats à ses assistants qui connaissaient aussi bien que lui ce qu'il fallait trouver pour que la théorie soit cohérente.* Cet argument est renforcé par le fait que l'on voit des altérations sur les feuilles de mesures de Mendel, afin que les calculs soient en accord avec ses espérances.

Jusqu'à quel point êtes-vous célèbre ?

Les scènes de lit de mort où une mère mourante reste accrochée à la vie juste assez longtemps pour que son fils, absent de longue date, revienne ne sont que trop fréquentes dans les films. Mais est-ce que cela arrive aussi dans la réalité ? Est-il possible d'ajourner sa mort jusqu'à un évènement important ? On a cru que les gens célèbres pouvaient le faire, dans le sens où ils attachaient une grande importance à leur anniversaire. Une étude menée par David P. Philipps (*Statistics : a Guide to the Unknown*, 1978, pp. 71-85) renforce cette idée avec des faits bien précis. Phillips recueillit les mois de naissance et de mort de 1 251 Américains célèbres, les dates des morts étant classées en trois parties : les mois avant, le mois même et les mois après la date de la personne (voir tableau).

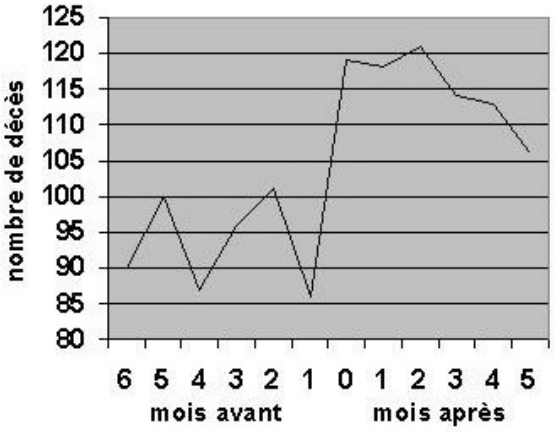
L'échantillon 1 correspond aux personnes très connues énumérées dans Four Hundred Notable Americans et l'échantillon 2 aux personnes remarquables identifiées à partir des trois tomes de *Who was who* pour les années 1951-1960, 1943-1950 et 1897-1947.

Si la date de la mort n'a rien à voir avec celle de la naissance, le nombre attendu de décès pour chacune des trois périodes du tableau devrait être 1 251/12, ce qui fait à peu près 104.

On voit que, pour les deux échantillons, la fréquence observée des décès survenant durant le mois de la naissance est nettement supérieure aux fréquences observées pour les deux autres périodes étudiées ; ces résultats sont illustrés ci-dessous.

Nombre de décès avant, pendant et après le mois de naissance

	Mois Avant						Mois de naissance	Mois Après					Total
	6	5	4	3	2	1		1	2	3	4	5	
Échantillon 1	24	31	20	23	34	16	26	36	37	41	26	34	348
Échantillon 2	66	69	67	73	67	70	93	82	84	73	87	72	903
Ensemble	90	100	87	96	101	86	119	118	121	114	113	106	1 251
1251/12=104	104	104	104	104	104	104	104	104	104	104	104	104	



Or, il est frappant de constater que nombreux sont les individus observés qui meurent au cours du mois de leur naissance, mais il est encore un phénomène plus étonnant : il existe, en effet, une forte diminution de la fréquence des décès précédant les dates de naissance, ce qui pourrait s'expliquer par un ajournement de la mort jusqu'à la date d'anniversaire.

Le rythme journalier

Si l'on vous demande votre taille, vous avez sans doute une réponse toute prête. Quelqu'un vous aura mesuré auparavant et vous l'aura communiquée. Mais vous ne vous êtes peut-être pas demandé pourquoi ce chiffre représente votre taille, et si vous l'aviez fait, on vous aurait

peut-être répondu qu'il a été obtenu en suivant une procédure *ad hoc*. En pratique, cette définition semble satisfaisante. Mais on peut se poser d'autres questions : la caractéristique que nous étudions dépend-elle de l'heure de la mesure ? Par exemple, y a-t-il une différence de taille entre le matin et le soir pour une personne donnée ? Si oui, quelle est l'amplitude de cette différence et quelle en est l'explication physiologique ?

Une enquête statistique simple peut fournir la réponse. Des mesures soigneusement récoltées sur 41 étudiants à Calcutta ont montré une différence de 9,3 mm en moyenne, la mesure du matin étant chaque fois supérieure (C.R. Rao, *Sankhya*, 19, 1957, pp. 96-98). Si, en réalité, la taille d'une personne était constante durant toute la journée, chaque différence de taille observée devrait être attribuée à des erreurs de mesure qui peuvent être soit positives, soit négatives avec une probabilité égale. Dans pareil cas, la probabilité que les 41 différences observées soient positives est de l'ordre de 2-41, ce qui correspond à un évènement qui arrive moins de cinq fois en 1 013 expériences. Ceci révèle que la probabilité que l'hypothèse de la différence de taille en cours de journée soit fausse est extrêmement faible. On peut donc accepter l'idée que nous nous allongeons d'un centimètre pendant que nous dormons et que nous nous rapetissons de la même longueur lorsque nous sommes au travail ! L'explication physiologique plausible est que, pendant la journée, les vertèbres s'écrasent les unes sur les autres, alors qu'une fois couché elles se mettent dans leur position normale.

CONCLUSION

Nous avons vu que la statistique joue un rôle essentiel dans diverses branches de la recherche pure ou appliquée. Son omniprésence a ses racines dans le système scientifique développé par les statisticiens pour la collecte, l'organisation, l'analyse, l'interprétation et la présentation de l'information qui peut être donnée sous forme chiffrée.

Le besoin que la recherche a de travailler avec des méthodes quantitatives a fait de la statistique un outil indispensable, surtout dans les sciences dites *douces* comme l'économie, la psychologie ou la sociologie.

L'importance du développement de l'enseignement et de la recherche dans le domaine des statistiques ne cesse de croître. Pour s'en convaincre, il n'y a qu'à se rappeler qu'en 1992, 2 254 professeurs ordinaires enseignaient cette discipline dans les quelques 289 départements de statistiques que comptent les universités nord-américaines. La même année, les diplômes délivrés par ces départements atteignaient le nombre impressionnant de 1 149 licences, 1 121 postgrades et 392 doctorats.

Aux États-Unis, l'association de statistique dénombre à elle seule 10 000 membres. L'importance des publications en statistique mérite également d'être soulignée, puisqu'au niveau mondial, on dénombre 415 périodiques.

Les avantages des applications statistiques dans l'industrie et dans l'administration sont bien connus. Dans les usines où les méthodes modernes de la statistique ont été utilisées, la production a augmenté de 10 à 100 %, sans recours à des investissements supplémentaires ou à une extension de l'usine. Ainsi, la connaissance statistique est une ressource naturelle. La statistique peut jouer un rôle vital dans les affaires nationales d'un gouvernement. Pour les statisticiens, la connaissance des technologies et celle de leur propre environnement professionnel constitue une exigence indispensable. Ils créent, entre les différentes disciplines, le lien qui permet aux responsables de prendre les décisions nécessaires au progrès. On peut voir que plus un pays est avancé, plus son système de statistique est performant et consulté.

BIBLIOGRAPHIE

Armitage, P. and Colton, T. (1998). *Encyclopedia of Biostatistics*. John Wiley & Sons, New York, 6 volumes.

Dodge, Y. (1993). *Statistique : Dictionnaire encyclopédique.* Dunod, Paris.

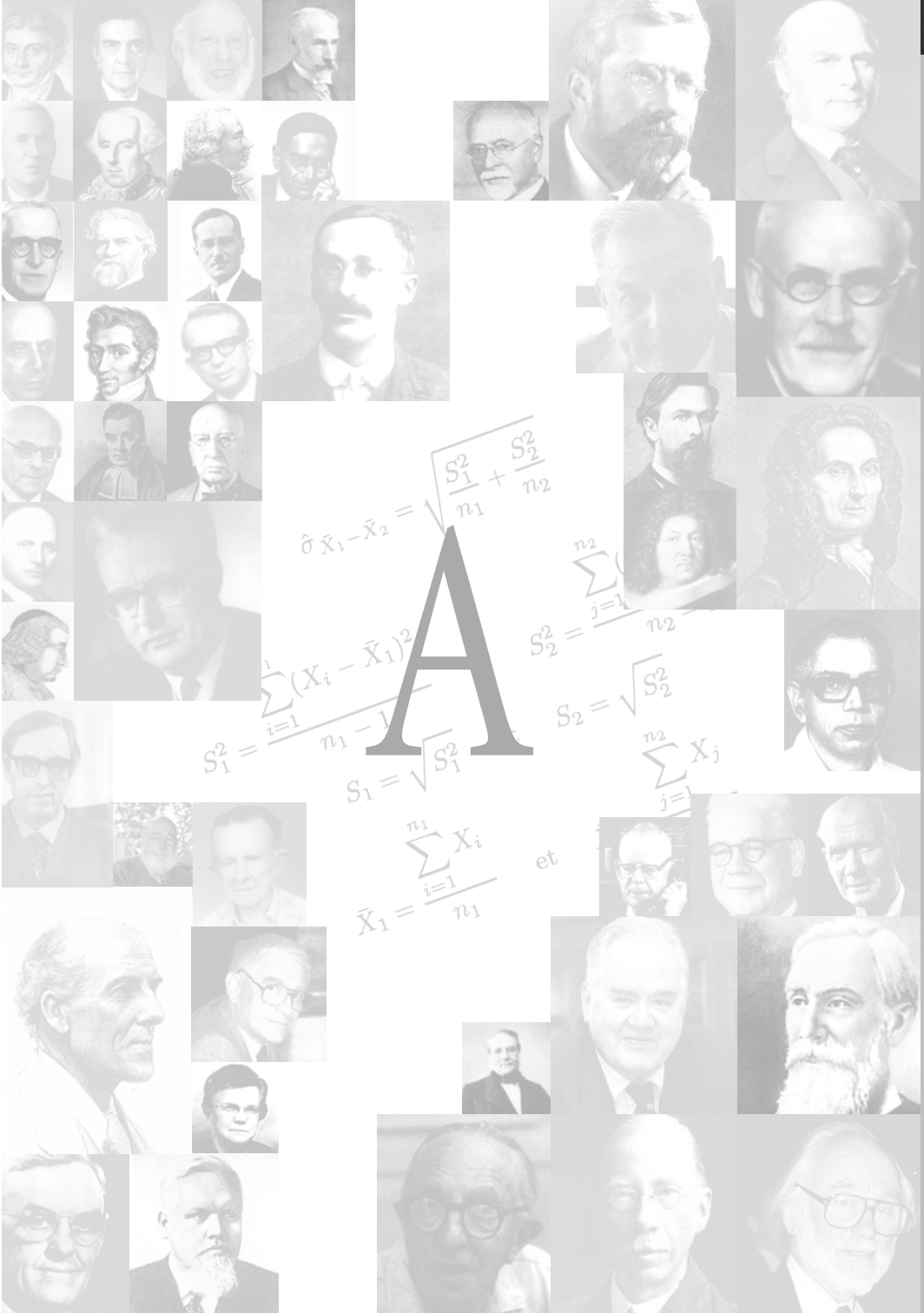
Dodge, Y. (2003). *The Oxford Dictionary of Statistical Terms.* Oxford : Oxford University Press.

El-Shaarawi, A.H., Piegorsch, W.W. (2001). *The Encyclopedia of Environmetrics.* John Wiley & Sons, New York, 4 volumes.

Kendall, M.G. and Bucklard, W.R. (1968). *Dictionary of Statistical Terms.* Longman Scientific & Technical, Essex. Fifth edition, Marriott, F.H.C. (1990).

Kotz, S. and Johnson, N.L. (1982). *Encyclopedia of Statistical Science.* John Wiley & Sons, New York, 10 volumes and update volumes 1 and 2.

Kruskal, W.H. and Tanur, J.M. (1978). *International Encyclopedia of Statistics.* The Free Press, New York.


A
$\hat{\sigma}_{\bar{X}_1-\bar{X}_2} = \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$
$S_1^2 = \frac{\sum_{i=1}^{n_1}(X_i - \bar{X}_1)^2}{n_1 - 1}$
$S_1 = \sqrt{S_1^2}$
$\bar{X}_1 = \frac{\sum_{i=1}^{n_1} X_i}{n_1}$
et
$S_2^2 = \frac{\sum_{j=1}^{n_2}}{n_2}$
$S_2 = \sqrt{S_2^2}$
$\sum_{j=1}^{n_2} X_j$

ADÉQUATION, TEST D'

Goodness of fit test

Voir **test d'adéquation**.

ALGORITHME

Algorithm

Un algorithme est un processus déterminé formé d'une séquence d'étapes bien définies menant à la solution d'un certain type de problèmes.

Ce processus peut être itératif, c'est-à-dire répété plusieurs fois. Il est généralement numérique.

HISTORIQUE

Le terme algorithme vient de la prononciation latine du nom de Abu Ja'far Muhammad ibn Musa Al-Khwarizmi, mathématicien du IX^{e} siècle vivant à Bagdad et précurseur de l'algèbre.

DOMAINES ET LIMITATIONS

Le mot algorithme a pris ces dernières années un autre sens avec l'arrivée des ordinateurs. Il fait alors référence à la description d'un processus dans une forme appropriée à l'exécution d'un programme sur ordinateur.

Le but principal des logiciels statistiques actuels est de dévélopper conjointement les langages de programmation ainsi que le développement des algorithmes statistiques de telle manière que ces algorithmes puissent être présentés sous une forme compréhensible par l'utilisateur. Ceci a pour avantage que l'utilisateur comprend les résultats produits par l'algorithme et qu'il a une plus grande confiance en l'exactitude des solutions. Parmi les revues statistiques traitant des algorithmes, citons notamment le *Journal of Algorithms* édité par l'*Academic Press*, la partie de *Applied Statistics* du *Journal of the Royal Statistical Society-serie C* traitant spécifiquement des algorithmes, *Computational Statistics* édité par *Physica-Verlag* (Heidelberg) et encore *Random Structures and Algorithms* édité par *Wiley* (New York).

EXEMPLE

Nous allons donner ci-dessous un algorithme permettant de calculer la valeur absolue d'un nombre différent de zéro, soit $|x|$.

Processus :

Étape 1. Identifier le signe algébrique du nombre donné.

Étape 2. Si le signe est négatif, passer à l'étape 3. Si le signe est positif, désigner la valeur absolue du nombre donné par le nombre lui-même :

$$|x| = x$$

et arrêter le processus.

Étape 3. Désigner la valeur absolue du nombre donné par l'opposé du nombre :

$$|x| = -x$$

et arrêter le processus.

MOTS CLÉS

Algorithme de Yates *(Yates' algorithm)*
Logiciel statistique *(Statistical software)*

RÉFÉRENCES

Chambers, J.M. (1977). *Computational Methods for Data Analysis.* John Wiley & Sons, New York.

Khwarizmi, Musa ibn Meusba (IX^{e} siècle). Jabr wa-al-muqeabalah. *The algebra of Mohammed ben Musa.* Edited and transl. by Frederic Rosen. Georg Olms Verlag.

Rashed, R. (1985). La naissance de l'algèbre, in *Le Matin des mathématiciens*, Émile Noël éditeur. Belin-Radio France.

ALGORITHME DE YATES

Yates' algorithm

L'**algorithme** de Yates est un processus utilisé pour calculer les **estimateurs** des **effets principaux** et des **interactions** dans une **expérience factorielle**.

HISTORIQUE

L'**algorithme** de Yates a été mis au point par Frank Yates en 1937.

ASPECTS MATHÉMATIQUES

On considère une **expérience factorielle** à n facteurs chacun ayant deux niveaux. On note cette **expérience** 2^n.

Les facteurs sont notés par des lettres majuscules. On emploie une lettre minuscule pour le niveau supérieur du facteur et on omet la lettre pour le niveau inférieur. Lorsque tous les facteurs sont au niveau inférieur, on utilise le symbole (1).

L'**algorithme** se présente sous la forme d'un tableau. Dans la première colonne, on trouve toutes les combinaisons des niveaux des facteurs dans un ordre standard (l'introduction d'une lettre est suivie par les combinaisons avec toutes les combinaisons précédentes). La deuxième colonne donne les observations de toutes les combinaisons.

Pour les n colonnes suivantes, la procédure est la même : la première moitié de la colonne est obtenue en sommant les paires de valeurs successives de la colonne précédente ; la seconde moitié de la colonne est obtenue en soustrayant la première valeur de la deuxième valeur de chaque paire.

La colonne n contient les contrastes des effets principaux et des interactions désignés dans la première colonne. Pour trouver les estimateurs de ces effets, on divise la n-ème colonne par $m \cdot 2^{n-1}$ où m est le nombre de répétitions de l'expérience et n le nombre de facteurs. En ce qui concerne la première ligne, on trouve dans la colonne n la somme de toutes les observations.

DOMAINES ET LIMITATIONS

L'algorithme de Yates peut être appliqué à des **expériences** comprenant des facteurs ayant plus de deux niveaux.

À l'aide de cet **algorithme**, on peut aussi trouver les sommes des carrés correspondant aux effets principaux et interactions de l'**expérience factorielle** 2^n. Elles sont données en élevant la n-ème colonne au carré et en divisant le résultat par $m \cdot 2^n$ (où m est le nombre de répétitions et n le nombre de facteurs).

EXEMPLES

Afin de calculer les estimations des effets principaux et celles des trois interactions, on considère une expérience factorielle 2^3 (c'est-à-dire trois facteurs à deux niveaux) relative à la production agricole de la betterave sucrière. Le premier facteur noté A concerne le type d'engrais (fumier ou non fumier), le second noté B la profondeur du labourage (20 cm ou 30 cm) et le dernier, soit C, la variété de betterave sucrière (variété 1 ou 2). Puisque chaque facteur compte deux niveaux, on peut parler d'une expérience $2 \times 2 \times 2$.

L'interaction entre le **facteur** A et le facteur B, par exemple, est notée par $A \times B$ tandis que l'interaction entre les trois facteurs est désignée par $A \times B \times C$.

Les observations effectuées au nombre de huit indiquent le poids de la récolte, exprimé en kilos, pour un demi-hectare. On a ainsi obtenu :

A		Fumier		Non fumier	
B		20 cm	30 cm	20 cm	30 cm
C	Var. 1	80	96	84	88
	Var. 2	86	96	89	93

En réalisant cette expérience, on cherche à quantifier non seulement l'effet particulier de l'engrais, du labourage et de la variété de betterave (effets simples) mais aussi l'impact sur la récolte des effets combinés (effets d'interaction).

Pour établir le tableau résultant de l'application de l'algorithme de Yates, il faut tout d'abord classer les huit observations selon l'ordre dit de Yates. Pour cela, l'usage du tableau de classement est pratique ; les symboles − et + représentent respectivement les niveaux inférieurs et supérieurs des facteurs. Par exemple, la première ligne indique l'observation "80" correspondant aux niveaux inférieurs pour chacun des trois facteurs (c'est-à-dire A : fumier, B : 20 cm et C : variété 1).

Tableau de classement des observations selon l'ordre de Yates

N° Obs.	Facteurs A	B	C	Observation correspondante
1	−	−	−	80
2	+	−	−	84
3	−	+	−	96
4	+	+	−	88
5	−	−	+	86
6	+	−	+	89
7	−	+	+	96
8	+	+	+	93

On peut désormais construire le tableau de l'algorithme de Yates :

Tableau correspondant à l'algorithme de Yates

Combinaisons	Obs.	Colonnes 1	2	3	Estimations	Effets
(1)	80	164	348	712	89	moyenne
a	84	184	364	−4	−1	A
b	96	175	−4	34	8,5	B
ab	88	189	0	−18	−4,5	$A \times B$
c	86	4	20	16	4	C
ac	89	−8	14	4	1	$A \times C$
bc	96	3	−12	−6	−1,5	$B \times C$
abc	93	−3	−6	6	1,5	$A \times B \times C$

On trouve les termes de la colonne numérotée 2, par exemple, en faisant :

- pour la première moitié de cette colonne (somme des éléments de chacune des paires de la colonne précédente soit 1) :

$$\begin{aligned} 348 &= 164 + 184 \\ 364 &= 175 + 189 \end{aligned}$$

$$\begin{aligned} -4 &= 4 + (-8) \\ 0 &= 3 + (-3) \end{aligned}$$

- pour la deuxième moitié (différence des éléments de chacune des paires, à l'inverse de précédemment) :

$$\begin{aligned} 20 &= 184 - 164 \\ 14 &= 189 - 175 \\ -12 &= -8 - 4 \\ -6 &= -3 - 3 \end{aligned}$$

L'estimation pour l'effet principal du facteur A est égale à -1. En effet, on a divisé la valeur -4 (colonne 3) par $m \cdot 2^{n-1} = 1 \cdot 2^{3-1} = 4$ où m vaut 1 puisque l'expérience n'a été réalisée qu'une seule fois (sans répétition) et n correspond au nombre de facteurs, soit 3 dans notre cas.

MOTS CLÉS

Analyse de variance *(Analysis of variance)*
Contraste *(Contrast)*
Effet principal *(Main effect)*
Expérience factorielle *(Factorial experiment)*
Interaction *(Interaction)*
Plan d'expérience *(Design of experiments)*

RÉFÉRENCE

Yates, F. (1937). The design and analysis of factorial experiments. *Technical Communication of the Commonwealth Bureau of Soils*, 35, Commonwealth Agricultural Bureaux, Farnham Royal.

ANALYSE COMBINATOIRE

Combinatory analysis

D'une façon générale, l'analyse combinatoire se réfère aux techniques permettant de déterminer le nombre d'éléments d'un ensemble particulier sans les compter individuellement.

Les éléments pourraient être les résultats d'une **expérience** scientifique ou les issues différentes d'un **événement** quelconque.

On peut souligner en particulier trois notions liées à l'analyse combinatoire :

— permutations ;

— combinaisons ;

— arrangements.

HISTORIQUE

Les problèmes d'analyse combinatoire ont depuis longtemps intéressé les mathématiciens. En effet, selon Lajos Takács (1982), ces questions remontent à la Grèce antique. Toutefois, les Hindous, les Perses avec Omar Khayyâm, poète et mathématicien, et surtout les Chinois y consacrèrent également leurs travaux. Un livre chinois vieux de 3 000 ans, *I Ching* donna les arrangements possibles d'un ensemble de n éléments, avec $n \leq 6$. En 1303, Shih-chieh Chu publia un ouvrage intitulé *Ssu Yuan Yü Chien* (Précieux miroir des quatre éléments) dont la couverture présente, arrangée en triangle, une **combinaison** de k éléments pris d'un ensemble de taille n, avec $0 \leq k \leq n$. Le problème du **triangle arithmétique** fut traité par plusieurs mathématiciens européens, dont Michael Stifel, Nicolo Fontana Tartaglia et Pierre Hérigone, mais surtout Blaise Pascal qui rédigea en 1654 un *Traité du triangle arithmétique* qui ne sera publié qu'en 1665.

Si l'on peut citer un autre écrit historique sur les combinaisons, publié en 1617, par Erycius Puteanus et intitulé *Erycii Puteani Pretatis Thaumata in Bernardi Bauhusii è Societate Jesu Proteum Parthenium*, ce ne fut qu'avec les travaux de Pierre de Fermat (1601-1665) et de Blaise Pascal (1623-1662) que l'analyse combinatoire prit véritablement toute son importance. En revanche, le terme « analyse combinatoire » fut introduit par Gottfried Wilhelm von Leibniz (1646-1716) en 1666. Dans son ouvrage *Dissertatio de Arte Combinatoria*, il étudia systématiquement les problèmes d'arrangements, de permutations et de combinaisons.

D'autres travaux méritent également d'être cités, comme ceux de John Wallis (1616-1703), rapportés dans *The Doctrine of Permutations and Combinations, being an essential and fundamental part of the Doctrines of Chances*, ou ceux de **Jakob Bernoulli**, d'**Abraham de Moivre**, de Girolamo Cardano (1501-1576), et de Galileo Galilei (Galilée) (1564-1642).

Dans la seconde moitié du XIX^{e} siècle, Arthur Cayley (1829-1895) résolut certains problèmes de cette analyse en utilisant des graphes qu'il désigna sous le nom d'« arbres » ; enfin on ne saurait évoquer ce sujet sans mentionner un ouvrage important, celui de Percy Alexander MacMahon (1854-1929), paru sous le titre *Combinatory Analysis* (1915-1916).

EXEMPLES

Voir **arrangement**, **combinaison** et **permutation.**

MOTS CLÉS

Arrangement *(Arrangement)*
Binôme *(Binomial)*
Combinaison *(Combination)*
Loi binomiale *(Binomial distribution)*
Permutation *(Permutation)*
Triangle arithmétique *(Arithmetic triangle)*

RÉFÉRENCES

MacMahon, P.A. (1915-1916). *Combinatory Analysis.* Vol. I et II. Cambridge University Press, Cambridge.

Stigler, S. (1986). *The History of Statistics, the Measurement of Uncertainty before 1900.* The Belknap Press of Harvard University Press. London, England.

Takács, L. (1982). Combinatorics. In Kotz S. and Johnson N.L. (editors). *Encyclopedia of Statistical Sciences.* Vol 2. John Wiley & Sons, New York.

ANALYSE DE COVARIANCE

Covariance analysis

L'analyse de covariance est une technique d'**estimation** et de test des effets des traitements. Elle permet

de déterminer s'il existe une différence significative entre plusieurs moyennes de traitements en tenant compte des valeurs observées de la **variable** avant le traitement.

L'**analyse de covariance** accroît la précision des comparaisons de traitement puisqu'elle implique l'ajustement de la variable de réponse Y à une variable concomitante X qui correspond aux valeurs observées avant le traitement.

HISTORIQUE

L'analyse de covariance date des années 1930. C'est **Ronald Aylmer Fisher** (1935) qui a développé cette méthode pour la première fois.

Par la suite, d'autres auteurs appliqueront l'analyse de covariance à des problèmes de type agricole ou médical. Nous citerons par exemple Maurice Stephenson Bartlett (1937) qui reprend des études qu'il avait faites sur la culture du coton en Égypte et sur le rendement laitier des vaches en hiver et auxquelles il applique l'analyse de covariance.

C'est Daniel Bertrand DeLury (1948) qui effectue une analyse de covariance pour comparer les effets de différents médicaments (atropine, quinidine, atrophine) sur les muscles des rats.

ASPECTS MATHÉMATIQUES

Considérons une analyse de covariance pour un **plan complètement randomisé** impliquant un seul **facteur**.

Le **modèle** linéaire que nous devons considérer est le suivant :

$$Y_{ij} = \mu + \tau_i + \beta X_{ij} + \varepsilon_{ij}, \qquad \begin{matrix} i = 1, 2, \ldots, t \\ j = 1, 2, \ldots, n_i \end{matrix}$$

où

- Y_{ij} représente la j-ème **observation** recevant le **traitement** i,
- μ la **moyenne** générale commune à tous les traitements,
- τ_i l'effet réel sur l'observation du traitement i,
- X_{ij} la **valeur** de la variable concomitante et
- ε_{ij} l'**erreur** expérimentale de l'observation Y_{ij}.

Calcul des sommes

Notons $\bar{X}_{i.}$ et $\bar{Y}_{i.}$ respectivement les moyennes des valeurs des X et des Y pour le traitement i et $\bar{X}_{..}$ et $\bar{Y}_{..}$ respectivement les moyennes de toutes les valeurs de X et de Y. Pour pouvoir effectuer le **test de Fisher** et calculer le ratio F qui va permettre de déterminer s'il existe une différence significative entre les traitements, nous avons besoin des sommes des carrés et des sommes des produits suivantes :

1. La somme des carrés totale pour X :

$$S_{XX} = \sum_{i=1}^{t} \sum_{j=1}^{n_i} \left(X_{ij} - \bar{X}_{..}\right)^2,$$

2. La somme des carrés totale pour Y (S_{YY}) :
elle se calcule de la même manière que S_{XX} en remplaçant les X par des Y,

3. La somme des produits totale de X et Y :

$$S_{XY} = \sum_{i=1}^{t} \sum_{j=1}^{n_i} \left(X_{ij} - \bar{X}_{..}\right)\left(Y_{ij} - \bar{Y}_{..}\right),$$

4. La somme des carrés des traitements pour X :

$$T_{XX} = \sum_{i=1}^{t} \sum_{j=1}^{n_i} \left(\bar{X}_{i.} - \bar{X}_{..}\right)^2,$$

5. La somme des carrés des traitements pour Y (T_{YY}) :
elle se calcule de la même manière que T_{XX} en remplaçant les X par des Y,

6. La somme des produits des traitements de X et Y :

$$T_{XY} = \sum_{i=1}^{t} \sum_{j=1}^{n_i} \left(\bar{X}_{i.} - \bar{X}_{..}\right)\left(\bar{Y}_{i.} - \bar{Y}_{..}\right),$$

7. La somme des carrés des erreurs pour X :

$$E_{XX} = \sum_{i=1}^{t} \sum_{j=1}^{n_i} \left(X_{ij} - \bar{X}_{i.}\right)^2,$$

8. La somme des carrés des erreurs pour Y : elle se calcule de la même manière que E_{XX} en remplaçant les X par des Y,

9. La somme des produits des erreurs X et Y :

$$E_{XY} = \sum_{i=1}^{t} \sum_{j=1}^{n_i} \left(X_{ij} - \bar{X}_{i.}\right)\left(Y_{ij} - \bar{Y}_{i.}\right).$$

Le calcul de ces sommes correspond à une **analyse de variance** effectuée pour X, pour Y et pour XY.
Les degrés de liberté associés à ces différentes sommes sont les suivants :

1. Pour les sommes totales : $\sum_{i=1}^{t} n_i - 1$,
2. Pour les sommes des traitements : $t - 1$,
3. Pour les sommes des erreurs : $\sum_{i=1}^{t} n_i - t$.

L'ajustement de la **variable** Y à la variable concomitante X donne deux nouvelles sommes de carrés :

1. La somme des carrés totale ajustée :

$$SC_{tot} = S_{YY} - \frac{S_{XY}^2}{S_{XX}},$$

2. La somme des carrés des erreurs ajustée :

$$SC_{err} = E_{YY} - \frac{E_{XY}^2}{E_{XX}}.$$

Les nouveaux degrés de liberté pour ces deux sommes sont :

1. $\sum_{i=1}^{t} n_i - 2$
2. $\sum_{i=1}^{t} n_i - t - 1$ a degré de liberté ayant été soustrait pour l'ajustement.

La troisième somme des carrés ajustée, à savoir la somme des carrés des traitements ajustée est donnée par :

$$SC_{tr} = SC_{tot} - SC_{err}.$$

Elle garde le même nombre de degrés de liberté $t - 1$.

Tableau d'analyse de covariance

Nous avons maintenant tous les éléments pour construire le tableau d'analyse de covariance, les moyennes des carrés étant obtenues par la division des sommes des carrés par les degrés de liberté.

Source de variation	Degrés de liberté	Somme des carrés et des produits $\sum_{i=1}^{t} x_i^2$	$\sum_{i=1}^{t} x_i y_i$	$\sum_{i=1}^{t} y_i^2$
Traitements	t - 1	T_{XX}	T_{XY}	T_{YY}
Erreurs	$\sum_{i=1}^{t} n_i - t$	E_{XX}	E_{XY}	E_{YY}
Total	$\sum_{i=1}^{t} n_i - 1$	S_{XX}	S_{XY}	S_{YY}

Remarque : les nombres dans les colonnes des $\sum_{i=1}^{t} x_i^2$ et des $\sum_{i=1}^{t} y_i^2$ ne peuvent pas être négatifs ; en revanche, les nombres dans la colonne des $\sum_{i=1}^{t} x_i y_i$ peuvent être négatifs.

Ajustement

Source de variation	Degrés de liberté	Somme des carrés	Moyenne des carrés
Traitements	t - 1	SC_{tr}	MC_{tr}
Erreurs	$\sum_{i=1}^{t} n_i - t - 1$	SC_{err}	MC_{err}
Total	$\sum_{i=1}^{t} n_i - 2$	SC_{tot}	

Ratio F Test concernant les traitements

Le ratio F (qui va servir à tester l'**hypothèse nulle** qu'il n'y a pas de différence significative entre les moyennes des traitements une fois la variable Y ajustée) est donné par :

$$F = \frac{MC_{tr}}{MC_{err}}.$$

Le ratio suit une **loi de Fisher** avec $t - 1$ et $\sum_{i=1}^{t} n_i - t - 1$ degrés de liberté.

L'hypothèse nulle :

$$H_0 \quad : \quad \tau_1 = \tau_2 = \ldots = \tau_t$$

sera alors rejetée au **seuil de signification** α si le ratio F est supérieur ou égal à la valeur de la **table de Fisher**, c'est-à-dire si :

$$F \geq F_{t-1,\sum_{i=1}^{t} n_i - t - 1,\alpha}.$$

Il est clair que, pour faire une analyse de covariance, le coefficient β est supposé différent de zéro. Si tel n'était pas le cas, une simple analyse de variance suffirait.

Test concernant la pente β

Nous désirons ainsi savoir s'il y avait une différence significative entre les variables concomitantes avant l'application du traitement.

Pour tester cette **hypothèse**, nous formulons l'hypothèse nulle :

$$H_0 \quad : \quad \beta = 0$$

et l'**hypothèse alternative** :

$$H_1 \quad : \quad \beta \neq 0.$$

Alors le ratio F :

$$F = \frac{E_{XY}^2 / E_{XX}}{MC_{err}}$$

suit une loi de Fisher avec 1 et $\sum_{i=1}^{t} n_i - t - 1$ degrés de liberté. L'hypothèse nulle sera rejetée au seuil de signification α si le ratio F est supérieur ou égal à la valeur de la table de Fisher, c'est-à-dire si :

$$F \geq F_{1,\sum_{i=1}^{t} n_i - t - 1,\alpha}.$$

DOMAINES ET LIMITATIONS

Les hypothèses de base qu'il faut respecter avant de faire une analyse de covariance sont les mêmes que pour l'**analyse de variance** et l'**analyse de régression**. Ce sont des hypothèses de normalité, d'homogénéité (homoscédasticité) de variances et d'indépendance des erreurs du modèle.

En analyse de covariance, comme en analyse de variance, l'**hypothèse nulle** stipule que les échantillons indépendants proviennent de différentes populations dont les moyennes sont identiques.

D'autre part, comme à toute technique statistique, un certain nombre de conditions d'application est associé à l'analyse de covariance :

1. Les distributions des populations doivent être approximativement normales, sinon normales.
2. Les populations d'où sont prélevés les échantillons doivent posséder la même variance σ^2, c'est-à-dire :
   $$\sigma_1^2 = \sigma_2^2 = \ldots = \sigma_k^2$$
   où k est le nombre de populations à comparer.
3. Les échantillons doivent être choisis aléatoirement et tous les échantillons doivent être indépendants.

   À cela s'ajoute une hypothèse de base spécifique à l'analyse de covariance, à savoir que les traitements effectués ne doivent pas influencer les valeurs de la **variable** concomitante X.

EXEMPLE

Considérons l'**expérience** qui consiste à comparer les effets de trois méthodes de nutrition sur une **population** de porcs.

Les données se présentent sous la forme d'un tableau comprenant trois méthodes de nutrition, chacune soumise à 5 porcs. Les poids initiaux sont notés par la variable concomitante X (en kg), et les gains de poids (après **traitement**) sont notés par Y :

Méthodes de nutrition

1		2		3	
X	Y	X	Y	X	Y
32	167	26	182	36	158
29	172	33	171	34	191
22	132	22	173	37	140
23	158	28	163	37	192
35	169	22	182	32	162

Calculons les moyennes $\bar{X}_{i.}, \bar{Y}_{i.}, \bar{X}_{..}$ et $\bar{Y}_{..}$:

$$\begin{aligned}
\bar{X}_{1.} &= \frac{32+29+22+23+35}{5} = 28,2 \\
\bar{X}_{2.} &= \frac{26+33+22+28+22}{5} = 26,2 \\
\bar{X}_{3.} &= \frac{36+34+37+37+32}{5} = 35,2 \\
\bar{X}_{..} &= \frac{1}{15}(32+...+35+26+...+22 \\
&\quad +36+...+32) \\
&= 29,87 \\
\bar{Y}_{1.} &= \frac{167+172+132+158+169}{5} = 159,6 \\
\bar{Y}_{2.} &= \frac{182+171+173+163+182}{5} = 174,2 \\
\bar{Y}_{3.} &= \frac{158+191+140+192+162}{5} = 168,6 \\
\bar{Y}_{..} &= \frac{1}{15}(167+...+169+182+...+182 \\
&\quad +158+...+162) \\
&= 167,47.
\end{aligned}$$

Nous calculons les différentes sommes des carrés et des produits :

1. La somme des carrés totale pour X :

$$\begin{aligned}
S_{XX} &= \sum_{i=1}^{3}\sum_{j=1}^{5}\left(X_{ij}-\bar{X}_{..}\right)^2 \\
&= (32-29,87)^2+\ldots \\
&\quad +(32-29,87)^2 \\
&= 453,73.
\end{aligned}$$

2. La somme des carrés totale pour Y :

$$\begin{aligned}
S_{YY} &= \sum_{i=1}^{3}\sum_{j=1}^{5}\left(Y_{ij}-\bar{Y}_{..}\right)^2 \\
&= (167-167,47)^2+\ldots \\
&\quad +(162-167,47)^2 \\
&= 3\,885,73.
\end{aligned}$$

3. La somme des produits totale de X et Y :

$$\begin{aligned}
S_{XY} &= \sum_{i=1}^{3}\sum_{j=1}^{5}\left(X_{ij}-\bar{X}_{..}\right)\left(Y_{ij}-\bar{Y}_{..}\right) \\
&= (32-29,87)\,(167-167,47)+ \\
&\quad \ldots+(32-29,87)\,(162-167,47) \\
&= 158,93.
\end{aligned}$$

4. La somme des carrés des traitements pour X :

$$\begin{aligned}
T_{XX} &= \sum_{i=1}^{3}\sum_{j=1}^{5}\left(\bar{X}_{i.}-\bar{X}_{..}\right)^2 \\
&= 5\,(28,2-29,87)^2 \\
&\quad +5\,(26,2-29,87)^2 \\
&\quad +5\,(35,2-29,87)^2 \\
&= 223,33.
\end{aligned}$$

5. La somme des carrés des traitements pour Y :

$$\begin{aligned}
T_{YY} &= \sum_{i=1}^{3}\sum_{j=1}^{5}\left(\bar{Y}_{i.}-\bar{Y}_{..}\right)^2 \\
&= 5\,(159,6-167,47)^2 \\
&\quad +5\,(174,2-167,47)^2 \\
&\quad +5\,(168,6-167,47)^2 \\
&= 542,53.
\end{aligned}$$

6. La somme des produits des traitements de X et Y :

$$\begin{aligned}
T_{XY} &= \sum_{i=1}^{3}\sum_{j=1}^{5}\left(\bar{X}_{i.}-\bar{X}_{..}\right)\left(\bar{Y}_{i.}-\bar{Y}_{..}\right) \\
&= 5\,(28,2-29,87)\,(159,6-167,47) \\
&\quad +\ldots+5\,(35,2-29,87)\,(168,6-167,47) \\
&= -27,67.
\end{aligned}$$

7. La somme des carrés des erreurs pour X :

$$\begin{aligned}
E_{XX} &= \sum_{i=1}^{3}\sum_{j=1}^{5}\left(X_{ij}-\bar{X}_{i.}\right)^2 \\
&= (32-28,2)^2+\ldots+(32-35,2)^2 \\
&= 230,40.
\end{aligned}$$

8. La somme des carrés des erreurs pour Y :

$$E_{YY} = \sum_{i=1}^{3}\sum_{j=1}^{5}\left(Y_{ij}-\bar{Y}_{i.}\right)^2$$

$$= (167 - 159,6)^2 + \ldots + (162 - 168,6)^2$$
$$= 3\,343,20.$$

9. La somme des produits des erreurs X et Y :

$$\begin{aligned} E_{XY} &= \sum_{i=1}^{3}\sum_{j=1}^{5}\left(X_{ij} - \bar{X}_{i.}\right)\left(Y_{ij} - \bar{Y}_{i.}\right) \\ &= (32 - 28,2)\,(167 - 159,6) + \ldots \\ &\quad + (32 - 35,2)\,(162 - 168,6) \\ &= 186,60. \end{aligned}$$

Les degrés de liberté associés à ces différentes sommes sont les suivants :

1. Pour les sommes totales :

$$\sum_{i=1}^{3} n_i - 1 = 15 - 1 = 14.$$

2. Pour les sommes des traitements :

$$t - 1 = 3 - 1 = 2.$$

3. Pour les sommes des erreurs :

$$\sum_{i=1}^{3} n_i - t = 15 - 3 = 12.$$

L'ajustement de la **variable** Y à la variable concomitante X donne deux nouvelles sommes des carrés :

1. La somme des carrés totale ajustée :

$$\begin{aligned} SC_{tot} &= S_{YY} - \frac{S_{XY}^2}{S_{XX}} \\ &= 3\,885,73 - \frac{158,93^2}{453,73} \\ &= 3\,830,06. \end{aligned}$$

2. La somme des carrés des erreurs ajustée :

$$\begin{aligned} SC_{err} &= E_{YY} - \frac{E_{XY}^2}{E_{XX}} \\ &= 3\,343,20 - \frac{186,60^2}{230,40} \\ &= 3\,192,07. \end{aligned}$$

Les nouveaux degrés de liberté pour ces deux sommes sont :

1. $\sum_{i=1}^{3} n_i - 2 = 15 - 2 = 13$
2. $\sum_{i=1}^{3} n_i - t - 1 = 15 - 3 - 1 = 11$

La somme des carrés des traitements ajustée est donnée par :

$$\begin{aligned} SC_{tr} &= SC_{tot} - SC_{err} \\ &= 3\,830,06 - 3\,192,07 \\ &= 637,99. \end{aligned}$$

Elle garde le même nombre de degrés de liberté :

$$t - 1 = 3 - 1 = 2.$$

Nous avons maintenant tous les éléments pour construire le tableau d'analyse de covariance, les moyennes des carrés étant obtenues par la division des sommes des carrés par les degrés de liberté.

Source de variation	Degrés de liberté	Somme des carrés et des produits $\sum_{i=1}^{3} x_i^2$	$\sum_{i=1}^{3} x_i y_i$	$\sum_{i=1}^{3} y_i^2$
Traitements	2	223,33	−27,67	543,53
Erreurs	12	230,40	186,60	3 343,20
Total	14	453,73	158,93	3 885,73

Ajustement

Source de variation	Degrés de liberté	Somme des carrés	Moyenne des carrés
Traitements	2	637,99	318,995
Erreurs	11	3 192,07	290,188
Total	13	3 830,06	

Le ratio F (qui va servir à tester l'**hypothèse nulle** qu'il n'y a pas de différence significative entre les moyennes des traitements une fois la variable Y ajustée) est donné par :

$$F = \frac{MC_{tr}}{MC_{err}} = \frac{318,995}{290,188} = 1,099.$$

Si nous choisissons un **seuil de signification** $\alpha = 0,05$, la valeur de F dans la **table de Fisher** est égale à :

$$F_{2,11,0,05} = 3,98.$$

Comme $F < F_{2,11,0,05}$, nous ne pouvons pas rejeter l'hypothèse nulle et concluons qu'il n'y a pas de différence significative entre les trois méthodes de nutrition, une fois la variable Y ajustée, aux poids initiaux X.

MOTS CLÉS

Analyse de régression *(Regression analysis)*
Analyse de variance *(Analysis of variance)*
Analyse de données manquantes *(Analysis of missing data)*
Plan d'expériences *(Design of experiments)*

RÉFÉRENCES

Bartlett, M.S. (1937). Some Examples of Statistical Methods of Research in Agriculture and Applied Biology. *Journal of the Royal Statistical Society*, (Suppl.), 4, pp. 137-183.

DeLury, D.B. (1948). The Analysis of Covariance. *Biometrics*, 4, pp. 153-170.

Fisher, R.A. (1935). *The Design of Experiments.* Oliver & Boyd, Edinburgh.

Huitema, B.E. (1980). *The analysis of covariance and alternatives.* John Wiley & Sons, New York.

Wildt, A.R. and Ahtola, O. (1978). *Analysis of Covariance.* Sage University Papers series on Quantitative Applications in the Social Sciences (12).

ANALYSE DE DONNÉES

Data analysis

De nombreuses activités scientifiques commencent par un recueil de données, qu'on peut généralement classer dans un tableau à double entrée de grande taille. L'analyse de données vise à permettre à l'utilisateur de ces tableaux d'extraire facilement le maximum d'informations qui lui sont nécessaires.

On peut considérer l'analyse de données au sens large, comme l'essence même des statistiques à laquelle tous les autres aspects sont subordonnés.

HISTORIQUE

Dans la mesure où l'analyse de données regroupe des méthodes très nombreuses et très différentes d'analyse statistique, il est difficile de rapporter un historique précis. On peut cependant mentionner les précurseurs de ces différents aspects :

- la première publication d'analyse de données exploratoire date de 1970-1971, elle est due à **John Wilder Tukey** ; il s'agit d'une première version de son ouvrage paru en 1977 ;
- les principes théoriques de l'**analyse factorielle des correspondances** sont dus à Hermann Otto Hartley (1935) (publiés sous son nom original allemand Hirschfeld) et à **Ronald Aylmer Fisher** (1940). Cependant elle ne fut développée que dans les années 1970 par Jean-Paul Benzécri ;
- les premiers travaux de **classification** furent effectués en biologie et en zoologie. La plus ancienne typologie est due à Galen (129-199), et précède de loin les recherches de Linné (XVIII^{e} siècle). Les méthodes de classifications numériques sont basées sur les idées d'Adanson (XVIII^{e} siècle) et ont été développées entre autres par Joseph Zubin (1938) et Robert L. Thorndike (1953).

DOMAINES ET LIMITATIONS

L'analyse de données regroupe des méthodes très nombreuses et très différentes d'analyse statistique.

L'analyse de données peut se décomposer de la façon suivante :

1. *L'analyse exploratoire de données*, qui consiste comme son nom l'indique à explorer les données, ce qui se résume à :
    - la **représentation graphique** des données ;

- la **transformation**, si nécessaire, des données ;
- la détection d'éventuelles observations aberrantes ;
- l'élaboration d'hypothèses de recherches imprévues au début de l'**expérience** ;
- l'**estimation robuste**.

2. *L'analyse de données initiale* traite :
   - du choix des méthodes statistiques à appliquer aux données.
3. *L'analyse multivariée* des données comprend :
   - le déploiement d'espaces multidimensionnels,
   - la transformation des données pour réduire les dimensions et faciliter l'interprétation ;
   - la recherche de structure.
4. *L'analyse de données* (lorsqu'on parle d'un aspect particulier de cette dernière) inclut :
   - l'**analyse factorielle des correspondances** ;
   - l'analyse des correspondances multiples ;
   - la **classification**.
5. *L'analyse des données confirmatoire* comprend :
   - l'**estimation** de paramètres ;
   - les tests d'hypothèses ;
   - la généralisation et la conclusion.

L'analyse de données est caractérisée par quelques idées de base, à savoir :

- utiliser de façon massive et efficace l'informatique ; (visualisation des résultats avec des moyens d'interprétation clairs et précis) ;
- rester proche des données ;
- réduire au maximum la subjectivité.

MOTS CLÉS

Analyse exploratoire de données *(Exploratory data analysis)*
Analyse factorielle des correspondances *(Correspondence analysis)*
Classification *(Classification)*
Classification automatique *(Cluster analysis)*
Donnée *(Data)*
Statistique *(Statistics)*
Test d'hypothèse *(Hypothesis testing)*
Transformation *(Transformation)*

RÉFÉRENCES

Benzécri, J.-P. (1976). *L'Analyse des données*, Tome 1 : *La Taxinomie* ; Tome 2 : *L'Analyse factorielle des correspondances.* Dunod, Paris.

Benzécri, J.-P. (1976). Histoire et préhistoire de l'analyse des données, *Les Cahiers de l'analyse des données 1*, n$^{os.}$ 1–4. Dunod, Paris.

Everitt, B.S. (1974). *Cluster Analysis.* Halstead Press, London.

Fisher, R.A. (1940). The precision of discriminant Functions. *Annals of Eugenics* (London), 10, pp. 422-429.

Gower, J.C. (1988). *Classification, Geometry and Data Analysis in Classification and Related Methods of Data Analysis*, Bock H.H. (editor). Elsevier Science Publishers B.V. (North-Holland).

Hirschfeld, H.O. (1935). A connection between correlation and contingency, *Proceedings of the Cambridge Philosophical Society*, 31, pp. 520-524.

Thorndike, R.L. (1953). Who belongs in a family ? *Psychometrika*, 18, pp. 267-276.

Tukey, J.W. (1970-1971). *Exploratory data analysis*, limited preliminary edition. Addison-Wesley, Reading, Mass.

Tukey, J.W. (1977). *Exploratory data analysis.* Addison-Wesley, Reading, Mass.

Zubin, J. (1938). A Technique for Measuring Like-mindedness. *Journal of Abnormal and Social Psychology*, 33, pp. 508-516.

ANALYSE DE DONNÉES BINAIRES

Analysis of binary data

L'analyse des **données binaires** suit normalement une des méthodes suivantes :

(a) l'étude d'adéquation (**test d'adéquation**) d'une distribution théorique à un ensemble de données empiriques, de nature catégorique ;

(b) la génération d'une **table de contingence** et son analyse, qui consiste à découvrir et à étudier les relations entre les attributs, si elles existent ;

(c) le **test d'homogénéité** de plusieurs populations par rapport à la distribution d'une variable qualitative binaire ;

(d) l'examen de l'hypothèse d'**indépendance**.

HISTORIQUE

Voir **analyse de données**.

ASPECTS MATHÉMATIQUES

Soient Y une **variable aléatoire** binaire et des variables aléatoires supplémentaires X_1, X_2, ..., X_n. Alors la **dépendance** de Y des variables X_1, X_2, ..., X_n est représentée à l'aide d'un des modèles suivants (dont les coefficients sont estimés à l'aide du **maximum de vraisemblance**) :

1. *modèle linéaire :* $P(Y=1)$ est exprimé comme fonction linéaire (dans les paramètres) des $X_{i.}$;
2. *modèle log-linéaire :* $\log P(Y=1)$ est exprimé comme fonction linéaire (dans les paramètres) des $X_{i.}$;
3. *modèle logistique :* $\log\left(\frac{P(Y=1)}{P(Y=0)}\right)$ est exprimé comme fonction linéaire (dans les paramètres) des $X_{i.}$

Les modèles 1 et 2 sont plus faciles à interpréter, la quantité à expliquer prend toutes les valeurs possibles du modèle linéaire. Il est néanmoins important de prêter attention à l'extrapolation en dehors du domaine dans lequel le modèle a été ajusté.

Il peut s'avérer que parmi les variables indépendantes $(X_1, X_2, \ldots, X_n)$, il y ait des variables catégoriques (éventuellement binaires). Dans ce cas, il est nécessaire de traiter les variables catégoriques non binaires de la façon suivante : Soit Z une variable aléatoire à m catégories, on numérote les catégories de 1 à m et on définit $m-1$ variables aléatoires $Z_1, Z_2, \ldots, Z_{m-1}$ prenant la valeur 1 dans le cas d'appartenance à la catégorie representée par l'indice. La variable Z sera remplacée par ces $m-1$ variables, dont les coefficients expriment l'influence de la catégorie considérée. La catégorie non représentée (pour éviter une situation de **colinéarité**) aura (pour la comparaison avec les autres catégories) un paramètre zéro.

MOT CLÉ

Analyse de données *(Data analysis)*

RÉFÉRENCE

Cox, D.R. (1970). *The analysis of binary data.* Methuen, London.

ANALYSE DE DONNÉES CATÉGORIQUES

Analysis of categorical data

L'analyse des **données catégoriques** suit normalement une des méthodes suivantes :

(a) l'étude d'adéquation (**test d'adéquation**) d'une distribution théorique à un ensemble de données empiriques, de nature catégorique ;

(b) la génération d'une **table de contingence** et son analyse, qui consiste à découvrir et à étudier les relations entre les attributs, si elles existent ;

(c) le **test d'homogénéité** de plusieurs populations par rapport à la distribution d'une **variable qualitative catégorielle** ;

(d) l'examen de l'hypothèse d'**indépendance**.

HISTORIQUE

Le terme de « contingence », utilisé en rapport avec des tableaux croisés de **données catégoriques** est vraisemblablement dû à **Karl Pearson** (1904). Quant au **test du chi-carré**, testant l'homogénéité, c'est M.S. Bartlett qui le proposa en 1937.

ASPECTS MATHÉMATIQUES

Voir **test d'adéquation** et **table de contingence**.

MOTS CLÉS

Analyse de données *(Data analysis)*
Analyse factorielle des correspondances *(Correspondence analysis)*
Donnée *(Data)*
Donnée catégorique *(Categorical data)*
Table de contingence *(Contingency table)*
Test d'adéquation *(Goodness of fit test)*
Test d'adéquation du chi-carré *(Chi-square goodness of fit test)*
Test d'homogénéité *(Homogeneity test)*
Test d'indépendence *(Test of independance)*

RÉFÉRENCES

Agresti, A. (1990). *Categorical Data Analysis.* John Wiley & Sons, New York.

Bartlett, M.S. (1937). Properties of sufficiency and statistical tests. *Proceedings of the Royal Society of London.* Series A, 160, pp. 268-282.

Cox, D.R. and Snell, E.J. (1989). *Analysis of binary data.* 2nd ed. Chapman & Hall, London.

Haberman, S.J. (1978). *Analysis of Qualitative Data*, Vol. I : Introductory Topics. Academic Press, New York.

Pearson, K. (1904). On the theory of contingency and its relation to association and normal correlation. *Drapers' Company Research Memoirs, Biometric Ser.* I. pp. 1-35.

ANALYSE DE DONNÉES MANQUANTES

Missing data analysis

Voir **observation manquante**.

ANALYSE DE RÉGRESSION

Regression analysis

L'analyse de régression est un outil qui permet d'étudier et de mesurer la **relation** existant entre deux ou plusieurs variables.

En se basant sur les données d'un **échantillon**, l'analyse de régression cherche à déterminer une **estimation** d'une relation mathématique entre deux variables (ou plus). Le but est d'estimer les valeurs d'une des variables à l'aide des valeurs de l'autre (ou des autres).

La variable estimée est appelée **variable dépendante**, symbolisée généralement par Y. Par contre, la ou les variables qui expliquent les variations de Y sont appelées variables indépendantes. Elles sont symbolisées par X.

Selon que la variable dépendante dépend d'une ou de plusieurs variables indépendantes, on parle de régression simple ou multiple. Si la relation entre les variables est linéaire, on parle alors de régression linéaire.

HISTORIQUE

Précurseur dans le domaine de la régression linéaire, **Roger Joseph Boscovich**, astronome et physicien, fut l'un des premiers à trouver une méthode permettant de déterminer les coefficients d'une droite. Afin que la droite passe le plus près possible de toutes les observations, les deux paramètres doivent être déterminés d'une façon telle que les deux conditions de R.J. Boscovich soient respectées :

1. La somme des déviations doit être égale à zéro.

2. La somme de la valeur absolue des déviations doit être minimale.

R.J. Boscovich établit ces conditions aux environs de 1755-1757. Il utilisa une méthode de résolution géométrique qu'il appliqua lorsqu'il dut calculer la longueur de cinq méridiens terrestres.

En 1789, le marquis **Pierre Simon de Laplace**, dans son ouvrage *Sur les degrés mesurés des méridiens, et sur les longueurs observées sur pendule* adopta les deux conditions de R.J. Boscovich et établit une méthode de résolution algébrique de l'**algorithme** de Boscovich.

Deux autres mathématiciens, **Carl Friedrich Gauss** et **Adrien Marie Legendre**, paraissent avoir découvert, sans se concerter, la méthode des **moindres carrés**. Il semble que C.F. Gauss l'ait utilisée dès 1795. Toutefois, c'est à A.M. Legendre que l'on doit la publication en 1805 des *Nouvelles méthodes pour la détermination des orbites des comètes*, dans laquelle figure un appendice intitulé *Sur la méthode des moindres carrés*. Une controverse quant au véritable auteur de la découverte éclata alors entre les deux hommes et, en 1809, C.F. Gauss publia sa méthode en se référant à ses précédents travaux. Il est à remarquer qu'à la même époque, un Américain nommé Robert Adrian, qui n'avait pas eu connaissance des travaux de C.F. Gauss et A.M. Legendre, se consacra également à la recherche sur les moindres carrés.

Le XVIII^{e} siècle marque donc de façon significative le développement de la méthode des moindres carrés. Selon Stigler (1986), ce développement est lié à trois problèmes que se posaient les savants de l'époque. Le premier concernait la détermination et la représentation mathématique des mouvements de la lune. Pour le second, il s'agissait d'expliquer l'apparente inégalité séculaire observée dans les mouvements des planètes de Jupiter et de Saturne. Le troisième enfin visait à déterminer la forme de la Terre.

Selon le même auteur, il faut voir dans les travaux de **Francis Galton** l'origine du mot régression linéaire. Les recherches de F. Galton portaient sur l'hérédité : la notion de régression naîtra de l'analyse de l'hérédité des membres d'une famille.

En 1875, F. Galton réalisa une **expérience** avec des petits pois dont il fit 7 groupes selon le poids. Pour chaque groupe, il prit 10 semences et demanda à 7 de ses amis d'en faire une culture. Il disposa ainsi de 490 germes. Il trouva que chaque groupe de semences classé par poids suivait une distribution normale et que les courbes, centrées en des poids différents, étaient dispersées de façon égale ; la variabilité des différents groupes était identique. F. Galton parla de « réversion » (en biologie on désigne ainsi un retour au type primitif). Il s'agissait de réversion linéaire car les 7 groupes de semences (progéniture des petits pois de départ) étaient distribués normalement, non par rapport au poids de leurs parents, mais autour d'une valeur proche de la moyenne de la population de leur poids.

L'hérédité humaine fut également un objet d'études pour F. Galton qui chercha à définir la relation entre la taille des parents et celle de leur enfant. En 1889, il publia un ouvrage intitulé *Natural Inheritance*. Dans ce domaine, il utilisa dès 1885 le terme de régression et ses travaux ultérieurs lui permirent d'aboutir au concept de **corrélation**.

ASPECTS MATHÉMATIQUES

Voir **régression LAD**, **régression linéaire généralisée**, **régression linéaire multiple** et **simple**.

DOMAINES ET LIMITATIONS

Le but d'une analyse de régression n'est pas uniquement de déterminer la relation entre la **variable dépendante** et la (les) **variable**(s) **indépendante**(s), mais également d'établir le degré de fiabilité de l'**estimation** et, par conséquent, des prédictions que l'on a obtenues grâce à cette relation.

L'analyse de régression permet aussi d'examiner si les résultats sont significatifs et si la **relation** entre les variables est réelle ou seulement apparente.

L'analyse de régression a des applications multiples dans presque tous les domaines des sciences. En effet, lorsqu'on arrive à déterminer la relation

existant entre deux ou plusieurs variables, on peut alors, par **extrapolation**, prévoir les valeurs futures de ces variables, étant entendu que les conditions demeurent identiques et qu'il existe toujours une marge d'**erreur**.

On vérifie la validité d'une **analyse de régression**, entres autres, par une analyse des résidus.

EXEMPLE

Voir **régression linéaire** (**multiple** ou **simple**), **régressions LAD** et **ridge**.

MOTS CLÉS

Analyse des résidus *(Analysis of residuals)*
Analyse de variance *(Analysis of variance)*
Coefficient de détermination *(Coefficient of determination)*
Équation normale *(Normal equation)*
Matrice H *(Hat matrix)*
Moindres carrés *(Least squares)*
Point levier *(Leverage point)*
Régression linéaire multiple *(Multiple linear regression)*
Régression linéaire simple *(Simple linear regression)*
Résidu *(Residual)*

RÉFÉRENCES

Eisenhart, C. (1977). Boscovich and the Combination of Observations. *Studies in the History of Statistics and Probability.* Vol. II. Edited by M. Kendall and R.L. Plackett. Charles Griffin & Company Ltd, London.

Galton, F. (1889). *Natural Inheritance.* The Macmillan Company, London.

Gauss, C.F. (1809). *Theoria motus corporum cœlestium.* Werke 7.

Laplace, P.S. de (1789). Sur les degrés mesurés des méridiens, et sur les longueurs observées sur pendule. *Histoire de l'Académie royale des inscriptions et belles lettres, avec les Mémoires de littérature tirées des registres de cette académie.* Paris.

Legendre, A.M. (1805). *Nouvelles méthodes pour la détermination des orbites des comètes*, Courcier, Paris.

Plackett, R.L. (1977). *Studies in the history of probability and statistics. The discovery of the method of least squares.* Vol. II. Edited by M. Kendall and R.L. Plackett. Charles Griffin & Company Ltd.

Stigler, S. (1986). *The History of Statistics, the Measurement of Uncertainty before 1900.* The Belknap Press of Harvard University Press. London, England.

ANALYSE DE VARIANCE

Analysis of variance

L'analyse de variance est une technique qui consiste à séparer la variation totale d'un ensemble de données en composantes raisonnées associées à des sources spécifiques de variation, dans le but de comparer trois ou plusieurs moyennes des populations considérées et de conclure à l'égalité ou à la non-égalité globale de toutes les moyennes.

Elle permet également de tester certaines hypothèses concernant les paramètres du **modèle**, ou d'estimer les composantes de la **variance**.

Les sources de variation se résument globalement en une composante appelée « **erreur** » et une autre composante que l'on pourrait désigner par le terme « effet » (bien que, selon les cas, ce terme prenne des noms plus précis et puisse encore se subdiviser).

HISTORIQUE

L'usage de l'analyse de variance remonte à **Ronald Aylmer Fisher** (1925). Il en fut le pionnier et en posa les principes fondamentaux. Ses premières applications seront faites en agriculture et en biologie.

ASPECTS MATHÉMATIQUES

La technique d'analyse de **variance** permet de comparer les moyennes de t (≥ 3) **échantillons**

aléatoires afin de déterminer s'il existe une différence significative entre les populations dont sont issus les échantillons. Cette technique d'analyse de variance ne peut s'appliquer que si, d'une part, les échantillons aléatoires sont indépendants et d'autre part, les distributions des populations sont approximativement normales et toutes de même variance σ^2.

L'**hypothèse nulle** d'égalité des moyennes étant posée, de même que l'**hypothèse alternative** affirmant que l'une d'entre elles au moins est différente, on se fixe un **seuil de signification**.

Ensuite, on effectue deux estimations de la variance inconnue σ^2:

- la première, notée s_I^2, consiste en une moyenne des variances à l'intérieur de chaque échantillon ;
- la seconde, s_E^2, est basée sur la variation entre les moyennes des échantillons.

Idéalement, si l'hypothèse nulle est vérifiée, ces deux estimations seront égales et le ratio,

$$F = \frac{s_E^2}{s_I^2},$$

utilisé par le **test de Fisher** vaudra 1. La **valeur** du ratio F, généralement supérieure à 1 à cause de la variation d'**échantillonnage**, doit être comparée à la valeur de la **table de Fisher** correspondant au seuil de signification fixé. La règle de décision consiste alors à rejeter l'hypothèse nulle si la valeur calculée est plus grande ou égale à la valeur tabulée, sinon on admet l'égalité des moyennes ce qui signifie que les échantillons proviennent d'une même population.

Lorsqu'on est en présence du **modèle** suivant :

$$\begin{aligned} Y_{ij} = \mu + \tau_i + \varepsilon_{ij}\ , \quad & i = 1, 2, \ldots, t \\ & j = 1, 2, \ldots, n_i \end{aligned}$$

où

Y_{ij}	représente la j-ème **observation** recevant le **traitement** i,
μ	la moyenne générale commune à tous les traitements,
τ_i	est l'effet sur l'observation du traitement i, et
ε_{ij}	est l'**erreur** expérimentale de l'observation Y_{ij},

l'hypothèse nulle s'énonce de la façon suivante :

$$H_0 \quad : \quad \tau_1 = \tau_2 = \ldots = \tau_t$$

ce qui signifie que les t traitements sont identiques.

L'hypothèse alternative est formulée comme suit :

$$H_1 \quad : \quad \text{les valeurs des } \tau_i \ (i = 1, 2, \ldots, t) \text{ ne sont pas toutes identiques}$$

On utilise le formulaire suivant :

$$\begin{aligned} SC_E &= \sum_{i=1}^{t} n_i \left(\bar{Y}_{i.} - \bar{Y}_{..}\right)^2 \\ SC_I &= \sum_{i=1}^{t} \sum_{j=1}^{n_i} \left(Y_{ij} - \bar{Y}_{i.}\right)^2 \end{aligned}$$

et $SC_T = \sum_{i=1}^{t} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{..})^2$ ou alors $SC_T = SC_E + SC_I$,

où

$\bar{Y}_{i.} = \sum_{j=1}^{n_i} \frac{Y_{ij}}{n_i}$	est la moyenne du i-ème groupe,
$\bar{Y}_{..} = \frac{1}{N} \sum_{i=1}^{t} \sum_{j=1}^{n_i} Y_{ij}$	est la moyenne globale prise sur toutes les observations, et
$N = \sum_{i=1}^{t} n_i$	est le nombre total d'obs.

On est donc maintenant en mesure calculer s_E^2 et s_I^2 puisque :

$$s_E^2 = \frac{SC_E}{t-1} \text{ et } s_I^2 = \frac{SC_I}{N-t}.$$

Il est de coutume de résumer les informations de l'analyse de variance dans un tableau appelé tableau d'analyse de variance qui se présente comme suit :

Source de variation	Degrés de liberté	Somme des carrés	Moyenne des carrés	F
Entre les groupes	$t-1$	SC_E	s_E^2	$\frac{s_E^2}{s_I^2}$
À l'intérieur des groupes	$N-t$	SC_I	s_I^2	
Total	$N-1$	SC_T		

DOMAINES ET LIMITATIONS

Une analyse de variance est toujours associée à un **modèle**. Il est donc clair qu'il y aura une analyse de variance différente pour chaque cas distinct. Par exemple, lorsque l'analyse de variance est applicable à des expériences factorielles comprenant un ou plusieurs facteurs et que ces expériences factorielles sont liées à plusieurs types de plans d'expérience.

On peut non seulement distinguer le nombre de facteurs intervenant dans l'**expérience** mais aussi le genre d'hypothèses liées aux effets des traitements. Ainsi, on a le modèle à effets fixes, le modèle à effets variables et le modèle à effets mixtes. Chacun de ces cas nécessitera une analyse particulière mais, quel que soit le modèle utilisé, il doit respecter des hypothèses de base d'additivité, de normalité, d'homoscédasticité et d'indépendance. En clair, cela signifie :

1. que les erreurs expérimentales du modèle sont des variables aléatoires indépendantes ;

2. que les erreurs suivent toutes une **loi normale** de **moyenne** zéro et de **variance** σ^2 inconnue.

Toutes les sortes de plans d'expérience peuvent être analysés par une analyse de variance. Parmi ces plans, les plus courants sont : les plans complètement randomisés, les plans randomisés à blocs et les plans carrés latins.

On peut aussi faire une analyse de variance dans les cas de **régression linéaire simple**, **multiple** ou encore **généralisée**.

Si au cours d'une analyse de variance on est amené à rejeter l'hypothèse nulle d'égalité entre les moyennes, on utilise le **test de la différence significative minimale** pour identifier les populations dont les moyennes sont significativement différentes, ce que l'analyse de variance, *a priori*, ne permet pas.

EXEMPLES

Voir **analyse de variance pour deux facteurs**, **analyse de variance pour un facteur**, **régression linéaire multiple** et **régression linéaire simple**.

MOTS CLÉS

Analyse de régression *(Regression analysis)*
Analyse de variance pour deux facteurs *(Two-way analysis of variance)*
Analyse de variance pour un facteur *(One-way analysis of variance)*
Facteur *(Factor)*
Loi de Fisher *(Fisher distribution)*
Plan d'expérience *(Design of experiments)*
Régression linéaire multiple *(Mulitple linear regression)*
Régression linéaire simple *(Simple linear regression)*
Table de Fisher *(Fisher table)*
Test de Fisher *(Fisher test)*
Test de la différence significative minimale *(Least significant difference test)*

RÉFÉRENCES

Fisher, R.A. (1925). *Statistical Methods for Research Workers.* Oliver & Boyd, Edinburgh.

Rao, C.R. (1952). *Advanced Statistical Methods in Biometric Research.* John Wiley & Sons, New York.

Scheffé, H. (1959). *The analysis of variance.* John Wiley & Sons, New York.

ANALYSE DE VARIANCE POUR DEUX FACTEURS

Two-way analysis of variance

L'**analyse de variance** pour deux facteurs A et B est une technique permettant d'étudier les effets moyens de deux facteurs sur une **variable** de réponse dépendante. Dans l'analyse de variance pour deux facteurs, la source de variation que l'on appelle « **erreur** » correspond à la variation à l'intérieur des groupes ; celle que l'on appelle « effet » correspond à la variation entre les groupes. La variation entre les groupes peut être décomposée en effets du facteur A et effets du facteur B.

HISTORIQUE

C'est **Ronald Aylmer Fisher** (1925) qui baptisa les « expériences complexes » expériences factorielles.

F. Yates (1935 et 1937) a développé le concept et l'analyse de ces expériences factorielles. Voir **analyse de variance**.

ASPECTS MATHÉMATIQUES

Soit une **expérience** impliquant deux facteurs, un facteur A comprenant a niveaux et un facteur B comprenant b niveaux.

Si le plan associé à cette **expérience factorielle** est un **plan complètement randomisé**, le **modèle** est le suivant :

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk},$$

$$\begin{array}{rcll} i & = & 1, 2, \ldots, a & \text{(niveaux du facteur } A\text{) ;} \\ j & = & 1, 2, \ldots, b & \text{(niveaux du facteur } B\text{) ;} \\ k & = & 1, 2, \ldots, c & \text{(nombre d'observations ;} \\ & & & \text{recevant le } \textbf{traitement}\ ij\text{)} \end{array}$$

où

μ est la **moyenne** générale commune à tous les traitements ;

α_i l'effet du i-ème niveau du facteur A ;
β_j l'effet du j-ème niveau du facteur B ;
$(\alpha\beta)_{ij}$ l'effet de l'interaction entre α_i et β_j et
ε_{ijk} l'**erreur** expérimentale de l'observation Y_{ijk}.

Ce modèle est soumis aux hypothèses de base associées à l'**analyse de variance**, à savoir que l'on suppose encore que les erreurs ε_{ijk} sont des variables aléatoires indépendantes suivant une **loi normale** $\mathcal{N}(0, \sigma^2)$.

On se propose d'effectuer l'analyse de variance pour deux facteurs sur ce modèle. On peut alors tester trois hypothèses :

1. H_0 : $\alpha_1 = \alpha_2 = \ldots = \alpha_a$
   H_1 : au moins un α_i est différent de α_j, $i \neq j$

2. H_0 : $\beta_1 = \beta_2 = \ldots = \beta_b$
   H_1 : au moins un β_i est différent de β_j, $i \neq j$

3. H_0 : $(\alpha\beta)_{11} = (\alpha\beta)_{12} = \ldots = (\alpha\beta)_{1b} = (\alpha\beta)_{21} = \ldots = (\alpha\beta)_{ab}$
   H_1 : au moins une des interactions est différente des autres.

Pour tester la première hypothèse, on utilise le **test de Fisher** pour lequel on forme un ratio dont le numérateur est une **estimation** de la **variance** du facteur A et le dénominateur une estimation de la variance à l'intérieur des groupes.

Ce ratio, noté F, suit une **loi de Fisher** avec $a - 1$ et $ab(c - 1)$ degrés de liberté.

L'**hypothèse nulle** :

$$H_0 \quad : \quad \alpha_1 = \alpha_2 = \ldots = \alpha_a$$

sera alors rejetée au **seuil de signification** α si le ratio F est supérieur ou égal à la **valeur** de la **table de Fisher**, c'est-à-dire si :

$$F \geq F_{a-1,ab(c-1),\alpha}.$$

Pour tester la deuxième hypothèse, on forme un ratio dont le numérateur est une estimation de la variance du facteur B et le dénominateur une estimation de la variance à l'intérieur des groupes.

Ce ratio, noté F, suit une loi de Fisher avec $b-1$ et $ab(c-1)$ degrés de liberté.

L'hypothèse nulle :

$$H_0 \quad : \quad \beta_1 = \beta_2 = \ldots = \beta_b$$

sera alors rejetée si le ratio F est supérieur ou égal à la valeur de la table de Fisher, c'est-à-dire si :

$$F \geq F_{b-1,ab(c-1),\alpha}.$$

Pour tester la troisième hypothèse, on forme un ratio dont le numérateur est une estimation de la variance de l'interaction entre les facteurs A et B et le dénominateur une estimation de la variance à l'intérieur des groupes.

Ce ratio, noté F, suit une loi de Fisher avec $(a-1)(b-1)$ et $ab(c-1)$ degrés de liberté.

L'hypothèse nulle :

$$H_0 \quad : \quad (\alpha\beta)_{11} = (\alpha\beta)_{12} = \ldots = (\alpha\beta)_{1b} = (\alpha\beta)_{21} = \ldots = (\alpha\beta)_{ab}$$

sera alors rejetée si le ratio F est supérieur ou égal à la valeur de la table de Fisher, c'est-à-dire si

$$F \geq F_{(a-1)(b-1),ab(c-1),\alpha}.$$

Variance du facteur A

Pour la variance du facteur A, il faut calculer la somme des carrés pour le facteur A (SC_A), qui s'obtient comme suit :

$$SC_A = b \cdot c \sum_{i=1}^{a} \left(\bar{Y}_{i..} - \bar{Y}_{...}\right)^2$$

où

- $\bar{Y}_{i..}$ est la moyenne de toutes les observations du niveau i du facteur A et
- $\bar{Y}_{...}$ la moyenne générale de toutes les observations.

Le nombre de degrés de liberté associé à cette somme est égale à $a-1$.

La variance du facteur A est donc égale à :

$$s_A^2 = \frac{SC_A}{a-1}\ .$$

Variance du facteur B

Pour la variance du facteur B, il faut calculer la somme des carrés pour le facteur B, qui s'obtient comme suit :

$$SC_B = a \cdot c \sum_{j=1}^{b} (\bar{Y}_{.j.} - \bar{Y}_{...})^2$$

où

- $\bar{Y}_{.j.}$ est la moyenne de toutes les observations du niveau j du facteur B et
- $\bar{Y}_{...}$ la moyenne générale de toutes les observations.

Le nombre de degrés de liberté associé à cette somme est égal à $b-1$.

La variance du facteur B est donc égale à :

$$s_B^2 = \frac{SC_B}{b-1}.$$

Variance de l'interaction AB

Pour la variance de l'**interaction** AB, il faut calculer la somme des carrés pour l'interaction AB, qui s'obtient comme suit :

$$SC_{AB} = c \sum_{i=1}^{a} \sum_{j=1}^{b} \left(\bar{Y}_{ij.} - \bar{Y}_{i..} - \bar{Y}_{.j.} + \bar{Y}_{...}\right)^2$$

où

- $\bar{Y}_{ij.}$ est la moyenne de toutes les observations du niveau ij de l'interaction AB,
- $\bar{Y}_{i..}$ est la moyenne de toutes les observations du niveau i du facteur A,
- $\bar{Y}_{.j.}$ est la moyenne de toutes les observations du niveau j du facteur B, et
- $\bar{Y}_{...}$ la moyenne générale de toutes les observations.

Le nombre de degrés de liberté associé à cette somme est égal à $(a-1)(b-1)$.

La variance de l'interaction AB est donc égale à :

$$s^2_{AB} = \frac{SC_{AB}}{(a-1)(b-1)}.$$

Variance à l'intérieur des groupes

Pour la variance à l'intérieur des groupes, il faut calculer la somme des carrés à l'intérieur des groupes qui s'obtient de la façon suivante :

$$SC_I = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c}\left(Y_{ijk} - \bar{Y}_{ij.}\right)^2$$

où

a et b	sont respectivement le nombre de niveaux des facteurs A et B,
c	est le nombre d'observations recevant le traitement ij (constant quels que soient i et j),
Y_{ijk}	est la k-ème observation du i-ème niveau du facteur A et du j-ème niveau de B, et
$\bar{Y}_{ij.}$	la moyenne de toutes les observations du niveau ij de l'interaction AB.

Le nombre de degrés de liberté associé à cette somme est égal à $ab(c-1)$.

La variance à l'intérieur des groupes est donc égale à :

$$s^2_I = \frac{SC_I}{ab(c-1)}.$$

La somme des carrés totale (SC_T) est égale à :

$$SC_T = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c} Y^2_{ijk}$$

Le nombre de degrés de liberté associé à cette somme est égal à N, c'est-à-dire au nombre total d'observations.

La somme des carrés pour la moyenne est égale à :

$$SC_M = N\bar{Y}^2_{...}$$

Le nombre de degrés de liberté associé à cette somme est égal à 1.

La somme des carrés totale (SC_T) peut être exprimée à l'aide de toutes les autres sommes des carrés de la façon suivante :

$$SC_T = SC_M + SC_A + SC_B + SC_{AB} + SC_I.$$

Tests de Fisher

On peut maintenant calculer les différents ratios F.

Pour tester l'hypothèse nulle

$$H_0 \quad : \quad \alpha_1 = \alpha_2 = \ldots = \alpha_a,$$

on forme le premier ratio F dont le numérateur est l'estimation de la variance du facteur A et le dénominateur l'estimation de la variance à l'intérieur des groupes :

$$F = \frac{s^2_A}{s^2_I}.$$

Si F est supérieur ou égal à la valeur de la table de Fisher pour $a-1$ et $ab(c-1)$ degrés de liberté, on rejette l'hypothèse nulle et on conclut que le facteur A a un effet significatif.

Pour tester l'hypothèse nulle

$$H_0 \quad : \quad \beta_1 = \beta_2 = \ldots = \beta_b,$$

on forme le deuxième ratio F dont le numérateur est l'estimation de la variance du facteur B et le dénominateur l'estimation de la variance à l'intérieur des groupes :

$$F = \frac{s^2_B}{s^2_I}.$$

Si F est supérieur ou égal à la valeur de la table de Fisher pour $b-1$ et $ab(c-1)$ degrés de liberté, on rejette l'hypothèse nulle et on conclut que le facteur B a un effet significatif.

Pour tester l'hypothèse nulle

$$H_0 \quad : \quad (\alpha\beta)_{11} = (\alpha\beta)_{12} = \ldots = (\alpha\beta)_{1b} = (\alpha\beta)_{21} = \ldots = (\alpha\beta)_{ab},$$

on forme le troisième ratio F dont le numérateur est l'estimation de la variance de l'interaction AB et le dénominateur l'estimation de la variance à l'intérieur des groupes :

$$F = \frac{s^2_{AB}}{s^2_I}.$$

Si F est supérieur ou égal à la valeur de la table de Fisher pour $(a-1)(b-1)$ et $ab(c-1)$ degrés de liberté, on rejette l'hypothèse nulle et on conclut que l'interaction B a un effet significatif.

Tableau d'analyse de variance

Toutes les informations nécessaires au calcul des deux ratios F peuvent être résumées sous forme d'un tableau d'analyse de variance :

Source de variation	Degrés de liberté	Somme des carrés	Moyenne des carrés	F
Moyenne	1	SC_M		
Facteur A	$a-1$	SC_A	s^2_A	$\frac{s^2_A}{s^2_I}$
Facteur B	$b-1$	SC_B	s^2_B	$\frac{s^2_B}{s^2_I}$
Interaction AB	$(a-1)\cdot(b-1)$	SC_{AB}	s^2_{AB}	$\frac{s^2_{AB}}{s^2_I}$
Intérieur des groupes	$ab(c-1)$	SC_I	s^2_I	
Total	abc	SC_T		

Si $c = 1$, c'est-à-dire qu'il n'y a qu'une seule observation par traitement ij (ou par cellule ij), il n'est pas possible d'attribuer une partie de l'erreur « intérieure des groupes » à l'interaction.

EXEMPLES

Le directeur d'une école veut faire tester quatre marques de machines à écrire. Il demande à cinq secrétaires professionnelles d'essayer chacune des quatre marques et de répéter l'exercice le lendemain. Elles doivent toutes taper le même texte et après quinze minutes, on relève le nombre de mots moyen tapés en une minute.

Voici les résultats :

	Secrétaires				
Machines à écrire	1	2	3	4	5
1	33	31	34	34	31
	36	31	36	33	31
2	32	37	40	33	35
	35	35	36	36	36
3	37	35	34	31	37
	39	35	37	35	40
4	29	31	33	31	33
	31	33	34	27	33

On fait une **analyse de variance** pour deux facteurs, le premier facteur étant les marques de machines à écrire et le deuxième facteur les secrétaires.

Les hypothèses nulles sont les suivantes :

1. H_0 : $\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4$
   H_1 : au moins un α_i est différent de α_j, $i \neq j$
2. H_0 : $\beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5$
   H_1 : au moins un β_i est différent de β_j, $i \neq j$
3. H_0 : $(\alpha\beta)_{11} = (\alpha\beta)_{12} = \ldots = (\alpha\beta)_{15} = (\alpha\beta)_{21} = \ldots = (\alpha\beta)_{45}$
   H_1 : au moins une des interactions est différentes des autres.

Variance du facteur A « marques des machines à écrire »

La somme des carrés pour le facteur A vaut :

$$\begin{aligned} SC_A &= 5 \cdot 2 \sum_{i=1}^{4} \left(\bar{Y}_{i..} - \bar{Y}_{...}\right)^2 \\ &= 5 \cdot 2[(33-34)^2 + (35,5-34)^2 + \\ &\quad (36-34)^2 + (31,5-34)^2] \\ &= 135. \end{aligned}$$

Le nombre de degrés de liberté associé à cette somme est égal à $4 - 1 = 3$.

La **variance** du facteur A (**moyenne** des carrés) est donc égale à :

$$\begin{aligned} s_A^2 &= \frac{SC_A}{a-1} \\ &= \frac{135}{3} \\ &= 45. \end{aligned}$$

Variance du facteur B « secrétaires »

La somme des carrés pour le facteur B vaut :

$$\begin{aligned} SC_B &= 4 \cdot 2 \sum_{j=1}^{5} \left(\bar{Y}_{\cdot j\cdot} - \bar{Y}_{\cdots}\right)^2 \\ &= 4 \cdot 2[(34-34)^2 + (33,5-34)^2 + \ldots \\ &\quad + (35,5-34)^2 + (34,5-34)^2] \\ &= 40. \end{aligned}$$

Le nombre de degrés de liberté associé à cette somme est égal à $5-1=4$.

La variance du facteur B (moyenne des carrés) est donc égale à :

$$\begin{aligned} s_B^2 &= \frac{SC_B}{b-1} \\ &= \frac{40}{4} \\ &= 10. \end{aligned}$$

Variance de l'interaction AB « marques des machines à écrire-secrétaires »

La somme des carrés pour l'**interaction** AB vaut :

$$\begin{aligned} SC_{AB} &= 2 \sum_{i=1}^{4} \sum_{j=1}^{5} \left(\bar{Y}_{ij\cdot} - \bar{Y}_{i\cdot\cdot} - \bar{Y}_{\cdot j\cdot} + \bar{Y}_{\cdots}\right)^2 \\ &= 2[(34,5-33-34+34)^2 + \\ &\quad (31-31,5-34,5+34)^2 + \ldots \\ &\quad + (33-31,5-34,5+34)^2] \\ &= 2[1,5^2 + (-1,5)^2 + \ldots + 1^2] \\ &= 83. \end{aligned}$$

Le nombre de degrés de liberté associé à cette somme est égal à $(4-1)(5-1)=12$.

La variance de l'interaction AB est donc égale à :

$$\begin{aligned} s_{AB}^2 &= \frac{SC_{AB}}{(a-1)(b-1)} \\ &= \frac{83}{12} \\ &= 6,92. \end{aligned}$$

Variance à l'intérieur des groupes

La somme des carrés à l'intérieur des groupes vaut :

$$\begin{aligned} SC_I &= \sum_{i=1}^{4} \sum_{j=1}^{5} \sum_{k=1}^{2} \left(Y_{ijk} - \bar{Y}_{ij\cdot}\right)^2 \\ &= (33-34,5)^2 + (36-34,5)^2 \\ &\quad + (31-31)^2 + \ldots + (33-33)^2 \\ &= 58. \end{aligned}$$

Le nombre de degrés de liberté associé à cette somme est égal à $4 \cdot 5\,(2-1) = 20$.

$$\begin{aligned} s_I^2 &= \frac{SC_I}{ab\,(c-1)} \\ &= \frac{58}{20} \\ &= 2,9. \end{aligned}$$

La somme des carrés totale est égale à :

$$\begin{aligned} SC_T &= \sum_{i=1}^{4} \sum_{j=1}^{5} \sum_{k=1}^{2} \left(Y_{ijk}\right)^2 \\ &= 33^2 + 36^2 + 31^2 + \ldots \\ &\quad + 27^2 + 33^2 + 33^2 \\ &= 46\,556. \end{aligned}$$

Le nombre de degrés de liberté associé à cette somme est égal au nombre N d'observations, soit 40.

La somme des carrés pour la **moyenne** est égale à :

$$\begin{aligned} SC_M &= N\bar{Y}_{\cdots}^2 \\ &= 40 \cdot 34^2 \\ &= 46\,240. \end{aligned}$$

Le nombre de degrés de liberté associé à cette somme est égal à 1.

Tests de Fisher

Pour tester l'hypothèse nulle

$$H_0 \quad : \quad \alpha_1 = \alpha_2 = \alpha_3 = \alpha_4,$$

on forme le premier ratio F dont le numérateur est l'**estimation** de la variance du facteur A et le dénominateur l'estimation de la variance à l'intérieur des groupes :

$$\begin{aligned} F &= \frac{s_A^2}{s_I^2} \\ &= \frac{45}{2,9} \\ &= 15,517. \end{aligned}$$

La **valeur** de la **table de Fisher** avec 3 et 20 degrés de liberté et pour un **seuil de signification** $\alpha = 0,05$ est égale à $3,10$.

Comme $F \geq F_{3,20,0,05}$, on rejette l'**hypothèse** H_0 et on conclut que les marques de machines à écrire ont un effet significatif sur le nombre de mots tapés en une minute.

Pour tester l'hypothèse nulle

$$H_0 \quad : \quad \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5,$$

on forme le deuxième ratio F dont le numérateur est l'estimation de la variance du facteur B et le dénominateur l'estimation de la variance à l'intérieur des groupes :

$$\begin{aligned} F &= \frac{s_B^2}{s_I^2} \\ &= \frac{10}{2,9} \\ &= 3,448. \end{aligned}$$

La valeur de la table de Fisher avec 4 et 20 degrés de liberté et pour $\alpha = 0,05$ est égale à $2,87$.

Comme $F \geq F_{4,20,0,05}$, on rejette l'hypothèse H_0 et on conclut que les secrétaires ont un effet significatif sur le nombre de mots tapés en une minute.

Pour tester l'hypothèse nulle

$$H_0 \quad : \quad (\alpha\beta)_{11} = (\alpha\beta)_{12} = \ldots = (\alpha\beta)_{15} = (\alpha\beta)_{21} = \ldots = (\alpha\beta)_{45},$$

on forme le troisième ratio F dont le numérateur est l'estimation de la variance de l'interaction AB et le dénominateur l'estimation de la variance à l'intérieur des groupes :

$$\begin{aligned} F &= \frac{s_{AB}^2}{s_I^2} \\ &= \frac{6,92}{2,9} \\ &= 2,385. \end{aligned}$$

La valeur de la table de Fisher avec 12 et 20 degrés de liberté et pour $\alpha = 0,05$ est égale à $2,28$.

Comme $F \geq F_{4,20,0,05}$, on rejette l'hypothèse H_0 et on conclut que l'effet de l'interaction entre les secrétaires et les types de machines à écrire est significatif sur le nombre de mots tapés en une minute.

Tableau d'analyse de variance

Toutes ces informations sont résumées dans le tableau d'analyse de variance ci-dessous :

Source de variation	Degrés de liberté	Somme des carrés	Moyenne des carrés	F
Moyenne	1	46 240		
Facteur A	3	135	45	15,517
Facteur B	4	40	10	3,448
Interaction AB	12	83	6,92	2,385
Intérieur des groupes	20	58	2,9	
Total	40	46 556		

MOTS CLÉS

Analyse de variance *(Analysis of variance)*

Analyse de variance pour un facteur *(One-way analysis of variance)*
Contraste *(Contrast)*
Interaction *(Interaction)*
Table de Fisher *(Fisher table)*
Test de Fisher *(Fisher test)*
Test de la différence significative minimale *(Least significant difference test)*

RÉFÉRENCES

Cox, D.R. and Reid, N. (2000). *Theory of the design of experiments.* Chapman & Hall, London.

Fisher, R.A. (1925). *Statistical Methods for Research Workers.* Oliver & Boyd, Edinburgh.

Yates, F. (1935). Complex Experiments. *Supplement to the Journal of the Royal Statistical Society*, 2, pp. 181-247.

Yates, F. (1937). The design and analysis of factorial experiments. *Technical Communication of the Commonwealth Bureau of Soils*, 35, Commonwealth Agricultural Bureaux, Farnham Royal.

ANALYSE DE VARIANCE POUR UN FACTEUR

One-way analysis of variance

L'**analyse de variance** pour un **facteur** à t niveaux (ou t traitements) est une technique permettant de déterminer s'il existe une différence significative entre les t traitements (populations). Le principe de l'analyse de variance pour un facteur est, à partir d'échantillons tirés de ces populations, de comparer la variabilité à l'intérieur de chaque échantillon avec la variabilité entre les échantillons. La source de variation que l'on appelle « **erreur** » correspond donc à la variabilité à l'intérieur des échantillons, celle que l'on appelle « effet » correspond à la variabilité entre les échantillons.

HISTORIQUE

Voir **analyse de variance**.

ASPECTS MATHÉMATIQUES

Le **modèle** linéaire pour un **facteur** à t niveaux (**traitements**) se présente comme suit :

$$Y_{ij} = \mu + \tau_i + \varepsilon_{ij} \ , \quad \begin{matrix} i = 1, 2, \ldots, t \\ j = 1, 2, \ldots, n_i \end{matrix}$$

où

- Y_{ij} représente la j-ème **observation** recevant le traitement i,
- μ est la **moyenne** générale commune à tous les traitements,
- τ_i est l'effet réel sur l'observation du traitement i, et
- ε_{ij} est l'**erreur** expérimentale de l'observation Y_{ij}.

Ce modèle est soumis aux hypothèses de base associées à l'**analyse de variance**, à savoir que l'on suppose encore que les erreurs ε_{ij} sont des variables aléatoires indépendantes suivant une **loi normale** $\mathcal{N}(0, \sigma^2)$.

Pour déterminer s'il existe une différence entre les t traitements, on effectue un **test d'hypothèse**. L'**hypothèse nulle** suivante :

$$H_0 \quad : \quad \tau_1 = \tau_2 = \ldots = \tau_t$$

signifie que les t traitements sont identiques.

L'**hypothèse alternative** est formulée comme suit:

$$H_1 \quad : \quad \text{les valeurs des } \tau_i \ (i = 1, 2, \ldots, t) \text{ ne sont pas toutes identiques.}$$

Le principe de l'analyse de variance est de comparer la variabilité à l'intérieur de chaque **échantillon** avec la variabilité entre les échantillons. Pour cela on

utilise le **test de Fisher** pour lequel on forme un ratio dont le numérateur est une **estimation** de la **variance** entre les groupes (échantillons) et le dénominateur une estimation de la variance à l'intérieur des groupes.

Ce ratio, noté F, suit une **loi de Fisher** avec $t-1$ et $N-t$ degrés de liberté (N étant le nombre total d'observations). L'hypothèse nulle H_0 sera alors rejetée au **seuil de signification** α si le ratio F est supérieur ou égal à la valeur de la **table de Fisher**, c'est-à-dire si

$$F \geq F_{t-1,N-t,\alpha}.$$

Si l'hypothèse nulle H_0 ne peut pas être rejetée, on peut considérer que les t échantillons proviennent de la même **population**.

Calcul de la variance entre les groupes

Pour la variance entre les groupes, il faut calculer la somme des carrés entre les groupes qui s'obtient comme suit :

$$SC_E = \sum_{i=1}^{t} n_i(\bar{Y}_{i.} - \bar{Y}_{..})^2$$

où $\bar{Y}_{i.}$ est la moyenne du i-ème groupe.

Le nombre de degrés de liberté associé à cette somme est égal à $t-1$, t étant le nombre d'échantillons (ou de groupes).

La variance entre les groupes est donc égale à :

$$s_E^2 = \frac{SC_E}{t-1}\ .$$

Calcul de la variance à l'intérieur des groupes

Pour la variance à l'intérieur des groupes, il faut calculer la somme des carrés à l'intérieur des groupes, qui s'obtient comme suit :

$$SC_I = \sum_{i=1}^{t}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_{i.})^2.$$

Le nombre de degrés de liberté associé à cette somme est égal à $N-t$, N étant le nombre total d'observations et t le nombre d'échantillons.

La variance à l'intérieur des groupes est donc égale à :

$$s_I^2 = \frac{SC_I}{N-t}.$$

La somme des carrés totale (SC_T) est la somme des carrés des déviations de chaque observation Y_{ij} par rapport à la moyenne générale $\bar{Y}_{..}$:

$$SC_T = \sum_{i=1}^{t}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_{..})^2.$$

Le nombre de degrés de liberté associé à cette somme est égal à $N-1$, N étant le nombre total d'observations. On peut aussi trouver SC_T de la façon suivante :

$$SC_T = SC_E + SC_I.$$

Test de Fisher

On peut maintenant calculer le ratio F et déterminer si les échantillons peuvent être vus comme provenant de la même population.

Le ratio F est égal à l'estimation de la variance entre les groupes divisée par l'estimation de la variance à l'intérieur des groupes, c'est-à-dire

$$F = \frac{s_E^2}{s_I^2}.$$

Si F est supérieur ou égal à la valeur de la table de Fisher pour $t-1$ et $N-t$ degrés de liberté, on peut conclure que la différence entre les échantillons provient des traitements expérimentaux et n'est pas due au seul hasard.

Tableau d'analyse de variance

Il est de coutume de résumer les informations de l'analyse de variance dans un tableau appelé tableau d'analyse de variance qui se présente comme suit :

Source de variation	Degrés de liberté	Somme des carrés	Moyenne des carrés	F
Entre les groupes	$t-1$	SC_E	s_E^2	$\frac{s_E^2}{s_I^2}$
À l'intérieur des groupes	$N-t$	SC_I	s_I^2	
Total	$N-1$	SC_T		

Comparaisons de moyennes prises deux à deux

Lorsque le test de Fisher conduit à rejeter l'hypothèse nulle, c'est-à-dire à considérer qu'il existe une différence significative entre les moyennes des échantillons, on peut se poser la question : où se trouve cette différence, entre quelles moyennes ?

Plusieurs tests ont été développés pour répondre à cette question ; il existe entre autres le **test de la différence significative minimale** (en anglais *Least Significant Difference* : LSD) qui effectue des comparaisons de moyennes prises deux à deux. Pour pouvoir employer ce test, il faut que le ratio F indique une différence significative entre les moyennes.

EXEMPLE

Pendant leur cuisson, les croissants absorbent de la graisse en quantité variable. On veut voir si la quantité absorbée dépend du type de graisse. On prépare donc quatre graisses différentes et on fait cuire six croissants par type de graisse.

Voici les données enregistrées (les chiffres sont des quantités de graisse absorbées par croissant) :

Graisse			
1	2	3	4
64	78	75	55
72	91	93	66
68	97	78	49
77	82	71	64
56	85	63	70
95	77	76	68

L'**hypothèse nulle** est qu'il n'y a pas de différence entre les traitements, c'est-à-dire que la quantité de graisse absorbée lors de la cuisson ne dépend pas du type de graisse :

$$H_0 \quad : \quad \tau_1 = \tau_2 = \tau_3 = \tau_4.$$

Pour tester cette **hypothèse**, on utilise le **test de Fisher** pour lequel il faut déterminer le ratio F. Il faut donc calculer la variance entre les groupes et la variance à l'intérieur des groupes.

Variance entre les groupes

Il faut calculer la **moyenne** générale $\bar{Y}_{..}$ et les moyennes des 4 groupes $\bar{Y}_{i.} = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{1i}$ $(i = 1, 2, 3, 4)$.

Comme tous les n_i valent 6 et que N est égal à 24, cela donne :

$$\begin{aligned} \bar{Y}_{..} &= \frac{64 + 72 + \ldots + 70 + 68}{24} \\ &= \frac{1770}{24} \\ &= 73,75. \end{aligned}$$

ainsi que :

$$\bar{Y}_{1.} = \frac{1}{6} \sum_{j=1}^{6} Y_{1j} = \frac{432}{6} = 72 \ ;$$

$$\bar{Y}_{2.} = \frac{1}{6} \sum_{j=1}^{6} Y_{2j} = \frac{510}{6} = 85 \ ;$$

$$\bar{Y}_{3.} = \frac{1}{6} \sum_{j=1}^{6} Y_{3j} = \frac{456}{6} = 76 \ ;$$

$$\bar{Y}_{4.} = \frac{1}{6} \sum_{j=1}^{6} Y_{4j} = \frac{372}{6} = 62.$$

On peut alors calculer la somme des carrés entre les groupes :

$$\begin{aligned} SC_E &= \sum_{i=1}^{4} n_i(\bar{Y}_{i.} - \bar{Y}_{..})^2 \\ &= 6\,(72 - 73,75)^2 + 6\,(85 - 73,75)^2 + \\ &\quad 6\,(76 - 73,75)^2 + 6\,(62 - 73,75)^2 \\ &= 1\,636,5. \end{aligned}$$

Le nombre de degrés de liberté associé à cette somme est égal à $4 - 1 = 3$.

La variance entre les groupes est donc égale à :

$$s_E^2 = \frac{SC_E}{t-1} = \frac{1\,636,5}{3} = 545,5.$$

Variance à l'intérieur des groupes

La somme des carrés à l'intérieur des groupes est égale à :

$$\begin{aligned} SC_I &= \sum_{i=1}^{4}\sum_{j=1}^{6} \left(Y_{ij} - \bar{Y}_{i.}\right)^2 \\ &= (64 - 72)^2 + (72 - 72)^2 + \ldots + \\ &\quad (70 - 62)^2 + (68 - 62)^2 \\ &= 2\,018. \end{aligned}$$

Le nombre de degrés de liberté associé à cette somme est égal à $24 - 4 = 20$.

La variance à l'intérieur des groupes est donc égale à :

$$s_I^2 = \frac{SC_I}{N-t} = \frac{2018}{20} = 100,9.$$

Test de Fisher

Tous les éléments sont connus pour calculer F :

$$F = \frac{s_E^2}{s_I^2} = \frac{545,5}{100,9} = 5,406.$$

Il faut trouver maintenant la **valeur** de la **table de Fisher** avec 3 et 20 degrés de liberté. Avec un **seuil de signification** $\alpha = 0,05$,

$$F_{3,20,0,05} = 3,10.$$

On constate que

$$\begin{aligned} F &\geq F_{3,20,0,05} \\ 5.406 &\geq 3,10 \end{aligned}$$

ce qui signifie que H_0 doit être rejetée. Il y a donc une différence significative entre les traitements. Par conséquent la quantité de graisse absorbée dépend du type de graisse utilisé.

Tableau d'analyse de variance

On résume ces informations dans le tableau d'**analyse de variance** suivant :

Source de variation	Degrés de liberté	Somme des carrés	Moyenne des carrés	F
Entre les groupes	3	1 636,5	545,5	5,406
À l'intérieur des groupes	20	2 018	100,9	
Total	23	3 654,5		

MOTS CLÉS

Analyse de variance *(Analysis of variance)*
Analyse de variance pour deux facteurs *(Two-way analysis of variance)*
Contraste *(Contrast)*
Table de Fisher *(Fisher table)*
Test de Fisher *(Fisher test)*
Test de la différence significative minimale *(Least significant difference test)*

RÉFÉRENCES

Cox, D.R. and Reid, N. (2000). *Theory of the design of experiments.* Chapman & Hall, London.

Montgomery, D.C. (1997). *Design and analysis of experiments.* 4th ed. John Wiley & Sons, Chichester.

ANALYSE DES RÉSIDUS

Analysis of residuals

L'analyse des résidus a pour objectif de tester la validité d'un **modèle** de **régression**, conformément aux hypothèses relatives aux résidus sous-jacentes au modèle choisi. L'analyse des résidus standardisés permet de détecter d'éventuelles observations aberrantes au sein de l'ensemble de données considéré.

HISTORIQUE

L'analyse des résidus remonte à Leonhard Euler (1749) et Johann Tobias Mayer (1750) qui se virent confrontés au milieu du XVIIIe siècle à un problème d'**estimation** de paramètres à partir d'observations dans le domaine de l'astronomie. La plupart des méthodes d'analyse des résidus utilisées actuellement sont basées sur les travaux de Francis John Anscombe (1961) et de F.J. Anscombe et **John Wilder Tukey** (1963). F.J. Anscombe présenta également en 1973 une discussion intéressante concernant les motivations d'utilisation des méthodes graphiques d'analyse. Un ouvrage complet est consacré à l'analyse des résidus par R. Dennis Cook et Sanford Weisberg (1982). Norman R. Draper et Harry Smith (1998) se sont également intéressés à ce problème dans un chapitre de leur ouvrage *Applied regression analysis.*

ASPECTS MATHÉMATIQUES

Considérons un modèle général de **régression linéaire multiple** :

$$Y_i = \beta_0 + \sum_{j=1}^{p-1} \beta_j X_{ij} + \varepsilon_i, \quad i = 1, \ldots, n$$

où ε_i est le terme d'**erreur** aléatoire non observable.

Les hypothèses sur les erreurs ε_i sont généralement les suivantes :

- les erreurs sont indépendantes,
- elles suivent une **loi normale**,
- leur **moyenne** est égale à zéro,
- leur **variance** est constante et égale à σ^2.

L'**analyse de régression** fournit une **estimation** de Y_i, notée $\hat{Y}_i$. Si le **modèle** choisi est adéquat, la distribution des résidus ou « erreurs observées » $e_i = Y_i - \hat{Y}_i$, doit confirmer ces hypothèses.

Les méthodes d'analyse des résidus sont principalement des méthodes graphiques, parmi lesquelles nous pouvons citer :

1. la représentation des résidus par un graphique de fréquences (par exemple un **diagramme en points**) ;
2. le graphique des résidus en fonction du temps (si l'ordre chronologique est connu) ;
3. le graphique des résidus en fonction des valeurs estimées $\hat{Y}_i$;
4. le graphique des résidus en fonction des variables indépendantes X_{ij} ;
5. le **Q-Q plot** des résidus.

DOMAINES ET LIMITATIONS

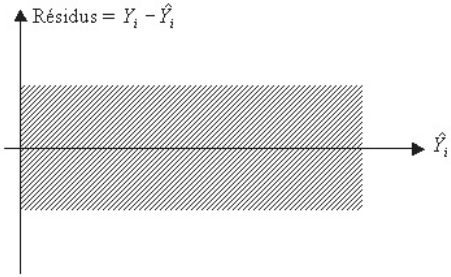
Pour la validité de l'analyse, il faut que certaines hypothèses (comme par exemple la normalité des résidus dans les estimations utilisant les **moindres carrés**) soient vérifiées.

Considérons le graphique des résidus en fonction des valeurs estimées $\hat{Y}_i$, ce graphique étant l'un des plus fréquemment utilisés pour vérifier la validité d'un **modèle**. Il consiste à placer :

- en ordonnée les résidus $e_i = Y_i - \hat{Y}_i$;

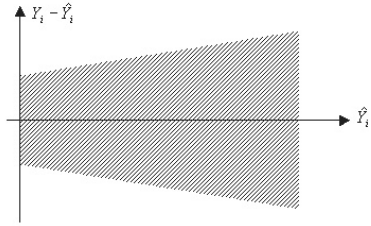
- en abscisse les valeurs estimées $\hat{Y}_i$.

Si le modèle choisi est adéquat, les résidus sont distribués uniformément sur une bande horizontale du graphique :

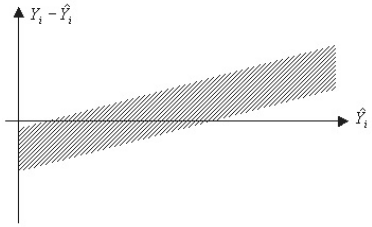


Par contre, si les hypothèses concernant les résidus ne sont pas vérifiées, le graphique peut prendre des allures différentes. Les trois figures ci-dessous montrent :

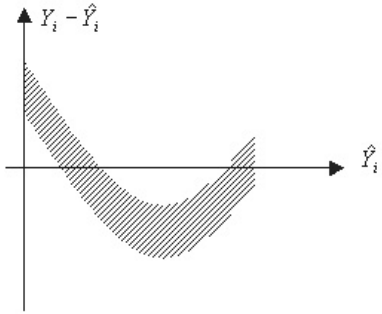
1. que la **variance** σ^2 n'est pas constante. Dans ce cas, il est nécessaire d'effectuer une **transformation** des données Y_i avant d'effectuer l'**analyse de régression** ;



2. que le modèle choisi est inadéquat (le modèle est linéaire, mais par exemple le terme constant a été omis alors qu'il était nécessaire) ;



3. que le modèle choisi est inadéquat (tendance parabolique).



Différentes statistiques ont été suggérées afin d'apporter une mesure numérique complémentaire aux techniques visuelles présentées ci-dessus, entre autres celles données par Anscombe (1961) et Anscombe et **John Wilder Tukey** (1963).

EXEMPLES

Au XIX^{e} siècle, un physicien écossais du nom de James D. Forbe espérait pouvoir estimer l'altitude au-dessus du niveau de la mer d'après la mesure du point d'ébullition de l'eau. Il savait que l'altitude pouvait être déterminée à partir de la pression atmosphérique. Il a donc étudié la relation entre la pression et le point d'ébullition. J.D. Forbe suggéra que dans l'**intervalle** des valeurs observées, le graphe du logarithme de la pression en fonction du point d'ébullition devait donner une droite. Comme le logarithme de ces pressions est petit et qu'il varie peu, nous avons multiplié ces valeurs par 100.

X point d'ébullition	Y $100 \cdot \log(\text{pression})$
194,5	131,79
194,3	131,79
197,9	135,02
198,4	135,55
199,4	136,46
199,9	136,83
200,9	137,82
201,1	138,00
201,4	138,06
201,3	138,05
203,6	140,04
204,6	142,44
209,5	145,47
208,6	144,34
210,7	146,30
211,9	147,54
212,2	147,80

Le **modèle** de **régression linéaire simple** relatif à ce problème s'écrit :

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad i = 1, \ldots, 17.$$

Nous trouvons par la méthode des **moindres carrés** la fonction d'estimation suivante :

$$\hat{Y}_i = -42,131 + 0,895 X_i$$

où $\hat{Y}_i$ est la valeur estimée de la **variable** Y pour un X donné.

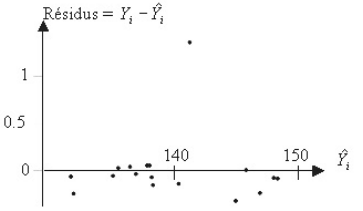
Pour chacune des 17 valeurs de X_i, nous avons donc une valeur estimée $\hat{Y}_i$. Nous pouvons calculer les résidus :

$$e_i = Y_i - \hat{Y}_i.$$

Ces résultats sont présentés dans le tableau suivant :

i	X_i	Y_i	$\hat{Y}_i$	$e_i = Y_i - \hat{Y}_i$
1	194,5	131,79	132,037	−0,247
2	194,3	131,79	131,857	−0,067
3	197,9	135,02	135,081	−0,061
4	198,4	135,55	135,529	0,021
5	199,4	136,46	136,424	0,036
6	199,9	136,83	136,872	−0,042
7	200,9	137,82	137,768	0,052
8	201,1	138,00	137,947	0,053
9	201,4	138,06	138,215	−0,155
10	201,3	138,05	138,126	−0,076
11	203,6	140,04	140,185	−0,145
12	204,6	142,44	141,081	1,359
13	209,5	145,47	145,469	0,001
14	208,6	144,34	144,663	−0,323
15	210,7	146,30	146,543	−0,243
16	211,9	147,54	147,618	−0,078
17	212,2	147,80	147,886	−0,086

Le graphique des résidus en fonction des valeurs estimées $\hat{Y}_i$ nous donne la représentation suivante :



Nous constatons qu'à l'exception d'une seule observation (la douzième) dont la valeur élevée du résidu, en valeur absolue, semble indiquer une observation aberrante, la répartition des résidus est parfaitement uniforme au sein d'une bande horizontale. Les résidus ne fournissent donc dans ce cas aucun élément nous permettant de douter de la validité du **modèle** choisi. Avec l'analyse de différents résidus standardisés que nous pouvons décider si oui ou non la douzième observation est une **observation aberrante** ou un **point levier**.

MOTS CLÉS

Analyse de régression *(Regression analysis)*
Graphique de dispersion *(Scatter plot)*

Moindres carrés *(Least squares)*
Régression linéaire multiple *(Multiple linear regression)*
Régression linéaire simple *(Simple linear regression)*
Résidu *(Residual)*
Test d'Anderson-Darling *(Anderson Darling test)*

RÉFÉRENCES

Anscombe, F.J. (1961). Examination of residuals. Proc. Fourth Berkeley Symp. *Math. Statist. Prob.*, I, pp. 1-36.

Anscombe, F.J. (1973). Graphs in statistical analysis. *The American Statistician*, 27, pp. 17-21.

Anscombe, F.J. and Tukey, J.W. (1963). Analysis of residuals. *Technometrics*, 5, pp. 141-160.

Cook, R.D. and Weisberg, S. (1982). *Residuals and influence in regression.* Chapman & Hall, London.

Cook, R.D. and Weisberg, S. (1994). *An Introduction to Regression Graphics.* John Wiley & Sons, New York.

Cook, R.D. and Weisberg, S. (1999). *Applied Regression Including Computing and Graphics.* John Wiley & Sons, New York.

Draper, N.R. and Smith, H. (1998). *Applied Regression Analysis*, 3rd edition. John Wiley & Sons, New York.

Euler, L. (1749). *Recherches sur la question des inégalités du mouvement de Saturne et de Jupiter, pièce ayant remporté le prix de l'année 1748, par l'Académie royale des sciences de Paris.* Réédité en 1960 in Leonhardi Euleri, *Opera Omnia*, 2e série, 25, pp. 47-157. Turici, Bâle.

Mayer, T. (1750). Abhandlung über die Umwalzung des Monds um seine Axe und die scheinbare Bewegung der Mondflecken. *Kosmographische Nachrichten und Sammlungen auf das Jahr 1748*, 1, pp. 52-183.

ANALYSE EXPLORATOIRE DE DONNÉES

Exploratory data analysis

L'analyse exploratoire de données s'occupe d'isoler les traits et caractéristiques des données et de les révéler à l'analyste. Elle fournit souvent le premier contact avec les données, précédant tout choix de modèles pour des composants structurels ou stochastiques et sert aussi à révéler des déviations des modèles familiers.

HISTORIQUE

Depuis à peu près 1970, l'analyse exploratoire de données signifie l'attitude, l'approche et les techniques développées, principalement par **John Wilder Tukey**, pour examiner les données avant d'utiliser un modèle probabiliste. Selon J.W. Tukey, l'analyse exploratoire de données est un *travail de détective*, un travail à la fois numérique et graphique, car de la même manière qu'un détective investigant un crime a besoin à la fois d'outils et de compréhension, un analyste de données a besoin à la fois d'outils et de compréhension. Le « box-and whisker plot » ou **box plot**, a été introduit par J.W. Tukey en 1972, parallèlement à d'autres méthodes de représentation semi-graphiques de données dont une des plus connues est le **diagramme stem and leaf**. L'origine de ce diagramme est associé à J.W. Tukey (1977). Le concept est basé sur l'histogramme qui date déjà du XVIIIe siècle.

ASPECTS MATHÉMATIQUES

Une méthode propre à l'analyse exploratoire des données pour transmettre l'information essentielle dans une distribution est le *résumé à cinq valeurs*. Il se présente sous la forme d'un tableau, comme présenté ci-dessous :

Médiane	
Premier quartile	Troisième quartile
Extrême inférieur	Extrême supérieur

Soit n le nombre d'observations ; on classe les données rangées par ordre croissant. On définit :

$$\text{rang médiane} = \frac{n+1}{2}$$
$$\text{rang quartile} = \frac{\lfloor \text{rang médiane} \rfloor + 1}{2}$$

où $\lfloor x \rfloor$ est la valeur de x tronquée à l'entier inférieur (c'est-à-dire la partie entière de x).

La **médiane** et les quartiles seront les données correspondantes aux rangs calculés, pour un ensemble de données classées par ordre croissant. Des rangs non entiers signifient que l'on calculera la moyenne entre les deux valeurs les plus proches pour obtenir la médiane ou les quartiles.

Dans son livre, J.W. Tukey appelle les premier et quatrième quartiles *hinges*.

DOMAINES ET LIMITATIONS

Dans l'analyse exploratoire de données telle que J.W. Tukey l'a définie, quatre thèmes principaux apparaissent et se combinent. Ceux-ci sont la **représentation graphique**, la réexpression (qui n'est rien d'autre que la **transformation** des variables), les résidus et la résistance (synonyme de la notion de robustesse). La résistance traite de l'insensibilité au mauvais comportement des données. Plus formellement, une analyse ou un résumé est *résistant(e)* si un changement arbitraire dans n'importe quelle partie des données produit un petit changement dans l'analyse ou le résumé. Cette attention à la résistance reflète le fait que de « bonnes » données ne contiennent que rarement moins de 5 % d'erreurs grossières et une protection contre les effets adverses de celles-ci devrait toujours être disponible. La *resistant line* de Tukey donne un ajustement robuste d'un nuage de points, ce qui veut dire que cette droite ne se laisse pas trop influencer par une observation particulière. La médiane est hautement résistante alors que la moyenne ne l'est pas.

Les représentations graphiques ont pour but l'analyse du comportement des données, des ajustements, des mesures de diagnostic et des résidus et donc de saisir les caractéristiques inattendues ainsi que les régularités familières.

Une contribution majeure dans les développements associés à l'analyse exploratoire de données a été l'accentuation des représentations visuelles et la variété de nouvelles techniques graphiques. Deux de celles-ci sont le stem and leaf et le **box plot**.

EXEMPLES

Résumé à 5 valeurs

Le tableau suivant présente les indices des revenus des cantons de la Suisse par habitant (Suisse = 100) en 1993 :

Indice des revenus des cantons 1993

Canton	Indice	Canton	Indice
Zurich	125, 7	Schaffhouse	99, 2
Berne	86, 2	Appenzell Rh.-Ext.	84, 3
Lucerne	87, 9	Appenzell Rh.-Int.	72, 6
Uri	88, 2	Saint-Gall	89, 3
Schwytz	94, 5	Grisons	92, 4
Obwald	80, 3	Argovie	98, 0
Nidwald	108, 9	Thurgovie	87, 4
Glaris	101, 4	Tessin	87, 4
Zoug	170, 2	Vaud	97, 4
Fribourg	90, 9	Valais	80, 5
Soleure	88, 3	Neuchâtel	87, 3
Bâle-Ville	124, 2	Genève	116, 0
Bâle-Campagne	105, 1	Jura	75, 1

On peut calculer le résumé à 5 valeurs :

$$\text{rang médiane} = \frac{n+1}{2} = \frac{26+1}{2} = 13, 5.$$

Ainsi, la médiane M_d sera la moyenne entre la treizième et la quatorzième observation, c'est-à-dire :

$$M_d = \frac{89,3+90,9}{2} = 90, 1.$$

Ensuite, on calcule :

$$\text{rang quartile} = \frac{13+1}{2} = 7.$$

Le premier quartile sera la septième observation depuis le bas et le troisième quartile la septième depuis le haut, c'est-à-dire :

$$\text{premier quartile} = 87, 3$$
$$\text{troisème quartile} = 101, 4.$$

Les extrêmes inférieur et supérieur sont respectivement $72, 6$ et $170, 2$.

Ainsi, on obtient le résumé à 5 valeurs suivant :

90, 1	
87, 3	101, 4
72, 6	170, 2

Résistance (ou robustesse) de la médiane par rapport à la moyenne

On considère les nombres suivants : 3, 3, 7, 7, 11, 11.

On calcule tout d'abord la moyenne de ces nombres :

$$\bar{x}= \frac{\sum_{i=1}^{n} x_i}{n} = \frac{3+3+7+7+11+11}{6} = 7.$$

Ensuite, on calcule la médiane M_d :

$$\begin{array}{l} \text{rang médiane} = \frac{n+1}{2} = \frac{6+1}{2} = 3,5 \\ M_d = \frac{7+7}{2} = 7. \end{array}$$

On constate que dans ce cas, la moyenne et la médiane valent 7.

On suppose maintenant qu'on ajoute un nombre : on ajoute $-1\,000$. On va recalculer la moyenne et la médiane des nombres suivants : $-1\,000$, 3, 3, 3, 7, 11, 11. La moyenne est :

$$\begin{aligned} \overline{x} &= \frac{\sum_{i=1}^{n} x_i}{n} \\ &= \frac{-1\,000+3+3+7+7+11+11}{7} \\ &= -136,86. \end{aligned}$$

On voit qu'en ajoutant un seul nombre, la moyenne a varié sensiblement.

La médiane est :

$$\begin{array}{l} \text{rang médiane} = \frac{n+1}{2} = \frac{7+1}{2} = 4 \\ M_d = 7. \end{array}$$

Contrairement à la moyenne, la médiane n'a pas changé, ce qui permet d'affirmer que la médiane est plus résistante aux valeurs extrêmes que la moyenne.

MOTS CLÉS

Box plot *(Box plot)*
Représentation graphique *(Graphical representation)*
Résidu *(Residual)*
Diagramme Stem and leaf *(Stem and leaf diagram)*
Transformation *(Transformation)*

RÉFÉRENCES

Tukey, J.W. (1972). Some graphical and Semigraphical Displays. In *Statistical Papers in Honor of George W. Snedecor*, ed. T.A. Bancroft, Arnes: The Iowa State University Press. Iowa, U.S.A, pp. 293-316.

Tukey, J.W. (1977). *Exploratory data analysis.* Addison-Wesley, Reading, Mass.

ANALYSE FACTORIELLE DES CORRESPONDANCES

Correspondence analysis

L'analyse factorielle des correspondances est une méthode d'**analyse de données** permettant la description de tables de contingence (ou tableaux croisés). Cette analyse se fait sous forme de **représentation graphique** des associations entre lignes et colonnes.

HISTORIQUE

Les principes théoriques de l'analyse factorielle des correspondances remontent aux travaux de Hermann Otto Hartley (1935) (publiés sous son nom original Hirschfeld) et de **Ronald Aylmer Fisher** (1940) sur les tables de contingence. Cette première présentation se situait dans un cadre de statistique inférentielle classique. Le terme d'analyse factorielle des correspondances apparut en automne 1962 et le premier exposé, sous ce nom, de la méthode fut donné par Jean-Paul Benzécri en hiver 1963. Depuis la parution en 1976 des ouvrages de J.-P. Benzécri retraçant les travaux de son laboratoire au cours des douze années précédentes, on utilise surtout les propriétés algébriques et géométriques de l'outil descriptif constituant l'analyse proprement dite.

ASPECTS MATHÉMATIQUES

Soit une **table de contingence** relative à deux variables qualitatives catégorielles X et Y possédant respectivement r et c catégories :

	Y_1	Y_2	...	Y_c	Total
X_1	n_{11}	n_{12}	...	n_{1c}	$n_{1.}$
X_2	n_{21}	n_{22}	...	n_{2c}	$n_{2.}$
...	...	...	...	...	...
X_r	n_{r1}	n_{r2}	...	n_{rc}	$n_{r.}$
Total	$n_{.1}$	$n_{.2}$	...	$n_{.c}$	$n_{..}$

où

- n_{ij} représente la **fréquence** observée pour la catégorie i de la **variable** X et la catégorie j de la variable Y;
- $n_{i.}$ représente la somme des fréquences observées pour la catégorie i de la variable X;
- $n_{.j}$ représente la somme des fréquences observées pour la catégorie j de la variable Y;
- $n_{..}$ représente le nombre total d'observations.

On suppose $r \geq c$, sinon on prend la transposée du tableau initial comme nouvelle table de contingence. Pour effectuer l'analyse factorielle des correspondances d'une table de contingence ayant un nombre de lignes supérieur à celui des colonnes, on procède selon les étapes suivantes :

1. Calcul du tableau des *profils-lignes* et des *profils-colonnes* désignés respectivement par X_I et X_J.

   Pour une ligne (colonne) fixée, on appelle profil-ligne (profil-colonne) la ligne (colonne) obtenue en divisant chaque élément de cette dernière par la somme de la ligne (colonne) considérée.

   Le profil-ligne de la ligne i s'obtient en divisant chaque terme de la ligne i par $n_{i.}$ qui n'est autre que la somme des fréquences observées pour la catégorie correspondant à cette ligne.

   On peut construire le tableau des profils-lignes en remplaçant chaque ligne de la table de contingence par son profil :

	Y_1	Y_2	...	Y_c	Total
X_1	$\frac{n_{11}}{n_{1.}}$	$\frac{n_{12}}{n_{1.}}$	...	$\frac{n_{1c}}{n_{1.}}$	1
X_2	$\frac{n_{21}}{n_{2.}}$	$\frac{n_{22}}{n_{2.}}$	...	$\frac{n_{2c}}{n_{2.}}$	1
...	...	...	...	...	...
X_r	$\frac{n_{r1}}{n_{r}}$	$\frac{n_{r2}}{n_{r}}$	...	$\frac{n_{rc}}{n_{r}}$	1
Total	$n'_{.1}$	$n'_{.2}$	...	$n'_{.c}$	r

   Il arrive que l'on multiplie encore chaque élément du tableau par 100 pour avoir les termes exprimés en pourcentage et ainsi la somme de chacune des c lignes vaudra 100 %.

   On construit le tableau des profils-colonnes de la même manière en remplaçant cette fois chaque colonne de la table de contingence par son profil : le profil-colonne de la colonne j s'obtient en divisant chaque terme de la colonne j par $n_{.j}$ qui n'est autre que la somme des fréquences observées pour la catégorie correspondant à cette colonne.

	Y_1	Y_2	...	Y_c	Total
X_1	$\frac{n_{11}}{n_{.1}}$	$\frac{n_{12}}{n_{.2}}$	...	$\frac{n_{1c}}{n_{.c}}$	$n'_{1.}$
X_2	$\frac{n_{21}}{n_{.1}}$	$\frac{n_{22}}{n_{.2}}$	...	$\frac{n_{2c}}{n_{.c}}$	$n'_{2.}$
...	...	...	...	...	...
X_r	$\frac{n_{r1}}{n_{.1}}$	$\frac{n_{r2}}{n_{.2}}$	...	$\frac{n_{rc}}{n_{.c}}$	$n'_{r.}$
Total	1	1	...	1	c

   Les tableaux des profils-lignes et profils-colonnes correspondent à une transformation de la table de contingence pour rendre les lignes et les colonnes comparables entre elles.

2. Détermination de la **matrice d'inertie** V.

   Pour cela, on procède de la façon suivante :

   - on calcule la moyenne pondérée des r points-lignes, c'est-à-dire :

     $$g_j = \sum_{i=1}^{r} \frac{n_{i.}}{n_{..}} \cdot \frac{n_{ij}}{n_{i.}} = \frac{n_{.j}}{n_{..}}, \quad j = 1, \ldots, c \ ;$$

     on écrit ensuite r fois en lignes les c valeurs obtenues g_j dans une matrice G ;

   - on construit une matrice diagonale D_I dont les éléments diagonaux sont: $\frac{n_{i.}}{n_{..}}$;

- finalement la matrice d'inertie V se calcule par la formule :

$$V = (X_I - G)' \cdot D_I \cdot (X_I - G).$$

3. À l'aide de la matrice M composée des termes $\frac{n_{..}}{n_{.j}}$ sur sa diagonale et de « 0 » ailleurs, on détermine la matrice C :

$$C = \sqrt{M} \cdot V \cdot \sqrt{M}$$

4. Recherche des valeurs propres notées k_l et des vecteurs propres notés v_l de cette matrice C.

   Les c valeurs propres $k_c,\ k_{c-1}, \ldots, k_1$ écrites dans un ordre décroissant sont les inerties expliquées. Les vecteurs propres correspondants sont appelés axes factoriels (ou axes d'inertie).

   Pour chaque valeur propre, on calcule l'inertie expliquée par l'axe factoriel correspondant. Par exemple, on dira que le premier axe factoriel explique :

$$\frac{100 \cdot k_1}{\sum_{l=1}^{c} k_l} (\text{en } \%) \text{ d'inertie.}$$

   De même, les deux premiers axes factoriels, par exemple, expliquent :

$$\frac{100 \cdot (k_1 + k_2)}{\sum_{l=1}^{c} k_l} \quad (\text{en } \%) \text{ d'inertie.}$$

   Si l'on veut par exemple connaître le nombre de valeurs propres et donc d'axes factoriels qu'il faut considérer pour expliquer au moins les trois quarts de l'inertie totale, on cumule l'inertie expliquée de chacune des valeurs propres jusqu'à obtenir 75 %.

   À partir des axes factoriels ainsi obtenus, on calcule les axes principaux d'inertie.

5. Les axes principaux d'inertie, notés u_l, sont alors donnés par :

$$u_l = \sqrt{M^{-1}} \cdot v_l$$

   c'est-à-dire que sa j-ème composante vaut :

$$u_{jl} = \sqrt{\frac{n_{.j}}{n_{..}}} \cdot v_{jl}$$

6. On calcule alors les composantes principales, notées y_k qui ne sont autres que les projections orthogonales des points-lignes sur les axes principaux d'inertie : la i-ème coordonnée de la l-ème composante principale prend la **valeur** suivante :

$$y_{il} = x_i \cdot M \cdot u_l$$

   c'est-à-dire que

$$y_{il} = \sum_{j=1}^{c} \frac{n_{ij}}{n_{i.}} \sqrt{\frac{n_{..}}{n_{.j}}} \cdot v_{jl}$$

   est la coordonnée du point-ligne i sur le l-ème axe.

7. Après avoir calculé les composantes principales y_l (des points-lignes) on peut déterminer celles des points-colonnes notées z_l à l'aide des y_l grâce aux formules de transition :

$$z_{jl} = \frac{1}{\sqrt{k_l}} \sum_{i=1}^{r} \frac{n_{ij}}{n_{.j}} \cdot y_{il}, \quad \text{pour } j = 1, 2, \ldots, c$$

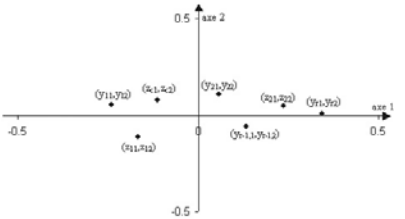
$$y_{il} = \frac{1}{\sqrt{k_l}} \sum_{i=1}^{c} \frac{n_{ij}}{n_{i.}} \cdot z_{jl}, \quad \text{pour } i = 1, 2, \ldots, r$$

   z_{jl} est la coordonnée du point-colonne j sur le l-ème axe factoriel.

8. La représentation simultanée des points-lignes et des points-colonnes sur un **graphique de dispersion** par rapport à deux axes factoriels permet de donner une signification de l'axe en fonction des points qui s'y opposent.

   La qualité de la représentation est liée à la part d'inertie expliquée par les deux axes principaux

utilisés. Cette dernière est d'autant meilleure que l'inertie expliquée se rapproche de 1.



9. Il peut être intéressant de situer des points-lignes ou points-colonnes supplémentaires (variables illustratives) par rapport aux variables actives utilisées pour l'analyse factorielle des correspondances.

   Soit par exemple un point-ligne supplémentaire :

   $$(n_{s1}, n_{s2}, n_{s3}, \ldots, n_{sc}),$$

   de profil-ligne

   $$\left(\frac{n_{s1}}{n_{s.}}, \frac{n_{s2}}{n_{s.}}, \ldots, \frac{n_{sc}}{n_{s.}}\right)$$

   la l-ème composante principale du point-ligne supplémentaire est donnée par la deuxième formule de 7., à savoir:

   $$y_{il} = \frac{1}{\sqrt{k_l}} \sum_{j=1}^{c} \frac{n_{sj}}{n_{s.}} \cdot z_{jl}, \quad \text{pour } i = 1, 2, \ldots, r$$

   On procéderait de même pour un point-colonne en appliquant cette fois-ci la première formule de 7.

DOMAINES ET LIMITATIONS

Lors de la représentation simultanée des points-lignes et des points-colonnes, il est déconseillé d'interpréter d'éventuelles proximités croisées entre lignes et colonnes, car les deux points ne sont pas dans le même espace de départ. Il est par contre intéressant d'interpréter la position d'un point-ligne par rapport à l'ensemble des points-colonnes (et vice-versa).

EXEMPLE

On prend l'exemple d'une entreprise qui par l'intermédiaire d'un statisticien souhaite connaître l'état de santé de son personnel. Un des sujets du questionnaire concerne le nombre de visites annuelles chez un médecin (les dentistes y compris). Trois possibilités de réponses sont proposées :

— entre 0 et 6 visites,

— entre 7 et 12 visites,

— plus de 12 visites par an.

Le personnel interrogé est réparti en 5 catégories :

— les cadres âgés de plus de 40 ans,

— les cadres âgés de moins de 40 ans,

— les employés de plus de 40 ans,

— les employés de moins de 40 ans,

— le personnel de bureau.

Les résultats sont rapportés dans la **table de contingence** suivante, à laquelle on a rajouté les marges :

	Nombre de visites annuelles			
	0 à 6	7 à 12	> 12	Total
Cadres > 40 ans	5	7	3	15
Cadres < 40 ans	5	5	2	12
Employés > 40 ans	2	12	6	20
Employés < 40 ans	24	18	12	54
Personnel de bureau	12	8	4	24
Total	48	50	27	125

À l'aide de ces données, on présente, en huit étapes principales, la démarche à suivre pour obtenir une **représentation graphique** générée par l'analyse factorielle des correspondances.

1. On détermine tout d'abord, d'une part, le tableau X_I des profils-lignes, en divisant chaque élément par le total de la ligne sur laquelle il est situé :

$0,333$	$0,467$	$0,2$	1
$0,417$	$0,417$	$0,167$	1
$0,1$	$0,6$	$0,3$	1
$0,444$	$0,333$	$0,222$	1
$0,5$	$0,333$	$0,167$	1
$1,794$	$2,15$	$1,056$	5

et, d'autre part, celui des profils-colonnes, en divisant chaque élément par la composante de la marge de la colonne correspondante :

$0,104$	$0,14$	$0,111$	$0,355$
$0,104$	$0,1$	$0,074$	$0,278$
$0,042$	$0,24$	$0,222$	$0,504$
$0,5$	$0,36$	$0,444$	$1,304$
$0,25$	$0,16$	$0,148$	$0,558$
1	1	1	3

2. On calcule ensuite la **matrice d'inertie** V pour les **fréquences** données respectivement par $\frac{n_{i.}}{n_{..}}$, pour i allant de 1 à 5, en procédant de la manière suivante :

   - on commence par calculer la **moyenne** pondérée des 5 points-lignes, c'est-à-dire :

   $$\begin{aligned} g_j &= \sum_{i=1}^{5} \frac{n_{i.}}{n_{..}} \cdot \frac{n_{ij}}{n_{i.}} \\ &= \frac{n_{.j}}{n_{..}} \quad \text{pour } j = 1,\ 2 \text{ et } 3 \ ; \end{aligned}$$

   - ensuite, on écrit ces 3 valeurs 5 fois en lignes dans une matrice G, ce qui donne :

   $$G = \begin{bmatrix} 0,384 & 0,400 & 0,216 \\ 0,384 & 0,400 & 0,216 \\ 0,384 & 0,400 & 0,216 \\ 0,384 & 0,400 & 0,216 \\ 0,384 & 0,400 & 0,216 \end{bmatrix}$$

   - on construit une **matrice** diagonale D_I à l'aide des $\frac{n_{i.}}{n_{..}}$;
   - et finalement, on est en mesure de calculer la matrice d'inertie V donnée par la formule suivante :

   $$V = (X_I - G)' \cdot D_I \cdot (X_I - G)$$

   qui vaut dans notre cas :

   $$V = \begin{bmatrix} 0,0175 & -0,0127 & -0,0048 \\ -0,0127 & 0,0097 & 0,0029 \\ -0,0048 & 0,0029 & 0,0019 \end{bmatrix}$$

3. On définit la matrice carrée M, d'ordre 3, qui porte les éléments $\frac{n_{..}}{n_{.j}}$ sur sa diagonale et des « 0 » ailleurs :

   $$M = \begin{bmatrix} 2,604 & 0 & 0 \\ 0 & 2,5 & 0 \\ 0 & 0 & 4,630 \end{bmatrix}$$

   La racine carrée de M, qu'on notera $\sqrt{M}$, s'obtient en prenant la racine carrée de chaque élément diagonal de M. À l'aide de cette dernière matrice, on détermine :

   $$C = \sqrt{M} \cdot V \cdot \sqrt{M}$$

   $$C = \begin{bmatrix} 0,0455 & -0,0323 & -0,0167 \\ -0,0323 & 0,0243 & 0,0100 \\ -0,0167 & 0,0100 & 0,0087 \end{bmatrix}$$

4. Les valeurs propres de C s'obtiennent par le calcul de diagonalisation de la matrice. En rangeant ces dernières par ordre décroissant, il vient :

   $$\begin{aligned} k_1 &= 0,0746 \\ k_2 &= 0,0039 \\ k_3 &= 0. \end{aligned}$$

   Pour chacune d'entre elles, l'inertie expliquée peut être déterminée. C'est ainsi que l'on a obtenu pour :

   $$k_1 : \frac{0,0746}{0,0746 + 0,0039} = 95,03\ \%,$$

   $$k_2 : \frac{0,0039}{0,0746 + 0,0039} = 4,97\ \%$$

   tandis que pour la dernière, k_3, l'inertie expliquée est nulle.

   Les deux premiers axes factoriels associés auxvaleurs propres k_1 et k_2 expliquent toute

l'inertie. Comme la troisième valeur propre est nulle il n'est pas nécessaire de calculer le **vecteur propre** qui lui est associé. Par contre, les deux premiers vecteurs propres normalisés sont :

$$v_1 = \begin{bmatrix} 0,7807 \\ -0,5576 \\ -0,2821 \end{bmatrix} \text{ et } v_2 = \begin{bmatrix} 0,0807 \\ 0,5337 \\ -0,8393 \end{bmatrix}$$

5. On détermine alors les axes principaux d'inertie par :

$$u_i = \sqrt{M^{-1}} \cdot v_i \text{ pour } i = 1 \text{ et } 2$$

où $\sqrt{M^{-1}}$ s'obtient en inversant les éléments diagonaux de $\sqrt{M}$.

On trouve :

$$u_1 = \begin{bmatrix} 0,4838 \\ -0,3527 \\ -0,1311 \end{bmatrix} \text{ et } v_2 = \begin{bmatrix} 0,0500 \\ 0,3401 \\ -0,3901 \end{bmatrix}$$

6. On calcule alors les composantes principales en projetant les points-lignes sur les axes principaux d'inertie. On construit une matrice auxiliaire U formée par les deux vecteurs u_1 et u_2, on pose :

$$Y = X_I \cdot M \cdot U = \begin{bmatrix} -0,1129 & 0,0790 \\ 0,0564 & 0,1075 \\ -0,5851 & 0,0186 \\ 0,1311 & -0,0600 \\ 0,2349 & 0,0475 \end{bmatrix}$$

On lit dans cette matrice Y, les coordonnées des 5 points-lignes écrites horizontalement et telles que la première colonne indique la composante suivant le premier axe factoriel et la deuxième celle suivant le second axe.

7. À partir des coordonnées des points-lignes, on peut trouver celles des points-colonnes à l'aide des formules de transition suivantes, écrites sous forme matricielle :

$$Z = K \cdot Y' \cdot X_J$$

où

- K est la matrice diagonale d'ordre 2 composée des $1/\sqrt{k_i}$ sur sa diagonale et de « 0 » ailleurs,
- Y' est la **transposée** de la matrice des coordonnées des points-lignes, et
- X_J est la matrice des profils-colonnes.

On obtient alors la matrice :

$$Z = \begin{bmatrix} 0,3442 & -0,2409 & -0,1658 \\ 0,0081 & 0,0531 & -0,1128 \end{bmatrix}$$

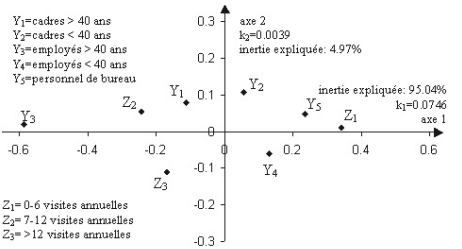
qui contient cette fois-ci en colonne, les composantes des trois points-colonnes, de sorte que la première ligne représente la coordonnée selon le premier axe factoriel et la deuxième ligne, la coordonnée selon le second axe.

On peut vérifier la formule de transition permettant de retrouver Y à partir de Z. En effet :

$$Y = X_I \cdot Z' \cdot K$$

avec les mêmes notations que précédemment.

8. On est alors en mesure de représenter les cinq catégories de personnes interrogées et les trois catégories de réponses proposées sur un même plan factoriel :



La lecture du plan factoriel ci-dessus peut s'effectuer à trois niveaux d'analyse différents selon que l'on considère :

- l'ensemble des catégories relatives au personnel,

- l'ensemble des modalités concernant les visites médicales,
- ou bien les deux ensembles à la fois.

Ainsi, dans le premier cas, la proximité entre deux points-lignes, c'est-à-dire entre deux catégories de personnel, signifie une similitude des profils du point de vue des visites chez le médecin. Tel est le cas, sur le plan factoriel, des « employés qui ont moins de 40 ans » (Y_4) et du « personnel de bureau » (Y_5). On peut vérifier, à partir du tableau des profils-lignes, que les pourcentages pour ces deux catégories sont en effet très voisins.

De même, la proximité entre deux points-colonnes (représentant deux catégories relatives au nombre de visites médicales) indique une similitude des profils par rapport à la répartition du personnel de l'entreprise. C'est ce qui se produit pour les modalités Z_2 (de 7 à 12 visites annuelles) et Z_3 (plus de 12 visites par an).

Si l'on considère maintenant les points-lignes et les points-colonnes simultanément (et non plus séparément comme précédemment), il sera possible d'identifier les modalités qui sont responsables de certaines proximités. Par exemple, « les employés âgés de moins de 40 ans » (Y_4) et « le personnel de bureau » (Y_5) semblent avoir le même comportement face à la santé : ils se rendent chez le médecin moins de 6 fois par an (Z_1) avec une forte proportion ($0,44$ et $0,5$ respectivement).

En conclusion, l'axe 1 oppose d'une part les catégories indiquant un nombre de visites moyen ou élevé (Z_2 et Z_3) concernant les employés ou les cadres de plus de 40 ans (Y_1 et Y_3), et d'autre part la modalité : nombre de visites faible (Z_1) avec les autres catégories de personnel (Y_2, Y_4 et Y_5). Le premier facteur peut donc s'interpréter comme l'importance du contrôle médical selon l'âge.

MOTS CLÉS

Analyse de données *(Data analysis)*
Axe factoriel *(Factorial axis)*
Graphique de dispersion *(Scatter plot)*
Matrice d'inertie *(Inertia matrix)*
Représentation graphique *(Graphical representation)*
Table de contingence *(Contingency table)*
Valeur propre *(Eigenvalue)*
Vecteur propre *(Eigenvector)*

RÉFÉRENCES

Benzécri, J.-P. (1976). *L'Analyse des données*, Tome 1 : *La Taxinomie* ; Tome 2 : *L'Analyse factorielle des correspondances.* Dunod, Paris.

Benzécri, J.-P. (1976). Histoire et préhistoire de l'analyse des données, *Les Cahiers de l'analyse des données 1*, n$^{os.}$ 1–4. Dunod, Paris.

Fisher, R.A. (1940). The precision of discriminant Functions.*Annals of Eugenics* (London), 10, pp. 422-429.

Hirschfeld, H.O. (1935). A connection between correlation and contingency, *Proceedings of the Cambridge Philosophical Society*, 31, pp. 520-524.

Saporta, G. (1990). *Probabilité, analyse de données et statistiques*, Technip, Paris.

ANALYSE MULTIVARIÉE

Multivariate analysis

Voir **analyse de données**, **analyse factorielle des correspondances** et **classification automatique**.

ANDERSON-DARLING, TEST D'

Anderson-Darling test

Voir **test d'Anderson-Darling**.

ANDERSON OSKAR

Anderson Oskar

Oskar Anderson (1887-1960) fut un membre important de la Continental School of Statistics ; sa contribution toucha une large étendue de sujets incluant la corrélation, l'analyse de séries temporelles (tôt dans ses recherches, il developpa, indépendamment mais de manière contemporaine à William Sealy Gosset (« Student »), la méthode des différences de Tintner pour analyser les séries temporelles), les méthodes non paramétriques et les enquêtes par sondage, aussi bien que l'économétrique et différentes autres applications aux sciences sociales.



Oskar Anderson obtint son baccalauréat avec mention au lycée de Kazan et étudia durant une année les mathématiques et la physique à l'Université de Kazan. Il entra à la faculté d'économie de l'Institut polytechnique de Saint-Pétersbourg pour y étudier les mathématiques, les statistiques et l'économie. C'est à la fin de 1912 qu'il devint « candidat d'économie » pour un mémoire sur le coefficient de corrélation et son application aux séries chronologiques. Peu après, il obtint un diplôme de droit à l'Université de Saint-Pétersbourg.

Une des premières recherches d'économétrie est abordée dans l'article *Ist die Qualitätstheorie des eldes statistisch nachweisbar ?* dans lequel Anderson traita le problème de la manière empirique de formuler des relations théoriques.

Les publications d'O. Anderson allient la tradition de la Continental School of Statistics avec les concepts de l'English Biometric School ; en particulier dans deux de ses ouvrages : *Einführung in die mathematische Statistik* et *Probleme des statistischen Methodenlehre in den Sozialwissenschaften*.

En 1949, il fonda avec Hans Kellerer et Hans Münzner le journal *Mitteilungsblatt für mathematische Statistik*.

Quelques ouvrages et articles principaux d'Oskar Anderson :

1935 *Einführung in die mathematische Statistik.* Wien : Julius Springer.

1954 *Probleme der statistischen Methodenlehre in den Sozialwissenschaften.* Würzberg : Physica-Verlag.

ANDERSON THEODORE W.

Anderson Theodore W.

Theodore Wilbur Anderson est né le 5 juin 1918 à Minneapolis (Minnesota) aux États-Unis. Docteur en mathématiques en 1945 à l'Université de Princeton, il devient en 1946 membre du département des statistiques mathématiques de l'Université de Colombie où ensuite il est nommé professeur en 1956. En 1967, il est nommé professeur de statistique et économie à l'Université de Stanford. Il est succes-

sivement membre de Guggenheim entre 1947-1948, éditeur des *Annals of Mathematical* Statistics de 1950 à 1952, président de l'Institut of Mathematical Statistics en 1963 et vice-président de l'American Statistical Association de 1971 à 1973. Il est membre de l'American Academy of Arts and Sciences, de la National Academy of Sciences, de l'Institut of Mathematical Statistics et de la Royal Statistical Society. La contribution la plus importante d'Anderson dans la statistique est certainement dans le domaine de l'**analyse multivariée**. En 1958, il publie le livre *An introduction to multivariate statistical analysis*, livre qui a été traduit en russe et fut la référence durant plus de quarante ans. Son nom est aussi lié au **test d'Anderson-Darling** (1952).



Depuis le début des années 80, il a écrit des livres et des publications dans le domaine de l'analyse statistique des données. Il a été sensible aux développements des outils informatiques, domaine dans lequel il a apporté sa contribution en écrivant un guide de l'analyse des données avec le logiciel MINITAB.

Quelques ouvrages et articles principaux de Theodore Wilbur Anderson :

1952 (and Darling D.A.) Asymptotic Theory of Certain Goodness of Fit Criteria Based on Stochastic Processes. *Annals of Mathematical Statistics*, 23, pp. 193-212.

1958 *An introduction to multivariate statistical analysis.* John Wiley & Sons, New York.

1971 *The Statistical Analysis of Time Series.* John Wiley & Sons, New York.

1989 Linear Latent Variable Models and Covariance Structures. *Journal of Econometrics*, 41, pp. 91-119.

1992 (and Kunitoma, N.) Asymptotic Distributions of Regression and Autoregression Coefficients with Martingale Difference Disturbances. *Journal of Multivariate Analysis*, 40, pp. 221-243.

1993 Goodness of Fit Tests for Spectral Distributions. *Annals of Statistics*, 21, pp. 830-847.

MOT CLÉ

Test d'Anderson-Darling *(Anderson-Darling test)*

ARRANGEMENT

Arrangement

Les arrangements sont une notion d'**analyse combinatoire**.

Le nombre d'arrangements est le nombre de manières de tirer k objets pris parmi n en tenant compte de l'ordre dans lequel les objets sont pris (par opposition aux combinaisons).

HISTORIQUE

Voir **analyse combinatoire**.

ASPECTS MATHÉMATIQUES

1. *Arrangements sans répétition*

   On parle d'arrangements sans répétition (ou sans remise) lorsque chaque objet tiré n'est pas remis à disposition pour le tirage suivant. Chaque objet ne peut donc intervenir qu'une seule fois parmi les k tirages.

Le nombre d'arrangements de k objets parmi n sans répétition est égal à :

$$A_n^k = \frac{n!}{(n-k)!}$$

2. *Arrangements avec répétitions*

   On parle d'arrangements avec répétitions (ou avec remise) lorsque chaque objet tiré est remis à disposition pour le tirage suivant. Chaque objet peut donc intervenir jusqu'à k fois parmi les k tirages.

   Le nombre d'arrangements de k objets parmi n avec répétitions est égal à :

$$A_n^k = n^k.$$

EXEMPLES

1. *Arrangements sans répétition*

   On considère une urne contenant 6 boules numérotées de 1 à 6. On tire successivement 4 boules de l'urne et on désire savoir combien de nombres il est possible de former avec les numéros des 4 boules sortantes.

   On veut connaître le nombre d'arrangements (puisque l'on tient compte de l'ordre des boules) sans répétition (puisque chaque boule ne peut être tirée qu'une seule fois) de 4 objets parmi 6.

   On a :

$$A_n^k = \frac{n!}{(n-k)!} = \frac{6!}{(6-4)!} = 360$$

   arrangements possibles. Il est donc possible de former 360 nombres différents de 4 chiffres avec les chiffres $1, 2, 3, 4, 5, 6$ lorsque chaque chiffre ne peut apparaître qu'une seule fois dans le nombre formé.

   Comme deuxième exemple, formons un arrangement sans répétition de deux lettres avec les trois lettres ABC. Avec $n = 3$ et $k = 2$, on trouve :

$$A_n^k = \frac{n!}{(n-k)!} = \frac{3!}{(3-2)!} = 6$$

   et, on trouve bien 6 arrangements différents sans répétitions :

$$AB, AC, BA, BC, CA, CB.$$

2. *Arrangements avec répétitions*

   On reprend la même urne, mais en effectuant à présent 4 tirages successifs en remettant à chaque fois la boule tirée dans l'urne.

   Le problème est de savoir combien de nombres (ou d'arrangements) possibles peuvent résulter de ce nouveau tirage.

   Dans ce cas, on cherche le nombre d'arrangements avec répétitions (puisque chaque boule tirée est remise dans l'urne avant le tirage suivant). Ce qui donne d'après les formules vues ci-dessus :

$$A_n^k = n^k = 6^4 = 1\,296$$

   arrangements différents. Il est donc possible de former 1296 nombres de 4 chiffres avec les chiffres $1, 2, 3, 4, 5, 6$ lorsque chaque chiffre peut apparaître plusieurs fois dans le nombre formé.

   Comme deuxième exemple, on considère à nouveau les 3 lettres ABC et on veut former un arrangement de deux lettres avec répétitions. Avec $n = 3$ et $k = 2$, on obtient :

$$A_n^k = n^k = 3^2 = 9$$

   Ce qui correspond bien à ce qu'on trouve :

$$AA, AB, AC, BA, BB, BC, CA, CB, CC.$$

MOTS CLÉS

Analyse combinatoire *(Combinatory analysis)*
Combinaison *(Combination)*
Permutation *(Permutation)*

RÉFÉRENCES

Voir **analyse combinatoire**.

AUTOCORRÉLATION ET AUTOCORRÉLATION PARTIELLE

Autocorrelation and partial autocorrelation

L'autocorrélation, notée ρ_k, est une mesure d'une corrélation de la **série chronologique** avec elle-même décalée de k périodes. Elle est obtenue en divisant la **covariance** entre deux observations décalées de k, d'une série chronologique (autocovariance) par l'**écart type** de y_t et y_{t-k}. La suite d'autocorrélations calculées pour des valeurs entières de l'espacement entre les observations est appelée *fonction d'autocorrélation*. Pour des séries chronologiques stationnaires, la fonction d'autocorrélation décroît exponentiellement vers 0.

HISTORIQUE

C'est à **George Yule**, dans les années 1920-1930, que l'on doit les premières recherches sur l'autocorrélation, l'autocorrélation partielle et le corrélogramme lorsqu'il développa la théorie des processus autorégressifs.

ASPECTS MATHÉMATIQUES

Autocorrélation

On définit l'autocorrélation d'une série chronologique Y_t par :

$$\begin{aligned}\rho_k &= \frac{cov\,(y_t, y_{t-k})}{\sigma_{y_t}\sigma_{y_{t-k}}} \\ &= \frac{\sum_{t=k+1}^{T}(y_t-\overline{y})\,(y_{t-k}-\overline{y})}{\sqrt{\sum_{t=k+1}^{T}(y_t-\overline{y})^2}\sqrt{\sum_{t=k+1}^{T}(y_{t-k}-\overline{y})^2}}\end{aligned}$$

avec $\overline{y}$ moyenne de la série calculée sur $T-k$ périodes, T désignant le nombre d'observations.

On déduit que :

$$\begin{aligned}\rho_0 &= 1 \text{ et} \\ \rho_k &= \rho_{-k}.\end{aligned}$$

Une estimation de l'autocorrélation (notée $\widehat{\rho}_k$) pour un nombre d'observations suffisamment grand ($T > 30$) est donnée par la formule suivante :

$$\widehat{\rho}_k = \frac{\sum_{t=k+1}^{T}(y_t-\overline{y})\,(y_{t-k}-\overline{y})}{\sum_{t=1}^{T}(y_t-\overline{y})^2}.$$

Autocorrélation partielle

La fonction d'autocorrélation partielle de retard k est définie comme le coefficient de corrélation partielle entre y_t et y_{t-k}, l'influence des autres variables décalée de k périodes $(y_{t-1}, y_{t-2}, \ldots, y_{t-k+1})$ ayant été retirée.

Test d'hypothèse

Lors d'une analyse de la fonction d'autocorrélation d'une série chronologique, il peut s'avérer utile de savoir quels sont les termes ρ_k significativement différents de 0. Le test d'hypothèse est le suivant :

$$\begin{aligned}H_0 &: \rho_k = 0 \\ H_1 &: \rho_k \neq 0.\end{aligned}$$

Pour un échantillon de grande taille ($T > 30$) le coefficient ρ_k tend de manière asymptotique vers une **loi normale** de moyenne 0 et d'écart type $\frac{1}{\sqrt{T}}$. Le **test de Student** est fondé sur la comparaison d'un t empirique et théorique.

L'intervalle de confiance du coefficient ρ_k est donné par :

$$\rho_k = 0 \pm t_{\alpha/2}\frac{1}{\sqrt{T}}.$$

Si le coefficient calculé $\widehat{\rho}_k$ est à l'extérieur de cet intervalle de confiance, il est significativement différent de 0 au seuil α (en général $\alpha = 0,05$ et $t_{\alpha/2} = 1,96$).

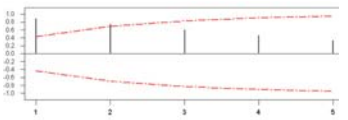
DOMAINES ET LIMITATIONS

Les fonctions d'autocorrélation et d'autocorrélation partielle sont essentiellement utilisées dans l'étude des séries chronologiques et plus précisement lorsqu'on veut ajuster un modèle ARMA. Ces fonctions peuvent être aussi rencontrées dans la **statistique spatiale** et on parle alors d'autocorrélation spatiale, si l'on observe une corrélation d'une variable avec elle-même dans l'espace. Si la présence d'une qualité dans une partie d'un territoire rend sa présence dans les zones voisines plus ou moins probable, le phénomène montre une autocorrélation spatiale ; autocorrélation positive : des régions voisines tendent à avoir des propriétés identiques ou des valeurs semblables (ex : régions homogènes, gradients réguliers) ; autocorrélation négative : des régions voisines ont des qualités différentes, ou bien les valeurs fortes et faibles alternent. Les mesures de l'autocorrélation dépendent de l'échelle d'analyse, du niveau de résolution de la grille à travers laquelle une distribution est observée.

EXEMPLES

On prend comme exemple le revenu national par habitant en Suisse de 1950 à 1994 mesuré tous les deux ans.

On calcule la fonction d'autocorrélation entre les données : on pense obtenir une autocorrélation positive entre les donneés. La figure suivante montre la présence de cette autocorrélation.



On remarque des pics significatifs pour les corrélations entre l'observation au temps t et l'observation au temps $t-1$, ainsi qu'entre l'observation au temps t et l'observation au temps $t-2$. Cette configuration des données est typique pour un processus autorégressif. Pour les deux premières valeurs, on remarque que cette autocorrélation est significative, car la statistique t de Student pour les $T = 23$ observations donne :

$$\rho_k = 0 \pm 1,96\frac{1}{\sqrt{23}}.$$

Année	Revenu national par habitant
50	11 999
52	12 491
54	13 696
56	15 519
58	17 128
60	19 948
62	23 362
64	26 454
66	28 231
68	30 332
70	33 955
72	40 320
74	40 839
76	37 846
78	39 507
80	41 180
82	42 108
84	44 095
86	48 323
88	51 584
90	55 480
92	54 348
94	54 316

Source : Office fédéral de la statistique suisse.

MOTS CLÉS

Série chronologique *(Time series)*
Test de Student *(Student test)*

RÉFÉRENCES

Bourbonnais, R. (1998). *Économétrie, manuel et exercices corrigés.* 2e édition. Dunod, Paris.

Box, G.E.P. and Jenkins, G.M. (1970). *Time series analysis: Forecasting and control, Holden-Day Series in Time Series Analysis*, San Francisco.

Chatfield, C. (1989). *The Analysis of Time Series : An Introduction.* 4th ed. Chapman & Hall.

AXE FACTORIEL

Factorial axis

Soit S une matrice carrée d'ordre n, si X_i est le **vecteur propre** de S associé à la **valeur propre** k_i, on dit que X_i est le i-ème axe factoriel ; plus précisément il s'agit de l'axe portant le vecteur X_i.

On classe généralement les valeurs propres dans l'ordre décroissant, ce qui explique la nécessité de numéroter les axes factoriels.

HISTORIQUE

Voir **analyse factorielle des correspondances**.

DOMAINES ET LIMITATIONS

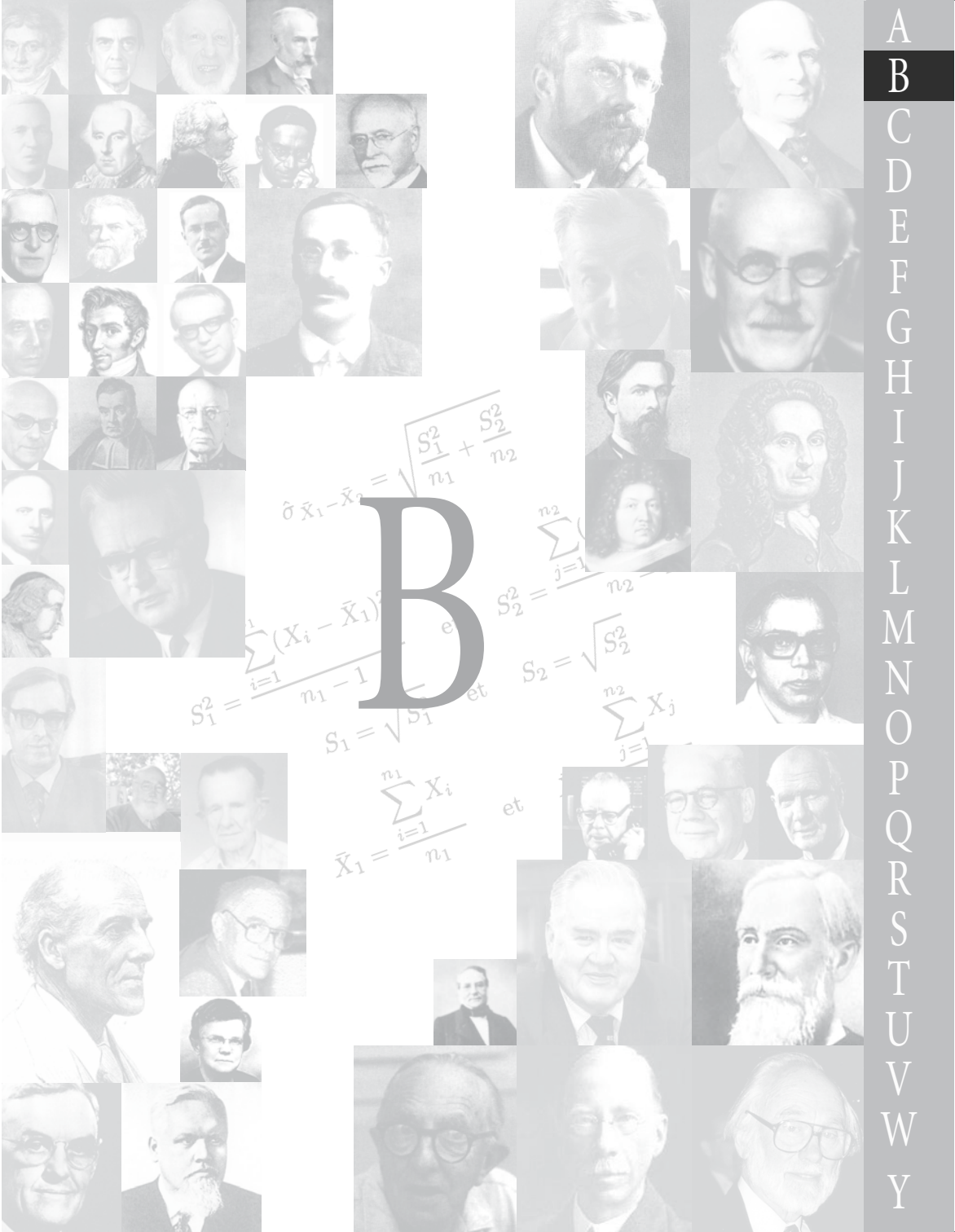
La notion d'axe factoriel est utilisé dans l'**analyse factorielle des correspondances.**

MOTS CLÉS

Analyse factorielle des correspondances *(Correspondence analysis)*
Inertie *(Inertia)*
Valeur propre *(Eigenvalue)*
Vecteur propre *(Eigenvector)*

RÉFÉRENCES

Voir **analyse factorielle des correspondances**.


B
A
B
C
D
E
F
G
H
I
J
K
L
M
N
O
P
Q
R
S
T
U
V
W
Y

BARNARD GEORGE A.

Barnard George A.



George Alfred Barnard est né en 1915, en Walhamstow, Angleterre. Il est diplômé en mathématiques à l'Université de Cambridge en 1936 et reçoit le titre de docteur à l'Université de Londres en 1965. Entre 1942 et 1945, il travaille au ministère de l'Approvisionnement comme consultant scientifique. G.A. Barnard devient membre du département des mathématiques de l'Imperial College de 1945 à 1966. De 1966 à 1975, il est professeur de mathématiques à l'Université d'Essex et, de 1975 jusqu'à sa retraite en 1981, il est professeur de statistique à l'Université de Waterloo. George A. Barnard a reçu de nombreuses distinctions parmi lesquelles la médaille d'or de la Royal Statistical Society et de l'Institute of Mathematics and Its Applications. En 1987, il est nommé membre honorifique de l'Institute of Statisticians.

George A. Barnard est décédé en 2002.

Quelques articles principaux de George A. Barnard :

1954 Sampling inspection and statistical decisions, *Journal of the Royal Statistical Society.* Series B, 16, pp. 151-174.

1958 Thomas Bayes - A biographical note, *Biometrika*, 45, pp. 293-315.

1989 On alleged gains in power from lower $p-$values. *Statistics in Medicine*, 8, pp. 1469-1477.

1990 Must clinical trials be large ? The interpretation of $p-$values and the combination of test results. *Statistics in Medicine*, 9, pp. 601-614.

BAYES, THÉORÈME DE

Bayes' theorem

Voir **théorème de Bayes**.

BAYES THOMAS

Bayes Thomas

Thomas Bayes (1702-1761) était le fils aîné de Joshua Bayes, l'un des six premiers ministres nonconformistes élus en Angleterre et membre de la « Royal Society ». Son instruction se fit par des professeurs privés selon la coutume des familles non conformistes. En 1731, il fut ordonné révérend à la chapelle presbytérienne de Tunbridge Wells, située à environ 150 km au sud-ouest de Londres. Grâce à quelques publications religieuses, il fut nommé membre de la Royal Society en 1742.

Son intérêt pour les mathématiques était connu par ses contemporains malgré le fait qu'il n'avait fait aucune publication technique, car il avait fait ses classes chez **Abraham de Moivre**, l'un des fondateurs de la théorie des probabilités. En 1763, R. Price fit publier parmi les papiers laissés par Bayes, l'œuvre principale de ce dernier :

1763 An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions*

of the Royal Society of London, 53, pp. 370-418. Réédité avec une biographie de **George A. Barnard** en 1958 et dans **Egon Sharpe Pearson** et Maurice George Kendall en 1970. *Studies in the History of Statistics and Probability.* pp. 131-153.



MOTS CLÉS

Théorème de Bayes *(Bayes' theorem)*

BERNOULLI, ÉPREUVE DE

Bernoulli trial

Voir **épreuve de Bernoulli**.

BERNOULLI, FAMILLE

Bernoulli family

Originaire de Bâle, la famille Bernoulli donna à la science plusieurs mathématiciens. **Jakob Bernoulli** (1654-1705), après avoir étudié la théologie à l'Université de Bâle selon les désirs de son père, entreprit de voyager pendant plusieurs années au cours desquelles il enseigna en même temps qu'il poursuivait ses propres études. S'étant résolument tourné vers les mathématiques, il devint professeur à l'Université de Bâle en 1687. Selon Stigler (1986), Jakob Bernoulli (appelé James Bernoulli dans la plupart des ouvrages anglais) est le père de la quantification de l'incertain.

Ce ne fut qu'en 1713, sept ans après sa mort et à l'instigation de son neveu Nikolaus, que fut publié son principal ouvrage « Ars Conjectandi ». Cet ouvrage est divisé en quatre parties : dans la première, l'auteur commente le célèbre traité de Huygens ; la deuxième est consacrée à la théorie des permutations et combinaisons, la troisième aux solutions de divers problèmes sur les jeux de hasard et enfin, dans la quatrième, sont exposées des propositions pour appliquer la théorie des probabilités aux questions d'intérêts moraux et de la science économique.

Un grand nombre des travaux de Jakob Bernoulli ne furent jamais publiés.

Johann Bernoulli (1667-1748), plus intéressé par les mathématiques que par la carrière médicale à laquelle le destinait son père, devint, lui aussi, professeur à l'Université, d'abord à Groningen aux Pays-Bas, puis à Bâle où il reprit la chaire laissée vacante par la mort de son frère.

Les deux frères avaient travaillé notamment sur le calcul différentiel et intégral, les études de fonctions et les problèmes de minimisation.

RÉFÉRENCE

Stigler, S. (1986). *The History of Statistics, the Measurement of Uncertainty before 1900.* The Belknap Press of Harvard University Press. London, England, pp. 62-71.

BERNOULLI, LOI DE

Bernoulli distribution

Voir **loi de Bernoulli**.

BERNOULLI, THÉORÈME DE

Bernoulli's theorem

Voir **théorème de Bernoulli**.

BERNOULLI JAKOB

Bernoulli Jakob

Jakob (ou Jacques ou Jacob ou James) Bernoulli (1654-1705) est, avec son frère Johann, un des pionniers de la forme leibnizienne du calcul. Il a transformé le calcul de l'espérance en calcul de probabilité à proprement parlé. Il a formulé et prouvé la loi faible des grands nombres qui est la base des probabilités modernes et des statistiques.



Jakob Bernoulli fut nommé maître ès lettres en 1671. En parallèle à sa licence de théologie qu'il obtint en 1676, il étudia secrètement les mathématiques et l'astronomie, contre le vœu de son père Nikolaus Bernoulli.

En 1682, Leibniz publia, dans *Acta Eruditorium*, une méthode pour déterminer l'intégrale de fonctions algébriques, une présentation courte de calcul différentiel sous une forme algorithmique et quelles remarques concernant l'idée fondamentale du calcul intégral. Ce papier éveilla l'attention de Jakob et de son frère Johann qui améliorèrent considérablement le travail de Leibniz, tant et si bien que Leibniz avoua que le calcul infinitésimal doit plus aux frères Bernoulli qu'à lui-même. C'est en 1690 que Jakob introduisit le terme « intégral ».

En 1687, il fut nommé professeur de mathématiques à l'Université de Bâle, où il restera jusqu'à sa mort en 1705.

Ars Conjectandi est l'ouvrage de Bernoulli généralement considéré comme étant le plus original. Il consiste en quatres parties : la première partie contient une réimpression du *De Ratiociniis in Ludo Aleae* de Huygens (publié en 1657) et est complétée par d'importantes modifications et généralisations. Dans la seconde partie, Bernoulli traite de la théorie des combinaisons. La troisième partie est composée de vingt-quatre exemples concernant la détermination du concept modifié d'espérance de Huygens pour différents jeux. Finalement, la quatrième partie est la plus intéressante et la plus originale, mais malheureusement Bernoulli n'eut pas le temps de la terminer ; c'est dans ce chaptire que Bernoulli distingue deux façons de déterminer, exactement ou approximativement, la mesure classique de probabilité.

D'autre part, Jakob Bernoulli s'intéressa aux stochastiques dans les années 1680. Le développement de son idée peut être clairement suivie dans son journal scientifique *Meditationes*.

Quelques ouvrages et articles principaux de Jakob Bernoulli :

1677 *Meditationes, Annotationes, Animadversiones Theologicæ & Philosophicæ*, a me JB. concinnatae & collectae ab anno 1677. MS Basel L I a 3.

1713 *Ars Conjectandi*, Opus Posthumum. Accedit

Tractatus de Seriebus infinitis, et Epistola Gallice scripta de ludo Pilae recticularis. Basileae, Impensis Thurnisiorum, fratrum.

BÊTA, LOI

Beta distribution

Voir **loi bêta**.

BIAIS

Bias

Au sens statistique, le biais est défini comme la différence existant entre l'**espérance mathématique** d'une **statistique** et la vraie **valeur** du **paramètre** correspondant. Le biais est donc une mesure de l'erreur systématique d'un **estimateur**. En calculant plusieurs fois des estimations d'un estimateur non biaisé, on va trouver « en moyenne » sur la bonne valeur. Le biais indique donc l'écartement de l'estimateur par rapport à la vraie valeur du paramètre.

HISTORIQUE

Le concept du caractère non biaisé d'un **estimateur** est dû à **Carl Friedrich Gauss** (1821) lorsqu'il travailla à la méthode des **moindres carrés**.

ASPECTS MATHÉMATIQUES

Soit une **statistique** T utilisée pour estimer un **paramètre** θ. Si $E[T] = \theta + b(\theta)$ (où $E[T]$ représente l'**espérance mathématique** de T), alors la quantité $b(\theta)$ est appelée le biais de la statistique T.

Si $b(\theta) = 0$, on a $E[T] = \theta$, et T est un **estimateur** sans biais de θ.

DOMAINES ET LIMITATIONS

On ne doit pas confondre le biais d'un estimateur d'un paramètre avec son degré de precision, qui est une mesure de l'erreur d'échantillonnage de celui-ci.

Il faut aussi citer qu'on utilise ce mot d'une manière non scientifique, mais qui reflète le concept de « biais » : on parle de biais lié à l'enquêteur, lors d'un sondage (par exemple), ou bien lors d'une étude sur l'efficacité des services de santé, on peut parler d'un biais de sélection (dû à la différence systématique des groupes que l'on compare), d'exclusion (dû à l'exclusion systématique de certains patients de l'étude) ou d'analyse (dû à la façon dont les résultats sont évalués) ; en effet, des défauts à certains niveaux de l'analyse statistique peuvent empêcher d'obtenir des résultats (des estimations) non biaisés. Le biais reste mathématiquement une proprieté d'un estimateur.

EXEMPLES

Soit $X_1, X_2, \ldots, X_n$ une séquence de variables aléatoires indépendantes distribuées selon la même **loi de probabilité** de **moyenne** μ et de **variance** σ^2 finie. On peut calculer le biais de l'**estimateur**

$$S^2 = \frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - \bar{x})^2$$

utilisé pour estimer la variance σ^2 de la **population** de la façon suivante :

$$\begin{aligned}
E[S^2] &= E\left[\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right] \\
&= E\left[\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2 - (\bar{x} - \mu)^2\right] \\
&= \frac{1}{n}E\left[\sum_{i=1}^{n}(x_i - \mu)^2\right] - E(\bar{x} - \mu)^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}E(x_i - \mu)^2 - E(\bar{x} - \mu)^2 \\
&= Var(x_i) - Var(\bar{x}) \\
&= \sigma^2 - \frac{\sigma^2}{n}
\end{aligned}$$

Le biais de S^2 est donc égal à $\frac{-\sigma^2}{n}$.

MOTS CLÉS

Espérance mathématique *(Expected value)*
Estimateur *(Estimator)*
Estimation *(Estimation)*

RÉFÉRENCES

Gauss, C.F. (1821, 1823 et 1826). Theoria Combinationis Observationum Erroribus Minimis Obnoxiæ, partie 1, 2 et suppl. *Werke* 4, pp. 1-108.

Gauss, C.F. (1855). *Méthode des moindres carrés. Mémoire sur la combinaison des observations.* Traduction française par J. Bertrand. Mallet-Bachelier, Paris.

BIENAYMÉ IRÉNÉE-JULES

Bienaymé Irénée-Jules

Irénée-Jules Bienaymé (1796-1878) entra au ministère des Finances et devint inspecteur général des Finances en 1836, en France. Il perdit son emploi en 1848, pour des raisons politiques. Peu après, il donna un cours sur les probabilités à la Faculté des Sciences de Paris. Or la chaire de probabilités fut donnée à Lamé en 1850, à nouveau pour des raisons politiques et, la même année, Bienaymé fut réintégré comme inspecteur général des Finances.

Ce disciple de **Pierre Simon de Laplace** prouva l'inégalité de Bienaymé-Tchebychev quelques années avant **Pafnouti Lvovitch Tchebychev**. La réédition de l'article de Bienaymé en 1867 précède la version française de la preuve de P.L. Tchebychev. Cette inégalité sera utilisée pour la preuve incomplète du **théorème central limite** faite par P.L. Tchebychev, puis achevée par **Andreï Andreïevitch Markov**.

De plus, il formula correctement le théorème critique de processus à branchement en 1845.



L'œuvre la plus fabuleuse, pour le domaine public, de sa vie de statisticien est peut-être son travail sur la correction de l'usage de la table de mortalité de Duvillard.

Quelques ouvrages et articles principaux d'Irénée-Jules Bienaymé :

1853 Considérations à l'appui de la découverte de Laplace sur la loi de probabilité dans la méthode des moindres carrés. *Comptes rendus de l'Académie des sciences*, Paris 37, 5-13 ; réédité en 1867 dans le journal de Liouville précédant la preuve des inégalités de Bienaymé-Tchebychev, *Journal de mathématiques pures et appliquées* 12, pp. 158-176.

BILATÉRAL, TEST

Two-tail test

Voir **test bilatéral**.

BINÔME

Binomial

Les sommes algébriques contenant des variables sont appelées polynômes (du grec *poly*, plusieurs). En particulier une expression qui contient deux termes est un binôme (du latin *bi*, double). On parle de monôme (du grec *mono*, unique) pour une expression à un seul terme et de trinôme (du latin *tri*, triple), lorsqu'elle comprend trois éléments.

HISTORIQUE

Voir **triangle arithmétique**.

ASPECTS MATHÉMATIQUES

Le carré d'un binôme se calcule facilement :

$$(a+b)^2 = a^2 + 2ab + b^2.$$

Il existe des formules du binôme pour des exposants plus grands que deux :

$$\begin{aligned}(a+b)^3 &= a^3 + 3a^2b + 3ab^2 + b^3\\ (a+b)^4 &= a^4 + 4a^3b + 6a^2b^2 + 4ab^3 + b^4.\end{aligned}$$

On peut écrire la formule du binôme généralisée de la façon suivante :

$$\begin{aligned}(a+b)^n &= a^n + na^{n-1}b + \frac{n\,(n-1)}{2}a^{n-2}b^2\\ &\quad + \cdots + \frac{n\,(n-1)}{2}a^2b^{n-2}\\ &\quad + nab^{n-1} + b^n\\ &= C_n^0 a^n + C_n^1 a^{n-1}b + C_n^2 a^{n-2}b^2\\ &\quad + \cdots + C_n^{n-1}ab^{n-1} + C_n^n b^n\\ &= \sum_{k=0}^{n} C_n^k a^{n-k}b^k\end{aligned}$$

où $C_n^k = \begin{pmatrix} n \\ k \end{pmatrix}$ est « n binomial k » et vaut $\frac{n!}{(n-k!)\cdot k!}$.

DOMAINES ET LIMITATIONS

Le binôme $(p+q)$ développé à la puissance n a donné son nom à la **loi binomiale** puisqu'il correspond à la **probabilité** totale obtenue après n épreuves de Bernoulli.

MOTS CLÉS

Combinaison *(Combination)*
Loi binomiale *(Binomial distribution)*
Triangle arithmétique *(Arithmetic triangle)*

BINOMIAL, TEST

Binomial test

Voir **test binomial**.

BINOMIALE, LOI

Binomial distribution

Voir **loi binomiale**.

BINOMIALE, TABLE

Binomial table

Voir **table binomiale**.

BINOMIALE NÉGATIVE, LOI

Negative binomial distribution

Voir **loi binomiale négative**.

BIOSTATISTIQUE

Biostatistics

La biostatistique est la branche scientifique qui utilise les méthodes statistiques dans le but de répondre à des questions qui sont soulevées dans les domaines de la biologie humaine et de la médecine. Le préfixe « bio » vient du grec *bios* qui signifie la vie. Les domaines de la biostatistique sont principalement l'**épidémiologie**, les essais cliniques, biologiques ainsi que l'étude de l'éthique de ces essais.

HISTORIQUE

La biostatistique a vu le jour au milieu du XVII^{e} siècle, lorsque Sir William Petty (1623-1687) et John Graunt (1620-1674) ont conçu de nouvelles méthodes pour analyser les « London Bills of Mortality ». Ils ont essentiellement appliqué ces nouvelles méthodes à l'analyse de la mortalité, de la natalité et de recensements, créant ainsi ce qu'on appelle aujourd'hui la biométrie. Par la suite, dans la seconde moitié du XIX^{e} siècle, les travaux de Gregor Mendel se sont dirigés vers l'étude de l'hérédité des plantes, observations fondées sur des données recueillies systématiquement, ainsi que sur l'application des méthodes numériques pour la description de la régularité de l'hérédité. **Francis Galton** et **Karl Pearson** ont été les deux figures les plus importantes dans le dévéloppement de cette science en s'intéressant aux nouveaux concepts et aux méthodes statistiques pour décrire et analyser les données de ressemblance physique, psychologique et comportementale entre les parents et leurs enfants. K. Pearson, F. Galton et Walter Frank Raphael Weldon (1860-1906) sont les co-fondateurs de la revue *Biometrika*. **Ronald Aylmer Fisher**, lors de ses travaux à la station d'études agricoles de Rothamsted, a proposé une méthode d'**échantillonnage** aléatoire pour le partage des animaux dans différents groupes et l'attribution de différents traitements. Cette phase marque le début de l'étude des essais cliniques.

DOMAINES ET LIMITATIONS

Les domaines de la biostatistique sont principalement l'**épidémiologie**, les essais cliniques ainsi que les questions éthiques qui sont soulevées, les essais biologiques. Un des domaines où la biostatistique a apporté des concepts statistiques pour analyser une question spécifique a été la mesure indirecte de l'intensité de la persistance d'un composé (ex : vitamines et hormones) administré à un être vivant. Il s'agit essentiellement d'un domaine de la statistique orienté vers la résolution de problèmes qui apparaissent dans les sciences biomédicales.

Parmi les méthodes statistiques les plus couramment employées dans le domaine de la biostatistique, on cite : les méthodes de **ré-échantillonnage** (**bootstrap**, **méthode du jackknife**), l'**analyse multivariée**, les différentes méthodes de régression, l'évaluation de l'incertitude liée à l'estimation, le traitement des données manquantes, etc.

Les revues scientifiques les plus connues traitant du domaine de la biostatistique sont : *Statistics in Medicine, Biometrical Journal, Controlled Clinical Trials, Journal of Biopharmaceutical Statistics* et *Statistical Methods in Medical Research.*

MOTS CLÉS

Démographie *(Demography)*
Épidémiologie *(Epidemiology)*

RÉFÉRENCES

Armitage P. et Colton T. (1998). *Encyclopedia of Biostatisics.* John Wiley & Sons, New York.

Graunt, J. (1662). *Natural and political observations mentioned in a following index, and made upon the bills of mortality : with reference to the government, religion, trade, growth, ayre, diseases, and the several changes of the said city.* Tho. Roycroft, for John Martin, James Allestry, and Tho. Dicas.

Greenwood, M. (1948). *Medical Statistics from Graunt to Farr.* Cambridge University Press, Cambridge.

Petty, W. (1682). *Another Essay in Political Arithmetic Concerning the Growth of the City of London.* Kelly, New York. (2th ed. 1963).

Petty, W. (1683). *Observations Upon the Dublin-Bills of Mortality, 1681, and the State of That City.* Kelly, New York. (2th ed. 1963).

BLOC

Block

On désigne par « blocs » les ensembles dans lesquels sont regroupées les unités expérimentales de telle sorte qu'elles soient le plus semblables possible.

On peut s'attendre ainsi à ce que l'**erreur** expérimentale soit plus petite que pour un même nombre d'unités aléatoirement situées à l'intérieur de la totalité de l'espace expérimental.

Les blocs sont généralement déterminés pour tenir compte, outre les causes contrôlables définies par les facteurs étudiés, d'autres causes qu'il peut être difficile, voire impossible, de maintenir constantes sur la totalité des unités expérimentales de l'**expérience**. Les variations entre les blocs sont alors éliminées lorsque l'on compare les effets des facteurs.

Plusieurs types de regroupements peuvent être utilisés pour réduire les effets d'une ou plusieurs sources d'erreurs. Ainsi, on parle de **plans randomisés à blocs** s'il n'y a qu'une source d'erreur.

HISTORIQUE

Le concept de bloc dans le domaine des plans d'expérience trouve son origine dans les travaux que **Ronald Aylmer Fisher** (1925) a entrepris lorsqu'il était à la tête de la « Rothamsted Experimental Station ». C'est en travaillant avec des chercheurs dans le domaine de l'agriculture que Fisher s'aperçut que la surface choisie pour l'**expérience** était manifestement hétérogène dans le sens où la fertilité variait d'une manière systématique d'un point à l'autre de la surface.

MOTS CLÉS

Facteur *(Factor)*
Plan carré gréco-latin *(Graeco-latin squares design)*
Plan carré latin *(Latin squares design)*
Plan d'expérience *(Design of experiments)*
Plan randomisé à blocs *(Randomized block design)*
Unité expérimentale *(Experimental unit)*

RÉFÉRENCE

Fisher, R.A. (1925). *Statistical Methods for Research Workers.* Oliver & Boyd, Edinburgh.

BONFERRONI CARLO E.

Bonferroni Carlo E.



Carlo Emilio Bonferroni naquit en 1892 à Bergame. Il étudia les mathématiques à Turin et fut diplômé en mathématiques. Il compléta sa formation durant une année à l'Université de Vienne et à l'Eidgenössiche Technische Hochschule de Zurich. Il fut officier durant la guerre de 1914-1918 et, après celle-ci, devint assistant professseur à l'École polytechnique de Turin. En 1923, il obtint la chaire de mathématiques

financières à l'Institut économique de Bari, où il fut recteur durant sept ans. Il fut finalement transféré à Florence où il occupa une chaire jusqu'à sa mort.

Bonferroni a traité différents sujets : les mathématiques actuaires, la probabilité et les mathématiques statistiques ainsi que l'analyse, la géométrie et la mécanique. Son nom est attaché aux inégalités de Bonferroni qui facilitent le traitement de la dépendance statistique et qui apparaissent pour la première fois en 1936, dans l'article *Teoria statistica delle classi e calcolo delle probabilità*. Les développements de ces inégalités ont engendré une large littérature.

Bonferroni mourut en 1960 à Florence (Italie).

Article principal de Carlo Emilio Bonferroni :

1936 Teoria statistica delle classi e calcolo delle probabilità. *Publicazioni del R Istituto Superiore di Scienze Economiche e Commerciali de Firenze*, 8, pp. 1-62.

BOOTSTRAP

Bootstrap

Le bootstrap est une famille de techniques qui servent principalement à estimer les erreurs-types, le **biais** et les **intervalles de confiance** d'un ou de plusieurs paramètres. Cela est fait à partir de n observations indépendantes d'une **variable aléatoire** avec **fonction de répartition** F inconnue et se revèle utile surtout lorsque le paramètre à estimer est une fonction compliquée de F.

L'idée de base du bootstrap (et qui ne diffère pas beaucoup de celle du jackknife) est d'estimer F par une distribution plausible $\hat{F}$ et ensuite de ré-échantillonner à partir de $\hat{F}$.

La possibilité d'utiliser les procédures bootstrap est strictement liée à l'ordinateur, qui permet un grand nombre de simulations. En effet, dans le bootstrap, les simulations automatiques prennent la place des calculs analytiques des méthodes « traditionnelles » d'estimation et dans certains cas donnent plus de liberté, par exemple quand nous ne désirons (ou ne pouvons) pas assumer des hypothèses concernant la structure de la distribution sous-jacente aux données. Des implémentations de certains types de bootstrap se trouvent dans des logiciels statistiques tels que S-PLUS, SAS, MATLAB.

HISTORIQUE

D'abord, relevons l'origine du mot « bootstrap », qui se refère au caractère réflexif des échantillons secondaires (ceux construits à partir de $\hat{F}$). Bootstrap signifie en français « lacets » et provient d'un conte sur le baron de Münchhausen qui, en se trouvant au fond d'un lac, en sortit en se prenant par ses lacets.

L'histoire du bootstrap, comme technique statistique en soi, commence en 1979 avec la publication d'un article par Efron (1979), qui eut un fort impact sur de nombreux chercheurs. Depuis, les articles sur la théorie et les nombreuses applications du bootstrap se multiplient de telle manière qu'il est difficile de citer seulement quelques-uns des chercheurs les plus importants dans ce domaine.

ASPECTS MATHÉMATIQUES

Divers types de bootstrap ont été proposés ; le but ici n'est pas l'exhaustivité de la tractation mais plutôt de donner une idée du type de méthode utilisé.

Soient $x_1, x_2, \ldots, x_n$ n observations indépendantes issues d'une variable aléatoire de fonction de répartition inconnue F.

Nous nous intéressons à l'estimation d'un paramètre inconnu δ dépendant de F et à sa fiabilité (c'est-à-dire à son biais, à sa **variance** et à ces intervalles de confiance). L'estimation de δ est donnée par une **statistique** $t = t(F)$ dépendante de F. F étant inconnue, nous recherchons une estimation $\hat{F}$ de F basée sur l'échantillon $x_1, x_2, \ldots, x_n$ et estimons δ par $t(\hat{F})$. Des exemples classiques pour t sont l'**espérance mathématique,** la variance et les quantiles de F. Notons T la variable aléatoire correspondante à la statistique t et dépendante de $x_1, x_2, \ldots, x_n$.

Distinguons :

- *bootstrap paramétrique.* Si $F = F_\theta$ fait partie d'une famille paramétrique de fonctions de répartition on peut estimer (le vecteur) θ par l'estimateur $\hat{\theta}$ du **maximum de vraisemblance**. On estime naturellement F par $F_{\hat{\theta}}$.
- *bootstrap non paramétrique.* Si F ne fait pas partie d'une famille paramétrique, elle est estimée par la fonction de répartition empirique calculée sur la base de l'échantillon considéré de taille n. La fonction de répartition empirique en x est définie par :

$$\hat{F}(x) = \frac{\text{nombre d'observations } x_i \leq x}{n}.$$

L'idée du bootstrap est la suivante.

1. Considérer R échantillons du type $x_1^*, x_2^*, \ldots, x_n^*$ issus indépendamment d'une variable aléatoire de distribution $\hat{F}$.

   Remarque : dans le cas non paramétrique cela est équivalent à engendrer des échantillons aléatoires avec remise à partir des données $x_1, x_2, \ldots, x_n$.

2. Calculer les valeurs estimées $t_1^*, t_2^*, \ldots, t_R^*$ du paramètre δ basé sur les échantillons ci-dessus. Calculer d'abord l'estimation $\hat{F}_i^*$ fondée sur l'échantillon i $(i = 1, \ldots, R)$ et ensuite poser :

$$t_i^* = t(\hat{F}_i^*).$$

   Suivant la méthode du « bootstrapping », le biais b et la variance v de t sont estimés par :

$$\begin{aligned} b_{boot} &= \bar{t}^* - t = \left(\frac{1}{R}\sum_{i=1}^{R} t_i^*\right) - t \\ v_{boot} &= \frac{1}{R-1}\sum_{i=1}^{R}(t_i^* - \bar{t}^*)^2. \end{aligned}$$

   Il faut remarquer que deux types d'erreur interviennent ici :

   - une erreur statistique, qui consiste dans le fait que nous considérons dans ce qui précède le biais et la variance étant donné $\hat{F}$ et pas le vrai F, inconnu ;
   - une erreur de simulation parce que le nombre de simulations n'est pas maximal. En effet, empiriquement, il est suffisant de considérer un nombre R de réechantillonnages compris entre 50 et 200.

En ce qui concerne les intervalles de confiance, il existe plusieurs méthodes dont nous vous présentons seulement une sélection :

1. *Intervalles normaux.* Supposons que T suit une loi de Gauss. En espérant que T^* suit aussi une loi normale, nous pouvons alors déduire l'intervalle de confiance suivant pour la valeur corrigée (par le biais) $t_b = t - b$ de t :

$$\left[t_b - \sqrt{v_{boot}} \cdot z_{1-\alpha/2},\ t_b + \sqrt{v_{boot}} \cdot z_{1-\alpha/2}\right]$$

   où z_γ est le γ quantile d'une loi de Gauss $N(0,1)$. Pour un seuil de tolérance de α, nous utilisons la valeur de la **table normale** $z_{1-\frac{\alpha}{2}}$, ce qui donne si $\alpha = 5\ \% : z_{1-\frac{0,05}{2}} = 1,96$.

2. *Intervalles bootstrap.* Supposant que $T - \delta$ ne dépend pas d'une inconnue, nous calculons le α-quantile a_α de $T - \delta$. Pour ce faire, classons les valeurs $t_1^* - t, t_2^* - t, \ldots, t_R^* - t$ dans l'ordre croissant :

$$t_{(1)}^* - t, t_{(2)}^* - t, \ldots, t_{(R)}^* - t$$

   et ensuite lisons les quantiles d'intérêt $a_\alpha = t_{((R+1)\alpha)}^* - t$ et $a_{1-\alpha} = t_{((R+1)(1-\alpha))}^* - t$.

   Et puisque :

$$\begin{aligned} 1 - 2\alpha &= \Pr\left(a_\alpha \leq T - \delta \leq a_{1-\alpha}\right) \\ &= \Pr(2t - t_{((R+1)(1-\alpha))}^* \leq \delta \\ &\quad \leq 2t - t_{((R+1)\alpha)}^*), \end{aligned}$$

   l'intervalle de confiance pour t au seuil de signification α est :

$$[2t - t_{((R+1)(1-\alpha))}^*, 2t - t_{((R+1)\alpha)}^*].$$

3. *Intervalles bootstrap studentisés.* Il est possible d'améliorer l'estimation précédente en considérant

$$Z = \frac{T - \delta}{\sqrt{v}}$$

où v est la variance de t et doit être estimée (normalement par des méthodes connues comme « Delta Methods » ou bien avec le jackknife). L'intervalle de confiance pour t est trouvé de manière analogue à ce qui précède :

$$\left[t - \sqrt{v} \cdot z^*_{((R+1)(1-\alpha))}, t + \sqrt{v} \cdot z^*_{((R+1)\alpha)}\right]$$

où $z^*_{((R+1)(1-\alpha))}$ et $z^*_{((R+1)\alpha)}$ sont les $(1-\alpha$ et $\alpha)$ quantiles « empiriques » de $\left(z_i^* = \frac{t_i^* - t}{v^*}\right)_{i=1,\ldots,R}$.

EXEMPLE

Nous considérons un exemple typique de bootstrap non paramétrique (pris de l'ouvrage de Davison, A. and Hinkley, D.V. [1997]). Les données concernent la population de 10 villes américaines en 1920 (U) et en 1930 (X) :

u	138	93	61	179	48
x	143	104	69	260	75

u	37	29	23	30	2
x	63	50	48	111	50

Nous nous intéressons à la valeur δ de la statistique $T = \frac{E(X)}{E(U)}$ qui permet de déduire la population de 1930 à partir de celle de 1920. Nous estimons δ par $t = \frac{\bar{x}}{\bar{u}} = 1,52$, mais quelle va être l'incertitude liée à cette estimation ?

Le bootstrap permet de simuler des valeurs t_i^* $(i = 1, \ldots, R)$ de δ issues d'une distribution bivariée $Y = (U, X)$ avec fonction de répartition empirique $\hat{F}$ qui met poids $\frac{1}{10}$ sur chaque observation.

En pratique, il s'agit de ré-échantillonner avec remise les 10 données et de calculer les

$$t_i^* = \frac{\sum_{j=1}^{10} x_{ij}^*}{\sum_{j=1}^{10} u_{ij}^*}, \quad i = 1, \ldots, R$$

correspondantes, avec x_{ij}^* et u_{ij}^* les valeurs des variables X et U. Cela permet d'approximer le biais et la variance de t suivant les formules données auparavant.

Le tableau suivant résume les simulations obtenues avec $R = 5$:

	j						
	1	2	3	4	5	6	7
u_{1j}^*	93	93	93	61	61	37	29
x_{1j}^*	104	104	104	69	69	63	50
u_{2j}^*	138	61	48	48	37	37	29
x_{2j}^*	143	69	75	75	63	63	50
u_{3j}^*	138	93	179	37	30	30	30
x_{3j}^*	143	104	260	63	111	111	111
u_{4j}^*	93	61	61	48	37	29	29
x_{4j}^*	104	69	69	75	63	50	50
u_{5j}^*	138	138	138	179	48	48	48
x_{5j^*}	143	143	143	260	75	75	75

	j			*Somme*
	8	9	10	
u_{1j}^*	29	30	2	$\sum u_{1j}^* = 528$
x_{1j}^*	50	111	50	$\sum x_{1j}^* = 774$
u_{2j}^*	30	30	2	$\sum u_{2j}^* = 460$
x_{2j}^*	111	111	50	$\sum x_{2j}^* = 810$
u_{3j}^*	30	2	2	$\sum u_{3j}^* = 571$
x_{3j}^*	111	50	50	$\sum x_{3j}^* = 1\,114$
u_{4j}^*	23	23	2	$\sum u_{4j}^* = 406$
x_{4j}^*	48	48	50	$\sum x_{4j}^* = 626$
u_{5j}^*	29	23	30	$\sum u_{5j}^* = 819$
x_{5j^*}	50	48	111	$\sum x_{5j}^* = 1\,123$

Ainsi, les t_i^* sont :

$$\begin{aligned} t_1^* &= 1,466, t_2^* = 1,761, t_3^* = 1,951, \\ t_4^* &= 1,542, t_5^* = 1,371. \end{aligned}$$

À partir de là, il est aisé de calculer le biais et la variance de t par les formules données auparavant :

$$b = 1,62 - 1,52 = 0,10 \text{ et } v = 0,0553.$$

Il est possible d'utiliser ces résultats pour calculer des intervalles de confiance normaux, en supposant que $T - \delta$ soit distribuée selon une loi normale $N(0,1\,;\,0,0553)$. Nous obtenons, pour un seuil de signification 5 %, l'intervalle de confiance suivant pour t : $[1,16\,;\,2,08]$.

Le nombre de villes considérées et le petit nombre de simulations ne permettent pas d'avoir beaucoup de confiance dans les valeurs obtenues.

MOTS CLÉS

Méthode du jackknife *(Jackknife method)*
Méthode de Monte-Carlo *(Monte Carlo method)*
Ré-échantillonnage *(Resampling)*
Simulation *(Simulation)*

RÉFÉRENCES

Davison, A.C. and Hinkley, D.V. (1997). *Bootstrap Methods and Their Application.* Cambridge University Press, Cambridge.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Annals of Statistics*, 7, pp. 1-26.

Efron, B. and Tibshirani, R.J. (1993). *An Introduction to the Bootstrap.* Chapman & Hall, New York.

Hall, P. (1992). *The Bootstrap and Edgeworth Expansion.* Springer-Verlag, New York.

Shao, J. and Tu, D. (1995). *The Jackknife and Bootstrap.* Springer-Verlag, New York.

BOSCOVICH ROGER J.

Boscovich Roger J.

Roger Joseph Boscovich naquit à Ragusa (actuellement Dubrovnik, Croatie) en 1711. Il y suivit le Collegium Ragusinum, puis alla à Rome étudier dans le Collegium Romanum, tous deux des collèges jésuites. Il mourut en 1787, à Milan.

Il développa, en 1757, une méthode géométrique pour trouver la ligne de regression simple L1 (afin de corriger les erreurs d'observations des astres). **Pierre Simon de Laplace** désapprouva le fait que R.J. Boscovich utilise une terminologie géométrique et traduisit cette méthode en une version algébrique en 1793.



Ouvrage principal de Roger Joseph Boscovich :

1757 *De Litteraria Expeditione per Pontificiam ditionem, et Synopsis amplioris Operis, ac habentur plura eius ex exemplaria etiam sensorum impressa.* Bononiensi Scientiarium et Artium Instituto Atque Academia Commentarii, Tomus IV, pp. 353-396.

RÉFÉRENCE

George Allen and Unwin Ltd. (1961). *Roger Joseph Boscovich, Studies of his Life and Work on 250$^{\text{th}}$ Anniversary of his Birth.* Lancelot Law Whyte.

BOX PLOT

Box plot

Le box plot est un moyen de représenter graphiquement les cinq valeurs suivantes : la **médiane**, le premier et le troisème quartiles, les valeurs extrêmes inférieure et supérieure.

Le box plot est donc un diagramme (boîte) qui donne des indications sur :

— une **mesure de tendance centrale** (en principe la médiane),

— la variabilité et

— la symétrie.

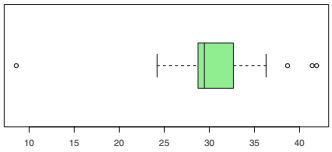
Il est souvent utilisé pour comparer plusieurs ensembles d'observations.

HISTORIQUE

Le « box-and-whisker plot », ou box plot, a été introduit par **John Wilder Tukey** en 1972, parallèlement à d'autres méthodes de représentation semi-graphiques de données dont une des plus connues est le **diagramme stem and leaf**.

ASPECTS MATHÉMATIQUES

Il existe plusieurs possibilités de représentation du box plot. Nous le présenterons sous la forme suivante :



Le rectangle central représente 50 % des observations. C'est l'**intervalle interquartile**. La limite inférieure est fixée au premier **quartile**, la limite supérieure au troisième quartile. Le rectangle est coupé par une ligne donnant la position de la **médiane**.

De chaque côté du rectangle, un segment de droite relie le quartile à la valeur extrême correspondante.

Dans cette représentation, les observations aberrantes sont traitées d'une manière particulière. Ainsi, lorsque les observations sont très dispersées, nous définissons deux valeurs, appelées limites intérieures, par :

$$lim.int.1 = \text{premier quartile-}(1,5\cdot \text{ intervalle interquartile})$$
$$lim.int.2 = \text{troisième quartile+}(1,5\cdot\text{intervalle interquartile})$$

Dans la construction du box plot, le segment de droite relie alors les quartiles aux valeurs observées les plus proches de ces bornes, mais à l'intérieur de l'intervalle déterminé. Les étoiles à l'extrémité du box plot indiquent les observations à l'extérieur de l'intervalle déterminé.

EXEMPLE

L'exemple suivant présente les indices des revenus des cantons de la Suisse par habitant (Suisse = 100) en 1993 :

Canton	Indice	Canton	Indice
Zurich	125,7	Schaffhouse	99,2
Berne	86,2	Appenzell Rh.-Ext.	84,2
Lucerne	87,9	Appenzell Rh.-Int.	72,6
Uri	88,2	Saint-Gall	89,3
Schwytz	94,5	Grisons	92,4
Obwald	80,3	Argovie	98,0
Nidwald	108,9	Thurgovie	87,4
Glaris	101,4	Tessin	87,4
Zoug	170,2	Vaud	97,4
Fribourg	90,9	Valais	80,5
Soleure	88,3	Neuchâtel	87,3
Bâle-Ville	124,2	Genève	116,0
Bâle-Campagne	105,1	Jura	75,1

Source : Office fédéral de la statistique (1993)

On calcule les différents quartiles :

Le box plot nous donne des informations sur la tendance centrale, sur la dispersion de la distribution et le degré de symétrie :

- tendance centrale :

  La **médiane** est égale à $90,10$.

- dispersion :

  L'**intervalle interquartile** indique dans quel **intervalle** se situent les 50 % des observations les plus au centre de la distribution.

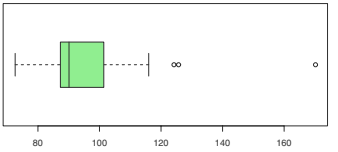
  Dans notre exemple, nous avons :

  – 50 % des cantons de la Suisse ont un indice se trouvant dans l'intervalle de $15,3$ ($[102,33 - 87,03]$).

Les deux limites du box plot sont les suivantes :

$$\begin{aligned} lim.int.1 &= \text{premier } \textbf{quartile}\text{-}(1,5\cdot \text{ intervalle interquartile}) \\ &= 87,03 - 1,5 \cdot 15,3 = 64,08. \\ lim.int.2 &= \text{troisème quartile}+(1,5\cdot\text{intervalle interquartile}) \\ &= 102,33 + 1,5 \cdot 15,3 = 125,28 \end{aligned}$$

Les valeurs supérieures à $125,28$ sont des observations aberrantes. Il y en a deux : $125,7$ (Zurich) et $170,2$ (Zoug). Le box plot est donc le suivant :

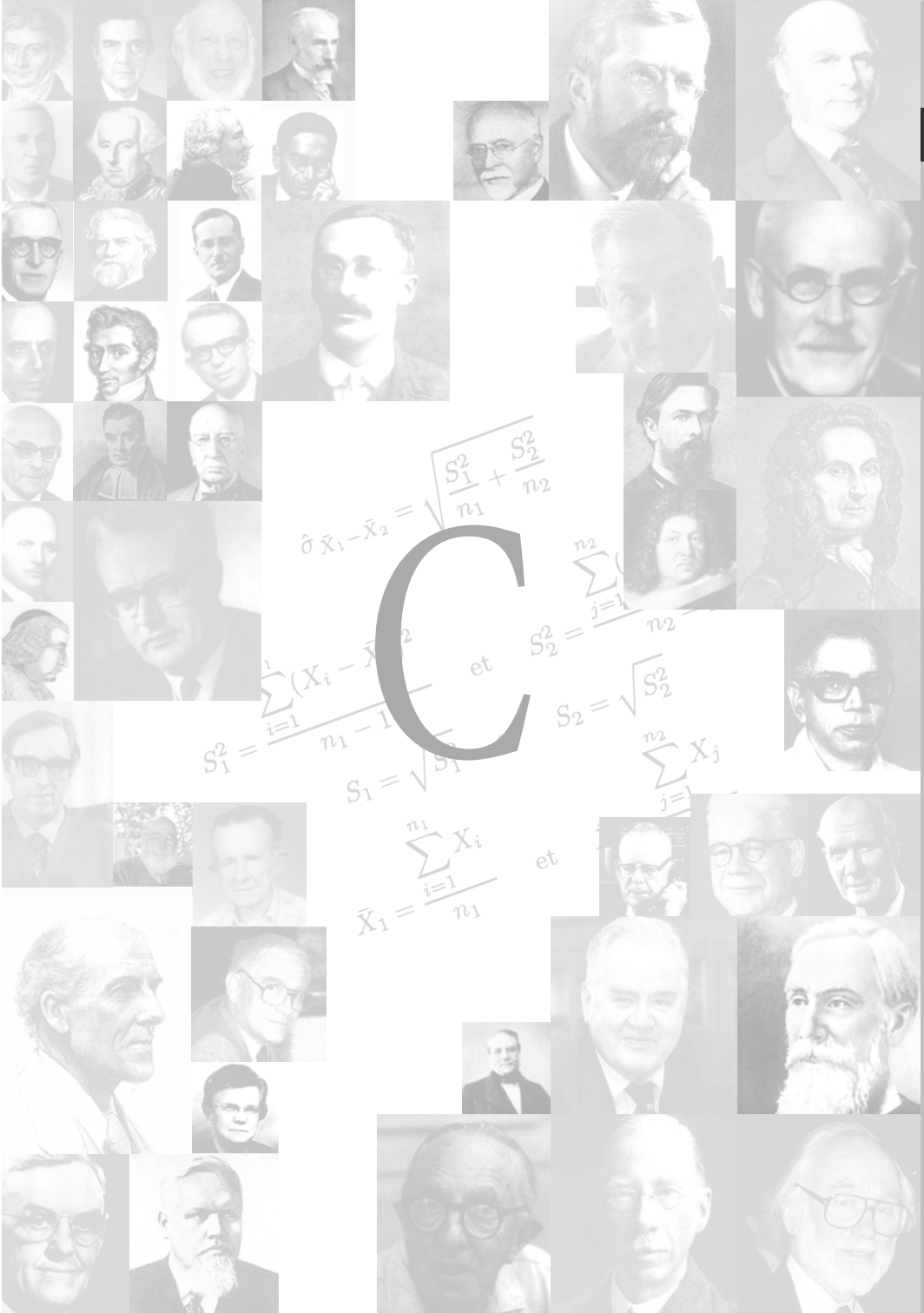


MOTS CLÉS

Analyse exploratoire des données *(Exploratory data analysis)*
Mesure de position *(Measure of location)*
Mesure de tendance centrale *(Measure of central tendency)*
Représentation graphique *(Graphical representation)*

RÉFÉRENCE

Tukey, J.W. (1972). Some graphical and Semigraphical Displays. In *Statistical Papers in Honor of George W. Snedecor*, ed. T.A. Bancroft, Arnes: The Iowa State University Press. Iowa, U.S.A, pp. 293-316.


C
et
et

CARRÉS GRÉCO-LATINS

Graeco-latin squares

Voir **plan carré gréco-latin**.

CARRÉS LATINS

Latin squares

Voir **plan carré latin**.

CATÉGORIE

Category

Une catégorie représente un ensemble de personnes ou d'objets ayant une caractéristique commune.

Si l'on cherche à étudier les personnes d'une **population**, on peut identifier des catégories « naturelles » telles que hommes et femmes ou bien d'autres catégories telles que ouvriers, employés, cadres, agriculteurs, etc. correspondant à la nomenclature des emplois.

MOTS CLÉS

Donnée binaire *(Binary data)*
Données catégoriques *(Categorical data)*
Population *(Population)*
Variable *(Variable)*
Variable aléatoire *(Random variable)*
Variable dichotomique *(Dichotomous variable)*
Variable qualitative catégorielle *(Categorial variable)*

CAUCHY, LOI DE

Cauchy distribution

Voir **loi de Cauchy**.

CAUSE ET EFFET EN ÉPIDÉMIOLOGIE

Cause and effect in epidemiology

En épidémiologie, la cause est un agent (germes microbiens, eau polluée, tabagisme, etc.) potentiellement pathogène (donc constituant un facteur de risque), qui est à l'origine d'une modification de l'état de santé ; l'effet est défini comme la modification de l'état de santé résultante.

L'effet est donc une comparaison de risque. On trouve deux effets différents :

- l'effet absolu d'une cause exprime le surcroît de risque ou encore le nombre additionnel de cas de maladie qui a résulté ou pourrait résulter de l'exposition à cette cause. Il se mesure par le **risque attribuable** et ses dérivés ;
- l'effet relatif d'une cause exprime la force de l'association causale entre l'agent causal et la maladie.

Une cause qui à elle seule produit un effet est dite *suffisante*.

HISTORIQUE

Les termes de causes et effets ont été définis tout au début de l'**épidémiologie**, que l'on peut localiser chronologiquement au XVII^{e} siècle.

ASPECTS MATHÉMATIQUES

Formellement, on a :

$$\begin{aligned}\text{Effet absolu} \quad &= \quad \text{risque des exposés} \\ &\quad\ - \text{risque des non-exposés.}\end{aligned}$$

Il exprime l'excès de risque ou de cas de maladie qui a (ou pourrait) résulté (er) de l'exposition à la cause en question.

$$\text{Effet relatif} \quad = \quad \frac{\text{risque des exposés}}{\text{risque des non-exposés}}.$$

Il traduit la force de l'association entre une maladie et la cause. Il se mesure par le **risque relatif** et l'**odds ratio**.

DOMAINES ET LIMITATIONS

En toute rigueur, la force d'une association entre un facteur d'exposition et une maladie ne suffit pas à établir un lien de causalité. Il est en effet raisonnable d'exiger en plus :

- un critère de temporalité : c'est-à-dire que l'on s'assure que la cause supposée précède bien l'effet ;
- des éléments de recherche fondamentale et expérimentale permettant de s'assurer que le facteur supposé causal n'est pas un facteur de confusion, c'est-à-dire un facteur non causal mais statistiquement lié à un facteur réellement causal non identifié.

Aux effets relatifs et absolus correspondent deux types de causalité. La causalité relative est indépendante de l'importance clinique ou de santé publique de l'effet ; elle ne permet pas en général de préjuger de l'importance clinique ou de santé publique de l'effet associé. Elle peut être forte, non seulement lorsque le risque chez les exposés est très élevé, mais également lorsque le risque chez les non-exposés est très faible.

La causalité absolue exprime l'impact clinique ou de santé publique de l'effet associé. Cette notion éclaire la question : si l'on venait à supprimer la cause, quel serait l'impact sur la population en terme de cas de maladie évitée ? Si tel patient cessait de fumer, quel serait la réduction de son risque de développer un cancer du poumon ou de faire un infarctus du myocarde ?

Cette tension entre causalité relative et absolue donne d'une certaine manière à l'épidémiologie son équilibre, l'identification de la relation causale, d'une part, et, de l'autre, l'importance de l'effet dans la population.

Un facteur de risque auquel est associé un effet relatif élevé mais concernant un petit nombre d'individus occasionera moins de cas de maladie et de décès qu'un facteur de risque auquel correspond un risque relatif plus faible mais qui concerne un grand nombre d'individus. Ainsi, il apparaît que la perception de l'importance d'une relation causale peut être fort différente selon que nous considérons celle-ci du point de vue de la causalité relative ou absolue.

Une autre remarque importante est liée à la notion d'interaction causale. Plusieurs causes d'une même maladie peuvent en effet agir au niveau du même processus pathogénique. Elles contribuent au même résultat, mais en interagissant. La conséquence principale de l'interaction d'effet est que l'on ne peut pas préjuger de l'effet de l'exposition simultanée aux causes A et B (noté $A+B$) sur la base de ce que l'on sait de l'effet de l'exposition à A seule ou de l'effet de l'exposition à B seule. Contrairement à ce qui se passe pour des causes indépendantes, il est impératif de ne pas se contenter d'analyser isolément les effets de causes interagissantes, mais d'en estimer l'effet conjoint.

EXEMPLE

L'effet relatif et l'effet absolu sont justiciables d'interprétations distinctes, comme le suggère l'exemple suivant.

Supposons deux populations P_1 et P_2, chacune de 100 000 habitants. Dans la population P_1, le risque (d'une maladie donnée) est de $0,2$ % chez les exposés et de $0,1$ % chez les non-exposés. Dans la population P_2, le risque des exposés est de 20 % et celui des non-exposés de 10 %, comme le montre le tableau suivant :

Population	Risque des exposés (%) A	Risque des non-exposés (%) B
P_1	$0,2$	$0,1$
P_2	20	10

L'effet relatif est identique dans les populations P_1 et P_2 (la cause multiplie le risque par un facteur 2), mais l'impact d'une même politique de prévention y serait très différent puisque l'effet absolu est dix fois plus important dans P_2 : le nombre de cas potentiellement évitables est de 100 dans la population P_1 et de 10 000 dans la population P_2.

Population	Effet relatif $\frac{A}{B}$	Effet absolu (%) $C = A - B$	Cas évitables $C \times 100\,000$
P_1	2,0	0,1	100
P_2	2,0	10	10 000

Considérons ensuite un **taux d'incidence** du cancer du poumon, chez une population de fumeurs de 35 cigarettes par jour ou plus, de $3,15/1\,000$/an. Ce taux, qui demeure somme toute modeste, masque cependant un puissant effet relatif (45 fois plus élevé chez les fumeurs que chez les non-fumeurs) du fait que le cancer du poumon est très rare chez les non-fumeurs (en admettant que le taux d'incidence est ici de $0,07/1\,000$/an).

MOTS CLÉS

Odds et odds ratio *(Odds and odds ratio)*
Risque relatif *(Relative risk)*
Risque attribuable *(Attributable risk)*
Risque évitable *(Avoidable risk)*
Taux d'incidence *(Incidence rate)*
Taux de prévalence *(Prevalence rate)*

RÉFÉRENCES

Lilienfeld A.M. and D.E. Lilienfeld. (1980). *Foundations of epidemiology.* 2nd ed New York: Clarendon Press, Oxford.

MacMahon, B. and Pugh, T.F. (1970). *Epidemiology: Principles and Methods.* Boston: Little Brown.

Morabia A. (1996). *Épidémiologie causale.* Éditions Médecine et Hygiène, Genève.

Morabia A. (1996). *L'Épidémiologie clinique.* Presses universitaires de France, Que sais-je ? Paris.

CENTILE

Percentile

Les centiles sont des **mesures de position** calculées sur un ensemble de données.

On appelle centiles les valeurs qui partagent une distribution en cent parties égales (chaque partie contient le même nombre d'observations).

Le x-ème centile est donc la valeur dont x % des observations lui sont inférieures et $(100 - x)$ % des observations lui sont supérieures. Par exemple, le centile 20 est la valeur qui sépare les 20 % inférieurs des 80 % supérieurs.

Nous aurons donc 99 centiles pour une distribution donnée :



Notons par exemple :

centile 10 = premier **décile**,
centile 25 = premier **quartile**,
centile 50 = **médiane**.

Cette notion fait partie de la famille des quantiles.

ASPECTS MATHÉMATIQUES

Le processus de calcul est similaire à celui de la **médiane**, des quartiles ou des déciles.

Lorsque nous possédons toutes les observations brutes, le processus de calcul des centiles est le suivant :

1. Organiser les n observations sous forme d'une **distribution de fréquences**.
2. Les centiles correspondent aux observations pour lesquelles la **fréquence** relative cumulée dépasse respectivement 1 %, 2 %, 3 %, ..., 98 %, 99 %.

   Certains auteurs proposent la formule suivante qui permet de déterminer avec précision la **valeur** des différents centiles :

   Calcul du j-ème centile :

   Soient i la partie entière de $\frac{j \cdot (n+1)}{100}$ et k la partie fractionnelle de $\frac{j \cdot (n+1)}{100}$, ainsi que x_i

et x_{i+1} les valeurs des observations classées respectivement en i-ème et $(i+1)$-ème position (lorsque les n observations sont classées par ordre croissant). Alors le j-ème centile est égal à :

$$C_j = x_i + k \cdot (x_{i+1} - x_i)$$

Lorsque nous possédons des observations groupées en classes, les centiles se déterminent de la manière suivante :

1. Déterminer la classe dans laquelle se trouve le centile voulu :

   - *premier centile* : première classe pour laquelle la fréquence relative cumulée dépasse 1 %.
   - *deuxième centile* : première classe pour laquelle la fréquence relative cumulée dépasse 2 %.

     ...
   - *quatre-vingt-dix–neuvième centile* : première classe pour laquelle la fréquence relative cumulée dépasse 99 %.

2. Calculer la valeur du centile en fonction de l'**hypothèse** selon laquelle les observations sont distribuées uniformément dans chaque classe, le j-ème centile C_j est :

$$C_j = L_j + \left[\frac{n \cdot \frac{j}{100} - \sum f_{inf}}{f_j} \right] \cdot c_j$$

où

L_j = borne inférieure de la classe du centile C_j,
n = nombre total d'observations,
$\sum f_{inf}$ = somme des fréquences inférieures à la classe du centile C_j,
f_j = fréquence de la classe du centile C_j,
c_j = dimension de l'**intervalle** de la classe du centile C_j.

DOMAINES ET LIMITATIONS

Le calcul des centiles n'a de sens que pour une **variable quantitative** pouvant prendre des valeurs sur un **intervalle** déterminé.

Le calcul des centiles ne peut se faire en pratique que lorsque le nombre d'observations est élevé puisque ce calcul consiste à diviser l'ensemble des observations en 100 parties.

Cette notion est peu utilisée car dans la plupart des analyses descriptives, la détermination des déciles ou même des quartiles est suffisante pour l'interprétation des résultats.

EXEMPLE

Considérons un exemple de calcul des centiles sur la **distribution de fréquences** d'une **variable** continue où les observations sont groupées en classes.

Le **tableau de fréquences** suivant représente les profits (en milliers d'euros) de 2 000 épiceries :

Profit (en milliers d'euros)	Fréquence	Fréquence cumulée	Fréquence relative cumulée
$100 - 150$	160	160	0, 08
$150 - 200$	200	360	0, 18
$200 - 250$	240	600	0, 30
$250 - 300$	280	880	0, 44
$300 - 350$	400	1 280	0, 64
$350 - 400$	320	1 600	0, 80
$400 - 450$	240	1 840	0, 92
$450 - 500$	160	2 000	1, 00
Total	2 000		

La classe comprenant le premier centile est la classe $100 - 150$ (celle dont la **fréquence** relative cumulée dépasse $0, 01 = 1$ %).

En considérant que les observations sont distribuées de manière uniforme dans chaque classe, nous obtenons pour le premier centile la **valeur** suivante :

$$\begin{aligned} \text{1er centile} &= 100 + \left[\frac{\left(2\,000 \cdot \frac{1}{100} \right) - 0}{160} \right] \cdot 50 \\ &= 106,25 \end{aligned}$$

La classe comprenant le dixème centile est la classe $150 - 200$. La valeur du dixème centile est égale à :

$$\begin{aligned} 10^{\text{e}} \text{ centile} &= 150 + \left[\frac{\left(2\,000 \cdot \frac{10}{100} \right) - 160}{200} \right] \cdot 50 \\ &= 160. \end{aligned}$$

Nous pouvons calculer tous les centiles de la même manière. Nous pourrons ensuite conclure par exemple que 1 % des 2 000 épiceries ont un profit compris entre 100 et 106, 25 milliers d'euros, ou que 10 % ont un profit compris entre 100 et 160 milliers d'euros, etc.

MOTS CLÉS

Décile *(Decile)*
Médiane *(Median)*
Mesure de position *(Measure of location)*
Quantile *(Quantile)*
Quartile *(Quartile)*

CENTRAL LIMITE, THÉORÈME

Central limit theorem

Voir **théorème central limite**.

CHI-CARRÉ, DISTANCE DU

Chi-square distance

Voir **distance du chi-carré**.

CHI-CARRÉ, LOI DU

Chi-square distribution

Voir **loi du chi-carré**.

CHI-CARRÉ, TABLE DU

Chi-square table

Voir **table du chi-carré**.

CHI-CARRÉ, TEST D'ADÉQUATION DU

Chi-square goodness of fit test

Voir **test d'adéquation du chi-carré**.

CHI-CARRÉ, TEST D'INDÉPENDANCE

Chi-square test of independence

Voir **test d'indépendance du chi-carré**.

CHI-CARRÉ, TEST DU

Chi-square test

Voir **test du chi-carré**.

CLASSIFICATION

Classification

La classification est une méthode d'analyse de données qui consiste au regroupement d'objets semblables. Chaque objet étant caractérisé par p variables, une classification se fait selon un critère

rationnel. Ainsi, selon le critère choisi, un objet peut appartenir à plusieurs classes.

HISTORIQUE

La classification des habitants d'une localité ou d'un pays selon leur sexe et selon d'autres caractéristiques physiques date de fort longtemps. Les Hindous, les Grecs anciens et les Romains développèrent de multiples typologies concernant les êtres humains. La plus ancienne est due au grec Galen (129-199).

Par la suite, l'utilisation de la classification s'étendit également aux domaines de la biologie et de la zoologie, pour ne citer que les travaux de Carl von Linné (1707-1778).

Les premières idées de classification se rapportant aux méthodes actuelles de **classification automatique** sont attribuées à Adanson (XVIII^{e} siècle). Joseph Zubin (1938) et Robert L. Thorndike (1953) tentèrent de développer quelques méthodes, mais le véritable développement des méthodes de classification correspond exactement à l'avènement de l'ordinateur.

ASPECTS MATHÉMATIQUES

Les méthodes de classification peuvent être distinguées en deux grandes catégories, l'une faisant appel aux probabilités et l'autre pas.

Dans la première catégorie, on trouve par exemple l'analyse discriminante. La seconde catégorie se divise en deux groupes : dans le premier, on trouve les méthodes de classification dites optimales utilisant la programmation mathématique ; dans le deuxième, on distingue plusieurs types de classification :

- le partitionnement, qui consiste à répartir les n objets en g groupes, de sorte qu'un objet appartienne exclusivement à un seul groupe. Le nombre de groupes g est fixé préalablement et le partitionnement satisfait de façon optimale le critère de classification utilisé ;

- le partitionnement avec classes empiétantes, qui admet l'appartenance simultanée d'un objet à plusieurs groupes ;

- la classification hiérarchique, qui s'intéresse également à la structure des données aux différents niveaux de la classification. Ainsi on peut se préoccuper de la **relation** existant entre les différents groupes lors du partitionnement. On distingue deux sortes de techniques hiérarchiques : la technique agglomérative et la technique divisive.

  Les méthodes agglomératives débutent lorsque tous les objets sont séparés, c'est-à-dire qu'à la première étape, les n objets sont répartis en n groupes. Puis à chaque étape, deux groupes sont agglomérés jusqu'à ce qu'il n'en reste plus qu'un seul.

  À l'inverse, les méthodes divisives débutent lorsque tous les objets sont ensemble, c'est-à-dire qu'à la première étape, les n objets ne forment qu'un seul groupe. Ensuite, à chaque étape, de nouveaux groupes sont formés jusqu'à l'obtention de n groupes ;

- la classification géométrique, qui consiste après avoir représenté les objets sur un **graphique de dispersion**, à déceler les groupes à former. On aura pris soin lors de la **représentation graphique**, de traduire la similitude des objets par une proximité sur le graphique.

Les trois premiers types de classification sont généralement englobés par le terme de **classification automatique**. Ils ont en commun le fait que les objets à classer doivent présenter une certaine structure, permettant la définition d'une mesure de ressemblance entre ces objets.
Chacun de ces types de classification comprend une multitude de méthodes permettant la formation de classes d'objets semblables.

DOMAINES ET LIMITATIONS

La classification peut être utilisée dans deux cas :

— les cas de description ;

— les cas de prédiction.

Dans le premier cas, la classification est faite à des fins de nomenclature et de résumés. Par exemple, les

professions seront classées en professions libérales, cadres, ouvriers, etc., et on pourra calculer des salaires moyens, la **fréquence** moyenne des problèmes de santé, etc., pour chaque classe.

Dans le deuxième cas, la classification mènera à une prédiction puis à une action. Par exemple, si les renards d'une région présentent les mêmes cas d'apathie et des traces de bave, on pourra conclure à une nouvelle épidémie de rage. Il sera alors possible de lancer une campagne de vaccination.

MOTS CLÉS

Analyse de données *(Data analysis)*
Classification automatique *(Cluster analysis)*
Méthode du chaînage complet *(Complete linkage method)*
Méthode du chaînage simple *(Single link method)*

RÉFÉRENCES

Everitt, B.S. (1974). *Cluster Analysis.* Halstead Press, London.

Gordon, A.D. (1981). *Classification. Methods for the exploratory analysis of multivariate data.* Chapman & Hall.

Kaufman, L. and Rousseeuw, P.J. (1990). *Finding Groups in Data, An Introduction to Cluster Analysis.* John Wiley & Sons, New York.

Thorndike, R.L. (1953). Who belongs in a family ? *Psychometrika*, 18, pp. 267-276.

Zubin, J. (1938). A Technique for Measuring Like-mindedness. *Journal of Abnormal and Social Psychology*, 33, pp. 508-516.

CLASSIFICATION AUTOMATIQUE

Cluster analysis

La nature offre un grand nombre de populations qu'il est souhaitable de répartir en catégories.

L'objectif d'une **classification** automatique sur un ensemble d'objets est de mettre en évidence des familles d'objets homogènes selon un certain critère.

Il existe plusieurs méthodes de classification automatique :

— le partitionnement ;

— le partitionnement avec classes empiétantes ;

— la classification hiérarchique.

Nous ne développerons ici que la classification hiérarchique.

HISTORIQUE

Voir **classification** et **analyse de données**.

ASPECTS MATHÉMATIQUES

Pour effectuer une **classification** automatique sur un ensemble de n objets, il est indispensable de pouvoir définir une **distance** ou plus généralement une mesure de ressemblance entre les objets à classer. Nous supposons donc l'existence d'une certaine structure entre les objets.

Pour effectuer une classification hiérarchique sur un ensemble E d'objets $\{X_1, X_2, \ldots, X_n\}$, il est indispensable d'avoir défini une distance sur E pour obtenir le **tableau de distances** entre les objets de E ; de même, il faut choisir une distance entre les sous-ensembles de E.

Dans la classification hiérarchique, il y a entre autres la méthode agglomérative. Celle-ci peut être résumée dans l'algorithme suivant :

1. Repérer la paire d'objets (X_i, X_j) telle que la distance les séparant soit la plus petite.

2. Agréger la paire d'objets (X_i, X_j) en un seul élément α et refaire un nouveau tableau de distances. Ce dernier est obtenu en supprimant les lignes et colonnes relatives à X_i et X_j et en les remplaçant par une ligne et une colonne relative à α. Ainsi, le nouveau tableau de distances aura une ligne et une colonne de moins.

3. Répéter ces deux opérations jusqu'à obtenir le nombre de classes souhaité ou jusqu'à ce que tous les objets soient réunis dans une même classe.

Notons que le choix de la distance entre le groupe formé d'éléments agrégés et les autres éléments donne lieu à des méthodes différentes. Citons par exemple la **méthode du chaînage simple** et la **méthode du chaînage complet**.

La méthode du chaînage simple est une méthode de classification hiérarchique utilisant la distance euclidienne pour construire le tableau de distances, et la distance entre deux classes est donnée par la distance euclidienne entre leurs deux plus proches éléments (distance minimum).

Par contre, dans la méthode du chaînage complet, la distance entre deux classes est donnée par la distance euclidienne entre leurs deux éléments les plus éloignés (distance maximum).

Étant donné que la seule différence entre ces deux méthodes réside dans le fait que la distance entre deux classes est respectivement la distance minimum et maximum, seule la méthode du chaînage simple sera considérée ici.

Nous commençons par construire le tableau de distances entre les éléments de E.

Comme ce tableau est symétrique et nul sur sa diagonale, nous ne considérons qu'une moitié du tableau :

$$\begin{array}{cccc} d(X_1,X_2) & d(X_1,X_3) & \dots & d(X_1,X_n) \\ & d(X_2,X_3) & \dots & d(X_2,X_n) \\ & & \ddots & \vdots \\ & & & d(X_{n-1},X_n) \end{array}$$

où $d(X_i,X_j)$ est la distance euclidienne entre X_i et X_j pour $i<j$ prenant leurs valeurs entre 1 et n.

L'**algorithme** de la méthode du chaînage simple est alors le suivant :

- chercher d'abord le minimum des $d(X_i,X_j)$ pour $i<j$;
- les éléments X_i et X_j sont alors agrégés en un nouveau groupe $C_k = X_i \cup X_j$;
- l'ensemble E est ensuite partitionné en :
  $\{X_1\},\dots,\{X_{i-1}\},\{X_i,X_j\},\{X_{i+1}\},\dots,$
  $\{X_{j-1}\},\{X_{j+1}\},\dots,\{X_n\}$;
- il faut alors refaire le tableau de distances où les lignes et les colonnes correspondant à X_i et X_j ont été supprimées et une ligne et une colonne représentant les distances entre X_m et C_k, $m=1,2,\dots,n,: m\neq i$ et $m\neq j$ données par :
  $$d(C_k,X_m) = \min\{d(X_i,X_m)\ ;\ d(X_j,X_m)\},$$
  ont été rajoutées.

Nous répétons alors l'algorithme jusqu'à ce que le nombre de groupes souhaité soit atteint ou jusqu'à ce qu'il n'y ait plus qu'un seul groupe formé de tous les éléments.

Dans le cas général, la distance entre deux groupes est donnée par :

$$\begin{array}{rcl} d(C_k,C_m) & = & \min\{d(X_i,X_j) \text{ tel que } X_i \\ & & \text{appartient à } C_k \text{ et } X_j \text{ à } C_m\}, \end{array}$$

la formule citée auparavant s'appliquant au cas particulier où les groupes sont respectivement composés de deux éléments et d'un seul élément.

La suite des regroupements effectués peut alors être représentée sur un **dendrogramme**, sur lequel nous reportons en abscisse la distance séparant les objets que nous rassemblons. Notons que, lors de la recherche de la paire d'éléments les plus rapprochés, il se peut que deux paires satisfassent le critère de minimalité. Le choix de la paire qui est agrégée en premier n'influence pas la suite (pour autant que l'algorithme ne s'arrête pas à cette étape) car, au pas suivant, ce sera la deuxième paire que nous regroupons. Sur le dendrogramme, l'ordre d'agrégation ne se verra pas puisque nous reportons la distance séparant les deux objets regroupés.

DOMAINES ET LIMITATIONS

Il est nécessaire de préciser qu'une **classification** automatique obtenue sur un ensemble n'est jamais

LA classification de cet ensemble, mais une classification (parmi beaucoup d'autres) établie à partir de variables et de méthodes choisies intentionnellement.

Le choix de la **distance** entre le groupe formé d'éléments agrégés et les autres éléments peut s'opérer de plusieurs façons, selon la méthode utilisée, comme par exemple dans la **méthode du chaînage simple** et dans la **méthode du chaînage complet**.

EXEMPLE

Illustrons la **méthode du chaînage simple** sur un exemple de notes obtenues par cinq étudiants à une session d'examens portant sur quatre cours, à savoir : l'anglais, le français, les mathématiques et la physique.

Nous désirons répartir ces cinq étudiants en deux groupes à l'aide de la méthode du chaînage simple pour tester une nouvelle méthode d'enseignement.

Ces notes, de 1 à 6, indiquées au demi-point sont résumées dans le tableau suivant :

	Anglais	Français	Maths	Physique
Alain	5,0	3,5	4,0	4,5
Jean	5,5	4,0	5,0	4,5
Marc	4,5	4,5	4,0	3,5
Paul	4,0	5,5	3,5	4,0
Pierre	4,0	4,5	3,0	3,5

Nous obtenons le **tableau de distances** en mettant les **distances** euclidiennes entre les étudiants dans le tableau suivant :

	Alain	Jean	Marc	Paul	Pierre
Alain	0	1,22	1,5	2,35	2
Jean	1,22	0	1,8	2,65	2,74
Marc	1,5	1,8	0	1,32	1,12
Paul	2,35	2,65	1,32	0	1,22
Pierre	2	2,74	1,12	1,22	0

En ne considérant que la moitié supérieure du tableau symétrique ci-dessus, nous trouvons :

	Jean	Marc	Paul	Pierre
Alain	$1,22$	$1,5$	$2,35$	2
Jean		$1,8$	$2,65$	$2,74$
Marc			$1,32$	$1,12$
Paul				$1,22$

La distance minimale vaut $1,12$ entre Marc et Pierre ; donc un premier groupe est formé avec ces deux étudiants. Il faut alors calculer les nouvelles distances.

Calculons par exemple la nouvelle distance entre Marc et Pierre d'une part et Alain d'autre part en prenant le minimum de la distance entre Marc et Alain et de celle entre Pierre et Alain, c'est-à-dire :

$$\begin{aligned} & d(\{\text{Marc, Pierre}\}, \text{Alain}) \\ &= \min\{d(\text{Marc, Alain}) \,;\, d(\text{Pierre, Alain})\} \\ &= \min\{1,5\,;\, 2\} = 1,5 \end{aligned}$$

de même :

$$\begin{aligned} & d(\{\text{Marc, Pierre}\}, \text{Jean}) \\ &= \min\{d(\text{Marc, Jean}) \,;\, d(\text{Pierre, Jean})\} \\ &= \min\{1,8\,;\, 2,74\} = 1,8 \end{aligned}$$

et :

$$\begin{aligned} & d(\{\text{Marc, Pierre}\}, \text{Paul}) \\ &= \min\{d(\text{Marc, Paul}) \,;\, d(\text{Pierre, Paul})\} \\ &= \min\{1,32\,;\, 1,22\} = 1,22. \end{aligned}$$

Le nouveau tableau de distances prend la forme suivante :

	Jean	Marc et Pierre	Paul
Alain	$1,22$	$1,5$	$2,35$
Jean		$1,8$	$2,65$
Marc et Pierre			$1,22$

La distance minimale vaut maintenant $1,22$ entre Alain et Jean (nous retrouvons ce minimum également entre le groupe Marc et Pierre d'une part et Paul d'autre part) ; choisissons de regrouper Alain et Jean en premier. L'autre agrégation se fera à l'étape suivante. Nous refaisons alors le nouveau tableau de distances, où, en utilisant le dernier tableau, nous obtenons :

$$\begin{aligned} & d(\{\text{Alain, Jean}\}, \{\text{Marc, Pierre}\}) \\ &= \min\{d(\text{Alain}, \{\text{Marc, Pierre}\}) \,; \\ &\quad\; d(\text{Jean}, \{\text{Marc, Pierre}\})\} \\ &= \min\{1,5\,;\, 1,8\} = 1,5 \end{aligned}$$

et de même :

$$\begin{aligned} & d(\{\text{Alain, Jean}\}, \text{Paul}) \\ &= \min\{d(\text{Alain, Paul}) \ ; \ d(\text{Jean, Paul})\} \\ &= \min\{2,35\,;\,2,65\} = 2,35. \end{aligned}$$

D'où le nouveau tableau de distances :

	Marc et Pierre	Paul
Alain et Jean	1,5	2,35
Marc et Pierre		1,22

Nous remarquons alors que maintenant Paul doit être intégré au groupe formé par Marc et Pierre et que la nouvelle distance sera :

$$\begin{aligned} & d(\{(\text{Marc, Pierre}), \text{Paul}\}, \{\text{Alain, Jean}\}) \\ &= \min\{d(\{\text{Marc, Pierre}\}, \{\text{Alain, Jean}\}) \ ; \\ &\quad d(\text{Paul}, \{\text{Alain, Jean}\})\} \\ &= \min\{1,5\,;\,2,35\} = 1,5 \end{aligned}$$

ce qui donne le tableau de distances suivant :

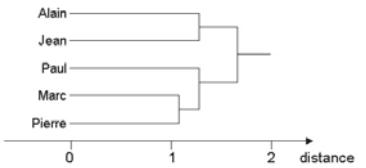
	Alain et Jean
Marc, Pierre et Paul	1,5

Nous avons finalement obtenu les deux groupes :

{Alain et Jean} et {Marc, Pierre et Paul}

qui sont séparés d'une distance de 1,5.

Le **dendrogramme** suivant illustre les agrégations successives :



MOTS CLÉS

Classification *(Classification)*
Dendrogramme *(Dendrogram)*
Distance *(Distance)*
Méthode du chaînage complet *(Complete linkage method)*
Méthode du chaînage simple *(Single link method)*
Tableau de distances *(Distance table)*

RÉFÉRENCES

Celeux, G., Diday, E., Govaert, G., Lechevallier, Y. et Ralambondrainy, H. (1989). *Classification automatique des données - aspects statistiques et informatiques.* Dunod, Paris.

Gordon, A.D. (1981). *Classification. Methods for the exploratory analysis of multivariate data.* Chapman & Hall.

Jambu, M. et Lebeaux, M.O. (1978). *Classification automatique pour l'analyse de données.* Dunod, Paris.

Lerman, L.C. (1981). *Classification et analyse ordinale des données.* Dunod, Paris.

Tomassone, R., Daudin, J.J., Danzart, M. et Masson, J.P. (1988). *Discrimination et classement.* Masson, Paris.

COEFFICIENT DE CORRÉLATION

Correlation coefficient

Le coefficient de corrélation simple est une mesure de l'intensité de la relation linéaire entre deux variables aléatoires.

Les valeurs possibles du coefficient de corrélation appartiennent à l'intervalle $[-1\,;\,1]$. Ces deux valeurs extrêmes représentent une **relation** linéaire parfaite entre les variables, « positive » dans le premier cas et « négative » dans l'autre. La valeur 0 (zéro) signifie l'absence de relation linéaire.

Le coefficient de corrélation présenté ici est également appelé « coefficient de corrélation de Bravais -Pearson ».

HISTORIQUE

Le concept de corrélation est né dans les années 1880 avec les travaux de **Francis Galton**. Dans son autobiographie, *Memories of My Life* (1890), celui-ci raconte qu'il découvrit ce concept lorsqu'au cours d'une promenade à Naworth Castle, une averse l'avait contraint à se mettre à l'abri.

Selon S.M. Stigler (1989), T.M. Porter (1986) effectuait des recherches historiques lorsqu'il retrouva un article oublié, écrit par F. Galton en 1890 dans *The North American Review*, sous le titre « Kinship and Correlation ». Dans cet article qu'il publia juste après sa découverte, Galton (1908) expliquait la nature et l'importance du concept de corrélation.

Cette découverte est liée aux activités antérieures du mathématicien et notamment à ses travaux sur l'hérédité et la régression linéaire. Dès 1860, F. Galton s'était intéressé à ce domaine d'étude. Il publia un ouvrage intitulé *Natural Inheritance* (1889) qui fut le point de départ de ses réflexions sur la corrélation.

En 1888, dans un article destiné à la Royal Statistical Society intitulé « Co-relations and their measurement chiefly from anthropometric data », Galton utilisa pour la première fois le terme de « corrélation ». Il hésitait encore entre les termes de « co-relation » et « corrélation » et parla d'« indice de corrélation ». En revanche, il ne faisait pas mention d'un concept de corrélation négative. Selon Stigler (1989), il semblerait que F. Galton n'ait envisagé la corrélation que comme une relation positive.

Karl Pearson écrivit en 1920 que la corrélation avait été créée par F. Galton dont l'ouvrage *Natural Inheritance* (1889) l'avait poussé à étudier lui-même ce concept avec deux autres chercheurs, Walter Frank Raphael Weldon et **Francis Y. Edgeworth**. K. Pearson et F.Y. Edgeworth développèrent la théorie de la corrélation.

W.F.R. Weldon proposa de donner au coefficient de corrélation le nom de « fonction de Galton ». Quant à F.Y. Edgeworth, il remplaça le terme « indice de co-relation » de F. Galton et celui de « fonction de Galton » de W.F.R. Weldon par « coefficient de corrélation ».

Selon Mudholkar (1982), Karl Pearson systématisa l'analyse de corrélation et établit la théorie de la corrélation pour trois variables. Des chercheurs de l'University College, et notamment **George Udny Yule** qui y fut son assistant, s'intéressèrent au développement de la corrélation multiple.

Puis ce fut en 1904 la publication par Charles Spearman de la première étude sur la corrélation des rangs.

Parmi les travaux menés dans ce domaine, on peut relever plus particulièrement ceux de G.U. Yule, qui dans un article intitulé « Why do we sometimes get non-sense-correlation between time-series » (1926) mit en lumière le problème de l'interprétation d'une analyse de corrélation. Les questions de corrélation robuste ont été abordées par Mosteller et **John Wilder Tukey** (1977).

ASPECTS MATHÉMATIQUES

Coefficient de corrélation linéaire simple

Le terme de corrélation linéaire simple est utilisé dans le cas d'une mesure de la **dépendance** linéaire entre deux **variables quantitatives** X et Y (voir aussi **régression linéaire simple**).

Si X et Y sont des variables aléatoires qui suivent une distribution conjointe (inconnue), nous pouvons alors définir le coefficient de corrélation entre X et Y comme suit :

Le coefficient de corrélation linéaire simple est égal au quotient de la **covariance** entre X et Y par le produit de leurs écarts-types :

$$\rho = \frac{Cov\,(X,Y)}{\sigma_X \sigma_Y}$$

où $Cov\,(X,Y)$ est la covariance mesurée entre X et Y ; σ_X et σ_Y sont les écarts-types respectifs de X et Y.

Si nous disposons d'un **échantillon** de taille n, (X_1, Y_1), $(X_2, Y_2), \ldots, (X_n, Y_n)$ tiré d'une distribution conjointe, la quantité :

$$r = \frac{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)\left(Y_i - \bar{Y}\right)}{\sqrt{\sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2 \sum_{i=1}^{n} \left(Y_i - \bar{Y}\right)^2}}$$

est une **estimation** de ρ ; c'est la corrélation d'**échantillonnage**.

Si nous symbolisons $\left(X_i - \bar{X}\right)$ par x_i et $\left(Y_i - \bar{Y}\right)$ par y_i, nous pouvons écrire cette équation sous la forme :

$$r = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\left(\sum_{i=1}^{n} x_i^2\right)\left(\sum_{i=1}^{n} y_i^2\right)}}.$$

Test d'hypothèse

Quand l'échantillon est extrait d'une **population** normale conjointe, il est naturel de vouloir tester des **hypothèses** concernant la véritable **valeur** de ρ.

Pour tester l'**hypothèse nulle**

$$H_0 \quad : \quad \rho = 0$$

contre l'**hypothèse alternative**

$$H_1 \quad : \quad \rho \neq 0,$$

nous calculons la **statistique** t :

$$t = \frac{r}{S_r}$$

où S_r est l'écart type de l'**estimateur** r :

$$S_r = \sqrt{\frac{1 - r^2}{n - 2}}.$$

Sous H_0, la statistique t suit une **loi de Student** avec $n - 2$ degrés de liberté.

Pour un **seuil de signification** α donné, on rejette H_0 si $|t| \geq t_{\frac{\alpha}{2}, n-2}$; la valeur de $t_{\frac{\alpha}{2}, n-2}$ est la **valeur critique** du test donnée dans la **table de Student**.

Coefficient de corrélation multiple

Le coefficient de corrélation multiple permet de déterminer si l'hyperplan estimé par une **régression linéaire multiple** est correctement ajusté par rapport au nuage de points.

La valeur du **coefficient de détermination** multiple R^2 est égale à :

$$\begin{aligned} R^2 &= \frac{\text{Variation expliquée}}{\text{Variation totale}} \\ &= \frac{\sum_{i=1}^{n} \left(\hat{Y}_i - \bar{Y}\right)^2}{\sum_{i=1}^{n} \left(Y_i - \bar{Y}\right)^2} \end{aligned}$$

Elle correspond au carré du coefficient de corrélation multiple. Nous notons que

$$0 \leq R^2 \leq 1.$$

Dans le cas de la **régression linéaire simple**, nous trouvons la relation suivante :

$$r = \text{sign}(\hat{\beta}_1)\sqrt{R^2}$$

où $\hat{\beta}_1$ est l'estimateur du coefficient de régression β_1 ayant pour valeur :

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (X_i - \bar{X})^2}.$$

DOMAINES ET LIMITATIONS

S'il existe une relation linéaire entre deux variables, le coefficient de corrélation est égal à 1 ou -1.

Une relation positive $(+)$ signifie que les deux variables varient dans le même sens. Si les individus obtiennent des scores élevés à la première variable (par exemple la **variable indépendante**), ils auront tendance à avoir également des scores élevés à la deuxième variable (**variable dépendante**). L'inverse est également vrai.

Une relation négative $(-)$ signifie que les individus qui ont des scores élevés pour la première

variable auront tendance à obtenir de petits scores pour la deuxième et inversement.

Il est à noter que si les variables sont indépendantes, alors le coefficient de corrélation est égal à zéro. Par contre, la réciproque n'est pas nécessairement vraie.

Le fait que deux ou plusieurs variables soient reliées entre elles au sens statistique du terme n'est pas suffisant pour conclure à l'existence d'une relation de cause à effet. L'existence d'une corrélation statistique n'est donc pas une preuve de causalité.

Les ouvrages de statistique proposent de nombreux coefficients de corrélation. Le choix de celui qu'il faut employer pour des données particulières repose sur différents facteurs, comme :

- le genre d'échelle de mesure utilisée pour exprimer la variable ;
- la nature de la distribution sous-jacente (continue ou discrète) ;
- les caractéristiques de la distribution des scores (linéaire ou non linéaire).

EXEMPLE

À l'aide des données des deux variables X et Y du tableau ci-dessous, voici les calculs nécessaires pour obtenir le coefficient de corrélation entre les deux variables :

N° d'ordre	X	Y	$x = X\text{-}\bar{X}$	$y = Y\text{-}\bar{Y}$	xy	x^2	y^2
1	174	64	-1,5	-1,3	1,95	2,25	1,69
2	175	59	-0,5	-6,3	3,14	0,25	36,69
3	180	64	4,5	-1,3	-5,85	20,25	1,69
4	168	62	-7,5	-3,3	24,75	56,25	10,89
5	175	51	-0,5	-14,3	7,15	0,25	204,49
6	170	60	-5,5	-5,3	29,15	30,25	28,09
7	170	68	-5,5	2,7	-14,85	30,25	7,29
8	178	63	2,5	-2,3	-5,75	6,25	5,29
9	187	92	11,5	26,7	307,05	132,25	712,89
10	178	70	2,5	4,7	11,75	6,25	22,09
Total	1755	653			358,5	284,5	1034,1

$$\bar{X} = 175,5 \quad \text{et} \quad \bar{Y} = 65,3$$

Appliquons la formule :

$$\begin{aligned} r &= \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\left(\sum_{i=1}^{n} x_i^2\right)\left(\sum_{i=1}^{n} y_i^2\right)}} \\ &= \frac{358,5}{\sqrt{284,5 \cdot 1034,1}} = 0,66. \end{aligned}$$

Test d'hypothèse

Nous pouvons calculer l'**écart type** estimé de r :

$$S_r = \sqrt{\frac{1-r^2}{n-2}} = \sqrt{\frac{0,56}{10-2}} = 0,266.$$

Calculons la **statistique** t :

$$t = \frac{r-0}{S_r} = \frac{0,66}{0,266} = 2,485.$$

Si nous choisissons un **seuil de signification** α de 5 %, la **valeur** de la **table de Student** $t_{0,025,8}$ est égale à $2,306$.

Puisque $|t| = 2,485 > t_{0,025,8} = 2,306$, l'**hypothèse nulle** H_0: $\rho = 0$ est rejetée.

MOTS CLÉS

Analyse de régression *(Regression analysis)*
Coefficient de corrélation des rangs de Kendall *(Kendall rank correlation coefficient)*
Coefficient de corrélation des rangs de Spearman *(Spearman rank correlation coefficient)*
Coefficient de détermination *(Coefficient of determination)*
Covariance *(Covariance)*
Dépendance *(Dependence)*
Régression linéaire multiple *(Multiple linear regression)*
Régression linéaire simple *(Simple linear regression)*

RÉFÉRENCES

Galton, F. (1888). *Co-relations and their measurement, chiefly from anthropological data. Proceedings of the Royal Society of London.* 45, pp. 135-145.

Galton, F. (1889). *Natural Inheritance.* The Macmillan Company, London.

Galton, F. (1890). Kinship and Correlation. *North American Review*, 150, pp. 419-431.

Galton, F. (1908). Memories of my Life. Methuen, London.

Mosteller, F. and Tukey, J.W. (1977). *Data Analysis and Regression : A Second Course in Statistics.* Addison-Wesley, Reading, Mass.

Mudholkar, G.S. (1982). Multiple correlation coefficient. In Kotz, S. and Johnson, N.L. (editors). *Encyclopedia of Statistical Sciences.* Vol. 5. John Wiley & Sons, New York.

Pearson, K. (1920). Studies in the history of Statistics and probability. *Biometrika*, 13, pp. 25-45. Reprinted in Studies in *The history of Statistics and probability.* Vol. I. Edited by Pearson, E.S. and Kendall, M.G., Charles Griffin & Company Ltd.

Porter, T.M. (1986). *The rise of statistical thinking, 1820-1900.* Princeton University Press, Princeton NJ.

Stigler, S. (1989). Francis Galton's Account of the Invention of Correlation.*Statistical Science*, 4, pp. 73-79.

Yule, G.U. (1897). On the Theory of Correlation. *Journal of the Royal Statistical Society*, 60, pp. 812-854.

Yule, G.U. (1926) Why do we sometimes get nonsense-correlations between time-series? A study sampling and the nature of time-series. *Journal of the Royal Statistical Society*, (2) 89, pp. 1-64.

COEFFICIENT DE CORRÉLATION DES RANGS DE KENDALL

Kendall rank correlation coefficient

Le coefficient de corrélation des rangs de Kendall (Kendall τ) est une mesure de corrélation non paramétrique. Il sert à déterminer la **relation** qui existe entre deux séries de données.

HISTORIQUE

Ce coefficient de corrélation des rangs fut déjà discuté dans les années 1900 par Fechner (1897), Lipps (1906) et Deuchler (1914).

Kendall (1938) ne l'a pas seulement redécouvert de manière indépendante, mais il l'a aussi étudié dans un esprit (non paramétrique) que l'on utilise encore aujourd'hui, ce qui justifie son appellation.

Sa monographie (1970) contient une exposition détaillée et complète de la théorie ainsi qu'une bibliographie.

ASPECTS MATHÉMATIQUES

Soit deux variables (X, Y) observées sur un **échantillon** de taille n fournissant n paires d'observations (X_1, Y_1), $(X_2, Y_2), \ldots, (X_n, Y_n)$. On obtient une indication de la corrélation entre X et Y en ordonnant les valeurs X_i en ordre croissant et en comptant le nombre de valeurs Y_i correspondantes ne satisfaisant pas cet ordre.

On note par Q le nombre d'inversions nécessaires parmi les valeurs de Y pour obtenir le même ordre (croissant) que celui des valeurs de X.

Comme il y a $\frac{n(n-1)}{2}$ paires distinctes pouvant être formées, on a : $0 \leq Q \leq \frac{n(n-1)}{2}$; la valeur 0 est atteinte lorsque toutes les valeurs Y_i sont déjà en ordre croissant et la valeur $\frac{n(n-1)}{2}$ est atteinte lorsque les valeurs Y_i sont totalement dans l'ordre

inverse des X_i, chaque paire devant être échangée pour obtenir l'ordre recherché.

Le coefficient de corrélation des rangs de Kendall, désigné par τ, est défini par :

$$\tau = 1 - \frac{4Q}{n\,(n-1)}.$$

Si toutes les paires sont dans l'ordre croissant,

$$\tau = 1 - \frac{4 \cdot 0}{n\,(n-1)} = 1.$$

Par contre, si toutes les paires sont dans l'ordre inverse,

$$\tau = 1 - \frac{4 \cdot \frac{1}{2} \cdot n\,(n-1)}{n\,(n-1)} = -1.$$

Il est possible de donner une définition équivalente du coefficient des rangs de Kendall de la manière suivante : deux observations sont dites concordantes si les deux membres de l'une des observations sont plus grands que le membre respectif de l'autre observation. Par exemple, $(0,9\,;\,1,1)$ et $(1,5\,;\,2,4)$ sont deux observations concordantes puisque $0,9 < 1,5$ et $1,1 < 2,4$. Deux observations sont dites discordantes si les deux membres d'une des observations sont d'ordre opposé par rapport au membre respectif de l'autre observation. Par exemple, $(0,8\,;\,2,6)$ et $(1,3\,;\,2,1)$ sont deux observations discordantes puisque $0,8 < 1,3$ et $2,6 > 2,1$.

Soit N_c et N_d le nombre de paires d'observations respectivement concordantes et discordantes.

Deux paires pour lesquelles $X_i = X_j$ ou $Y_i = Y_j$ ne sont ni concordantes ni discordantes et ne sont donc comptabilisées ni dans N_c ni dans N_d.

Avec les notations introduites, le coefficient des rangs de Kendall est défini par :

$$\tau = \frac{2\,(N_c - N_d)}{n\,(n-1)}.$$

On note que lorsqu'il n'y a pas de paires pour lesquelles $X_i = X_j$ ni $Y_i = Y_j$, les deux formulations de τ sont exactement les mêmes. Dans le cas contraire, les valeurs fournies par les deux formules peuvent être différentes.

Test d'hypothèse

Le coefficient de corrélation des rangs de Kendall est souvent utilisé comme test statistique afin de déterminer s'il existe une relation entre deux variables aléatoires. Le test peut être un **test bilatéral** ou un **test unilatéral**. Les hypothèses sont alors :

A. *Cas bilatéral :*

H_0 : les X et les Y sont mutuellement indépendants ;
H_1 : il y a soit une corrélation positive soit une corrélation négative entre les X et les Y.

On parle de corrélation positive si les grandes valeurs de X ont tendance à être associées avec les grandes valeurs de Y et les petites valeurs de X avec les petites valeurs de Y. On parle de corrélation négative si les grandes valeurs de X ont tendance à être associées avec les petites valeurs de Y et vice-versa.

B. *Cas unilatéral :*

H_0 : les X et les Y sont mutuellement indépendants ;
H_1 : il y a une corrélation positive entre les X et les Y.

C. *Cas unilatéral :*

H_0 : les X et les Y sont mutuellement indépendants ;
H_1 : il y a une corrélation négative entre les X et les Y.

Le test statistique est donc défini de la façon suivante :

$$T = N_c - N_d.$$

Règles de décision

Les règles de décision seront différentes selon les hypothèses posées. C'est ainsi que l'on a les règles de décision A, B et C relatives aux cas précédents.

Règle de décision A

On rejette H_0 au **seuil de signification** α si

$$T > t_{n,1-\frac{\alpha}{2}} \text{ ou } T < t_{n,\frac{\alpha}{2}}$$

où t est la **valeur critique** du test donnée par la table de Kendall.

Sinon on accepte H_0.

Règle de décision B

On rejette H_0 au seuil de signification α si

$$T > t_{n,1-\alpha}.$$

Sinon on accepte H_0.

Règle de décision C

On rejette H_0 au seuil de signification α si

$$T < t_{n,\alpha}.$$

Sinon on accepte H_0.

Il est également possible d'utiliser

$$\tau = 1 - \frac{4Q}{n\,(n-1)}$$

comme test statistique.

Lorsque X et Y sont distribuées indépendamment dans la **population**, la distribution exacte de τ a pour **espérance mathématique** zéro et pour **variance** :

$$\sigma_\tau^2 = \frac{2\,(2n+5)}{9n\,(n-1)}$$

et tend vers une distribution normale très rapidement, l'approximation étant suffisamment bonne pour $n \geq 10$.

Dans ce cas, pour tester l'**indépendance** avec un seuil de signification α, il suffit de vérifier si τ se situe hors des limites :

$$\pm z_{1-\alpha} \cdot \sigma_\tau$$

où $z_{1-\alpha}$ est la valeur de la **table normale** et de rejeter l'hypothèse d'indépendance si c'est le cas.

DOMAINES ET LIMITATIONS

On utilise le coefficient de corrélation des rangs de Kendall comme **test d'hypothèse** en vue d'étudier la dépendance entre deux variables aléatoires. Il peut donc être considéré comme un **test d'indépendance**.

En tant que mesure de corrélation non paramétrique, il peut aussi être utilisé avec des données nominales et ordinales.

Une mesure de corrélation entre deux variables doit satisfaire les points suivants :

1. elle prend ses valeurs entre -1 et $+1$;
2. on parle de corrélation positive entre X et Y si la valeur du coefficient de corrélation est positive ; une corrélation « positive » parfaite correspond à une valeur de $+1$;
3. on parle de corrélation négative entre X et Y si la valeur du coefficient de corrélation est négative ; une corrélation parfaite correspond à une valeur de -1 ;
4. on parle de corrélation nulle entre X et Y si le coefficient de corrélation est proche de zéro ; on dit aussi que X et Y ne sont pas corrélées.

Le coefficient de corrélation des rangs de Kendall présente les avantages suivants :

- les données peuvent être des observations non numériques pour autant qu'elles puissent être classées selon un critère déterminé ;
- il est simple à calculer ;
- le test statistique associé ne fait pas d'**hypothèse** de base sur la forme de la distribution de la **population** d'où sont tirés les échantillons.

La table de Kendall donne les valeurs théoriques de la **statistique** τ du coefficient de corrélation des rangs de Kendall employé comme test statistique sous l'hypothèse d'**indépendance** de deux variables aléatoires.

On peut trouver une table de Kendall chez Kaarsemaker et van Wijngaarden (1953).

Voici un extrait de la table de Kendall pour $n = 4, \ldots, 10$ et $\alpha = 0,01$ et $0,05$:

n	$\alpha = 0,01$	$\alpha = 0,05$
4	6	4
5	8	6
6	11	9
7	15	11
8	18	14
9	22	16
10	25	19

EXEMPLE

Dans cet exemple, on fait passer des tests d'intelligence à 8 couples de vrais jumeaux. Le but est de voir s'il y a **indépendance** entre les tests de celui qui est né en premier et ceux de celui qui est né en second.

Les données se trouvent dans le tableau ci-dessous, les scores plus élevés correspondant à de meilleurs résultats aux tests.

Couple de jumeaux	Né le 1er X_i	Né en second Y_i
1	90	88
2	75	79
3	99	98
4	60	66
5	72	64
6	83	83
7	83	88
8	90	98

On classe ensuite les couples par ordre croissant des X et on détermine les couples concordants et les couples discordants. Cela donne :

Couple de jumeaux (X_i, Y_i)	Couples concordants	Couples discordants
$(60, 66)$	6	1
$(72, 64)$	6	0
$(75, 79)$	5	0
$(83, 83)$	3	0
$(83, 88)$	2	0
$(90, 88)$	1	0
$(90, 98)$	0	0
$(99, 98)$	0	0
	$N_c = 23$	$N_d = 1$

Le coefficient de corrélation des rangs de Kendall est donné par :

$$\begin{aligned} \tau &= \frac{2\,(N_c - N_d)}{n\,(n-1)} \\ &= \frac{2\,(23-1)}{8 \cdot 7} \\ &= 0,7857. \end{aligned}$$

Comme il y a plusieurs observations pour lesquelles $X_i = X_j$ ou $Y_i = Y_j$, la valeur du coefficient défini par :

$$\begin{aligned} \tau &= 1 - \frac{4Q}{n\,(n-1)} \\ &= 1 - \frac{4 \cdot 1}{56} \\ &= 0,9286 \end{aligned}$$

est différente.

Dans les deux cas, on constate une corrélation positive entre les tests d'intelligence.

On effectue maintenant le **test d'hypothèse**. On pose :

H_0 : il y a indépendance entre les tests d'intelligence d'un couple de jumeaux ;
H_1 : il y a une corrélation positive entre les tests d'intelligence.

On choisit un **seuil de signification** $\alpha = 0,05$. Puisqu'on se trouve dans le cas B, on rejette H_0 si

$$T > t_{8,0,95}$$

où $T = N_c - N_d$ et $t_{8,0,95}$ est la **valeur** de la table de Kendall. Comme $T = 22$ et $t_{8,0,95} = 14$, on rejette H_0.

On peut donc conclure qu'il y a une corrélation positive entre les résultats des tests d'intelligence d'un couple de jumeaux.

MOTS CLÉS

Test d'hypothèse *(Hypothesis testing)*
Test d'indépendance *(Test of independence)*
Test non paramétrique *(Nonparametric test)*

RÉFÉRENCES

Deuchler, G. (1914). Über die Methoden der Korrelationsrechnung in der Pädagogik und Psychologie. *Zeitung für Pädagogische Psycholologie und Experimentelle Pädagogik*, 15, pp. 114-131, pp. 145-159, pp. 229-242.

Fechner, G.T. (1897). *Kollektivmasslehre.* W. Engelmann, Leipzig.

Kaarsemaker, L. and van Wijngaarden, A. (1953). Tables for use in rank correlation. *Statistica Neerlandico*, 7, p. 53.

Kendall, M.G. (1938). A new measure of rank correlation, *Biometrika*, 30, pp. 81-93.

Kendall, M.G. (1948). *Rank correlation methods.* Charles Griffin & Company Ltd.

Lipps, G.F. (1906). *Die Psychischen Mass-methoden.* F. Vieweg und Sohn, Braunschweig, Germany.

COEFFICIENT DE CORRÉLATION DES RANGS DE SPEARMAN

Spearman rank correlation coefficient

Le coefficient de corrélation des rangs de Spearman est une mesure de corrélation non paramétrique. Il sert à déterminer la **relation** qui existe entre deux séries de données.

HISTORIQUE

Charles Spearman était un psychologue. C'est lui qui, pour la première fois en 1904, a introduit le coefficient de corrélation des rangs. Souvent appelé le ρ de Spearman, c'est l'une des plus anciennes statistiques de rang utilisées encore aujourd'hui.

ASPECTS MATHÉMATIQUES

Soient $(X_1, X_2, \ldots, X_n)$ et $(Y_1, Y_2, \ldots, Y_n)$ deux échantillons de taille n. On note R_{X_i} le rang que prend X_i par rapport aux autres valeurs de l'échantillon des X, pour $i = 1, 2, \ldots, n$. Ainsi $R_{X_i} = 1$ si X_i est la plus petite valeur des X, $R_{X_i} = 2$ si X_i est la deuxième plus petite valeur, etc., jusqu'à $R_{X_i} = n$ si X_i est la plus grande valeur des X. De même, on note R_{Y_i} le rang de Y_i, pour $i = 1, 2, \ldots, n$.

Le coefficient de corrélation des rangs de Spearman, généralement désigné par ρ, est défini par :

$$\rho = 1 - \frac{6 \sum_{i=1}^{n} d_i^2}{n\left(n^2 - 1\right)}$$

où $d_i = R_{X_i} - R_{Y_i}$.

Si plusieurs observations ont exactement la même valeur, on attribue un rang moyen à ces observations. S'il y a beaucoup de rangs moyens, il convient de faire une correction et de calculer :

$$\rho = \frac{S_x + S_y - \sum_{i=1}^{n} d_i^2}{2\sqrt{S_x \cdot S_y}}$$

où

$$S_x = \frac{n\left(n^2 - 1\right) - \sum_{i=1}^{g} \left(t_i^3 - t_i\right)}{12}$$

avec g le nombre de groupes de rangs moyens et t_i la taille du groupe de rang i pour l'échantillon des X, et

$$S_y = \frac{n\left(n^2-1\right) - \sum_{j=1}^{h}\left(t_j^3 - t_j\right)}{12}$$

avec h le nombre de groupes de rangs moyens et t_j la taille du groupe de rang j pour l'échantillon des Y.

(S'il n'y a pas de rangs moyens, les observations sont vues comme autant de groupes de taille 1. Par conséquent, $g = h = n$ et $t_i = t_j = 1$ pour $i, j = 1, 2, \ldots, n$ et $S_x = S_y = \frac{n\left(n^2-1\right)}{12}$.)

Test d'hypothèse

Le coefficient de corrélation des rangs de Spearman est souvent utilisé comme test statistique afin de déterminer s'il existe une relation entre deux variables aléatoires. Le test peut être un **test bilatéral** ou un **test unilatéral**. Les hypothèses sont alors :

A. *Cas bilatéral :*

H_0 : X et Y sont mutuellement indépendants ;
H_1 : il y a soit une corrélation positive soit une corrélation négative entre X et Y.

On parle de corrélation positive si les grandes valeurs de X ont tendance à être associées avec les grandes valeurs de Y et les petites valeurs de X avec les petites valeurs de Y.

On parle de corrélation négative si les grandes valeurs de X ont tendance à être associées avec les petites valeurs de Y et vice-versa.

B. *Cas unilatéral :*

H_0 : X et Y sont mutuellement indépendants ;
H_1 : il y a une corrélation positive entre X et Y.

C. *Cas unilatéral :*

H_0 : X et Y sont mutuellement indépendants ;
H_1 : il y a une corrélation négative entre X et Y.

Règles de décision

Les règles de décision seront différentes selon les hypothèses posées. C'est ainsi que l'on a les règles de décision A, B et C relatives aux cas précédents.

Règle de décision A

On rejette H_0 au **seuil de signification** α si

$$\rho > t_{n,1-\frac{\alpha}{2}} \text{ ou } \rho < t_{n,\frac{\alpha}{2}}$$

où t est la **valeur critique** du test donnée par la table de Spearman.

Règle de décision B

On rejette H_0 au seuil de signification α si

$$\rho > t_{n,1-\alpha}.$$

Règle de décision C

On rejette H_0 au seuil de signification α si

$$\rho < t_{n,\alpha}.$$

Remarque : la notation t que Spearman utilise n'est aucunement liée à celle de Student.

DOMAINES ET LIMITATIONS

On utilise le coefficient de corrélation des rangs de Spearman comme **test d'hypothèse** en vue d'étudier la dépendance entre deux variables aléatoires. Il peut donc être considéré comme un **test d'indépendance**.

En tant que mesure de corrélation non paramétrique, il peut aussi être utilisé avec des données nominales et ordinales.

Une mesure de corrélation entre deux variables doit satisfaire les points suivants :

1. elle prend ses valeurs entre -1 et $+1$;
2. on parle de corrélation positive entre X et Y si la valeur du coefficient est positive ; une corrélation parfaite correspond à une valeur de $+1$;
3. on parle de corrélation négative entre X et Y si la valeur du coefficient est négative ; une corrélation parfaite correspond à une valeur de -1 ;
4. on parle de corrélation nulle entre X et Y si le coefficient de corrélation est proche de zéro ; on dit aussi que X et Y ne sont pas corrélées.

Le coefficient de corrélation des rangs de Spearman présente les avantages suivants :

- les données peuvent être des observations non numériques pour autant qu'elles puissent être classées selon un critère déterminé ;
- il est simple à calculer ;
- le test statistique associé ne fait pas d'**hypothèse** de base sur la forme de la distribution de la **population** d'où sont tirés les échantillons.

La table de Spearman donne les valeurs théoriques du coefficient de corrélation des rangs de Spearman sous l'hypothèse d'**indépendance** de deux variables aléatoires.

Voici un extrait de la table de Spearman pour $n = 6$, 7, 8 et $\alpha = 0,05$ et $0,02$:

n	$\alpha = 0,05$	$\alpha = 0,02$
6	$0,886$	$0,943$
7	$0,786$	$0,893$
8	$0,738$	$0,833$

Une table complète de Spearman peut être trouvée dans Glasser and Winter (1961).

EXEMPLE

Dans cet exemple, on fait passer des tests d'intelligence à huit couples de vrais jumeaux. Le but est de voir s'il y a **indépendance** entre les tests de celui qui est né en premier et ceux de celui qui est né en second.

Les données se trouvent dans le tableau ci-dessous, les scores plus élevés correspondant à de meilleurs résultats aux tests.

Couple de jumeaux	Né le 1er X_i	Né en second Y_i
1	90	88
2	75	79
3	99	98
4	60	66
5	72	64
6	83	83
7	86	86
8	92	95

On classe les X entre eux et les Y entre eux et on calcule d_i. Cela donne :

Couple de jumeaux	Né le 1er X_i	R_{X_i}	Né en second Y_i	R_{Y_i}	d_i
1	90	6	88	6	0
2	75	3	79	3	0
3	99	8	98	8	0
4	60	1	66	2	-1
5	72	2	64	1	1
6	83	4	83	4	0
7	86	5	86	5	0
8	92	7	95	7	0

On calcule ensuite le coefficient de corrélation des rangs de Spearman :

$$\begin{aligned} \rho &= 1 - \frac{6 \cdot \sum_{i=1}^{8} d_i^2}{n\,(n^2-1)} = 1 - \frac{6 \cdot 2}{8\,(8^2-1)} \\ &= 0,9762. \end{aligned}$$

Il y a donc une corrélation positive (presque parfaite) entre les tests d'intelligence.

On suppose maintenant que les résultats pour les couples 7 et 8 soient changés. On aura alors le nouveau tableau ci-dessous :

Couple de jumeaux	Né le 1er X_i	R_{X_i}	Né en second Y_i	R_{Y_i}	d_i
1	90	6,5	88	5,5	1
2	75	3	79	3	0
3	99	8	98	7,5	−0,5
4	60	1	66	2	−1
5	72	2	64	1	1
6	83	4,5	83	4	0,5
7	83	4,5	88	5,5	−1
8	90	6,5	98	7,5	−1

Comme il y a des rangs moyens, on utilise la formule

$$\rho = \frac{S_x + S_y - \sum_{i=1}^{8} d_i^2}{2\sqrt{S_x \cdot S_y}}$$

pour calculer le coefficient de corrélation des rangs de Spearman. On calcule d'abord S_x :

$$\begin{aligned} S_x &= \frac{n\left(n^2-1\right) - \sum_{i=1}^{g}\left(t_j^3 - t_j\right)}{12} \\ &= \frac{8\left(8^2-1\right) - \left(2^3 - 2 + 2^3 - 2\right)}{12} \\ &= 41 \end{aligned}$$

puis S_y :

$$\begin{aligned} S_y &= \frac{n\left(n^2-1\right) - \sum_{j=1}^{h}\left(t_j^3 - t_j\right)}{12} \\ &= \frac{8\left(8^2-1\right) - \left(2^3 - 2 + 2^3 - 2\right)}{12} \\ &= 41. \end{aligned}$$

La valeur de ρ est alors la suivante :

$$\rho = \frac{41 + 41 - (1 + 0 + \ldots + 1)}{2\sqrt{41 \cdot 41}} = \frac{76,5}{82} = 0,9329.$$

Là aussi, on voit qu'il y a une corrélation positive presque parfaite.

On effectue maintenant le **test d'hypothèse**. On pose :

H_0 : il y a indépendance entre les tests d'intelligence d'un couple de jumeaux ;
H_1 : il y a une corrélation positive entre les tests d'intelligence.

On choisit un **seuil de signification** $\alpha = 0,05$. Puisqu'on se trouve dans le cas B, on rejette H_0 si $\rho > t_{8,0,95}$ où $t_{8,0,95}$ est la **valeur** de la table de Spearman, c'est-à-dire si :

$$\rho > 0,738.$$

Dans les deux cas ($\rho = 0,9762$ et $\rho = 0,9329$), on rejette H_0. Nous pouvons donc conclure qu'il y a une corrélation positive entre les résultats du test d'intelligence d'un couple de jumeaux.

MOTS CLÉS

Test d'hypothèse *(Hypothesis testing)*
Test d'indépendance *(Test of independence)*
Test non paramétrique *(Nonparametric test)*

RÉFÉRENCES

Glasser, G.J. and Winter, R.F. (1961). Critical values of the coefficient of rank correlation for testing thehypothesis of independance. *Biometrika*, 48, pp. 444-448.

Spearman, C. (1904). The Proof and Measurement of Association between Two Things. *Amer. J. Psych.*, Vol. 15, pp. 72-101.

COEFFICIENT DE CORRÉLATION MULTIPLE

Multiple correlation coefficient

Voir **coefficient de corrélation**.

COEFFICIENT DE DÉTERMINATION

Coefficient of determination

Le coefficient de détermination, noté R^2 est le quotient entre la variation expliquée et la variation totale dans un **modèle** de **régression linéaire simple** ou **multiple** :

$$R^2 = \frac{\text{Variation expliquée}}{\text{Variation totale}}.$$

Il est égal au carré du **coefficient de corrélation** et varie entre 0 et 1. Il s'exprime souvent en **pourcentage**.

HISTORIQUE

Voir **coefficient de corrélation**.

ASPECTS MATHÉMATIQUES

Considérons le **modèle** de régression multiple suivant :

$$Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip} + \varepsilon_i$$

pour $i = 1, \ldots, n$, où

Y_i	sont les variables dépendantes,
X_{ij}	$(i = 1, \ldots, n, j = 1, \ldots, p)$ sont les variables indépendantes,
ε_i	sont les termes d'**erreur** aléatoire non observables,
β_j	$(j = 1, \ldots, p)$ sont les paramètres à estimer.

L'**estimation** des paramètres $\beta_0, \beta_1, \ldots, \beta_p$ donne l'estimation

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{i1} 1 + \cdots + \hat{\beta}_p X_{ip}.$$

Le coefficient de détermination permet de mesurer la qualité de l'ajustement.

Pour déterminer la qualité des estimations de l'équation de **régression**, considérons l'écart entre la **valeur** observée et la valeur estimée de chaque **observation** de l'**échantillon**. Cet écart (ou **résidu**) s'exprime aussi de la manière suivante :

$$Y_i - \hat{Y}_i = \left(Y_i - \bar{Y}\right) - \left(\hat{Y}_i - \bar{Y}\right).$$

Nous démontrons (voir plus loin) que l'égalité suivante est également vraie :

$$\sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2 = \sum_{i=1}^{n} \left(Y_i - \bar{Y}\right)^2 - \sum_{i=1}^{n} \left(\hat{Y}_i - \bar{Y}\right)^2.$$

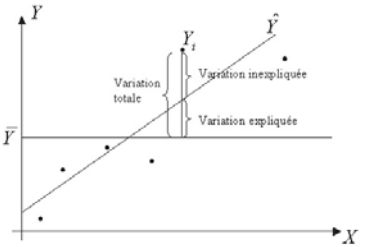
Cette égalité peut aussi s'exprimer différemment :

$$\begin{array}{lll} \sum\limits_{i=1}^{n} \left(Y_i - \bar{Y}\right)^2 & = \sum\limits_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2 & + \sum\limits_{i=1}^{n} \left(\hat{Y}_i - \bar{Y}\right)^2 \\ SC_{tot} & = SC_{res} & + SC_{reg} \end{array}$$

où

SC_{tot}	est la somme des carrés totale,
SC_{res}	la somme des carrés résiduelle et
SC_{reg}	la somme des carrés de la régression.

Ces concepts et leurs relations sont représentés dans le graphique ci-dessous :



À partir de ces concepts, nous pouvons définir R^2, qui est le coefficient de détermination. Il mesure la proportion de la variation dans la **variable** Y, qui est expliquée par l'équation de régression. Ainsi :

$$\begin{aligned} R^2 &= \frac{\text{Variation expliquée}}{\text{Variation totale}} \\ &= \frac{SC_{reg}}{SC_{tot}} \\ &= \frac{\sum\limits_{i=1}^{n} \left(\hat{Y}_i - \bar{Y}\right)^2}{\sum_{i=1}^{n} \left(Y_i - \bar{Y}\right)^2}. \end{aligned}$$

Si la fonction de régression doit servir à faire des prévisions, il est souhaitable que la valeur de R^2 soit très grande, car plus la valeur de R^2 est grande, plus celle de la variation inexpliquée est petite.

EXEMPLE

Le tableau suivant représente le produit national brut (PNB) et la demande des biens de première nécessité obtenus pour la période 1969-1980 dans un certain pays.

Année	PNB	Demande des biens de première nécessité
	X	Y
1969	50	6
1970	52	8
1971	55	9
1972	59	10
1973	57	8
1974	58	10
1975	62	12
1976	65	9
1977	68	11
1978	69	10
1979	70	11
1980	72	14

Nous cherchons à estimer la demande des biens de première nécessité en fonction du PNB suivant le **modèle**

$$Y_i = a + b \cdot X_i + \varepsilon_i, \quad i = 1, \ldots, 12.$$

L'**estimation** des paramètres a et b par la méthode des moindres carrés nous donne les estimateurs suivants :

$$\begin{aligned} \hat{b} &= \frac{\sum_{i=1}^{12} \left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{\sum_{i=1}^{12} \left(X_i - \bar{X}\right)^2} \\ &= 0,226 \end{aligned}$$

$$\begin{aligned} \hat{a} &= \bar{Y} - \hat{b} \cdot \bar{X} \\ &= -4,047 \end{aligned}$$

La droite estimée peut donc s'écrire

$$\hat{Y} = -4,047 + 0,226 \cdot X.$$

La qualité de l'ajustement des points par la droite nous est donnée par le coefficient de détermination

$$R^2 = \frac{\text{Variation expliquée}}{\text{Variation totale}} = \frac{\sum_{i=1}^{12} \left(\hat{Y}_i - \bar{Y}\right)^2}{\sum_{i=1}^{12} \left(Y_i - \bar{Y}\right)^2}.$$

Nous pouvons calculer la **moyenne**

$$\bar{Y} = \frac{\sum_{i=1}^{12} Y_i}{n} = 9,833,$$

Y_i	$\hat{Y}_i$	$\left(Y_i - \bar{Y}\right)^2$	$\left(\hat{Y}_i - \bar{Y}\right)^2$
6	7,260	14,694	6,622
8	7,712	3,361	4,500
9	8,390	0,694	2,083
10	9,294	0,028	0,291
8	8,842	3,361	0,983
10	9,068	0,028	0,586
12	9,972	4,694	0,019
9	10,650	0,694	0,667
11	11,328	1,361	2,234
10	11,554	0,028	2,961
11	11,780	1,361	3,789
14	12,232	17,361	4,754
Total		47,667	30,489

Nous avons donc

$$R^2 = \frac{30,489}{47,667} = 0,6396$$

ou en pourcentage :

$$R^2 = 63,96\ \%.$$

Nous pouvons conclure que selon le modèle choisi, 63,96 % de la variation de la demande des biens de première nécessité s'explique par la variation du PNB.

Il est évident que la **valeur** de R^2 ne peut pas dépasser 100 % ; la valeur 63,96 % est assez grande mais elle n'est pourtant pas suffisamment proche de 100 % pour qu'on ne tente pas d'améliorer le modèle.

Cela peut signifier qu'en plus du PNB, d'autres variables devraient être prises en considération pour une détermination plus fine de la fonction de demande en biens de première nécessité, donc que le PNB n'explique qu'une partie de la variation.

MOTS CLÉS

Analyse de régression *(Regression analysis)*
Coefficient de corrélation *(Correlation coefficient)*
Régression linéaire multiple *(Multiple linear regression)*
Régression linéaire simple *(Simple linear regression)*

COEFFICIENT DE VARIATION

Coefficient of variation

Le coefficient de variation est une **mesure de dispersion** relative décrivant l'**écart type** comme pourcentage de la **moyenne arithmétique**.

Ce coefficient permet de comparer la dispersion de variables quantitatives qui ne sont pas exprimées dans la même unité (par exemple, comparer les salaires dans différents pays, donc en différentes monnaies) ou qui ont des moyennes très différentes.

ASPECTS MATHÉMATIQUES

Le coefficient de variation CV est défini comme le rapport entre l'**écart type** et la **moyenne arithmétique** d'un ensemble d'observations :

$$CV = \frac{S}{\bar{x}} \cdot 100$$

pour un **échantillon**, où :

- S est l'écart type de l'échantillon et
- $\bar{x}$ la moyenne arithmétique de l'échantillon

ou alors :

$$CV = \frac{\sigma}{\mu} \cdot 100$$

pour une **population**, avec

- σ est l'écart type de la population et
- μ la moyenne de la population

C'est une valeur indépendante de l'unité de mesure.

EXEMPLE

Étudions la distribution des salaires dans deux entreprises de deux pays différents.

D'après un **sondage**, les moyennes arithmétiques et les écarts types des salaires sont les suivants :

$$\begin{array}{ll} \textit{Entreprise } A: & \bar{x} = 2\,500 \text{ EUR} \\ & S = 200 \text{ EUR} \end{array}$$

$$CV = \frac{200}{2\,500} \cdot 100 = 8$$

$$\begin{array}{ll} \textit{Entreprise } B: & \bar{x} = 1\,000 \text{ EUR} \\ & S = 50 \text{ EUR} \end{array}$$

$$CV = \frac{50}{1\,000} \cdot 100 = 5.$$

L'écart type représente donc 8 % de la moyenne arithmétique dans l'entreprise A, et 5 % dans l'entreprise B. La distribution des salaires est donc un peu plus homogène dans l'entreprise B que dans l'entreprise A.

MOTS CLÉS

Écart type *(Standard deviation)*
Mesure de dispersion *(Measure of dispersion)*
Moyenne arithmétique *(Arithmetic mean)*

RÉFÉRENCE

Johnson, N.L. and Leone, F.C. (1964). *Statistics and experimental design in engineering and the physical sciences.* John Wiley & Sons, New York.

COEFFICIENT DE YULE ET KENDALL

Yule and Kendall coefficient

Le coefficient de Yule et Kendall sert à mesurer l'asymétrie d'une **distribution de fréquences**. Il tient compte des positions relatives des quartiles par rapport à la **médiane**, c'est-à-dire qu'il compare l'étalement de la courbe à droite et à gauche de la médiane. Il appartient aux mesures d'asymétrie.

ASPECTS MATHÉMATIQUES

Dans une distribution symétrique, les quartiles sont à égale distance de chaque côté de la **médiane**. Par conséquent,

$$(Q_3 - M_d) - (M_d - Q_1) = 0$$

où

M_d	est la médiane,
Q_1	le premier quartile et
Q_3	le troisième quartile.

Si la distribution est asymétrique, l'égalité ci-dessus n'est plus vraie.

Le membre de gauche de l'équation ci-dessus peut être réarrangé de la façon suivante :

$$Q_3 + Q_1 - 2M_d.$$

Pour obtenir un coefficient d'asymétrie qui soit indépendant de l'unité de mesure, il faut encore diviser l'équation par la différence $(Q_3 - Q_1)$.

Le coefficient de Yule est alors :

$$C_Y = \frac{Q_3 + Q_1 - 2M_d}{Q_3 - Q_1}.$$

Si C_Y est positif, la distribution est étirée vers la droite, si C_Y est négatif, la distribution est étirée vers la gauche.

DOMAINES ET LIMITATIONS

Si le coefficient est positif, il traduit une courbe étirée vers la droite.

S'il est négatif, il traduit une courbe étirée vers la gauche.

Un coefficient proche de zéro signifie que la courbe est approximativement symétrique.

Le coefficient de Yule et Kendall peut varier entre -1 et $+1$. En effet, la **médiane** M_d se situant toujours entre le premier et le troisième **quartile**, on obtient les deux possibilités extrêmes suivantes :

1. $M_d = Q_1$;
2. $M_d = Q_3$.

Dans le premier cas, le coefficient devient :

$$C_Y = \frac{Q_3 + Q_1 - 2Q_1}{Q_3 - Q_1} = 1$$

et dans le deuxième cas, il devient :

$$C_Y = \frac{Q_3 + Q_1 - 2Q_3}{Q_3 - Q_1} = -1.$$

Dans ces conditions, il faut que le coefficient soit vraiment proche de zéro pour que la distribution puisse être considérée comme symétrique.

Ce coefficient ainsi que les autres mesures d'asymétrie n'ont d'intérêt que s'ils permettent de comparer les formes de deux ou plusieurs distributions. En effet, les résultats varient considérablement d'une formule à l'autre. Il est alors évident que les comparaisons doivent être faites avec la même formule.

EXEMPLE

On veut comparer la forme de la distribution du chiffre d'affaires journalier dans soixante-quinze épiceries sur deux années. On calcule alors dans les deux cas le coefficient de Yule et Kendall. Les données sont les suivantes :

Chiffre d'affaires (EUR)	Fréquences année 1	Fréquences année 2
$215 - 235$	4	25
$235 - 255$	6	15
$255 - 275$	13	9
$275 - 295$	22	8
$295 - 315$	15	6
$315 - 335$	6	5
$335 - 355$	5	4
$355 - 375$	4	3

Pour l'année 1, le premier **quartile**, la **médiane** et le troisième quartile valent :

$$\begin{aligned} Q_1 &= 268,46 \\ M_d &= 288,18 \\ Q_3 &= 310. \end{aligned}$$

Le coefficient est donc le suivant :

$$\begin{aligned} C_Y &= \frac{Q_3 + Q_1 - 2M_d}{Q_3 - Q_1} \\ &= \frac{310 + 268,46 - 2 \cdot 288,18}{310 - 268,46} \\ &= 0,05. \end{aligned}$$

Pour l'année 2, ces valeurs sont respectivement :

$$\begin{aligned} Q_1 &= 230 \\ M_d &= 251,67 \\ Q_3 &= 293,12. \end{aligned}$$

Le coefficient de Yule et Kendall est donc égal à :

$$\begin{aligned} C_Y &= \frac{Q_3 + Q_1 - 2M_d}{Q_3 - Q_1} \\ &= \frac{293,12 + 230 - 2 \cdot 251,67}{293,12 - 230} \\ &= 0,31. \end{aligned}$$

Pour l'année 1, le coefficient donnant un résultat très proche de zéro, on peut admettre que la distribution du chiffre d'affaires journalier dans les 75 épiceries est très voisine d'une distribution symétrique. Pour l'année 2, par contre, le coefficient est plus élevé ; cela signifie que la distribution est étirée vers la droite.

MOTS CLÉS

Mesure d'asymétrie *(Measure of skewness)*
Mesure de forme *(Measure of shape)*

RÉFÉRENCE

Yule, G.V. and Kendall, M.G. (1968). *An Introduction to the Theory of Statistics.* 14th edition. Charles Griffin & Company Ltd.

COEFFICIENT β_1 DE PEARSON

Coefficient of skewness

Le coefficient β_1 de Pearson sert à mesurer l'asymétrie d'une distribution. Il est basé sur la notion de **moment** de la distribution. Ce coefficient appartient aux mesures d'asymétrie.

HISTORIQUE

Entre la fin du XIXe siècle et le début du XXe siècle, **Karl Pearson** s'intéressa à de grands ensembles de données qui s'éloignaient parfois considérablement de la normalité, présentant notamment une asymétrie importante.

Il utilisa tout d'abord comme **mesure d'asymétrie** le coefficient suivant :

$$\text{asymétrie} = \frac{\bar{x} - M_o}{S},$$

où $\bar{x}$ représente la **moyenne arithmétique**, S l'**écart type** et M_o le **mode**.

Cette mesure vaut zéro si la distribution des données est symétrique.

Il découvrit empiriquement que pour une distribution modérément asymétrique (distribution gamma) :

$$M_o - \bar{x} \approx 3 \cdot (M_d - \bar{x}),$$

où M_d est la médiane. En substituant cette expression dans le coefficient précédent, on trouve la formule alternative suivante :

$$\text{asymétrie} = \frac{3 \cdot (\bar{x} - M_d)}{S}.$$

Par la suite, Karl Pearson (1894, 1895) introduisit un coefficient d'asymétrie, connu sous le nom de coefficient β_1, basé sur le calcul des moments centrés. Ce coefficient est plus difficile à calculer, mais il est plus élaboré et mieux adapté lorsque le nombre d'observations est grand.

On doit aussi à Karl Pearson le **coefficient β_2 de Pearson** qui sert à mesurer l'aplatissement d'une courbe. Ce coefficient est également fondé sur les moments de la distribution à étudier.

Des tables donnant les valeurs limites des coefficients β_1 et β_2 se trouvent dans l'ouvrage de Pearson et Hartley (1966, 1972).

ASPECTS MATHÉMATIQUES

Le coefficient d'asymétrie β_1 de Pearson est basé sur le **moment** centré d'ordre 3 d'une distribution égal à :

$$\mu_3 = E\left[(X-\mu)^3\right].$$

Pour obtenir un coefficient d'asymétrie qui soit indépendant de l'unité de mesure, on va diviser ce moment d'ordre 3 par l'**écart type** σ de la **population,** élevé à la puissance 3. Le coefficient que l'on obtient, désigné par $\sqrt{\beta_1}$, est donc égal à :

$$\sqrt{\beta_1} = \frac{\mu_3}{\sigma^3}.$$

L'**estimateur** de ce coefficient calculé sur un **échantillon** $(x_1, x_2, \ldots x_n)$ est noté par $\sqrt{b_1}$. Il est égal à :

$$\sqrt{b_1} = \frac{m_3}{S^3}$$

où n le nombre total d'observations, S^3 est l'écart type de l'échantillon élevé à la puissance 3 et m_3 est le moment centré d'ordre 3 de l'échantillon donné par :

$$m_3 = \frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - \bar{x})^3$$

avec $\bar{x}$ la **moyenne arithmétique**.

Pour le cas où une **variable aléatoire** X prend ses valeurs x_i avec les fréquences f_i, $i = 1, 2, \ldots, h$, le moment centré d'ordre 3 de l'échantillon est donné par la formule :

$$m_3 = \frac{1}{n} \cdot \sum_{i=1}^{h} f_i \cdot (x_i - \bar{x})^3 .$$

Si le coefficient est positif, la distribution est étirée à droite. S'il est négatif, il traduit une distribution étirée à gauche. S'il est proche de zéro, cela signifie que la distribution est approximativement symétrique.

Si l'échantillon est tiré d'une population normale, la **statistique** $\sqrt{b_1}$ suit approximativement une **loi normale** de **moyenne** égale à 0 et d'écart type égal à $\sqrt{\frac{6}{n}}$. Si la **taille d'échantillon** n est supérieure à 150, on peut utiliser la **table normale** pour tester l'**hypothèse** d'asymétrie.

DOMAINES ET LIMITATIONS

Ce coefficient ainsi que les autres mesures d'asymétrie n'ont d'intérêt que s'ils permettent de comparer les formes de deux ou plusieurs distributions.

EXEMPLE

On veut comparer la forme de la distribution du chiffre d'affaires journalier dans soixante-quinze épiceries pour deux années. On calcule alors dans les deux cas le coefficient d'asymétrie.

Les données sont les suivantes :

Chiffre d'affaires (EUR)	Fréquences année 1	Fréquences année 2
$215-235$	4	25
$235-255$	6	15
$255-275$	13	9
$275-295$	22	8
$295-315$	15	6
$315-335$	6	5
$335-355$	5	4
$355-375$	4	3

Le **moment** d'ordre 3 de l'**échantillon** est donné par :

$$m_3 = \frac{1}{n} \cdot \sum_{i=1}^{h} f_i \cdot (x_i - \bar{x})^3 .$$

Pour l'année 1, $n = 75$, $\bar{x} = 290,60$ et x_i est le centre de chaque **intervalle** de classe. Nous résumons les calculs dans le tableau suivant :

x_i	$x_i - \bar{x}$	f_i	$f_i\,(x_i - \bar{x})^3$
225	$-65,60$	4	$-1\,129\,201,664$
245	$-45,60$	6	$-568\,912,896$
265	$-25,60$	13	$-218\,103,808$
285	$-5,60$	22	$-3\,863,652$
305	$14,40$	15	$44\,789,760$
325	$34,40$	6	$244\,245,504$
345	$54,40$	5	$804\,945,920$
365	$74,40$	4	$1\,647\,323,136$
			$821\,222,410$

Comme $S = 33,88$, le coefficient d'asymétrie est égal à :

$$\sqrt{b_1} = \frac{\frac{1}{75} \cdot 821\,222,41}{(33,88)^3} = \frac{10\,949,632}{38\,889,307} = 0,282.$$

Pour l'année 2, $n = 75$ et $\bar{x} = 265,27$. Nous résumons les calculs dans le tableau suivant :

x_i	$x_i - \bar{x}$	f_i	$f_i\,(x_i - \bar{x})^3$
225	$-40,27$	25	$-1\,632\,213,81$
245	$-20,27$	15	$-124\,864,28$
265	$-0,27$	9	$-0,17$
285	19,73	8	61 473,98
305	39,73	6	376 371,09
325	59,73	5	1 065 663,90
345	79,73	4	2 027 588,19
365	99,73	3	2 976 063,94
			4 750 082,83

Comme $S = 42,01$, le coefficient d'asymétrie est égal à :

$$\sqrt{b_1} = \frac{\frac{1}{75} \cdot 4\,750\,082,83}{(42,01)^3} = \frac{63\,334,438}{74\,140,933} = 0,854\,.$$

Pour l'année 1, le coefficient donnant un résultat proche de zéro, ($\sqrt{b_1} = 0,282$), on peut admettre que la distribution du chiffre d'affaires journalier dans les soixante-quinze épiceries est très voisine d'une distribution symétrique. Pour l'année 2, par contre, le coefficient est plus élevé ; cela signifie que la distribution est étirée vers la droite.

MOTS CLÉS

Mesure d'asymétrie *(Measure of skewness)*
Mesure de forme *(Measure of shape)*

RÉFÉRENCES

Pearson, E.S. and Hartley, H.O. (1966,1972). *Biometrika Tables for Statisticians*, Vol. I and II. Cambridge University Press, Cambridge.

Pearson, K. (1894). 1948. Contributions to the Mathematical Theory of Evolution. I. pp. 1-40 in Karl Pearson, *Karl Pearson's Early Statistical Papers.* Cambridge University Press, Cambridge. First published as: On the Dissection of Asymmetrical Frequency Curves, *Philosophical Transactions of the Royal Society of London.* Series A, Vol. 185, pp. 71-110.

Pearson, K. (1895). 1948. Contributions to the Mathematical Theory of Evolution. II: Skew Variation in Homogeneous Material. pp. 41-112 in Karl Pearson, *Karl Pearson's Early Statistical Papers.* Cambridge University Press, Cambridge. First published in *Philosophical Transactions of the Royal Society of London.* Series A, Vol. 186, pp. 343-414.

COEFFICIENT β_2 DE PEARSON

Coefficient of kurtosis

Le coefficient β_2 de Pearson sert à mesurer l'aplatissement d'une courbe. Il est fondé sur les moments de la distribution à étudier. Il fait partie des mesures d'aplatissement.

HISTORIQUE

Voir **coefficient β_1 de Pearson**.

ASPECTS MATHÉMATIQUES

Le coefficient β_2 de Pearson est basé sur le **moment** centré d'ordre 4 d'une **distribution** qui est égal à :

$$\mu_4 = E\left[(X - \mu)^4\right].$$

Pour obtenir un coefficient d'aplatissement qui soit indépendant de l'unité de mesure, on divise ce moment d'ordre 4 par l'**écart type** de la **population** σ élevé à la puissance 4. Le coefficient β_2 de Pearson est donc égal à :

$$\beta_2 = \frac{\mu_4}{\sigma^4}.$$

L'**estimateur** de ce coefficient calculé sur un **échantillon** $(x_1, x_2, \ldots, x_n)$ est noté par b_2. Il est égal à :

$$b_2 = \frac{m_4}{S^4}$$

où n le nombre total d'observations et où S^4 est l'écart type de l'échantillon élevé à la puissance 4 et m_4 est le moment centré d'ordre 4 de l'échantillon donné par :

$$m_4 = \frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - \bar{x})^4$$

avec $\bar{x}$ la **moyenne arithmétique**.

Pour le cas où une **variable aléatoire** X prend ses valeurs x_i avec les fréquences f_i, $i = 1, 2, \ldots, h$, le moment centré d'ordre 4 de l'échantillon est donné par la formule :

$$m_4 = \frac{1}{n} \cdot \sum_{i=1}^{h} f_i \cdot (x_i - \bar{x})^4 .$$

DOMAINES ET LIMITATIONS

Pour une **loi normale**, le coefficient d'aplatissement β_2 est égal à 3. Par conséquent, la courbe sera qualifiée de platicurtique (c'est-à-dire plus plate que la loi normale) si on trouve un coefficient β_2 de Pearson inférieur à 3. Elle sera leptocurtique (c'est-à-dire plus pointue que la loi normale) si β_2 est supérieur à 3.

On va montrer que le coefficient β_2 de Pearson est égal à 3 pour la loi normale.

On sait que $\beta_2 = \frac{\mu_4}{\sigma^4}$, c'est-à-dire le **moment** centré d'ordre 4 divisé par l'**écart type** à la puissance 4. On peut montrer pour la loi normale que le moment centré d'ordre s noté μ_s satisfait la relation suivante :

$$\mu_s = (s-1)\,\sigma^2 \cdot \mu_{s-2}.$$

Cette formule est une formule de récurrence qui permet d'exprimer les moments d'ordre supérieur en fonction des moments d'ordre inférieur.

Compte tenu que $\mu_0 = 1$ (le moment d'ordre zéro de toute **variable aléatoire** est égal à l'unité en tant qu'**espérance mathématique** de la puissance zéro de cette **variable**) et que $\mu_1 = 0$ (le moment centré d'ordre 1 est nul pour toute variable aléatoire), on a :

$$\begin{aligned} \mu_2 &= \sigma^2 \\ \mu_3 &= 0 \\ \mu_4 &= 3\sigma^4 \\ &\textit{etc.} \end{aligned}$$

Le coefficient β_2 de Pearson est alors égal à :

$$\beta_2 = \frac{\mu_4}{\sigma^4} = \frac{3\sigma^4}{\sigma^4} = 3.$$

EXEMPLE

On désire étudier l'aplatissement de la distribution du chiffre d'affaires journalier de soixante-quinze épiceries. Calculons le coefficient β_2 de Pearson sur le relevé des **données** suivantes :

Tableau du chiffre d'affaires journalier de soixante-quinze épiceries

Chiffre d'affaires (EUR)	Fréquences
$215 - 235$	4
$235 - 255$	6
$255 - 275$	13
$275 - 295$	22
$295 - 315$	15
$315 - 335$	6
$335 - 355$	5
$355 - 375$	4

Le **moment** d'ordre 4 de l'**échantillon** est donné par :

$$m_4 = \frac{1}{n} \sum_{i=1}^{h} f_i \,(x_i - \bar{x})^4$$

où $n = 75$, $\bar{x} = 290,60$ et x_i est le centre de chaque **intervalle** de classe. On peut résumer le tout dans le tableau ci-dessous :

x_i	$x_i - \bar{x}$	f_i	$f_i(x_i - \bar{x})^4$
225	$-65,60$	4	74 075 629, 16
245	$-45,60$	6	25 942 428, 06
265	$-25,60$	13	5 583 457, 48
285	$-5,60$	22	21 635, 89
305	14, 40	15	644 972, 54
325	34, 40	6	8 402 045, 34
345	54, 40	5	43 789 058, 05
365	74, 40	4	122 560 841, 32
			281 020 067, 84

Comme $S = 33,88$, le coefficient β_2 de Pearson est égal à :

$$\beta_2 = \frac{\frac{1}{75} \cdot 281\,020\,067,84}{(33,88)^4} = 2,84.$$

Puisque β_2 est inférieur à 3, on peut conclure que la distribution du chiffre d'affaires journalier dans soixante-quinze épiceries est platicurtique, c'est-à-dire plus aplatie que la distribution de la **loi normale**.

MOTS CLÉS

Mesure d'aplatissement *(Measure of kurtosis)*
Mesure de forme *(Measure of shape)*

RÉFÉRENCES

Voir **coefficient β_1 de Pearson**.

COLINÉARITÉ

Collinearity

Des variables sont dites mathématiquement colinéaires si l'une d'entre elles est exactement une combinaison linéaire des autres variables. Elles sont dites statistiquement colinéaires si l'une d'entre elles est approximativement une combinaison linéaire des autres variables. Dans le cas d'un modèle de régression où les variables explicatives sont fortement corrélées entre elles, on dit que l'on a de la colinéarité (ou de la multicolinéarité) entre les variables explicatives. Dans le premier cas, il est tout simplement impossible de définir les estimateurs des moindres carrés, alors que dans le second cas, ces estimateurs ont des variances qui peuvent être très importantes.

HISTORIQUE

Le terme de colinéarité est entré dans l'histoire des mathématiques au début du XX^{e} siècle, lors de la redécouverte du théorème de Pappus d'Alexandrie, mathématicien de l'Antiquité (III^{e} siècle). Soit trois points A, B, C sur une droite et 3 points A', B', C' sur une autre droite distincte. Si on joint les paires de points AB'.A'B, CA'.AC', et BC'.B'C leurs intersections seront toutes sur une droite, donc que les trois points d'intersection vont être colinéaires.

ASPECTS MATHÉMATIQUES

Matriciellement, pour la matrice des variables explicatives $\mathbf{X}$, la colinéarité signifie qu'une des colonnes de $\mathbf{X}$ est approximativement une combinaison linéaire des autres colonnes. Ceci implique que $\mathbf{X'X}$ est presque singulière. Par conséquent, l'**estimateur** obtenu par la méthode des moindres carrés $\widehat{\beta} = (\mathbf{X'X})^{-1}\mathbf{X'Y}$ s'obtient en inversant une **matrice** presque singulière, menant à l'instabilité de ses composantes. La **régression ridge** fut conçue dans le but de traiter les problèmes de colinéarité.

Une relation colinéaire impliquant plus de deux variables ne sera pas forcément décelée par les corrélations observées entre deux de ces variables. Une meilleure indication de la présence d'un problème de colinéarité est donnée par les facteurs d'inflation de la variance (en anglais, *variance inflation factors*) symbolisés par VIF. Le facteur d'inflation de la variance d'une variable explicative X_j est défini par :

$$VIF_j = \frac{1}{1 - R_j^2}$$

où R_j^2 est le **coefficient de détermination** du modèle

$$X_j = \beta_0 + \sum_{k \neq j} \beta_k X_k + \varepsilon$$

c'est-à-dire du modèle où les X_j sont régressés par rapport aux variables explicatives restantes X_k ($k \neq j$). Le coefficient VIF prend des valeurs entre 1 et ∞. Si les X_j sont mathématiquement colinéaires avec les autres variables, on aura $R_j^2 = 1$ et donc $VIF_j = \infty$. Si par contre les X_j sont réciproquement indépendants, on aura $R_j^2 = 0$ et $VIF_j = 1$. En pratique, nous considérons qu'il y a un réel problème de colinéarité lorsque VIF_j est supérieur à 100, ce qui correspond à un R_j^2 supérieur à $0,99$.

DOMAINES ET LIMITATIONS

Inverser une matrice singulière est similaire à inverser le nombre 0, ce n'est pas une opération valide. Sur le même principe, inverser une matrice presque singulière est similaire à inverser un nombre très petit. De ce fait, certains des éléments de la matrice inversée seront très grands. Par conséquent, lorsque des variables explicatives sont colinéaires, certains

éléments de la matrice $(\mathbf{X}'\mathbf{X})^{-1}$ de $\widehat{\beta}$ seront vraisemblablement élevées. C'est la raison pour laquelle la colinéarité mène à l'instabilité des estimateurs de la régression. Outre l'estimation de paramètres instables, la colinéarité pose également un problème relatif au calcul, il est difficile en effet de calculer avec précision l'inverse d'une matrice presque singulière.

EXEMPLE

Dans l'exemple suivant, les observations sont relatives à treize fournées de ciment. Chaque fournée est composée de quatres ingrédients cités dans le tableau. Le but de l'expérience est de déterminer comment les quantités X_1, X_2, X_3 et X_4 de ces 4 ingrédients affectent la quantité Y de chaleur émise par le durcissement du ciment.

- Y_i quantité de chaleur émise par le durcissement de la i-ème fournée (en joules) ;
- X_{i1} quantité de l'ingrédient 1 (aluminate de tricalcium) dans la i-ème fournée ;
- X_{i2} quantité de l'ingrédient 2 (silicate de tricalcium) dans la i-ème fournée ;
- X_{i3} quantité de l'ingrédient 3 (alumino-ferrite de tetracalcium) dans la i-ème fournée ;
- X_{i4} quantité de l'ingrédient 4 (silicate de dicalcium) dans la i-ème fournée.

Chaleur émise par le ciment

Fournée i	Ingrédient 1 X_1	Ingrédient 2 X_2	Ingrédient 3 X_3	Ingrédient 4 X_4	Chaleur Y
1	7	26	6	60	78,5
2	1	29	12	52	74,3
3	11	56	8	20	104,3
4	11	31	8	47	87,6
5	7	54	6	33	95,9
6	11	55	9	22	109,2
7	3	71	17	6	102,7
8	1	31	22	44	72,5
9	2	54	18	22	93,1
10	21	48	4	26	115,9
11	1	40	23	34	83,9
12	11	66	9	12	113,3
13	10	68	8	12	109,4

Source : Birkes & Dodge (1993)

Nous effectuons tout d'abord une **régression linéaire multiple**. Le modèle de la régression linéaire s'écrit :

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \varepsilon.$$

Nous obtenons les résultats suivants :

Variable	Coefficient $\hat{\beta}_i$	Écart type	t_c calculé
Constante	62,41	70,07	0,89
X_1	1,5511	0,7448	2,08
X_2	0,5102	0,7238	0,70
X_3	0,1019	0,7547	0,14
X_4	−0,1441	0,7091	−0,20

et constatons que seul le coefficient de la variable X_1 est significativement différent de zéro, puisque dans les autres cas, $t_c < t_{(\alpha/2,n-2)}$ (valeur prise dans la table de Student) pour un **seuil de signification** de $\alpha = 0,1$. De plus, les écarts types des coefficients estimés $\widehat{\beta}_2$, $\widehat{\beta}_3$ et $\widehat{\beta}_4$ sont plus grands que les coefficients eux-mêmes. Il est connu que lorsque des variables sont fortements corrélées entre elles, l'effet de l'une peut masquer l'effet de l'autre. De ce fait, les coefficients peuvent apparaître comme n'étant pas significativement différents de zéro.

Pour vérifier la présence de *multicolinéarité* par couple de variable, nous calculons la matrice des corrélations.

	X_1	X_2	X_3	X_4
X_1	1	0,229	−0,824	0,245
X_2	0,229	1	−0,139	−0,973
X_3	−0,824	−0,139	1	0,030
X_4	0,245	−0,973	0,030	1

Nous remarquons l'existence d'une corrélation négative très forte (−0,973) entre X_2 et X_4. En regardant les données, nous pouvons en voir la raison. En effet, à part les fournées 1 et 5, la quantité totale de composés de silicium $(X_2 + X_4)$ est presque constante et vaut environ 77 ; ainsi, X_4 est approximativement égal à $77 - X_2$. Cette situation nous empêche de distinguer les effets individuels de X_2 de ceux de X_4. Par exemple, nous constatons qu'aux quatre valeurs les plus élevées de X_4, qui sont 60, 52, 47 et 44, correspondent des valeurs de Y qui sont en-dessous de la chaleur moyenne de 95,4. Ainsi, il

semble à première vue que de grandes quantités de l'ingrédient 4 entraînent un faible dégagement de chaleur. Or on remarque également que les quatre valeurs les plus élevées de X_4 correspondent des quatre valeurs les plus faibles de X_2, entraînant la corrélation négative élevée entre X_2 et X_4. Cela peut signifier qu'en réalité l'ingrédient 4 n'a que peu d'effet sur la variable Y et que seules les faibles quantités d'ingrédient 2 expliquent une émission de chaleur plus faible. C'est ainsi que la colinéarité entre deux variables explicatives rend plus difficile la distinction entre l'effet de l'une ou l'autre des variables.

MOTS CLÉS

Régression ridge *(Ridge regression)*
Régression linéaire multiple *(Multiple linear model)*

RÉFÉRENCES

Birkes, D. and Dodge, Y. (1993). *Alternative methods of Regression.* John Wiley & Sons, New York.

Bock, R.D. (1975). *Multivariate Statistical Methods in Behavioral Research.* McGraw-Hill, New York.

COLLECTE DES DONNÉES

Data collection

Chaque fois qu'il est nécessaire de récolter des **données**, il faut considérer plusieurs aspects de la démarche. Il convient d'abord de définir pourquoi des données doivent être récoltées et à quelles fins elles vont être utiliser. Un deuxième aspect de la question nous amène à déterminer quelle sorte de données il faut récolter. Il est essentiel que les données récoltées soient en rapport avec le but recherché. Enfin, nous pouvons nous demander comment récolter des données. Il existe notamment trois types de collecte des données :

1. relevé de phénomènes sociaux ou naturels ;
2. **échantillonnage** et **recensement** ;
3. recherche expérimentale.

1. *Relevé*

Tout ce qui est accessible fait l'objet de relevés. Nous avons ainsi des relevés de naissances, de morts, de mariages, de températures journalières, de chutes de pluie mensuelles, de ventes d'automobiles, etc.

Il n'existe pas de méthodes statistiques générales permettant de tirer des conclusions valables à partir de ces données.

2. *Échantillonnage et recensement*

Les méthodes d'échantillonnage peuvent être divisées en deux groupes : les méthodes aléatoires et les méthodes non aléatoires.

Les *méthodes non aléatoires* consistent à construire par des procédures empiriques un **échantillon** qui ressemble autant que possible à la **population** de laquelle il est tiré. La méthode la plus utilisée est celle de l'**échantillonnage par quotas**.

Les *méthodes aléatoires* font appel à un procédé probabiliste pour former l'échantillon issu de la population, de telle sorte que l'on connaît à l'avance la **probabilité** qu'une unité donnée de la population appartienne à l'échantillon. Ainsi, il est possible de calculer l'**erreur** due à l'échantillonnage. Les principales méthodes aléatoires sont l'**échantillonnage aléatoire simple**, l'**échantillonnage stratifié**, l'**échantillonnage systématique** et l'**échantillonnage par grappes**.

Le recensement consiste à obtenir des informations en observant tous les objets d'une population. Il est clair que lorsque la taille de la population est élevée, ces enquêtes sont très coûteuses. C'est pourquoi le recensement est très peu utilisé.

3. *Recherche expérimentale*

Chaque **expérience** implique une collecte des données et fait appel à un **plan d'expérience**. Les plans d'expérience sont utilisés pour récolter des données dans tous les domaines de la recherche.

HISTORIQUE

Il est vraisemblable que le type de données le plus ancien remonte à l'Antiquité et notamment au **recensement** de la **population**.

Antille et Ujvari (1991) signalent l'existence d'un poste de statisticien officiel, en Chine, durant le règne de la dynastie Chow (111-211 av. J.-C.).

L'auteur latin Tacite nous apprend que l'empereur Auguste donna l'ordre de compter tous les soldats, tous les navires et toutes les richesses du royaume.

Nous trouvons la trace de recensements dans la Bible, dans laquelle saint Luc rapporte que « César Auguste » prit un décret prescrivant le recensement de toute la terre [...] et tous allaient se faire inscrire, chacun dans sa propre ville.

Déjà à cette époque, il existe une forme de **statistique** dont le nom même trahit l'origine administrative, puisqu'il est dérivé du latin *status*, l'État.

MOTS CLÉS

Échantillonnage *(Sampling)*
Plan d'expérience *(Design of experiments)*
Recensement *(Census)*
Sondage *(Survey)*

RÉFÉRENCES

Antille, G. et Ujvari, A. (1991). *Pratique de la statistique inférentielle.* PAN, Neuchâtel.

Boursin, J.-L. (1986). *Les Structures du hasard.* Éditions du Seuil, collection Points, série Sciences.

COMBINAISON

Combination

Les combinaisons sont une notion d'**analyse combinatoire**. Le nombre de combinaisons est le nombre de groupes de k objets qu'il est possible de former à partir d'un total de n objets.

Dans le calcul du nombre de combinaisons, l'ordre des objets à l'intérieur de chaque groupe ne nous intéresse pas (par opposition aux arrangements).

HISTORIQUE

Voir **analyse combinatoire**.

ASPECTS MATHÉMATIQUES

1. *Combinaisons sans répétition*

   Nous parlons de combinaisons sans répétition (ou sans remise) lorsque chaque objet tiré n'est pas remis à disposition pour le tirage suivant. Chaque objet ne peut donc intervenir qu'une seule fois dans chaque groupe.

   Le nombre de combinaisons sans répétition de k objets parmi n est égal à :

$$C_n^k = \left(\begin{array}{c} n \\ k \end{array} \right) = \frac{n!}{k! \cdot (n-k)!}$$

2. *Combinaisons avec répétitions*

   Nous parlons de combinaisons avec répétitions (ou avec remise) lorsque chaque objet tiré est remis à disposition pour le tirage suivant. Chaque objet peut donc intervenir r fois dans chaque groupe, $r = 0, 1, \ldots, k$.

   Le nombre de combinaisons avec répétitions de k objets parmi n est égal à :

$$K_n^k = \left(\begin{array}{c} n+k-1 \\ k \end{array} \right) = \frac{(n+k-1)!}{k! \cdot (n-1)!}$$

EXEMPLES

1. *Combinaisons sans répétition*

Considérons une assemblée de 8 personnes parmi lesquelles nous devons élire un comité de 3 personnes.

Le problème est donc de savoir combien de groupes de 3 personnes nous pouvons former à partir d'un ensemble de 8 personnes, chaque personne ne pouvant intervenir qu'une fois dans chaque groupe.

Nous allons donc calculer le nombre de combinaisons possibles de 3 objets parmi 8 :

$$\begin{aligned} C_n^k &= \frac{n!}{k! \cdot (n-k)!} = \frac{8!}{3! \cdot (8-3)!} \\ &= \frac{40320}{6 \cdot 120} = 56. \end{aligned}$$

Il est donc possible de former 56 comités différents de trois personnes parmi une assemblée de 8 personnes.

2. *Combinaisons avec répétitions*

Considérons une urne contenant 6 boules numérotées. Nous effectuons 4 tirages successifs en remettant à chaque fois la boule tirée dans l'urne.

Le problème est de savoir combien de combinaisons possibles peuvent découler d'un tel tirage.

Dans ce cas, nous cherchons le nombre de combinaisons avec répétitions (puisque chaque boule tirée est remise dans l'urne avant le tirage suivant). Nous aurons donc

$$\begin{aligned} K_n^k &= \frac{(n+k-1)!}{k! \cdot (n-1)!} = \frac{9!}{4! \cdot (6-1)!} \\ &= \frac{362\,880}{24 \cdot 120} = 126 \end{aligned}$$

combinaisons différentes.

MOTS CLÉS

Analyse combinatoire *(Combinatory analysis)*
Arrangement *(Arrangement)*
Loi binomiale *(Binomial distribution)*
Permutation *(Permutation)*

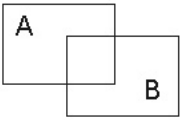
COMPATIBILITÉ

Compatibility

Deux événements sont dits compatibles si la réalisation du premier n'empêche pas la réalisation du second (ou en d'autres termes si l'intersection entre les deux événements est de probabilité non nulle) :

$$P(A \cap B) \neq 0.$$

Schématiquement, nous pouvons représenter deux événements A et B compatibles de la façon suivante :



Deux événements A et B sont alors dits incompatibles (ou mutuellement exclusifs) si la réalisation de A empêche la réalisation de B, ou inversement. Nous pouvons les représenter de la façon suivante :



ce qui signifie que la **probabilité** que les deux événements se réalisent en même temps est nulle :

$$A \cap B = \phi \quad \longrightarrow \quad P(A \cap B) = 0.$$

ASPECTS MATHÉMATIQUES

Si deux événements A et B sont compatibles, la **probabilité** qu'au moins un des deux événements se réalise est obtenue par la formule suivante :

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Par contre, si les deux événements A et B sont incompatibles, la probabilité que les événements A et B se réalisent simultanément est nulle,

$$P(A \cap B) = 0.$$

La probabilité qu'au moins un des deux événements se réalise peut donc être obtenue par simple addition des probabilités individuelles de A et B :

$$P(A \cup B) = P(A) + P(B)$$

EXEMPLE

Considérons l'**expérience aléatoire** qui consiste à tirer une carte d'un paquet de 52 cartes et les trois événements suivants :

$$\begin{aligned} A &= \text{tirer un cœur} \\ B &= \text{tirer une dame} \\ C &= \text{tirer un trèfle.} \end{aligned}$$

Les probabilités associées à chacun de ces événements sont :

$$P(A) = \frac{13}{52}$$
$$P(B) = \frac{4}{52}$$
$$P(C) = \frac{13}{52}.$$

Les événements A et B sont compatibles, car le fait de tirer un cœur n'empêche pas la réalisation de l'événement B qui est de tirer une dame. L'intersection entre A et B est égale à la dame de cœur. La probabilité de cet événement est donnée par :

$$P(A \cap B) = \frac{1}{52}.$$

La probabilité de la réunion des deux événements A et B (tirer un cœur ou une dame) est donc égale à :

$$\begin{aligned} P(A \cup B) &= P(A) + P(B) - P(A \cap B) \\ &= \frac{13}{52} + \frac{4}{52} - \frac{1}{52} \\ &= \frac{4}{13}. \end{aligned}$$

Par contre, les événements A et C sont incompatibles, car si la carte tirée est un cœur, elle ne peut pas être simultanément un trèfle ! L'intersection entre A et C est l'ensemble vide :

$$A \cap C = \phi.$$

La probabilité de la réunion des deux événements A et C (tirer un cœur ou un trèfle) est donc simplement donnée par la somme des probabilités de chacun des deux événements :

$$\begin{aligned} P(A \cup C) &= P(A) + P(C) \\ &= \frac{13}{52} + \frac{13}{52} \\ &= \frac{1}{2}. \end{aligned}$$

MOTS CLÉS

Événement *(Event)*
Indépendance *(Independence)*
Probabilité *(Probability)*

COMPLÉMENTAIRE

Complementary

Considérons l'**ensemble fondamental** Ω d'une **expérience aléatoire** quelconque.

Pour tout **événement** A élément de Ω, nous pouvons déterminer un nouvel événement B contenant tous les éléments de l'ensemble fondamental Ω qui ne sont pas inclus dans A.

Cet événement B est appelé complémentaire de A par rapport à Ω et est obtenu par négation de A.

ASPECTS MATHÉMATIQUES

Soit un **événement** A élément de l'**ensemble fondamental** Ω.

Le complémentaire de A par rapport à Ω est noté $\bar{A}$. Il est donné par la négation de A :

$$\begin{aligned} \bar{A} &= \Omega - A \\ &= \{w \in \Omega\,;\, w \notin A\}. \end{aligned}$$

EXEMPLE

Considérons l'**expérience aléatoire** qui consiste à jeter trois fois une pièce de monnaie.

L'**ensemble fondamental** de cette expérience est

$$\begin{aligned} \Omega = \{&PPP, PPF, PFP, PFF, \\ &FPP, FPF, FFP, FFF\} \end{aligned}$$

Soit l'**événement**

$$\begin{aligned} A &= \text{Face } (F) \text{ est apparue deux fois} \\ &= \{PFF, FPF, FFP\} \end{aligned}$$

Le complémentaire de A par rapport à Ω est égal à

$$\begin{aligned} \bar{A} &= \{PPP, PPF, PFP, FPP, FFF\} \\ &= \text{Face } (F) \text{ n'est pas apparue} \\ &\quad \text{exactement deux fois} \end{aligned}$$

MOTS CLÉS

Ensemble fondamental *(Sample space)*
Événement *(Event)*
Expérience aléatoire *(Random experiment)*

CONTINGENCE, TABLE DE

Contingency table

Voir **table de contingence**.

CONTRASTE

Contrast

Un contraste est une fonction linéaire des observations d'une **expérience factorielle**, telle que la somme des coefficients est nulle.

HISTORIQUE

Selon Henry Scheffé (1953), **John Wilder Tukey** (1949 et 1951) fut le premier à proposer une méthode pour l'**estimation** simultanée de tous les contrastes.

ASPECTS MATHÉMATIQUES

Soient $T_1, T_2, \ldots, T_k$ les sommes de $n_1, n_2, \ldots, n_k$ observations. La fonction linéaire

$$c_j = c_{1j} \cdot T_1 + c_{2j} \cdot T_2 + \cdots + c_{kj} \cdot T_k$$

est un contraste si et seulement si

$$\sum_{i=1}^{k} n_i \cdot c_{ij} = 0.$$

Si chaque $n_i = n$, c'est-à-dire si chaque T_i est la somme d'un même nombre d'observations, la condition se réduit à :

$$\sum_{i=1}^{k} c_{ij} = 0.$$

DOMAINES ET LIMITATIONS

Dans la plupart des expériences impliquant plusieurs traitements, il est intéressant pour l'expérimentateur de faire des comparaisons entre les différents traitements. Pour effectuer ce genre de comparaisons, le statisticien a recours à ce qu'il appelle les contrastes.

EXEMPLE

Lorsqu'on fait une **analyse de variance pour un facteur** à trois niveaux (traitements), les quelques contrastes que l'on peut former sont :

$$\begin{aligned} c_1 &= T_1 - T_2 \\ c_2 &= T_1 - T_3 \\ c_3 &= T_2 - T_3 \\ c_4 &= T_1 - 2 \cdot T_2 + T_3. \end{aligned}$$

MOTS CLÉS

Analyse de variance *(Analysis of variance)*
Expérience *(Experiment)*
Expérience factorielle *(Factorial experiment)*

RÉFÉRENCES

Ostle, B. (1954). *Statistics in Research ; Basic concepts and techniques for research workers.* Iowa State College Press, Ames.

Scheffé, H. (1953). A method for judging all contrasts in the analysis of variance. *Biometrika*, 40, pp. 87-104.

Tukey, J.W. (1949). Comparing individual means in the analysis of variance. *Biometrics*, 5, pp. 99-114.

Tukey, J.W. (1951). *Quick and dirty methods in statistics. Part II - Simple analyses for standard designs.* Quality Control Conference Papers 1951. pp. 189-197. American Society for Quality Control, New York.

CONVERGENCE

Convergence

En statistique, le terme convergence implique la notion de **probabilité**. Nous utilisons volontiers le

terme de convergence stochastique pour la distinguer de la convergence au sens classique du terme.

ASPECTS MATHÉMATIQUES

Nous définissons ici les types de convergence stochastique les plus importants. Soit $\{X_n\}_{n\in\mathbb{N}}$ une suite de variables aléatoires.

1. $\{X_n\}_{n\in\mathbb{N}}$ converge *en loi* vers une variable aléatoire X si :

$$\lim_{n\to\infty} F_{X_n}(z) = F_X(z) \quad \text{pour } z \text{ réel}$$

où F_{X_n} et F_X désignent les fonctions de répartition de X_n et de X.

Cette convergence n'est rien d'autre que la convergence ponctuelle (bien connue en mathématique) de la suite des fonctions de répartition des X_n.

2. $\{X_n\}_{n\in\mathbb{N}}$ converge *en probabilité* vers une variable aléatoire X si :

$$\lim_{n\to\infty} P(|X_n - X| > \varepsilon) = 0 \quad \text{pour tout } \varepsilon > 0.$$

3. $\{X_n\}_{n\in\mathbb{N}}$ converge *presque sûre* vers une variable aléatoire X si :

$$P\left(\left\{w | \lim_{n\to\infty} X_n(w) = X(w)\right\}\right) = 1.$$

4. Supposons que tous les X_n aient une **espérance** finie. La suite $\{X_n\}_{n\in\mathbb{N}}$ converge *en moyenne quadratique* vers X si :

$$\lim_{n\to\infty} E\left[(X_n - X)^2\right] = 0.$$

Citons quelques résultats intéressants :

- la convergence presque sûrement et la convergence en moyenne quadratique impliquent toutes deux la convergence en probabilité ;
- la convergence en probabilité implique la convergence en loi, qui est donc la plus « forte ».

EXEMPLE

Exemple 1 : soient X_i des variables aléatoires indépendantes suivantes une **loi uniforme** sur $[0,1]$. À partir des X_i nous définissons la suivante suite de variables aléatoires :

$$Z_n = n \cdot \min_{i=1,\ldots,n} X_i.$$

La suite $\{Z_n\}_{n\in\mathbb{N}}$ converge en loi vers une **loi exponentielle** Z de paramètre 1 car :

$$\begin{aligned} F_{Z_n}(t) &= P(Z_n > t) \\ &= P\left(\min_{i=1,\ldots,n} X_i > \frac{t}{n}\right) \\ &= P\left(X_1 > \frac{t}{n} \text{ et } X_2 > \frac{t}{n} \text{ et } \ldots X_n > \frac{t}{n}\right) \\ &\overset{ind.}{=} \prod_{i=1}^{n} P(X_i > \frac{t}{n}) = (1 - \frac{t}{n})^n. \end{aligned}$$

Comme $\lim_{n\to\infty}\left(1 - \frac{t}{n}\right)^n = \exp(-t)$, il suit :

$$\lim_{n\to\infty} F_{Z_n} = \lim_{n\to\infty} P(Z_n \leq t) = 1 - \exp(-t) = F_Z.$$

Exemple 2 : soit S_n le nombre de succès obtenu lors de n épreuves de Bernoulli avec probabilité de succès p. Le **théorème de Bernoulli** stipule la convergence en probabilité de $\frac{S_n}{n}$ vers la variable « aléatoire » qui prend la valeur p avec probabilité 1.

MOTS CLÉS

Loi des grands nombres (*Law of large numbers*)
Probabilité *(Probability)*
Processus stochastique *(Stochastic process)*
Théorème central limite *(Central limit theorem)*
Théorème de convergence *(Convergence theorem)*
Théorème de Bernoulli *(Bernoulli's theorem)*
Théorème de Moivre-Laplace *(Moivre-Laplace theorem)*
Variable aléatoire *(Random variable)*

RÉFÉRENCES

Le Cam, L.M. and Yang, C.L. (1990). *Asymptotics in statistics : some basic concepts.* Springer-Verlag.

Staudte, R.G. and Sheater, S.J. (1990). *Robust estimation and testing.* John Wiley & Sons, New York.

CONVERGENCE, THÉORÈME DE

Convergence theorem

Voir **théorème de convergence**.

CORRÉLATION PARTIELLE

Partial correlation

La corrélation partielle entre deux variables se définit lorsque deux variables présentent une forte corrélation sans qu'on puisse interpréter la corrélation comme une relation directe entre les deux variables. Une *mesure de la corrélation partielle* entre les variables X et Y étant donnée, une troisième variable Z est une mesure de la relation directe entre X et Y, qui ne tient plus compte des conséquences des relations linéaires de ces deux variables avec Z. Le calcul de la corrélation partielle entre X et Y étant donné, Z sert à identifier en quelle mesure la relation linéaire entre X et Y peut être due à la corrélation des deux variables avec Z.

HISTORIQUE

L'interprétation de l'analyse du coefficient de corrélation entre deux variables est devenue une question importante dans la statistique lors de l'utilisation de plus en plus large des méthodes de corrélation au début du vingtième siècle. **Karl Pearson** était conscient qu'une grande corrélation entre deux variables peut être due à leur corrélation respective avec une troisième variable. Cependant, ce phénomène n'a pas été reconnu jusqu'en 1926 date à laquelle **George Udny Yule** le démontra à travers l'exemple de l'obtention des coefficients de corrélations insensés entre des séries temporelles.

ASPECTS MATHÉMATIQUES

À partir d'un échantillon de taille n, (x_1, y_1, z_1), (x_2, y_2, z_2), ..., (x_n, y_n, z_n) provenant de la distribution inconnue de trois variables aléatoires X, Y et Z, le *coefficient de corrélation partielle*, que l'on note $r_{xy.z}$, est défini comme étant le **coefficient de corrélation** calculé entre les $\widehat{x}_i$ et les $\widehat{y}_i$ avec

$$\begin{aligned} \widehat{x}_i &= \widehat{\beta}_{0x} + \widehat{\beta}_{1x} z_i \\ \widehat{y}_i &= \widehat{\beta}_{0y} + \widehat{\beta}_{1y} z_i, \end{aligned}$$

où les $\widehat{\beta}_{0x}$ et $\widehat{\beta}_{1x}$ sont les estimateurs des **moindres carrés** obtenus en effectuant une régression des x_i sur les z_i, et les $\widehat{\beta}_{0y}$ et $\widehat{\beta}_{1y}$ les estimateurs des moindres carrés obtenus en effectuant une régression des y_i sur les z_i. On a ainsi par définition :

$$r_{xy.z} = \frac{\sum (\widehat{x}_i - \overline{x})(\widehat{y}_i - \overline{y})}{\sqrt{\sum (\widehat{x}_i - \overline{x})^2} \sqrt{\sum (\widehat{y}_i - \overline{y})^2}}.$$

On peut montrer le résultat suivant :

$$r_{xy.z} = \frac{r_{xy} - r_{xz} \cdot r_{yz}}{\sqrt{1 - r_{xz}^2} \cdot \sqrt{1 - r_{yz}^2}},$$

où r_{xy}, r_{xz} et r_{yz} sont les coefficients de corrélation respectivement entre les x_i et les y_i, entre les x_i et les z_i et entre les y_i et les z_i. On remarque ainsi que lorsque les x_i et les y_i ne sont pas corrélés avec les z_i, c'est-à-dire lorsque $r_{xz} = r_{yz} = 0$, on a :

$$r_{xy.z} = r_{xy},$$

c'est-à-dire que le coefficient de corrélation est égal au coefficient de corrélation partielle, ou autrement dit, que la valeur du coefficient de corrélation entre les x_i et les y_i n'est pas du tout due à la présence des z_i. Par contre, si ces corrélations avec les z_i sont importantes, on peut avoir $r_{xy.z} \cong 0$, ce qui indique que la corrélation observée entre les x_i et les y_i est uniquement due aux corrélations entre ces variables et la variable z_i. On dit alors qu'il n'y a pas de relation directe entre les x_i et les y_i (mais une relation indirecte, étant donné les z_i).

Plus généralement, on peut calculer la corrélation partielle entre les x_i et les y_i étant donné des z_{i1} et z_{i2} relatifs à deux variables Z_1 et Z_2 de la manière suivante :

$$r_{xy.z_1z_2} = \frac{r_{xy\cdot z_1} - r_{xz_2.z_1} \cdot r_{yz_2.z_1}}{\sqrt{1 - r^2_{xz_2.z_1}} \cdot \sqrt{1 - r^2_{yz_2.z_1}}}.$$

DOMAINES ET LIMITATIONS

On doit insister sur le fait qu'une corrélation ne mesure que la liaison linéaire entre deux variables, sans se préoccuper de modèles fonctionnels ou de la capacité de prédiction ou de prévision.

EXEMPLE

On a observé, par exemple, une forte corrélation positive entre le nombre d'infarctus enregistrés durant une certaine période et le nombre de glaces vendues durant cette même période. Il ne faut pourtant pas en conclure qu'une forte consommation de glaces provoque des infarctus ou inversement qu'un infarctus provoque une envie particulière de glaces. En effet, dans cet exemple, le nombre élevé d'infarctus provient non pas du nombre de glaces vendues, mais de la chaleur. On a donc une forte corrélation entre le nombre d'infarctus et la température. De même, le nombre élevé de glaces vendues est également expliqué par la chaleur, ce qui implique aussi une forte corrélation entre le nombre de glaces vendues et la température. La forte corrélation observée entre le nombre d'infarctus et le nombre de glaces vendues est donc une conséquence de deux autres fortes corrélations. Une troisième variable est en fait cachée derrière la relation apparente des deux premières.

MOT CLÉ

Coefficient de corrélation *(Correlation coefficient)*

RÉFÉRENCE

Yule, G.U. (1926) Why do we sometimes get nonsense-correlations between time-series? A study sampling and the nature of time-series. *Journal of the Royal Statistical Society*, (2) 89, pp. 1-64.

CORRÉLATION SÉRIELLE

Serial correlation

On appelle corrélation serielle ou **autocorrélation**, la dépendance entre les observations d'une **série chronologique**.

HISTORIQUE

Voir **série chronologique**.

ASPECTS MATHÉMATIQUES

Soient $x_1, \ldots, x_n$ n observations d'une variable aléatoire X qui dépendent du temps (on parle alors d'une série chronologique). Pour mesurer la dépendance entre les observations et le temps, on utilise le coefficient :

$$d = \frac{\sum_{t=2}^{n} (x_t - x_{t-1})^2}{\sum_{t=1}^{n} x_t^2}$$

qui traduit la relation entre deux observations consécutives. Ce coefficient prend des valeurs comprises entres 0 et 4. S'il est proche de 0, alors la différence entre deux observations succesives est minime, on parle alors d'une autocorrélation positive. S'il est proche de 4, on a alors une autocorrélation négative : une observation avec une valeur élevée a tendance a être suivie par une observation avec une valeur basse et réciproquement. Dans le cas où les observations sont indépendantes du temps, le coefficient d a une valeur proche de 2.

On peut aussi mesurer la corrélation sérielle des observations $x_1, \ldots, x_n$ en calculant le **coefficient de corrélation** entre la série des x_t et celle des x_t « décalée de k » :

$$r_k = \frac{\sum_{t=k+1}^{n} (x_t - \overline{x})(x_{t-k} - \overline{x})}{\sum_{t=1}^{n} (x_t - \overline{x})^2}$$

et en l'analysant de la même façon que pour le coefficient de corrélation de deux séries différentes.

DOMAINES ET LIMITATIONS

Modèles auto-régressifs

Lorsque une corrélation sérielle est observée entre les résidus d'un modèle de regression linéaire simple, des modèles provenant du domaine des séries chronologiques peuvent être utilisés avec succès. Il s'agit par exemple de considérer des modèles du type :

$$Y_t = \beta_0 + \beta_1 X_t + \varepsilon_t$$

où les termes d'erreur ε_t ne sont pas indépendants, mais sont décrits par une relation du type

$$\varepsilon_t = \rho_s \varepsilon_{t-1} + \delta_t$$

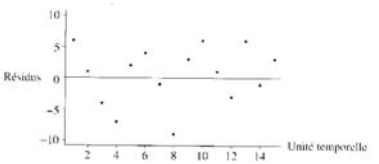
où ρ_s peut être estimé par la corrélation sérielle des ε_t et où les δ_t sont des erreurs indépendantes.

EXEMPLE

Graphique des résidus en fonction du temps

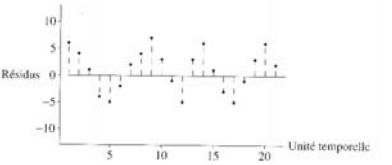
Ce type de graphique est surtout utilisé dans le cadre de l'analyse des séries chronologiques, c'est-à-dire lorsque le temps est une variable explicative du modèle.

Résidus par rapport au temps.



Une représentation des résidus par rapport au temps fait souvent apparaître la présence de corrélation sérielle, positive ou négative, comme c'est le cas dans la figure ci-dessous. Une corrélation sérielle positive ou négative implique que les erreurs ne sont pas indépendantes.

Résidus présentant une corrélation sérielle.



MOTS CLÉS

Autocorrélation *(Autocorrelation)*
Série chronologique *(Time series)*
Test d'hypothèse *(Hypothesis test)*

RÉFÉRENCES

Bartlett, M.S. (1946). On the theoretical specification of sampling properties of autocorrelated time series. *Journal of the Royal Statistical Society.* Series B, 8, pp. 27-41.

Box, G.E.P. and Jenkins, G.M. (1970). *Time series analysis: Forecasting and control, Holden-Day Series in Time Series Analysis*, San Francisco.

Durbin, J. and Watson, G.S. (1951). Testing for Serial Correlation in Least Squares Regression. II. *Biometrika*, 38, pp. 159-177.

COUPLE DE VARIABLES ALÉATOIRES

Pair of random variables

Un couple de variables dont les valeurs sont déterminées en fonction du résultat d'une **expérience aléatoire** en tant que paire de nombres ordonnés est appelé couple de variables aléatoires. On parle de :

- couple de variables aléatoires discrètes si l'ensemble des valeurs prises par chacune des variables aléatoires est un ensemble fini ou infini dénombrable ;
- couple de variables aléatoires continues si l'ensemble des valeurs prises par chacune des

variables aléatoires est un ensemble infini non dénombrable.

Dans le cas d'un couple de variables aléatoires, l'ensemble de tous les résultats possibles de l'expérience forme l'**ensemble fondamental** qui se situe dans un espace à deux dimensions.

ASPECTS MATHÉMATIQUES

Un couple de variables aléatoires est généralement désigné par les dernières lettres de l'alphabet (majuscules) : (X, Y). C'est une fonction de deux variables à valeurs réelles dans $\mathbb{R}^2$.

Si A est un sous-ensemble de l'**ensemble fondamental**, on écrit :

$$\begin{aligned} P(A) &= \textbf{probabilité que } (X,Y) \in A \\ &= P[(X,Y) \in A]. \end{aligned}$$

Le cas d'un couple de variables aléatoires peut se généraliser au cas de n variables aléatoires : s'il faut n nombres ordonnés pour décrire la réalisation d'une **expérience aléatoire**, une telle réalisation se représente par n variables aléatoires $X_1, X_2, \ldots, X_n$. L'ensemble fondamental se situe alors dans un espace à n dimensions. Si A est un sous-ensemble de l'ensemble fondamental, on écrit :

$$P(A) = P[(X_1, X_2, \ldots, X_n) \in A].$$

DOMAINES ET LIMITATONS

Certaines situations aléatoires obligent à considérer non pas une seule mais plusieurs entités numériques simultanément. Soit par exemple un système constitué de deux éléments disposés en série et de durées de vie aléatoires X et Y ; l'étude de la durée de vie de l'ensemble met en jeu des événements concernant à la fois X et Y. Soit encore une pièce cylindrique usinée par une machine, de diamètre X et de longueur Y aléatoires ; les tolérances imposées porteront en général simultanément sur les deux entités dont il faudra connaître la structure de dépendance probabiliste.

MOTS CLÉS

Covariance *(Covariance)*
Fonction de densité conjointe *(Joint density function)*
Fonction de densité marginale *(Marginal density function)*
Variable aléatoire *(Random variable)*

RÉFÉRENCE

Hogg, R.V. and Craig, A.T. (1965). *Introduction to mathematical statistics.* 2th ed. The Macmillan Company, New York.

COURBE DE FRÉQUENCES

Frequency curve

Si l'on fait subir à un **polygone de fréquences** une sorte de polissage, on obtient une courbe lisse que l'on appelle courbe de fréquences.

On peut effectuer ce genre de polissage si le nombre d'observations dans la **distribution de fréquences** devient infiniment grand et si l'amplitude des classes devient infiniment petite.

La courbe de fréquences correspond donc à la forme limite d'un polygone de fréquences.

HISTORIQUE

Durant la dernière partie du XIXe siècle, la tendance était de considérer toutes les distributions comme des lois normales. Les histogrammes qui présentaient plusieurs modes étaient ajustés pour former un ensemble de courbes de fréquences normales et ceux qui présentaient une asymétrie étaient analysés en transformant les données de telle façon que l'**histogramme** résultant de cette **transformation** puisse être comparé à une courbe normale.

C'est **Karl Pearson** (1895) qui eut l'idée de construire un système de courbes de fréquences théoriques ayant des formes très variées afin d'obtenir de meilleures approximations des histogrammes.

MOTS CLÉS

Distribution de fréquences *(Frequency distribution)*
Loi de probabilité continue *(Continuous probability distribution)*

Loi normale *(Normal distribution)*
Polygone de fréquences *(Frequency polygon)*

RÉFÉRENCE

Pearson, K. (1895). 1948. Contributions to the Mathematical Theory of Evolution. II: Skew Variation in Homogeneous Material. pp. 41-112 in Karl Pearson, *Karl Pearson's Early Statistical Papers.* Cambridge University Press, Cambridge. First published in *Philosophical Transactions of the Royal Society of London.* Series A, Vol. 186, pp. 343-414.

COVARIANCE

Covariance

La covariance de deux variables aléatoires X et Y est une mesure de la variation simultanée entre ces deux variables.

HISTORIQUE

Voir **analyse de variance** et **coefficient de corrélation**.

ASPECTS MATHÉMATIQUES

Soient X et Y deux variables aléatoires définies sur le même **ensemble fondamental** Ω. On définit la covariance de X et Y, que l'on note par $Cov\,(X,Y)$, par

$$Cov\,(X,Y) = E\left[(X - E\,[X])\,(Y - E\,[Y])\right].$$

où $E\,[.]$ représente l'**espérance mathématique**.
Le développement du membre de droite donne :

$$\begin{aligned} Cov\,(X,Y) &= E(XY - E\,[X]\,Y - XE\,[Y] \\ &\quad + E\,[X]\,E\,[Y]) \\ &= E\,[XY] - E\,[X]\,E\,[Y] \\ &\quad - E\,[X]\,E\,[Y] + E\,[X]\,E\,[Y] \\ &= E\,[XY] - E\,[X]\,E\,[Y]\,. \end{aligned}$$

Propriétés de la covariance

Soient X, Y, Z des variables aléatoires définies sur le même ensemble fondamental Ω et a, b, c, d des constantes, on a :

1. $Cov\,(X,Y) = Cov\,(Y,X)$;
2. $Cov\,(X,c) = 0$;
3. $Cov\,(aX + bY, Z) = aCov\,(X,Z) + bCov\,(Y,Z)$;
   $Cov\,(X, cY + dZ) = cCov\,(X,Y) + dCov\,(X,Z)$;
4. $Cov\,(aX + b, cY + d) = acCov(X,Y)$.

Conséquences de la définition

1. Si X et Y sont des variables aléatoires indépendantes,

   $$Cov\,(X,Y) = 0.$$

   En effet, $E\,[XY] = E\,[X]\,E\,[Y]$; par conséquent

   $$Cov\,(X,Y) = E\,[XY] - E\,[X]\,E\,[Y] = 0.$$

   La réciproque n'est en général pas vraie : $Cov\,(X,Y) = 0$ n'implique pas forcément que X et Y sont indépendantes.

2. $Cov\,(X,X) = Var\,(X)$
   où $Var\,(X)$ représente la **variance** de X.

   En effet :

   $$\begin{aligned} Cov\,(X,X) &= E\,[XX] - E\,[X]\,E\,[X] \\ &= E\left[X^2\right] - (E\,[X])^2 \\ &= Var\,(X)\,. \end{aligned}$$

DOMAINES ET LIMITATIONS

Soient deux variables aléatoires X et Y. On peut s'intéresser à leur somme $X + Y$. On a alors :

$$\begin{aligned} E\,[X+Y] &= E\,[X] + E\,[Y] \quad \text{et} \\ Var\,(X+Y) &= Var\,(X) + Var\,(Y) + 2Cov\,(X,Y)\,. \end{aligned}$$

On démontre ces résultats pour des variables discrètes. Si $P_{ji} = P\,(X = x_i, Y = y_j)$, on a :

$$\begin{aligned}
E[X+Y] &= \sum_i \sum_j (x_i + y_j) P_{ji} \\
&= \sum_i \sum_j x_i P_{ji} + \sum_i \sum_j y_j P_{ji} \\
&= \sum_i x_i \left(\sum_j P_{ji} \right) + \sum_j y_j \left(\sum_i P_{ji} \right) \\
&= \sum_i x_i P_i + \sum_j y_j P_j \\
&= E[X] + E[Y].
\end{aligned}$$

De plus, on a :

$$\begin{aligned}
Var(X+Y) &= E\left[(X+Y)^2\right] - (E[X+Y])^2 \\
&= E\left[X^2\right] + 2E[XY] + E\left[Y^2\right] \\
&\quad - (E[X+Y])^2 \\
&= E\left[X^2\right] + 2E[XY] + E\left[Y^2\right] \\
&\quad - (E[X] + E[Y])^2 \\
&= E\left[X^2\right] + 2E[XY] + E\left[Y^2\right] \\
&\quad - (E[X])^2 - 2E[X]E[Y] - (E[Y])^2 \\
&= Var(X) + Var(Y) \\
&\quad + 2(E[XY] - E[X]E[Y]) \\
&= Var(X) + Var(Y) \\
&\quad + 2Cov(X,Y).
\end{aligned}$$

On peut généraliser ces résultats pour n variables aléatoires $X_1, X_2, \ldots, X_n$, ayant chacune une **espérance mathématique** égale à $E(X_i)$ et une **variance** égale à $Var(X_i)$. Nous avons alors :

$$\begin{aligned}
&E[X_1 + X_2 + \cdots + X_n] = \\
&\quad E[X_1] + E[X_2] + \cdots + E[X_n] \\
&= \sum_{i=1}^{n} E[X_i].
\end{aligned}$$

$$\begin{aligned}
&Var(X_1 + X_2 + \cdots + X_n) = \\
&\quad Var(X_1) + Var(X_2) + \cdots + Var(X_n) \\
&\quad + 2[Cov(X_1, X_2) + \cdots + Cov(X_1, X_n) \\
&\quad + Cov(X_2, X_3) + \cdots + Cov(X_2, X_n) \\
&\quad + \cdots + Cov(X_{n-1}, X_n)] \\
&= \sum_{i=1}^{n} Var(X_i) + 2\sum_{i=1}^{n} \sum_{j>i} Cov(X_i, X_j).
\end{aligned}$$

EXEMPLE

Soient deux tests psychologiques effectués successivement et pour lesquels un sujet quelconque reçoit une note X de 0 à 3 pour le premier test et une note Y de 0 à 2 pour le second test.

Sachant que les probabilités que X soit égal à 0, 1, 2, 3 sont respectivement $0,16$, $0,3$, $0,41$ et $0,13$, et que les probabilités que Y soit égal à 0, 1, 2 sont respectivement $0,55$, $0,32$ et $0,13$, on a :

$$\begin{aligned}
E[X] &= 0 \cdot 0,16 + 1 \cdot 0,3 + 2 \cdot 0,41 + 3 \cdot 0,13 \\
&= 1,51 \\
E[Y] &= 0 \cdot 0,55 + 1 \cdot 0,32 + 2 \cdot 0,13 \\
&= 0,58 \\
E[XY] &= 0 \cdot 0 \cdot 0,16 \cdot 0,55 + 0 \cdot 1 \cdot 0,16 \cdot 0,32 \\
&\quad + \ldots + 3 \cdot 2 \cdot 0,13 \cdot 0,13 \\
&= 0,88.
\end{aligned}$$

On peut calculer la covariance de X et Y :

$$\begin{aligned}
Cov(X,Y) &= E[XY] - E[X]E[Y] \\
&= 0,88 - (1,51 \cdot 0,58) \\
&= 0,00428.
\end{aligned}$$

MOTS CLÉS

Coefficient de corrélation *(Correlation coefficient)*
Espérance mathématique *(Expected value)*
Variable aléatoire *(Random variable)*
Variance d'une variable aléatoire *(Variance of a random variable)*

COVARIATION

Covariation

Il est souvent intéressant, en particulier en économie, de comparer deux séries chronologiques.

La mesure des liens de **dépendance** entre deux variables nous renvoie à la notion de corrélation. Cependant, dans le cas présent, puisque les séries chronologiques sont toutes liées par une troisième variable, le temps, l'interprétation du **coefficient**

de corrélation ne pourrait faire apparaître qu'une liaison artificielle.

En effet, si on considère deux séries chronologiques x_t et y_t représentant des phénomènes complètement indépendants et dont les fonctions sont linéaires par rapport au temps :

$$\begin{aligned} x_t &= a \cdot t + b \\ y_t &= c \cdot t + d \end{aligned}$$

où a, b, c et d sont des constantes, il est toujours possible d'éliminer le facteur temps t entre les deux équations et d'aboutir à une relation fonctionnelle de type $y = e \cdot x + f$. Cette relation montre une dépendance linéaire entre les deux séries chronologiques, alors que la réalité est par **hypothèse** exactement l'inverse.

En conclusion, la mesure de la corrélation entre l'évolution dans le temps de deux phénomènes n'implique pas l'existence d'un lien entre eux. On utilise alors le terme de covariation et on mesure cette dépendance à l'aide d'un coefficient de covariation. On distingue :

- le coefficient de covariation linéaire ;
- le coefficient de covariation rapporté à la tendance.

HISTORIQUE

Voir **coefficient de corrélation** et **série chronologique**.

ASPECTS MATHÉMATIQUES

Pour comparer deux séries chronologiques y_t et x_t, il convient d'abord de les représenter sur un même graphique.

Cependant, la comparaison visuelle directe est généralement malaisée. On procède alors aux changements de **variables** suivants :

$$Y_t = \frac{y_t - \bar{y}}{S_y} \text{ et } X_t = \frac{x_t - \bar{x}}{S_x}$$

qui ne sont autres que les variables centrées réduites où S_y et S_x sont les écarts types des séries chronologiques respectives.

On distingue les coefficients de covariation suivants :

- *le coefficient de covariation linéaire*

  Son expression est, dans sa forme, identique à celle du **coefficient de corrélation** r, mais ici les calculs n'ont plus la même portée puisqu'il s'agit de détecter l'existence d'un éventuel lien entre des variations elles-mêmes liées au temps et d'en mesurer l'ordre de grandeur

  $$C = \frac{\sum_{t=1}^{n} (x_t - \bar{x}) \cdot (y_t - \bar{y})}{\sqrt{\sum_{t=1}^{n} (x_t - \bar{x})^2 \cdot \sum_{t=1}^{n} (y_t - \bar{y})^2}}.$$

  Il est compris entre -1 et $+1$. S'il est proche de ± 1, il existe une **relation** linéaire entre les évolutions dans le temps des deux variables. Il est à noter que :

  $$C = \frac{\sum_{t=1}^{n} X_t \cdot Y_t}{n},$$

  n étant le nombre de dates d'observations, Y_t, X_t les séries centrées réduites obtenues par changement de variable.

- *le coefficient de covariation rapporté à la tendance*

  On élimine l'influence des moyennes, en calculant :

  $$K = \frac{\sum_{t=1}^{n} (x_t - T_{x_t}) \cdot (y_t - T_{y_t})}{\sqrt{\sum_{t=1}^{n} (x_t - T_{x_t})^2 \cdot \sum_{t=1}^{n} (y_t - T_{y_t})^2}}.$$

  On a simplement remplacé les moyennes $\bar{x}$ et $\bar{y}$ par les valeurs de la **tendance séculaire** T_{x_t}, et T_{y_t} de chacune des séries chronologiques.

Le coefficient de covariation rapporté à la tendance K varie également de -1 à $+1$ et, plus il est proche de ± 1, plus la covariation entre les séries chronologiques est forte.

DOMAINES ET LIMITATIONS

Il est souvent intéressant de comparer deux séries chronologiques. Les exemples en économie sont nombreux : comparaison de l'évolution du prix d'un produit avec celle des quantités de ce même produit, évolution du revenu national et évolution des transactions immobilières, etc. Il est donc important de repérer s'il existe un lien de dépendance entre deux phénomènes évoluant dans le temps : c'est l'objet de la mesure de covariation.

La comparaison visuelle entre deux séries chronologiques est une opération capitale, mais elle est très souvent difficile car :

- les données peuvent provenir de domaines très variés et présenter de ce fait des ordres de grandeurs très différents ; on préfère alors étudier les écarts à la **moyenne** ;
- les pics et les creux des deux séries chronologiques ont des amplitudes trop différentes ; il vaut mieux homogénéiser les dispersions, en ramenant les variations à l'**écart type** de la série chronologique.

On facilite donc la comparaison visuelle en considérant les variables centrées réduites obtenues par les changements de variables suivants :

$$Y_t = \frac{y_t - \bar{y}}{S_y} \text{ et } X_t = \frac{x_t - \bar{x}}{S_x}.$$

Par ailleurs, comme dans le cas du **coefficient de corrélation**, il faut se méfier des relations non linéaire s pouvant exister entre les deux variables, mais laissant la **valeur** de C proche de zéro.

La prudence lors de l'interprétation est donc de rigueur.

On donne la préférence d'utilisation au coefficient de covariation rapporté à la tendance lorsque la relation entre les séries chronologiques n'est pas linéaire.

EXEMPLE

Proposons-nous de faire l'étude de la covariation entre deux séries chronologiques.

La **variable** x_t représente la production annuelle d'un produit agricole ; la variable y_t, son prix unitaire moyen annuel en euros constants.

t	x_t	y_t
1	320	5,3
2	660	3,2
3	300	2,2
4	190	3,4
5	320	2,7
6	240	3,5
7	360	2,0
8	170	2,5

$$\sum_{t=1}^{8} x_t = 2560 \quad \sum_{t=1}^{8} y_t = 24,8$$

d'où $\bar{x} = 320$ et $\bar{y} = 3,1$.

$x_t - \bar{x}$	$(x_t - \bar{x})^2$	$y_t - \bar{y}$	$(y_t - \bar{y})^2$
0	0	2,2	4,84
340	115 600	0,1	0,01
−20	400	−0,9	0,81
−130	16 900	0,3	0,09
0	0	−0,4	0,16
−80	6 400	0,4	0,16
40	1 600	−1,1	1,21
−150	22 500	0,6	0,36

$$\sum_{t=1}^{8} (x_t - \bar{x})^2 = 163\,400 \quad \sum_{t=1}^{8} (y_t - \bar{y})^2 = 7,64$$

d'où $\sigma_x = 142,9$ et $\sigma_y = 0,98$.

On trouve alors les valeurs centrées réduites X_t et Y_t.

X_t	Y_t	$X_t \cdot Y_t$	$X_t \cdot Y_{t-1}$
0,00	2,25	0,00	5,36
2,38	0,10	0,24	−0,01
−0,14	−0,92	0,13	0,84
−0,91	0,31	−0,28	0,00
0,00	−0,41	0,00	0,23
−0,56	0,41	−0,23	0,11
0,28	−1,13	−0,32	1,18
−1,05	−0,61	0,64	

On calcule alors le coefficient de covariation linéaire :

$$C = \frac{\sum_{t=1}^{8} X_t \cdot Y_t}{8} = \frac{0,18}{8} = 0,0225.$$

Si l'on compare les observations X_t avec Y_{t-1}, c'est-à-dire la production de l'année avec les prix de l'année précédente, on trouve :

$$C = \frac{\sum_{t=2}^{8} X_t \cdot Y_{t-1}}{8} = \frac{7,71}{8} = 0,964.$$

Le coefficient de covariation linéaire avec déphasage d'une année est donc fort. Il y a une covariation (positive) forte avec déphasage d'un an entre les deux variables.

MOTS CLÉS

Coefficient de corrélation *(Correlation coefficient)*
Écart type *(Standard deviation)*
Moyenne mobile *(Moving average)*
Série chronologique *(Time series)*
Tendance séculaire *(Secular trend)*

RÉFÉRENCES

Kendall, M.G. (1973). *Time Series.* Charles Griffin & Company Ltd.

Py, B. (1987). *Statistique descriptive.* Economica, Paris.

COX DAVID R.

Cox David R.

David Roxbee Cox est né en 1924 à Birmingham en Angleterre. Il a étudié les mathématiques à l'Université de Cambridge et obtenu son doctorat en mathématiques appliquées à l'Université de Leeds en 1949. Il a travaillé par la suite aussi bien dans la recherche industrielle que dans les milieux académiques et de l'édition scientifique. De 1966 à 1988, il a été professeur de statistiques à l'Imperial College of Sciences and Technology de Londres, puis de 1988 à 1994, il a enseigné au Nuffield College, à Oxford.



David Cox est un éminent statisticien. Il a été consacré chevalier en 1982 par la reine d'Angleterre en reconnaissance de ses contributions à la science statistique et a été nommé docteur honoris causa par des nombreuses universités en Angleterre et d'ailleurs. Il a également été honoré comme membre illustre par plusieurs académies de sciences : de 1981 à 1983, président de la Royal Statistical Society, président de la Société Bernoulli de 1972 à 1983 et président de l'Institut international de statistique de 1995 à 1997.

Par la variété des sujets qu'il a abordés et développés, le professeur D. Cox a profondément marqué sa profession. Il a été nommé docteur honoris causa de l'Université de Neuchâtel en 1992.

Sir David Cox est l'auteur ou le co-auteur de plus de 250 articles et 16 livres et entre 1966 et 1991 il a été éditeur de la revue Biometrika.

Quelques ouvrages et articles principaux de Sir David Cox :

1964 (and Box, G.E.P.) An analysis of transformations (with discussion) *Journal of the Royal Statistical Society.* Series B, 26, pp. 211-243.

1970 *The Analysis of Binary Data.* Methuen, London.

1973 (and Hinkley, D.V.) *Theoretical Statistics.* Chapman & Hall, London.

1974 (and Atkinson, A.C.) Planning experiments for discriminating between models. *Journal of the Royal Statistical Society.* Series B, 36, pp. 321-348.

1978 (and Hinkley, D.V.) *Problems and Solutions in Theoretical Statistics.* Chapman & Hall, London.

1981 (and Snell, E.J.) *Applied Statistics : principles and examples.* Chapman & Hall.

1981 Theory and general principles in statistics. *Journal of the Royal Statistical Society.* Series A, 144, pp. 289-297.

2000 (and Reid, N.) *Theory of Design Experiments.* Chapman & Hall, London.

COX GERTRUDE

Cox Gertrude

Gertrude Mary Cox naquit en 1900, près de Dayton (Iowa) aux États-Unis. Son ambition étant d'aider ses prochains, elle s'inscrivit à un cours de sciences sociales, qu'elle suivit pendant deux ans. Ensuite elle travailla deux ans dans un orphelinat pour jeunes garçons à Montana. C'est pour devenir directrice de l'orphelinat qu'elle décida d'entrer au Iowa State College qu'elle quitta en 1929. Afin de payer ses études, Gertrude Cox travailla avec **George Waddel Snedecor**, son professeur d'analyse, dans son laboratoire de statistique ; c'est peut-être à cause de cette expérience qu'elle s'intéressa aux statistiques. À la fin de ses études, elle entreprit un doctorat en statistique psychologique. En 1933, tandis qu'elle n'avait pas terminé son doctorat, George Snedecor, alors directeur du Statistical Laboratory de l'État d'Iowa, la convainquit de travailler avec lui, ce qu'elle accepta, tout en restant dans le domaine de la psychologie puisqu'elle travailla sur l'évaluation de tests d'aptitude et l'analyse de données psychologiques.



Le premier novembre 1940, elle devint directrice du département de statistiques expérimentales de l'État de Caroline du Nord.

En 1945, le General Education Board lui alloua des subventions afin de créer un institut de statistiques à l'Université de Caroline du Nord avec un département de statistiques mathématiques à Chapel Hill.

Elle fut un membre fondateur du journal *International Biometric Society* en 1947 dont elle fut éditrice de 1947 à 1955 et présidente de 1968 à 1969. Elle présida également l'American Statistical Association (ASA) en 1956.

Elle mourut d'une leucémie en 1978.

Ouvrage principal de Gertrude Cox :

1957 (and Cochran, W.) *Experimental Designs.* 2nd edition. John Wiley & Sons, New York

C_p *DE MALLOWS*

C_p *criterion*

Le critère de choix C_p est une méthode de choix du modèle de régression en fonction du nombre de **variables explicatives** entrant dans le cas d'une **régression linéaire**. Un faible C_p correspond à un meilleur modèle dans le sens qu'il présente la *somme des carrés des résidus* (SC_{res}) la plus faible et la moins pénalisée par le nombre de paramètres entrant dans le modèle.

HISTORIQUE

Introduit par Colin L. Mallows en 1964, le critère de choix C_p est utilisé dès lors comme critère d'évaluation de l'adéquation d'un modèle de regression.

ASPECTS MATHÉMATIQUES

Soit $Y_i = \beta_0 + \sum_{j=1}^{p-1} X_{ji}\beta_j + \varepsilon_i, \quad i = 1, \ldots, n$ un modèle de **régression linéaire multiple**. Notons $EQM(.)$ l'**erreur quadratique moyenne**. L'idée du critère introduit dans cette section est de choisir un modèle pour lequel la somme des erreurs quadratiques moyennes est minimale :

$$\begin{aligned}\sum EQM\left(\widehat{y}_i\right) &= \sum\left(E\left(\left(\widehat{y}_i - y_i\right)^2\right) - \sigma^2\left(1 - 2h_{ii}\right)\right)\\ &= E\left(\sum\left(\widehat{y}_i - y_i\right)^2\right) - \sigma^2\sum\left(1 - 2h_{ii}\right)\\ &= E\left(SC_{res}\right) - \sigma^2\left(n - 2p\right)\end{aligned}$$

avec

- n le nombre d'observations,
- p, le nombre de paramètres estimés,
- h_{ii} désignant les éléments diagonaux de la **matrice** H et
- $\widehat{y}_i$ l'estimateur de y_i.

Rappellons la propriété des h_{ii} :

$$\sum h_{ii} = p.$$

Définissons le coefficient

$$\begin{aligned}\Gamma_p &= \frac{\sum EQM\left(\widehat{y}_i\right)}{\sigma^2}\\ &= \frac{E\left(SC_{res}\right) - \sigma^2\left(n - 2p\right)}{\sigma^2}\\ &= \frac{E\left(SC_{res}\right)}{\sigma^2} - n + 2p.\end{aligned}$$

Si le modèle est correct, nous devons avoir :

$$E\left(SC_{res}\right) = \left(n - p\right)\sigma^2$$

ce qui implique

$$\Gamma_p = p.$$

En pratique, nous estimons Γ_p par

$$C_p = \frac{SC_{res}}{\widehat{\sigma}^2} - n + 2p$$

où $\widehat{\sigma}^2$ est un estimateur de σ^2. Nous estimons ici σ^2 par le s^2 du modèle qui fait intervenir *toutes* les k variables explicatives à disposition. Pour ce modèle complet, nous obtenons effectivement $C_p = p$, ce n'est pas un résultat intéressant. Mais pour tous les autres modèles où nous ne faisons intervenir qu'un *sous-ensemble* des variables explicatives à disposition, le coefficient C_p peut prendre d'autres valeurs que p. Nous choisissons alors parmi les modèles qui ne font intervenir qu'un sous-ensemble des variables explicatives celui pour lequel la valeur de C_p est la plus proche de p.

Si nous avons k variables explicatives à disposition, nous pouvons ainsi définir le coefficient C_p pour un modèle faisant intervenir un sous-ensemble $X_1, \ldots, X_{p-1}$, $p \leq k$, des k variables explicatives à disposition, de la manière suivante :

$$C_p = \frac{(n-k)\, SC_{res}\,(X_1, \ldots, X_{p-1})}{SC_{res}\,(X_1, \ldots, X_k)} - n + 2p$$

où $SC_{res}\,(X_1, \ldots, X_{p-1})$ désigne la somme des carrés des résidus relative au modèle avec $p-1$ variables explicatives, alors que $SC_{res}\,(X_1, \ldots, X_k)$ désigne la somme des carrés des résidus relative au modèle avec k variables explicatives. Entre deux modèles faisant intervenir $p-1$ variables explicatives, nous choisissons donc, d'après ce critère, celui pour lequel la valeur du coefficient C_p est la plus proche de p.

DOMAINES ET LIMITATIONS

Le critère Cp est utilisé dans la procédure de sélection des variables. C'est un critère qui en plus du critère R^2 peut fournir une réponse sur l'adéquation du modèle choisi. L'idée sous-jacente à de telles procédures est la suivante : plutôt que de chercher à expliquer une variable par toutes les variables explicatives disponibles, il est parfois possible de déterminer un modèle avec un sous-ensemble de ces variables, dont le pouvoir explicatif est presque aussi satisfaisant que celui du modèle comprenant toutes les variables explicatives. Ceci est d'autant plus avantageux que la présence de colinéarité dans les variables explicatives a tendance à diminuer la précision des estimateurs, d'où la nécessité de supprimer certaines variables superflues.

EXEMPLES

Considérons les donnée relatives aux incendies dans les quartiers de Chicago. Notons Y la variable correspondant au logarithme du nombre d'incendies pour 1 000 ménages par quartier i de Chicago en 1975, ainsi que X_1, X_2 et X_3 les variables correspondant respectivement à la proportion d'habitations construites avant 1940, au nombre de vols et au revenu médian par quartier i.

Comme cet ensemble comprend trois variables explicatives, on a $2^3 = 8$ modèles possibles. Nous répartissons les huit équations possibles et quatre ensembles :

1. L'ensemble A contient l'unique équation sans variable explicative :
$$Y = \beta_0 + \varepsilon.$$
2. L'ensemble B contient les trois équations avec une seule variable explicative :
$$\begin{aligned} Y &= \beta_0 + \beta_1 X_1 + \varepsilon \\ Y &= \beta_0 + \beta_2 X_2 + \varepsilon \\ Y &= \beta_0 + \beta_3 X_3 + \varepsilon. \end{aligned}$$
3. L'ensemble C contient les trois équations avec deux variables explicatives :
$$\begin{aligned} Y &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon \\ Y &= \beta_0 + \beta_1 X_1 + \beta_3 X_3 + \varepsilon \\ Y &= \beta_0 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon. \end{aligned}$$
4. L'ensemble D contient l'unique équation avec trois variables explicatives :
$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon.$$

Quartier i	X_1 x_{i1}	X_2 x_{i2}	X_3 x_{i3}	Y y_i
1	0,604	29	11,744	1,825
2	0,765	44	9,323	2,251
3	0,735	36	9,948	2,351
4	0,669	37	10,656	2,041
5	0,814	53	9,730	2,152
6	0,526	68	8,231	3,529
7	0,426	75	21,480	2,398
8	0,785	18	11,104	1,932
9	0,901	31	10,694	1,988
10	0,898	25	9,631	2,715
11	0,827	34	7,995	3,371
12	0,402	14	13,722	0,788
13	0,279	11	16,250	1,740
14	0,077	11	13,686	0,693
15	0,638	22	12,405	0,916
16	0,512	17	12,198	1,099
17	0,851	27	11,600	1,686
18	0,444	9	12,765	0,788
19	0,842	29	11,084	1,974
20	0,898	30	10,510	2,715
21	0,727	40	9,784	2,803
22	0,729	32	7,342	2,912
23	0,631	41	6,565	3,589
24	0,830	147	7,459	3,681
25	0,783	22	8,014	2,918
26	0,790	29	8,177	3,148
27	0,480	46	8,212	2,501

Quartier i	X_1 x_{i1}	X_2 x_{i2}	X_3 x_{i3}	Y y_i
28	0,715	23	11,230	1,723
29	0,731	4	8,330	3,082
30	0,650	31	5,583	3,073
31	0,754	39	8,564	2,197
32	0,208	15	12,102	1,281
33	0,618	32	11,876	1,609
34	0,781	27	9,742	3,353
35	0,686	32	7,520	2,856
36	0,734	34	7,388	2,425
37	0,020	17	13,842	1,224
38	0,570	46	11,040	2,477
39	0,559	42	10,332	2,351
40	0,675	43	10,908	2,370
41	0,580	34	11,156	2,380
42	0,152	19	13,323	1,569
43	0,408	25	12,960	2,342
44	0,578	28	11,260	2,747
45	0,114	3	10,080	1,946
46	0,492	23	11,428	1,960
47	0,466	27	13,731	1,589

Source : Andrews et Herzberg (1985)

Nous symbolisons les modéles de la façon suivante : **1** pour X_1, **2** pour X_2, **3** pour X_3, **12** pour X_1 et X_2, **13** pour X_1 et X_3, **23** pour X_2 et X_3 et **123** pour le modèle complet. Le tableau suivant représente les résultats obtenus pour le C_p de Mallows pour chacun des modèles :

Modèle	C_p
1	32,356
2	29,937
3	18,354
12	18,936
13	14,817
23	3,681
123	4,000

Dans un modèle du groupe B (comprenant une seule variable explicative), le nombre de paramètres à estimer est $p = 2$ et le C_p d'aucun des trois modèles ne s'approche de 2. Pour un modèle du groupe C (à trois variables explicatives), p vaut 3 et le C_p du modèle **23** s'en approche. Pour le modèle complet, nous trouvons $C_p = 4$, ce qui est aussi le nombre de paramètres à estimer, mais n'est pas un résultat intéressant comme expliqué auparavant. Ainsi, le choix du modèle :

$$Y = \beta_0 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$$

semble être le plus judicieux.

MOTS CLÉS

Analyse de régression *(Regression analysis)*
Colinéarité *(Collinearity)*
Coefficient de détermination *(Coefficient of determination)*
Erreur quadratique moyenne *(Mean squared error)*
Matrice H *(Hat matrix)*

RÉFÉRENCES

Andrews D.F., Hertzberg, A.M. (1985). *Data: A Collection of Problems from many fields for Students and Reserch Worker.* Springer-Verlag, New York.

Mallows C.L. (1964). *Choosing Variables in a linear Regression : A Graphical Aid.* Presented at the Central Regional Meeting of the Institute of Mathematical Statistics, Manhattan, Kansas, May 7-9.

Mallows C.L. (1973). Some comments on *Cp. Technometrics*, 15, pp. 661-675.

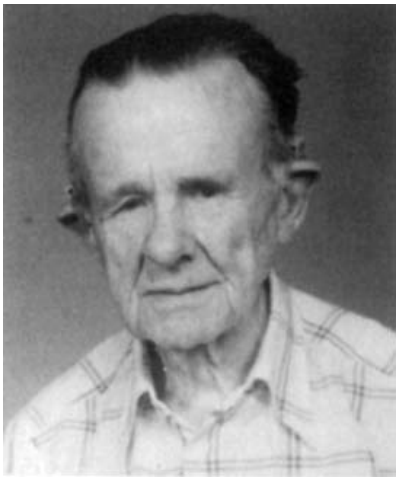
CRAMÉR HARALD

Cramér Harald

Harald Cramér (1893-1985) entra à l'Université de Stockholm en 1912 pour y étudier la chimie et les mathématiques ; il devint l'étudiant de Mittag Leffler et Marcel Riesz. En 1919, Cramér fut nommé assistant professeur à l'Université de Stockholm. À côté, il travailla en tant qu'actuaire pour la Svenska Life Assurance Company, ce qui lui permit d'étudier les probabilités et les statistiques.

Son œuvre classique en mathématiques actuaires s'intitule *Collective Risk Theory.* En 1929, il fut nommé pour créer une nouvelle chaire à Stockholm, devenant le premier professeur suédois en

mathématiques actuaires et mathématiques statistiques. C'est à la fin de la Seconde Guerre mondiale qu'il écrivit son ouvrage majeur : *Mathematical Methods of Statistics* qui fut publié pour la première fois en 1945 et réédité en 1999.



Quelques ouvrages et articles principaux de Harald Cramér :

1946 *Mathematical Methods of Statistics.* Princeton University Press, Princeton NJ.

1946 *Collective Risk Theory : A Survey of the Theory from the Point of View of the Theory of Stochastic Processes.* Skandia Jubilee Volume, Stockholm.



D

DANIELS HENRY E.

Daniels Henry E.



Henry Ellis Daniels (1912-2000) obtient son diplôme à l'Université d'Edinburgh en 1933 et continue ses études à l'Université de Cambridge. Après avoir reçu le titre de docteur de l'Université d'Edinburgh, il retourne à Cambridge comme maître assistant en mathématiques en 1947. En 1957, Henry Daniels devient le premier professeur de mathématique statistique à l'Université de Birmingham, poste qu'il occupe jusqu'en 1978, moment où il prend sa retraite. Les domaines de recherche de prédilection de Henry Daniels ont été la statistique inférentielle, les approximations *saddlepoint* en statistique, la modélisation épidémiologique et la théorie statistique appliquée à la technologie textile. Lorsqu'il était à Birmingham, il a instauré les rencontres annuelles des statisticiens à Gregynog Hall en Powys, réunions où la musique de chambre s'intercale aux présentations. Par la suite, cette rencontre annuelle est devenue une des plus appréciées des statisticiens anglais.

Entre 1974 et 1975, il a été président de la Royal Statistical Society qui lui avait décerné en 1957 la Guy Medal en argent. En 1984, il reçoit la Guy Medal en or. En 1980, il est élu membre de la Royal Society et, en 1985, membre honorifique de l'International Statistical Society.

Quelques ouvrages et articles principaux de Henry E. Daniels :

1954 Saddlepoint Approximations in Statistics. *Annals of Mathematical Statistics*, 25, pp. 631-650.

1955 Discussion of a paper by G.E.P. Box and S.L. Anderson. *Journal of the Royal Statistical Society.* Series B, 17, pp. 27-28.

1958 Discussion of paper by D.R. Cox. *Journal of the Royal Statistical Society.* Series B, 20, pp. 236-238.

1987 Tail Probability Approximations. *International Statistical Review.* 55, pp. 137-48.

DÉCILE

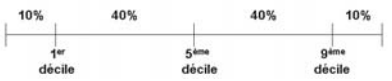
Decile

Les déciles sont des mesures de position calculées sur un ensemble de données. Ce sont les valeurs qui partagent une distribution en dix parties égales (chaque partie contient le même nombre d'observations).

Le x-ème décile indique la valeur dont au plus x % des observations lui sont inférieures et $(100 - x)$ % des observations lui sont supérieures. Par exemple, le décile 8 est la valeur qui sépare les 80 % inférieures des 20 % supérieures.

Le décile 5 représente la **médiane**. Cette notion fait partie de la famille des **quantiles**.

Nous aurons donc neuf déciles pour une distribution donnée :



ASPECTS MATHÉMATIQUES

Le processus de calcul est similaire à celui de la **médiane** ou des quartiles.

Lorsque nous possédons toutes les observations brutes, le processus de calcul des déciles est le suivant :

1. Organiser les n observations sous forme d'une **distribution de fréquences**.
2. Les déciles correspondent aux observations pour lesquelles la **fréquence** relative cumulée dépasse respectivement 10 %, 20 %, 30 %, ..., 80 %, 90 %.

   Certains auteurs proposent la formule suivante qui permet de déterminer avec précision la **valeur** des différents déciles :

   Calcul du j-ème décile :

   — soit i la partie entière de $j \cdot \frac{n+1}{10}$ et k la partie fractionnaire de $j \cdot \frac{n+1}{10}$;

   — soient x_i et x_{i+1} les valeurs des observations classées respectivement en i-ème et $(i+1)$-ème position (lorsque les observations sont classées par ordre croissant). Le j-ème décile est égal à :

$$D_j = x_i + k \cdot (x_{i+1} - x_i)\,.$$

Lorsque nous possédons des observations groupées en classes, les déciles se déterminent de la manière suivante :

1. Déterminer la classe dans laquelle se trouve le décile voulu :
   - 1^{er} *décile* : première classe pour laquelle la fréquence relative cumulée dépasse 10 %.
   - 2^{e} *décile* : première classe pour laquelle la fréquence relative cumulée dépasse 20 %.

     ...
   - 9^{e} *décile* : première classe pour laquelle la fréquence relative cumulée dépasse 90 %.
2. Calculer la valeur du décile en fonction de l'**hypothèse** selon laquelle les observations sont distribuées uniformément dans chaque classe le j-ème d'ecile D_j est :

$$D_j = L_j + \left[\frac{n \cdot \frac{j}{10} - \sum f_{inf}}{f_j}\right] \cdot c_j$$

où

L_j	=	borne inférieure de la classe du décile D_j,
n	=	nombre total d'observations,
$\sum f_{inf}$	=	somme des fréquences inférieures à la classe du décile D_j,
f_j	=	fréquence de la classe du décile D_j,
c_j	=	dimension de l'**intervalle** de la classe du décile D_j.

DOMAINES ET LIMITATIONS

Le calcul des déciles n'a de sens que pour une **variable quantitative** pouvant prendre des valeurs sur un **intervalle** déterminé.

Le nombre d'observations doit être relativement élevé puisque le calcul des déciles consiste à diviser l'ensemble des observations en dix parties.

L'utilisation des déciles est relativement fréquente dans la pratique, par exemple pour l'interprétation de la distribution des revenus d'une ville ou d'un canton.

EXEMPLE

Considérons un exemple de calcul des déciles sur la **distribution de fréquences** d'une **variable** continue où les observations sont groupées en classes.

Le **tableau de fréquences** suivant représente les profits (en milliers d'euros) de 500 épiceries :

Profit (en milliers d'euros)	Fréquence	Fréquence cumulée	Fréquence relative cumulée
$100-150$	40	40	$0,08$
$150-200$	50	90	$0,18$
$200-250$	60	150	$0,30$
$250-300$	70	220	$0,44$
$300-350$	100	320	$0,62$
$350-400$	80	400	$0,80$
$400-450$	60	460	$0,92$
$450-500$	40	500	$1,00$
Total	500		

La classe comprenant le 1^{er} décile est la classe $150-200$ (celle dont la **fréquence** relative cumulée dépasse 10 %).

En considérant que les observations sont distribuées de manière uniforme dans chaque classe, nous obtenons pour le premier décile la **valeur** suivante :

$$\begin{aligned} 1^{\text{er}} \text{ décile} &= 150 + \left[\frac{\left(500 \cdot \frac{1}{10}\right) - 40}{50}\right] \cdot 50 \\ &= 160. \end{aligned}$$

La classe comprenant le 2^{e} décile est la classe $200-250$. La valeur du 2^{e} décile est égale à

$$\begin{aligned} 2^{\text{e}} \text{ décile} &= 200 + \left[\frac{\left(500 \cdot \frac{2}{10}\right) - 90}{60}\right] \cdot 50 \\ &= 208,33. \end{aligned}$$

Nous pouvons calculer les autres déciles de la même manière, ce qui nous donne :

Décile	Classe	Valeur
1	$150-200$	$160,00$
2	$200-250$	$208,33$
3	$200-250$	$250,00$
4	$250-300$	$285,71$
5	$300-350$	$315,00$
6	$300-350$	$340,00$
7	$350-400$	$368,75$
8	$350-400$	$400,00$
9	$400-450$	$441,67$

Nous pouvons donc conclure par exemple que 10 % des 500 épiceries ont un profit compris entre 100 et 160 milliers de francs, 50 % ont un profit inférieur à 315 milliers de francs, 10 % ont un profit compris entre $441,67$ et 500 milliers de francs, etc.

MOTS CLÉS

Centile *(Percentile)*
Médiane *(Median)*
Mesure de position *(Measure of location)*
Quantile *(Quantile)*
Quartile *(Quartile)*

DEGRÉ DE LIBERTÉ

Degree of freedom

Le nombre de degrés de liberté est un **paramètre** de la **loi du chi-carré** et d'autres lois de probabilité apparentées à la loi du chi-carré telles que la **loi de Student** et la **loi de Fisher**.

On peut voir le nombre de degrés de liberté comme le nombre de termes linéairement indépendants impliqués dans le calcul d'une somme de carrés basée sur n observations indépendantes.

HISTORIQUE

Le terme degré de liberté a été introduit par **Ronald Aylmer Fisher** en 1925.

ASPECTS MATHÉMATIQUES

Soit $Y_1, Y_2, \ldots, Y_n$ un échantillon aléatoire de taille n tiré d'une population de moyenne $\bar{Y}$ inconnue. La somme des déviations de n observations par rapport à leur **moyenne arithmétique** est toujours égale à zéro :

$$\sum_{i=1}^{n} \left(Y_i - \bar{Y}\right) = 0.$$

Cette exigence constitue une contrainte sur chacune des déviations $Y_i - \bar{Y}$ utilisées dans le calcul de la **variance** :

$$S^2 = \frac{\sum\limits_{i=1}^{n} \left(Y_i - \bar{Y}\right)^2}{n-1}.$$

Cette contrainte implique que $n-1$ déviations déterminent complètement la n-ème. On dit alors que les

n déviations (et donc également leur somme de carrés et la variance de l'**échantillon** S^2) ont $n-1$ degrés de liberté.

MOTS CLÉS

Loi de Fisher *(Fisher distribution)*
Loi de Student *(Student distribution)*
Loi du chi-carré *(Chi-square distribution)*
Paramètre *(Parameter)*
Variance *(Variance)*

RÉFÉRENCE

Fisher, R.A. (1925). Applications of "Student's" Distribution. *Metron*, 5, pp. 90-104.

DEMING W. EDWARDS

Deming W. Edwards

W. Edwards Deming naquit en 1900, à Sioux City, en Iowa. Il étudia dans les universités du Wyoming (où il fut licencié ès sciences en 1921) et du Colorado (où il fit un postgrade en mathématiques et physique qu'il termina en 1924) avant d'aller à l'Université de Yale où il obtintt son doctorat de physique en 1928. Après son doctorat, il travailla durant dix ans dans un laboratoire du ministère de l'Agriculture. En 1939, Deming rejoignit le *Bureau of Census* à Washington (un organisme comparable à l'INSEE). Il y utilisa ses connaissances théoriques pour mettre au point les premières enquêtes par échantillonnage, avec des techniques qui feront école dans le monde entier. Il quitta ce poste en 1946 pour devenir consultant en études statistiques et professeur de statistiques à l'Université de New York. Il mourut à Washington en 1993.

Jusqu'en 1980 où elles connurent un véritable succès, les théories de W.E. Deming avaient été bannies des entreprises américaines. En 1947, W.E. Deming est envoyé à Tokyo comme conseiller de l'état-major des forces alliées pour appliquer ses techniques d'échantillonnage ; dès 1950, l'industrie japonaise adopte les théories de W.E. Deming sur le management et dix ans plus tard les produits japonais commencent à déferler en Amérique. Le public américain ne s'y trompe pas : ils sont meilleurs et moins chers.



Quelques ouvrages et articles principaux de W. Edwards Deming :

1938 *Statistical Adjustment of Data.* John Wiley & Sons, New York.

1950 *Some Theory of Sampling.* John Wiley & Sons, New York.

1960 *Sample Design in Business Research.* John Wiley & Sons, New York.

DÉMOGRAPHIE

Demography

La démographie est l'étude de la **population** humaine. Elle consiste à analyser des phénomènes tels

que les naissances, les décès, les mouvements migratoires, les mariages, les divorces, le taux de fécondité, le taux de mortalité, la pyramide des âges, etc.

Ces phénomènes sont traités aussi bien dans leur contexte biologique que socio-économique.

Les méthodes qu'utilise la démographie sont de nature empirique et statistique. Elles font appel aux mathématiques avancées ainsi qu'à toute branche des sciences sociales.

HISTORIQUE

On peut situer au milieu du XVII^{e} siècle les premières études démographiques. À cette époque, John Graunt, un scientifique anglais, entreprit d'analyser les seules données dont il disposait : la liste des décès au voisinage de Londres, classés selon leur cause. En 1662, il fit paraître une étude dans laquelle il tentait d'évaluer la taille moyenne des familles, l'importance des mouvements migratoires et d'autres éléments constituant la structure de la population.

En collaboration avec Sir William Petty, J. Graunt fit des propositions visant à obtenir une étude plus sérieuse de la population et la fondation d'un centre où seraient regroupées les données statistiques.

Au cours du XVIII^{e} siècle, on réalisa des analyses du même genre en les améliorant progressivement, mais il fallut attendre le XIX^{e} pour que plusieurs pays européens ainsi que les États-Unis entreprennent des recensements nationaux et mettent en place l'enregistrement des naissances, des mariages et des décès.

Ces relevés permirent d'observer que toutes les régions n'offraient pas à leurs habitants les mêmes chances de survie. À partir de ces conclusions, on peut prouver la nécessité d'amélioration des conditions de travail et d'hygiène.

Des prévisions devenaient possibles et vers le tournant du siècle, on vit apparaître les premiers journaux et revues consacrés à la démographie. Ce furent par exemple *Demography* aux États-Unis, *Population* en France et *Population Studies* en Grande-Bretagne.

Le milieu du XX^{e} siècle marqua une nouvelle étape lorsque les études des démographes portèrent sur la population mondiale considérée comme un tout. Il devenait en effet urgent de s'intéresser aux problèmes démographiques du tiers-monde.

MOTS CLÉS

Population *(Population)*
Recensement *(Census)*

RÉFÉRENCES

Cox, P.R. (1976). *Demography*, 5th ed. Cambridge University Press, Cambridge.

Hauser, P.M. and Duncan, O.D. (1959). *The Study of Population: An Inventory and Appraisal.* Chicago University Press, Chicago.

DENDROGRAMME

Dendrogram

Un dendrogramme est une **représentation graphique** des différentes agrégations effectuées au cours d'une **classification automatique**. Il est formé de nœuds correspondant aux groupes et de branches désignant les associations faites à chaque pas. La structure du dendrogramme est déterminée par l'ordre dans lequel ont été exécutées les agrégations.

En plaçant une échelle le long du dendrogramme, il est possible de représenter les distances auxquelles les agrégations ont lieu.

HISTORIQUE

De façon générale, le dendrogramme (du grec *dendron*, arbre) est un diagramme en arbre illustrant les relations de nature quelconque existant entre les objets d'un ensemble.

Les premiers exemples de dendrogramme sont les arbres phylogéniques utilisés par les spécialistes de systématique. Le terme de dendrogramme semble avoir été utilisé pour la première fois dans l'ouvrage de Mayr *et al.* (1953).

ASPECTS MATHÉMATIQUES

Lors d'une **classification automatique** sur un ensemble quelconque d'objets, nous effectuons le nombre d'agrégations souhaité à partir du **tableau de distances** et selon la méthode choisie (**méthode du chaînage simple**, **méthode du chaînage complet**, etc.).

Chaque extrémité du dendrogramme représente les différentes classes obtenues à la fin de la classification automatique.

Nous convenons que chaque nœud du dendrogramme représente un groupe et est placé par rapport à une échelle horizontale selon la distance à laquelle il a été formé.

Les objets sont placés au niveau 0 de l'échelle. Pour éviter que les branches reliant les nœuds rassemblés ne s'entremêlent, nous procéderons de façon systématique.

- Les deux premiers éléments regroupés sont placés l'un en dessous de l'autre, au niveau 0 de l'échelle. Nous dessinons alors à côté de chaque élément un segment horizontal partant de 0 et dont l'extrémité droite sera déterminée par la **distance** d'agrégation. La classe résultante sera représentée par un palier vertical reliant les deux extrémités droites. Le milieu du palier servira de point de départ pour un nouveau segment horizontal lors d'une agrégation postérieure. Les branches sont construites uniquement à l'aide de segments horizontaux suivis de segments verticaux au niveau d'agrégation.

- Deux cas peuvent se présenter à l'étape suivante :

  1. nous regroupons à nouveau deux classes formées d'un seul élément, ces dernières sont placées au niveau 0 à la suite et le nœud résultant sera situé de la même manière qu'auparavant ;

  2. nous rattachons un objet à une classe formée précédemment. L'élément est rajouté au niveau 0 à la suite de la classe (au besoin, l'intercaler entre des classes déjà existantes). Nous plaçons le nouveau nœud à la bonne distance (horizontale) et à mi-hauteur entre les deux sous-ensembles agrégés.

Nous procédons ainsi jusqu'à obtenir la configuration souhaitée. S'il reste, à la fin des opérations, des classes formées d'un seul élément, il suffira de les rajouter au dendrogramme au niveau 0. Chaque classe, après la **classification**, correspond à une partie disjointe du dendrogramme.

DOMAINES ET LIMITATIONS

Le dendrogramme a la propriété de décrire un parcours ordonné de l'ensemble des opérations effectuées lors d'une **classification automatique**. Il illustre de façon très précise la **classification** en tant que telle.

La méthode de construction d'un dendrogramme ne peut pas toujours être appliquée tel quel. En effet, comme nous le voyons dans l'exemple proposé, il se peut que les distances d'agrégation à deux ou plusieurs pas successifs soient les mêmes ; il faut alors changer la procédure pour éviter que les branches du dendrogramme ne s'entremêlent.

Nous avons ici déployé le dendrogramme horizontalement, il va de soi qu'on pourrait également le construire de haut en bas, voire de bas en haut. D'autre part, les branches sont parfois également représentées par des segments obliques.

EXEMPLE

Sur un exemple arbitraire, proposons-nous d'effectuer une **classification automatique** en utilisant la **méthode du chaînage simple** et représentons le dendrogramme illustrant les différentes agrégations effectuées.

Prenons l'exemple des notes obtenues par cinq étudiants à une session d'examens portant sur quatre cours à savoir : l'anglais, le français, les mathématiques et la physique.

Nous désirons répartir ces cinq étudiants en deux groupes à l'aide de la méthode du chaînage simple pour tester une nouvelle méthode d'enseignement.

Ces notes, de 1 à 6, indiquées au demi-point, sont résumées dans le tableau suivant :

	Anglais	Français	Maths	Physique
Alain	5,0	3,5	4,0	4,5
Jean	5,5	4,0	5,0	4,5
Marc	4,5	4,5	4,0	3,5
Paul	4,0	5,5	3,5	4,0
Pierre	4,0	4,5	3,0	3,5

Nous obtenons le **tableau de distances** en mettant les distances euclidiennes entre les étudiants dans le tableau suivant :

	Alain	Jean	Marc	Paul	Pierre
Alain	0	1,22	1,5	2,35	2
Jean	1,22	0	1,8	2,65	2,74
Marc	1,5	1,8	0	1,32	1,12
Paul	2,35	2,65	1,32	0	1,22
Pierre	2	2,74	1,12	1,22	0

À l'aide de la méthode du chaînage simple, qui utilise la distance minimale entre les objets de deux classes, nous obtenons les partitions suivantes, à chaque étape :

*1*re *étape* : nous regroupons Marc et Pierre à une distance de 1,12 :

{Alain}, {Jean}, {Marc, Pierre}, {Paul}

*2*e *étape* : le nouveau tableau de distances est le suivant :

	Jean	Marc et Pierre	Paul
Alain	1,22	1,5	2,35
Jean		1,8	2,65
Marc et Pierre			1,22

Nous regroupons donc Alain et Jean à une distance de 1,22 et obtenons la partition suivante :

{Alain, Jean}, {Marc, Pierre}, {Paul}

*3*e *étape* : le nouveau tableau de distances est le suivant :

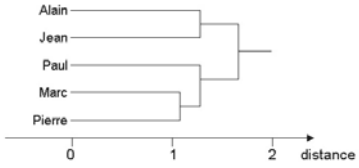
	Marc et Pierre	Paul
Alain et Jean	1,5	2,35
Marc et Pierre		1,22

Nous regroupons alors Paul avec Marc et Pierre à une distance de 1,22 pour obtenir la partition suivante :

{Alain, Jean}, {Marc, Pierre, Paul}

*4*e *étape* : les deux groupes cherchés peuvent encore être rassemblés à une distance de 1,5.

Cela donne le **dendrogramme** suivant :



Nous remarquons qu'il a fallu placer Paul avant le groupe Marc et Pierre pour ne pas entremêler les branches de notre diagramme.

MOTS CLÉS

Classification *(Classification)*
Classification automatique *(Cluster analysis)*
Distance *(Distance)*
Méthode du chaînage complet *(Complete linkage method)*
Méthode du chaînage simple *(Single link method)*
Tableau de distances *(Distance table)*

RÉFÉRENCES

Mayr, E., Linsley, E.G. and Usinger, R.L. (1953). *Methods and Principles of Systematic Zoology.* McGraw-Hill, New York.

Sneath, P.H.A. and Sokal, R.R. (1973). *Numerical Taxonomy. The principles and practice of numerical classification. A Series of Books in Biology.* San Francisco: W.H. Freeman and Comp.

Sokal, R.R. and Sneath, P.H. (1963). *Principles of Numerical Taxonomy.* W.H. Freeman, San Francisco.

DÉPENDANCE

Dependence

La notion de dépendance peut prendre deux sens : le premier concerne les événements d'une **expérience aléatoire**. On dit alors que deux événements sont dépendants si la réalisation de l'un dépend de la réalisation de l'autre.

Le deuxième sens du mot dépendance concerne la **relation**, en général fonctionnelle, qui peut exister entre des variables aléatoires. On peut mesurer cette dépendance et, dans la plupart des cas, ces mesures font intervenir la **covariance** des variables aléatoires.

La mesure de dépendance la plus utilisée est le **coefficient de corrélation**. On peut aussi tester l'**hypothèse nulle** selon laquelle les deux variables aléatoires X et Y sont indépendantes, contre l'**hypothèse alternative** de dépendance. Des tests utilisés dans ce but sont notamment le **test d'indépendance du chi-carré**, de même que les tests basés sur le **coefficient de corrélation des rangs de Spearman** et sur le **coefficient de corrélation des rangs de Kendall**.

MOTS CLÉS

Coefficient de corrélation *(Correlation coefficient)*
Coefficient de corrélation des rangs de Kendall *(Kendall rank correlation coefficient)*
Coefficient de corrélation des rangs de Spearman *(Spearman rank correlation coefficient)*
Covariance *(Covariance)*
Test d'indépendance *(Test of independence)*
Test d'indépendance du chi-carré *(Chi-square test of independence)*

DÉTERMINANT

Determinant

À toute **matrice** carrée $\mathbf{A}$ d'ordre n, on peut associer un nombre appelé déterminant de la matrice $\mathbf{A}$.

ASPECTS MATHÉMATIQUES

Soit $\mathbf{A} = (a_{ij})$ une **matrice** carrée d'ordre n, le déterminant de $\mathbf{A}$, noté dét($\mathbf{A}$) ou $|\mathbf{A}|$ est donné par la somme suivante :

$$\text{dét}\,(\mathbf{A}) = \sum (\pm)\, a_{1i} \cdot a_{2j} \cdot \ldots \cdot a_{nr}$$

où la somme est prise sur toutes les permutations du second indice. On assigne à un terme le signe + si le nombre d'inversions dans $(i, j, \ldots, r)$ est pair (la permutation est alors dite paire) et − si le nombre est impair (permutation impaire).

On peut également définir le déterminant en « développant » suivant une ligne ou suivant une colonne. Ainsi, par exemple, en développant la matrice suivant la première ligne :

$$\text{dét}\,(\mathbf{A}) = \sum_{j=1}^{n} (-1)^{1+j} \cdot a_{1j} \cdot A_{1j}$$

où A_{1j} est le déterminant de la « sous-matrice » carrée d'ordre $n-1$ obtenue de $\mathbf{A}$ en biffant la première ligne et la j-ème colonne, A_{1j} est appelé cofacteur de l'élément a_{1j}.

Le choix de la ligne ou de la colonne étant arbitraire, on peut écrire :

$$\text{dét}\,(\mathbf{A}) = \sum_{j=1}^{n} (-1)^{i+j} \cdot a_{ij} \cdot A_{ij}$$

pour i fixé (développement suivant la i-ème ligne) ou

$$\text{dét}\,(\mathbf{A}) = \sum_{i=1}^{n} (-1)^{i+j} \cdot a_{ij} \cdot A_{ij}$$

pour j fixé (suivant la j-ème colonne).

Cette deuxième façon de définir le déterminant est récursive car on calcule le déterminant d'une matrice carrée d'ordre n à l'aide de déterminants de matrices d'ordre $n-1$.

Propriétés du déterminant

- Le déterminant d'une matrice carrée $\mathbf{A}$ est égal au déterminant de sa **transposée** $\mathbf{A}'$:

$$\text{dét}\,(\mathbf{A}) = \text{dét}\,(\mathbf{A}')\,.$$

- Si on échange deux lignes (ou deux colonnes) de la matrice $\mathbf{A}$, alors le déterminant change de signe.
- Si tous les éléments d'une ligne (ou d'une colonne) sont nuls, alors le déterminant est nul.
- Si on ajoute à une ligne de la matrice $\mathbf{A}$ un multiple quelconque des éléments correspondants d'une autre ligne de $\mathbf{A}$, alors le déterminant reste inchangé. Ainsi, si deux lignes de $\mathbf{A}$ sont identiques, il suffit de soustraire l'une à l'autre pour obtenir une ligne dont tous les éléments sont nuls et donc, par la propriété précédente, le déterminant de $\mathbf{A}$ est nul. Il en va de même lorsqu'une ligne est multiple de l'autre.
- Le déterminant d'une matrice qui n'a que des « 0 » en-dessous (ou en-dessus) de la diagonale est égal au produit des éléments de la diagonale (la matrice est dite triangulaire).
- Soient $\mathbf{A}$ et $\mathbf{B}$ deux matrices carrées d'ordre n, le déterminant du produit des deux matrices est égal au produit des déterminants des deux matrices :

$$\text{dét}\,(\mathbf{A}\cdot\mathbf{B}) = \text{dét}\,(\mathbf{A})\cdot\text{dét}\,(\mathbf{B})\,.$$

EXEMPLE

1. Soit $\mathbf{A}$ une **matrice** carrée d'ordre 2 :

$$\mathbf{A} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

$$\text{dét}(\mathbf{A}) = a\cdot d - b\cdot c$$

2. Soit $\mathbf{B}$ une matrice carrée d'ordre 3 :

$$\mathbf{B} = \begin{bmatrix} 3 & 2 & 1 \\ 4 & -1 & 0 \\ 2 & -2 & 0 \end{bmatrix}$$

En développant suivant la dernière colonne :

$$\begin{aligned} \text{dét}\,(\mathbf{B}) &= 1\cdot\begin{vmatrix} 4 & -1 \\ 2 & -2 \end{vmatrix} + 0\cdot\begin{vmatrix} 3 & 2 \\ 2 & -2 \end{vmatrix} \\ &\quad + \cdot\begin{vmatrix} 3 & 2 \\ 4 & -1 \end{vmatrix} \\ &= 1\cdot[4\cdot(-2) - (-1)\cdot 2] = -8+2 \\ &= -6 \end{aligned}$$

MOTS CLÉS

Inversion *(Inversion)*
Matrice *(Matrix)*
Permutation *(Permutation)*

RÉFÉRENCE

Dodge, Y. (2002). *Mathématiques de base pour économistes.* Springer-Verlag France, Paris.

DÉVIATION

Deviation

La notion de déviation indique la différence entre une **valeur** observée et une valeur fixe de l'ensemble des valeurs possibles d'une **variable quantitative**.

Cette valeur fixe est souvent la **moyenne arithmétique** de l'ensemble des valeurs ou la **médiane**.

ASPECTS MATHÉMATIQUES

Soit l'ensemble de valeurs $x_1, x_2, \ldots, x_n$. Notons $\bar{x}$ la **moyenne arithmétique** de ces valeurs. La déviation d'une valeur particulière x_i par rapport à la moyenne arithmétique est égale à :

$$\text{déviation} = (x_i - \bar{x})\,.$$

De façon analogue, si on note M_d la **médiane** de ces mêmes valeurs, la déviation d'une valeur x_i par rapport à la médiane vaut :

$$\text{déviation} = (x_i - M_d)\,.$$

MOTS CLÉS

Écart moyen *(Mean deviation)*
Écart type *(Standard deviation)*
Variance *(Variance)*

DIAGRAMME CIRCULAIRE

Pie chart

Le diagramme circulaire est une sorte de **graphique quantitatif**. Il est constitué d'un cercle divisé en

secteurs, chaque secteur ayant un angle au centre proportionnel à la grandeur représentée.

HISTORIQUE

Voir **représentation graphique**.

ASPECTS MATHÉMATIQUES

Pour construire un diagramme circulaire, il faut tout d'abord faire le total des fréquences à représenter, puis calculer les fréquences relatives que représentent les différents secteurs.

Pour dessiner ces secteurs sur le cercle, il faut convertir les fréquences relatives en degrés par la transformation suivante :

$$\alpha = f \cdot 360^o$$

où α représente l'angle au centre et f la fréquence relative.

DOMAINES ET LIMITATIONS

Les diagrammes circulaires sont utilisés pour donner une représentation visuelle de données formant les différentes parties d'un tout. On les trouve surtout dans les médias (télévision, journaux, magazines...) où ils expliquent de manière simple et synthétique une situation de fait difficile à comprendre sous forme de chiffres.

EXEMPLE

Nous allons représenter la répartition de l'état matrimonial en Australie au 30 juin 1981 à l'aide d'un diagramme circulaire. Les données sont les suivantes :

Distribution de l'état matrimonial en Australie au 30 juin 1981 (en milliers)

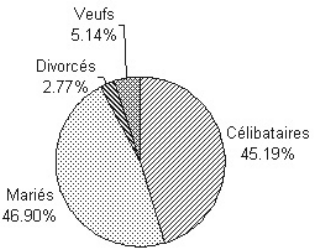
État matrimonial	Fréquence	Fréquence relative
Célibataires	6 587, 3	0, 452
Mariés	6 836, 8	0, 469
Divorcés	403, 5	0, 028
Veufs	748, 7	0, 051
Total	14 576, 3	1, 000

Source : Bureau de statistique australien, Australian Pocket Year Book 1984, p. 11

Nous transformons les fréquences relatives en degrés :

a) Célibataires $\alpha = 0,452 \cdot 360^o = 162,72^o$
b) Mariés $\alpha = 0,469 \cdot 360^o = 168,84^o$
c) Divorcés $\alpha = 0,028 \cdot 360^o = 10,08^o$
d) Voeufs $\alpha = 0,051 \cdot 360^o = 18,36^o$

Le diagramme circulaire est alors :



MOTS CLÉS

Diagramme en barres *(Bar chart)*
Diagramme figuratif *(Pictogram)*
Graphique quantitatif *(Quantitative graph)*
Représentation graphique *(Graphical representation)*

DIAGRAMME EN BARRES

Bar chart

Un diagramme en barres est une sorte de **graphique quantitatif**. Il consiste en une suite de barres verticales ou horizontales de largeur identique et de longueur égale aux quantités à représenter.

Les diagrammes en barres sont utilisés soit pour comparer les catégories d'une **variable qualitative catégorielle**, soit pour comparer des ensembles de données d'une **variable** portant sur différentes années ou différents endroits.

HISTORIQUE

Voir **représentation graphique**.

ASPECTS MATHÉMATIQUES

Pour construire un diagramme en barres verticales, il faut fixer deux axes : un vertical et un horizontal.

Sur l'axe horizontal, figurent les différentes catégories ; sur l'axe vertical, on dispose les différentes quantités.

Si l'on veut construire un diagramme en barres horizontales, il faut inverser les axes.

Les barres doivent toutes avoir la même largeur puisque seule leur longueur est utilisée pour les comparaisons.

On peut ajouter, pour faciliter la compréhension du graphique, des zones d'ombre, des zones hachurées ou de la couleur.

DOMAINES ET LIMITATIONS

Un diagramme en barres peut aussi être utilisé pour représenter des quantités négatives. Pour cela, l'échelle des quantités doit descendre au-dessous de zéro.

Il existe plusieurs sortes de diagrammes en barres. Celui qui est décrit ci-dessus s'intitule diagramme en barres simples.

Le diagramme en barres multiples sert à comparer plusieurs variables.

Le diagramme en barres composées est un diagramme en barres multiples où les différents ensembles de données sont mis les uns sur les autres. Ce genre de diagramme est employé quand les différentes données forment un tout que l'on peut comparer dans le temps.

Il existe une autre façon d'expliciter un ensemble de données formant un tout qui consiste à les représenter sous la forme d'un diagramme en barres divisées en pourcentage. L'ensemble des données représente 100 % et les différentes composantes sont calculées en **pourcentage** du tout (voir aussi **diagramme circulaire**).

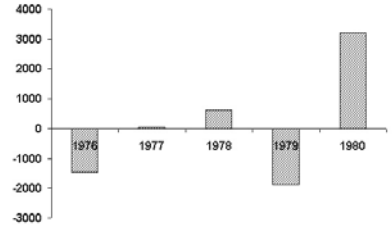
EXEMPLES

Voici un exemple de *diagramme en barres simples* :

Balance des paiements du Royaume-Uni de 1976 à 1980

Année	Balance des paiements (millions de £)
1976	−1476
1977	54
1978	620
1979	−1863
1980	3206

Source : The Pink Book

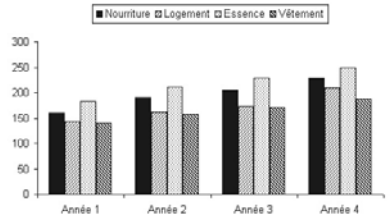


Voici un exemple de *diagramme en barres multiples* :

Prix de détail indexé sur 4 *années (année de base : année 0 =* 100*)*

	Année 1	Année 2	Année 3	Année 4
Nourriture	160	190	204	228
Logement	143	162	173	209
Essence	182	211	228	250
Vêtements	139	157	171	187

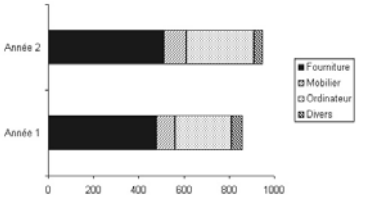
Source : Monthly Digest of Statistics



Pour illustrer un *diagramme en barres composées*, considérons les données suivantes :

Achats	Année 1	Année 2
Fourniture	480	510
Mobilier	80	100
Ordinateur	250	300
Divers	50	40
Total	860	950

Le diagramme obtenu est le suivant :

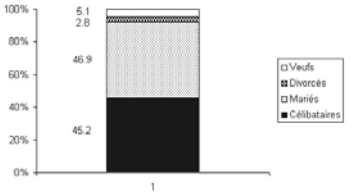


Construisons un *diagramme en barres divisées en pourcentage* sur la base des données du **tableau de fréquences** suivant :

Distribution de l'état matrimonial en Australie au 30 juin 1981 (en milliers)

État matrimonial	Fréquence	Fréquence relative	Pourcentage
Célibataires	6587, 3	0, 452	45, 2
Mariés	6836, 8	0, 469	46, 9
Divorcés	403, 5	0, 028	2, 8
Veufs	748, 7	0, 051	5, 1
Total	14576, 3	1, 000	100, 0

Source : Bureau de Statistique Australien, Australian Pocket Year Book 1984, p.11



MOTS CLÉS

Graphique quantitatif *(Quantitative graph)*
Représentation graphique *(Graphical representation)*

DIAGRAMME EN BÂTONS

Line chart

Le diagramme en bâtons est une sorte de graphique de fréquences. Il se prête bien à la représentation d'une **distribution de fréquences** d'une **variable aléatoire** discrète que l'on veut représenter. Il se présente sous la forme d'un graphe où l'on met en abscisse les valeurs de la **variable** et en ordonnée les fréquences correspondant à ces valeurs.

HISTORIQUE

Voir **représentation graphique**.

ASPECTS MATHÉMATIQUES

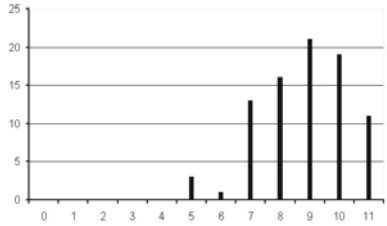
Le diagramme en bâtons possède deux axes. On place en abscisse les valeurs x_i de la **variable aléatoire** discrète *que l'on veut représenter*. On élève ensuite au-dessus de chaque valeur x_i un bâton dont la hauteur est égale à la **fréquence** n_i (ou fréquence relative f_i) correspondant à cette valeur.

EXEMPLE

Dans le tableau suivant, on trouve la distribution des pointures de souliers pour un **échantillon** de 84 hommes.

Pointure	Fréquence
5	3
6	1
7	13
8	16
9	21
10	19
11	11
Total	84

Voici le diagramme en bâtons construit à partir de ce tableau :



MOTS CLÉS

Diagramme en points *(Dot plot)*
Distribution de fréquences *(Frequency distribution)*
Histogramme *(Histogram)*
Représentation graphique *(Graphical representation)*

DIAGRAMME EN POINTS

Dot plot

Le diagramme en points est une forme de graphique de fréquences.

Quand le nombre de données est relativement petit (jusqu'à 25 environ), on peut représenter ces données et leurs fréquences par un diagramme en points.

Ce type de **représentation graphique** est tout spécialement utile pour attirer l'attention sur des observations aberrantes.

ASPECTS MATHÉMATIQUES

Le diagramme en points se construit de la façon suivante : on place une échelle comprenant toutes les valeurs de l'ensemble de données ; puis on marque les données prises individuellement par un point au-dessus de la valeur correspondante ; s'il y a plusieurs données avec la même valeur, on aligne les points les uns au-dessus des autres.

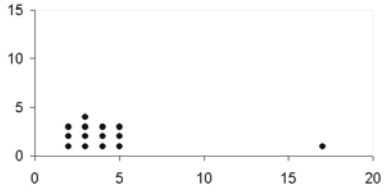
EXEMPLE

On considère les données suivantes :

5, 2, 4, 5, 3, 2, 4, 3, 5, 2, 3, 4, 3, 17.

Pour construire un diagramme en points, on fixe une droite échelonnée de 0 à 20.

Puis on représente les données par des points au-dessus de chaque **valeur**. Cela donne :



On remarque que la série comprend une **observation aberrante** de valeur égale à 17.

MOTS CLÉS

Diagramme en bâtons *(Line chart)*
Distribution de fréquences *(Frequency distribution)*
Histogramme *(Histogram)*
Représentation graphique *(Graphical representation)*

DIAGRAMME FIGURATIF

Pictogram

Le diagramme figuratif est un **graphique quantitatif**, c'est-à-dire qu'il est utilisé pour représenter des quantités.

Les diagrammes figuratifs sont construits à l'aide de figures (bouteilles, automobiles, sacs de blé, etc.) de dimensions variables proportionnelles aux grandeurs représentées, ou bien à l'aide de figures de même taille (sorte d'étalon de mesure) que l'on reproduit un certain nombre de fois selon les grandeurs à représenter.

Comme le **diagramme en barres**, le diagramme figuratif est utilisé soit pour comparer les catégories d'une **variable qualitative catégorielle**, soit pour comparer des ensembles de données d'une **variable** portant sur différentes années ou différents endroits.

Les diagrammes figuratifs sont surtout employés dans les revues comme graphiques publicitaires à des fins de comparaison et sont de peu d'intérêt statistique car ils fournissent des approximations assez grossières.

ASPECTS MATHÉMATIQUES

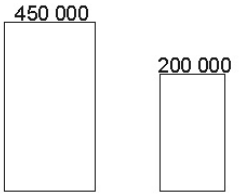
La figure choisie pour illustrer les différentes quantités dans un diagramme figuratif doit si possible être divisible en quatre parties, de sorte que des moitiés et des quarts puissent être représentés.

L'échelle est choisie en fonction de la place disponible. Par conséquent, la figure n'est pas forcément égale à l'unité.

EXEMPLE

Comparons, à l'aide d'un diagramme figuratif, la production de blé de deux pays A et B. La production du pays A est égale à 450 000 quintaux et celle du pays B à 200 000 quintaux.

Cela donne :



La surface de chaque figure est proportionnelle à la quantité à représenter.

MOTS CLÉS

Diagramme en barres *(Bar chart)*
Graphique quantitatif *(Quantitative graph)*
Représentation graphique *(Graphical representation)*

DIAGRAMME STEM AND LEAF

Stem and leaf diagram

Le diagramme stem and leaf est un graphique de fréquences. L'idée de base est de fournir une information sur la **distribution de fréquences**, tout en retenant les valeurs mêmes des données. En effet, *stem* ou « tige » correspond aux intervalles de classes et *leaf* ou « feuille » correspond aux nombres d'observations dans la classe, représentées par les différentes données. On peut donc y lire directement les valeurs des données.

HISTORIQUE

L'origine du diagramme stem and leaf est très souvent associée à **John Wilder Tukey** (1970-71). Son concept est fondé sur l'**histogramme** qui date déjà du XVIII^{e} siècle. Dans les travaux de Emerson et Hoaglin (1983), on trouve des variations du diagramme de base, sa relation avec l'histogramme et de nombreux exemples.

ASPECTS MATHÉMATIQUES

Pour construire un diagramme stem and leaf, il faut couper chaque nombre en une partie principale (*stem*) et une partie secondaire (*leaf*).

Par exemple, $4, 2$, 42 ou 420 peuvent être séparés en 4 pour le *stem* et 2 pour le *leaf*.

Ensuite, les « tiges » sont listées verticalement (une ligne par « tige ») par ordre de grandeur et les « feuilles » sont écrites à côté, également par ordre de grandeur. Si un groupe de données comprend les valeurs $4, 2$, $4, 4$, $4, 8$, $4, 4$, $4, 1$, le stem and leaf se présente ainsi :

$$4|12448.$$

DOMAINES ET LIMITATIONS

Le diagramme stem and leaf donne des renseignements sur la distribution, la symétrie, la concentration, les vides et les observations aberrantes.

Quelquefois, il se peut qu'il y ait trop de nombres ayant la même « tige ». Dans ce cas-là, on ne peut pas tous les mettre sur la même ligne. La solution la plus utilisée consiste à séparer les 12 « feuilles » en deux fois 6 « feuilles ». Par exemple, 4|012244 et 4|567889 à la place de 4|012244567889.

On peut comparer deux séries de données en mettant les deux « feuilles » à droite et à gauche de la même « tige » :

012244|4|567889.

EXEMPLE

Le tableau suivant nous donne pour la plupart des pays d'Afrique le produit intérieur brut, par habitant, en dollars pour l'année 1970.

Produit intérieur brut par habitant - 1970 (en dollars)

AFRIQUE			
Algérie	324	Maurice	233
Angola	294	Maroc	224
Bénin	81	Mozambique	228
Botswana	143	Niger	90
Burundi	67	Nigéria	145
Rép. centrafric.	127	Réunion	769
Tchad	74	Rwanda	60
Comores	102	Sénégal	236
Congo	238	Sierra Leone	167
Égypte	217	Somalie	89
Guinée équat.	267	Afrique du Sud	758
Éthiopie	72	Rhodésie (Sud)	283
Gabon	670	Soudan	117
Gambie	101	Swaziland	270
Ghana	257	Togo	135
Guinée	82	Tunisie	281
Guinée-Bissau	259	Ouganda	135
Côte d'Ivoire	347	Rép.-Unie	
Kenya	143	du Cameroun	185
Lesotho	74	Rép.-Unie	
Libéria	268	de Tanzanie	100
Madagascar	133	Haute-Volta	59
Malawi	74	Zaïre	88
Mali	54	Zambie	421
Mauritanie	165		

Source : Annuaire statistique de l'ONU, 1976

Pour construire un diagramme stem and leaf à partir de ces données, on commence par les classer par ordre de grandeur. Puis on choisit le *stem* : dans cet exemple, on prendra les centaines. On prend ensuite les chiffres des dizaines de chaque donnée, que l'on place à droite de la centaine correspondante, ce qui nous donne la partie *leaf* du diagramme. Les unités ne sont pas prises en compte.

Le stem and leaf est le suivant :

Unité : 1 | 1 = 110

0	5566777788889
1	00012333444668
2	12233355667889
3	24
4	2
5	
6	7
7	56

MOTS CLÉS

Distribution de fréquences *(Frequency distribution)*
Histogramme *(Histogram)*
Représentation graphique *(Graphical representation)*

RÉFÉRENCES

Emerson, J.D. and Hoaglin, D.C. (1983). In *Understanding Robust and Exploratory Data Analysis*, D.C. Hoaglin, F. Mosteller and J.W. Tukey. Eds John Wiley & Sons, New York.

Tukey, J.W. (1970-1971). *Exploratory data analysis*, limited preliminary edition. Addison-Wesley, Reading, Mass.

DICHOTOMIE

Dichotomy

La dichotomie est la division des individus d'une **population** ou d'un **échantillon** en deux groupes, en fonction d'un critère déterminé.

On parle de **variable dichotomique** quand une **variable** ne peut prendre que deux valeurs. On est alors en présence de données binaires.

MOTS CLÉS

Donnée binaire *(Binary data)*
Variable dichotomique *(Dichotomous variable)*

DIFFÉRENCE SIGNIFICATIVE MINIMALE, TEST DE LA

Least significant difference test

Voir **test de la différence significative minimale**.

DISPERSION

Dispersion

Voir **mesure de dispersion**.

DISTANCE

Distance

On appelle distance entre deux objets la mesure de l'éloignement entre ces deux objets. Une distance correspond donc à un nombre réel :

- nul, si les deux objets sont confondus ;
- strictement positif sinon.

ASPECTS MATHÉMATIQUES

Soient X, Y et Z trois objets quelconques ; notons $d(X,Y)$ la distance entre X et Y.

Une distance vérifie les propriétés suivantes :

- elle est positive

$$d(X,Y) > 0$$

et nulle si et seulement si les objets sont confondus

$$d(X,Y) = 0 \quad \Leftrightarrow \quad X = Y \ ;$$

- elle est symétrique, c'est-à-dire

$$d(X,Y) = d(Y,X) \ ;$$

- elle vérifie l'inégalité, dite inégalité triangulaire

$$d(X,Z) \leq d(X,Y) + d(Y,Z)$$

qui se résume en disant que la distance d'un objet à l'autre est inférieure ou égale à celle obtenue en passant par un troisième objet.

Soient X et Y deux vecteurs à n composantes,

$X = (x_i)'$ et $Y = (y_i)'$ pour $i = 1, 2, \ldots, n$.

Une famille de distances souvent utilisée est la suivante :

$$d(X,Y) = \sqrt[p]{\sum_{i=1}^{n} |x_i - y_i|^p} \text{ avec } p \geq 1.$$

Ces distances portent le nom de *distances de Minkovski*.

En posant $p = 1$, on obtient ce que l'on appelle une *distance absolue* ou encore *norme L1*.

Pour $p = 2$, on obtient la *distance euclidienne* définie par :

$$d(X,Y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}.$$

La distance euclidienne est la distance la plus fréquemment utilisée.

Généralement, lorsqu'on parle de distance sans spécification, on sous-entend la distance euclidienne.

On mentionne encore les distances de Minkovski pondérées définies par :

$$d(X,Y) = \sqrt[p]{\sum_{i=1}^{n} w_i \cdot |x_i - y_i|^p} \text{ avec } p \geq 1$$

et où w_i, $i = 1, \ldots, n$ représentent les différents poids, dont la somme vaut 1.

Lorsque $p = 1$ ou 2, on obtient la distance absolue et euclidienne pondérée, respectivement.

DOMAINES ET LIMITATIONS

Pour un ensemble de r objets distincts X_1, X_2, ..., X_r, on peut calculer pour chaque paire (X_i, X_j) la distance entre l'objet i et l'objet j. Ces distances permettent de construire le **tableau de distances** ou plus généralement un tableau de dissimilarité de terme général

$$d_{ij} = d(X_i, X_j) \quad \text{pour } i, j = 1, 2, \ldots, r$$

utilisé par les méthodes de **classification automatique**.

On note que si l'inégalité triangulaire n'est pas satisfaite, on parle d'une **mesure de dissimilarité**.

EXEMPLE

On prend l'exemple des notes obtenues par cinq étudiants à une session d'examens portant sur quatre cours à savoir :

l'anglais, le français, les mathématiques et la physique.

Ces notes, de 1 à 6, indiquées au demi-point sont résumées dans le tableau suivant :

	Anglais	Français	Maths	Physique
Alain	5,0	3,5	4,0	4,5
Jean	3,5	4,0	5,0	5,5
Marc	4,5	4,5	4,0	3,5
Paul	6,0	5,5	5,5	5,0
Pierre	4,0	4,5	2,5	3,5

La distance euclidienne d'Alain à Jean, qui sera notée d_2, est donnée par :

$$\begin{aligned} & d_2(Alain, Jean) \\ &= \sqrt{(5,0-3,5)^2 + \ldots + (4,5-5,5)^2} \\ &= \sqrt{2,25+0,25+1+1} \\ &= \sqrt{4,5} = 2,12 \end{aligned}$$

Tandis que la distance absolue d'Alain à Jean, notée d_1, est donnée par :

$$\begin{aligned} d_1(Alain, Jean) &= |5,0-3,5| + \ldots + |4,5-5,5| \\ &= 1,5+0,5+1+1 = 4,0. \end{aligned}$$

Et ainsi de suite on trouve :

$$\begin{array}{lcll} d_2(Alain, Marc) & = & \sqrt{2,25} & = 1,5 \\ d_2(Alain, Paul) & = & \sqrt{7,5} & = 2,74 \\ d_2(Alain, Pierre) & = & \sqrt{5,25} & = 2,29 \\ d_2(Jean, Marc) & = & \sqrt{6,25} & = 2,5 \\ d_2(Jean, Paul) & = & \sqrt{9} & = 3 \\ d_2(Jean, Pierre) & = & \sqrt{10,75} & = 3,28 \\ d_2(Marc, Paul) & = & \sqrt{7,75} & = 2,78 \\ d_2(Marc, Pierre) & = & \sqrt{2,5} & = 1,58 \\ d_2(Paul, Pierre) & = & \sqrt{16,25} & = 4,03 \\ d_1(Alain, Marc) & = & 2,5 & \\ d_1(Alain, Paul) & = & 5,0 & \\ d_1(Alain, Pierre) & = & 4,5 & \end{array}$$

$$\begin{array}{lcll} d_2(Paul, Pierre) & = & \sqrt{16,25} & = 4,03 \\ d_1(Alain, Marc) & = & 2,5 & \\ d_1(Alain, Paul) & = & 5,0 & \\ d_1(Alain, Pierre) & = & 4,5 & \\ d_1(Jean, Marc) & = & 4,5 & \\ d_1(Jean, Paul) & = & 5,0 & \\ d_1(Jean, Pierre) & = & 5,5 & \\ d_1(Marc, Paul) & = & 6,5 & \\ d_1(Marc, Pierre) & = & 2,0 & \\ d_1(Paul, Pierre) & = & 7,5 & \end{array}$$

On obtient le **tableau de distances** euclidiennes en mettant ces résultats dans le tableau suivant :

	Alain	Jean	Marc	Paul	Pierre
Alain	0	2,12	1,5	2,74	2,29
Jean	2,12	0	2,5	3	3,28
Marc	1,5	2,5	0	2,78	1,58
Paul	2,74	3	2,78	0	4,03
Pierre	2,29	3,28	1,58	4,03	0

De même, on trouve le tableau de distances absolues suivant :

	Alain	Jean	Marc	Paul	Pierre
Alain	0	4	2,5	5	4,5
Jean	4	0	4,5	5	5,5
Marc	2,5	4,5	0	6,5	2
Paul	5	5	6,5	0	7,5
Pierre	4,5	5,5	2	7,5	0

On remarque que selon la distance choisie, l'ordre de proximité des individus peut être différent : par la distance euclidienne, c'est Alain qui est le plus proche de Marc ($d_2(Alain, Marc) = 1,5$), alors que par

la distance absolue, Alain ne vient qu'en deuxième position ($d_1(Alain, Marc) = 2, 5$), et c'est Pierre qui est le plus proche de Marc avec une distance de 2 ($d_1(Pierre, Marc) = 2$).

MOTS CLÉS

Classification automatique *(Cluster analysis)*
Distance du chi-carré *(Chi-square distance)*
Mesure de dissimilarité *(Dissimilarity measure)*
Tableau de distances *(Distance table)*
Vecteur *(Vector)*

DISTANCE DE MAHALANOBIS

Mahalanobis distance

On appelle distance de Mahalanobis de X à la quantité μ :

$$d_M(X, \mu) = (X - \mu)^t C^{-1} (X - \mu).$$

Cette distance permet de mesurer la distance d'une réalisation d'un vecteur aléatoire à un échantillon caractérisé par sa position (vecteur moyenne) et sa dispersion (matrice de covariance).

HISTORIQUE

La distance de Mahalanobis a été introduite par **Prasanta Chandra Mahalanobis** en 1930.

ASPECTS MATHÉMATIQUES

Si μ_i désigne $E(X_i)$, on appelle par définition espérance de $\mathbf{X} = (X_1, ..., X_p)$ le vecteur certain :

$$E[\mathbf{X}] = \boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ . \\ . \\ . \\ \mu_p \end{bmatrix}$$

La matrice de variance-covariance $\boldsymbol{\Sigma}$ de $\mathbf{X}$ est définie par :

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & cov(X_1, X_2) & ... & cov(X_1, X_p) \\ . & \sigma_2^2 & & . \\ . & & & . \\ . & & & . \\ . & & & . \\ . & & & \sigma_p^2 \end{bmatrix}$$
$$= E[\mathbf{X}\mathbf{X}'] - \boldsymbol{\mu}\boldsymbol{\mu}'$$

C'est une matrice carrée symétrique d'ordre p. Si les variables X_i sont réduites, $\boldsymbol{\Sigma}$ s'identifie avec la matrice de corrélation :

$$\begin{bmatrix} 1 & \rho_{1\,2} & \rho_{1\,3} & ... & \rho_{1\,p} \\ & 1 & \rho_{2\,3} & ... & \rho_{2\,p} \\ & & 1 & ... & \rho_{3\,p} \\ & & & \ddots & \vdots \\ & & & & 1 \end{bmatrix}$$

On appelle transformation de Mahalanobis la transformation définie par $\boldsymbol{\Sigma}^{-1/2}$.

$\mathbf{Y} = \boldsymbol{\Sigma}^{-1/2}(\mathbf{X} - \boldsymbol{\mu})$ est alors un vecteur aléatoire centré-réduit à composantes non correlées. La variable aléatoire $(\mathbf{X} - \boldsymbol{\mu})' \ \boldsymbol{\Sigma}^{-1}(\mathbf{X} - \boldsymbol{\mu}) = D^2$ a pour espérance p. En effet, $D^2 = \sum_{i=1}^{p} Y_i^2$ où les Y_i sont d'espérance nulle et de variance 1. D est appelée distance de Mahalanobis de $\mathbf{X}$ à $\boldsymbol{\mu}$.

MOTS CLÉS

Analyse factorielle des correspondances *(Correspondence analysis)*
Analyse de covariance *(Covariance analysis)*
Analyse de variance *(Variance analysis)*
Classification *(Classification)*
Distance *(Distance)*

RÉFÉRENCE

Mahalanobis, P.C. (1930). On Tests and Meassures of Groups Divergence. Part I. Theoretical formulæ. *Journal of the Asiatic Society of Bengal*, 26, pp. 541-588.

DISTANCE DU CHI-CARRÉ

Chi-square distance

Dans un **tableau de fréquences** relatives à n lignes et p colonnes, on a la possibilité de calculer les profils-lignes et les profils-colonnes. Ces derniers permettent de définir des points dans des espaces à n ou p dimensions, dans lesquels seront définies les distances entre ces points. On appelle distance du chi-carré, la distance euclidienne entre les composantes des profils sur laquelle on fait intervenir une pondération (le poids correspondant à chaque terme étant l'inverse de sa **fréquence**).

ASPECTS MATHÉMATIQUES

Soit (f_{ij}) un **tableau de fréquences** relatives à n lignes et p colonnes. La distance du chi-carré entre deux lignes i et i' est donnée par la formule :

$$d\left(i,i'\right)=\sqrt{\frac{1}{f_{.j}}\cdot\sum_{j=1}^{p}\left(\frac{f_{ij}}{f_{i.}}-\frac{f_{i'j}}{f_{i'.}}\right)^{2}}$$

où

- $f_{i.}$ est la somme des composantes de la i-ème ligne ;
- $f_{.j}$ est la somme des composantes de la j-ème colonne ;
- $\left[\frac{f_{ij}}{f_{i.}}\right]$ pour $j=1,2,\ldots,p$ est le i-ème profil-ligne.

De la même façon, la distance entre 2 colonnes j et j' est donnée par :

$$d\left(j,j'\right)=\sqrt{\frac{1}{f_{i.}}\cdot\sum_{i=1}^{n}\left(\frac{f_{ij}}{f_{.j}}-\frac{f_{ij'}}{f_{.j'}}\right)^{2}}$$

où $\left[\frac{f_{ij}}{f_{.j}}\right]$ pour $j=1,\ldots,n$ est le j -ème profil-colonne.

DOMAINES ET LIMITATIONS

La distance du chi-carré qui introduit une pondération inversement proportionnelle aux totaux des lignes (ou des colonnes) permet d'augmenter l'importance des petits écarts dans les lignes ou colonnes de faible somme par rapport à celles de somme plus importante.

La distance du chi-carré a la propriété dite d'équivalence distributionnelle, c'est-à-dire qu'elle assure une invariance des distances entre lignes et colonnes lorsque l'on agrège deux colonnes (ou deux lignes) ayant des profils identiques.

EXEMPLE

À partir d'une **table de contingence** représentant la satisfaction au travail des employés de trois entreprises différentes, proposons-nous de construire le **tableau de distances** utilisant la distance du chi-carré. La **variable** X étudiée est définie par les trois catégories :

X_1 : « grande satisfaction »
X_2 : « satisfaction moyenne »
X_3 : « faible satisfaction »

Les observations recueillies sur des échantillons d'individus provenant des trois entreprises sont les suivantes :

	entr. 1	entr. 2	entr. 3	Total
X_1	20	55	30	105
X_2	18	40	15	73
X_3	12	5	5	22
Total	50	100	50	200

Le **tableau de fréquences** relatives s'obtient en divisant tous les éléments du tableau par 200, le nombre total d'observations :

	entr. 1	entr. 2	entr. 3	Total
X_1	$0,1$	$0,275$	$0,15$	$0,525$
X_2	$0,09$	$0,2$	$0,075$	$0,365$
X_3	$0,06$	$0,025$	$0,025$	$0,11$
Total	$0,25$	$0,5$	$0,25$	1

On peut chercher à calculer les distances entre les entreprises. Le tableau des profils-colonnes sera donné par :

	entr. 1	entr. 2	entr. 3	Total
X_1	$\frac{0,1}{0,25}=0,4$	$\frac{0,275}{0,5}=0,55$	$\frac{0,15}{0,25}=0,6$	$1,55$
X_2	$\frac{0,09}{0,25}=0,36$	$\frac{0,2}{0,5}=0,4$	$\frac{0,075}{0,25}=0,3$	$1,06$
X_3	$\frac{0,06}{0,25}=0,24$	$\frac{0,025}{0,5}=0,05$	$\frac{0,025}{0,25}=0,1$	$0,39$
Total	1	1	1	3

Ce qui permet de calculer les distances entre les différentes colonnes :

$$\begin{aligned} d^2(1,2) &= \frac{1}{0,525}\cdot(0,4-0,55)^2 \\ &\quad +\frac{1}{0,365}\cdot(0,36-0,4)^2 \\ &\quad +\frac{1}{0,11}\cdot(0,24-0,05)^2 \\ &= 0,375423 \\ d(1,2) &= 0,613. \end{aligned}$$

De manière analogue, on calcule $d(1,3)$ et $d(2,3)$. Les distances ainsi obtenues peuvent se résumer dans le tableau de distances suivant :

	entr. 1	entr. 2	entr. 3
entr. 1	0	0,613	0,514
entr. 2	0,613	0	0,234
entr. 3	0,514	0,234	0

On peut également calculer les distances entre les lignes, c'est-à-dire entre les degrés de satisfaction ; pour ce faire il faut déterminer le tableau des profils-lignes :

	entr. 1	entr. 2	entr. 3	Total
X_1	$\frac{0,1}{0,525}=0,19$	$\frac{0,275}{0,525}=0,524$	$\frac{0,15}{0,525}=0,286$	1
X_2	$\frac{0,09}{0,365}=0,246$	$\frac{0,2}{0,365}=0,548$	$\frac{0,075}{0,365}=0,206$	1
X_3	$\frac{0,06}{0,11}=0,546$	$\frac{0,025}{0,11}=0,227$	$\frac{0,025}{0,11}=0,227$	1
Total	0,982	1,299	0,719	3

Ce qui permet de calculer les distances entre les différentes lignes :

$$\begin{aligned} d^2(1,2) &= \frac{1}{0,25}\cdot(0,19-0,246)^2 \\ &\quad +\frac{1}{0,5}\cdot(0,524-0,548)^2 \\ &\quad +\frac{1}{0,25}\cdot(0,286-0,206)^2 \\ &= 0,039296 \\ d(1,2) &= 0,198. \end{aligned}$$

On calcule de manière analogue $d(1,3)$ et $d(2,3)$. On résume finalement les distances entre les degrés de satisfaction dans le tableau de distances suivant :

	X_1	X_2	X_3
X_1	0	0,198	0,835
X_2	0,198	0	0,754
X_3	0,835	0,754	0

MOTS CLÉS

Distance *(Distance)*
Fréquence *(Frequency)*
Table de contingence *(Contingency table)*
Tableau de distances *(Distance table)*

DISTANCES, TABLEAU DE

Distance table

Voir **tableau de distances**.

DISTRIBUTION D'ÉCHANTILLONNAGE

Sampling distribution

Une distribution d'échantillonnage est la distribution d'une **statistique** $T(X_1,\ldots,X_n)$, où $T(.)$ est une fonction des valeurs observées des variables aléatoires $X_1,\ldots,X_n$.

Nous parlons par exemple de la distribution d'échantillonnage des moyennes, ou de la distribution d'échantillonnage des variances.

HISTORIQUE

Voir **échantillonnage**.

ASPECTS MATHÉMATIQUES

Considérons une **population** de taille N et de paramètres μ et σ (μ étant la **moyenne** inconnue de la population, et σ son **écart type**).

Nous prélevons de cette population un **échantillon** de taille n. Le nombre d'échantillons de taille n qu'il

est possible de prélever est donné par le nombre de combinaisons de n objets parmi N :

$$k = C_N^n = \frac{N!}{(N-n)! \cdot n!}.$$

Chaque échantillon peut être caractérisé par exemple par sa moyenne échantillonnale que nous noterons :

$$\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_k.$$

L'ensemble de ces k moyennes forme alors la distribution d'échantillonnage des moyennes.

Nous adopterons les notations suivantes pour caractériser la moyenne et l'écart type de la distribution d'échantillonnage des moyennes :

$$\text{moyenne : } \mu_{\bar{x}} \text{ et écart type : } \sigma_{\bar{x}}.$$

Moyenne de la distribution d'échantillonnage des moyennes

Soit $X_1, \ldots, X_n$ des variables aléatoires indépendantes distribuées selon une loi quelconque de moyenne μ et de **variance** σ^2.

Par la définition de la **moyenne arithmétique**, la moyenne de la distribution d'échantillonnage des moyennes est égale à :

$$\mu_{\bar{x}} = \frac{\sum_{i=1}^{k} \bar{x}_i}{k},$$

où k est le nombre d'échantillons de taille n qu'il est possible de former à partir de la population.

Nous pouvons démontrer que l'**espérance mathématique** de la moyenne de la distribution d'échantillonnage des moyennes est égale à la moyenne de la population μ :

$$E[\mu_{\bar{x}}] = \mu.$$

Démonstration :

$$\begin{aligned}
E[\mu_{\bar{x}}] &= E\left[\frac{\sum_{i=1}^{k} \bar{x}_i}{k}\right] = \frac{1}{k} \cdot E\left[\sum_{i=1}^{k} \bar{x}_i\right] \\
&= \frac{1}{k} \cdot \sum_{i=1}^{k} E[\bar{x}_i] \\
&= \frac{1}{k} \cdot \sum_{i=1}^{k} E\left[\frac{\sum_{j=1}^{n} x_j}{n}\right] \\
&= \frac{1}{k} \cdot \sum_{i=1}^{k} \left(\frac{1}{n} \cdot E\left[\sum_{j=1}^{n} x_j\right]\right) \\
&= \frac{1}{k} \cdot \sum_{i=1}^{k} \left(\frac{1}{n} \cdot \sum_{j=1}^{n} E[x_j]\right) \\
&= \frac{1}{k} \cdot \sum_{i=1}^{k} \left(\frac{1}{n} \cdot n \cdot \mu\right) \\
&= \frac{1}{k} \cdot k \cdot \mu = \mu.
\end{aligned}$$

Écart type de la distribution d'échantillonnage des moyennes

Par la définition de l'écart type, l'écart type de la distribution d'échantillonnage des moyennes, appelé aussi **erreur type** est égal à :

$$\sigma_{\bar{x}} = \sqrt{\frac{\sum_{i=1}^{k} (\bar{x}_i - \mu_{\bar{x}})^2}{k}}.$$

De même qu'il existe une relation entre la moyenne de la distribution d'échantillonnage des moyennes et la moyenne de la population, il existe une relation entre l'erreur type et l'écart type de la population :

- dans le cas de l'étude d'un **échantillonnage** exhaustif (sans remise) sur une population finie, l'erreur type de la moyenne est donné par :

  $$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}}$$

  où σ est l'écart type de la population

- si la population est infinie, ou si l'échantillonnage étudié est non exhaustif (avec remise), l'erreur type de la moyenne est égale à l'écart type de la

population divisé par la racine carrée de la taille de l'échantillon :

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}.$$

Caractéristiques de la distribution d'échantillonnage des moyennes

- Si n est suffisamment grand ($n \geq 30$), la distribution d'échantillonnage des moyennes suit approximativement une **loi normale**, quelle que soit la distribution de la population.
- Si la population est distribuée selon une loi normale, la distribution d'échantillonnage des moyennes suit toujours une loi normale, quelle que soit la taille de l'échantillon.
- La distribution d'échantillonnage des moyennes est toujours symétrique autour de sa moyenne.

Des considérations similaires peuvent être effectuées pour d'autres statistiques que la moyenne arithmétique.

Considérons la distribution des variances calculées à partir de tous les échantillons possibles de taille n, tirés à partir d'une population normale de variance σ^2. Cette distribution est appelée distribution d'échantillonnage des variances.

Une des caractéristiques de cette distribution est qu'elle ne peut prendre que des valeurs positives ou nulles, la variance s^2 étant définie comme une somme de carrés. Il existe la relation suivante entre la distribution d'échantillonnage des variances et la **loi du chi-carré** :

$$s^2 = \frac{X^2\sigma^2}{v} = \frac{X^2\sigma^2}{n-1}$$

où X^2 est une **variable aléatoire** distribuée selon une loi du chi-carré avec $v = n - 1$ degrés de liberté.

Connaissant l'espérance mathématique et la variance d'une loi du chi-carré :

$$\begin{aligned} E\left(X^2\right) &= v \\ Var\left(X^2\right) &= 2 \cdot v, \end{aligned}$$

nous pouvons déterminer ces mêmes caractéristiques pour la distribution de s^2, dans le cas où les échantillons proviennent d'une population infinie.

Moyenne de la distribution d'échantillonnage des variances

La moyenne de la distribution de s^2 est égale à la variance de la population :

$$\begin{aligned} \mu_{s^2} &= E\left(s^2\right) = E\left[\frac{X^2\sigma^2}{v}\right] \\ &= \left(\frac{\sigma^2}{v}\right) \cdot E\left(X^2\right) \\ &\quad (\text{car } \tfrac{\sigma^2}{v} \text{ est une constante}) \\ &= \left(\frac{\sigma^2}{v}\right) \cdot v \\ &= \sigma^2. \end{aligned}$$

Variance de la distribution d'échantillonnage des variances

La variance de la distribution de s^2 dépend également de la variance de la population σ^2, mais aussi de la taille des échantillons $n = v + 1$:

$$\begin{aligned} \sigma^2_{s^2} &= Var\left(s^2\right) = Var\left[\frac{X^2\sigma^2}{v}\right] \\ &= \left(\frac{\sigma^4}{v^2}\right) \cdot Var\left(X^2\right) \\ &= \left(\frac{\sigma^4}{v^2}\right) \cdot 2v \\ &= 2 \cdot \frac{\sigma^4}{v}. \end{aligned}$$

DOMAINES ET LIMITATIONS

L'étude de la distribution d'échantillonnage d'une **statistique** permet de juger de la proximité de la statistique mesurée sur un seul **échantillon** avec le **paramètre** inconnu de la **population**. Elle permet

donc également de préciser les marges d'**erreur** des estimateurs, calculés à partir de l'échantillon lui-même.

La notion de distribution d'échantillonnage est donc à la base de la construction d'**intervalles de confiance** indiquant le degré de précision d'un estimateur, de même que de la réalisation de tests d'hypothèses sur les paramètres inconnus de la population.

EXEMPLE

Par un exemple, nous allons montrer la relation existant d'une part entre la **moyenne** de la distribution d'échantillonnage des moyennes et la moyenne de la **population**, et d'autre part entre l'**erreur type** et l'**écart type** de la population.

Considérons une population fictive de cinq magasins sur laquelle nous désirons connaître le prix moyen de vente d'un certain produit. Le tableau suivant nous donne le prix de vente du produit pour chaque magasin appartenant à la population :

Nombre de magasins	Prix
1	$30,50$
2	$32,-$
3	$37,50$
4	$30,-$
5	$33,-$

La moyenne de la population est égale à :

$$\begin{aligned}\mu &= \frac{\sum_{i=1}^{5} x_i}{5} \\ &= \frac{30,50+32+37,50+30+33}{5} \\ &= 32,60.\end{aligned}$$

En prélevant des échantillons de taille 3, nous pouvons déterminer les différentes moyennes échantillonnales pouvant résulter du tirage aléatoire de l'échantillon. Nous aurons :

$$k = C_5^3 = \frac{5!}{3! \cdot 2!} = 10 \text{échantillons possibles.}$$

Le tableau suivant représente les différents échantillons et leur moyenne échantillonnale respective :

Nombre de magasins faisant partie de l'échantillon		Moyennes échantillonnales $\bar{x}_i$
1.	$1,2,3$	$33,\overline{3}$
2.	$1,2,4$	$30,8\overline{3}$
3.	$1,2,5$	$31,8\overline{3}$
4.	$1,3,4$	$32,6\overline{6}$
5.	$1,3,5$	$33,6\overline{6}$
6.	$1,4,5$	$31,1\overline{6}$
7.	$2,3,4$	$33,1\overline{6}$
8.	$2,3,5$	$34,1\overline{6}$
9.	$2,4,5$	$31,6\overline{6}$
10.	$3,4,5$	$33,50$
	Total	$326,0$

À partir des données de ce tableau, nous sommes en mesure de calculer la moyenne de la distribution d'échantillonnage des moyennes :

$$\mu_{\bar{x}} = \frac{\sum_{i=1}^{k} \bar{x}_i}{k} = \frac{326,00}{10} = 32,60.$$

Sur cet exemple, nous pouvons constater que la moyenne de la distribution d'échantillonnage des moyennes est égale à la moyenne de la population :

$$\mu_{\bar{x}} = \mu.$$

Nous pouvons procéder de la même manière pour vérifier la relation existant entre l'erreur type et l'écart type de la population.

Selon la définition, l'écart type de la population se calcule comme suit :

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}}.$$

Selon les données du premier tableau, nous obtenons :

$$\sigma = \sqrt{\frac{35,70}{5}} = 2,6721.$$

Selon la même définition, nous pouvons à présent calculer l'écart type de la distribution

d'échantillonnage des moyennes :

$$\sigma_{\bar{x}} = \sqrt{\frac{\sum_{i=1}^{k} (\bar{x}_i - \mu_{\bar{x}})^2}{k}} = \sqrt{\frac{11.899}{10}} = 1,0908.$$

Le tableau suivant présente le détail des calculs effectués :

Nombre de magasins faisant partie de l'échantillon		Moyennes échantillonnales $\bar{x}_i$	$(\bar{x}_i - \mu_{\bar{x}})^2$
1.	1, 2, 3	$33,3\overline{3}$	0,538
2.	1, 2, 4	$30,8\overline{3}$	3,121
3.	1, 2, 5	$31,8\overline{3}$	0,588
4.	1, 3, 4	$32,6\overline{6}$	0,004
5.	1, 3, 5	$33,6\overline{6}$	1,138
6.	1, 4, 5	$31,1\overline{6}$	2,054
7.	2, 3, 4	$33,1\overline{6}$	0,321
8.	2, 3, 5	$34,1\overline{6}$	2,454
9.	2, 4, 5	$31,6\overline{6}$	0,871
10.	3, 4, 5	33,50	0,810
	Total	326,0	11,899

En fonction des deux résultats obtenus :

$$\sigma = 2,6721 \text{ et } \sigma_{\bar{x}} = 1,0908,$$

nous pouvons vérifier l'égalité suivante :

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}}$$

puisque $\frac{\sigma}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}} = \frac{2,6721}{\sqrt{3}} \cdot \sqrt{\frac{5-3}{5-1}} = 1,0908.$

Ces résultats nous permettent de conclure qu'en connaissant les paramètres μ et σ de la population, nous sommes en mesure d'évaluer les caractéristiques correspondantes de la distribution d'échantillonnage des moyennes. Inversement, nous sommes également en mesure de déterminer les paramètres de la population en connaissant les caractéristiques de la distribution d'échantillonnage.

MOTS CLÉS

Erreur type *(Standard error)*
Estimation *(Estimation)*
Intervalle de confiance *(Confidence interval)*
Test d'hypothèse *(Hypothesis testing)*

DISTRIBUTION DE FRÉQUENCES

Frequency distribution

La répartition de la **population** peut être exprimée en terme relatif, par exemple en **pourcentage**, en fraction ou en terme absolu. Le résultat est appelé la distribution de fréquences.

Il s'agit également du résultat du regroupement d'observations d'une **variable** étudiée, soit en fonction des valeurs particulières (dans le cas d'une variable discrète), soit à l'intérieur d'intervalles de classe (dans le cas d'une variable continue).

Le nombre d'apparitions de chaque valeur, de même que le nombre d'observations à l'intérieur de chaque classe, est appelé fréquence.

La distribution de fréquences ainsi obtenue peut être représentée soit sous forme de **tableau de fréquences**, soit sous forme graphique, notamment par un **histogramme**, un **polygone de fréquences**, un **diagramme en bâtons** ou une courbe de fréquence.

Il est possible d'associer à une distribution de fréquences une distribution de fréquences relatives qui correspond à la distribution de chaque valeur ou de chaque classe par rapport au nombre d'observations.

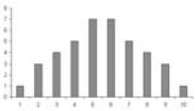
HISTORIQUE

Voir **représentation graphique**.

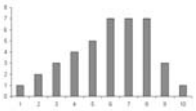
DOMAINES ET LIMITATIONS

Les distributions de fréquences que l'on rencontre en pratique peuvent varier considérablement :

- certaines sont symétriques, d'autres sont asymétriques

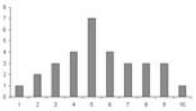


Distribution symétrique

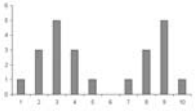


Distribution asymétrique

- certaines n'ont qu'un seul **mode** (distributions unimodales), d'autres en ont plusieurs (distributions plurimodales).



Distribution unimodale



Distribution bimodale

Nous pouvons en fait les classer en quatre types principaux :

1. distribution symétrique avec un seul mode,
2. distribution asymétrique avec un seul mode,
3. distribution en forme de J,
4. distribution en forme de U.

Il apparaît fréquemment dans une étude statistique que nous devions comparer deux distributions. Si les deux distributions sont du même type, il est possible de les comparer en examinant leurs caractéristiques principales, telles que : la **mesure de tendance centrale**, la **mesure de dispersion** ou la **mesure de forme**.

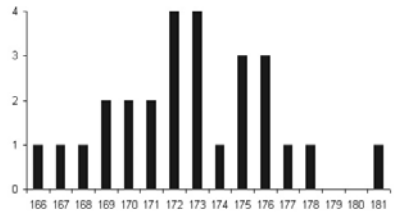
EXEMPLE

Les données suivantes représentent la taille, mesurée en centimètres, observée sur une **population** de vingt-sept élèves d'une classe du degré secondaire :

169	177	178	181	173
172	175	171	175	172
173	174	176	170	172
173	172	171	176	176
175	168	167	166	170
173	169			

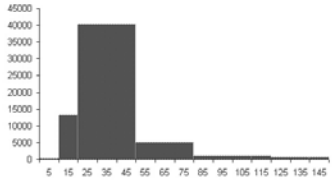
En ordonnant et en regroupant ces observations par **valeur**, nous obtenons la distribution de fréquences suivante :

Taille	Fréquence	Taille	Fréquence
166	1	174	1
167	1	175	3
168	1	176	3
169	2	177	1
170	2	178	1
171	2	179	0
172	4	180	0
173	4	181	1



Voici un autre exemple qui concerne la distribution de fréquences sur une variable continue : après recueil et distribution des revenus de la population du canton de Neuchâtel (CH), période 1975-1976, nous obtenons le **tableau de fréquences** suivant :

Classes de revenu net (par CHF 1000,-)	Fréquences	Fréquences relatives
$0 - 10$	283	$0,0047$
$10 - 20$	13 175	$0,2176$
$20 - 50$	40 316	$0,6661$
$50 - 80$	5 055	$0,0835$
$80 - 120$	1 029	$0,0170$
120 et plus	670	$0,0111$



MOTS CLÉS

Courbe de fréquences *(Frequency curve)*
Diagramme en bâtons *(Line chart)*
Fréquence *(Frequency)*
Histogramme *(Histogram)*
Intervalle *(Interval)*
Polygone de fréquences *(Frequency polygon)*
Tableau de fréquences *(Frequency table)*

DISTRIBUTION PLURIMODALE

Multimodal distribution

Voir **distribution de fréquences**.

DISTRIBUTION UNIMODALE

Unimodal distribution

Voir **distribution de fréquences**.

DONNÉE

Data

Une donnée est le résultat d'une **observation** faite sur une **population** ou sur un **échantillon**.

Le mot « donnée », du latin *datum*, est défini comme étant un fait (pas forcément numérique) à partir duquel on peut tirer une conclusion.

Il ne faut pas confondre un nombre (ou tout autre forme de description) qui ne contient pas forcément une information avec une donnée qui, elle, contient de l'information.

Les données sont liées à la **variable** étudiée. On dit ainsi, que les données sont quantitatives, qualitatives, discrètes ou continues, si la variable associée est elle-même quantitative, qualitative, discrète ou continue.

MOTS CLÉS

Donnée binaire *(Binary data)*
Données catégoriques *(Categorical data)*
Données incomplètes *(Incomplete data)*
Données spatiales *(Spatial data)*
Données standardisées *(Standardised data)*
Échantillon *(Sample)*
Observation *(Observation)*
Population *(Population)*
Valeur *(Value)*
Variable *(Variable)*

RÉFÉRENCE

Federer, W.T. (1982). Data Collection. In Kotz, S. and Johnson, N.L. (editors). *Encyclopedia of Statistical Sciences*. Vol. 2. John Wiley & Sons, New York.

DONNÉE BINAIRE

Binary data

On parle de donnée binaire quand la **variable** étudiée ne peut prendre que deux valeurs. On représente généralement ces deux valeurs par 0 et 1 même si la variable n'est pas quantitative.

Le sexe, la présence ou l'absence d'un caractère, le succès ou l'échec d'une expérience, etc., sont des variables impliquant des données binaires. Ces variables sont dites variables dichotomiques.

EXEMPLE

Un météorologue désire savoir dans quelle mesure ses prévisions sont fondées ; pour cela, il s'intéresse à la **variable aléatoire** représentant la prévision. Cette **variable** ne peut prendre que deux valeurs :

$$X = \begin{cases} 0 & \text{si la prévision s'est avérée incorrecte.} \\ 1 & \text{si la prévision était correcte.} \end{cases}$$

Les prévisions que ce météorologue a fait durant 50 jours consécutifs sont les suivantes : 32 fois la

prévision a été correcte, alors que 18 fois, elle s'est avérée incorrecte.

Pour savoir si ses prévisions sont meilleures que celles qu'il aurait pu obtenir en lançant une pièce de monnaie et en prédisant le temps selon que la pièce est tombée sur pile ou face, il décide de tester l'**hypothèse nulle** H_0 que la proportion p des prédictions correctes vaut $0,5$ contre l'**hypothèse alternative** H_1 que cette proportion est différente de $0,5$:

$$\begin{array}{rcl} H_0 & : & p = 0,5 \\ H_1 & : & p \neq 0,5. \end{array}$$

Il calcule la valeur de

$$\chi^2 = \sum \frac{(\text{fréquence observée - fréquence théorique})^2}{\text{fréquence théorique}}$$

avec les fréquences observées 18 et 32 et les fréquences théoriques 25 et 25 respectivement :

$$\begin{aligned} \chi^2 &= \frac{(32-25)^2 + (18-25)^2}{25} \\ &= \frac{49+49}{25} = \frac{98}{25} = 3,92. \end{aligned}$$

En supposant que le **théorème central limite** s'applique, il compare cette valeur à celle de la **loi du chi-carré** avec 1 **degré de liberté**. À un **seuil de signification** de 5 %, il trouve dans la **table du chi-carré** :

$$\chi^2_{1,0.95} = 3,84.$$

Ainsi, comme $3,92 > 3,84$, il rejette l'**hypothèse nulle**, par conséquent, le météorologue prédit mieux le temps qu'une pièce de monnaie.

MOTS CLÉS

Donnée *(Data)*
Données catégoriques *(Categorical data)*
Loi de Bernoulli *(Bernoulli distribution)*
Regréssion logistique *(Logistic regression)*
Table de contingence *(Contingency table)*
Test du rapport de vraisemblance *(Likelihood ratio test)*
Variable *(Variable)*
Variable dichotomique *(Dichotomous variable)*

RÉFÉRENCES

Cox, D.R. (1970). *The analysis of binary data.* Methuen, London.

Bishop, Y.M.M., Fienberg, S.E., and Holland, P.W. (1975). *Discrete Multivariate Analysis : Theory and Practice.* MIT Press, Cambridge, Mass.

DONNÉES CATÉGORIQUES

Categorical data

Les données catégoriques sont des données qui peuvent être classées en catégories, c'est-à-dire en classes de données vérifiant toutes une même propriété.

On peut distinguer plusieurs types de données catégoriques :

- binaires, caractérisant la présence ou l'absence d'une proprieté ;
- multi-catégorielles non-ordonnées (dites aussi « nominales ») ;
- multi-catégorielles ordonnées (dites aussi « ordinales ») ;
- variables à nombre entier.

On représente des données catégoriques sous forme de table de contingence.

DOMAINES ET LIMITATIONS

Il existe des variables essentiellement continues, mais qui peuvent se présenter sous forme de variables catégoriques. Un exemple est l'« âge », qui est une variable continue, mais les données relatives auquel peuvent être groupées en classes d'âge et donc devenir catégoriques.

EXEMPLE

Lors d'une votation, les citoyens concernés doivent approuver ou désapprouver un projet de loi. Les réponses peuvent donc être « oui » ou « non » et donnent lieu à des données catégoriques binaires. Sous forme de table de contingence, on pourrait avoir par exemple :

	oui	non
choix	8 546	5 455

La répartition par profession des employés d'une entreprise (où sont representés au moins 3 professions) est un exemple de données multi-catégorielles non ordonnées : les métiers de secrétaire, archiviste et mécanicien ne peuvent être ordonnés.

En revanche, si on s'intéresse au niveau d'éducation des employés, il y aura probablement un ordre naturel dans les catégories, par exemple : « primaire », « secondaire », « gymnase » et « universitaire », et on traitera des données multi-catégorielles ordonnées.

La taille des familles des employés donne lieu à des données catégoriques à nombres entiers.

MOTS CLÉS

Analyse de données catégoriques *(Categorical data analysis)*
Catégorie *(Category)*
Donnée *(Data)*
Donnée binaire *(Binary data)*
Variable dichotomique *(Dichotomic variable)*
Variable qualitative catégorielle *(Categorical qualitative variable)*
Variable aléatoire *(Random variable)*

RÉFÉRENCES

Voir **analyse de données catégoriques.**

DONNÉES INCOMPLÈTES

Incomplete data

Voir **observation manquante**.

DONNÉES SPATIALES

Spatial data

Les données spatiales sont les attributs des phénomènes se référant à une localisation spatiale, comme par exemple la **classification** de l'utilisation des sols, la concentration en polluants à différents endroits, les nuages, etc., en bref toute information qui peut être représentée sous forme de carte. En termes simples, les données spatiales peuvent être vues comme des données situées ou référencées à la surface terrestre. Dans la famille des données spatiales, on trouve les données ayant une collection dénombrable de sites spatiaux, par exemple la distribution de la mortalité infantile dans différentes villes et pays, et des données spatiales à configuration ponctuelle (c'est-à-dire des données pour lesquelles la localisation et la magnitude sont aléatoires, par exemple la sedimentation des minérais dans l'espace).

On peut aussi considérer qu'une donnée spatiale est la réalisation d'une variable régionalisée. La *théorie de la variable régionalisée* suppose que toute mesure peut être modélisée (visualisée) comme étant la réalisation d'une fonction aléatoire (ou d'un processus aléatoire). Par exemple, les échantillons pris sur le terrain peuvent être perçus comme des réalisations d'une variable régionalisée (voir sous le mot **géostatistique**).

HISTORIQUE

Avec le développement des systèmes d'information du territoire (SIT) ou géographiques (SIG), la notion de donnée spatiale est devenue très repandue dans la statistique.

DOMAINES ET LIMITATIONS

Le système d'information géographique est une collection de programmes informatiques qui facilitent, du fait de la géoréférence, l'intégration des données spatiales, non spatiales, qualitatives et quantitatives dans des bases de données qui peuvent être organisées sous un unique systême. Les données spatiales proviennent des domaines très vastes et souvent nécessitent une sélection et une analyse selon des critères spécifiques à chaque type de recherche. Les satellites récoltent des quantités énormes de données dont seulement une infime fraction est analysée. L'incovénient de cette richesse des données doit être contrôlé par la sélection. La sélection des données peut être faite selon le type des problèmes à résoudre et les modèles utilisés pour faire l'analyse.

EXEMPLES

De manière générale, les données spatiales peuvent être traitées comme des réalisations de variables aléatoires comme c'est usuel dans toute analyse statistique. Ci-dessous on présente comme exemple des données spatiales recueillies dans le sol du Jura suisse. Il s'agit d'un jeu de 259 données sur la pollution avec des métaux lourds sur une région de $14,5$ km^2, dont dix sont présentés ci-dessous :

X	Y	Sol	Roche	Cd	Co
2,386	3,077	3	3	1,740	9,32
2,544	1,972	2	2	1,335	10,00
2,807	3,347	2	3	1,610	10,60
4,308	1,933	3	2	2,150	11,92
4,383	1,081	3	5	1,565	16,32
3,244	4,519	3	5	1,145	3,50
3,925	3,785	3	5	0,894	15,08
2,116	3,498	3	1	0,525	4,20
1,842	0,989	3	1	0,240	4,52

Cr	Cu	Ni	Pb	Zn
38,32	25,72	21,32	77,36	92,56
40,20	24,76	29,72	77,88	73,56
47,00	8,88	21,40	30,80	64,80
43,52	22,70	29,72	56,40	90,00
38,52	34,32	26,20	66,40	88,40
40,40	31,28	22,04	72,40	75,20
30,52	27,44	21,76	60,00	72,40
25,40	66,12	9,72	141,00	72,08
27,96	22,32	11,32	52,40	56,40

Les données « sol », qui se réfèrent à l'utilisation du sol et les données « roche » qui décrivent le type de roche sous-jacente à l'endroit échantillonné sont des données catégoriques (**variable qualitative catégorielle**). Les différents types de roche sont :

1 : Argovien, 2 : Kimmeridgien, 3 : Séquanien, 4 : Portlandien, 5 : Quatérnaire.

L'utilisation du sol correspond à :

1 : forêt, 2 : pâturage, 3 : prairie, 4 : champ labouré.

MOTS CLÉS

Classification *(Classification)*
Échantillonnage *(Sampling)*
Géostatistique *(Geostatistics)*
Histogramme *(Histogram)*
Statistique spatiale *(Spatial Statistics)*

RÉFÉRENCES

Atteia, O., Dubois, J.-P. and Webster, R. (1994). Geostatistical analysis of soil contamination in the Swiss Jura. *Environmental Pollution*, 86: 315-327.

Goovaerts, P. (1997). *Geostatistics for Natural Resources Evaluation.* Oxford : Oxford University Press.

Matheron, G. (1965). *Les variables régionalisées et leur estimation.* Masson. Paris.

DONNÉES STANDARDISÉES

Standardised data

Une donnée standardisée est une donnée à laquelle on a soustrait la moyenne des observations liées à la **variable aléatoire** (associée aux données) puisque l'on a divisé par l'écart type des observations. Ainsi les données standardisées ont une moyenne de 0 et une variance égale à 1.

ASPECTS MATHÉMATIQUES

La standardisation d'une série de donnée X avec les termes x_i, $i = 1, \dots n$, revient à remplacer le terme x_i de la série par x_i^s :

$$x_i^s = \frac{x_i - \overline{x}_i}{s}$$

où $\overline{x}_i$ et s sont respectivement la moyenne et l'écart type des données x_i.

DOMAINES ET LIMITATIONS

Certains statisticiens ont l'habitude de standardiser les variables explicatives dans un modèle de régression linéaire de manière à simplifier les difficultés numériques survenant lors des calculs matriciels et à faciliter l'interprétation et la comparaison

des coefficients de régression. L'application de la méthode des moindres carrés en utilisant des données standardisées fournit des résultats équivalents à ceux obtenus en utilisant les données originales.

Une régression est généralement modélisée par

$$Y_i = \sum_{j=1}^{p} \beta_j X_{ij} \quad i = 1, \ldots, n$$

où les X_j sont des données non standardisées. Si les X_{ij} sont standardisés (de moyenne 0 et de variance 1), les β_j sont appelés des coefficients de régression standardisées (*standardised regression coefficients*) et sont équivalents à la corrélation pour autant que Y soit aussi standardisé. La même terminologie peut être utilisée si les X_j sont réduits à avoir une variance égale à 1, mais pas nécessairement une moyenne nulle. Les coefficients β_0 absorbent la différence.

EXEMPLE

Dans l'exemple suivant, les observations sont relatives à treize fournées de ciment. Chaque fournée est composée de quatres ingrédients cités dans le tableau. Le but de l'expérience est de déterminer comment les quantités X_1, X_2, X_3 et X_4 de ces 4 ingrédients affectent la quantité Y de chaleur émise par le durcissement du ciment.

Chaleur émise par le ciment

Fournée	Ingrédient				Chaleur
i	X_1	X_2	X_3	X_4	Y
1	7	26	6	60	78.5
2	1	29	15	52	74.3
3	11	56	8	20	104.3
4	11	31	8	47	87.6
5	7	52	6	33	95.9
6	11	55	9	22	109.2
7	3	71	17	6	102.7
8	1	31	22	44	72.5
9	2	54	18	22	93.1
10	21	47	4	26	115.9
11	1	40	23	34	83.9
12	11	66	9	12	113.3
13	10	68	8	12	109.4

Source : Birkes & Dodge (1993)

x_i quantité de chaleur émise par le durcissement de la i-ème fournée (en joules) ;
x_{i1} quantité de l'ingrédient 1 (aluminate de tricalcium) dans la i-ème fournée ;
x_{i2} quantité de l'ingrédient 2 (silicate de tricalcium) dans la i-ème fournée ;
x_{i3} quantité de l'ingrédient 3 (alumino-ferrite de tetracalcium) dans la i-ème fournée ;
x_{i4} quantité de l'ingrédient 4 (silicate de dicalcium) dans la i-ème fournée.

Le modèle obtenu de la régression linéaire simple est :

$$\widehat{Y} = 62,4 + 1,55X_1 + 0,51X_2 + 0,102X_3 - 0,144X_4 + \varepsilon.$$

Le tableau ci-dessous présente les données sur le ciment avec les données relatives aux variables explicatives standardisées notées X_1^s, X_2^s, X_3^s et X_4^s.

Données relatives au ciment avec des variables explicatives standardisées

Fournée	Ingrédient standardisé				Chaleur
i	X_1^s	X_2^s	X_3^s	X_4^s	Y
1	$-0,079$	$-1,424$	$-0,901$	$1,792$	$78,5$
2	$-1,099$	$-1,231$	$0,504$	$1,314$	$74,3$
3	$0,602$	$0,504$	$-0,589$	$-0,597$	$104,3$
4	$0,602$	$-1,102$	$-0,589$	$1,016$	$87,6$
5	$-0,079$	$0,247$	$-0,901$	$0,179$	$95,9$
6	$0,602$	$0,440$	$-0,432$	$-0,478$	$109,2$
7	$-0,759$	$1,468$	$0,817$	$-1,434$	$102,7$
8	$-1,099$	$-1,102$	$1,597$	$0,836$	$72,5$
9	$-0,929$	$0,376$	$0,973$	$-0,478$	$93,1$
10	$2,302$	$-0,074$	$-1,213$	$-0,239$	$115,9$
11	$-1,099$	$-0,524$	$1,753$	$0,239$	$83,9$
12	$0,602$	$1,147$	$-0,432$	$-1,075$	$113,3$
13	$0,432$	$1,275$	$-0,589$	$-1,075$	$109,4$

L'estimation par la méthode des moindres carrés du modèle complet en utilisant ces variables explicatives standardisées donne :

$$\widehat{Y} = 95,4 + 9,12X_1^s + 7,94X_2^s + 0,65X_3^s - 2,41X_4^s.$$

On remarque par exemple que le coefficient $\widehat{\beta}_1^s = 9,12$ de X_1^s peut se calculer en multipliant le coefficient $\widehat{\beta}_1 = 1,55$ de X_1 dans le modèle initial par l'écart type $s_1 = 5,882$ des X_1. On remarque que la moyenne $\overline{X}_1 = 7,46$ des X_1 n'entre pas en considération

dans la relation entre $\widehat{\beta}_1^s$ et $\widehat{\beta}_1$ mais dans la relation entre $\widehat{\beta}_0^s$ et $\widehat{\beta}_0$. En ne retenant que les deux premières variables explicatives standardisées, on obtient par ailleurs la solution

$$\widehat{Y} = 95,4 + 8,64X_1^s + 10,3X_2^s.$$

MOTS CLÉS

Donnée *(Data)*
Regression ridge *(Ridge regression)*
Variables explicatives *(Predictor variable)*

RÉFÉRENCE

Marriott, F.H.C. (1990). *A dictionnary of statistical terms.* 5th edition. International Statistical Institute, Longman Scientific & Technical.

DURBIN-WATSON, TEST DE

Durbin-Watson test

Voir **test de Durbin-Watson**.



E

ÉCART GÉOMÉTRIQUE

Geometric standard deviation

L'écart géométrique d'un ensemble d'observations quantitatives est une **mesure de dispersion**. Il correspond à l'écart des observations autour de la **moyenne géométrique**.

HISTORIQUE

Voir **estimation L1**.

ASPECTS MATHÉMATIQUES

Soit $x_1, x_2, \ldots, x_n$ un ensemble de n observations relatives à une **variable quantitative** X. L'écart géométrique, noté $E_{géom}$, se calcule comme suit :

$$\log E_{géom} = \frac{1}{n} \sum_{i=1}^{n} (\log x_i - \log G),$$

où $G = \sqrt[n]{x_1 \cdot x_2 \cdot \ldots \cdot x_n}$ est la moyenne géométrique des observations.

MOTS CLÉS

Écart type *(Standard deviation)*
Estimation L1 *(L1 estimation)*
Mesure de dispersion *(Measure of dispersion)*
Variance *(Variance)*

ÉCART MÉDIAN

Median absolute deviation

L'écart médian d'un ensemble d'observations quantitatives est une **mesure de dispersion**. Il correspond à la **moyenne** des valeurs absolues des écarts de chaque observation par rapport à la **médiane**.

HISTORIQUE

Voir **estimation L1**.

ASPECTS MATHÉMATIQUES

Soit $x_1, x_2, \ldots, x_n$ un ensemble de n observations relatives à une **variable quantitative** X. L'écart médian, noté $E_{méd}$, se calcule comme suit :

$$E_{méd} = \frac{\sum_{i=1}^{n} |x_i - Md|}{n},$$

où Md est la médiane des observations.

DOMAINES ET LIMITATIONS

L'écart médian est une notion fondamentale de l'**estimation L1**. En effet, l'estimation L1 d'un **paramètre** de **mesure de tendance centrale** θ est une méthode d'**estimation** basée sur la recherche du minimum des déviations absolues des observations x_i par rapport à ce paramètre θ :

$$\sum_{i=1}^{n} |x_i - \theta|.$$

La valeur qui minimise cette expression est la **médiane** de l'**échantillon**. Dans le cas où il y a des observations aberrantes, cette méthode est plus efficace que la méthode des **moindres carrés**.

MOTS CLÉS

Écart type *(Standard deviation)*
Estimation L1 *(L1 estimation)*
Mesure de dispersion *(Measure of dispersion)*
Variance *(Variance)*

ÉCART MOYEN

Mean absolute deviation

L'écart moyen d'un ensemble d'observations quantitatives est une **mesure de dispersion**. Il correspond à la **moyenne** des valeurs absolues des écarts de chaque observation par rapport à la moyenne arithmétique.

HISTORIQUE

Voir **estimation L1**.

ASPECTS MATHÉMATIQUES

Soit $x_1, x_2, \ldots, x_n$ un ensemble de n observations relatives à une **variable quantitative** X. L'écart

moyen, noté EM, se calcule comme suit :

$$EM = \frac{\sum_{i=1}^{n} |x_i - \bar{x}|}{n},$$

où $\bar{x}$ est la **moyenne arithmétique** des observations.

Lorsque les observations sont ordonnées sous forme d'une **distribution de fréquences**, le calcul de l'écart moyen se fait de la façon suivante :

$$EM = \frac{\sum_{i=1}^{k} f_i \cdot |x_i - \bar{x}|}{\sum_{i=1}^{k} f_i},$$

où

- x_i représente les différentes valeurs de la **variable**,
- f_i les fréquences associées à ces valeurs et
- k le nombre de valeurs différentes.

DOMAINES ET LIMITATIONS

L'écart moyen est une mesure de dispersion plus robuste que l'**écart type**.

EXEMPLE

Cinq étudiants ont passé successivement deux examens auxquels ils ont obtenu les notes suivantes :

Examen 1 :

$3,5$	4	$4,5$	$3,5$	$4,5$	$\bar{x} = \frac{20}{5} = 4.$

Examen 2 :

$2,5$	$5,5$	$3,5$	$4,5$	4	$\bar{x} = \frac{20}{5} = 4.$

La **moyenne arithmétique** $\bar{x}$ des notes est identique pour les deux examens. Toutefois, la variabilité des notes (ou la dispersion) n'est pas la même :

Pour l'examen 1, l'écart moyen est égal à :

$$\begin{aligned} \text{écart moyen} &= \frac{|3,5-4|}{5} + \frac{|4-4|}{5} \\ &\quad + \frac{|4,5-4|}{5} + \frac{|3,5-4|}{5} \\ &\quad + \frac{|4,5-4|}{5} \\ &= \frac{2}{5} = 0,4. \end{aligned}$$

Pour l'examen 2, l'écart moyen est beaucoup plus élevé :

$$\begin{aligned} \text{écart moyen} &= \frac{|2,5-4|}{5} + \frac{|5,5-4|}{5} \\ &\quad + \frac{|3,5-4|}{5} + \frac{|4,5-4|}{5} \\ &\quad + \frac{|4-4|}{5} \\ &= \frac{4}{5} = 0,8. \end{aligned}$$

Les notes du deuxième examen sont donc plus dispersées autour de la moyenne arithmétique que les notes du premier examen.

MOTS CLÉS

Écart type *(Standard deviation)*
Estimation L1 *(L1 estimation)*
Mesure de dispersion *(Measure of dispersion)*
Variance *(Variance)*

ÉCART TYPE

Standard deviation

L'écart type est une **mesure de dispersion**. Il correspond à la racine carrée de la **variance**. Il exprime la même caractéristique que la variance, mais en tenant compte de l'unité de mesure.

Il est généralement dénoté par σ lorsqu'il est relatif à une **population** et par S lorsqu'il est relatif à un **échantillon**.

En pratique, l'écart type σ de la population sera estimé par l'écart type S d'un échantillon issu de cette population.

HISTORIQUE

Le terme « écart type » ou *standard deviation* est étroitement lié aux travaux de deux mathématiciens anglais **Karl Pearson** et **William Sealy Gosset**.

C'est en effet au cours d'une conférence qu'il donna devant la Royal Society de Londres en 1893, que K. Pearson l'utilisa pour la première fois. Il le reprit dans son article intitulé « On the Dissection of Asymmetrical Frequency Curves » en 1894 et on lui doit également l'introduction du symbole σ pour désigner l'écart type. W.S. Gosset, dit Student, se consacra également à ces problèmes et formalisa les travaux dans ce domaine. Il s'attacha notamment à expliquer pourquoi il importe de distinguer S (écart type relatif à un **échantillon**) de σ (écart type relatif à la **population**).

Dans son article du mois de mars 1908, W.S. Gosset définit l'écart type d'un échantillon par :

$$S = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}}.$$

Une question qui survint était de savoir si l'expression ci-dessus devait être divisée par n ou par $n-1$. Pearson demanda l'opinion de Gosset et ce dernier lui répondit dans une lettre datée du 13 mars 1927 que les deux formules se trouvaient dans la littérature. Celle dont le dénominateur est $n-1$ donne une **valeur** moyenne indépendante de la taille de l'échantillon et donc de la population. En outre, avec des échantillons de grande taille, la différence entre les deux formules est négligeable alors que par la suite les calculs sont simplifiés si l'on utilise la formule où n apparaît au dénominateur.

En résumé, W.S. Gosset indique que l'utilisation de $n-1$ est probablement mieux appropriée avec de petits échantillons, tandis que l'utilisation de n est préférable pour de grands échantillons.

Il est à noter que la découverte de l'écart type est à mettre en relation avec le développement de la théorie de l'**estimation** et des **tests d'hypothèses**.

Par ailleurs, l'étude de la variabilité fut étroitement liée aux travaux des astronomes, dans la mesure où ils étaient intéressés par les découvertes relatives à la distribution des erreurs.

ASPECTS MATHÉMATIQUES

Soit une **variable aléatoire** X d'**espérance mathématique** $E[X]$. La **variance** σ^2 de la **population** est définie par :

$$\sigma^2 = E\left[(X - E[X])^2\right]$$

ou

$$\sigma^2 = E\left[X^2\right] - E[X]^2.$$

L'écart type σ est la racine carrée positive de σ^2 :

$$\sigma = \sqrt{E\left[(X - E[X])^2\right]}.$$

Pour estimer l'écart type σ d'une population, il faut connaître sa distribution, mais, en pratique, il est commode d'estimer σ à partir d'un **échantillon** aléatoire de taille n : $x_1, \ldots, x_n$ relatif à la variable X ; on calcule alors l'estimateur S de σ avec la définition suivante :

$$S = \sqrt{\frac{\sum_{i=1}^{n} (x_1 - \bar{x})^2}{n}},$$

où $\bar{x}$ est la **moyenne arithmétique** de l'échantillon. Mais cet estimateur est biaisé, à savoir :

$$E\left(S^2\right) \neq \sigma.$$

Pour obtenir un estimateur non biaisé, on utilise la formule :

$$S = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}}.$$

On peut également calculer l'écart type par la formule équivalente suivante :

$$S = \sqrt{\frac{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}{n-1}} = \sqrt{\frac{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}{n(n-1)}}.$$

DOMAINES ET LIMITATIONS

L'écart type est utilisé comme échelle de mesure dans les tests d'hypothèse et les intervalles de confiance. D'autres **mesures de dispersion** telles que l'**empan** ou l'**écart moyen** sont également utilisées en **statistique descriptive** mais jouent un rôle moins important dans les tests d'hypothèse. On peut vérifier que :

$$\text{écart moyen} \leq S \leq \frac{\text{empan}}{2} \cdot \sqrt{\frac{n}{n-1}}.$$

Si l'on considère une **variable aléatoire** suivant une **loi normale** de **moyenne** μ et d'écart type σ, nous pouvons vérifier que $68,26$ % des observations se situent dans l'**intervalle** $\mu \pm \sigma$. De même, $95,44$ % des observations se situent dans l'intervalle $\mu \pm 2\sigma$, et $99,74$ % des observations se situent dans l'intervalle $\mu \pm 3\sigma$.

EXEMPLE

Cinq étudiants ont passé successivement deux examens auxquels ils ont obtenu les notes suivantes :

Examen 1 :

$3,5$	4	$4,5$	$3,5$	$4,5$	$\bar{x} = \frac{20}{5} = 4.$

Examen 2 :

$2,5$	$5,5$	$3,5$	$4,5$	4	$\bar{x} = \frac{20}{5} = 4.$

La **moyenne arithmétique** $\bar{x}$ de ces deux ensembles d'observations est identique. Toutefois, la dispersion des observations autour de la moyenne n'est pas la même.

Pour calculer l'écart type, nous devons tout d'abord calculer les déviations de chaque observation par rapport à la moyenne arithmétique, puis élever ces déviations au carré :

Examen 1 :

Notes	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$
$3,5$	$-0,5$	$0,25$
4	$0,0$	$0,00$
$4,5$	$0,5$	$0,25$
$3,5$	$-0,5$	$0,25$
$4,5$	$0,5$	$0,25$
$\sum_{i=1}^{5} (x_i - \bar{x})^2$		$1,00$

Examen 2 :

Notes	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$
$2,5$	$-1,5$	$2,25$
$5,5$	$1,5$	$2,25$
$3,5$	$-0,5$	$0,25$
$4,5$	$0,5$	$0,25$
4	$0,0$	$0,00$
$\sum_{i=1}^{5} (x_i - \bar{x})^2$		$5,00$

L'écart type pour chaque examen est alors égal à :

Examen 1 :

$$S = \sqrt{\frac{\sum_{i=1}^{5} (x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{1}{4}} = 0,5.$$

Examen 2 :

$$S = \sqrt{\frac{\sum_{i=1}^{5} (x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{5}{4}} = 1,118.$$

L'écart type des notes du deuxième examen étant plus grand, cela signifie qu'au deuxième examen les notes sont plus dispersées autour de la moyenne arithmétique qu'au premier. La variabilité du deuxième examen est donc plus grande que celle du premier.

MOTS CLÉS

Coefficient de variation *(Coefficient of variation)*
Écart moyen *(Mean deviation)*
Mesure de dispersion *(Measure of dispersion)*
Variance *(Variance)*

RÉFÉRENCES

Barnard, G.A., Plackett, R.L. and Pearson, S. (1990). *'Student': a statistical biography of William Sealy Gosset: based on writings by E.S. Pearson.* Oxford : Oxford University Press.

Pearson, K. (1894). 1948. Contributions to the Mathematical Theory of Evolution. I. pp. 1-40 in Karl Pearson, *Karl Pearson's Early Statistical Papers.* Cambridge University Press, Cambridge. First published as: On the Dissection of Asymmetrical Frequency Curves, *Philosophical Transactions of the Royal Society of London.* Series A, Vol. 185, pp. 71-110.

Gosset, S.W. « **Student** » (1908). The Probable Error of a Mean. *Biometrika*, 6, pp. 1-25.

ÉCHANTILLON

Sample

Un échantillon est un sous-ensemble d'une **population** sur lequel on effectue une étude statistique.

Une étude sur un échantillon vise généralement à tirer des conclusions relatives à la population dont il est issu.

EXEMPLE

Une ménagère est en train de préparer une soupe. Pour déterminer si la soupe est assez salée, elle en goûte une cuillerée. En fonction de ce qu'elle aura observé sur l'échantillon représenté dans ce cas par la cuillerée, elle pourra décider s'il est nécessaire de rajouter du sel au potage, symbolisant la **population** étudiée.

MOTS CLÉS

Échantillonnage *(Sampling)*
Estimation *(Estimation)*
Inégalité de Tchebychev *(Chebyshev's inequality)*
Population *(Population)*
Taille d'échantillon *(Sample size)*

ÉCHANTILLONNAGE

Sampling

De façon générale, le terme échantillonnage signifie la sélection d'une partie d'une **population**, en l'occurrence d'un **échantillon**, l'étude de certaines caractéristiques de cet échantillon et le fait d'en tirer des inférences relatives à la population.

D'autre part, l'échantillonnage représente également l'ensemble des opérations destinées à former un échantillon à partir d'une population donnée.

HISTORIQUE

Le concept de la représentativité d'un **échantillon** est une notion très récente.

Si les premières tentatives d'**extrapolation** sont bien apparues au XIII^{e} siècle, notamment en France où le nombre de « feux » (famille, foyer, habitation) servait à estimer la **population**, le **recensement** resta jusqu'au XIX^{e} siècle plus fréquemment utilisé que l'échantillonnage.

Le principe d'échantillonnage avec ou sans remise est apparu pour la première fois dans l'ouvrage intitulé *De ratiociniis in Aleae Ludo* publié en 1657 par le scientifique hollandais Christiaan Huygens (1629-1695).

C'est en 1895, à Berne, lors du congrès de l'Institut international de la statistique, qu'Anders Nicolai Kiaer présenta un mémoire intitulé *Observations et expériences concernant des dénombrements représentatifs.* N. Kiaer, alors directeur du Bureau central de statistique du Royaume de Norvège, comparait dans son exposé la structure de l'échantillon à celle de la population obtenue par recensement. Il définissait ainsi le contrôle *a posteriori* et employait pour la première fois le terme « représentatif ».

N. Kiaer rencontra l'opposition presque unanime des participants au congrès, partisans convaincus de l'énumération complète.

En 1897 toutefois, N. Kiaer confirma la notion de représentativité et reçut l'appui de la Conférence des statisticiens scandinaves. Il eut par la suite l'occasion de défendre à nouveau son point de vue au cours de plusieurs conférences : Saint-Pétersbourg (1897), Stockholm (1899), Budapest (1901) et Berlin (1903).

Un nouveau pas fut franchi avec les travaux de Lucien March sur le rôle de l'aléatoire dans l'échantillonnage ; il fut en effet le premier à développer l'idée d'un échantillonnage probabiliste appelé aussi échantillonnage aléatoire.

D'autres statisticiens s'intéressèrent également au problème. Ladislaus Josephowitsch (von) Bortkiewicz, professeur à Berlin, suggéra le calcul des probabilités pour tester l'écart entre les répartitions de l'échantillon et de la population totale sur les variables clés. A.L. Bowley (1906) développa notamment l'échantillonnage aléatoire, la stratification. Selon Desrosières (1988), il s'intéressa également à la notion d'**intervalle de confiance** dont il présenta les premiers calculs en 1906 devant la Royal Statistical Society.

1925 marque une nouvelle étape dans l'histoire de l'échantillonnage. C'est en effet l'année du congrès de Rome de l'Institut international de statistique au cours duquel on distingua : l'échantillonnage aléatoire (choix au hasard, *random sampling*) et l'échantillonnage raisonné (choix judicieux, *purposive sampling*). Après cette date, le problème ne se posa plus en termes de choix entre l'échantillon et le dénombrement total, mais entre les diverses manières d'effectuer l'échantillonnage.

Des différents usages de l'échantillonnage, le plus fréquent fut aux États-Unis celui qu'en firent les instituts de sondage pour les enquêtes d'opinion publique et notamment les enquêtes pré-électorales. Le 3 novembre 1936, jour de la publication des résultats des élections présidentielles, marque à cet égard l'une des dates les plus importantes. Pour ce scrutin, le *Litterary Digest* avait prédit la victoire de Landon, alors que, sans se concerter, Crossley, Roper et Gallup avaient annoncé celle de son adversaire, en l'occurence Roosevelt. Ce fut Roosevelt qui fut élu. L'erreur du *Litterary Digest* avait été de procéder à un sondage téléphonique pour établir ses prévisions, or, à cette époque, seuls les gens aisés possédaient le téléphone : l'échantillon était donc biaisé !

1938 marque en France la naissance de l'IFOP (Institut français d'opinion publique) créé par Jean Stoetzel.

Il est intéressant de remonter un peu dans le temps pour souligner le rôle des statisticiens russes dans l'évolution des techniques de l'échantillonnage. En effet, dès le XIX^{e} siècle, elles étaient connues et utilisées dans leur pays. Selon Tassi (1988), A.I. Tchuprov (1842-1908) en fut l'un des précurseurs et, dès 1910, son fils A.A. Tchuprov utilisa l'échantillonnage aléatoire, fit référence au tirage par grappes, à l'**échantillonnage stratifié** avec et sans remise, et, sous l'influence d'Andrei Andreyevich Markov, encouragea l'utilisation des probabilités en **statistique**. Bien que l'école russe n'ait été connue que tardivement pour ses travaux d'échantillonnage, il semble donc que les statisticiens de cette école aient eu connaissance de ces techniques bien avant ceux d'Europe occidentale.

On remarquera qu'en 1924, A.G. Kovalevsky avait déjà traité rigoureusement le **sondage** par stratification et l'allocation optimale par strate, alors que ce résultat ne sera redécouvert que quelque dix ans plus tard par Jersey Neyman.

Enfin cet historique ne saurait omettre les contributions de **Jerzy Neyman** aux travaux sur l'échantillonnage. Considéré comme l'un des fondateurs de la théorie statistique des sondages, il prit position en faveur de l'échantillonnage aléatoire contre l'échantillonnage raisonné, et travailla entre autres sur l'échantillonnage avec ou sans remise, l'échantillonnage stratifié et l'**échantillonnage par grappes**. Il utilisait déjà les estimateurs par le quotient, la différence et la régression, des techniques qui seront développées par William Gemmell Cochran, Hansen, Adolf Hurwitz et Madow.

ASPECTS MATHÉMATIQUES

Il existe plusieurs méthodes d'échantillonnage. Elles peuvent être divisées en deux groupes : des méthodes non aléatoires et des méthodes aléatoires.

Les méthodes non aléatoires (ou par choix raisonné) consistent à construire, à partir d'informations déjà connues sur la **population**, un **échantillon** qui présente le maximum de traits caractéristiques de cette population. Il s'agit de procédés empiriques qui comportent une part d'arbitraire.

La procédure non aléatoire la plus utilisée est la méthode d'**échantillonnage par quotas**.

Les méthodes aléatoires (ou échantillonnages scientifiques, parfois également appelés échantillonnages statistiques) font appel à un mécanisme probabiliste pour former l'échantillon issu de la population ; un tel mécanisme permet de connaître à l'avance la **probabilité** (différente de zéro) qu'une unité donnée de la population appartienne à l'échantillon.

Les variables observées sur l'échantillon sont des variables aléatoires. À partir de celles-ci, on peut estimer les grandeurs correspondantes de la population et calculer l'erreur due à l'échantillonnage.

Les principales méthodes aléatoires sont :

— l'**échantillonnage aléatoire simple**,

— l'**échantillonnage stratifié**,

— l'**échantillonnage systématique**,

— l'**échantillonnage par grappes**.

DOMAINES ET LIMITATIONS

Le but de l'échantillonnage est de fournir suffisamment d'informations pour effectuer de l'**inférence** sur les caractéristiques d'une **population**. Il a donc pour objectif de sélectionner un **échantillon** reproduisant le plus fidèlement possible ces caractéristiques.

On tente de réduire au maximum les erreurs potentiellement dues à l'échantillonnage. Même si cette erreur échantillonnale est inévitable, elle est tolérée car elle est très largement compensée par les avantages présentés par les méthodes d'échantillonnage. Ces derniers se résument à :

- *Une réduction du coût*

  Au vu de tous les frais que peut encourir une récolte de données, il est évident que l'on peut réduire le coût en effectuant un échantillonnage.

- *Un gain de temps*

  La rapidité de prise de décisions n'est assurée que si on peut restreindre l'étude à un échantillon.

- *Une augmentation des possibilités*

  Dans certains domaines, une analyse complète sur la population est impossible, le choix se restreint alors à recourir à un échantillonnage ou à renoncer à l'étude.

- *Une précision des résultats obtenus par l'échantillonnage*

  Puisque de plus grands échantillons n'apporteraient pas de résultats significativement plus précis.

Seules les méthodes aléatoires d'échantillonnage permettent d'évaluer l'erreur échantillonnale, car le calcul de cette dernière est basé sur les probabilités.

MOTS CLÉS

Collecte des données *(Data collection)*
Échantillonnage aléatoire simple *(Simple random sampling)*
Échantillonnage par grappes *(Cluster sampling)*
Échantillonnage par quotas *(Quota sampling)*
Échantillonnage stratifié *(Stratified sampling)*
Échantillonnage systématique *(Systematic sampling)*
Estimateur *(Estimator)*
Estimation *(Estimation)*

RÉFÉRENCES

Bowley, A.L. (1906). Presidential address to the economic section of the British Association. *Journal of the Royal Statistical Society*, 69, pp. 540-558.

Delanius, T. (1985). *Elements of the survey sampling.* SAREC, Stockholm.

Desrosières, A. (1988). La partie pour le tout : comment généraliser ? La préhistoire de la con-

trainte de représentativité. *Estimation et sondages, cinq contributions à l'histoire de la statistique.* Economica, Paris.

Hedayat, A.S. and Sinha, B.K. (1991). *Design and Inference in Finite Population Sampling.* John Wiley & Sons, New York.

Tassi, Ph. (1988). De l'exhaustif au partiel : un peu d'histoire sur le développement des sondages. *Estimation et sondages, cinq contributions à l'histoire de la statistique.* Economica, Paris.

ÉCHANTILLONNAGE ALÉATOIRE SIMPLE

Simple random sampling

Un échantillonnage aléatoire simple est une méthode d'**échantillonnage** pour choisir n unités parmi les N de la population, de sorte que chacun des C_N^n échantillons possibles ait la même **probabilité** d'être sélectionné.

HISTORIQUE

Voir **échantillonnage**.

ASPECTS MATHÉMATIQUES

Pour obtenir un **échantillon** aléatoire simple de taille n, on commence par numéroter les individus de la **population** de 1 à N ; puis on tire n nombres compris entre 1 et N. Les tirages se font en principe sans remise.

On peut également utiliser des tables de nombres aléatoires pour obtenir un échantillon aléatoire simple.

Le but d'un échantillon aléatoire simple est de fournir une **estimation** sans **biais** de la **moyenne** et de la **variance** de la population. En effet, si on note $y_1, y_2, \ldots, y_N$ les caractéristiques d'une population et $y_1, \ldots, y_n$ les valeurs correspondantes figurant dans l'échantillon aléatoire simple, on a les résultats suivants :

	Population	Échantillon
Total	$T = \sum_{i=1}^{N} y_i$	$t = \sum_{i=1}^{n} y_i$
Moyenne	$\bar{Y} = \frac{T}{N}$	$\bar{y} = \frac{t}{n}$
Variance	$\sigma^2 = \frac{\sum_{i=1}^{N} \left(y_i - \bar{Y}\right)^2}{N}$	$s^2 = \frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{n-1}$

DOMAINES ET LIMITATIONS

L'échantillonnage aléatoire simple peut s'effectuer avec ou sans remise.

Si on effectue un tirage avec remise, on se trouve toujours devant la même **population**. L'**échantillon** est alors caractérisé par une suite de variables aléatoires indépendantes et identiquement distribuées. Si la population est suffisamment grande et si la taille de l'échantillon est relativement petite par rapport à celle de la population, on peut considérer qu'il y a **indépendance** des variables aléatoires même si les tirages sont effectués sans remise.

On est alors en mesure d'estimer certaines caractéristiques de la population en les déterminant à partir de cet échantillon.

L'échantillonnage aléatoire simple est à la base de l'ensemble de la théorie d'**échantillonnage**.

EXEMPLE

On considère une **population** de 658 individus. Pour former un **échantillon** aléatoire simple, on attribue un numéro à trois chiffres à tous les individus : 001, 002, ..., 658. Puis à l'aide d'une table de nombres aléatoires, on prélève un échantillon de 10 individus. On choisit aléatoirement un premier chiffre et un sens de lecture. Puis on lit les trois premiers chiffres jusqu'à ce qu'on obtienne 10 nombres distincts compris entre 001 et 658 pour définir notre échantillon.

MOTS CLÉS

Distribution d'échantillonnage *(Sampling distribution)*
Échantillonnage *(Sampling)*

Échantillonnage par grappes *(Cluster sampling)*
Échantillonnage stratifié *(Stratified sampling)*
Échantillonnage systématique *(Systematic sampling)*
Estimateur *(Estimator)*
Estimation *(Estimation)*

RÉFÉRENCE

Cochran, W.G. (1963). *Sampling Techniques*, 2nd ed. John Wiley & Sons, New York.

ÉCHANTILLONNAGE PAR GRAPPES

Cluster sampling

Pour effectuer un **échantillonnage** par grappes, il s'agit dans un premier temps de diviser la **population** en sous-ensembles appelés grappes.

Chacune de ces grappes est constituée d'individus supposés être représentatifs de la population.

L'échantillonnage par grappes consiste alors à choisir un **échantillon** aléatoire de grappes puis à observer, pour chacune de ces grappes, tous les individus y appartenant.

HISTORIQUE

Voir **échantillonnage**.

ASPECTS MATHÉMATIQUES

L'échantillonnage par grappes consiste à tirer au hasard des ensembles d'unités de la population, ou grappes, et ensuite à mener l'enquête sur toutes les unités de ces grappes. Les grappes sont souvent constituées par des unités de type géographique comme les quartiers d'une ville. La méthode consiste à diviser une ville en quartiers, puis à sélectionner les quartiers qui feront partie de l'échantillon. On mènera ensuite l'enquête sur toutes les personnes ou ménages, habitant dans les quartiers choisis.

Il y a deux raisons principales de procéder à un échantillonnage par grappes.

Dans beaucoup d'enquêtes, il se trouve qu'il n'existe pas une liste complète et fiable des unités de la population pour baser l'échantillonnage, et qu'il est excessivement coûteux de construire une telle liste. Par exemple, dans beaucoup de pays, y compris les pays industrialisés, il est rare que des listes complètes et à jour de la population, des logements ou des exploitations agricoles par exemple soient disponibles. Dans ces situations, l'échantillonnage peut s'effectuer à partir de cartes géographiques où chaque région urbaine est divisée en quartiers et chaque région rurale en groupements de terrains. Les quartiers et les superficies agricoles sont considérés comme des grappes et on travaille à partir de la liste complète des grappes à défaut d'une liste complète et à jour des unités de base. Ainsi, on échantillonne un nombre de grappes nécessaires à partir de la liste et ensuite on mène l'enquête sur toutes les unités de la grappe sélectionnée.

DOMAINES ET LIMITATIONS

L'avantage de l'**échantillonnage** par grappes réside dans le fait qu'il n'est pas nécessaire de numéroter tous les individus de la **population**.

En effet, dans beaucoup de pays, il n'y a pas de liste à jour disponible des gens ou des maisons par exemple. Le coût de la mise en place d'une telle liste est souvent prohibitif.

Il est donc plus facile de dresser une liste des sous-ensembles de la population (grappes).

En général, l'échantillonnage par grappes fournit des estimations moins précises que l'**échantillonnage aléatoire simple**, mais cette diminution de l'information est compensée par un coût beaucoup moins élevé.

Pour obtenir un échantillonnage par grappes aussi efficace que possible, il faut que :

- les grappes ne soient pas trop grosses, de façon à ce que leur nombre soit suffisant ;
- leur taille soit aussi uniforme que possible ;
- les individus qui les composent soient aussi hétérogènes que possible par rapport au caractère observé.

Une autre raison de procéder à un échantillonnage par grappes est une question de coût. Même quand il existe une liste complète et à jour des unités de base, il se peut que, pour des motifs d'ordre économique, il soit préférable de procéder à un échantillonnage par grappes. Ainsi, on diminue les frais de transport, de recrutement d'enquêteurs dans différentes régions, etc. L'échantillonnage par grappes est plus avantageux si la réduction des frais d'enquête est plus importante que l'augmentation de la variance échantillonnale qui en résulte. Le choix doit se faire en comparant les avantages liés à des coûts moindres et les inconvénients dus à une précision plus faible.

EXEMPLE

Soit N la **population** de la ville X. On veut étudier sa répartition selon l'âge sans faire de recensement. On divise alors la population en G quartiers. On effectue un **échantillonnage aléatoire simple** parmi ces quartiers. On obtient ainsi g quartiers.

L'**échantillon** final sera composé de tous les individus faisant partie des g quartiers sélectionnés.

MOTS CLÉS

Échantillonnage *(Sampling)*
Échantillonnage aléatoire simple *(Simple random sampling)*

RÉFÉRENCE

Hansen, M.H., Hurwitz, W.N. and Madow, M.G. (1953). *Sample Survey Methods and Theory. Vol. I : Methods and applications. Vol. II : Theory.* John Wiley & Sons, New York. Chapman & Hall, London.

ÉCHANTILLONNAGE PAR QUOTAS

Quota sampling

L'échantillonnage par quotas est une méthode d'**échantillonnage** non aléatoire. L'**échantillon** est choisi de façon à constituer une image aussi fidèle que possible de la **population**.

La méthode des quotas se base sur la répartition connue de la population pour un certain nombre de caractères (sexe, âge, catégorie socio-professionnelle...). On fait ensuite l'**hypothèse** que les différents caractères de la population sont corrélés, c'est-à-dire que si l'échantillon est représentatif pour les caractères de contrôle, il le sera aussi pour la **variable** étudiée.

L'échantillon est construit en respectant la distribution de la population et en choisissant un certain taux de **sondage** (quotas). Ensuite la désignation des personnes à interroger est laissée au choix de l'enquêteur.

HISTORIQUE

Voir **échantillonnage**.

ASPECTS MATHÉMATIQUES

Soit une **population** composée de N individus. On connaît la répartition de cette population pour les caractères de contrôle $A, B, \ldots, Z$.

Les caractères de contrôle $A, B, \ldots, Z$ ayant les modalités respectives A_1, A_2, ..., A_a ; B_1, B_2, ..., B_b ; ... ; Z_1, Z_2, ..., Z_z, chacune de taille respective X_{A_1}, X_{A_2}, ..., X_{A_a} ; X_{B_1}, X_{B_2}, ..., X_{B_b} ; ... ; X_{Z_1}, X_{Z_2}, ..., X_{Z_z}. On obtient l'**échantillon** désiré en multipliant les effectifs des différentes modalités de tous les caractères de contrôle par le taux de **sondage** q, c'est-à-dire le **pourcentage** de la population que l'on désire sonder.

L'échantillon sera de taille $n = N \cdot q$ et aura la distribution suivante :

$$q \cdot X_{A_1}, q \cdot X_{A_2}, \ldots, q \cdot X_{A_a},$$
$$q \cdot X_{B_1}, q \cdot X_{B_2}, \ldots, q \cdot X_{B_b}, \ldots,$$
$$q \cdot X_{Z_1}, q \cdot X_{Z_2}, \ldots, q \cdot X_{Z_z},$$

où $X_{\cdot}$ représente les effectifs des différentes modalités pour tous les caractères de contrôle.

DOMAINES ET LIMITATIONS

Un caractère de contrôle qui va servir de base pour la répartition de l'**échantillon** doit répondre aux conditions suivantes :

- il doit être en corrélation étroite avec les variables étudiées ;
- il doit avoir une distribution connue pour l'ensemble de la **population**.

Les caractères de contrôle les plus utilisés sont : le sexe, l'âge, la catégorie socio-professionnelle, la région, etc.

Les avantages de l'**échantillonnage** par quotas sont les suivants :

- la méthode n'exige pas, à la différence des échantillonnages aléatoires, l'existence d'une base de **sondage** ;
- le coût est nettement moins élevé que pour les méthodes aléatoires.

Les principaux inconvénients sont :

- la méthode des quotas n'a pas de fondement théorique suffisant. Elle repose sur l'**hypothèse** suivante : une distribution correcte des caractères de contrôle assure la représentativité de la distribution des caractères étudiés ;
- la méthode des quotas ne permet pas de calculer la précision des **estimations** obtenues à partir de l'échantillon. Les personnes interrogées étant choisies par les enquêteurs, il est impossible de savoir avec quelle **probabilité** l'individu de la population aurait fait partie de l'échantillon. On ne peut donc pas appliquer le calcul des probabilités qui, dans le cas d'un échantillonnage aléatoire, permet d'associer à chaque estimation une mesure de l'**erreur** qui peut avoir été commise.

EXEMPLE

On désire étudier les possibilités de marché pour un produit de consommation dans la ville X. Les variables de contrôle choisies sont le sexe, l'âge et la catégorie socio-professionnelle.

La répartition de la **population** (en milliers de personnes) selon les caractères de contrôle est la suivante :

Sexe		Âge		Catégorie socio-professionnelle	
Masculin	162	0–20	70	Patrons	19,25
Féminin	188	21–60	192,5	Prof.libérales	14,7
		61 et plus	87,5	Cadres sup.	79,1
				Ouvriers	60,9
				Inactifs	176,05
Total	350	Total	350	Total	350

On choisit un taux de sondage égal à $\frac{1}{300}$. L'**échantillon** a donc la taille et la distribution suivante :

$$\begin{aligned} n &= N \cdot q \\ &= 350\,000 \cdot \frac{1}{300} \\ &= 1\,167. \end{aligned}$$

Sexe		Âge		Catégorie socio-professionnelle	
Masculin	540	0–20	233	Patrons	64
Féminin	627	21–60	642	Prof.libérales	49
		61 et plus	292	Cadres sup.	264
				Ouvriers	203
				Inactifs	587
Total	1167	Total	1167	Total	1167

Le choix des individus se fait ensuite par un **échantillonnage aléatoire simple** dans chaque catégorie.

MOTS CLÉS

Échantillonnage *(Sampling)*
Estimation *(Estimation)*

RÉFÉRENCE

Sukhatme, P.V. (1954). *Sampling Theory of Surveys with Applications.* The Iowa State University Press. Iowa, U.S.A.

ÉCHANTILLONNAGE STRATIFIÉ

Stratified sampling

Dans un **échantillonnage** stratifié, on divise dans un premier temps la **population** en sous-populations

appelées *strates*. Ces strates ne doivent pas s'interpénétrer et l'ensemble de ces strates doit constituer l'ensemble de la population.

Une fois que les strates ont été déterminées, on tire un **échantillon** aléatoire (pas forcément de même taille) de chacune des strates, cet échantillonnage étant fait indépendamment dans les différentes strates.

HISTORIQUE

Voir **échantillonnage**.

ASPECTS MATHÉMATIQUES

En ce qui concerne la détermination du nombre d'individus à prélever dans chaque strate, il y a principalement deux approches : l'**échantillon** stratifié représentatif et l'échantillon stratifié optimal.

Échantillon stratifié représentatif

On tire, dans chaque strate, un nombre d'individus proportionnel à la grandeur de la strate.

Soit N_i la grandeur de la strate i, N le nombre d'individus de la **population** et n la taille de l'échantillon global, on définit la taille d'un sous-échantillon n_i par :

$$n_i = \frac{N_i}{N} \cdot n \quad i = 1, 2, \ldots, k$$

k étant le nombre de strates de la population.

Échantillon stratifié optimal

On se donne comme objectif de déterminer les tailles des sous-échantillons à tirer dans les différentes strates de façon à obtenir la meilleure **estimation** possible. Cela revient à minimiser la **variance** de l'**estimateur** m sous la contrainte

$$\sum_{i=1}^{k} n_i = n,$$

m étant la **moyenne** de l'échantillon définie par :

$$m = \frac{1}{N} \sum_{i=1}^{k} N_i \bar{x}_i$$

où $\bar{x}_i$ est la moyenne du sous-échantillon i. C'est un problème d'**optimisation** classique qui donne le résultat suivant :

$$n_i = \frac{n}{\sum_{i=1}^{k} N_i \cdot \sigma_i} N_i \cdot \sigma_i$$

où σ_i est l'**écart type** de la strate i.

Comme $n \sum_{i=1}^{k} N_i \cdot \sigma_i$ est un facteur constant, la taille du sous-échantillon est proportionnelle à la grandeur de la strate et à l'écart type.

La taille du sous-échantillon n_i devra être d'autant plus élevée que :

- la dispersion de la variable étudiée à l'intérieur de la strate sera grande ;
- la taille de N_i sera importante.

DOMAINES ET LIMITATIONS

L'**échantillonnage** stratifié est basé sur l'idée d'obtenir, grâce au contrôle de certaines variables, un **échantillon** qui soit une image aussi fidèle que possible de la **population**.

Il se peut que la dispersion au sein de la population soit grande ; dans ce cas, pour maintenir l'**erreur** d'**échantillonnage** à un niveau acceptable, il faut prélever un échantillon de taille très importante, ce qui peut augmenter considérablement le coût. Pour éviter ce problème, l'échantillonnage stratifié peut s'avérer judicieux. En effet, on peut subdiviser la population en strates de manière à ce que la variation à l'intérieur des strates soit relativement faible, ce qui permet de réduire la **taille d'échantillon** requise pour une erreur d'échantillonnage donnée.

Plusieurs raisons justifient la nécessité d'avoir recours à un échantillonnage stratifié :

1. si la population est trop hétérogène, il est préférable de constituer des groupes plus homogènes afin d'obtenir un échantillon plus représentatif de la population ;

2. si l'on désire obtenir des informations sur certains aspects de la population (par exemple, par rapport aux différents cantons suisses).

Un caractère de contrôle qui va servir à définir les strates doit obéir aux règles suivantes :

- il doit être en corrélation étroite avec les variables étudiées ;
- il doit avoir une **valeur** connue pour toutes les unités de la population.

Voici quelques-unes des variables de contrôle les plus utilisées : le sexe, l'âge, la région, le revenu, les catégories socio-professionnelles.

EXEMPLE

On considère une **population** de $N = 600$ étudiants divisée en $k = 4$ strates. On étudie les résultats d'un test dont les points vont de 0 à 100.

On possède les renseignements suivants :

	Taille N_i	Écart type σ_i
Strate 1	285	$8,6$
Strate 2	150	$5,2$
Strate 3	90	$2,2$
Strate 4	75	$1,4$

Si on désire prélever un **échantillon** stratifié de $n = 30$ individus, on obtient les sous-échantillons suivants :

1. *Échantillon stratifié représentatif*

$$n_i = \frac{N_i}{N} \cdot n.$$

On obtient alors les valeurs suivantes, arrondies à l'entier le plus proche :

$$\begin{aligned} n_1 &= \frac{N_1}{N} \cdot n \\ &= \frac{285}{600} \cdot 30 \\ &= 14,25 \\ &\approx 14 \end{aligned}$$

$$\begin{aligned} n_2 &= 7,5 \approx 7 \\ n_3 &= 4,5 \approx 5 \\ n_4 &= 3,75 \approx 4. \end{aligned}$$

2. *Échantillon stratifié optimal*

$$n_i = \frac{n}{\sum_{i=1}^{k} N_i \cdot \sigma_i} \cdot N_i \cdot \sigma_i.$$

On trouve :

$$\begin{aligned} N_1 \cdot \sigma_1 &= 2\,451 \\ N_2 \cdot \sigma_2 &= 780 \\ N_3 \cdot \sigma_3 &= 198 \\ N_4 \cdot \sigma_4 &= 105. \end{aligned}$$

Le facteur constant $\frac{n}{\sum_{i=1}^{k} N_i \cdot \sigma_i}$ est égal à :

$$\frac{n}{\sum_{i=1}^{k} N_i \cdot \sigma_i} = \frac{30}{3\,534} = 0,0085.$$

On obtient alors les valeurs suivantes, arrondies à l'entier le plus proche :

$$\begin{aligned} n_1 &= 0,0085 \cdot 2\,451 = 20,81 \approx 21 \\ n_2 &= 0,0085 \cdot 780 = 6,62 \approx 6 \\ n_3 &= 0,0085 \cdot 198 = 1,68 \approx 2 \\ n_4 &= 0,0085 \cdot 105 = 0,89 \approx 1. \end{aligned}$$

Si l'on compare les deux méthodes, on voit que la seconde prélève un plus grand nombre d'individus dans la première strate (la plus dispersée) et d'autant moins d'individus que la strate est homogène.

MOTS CLÉS

Échantillonnage *(Sampling)*
Échantillonnage aléatoire simple *(Simple random sampling)*
Estimateur *(Estimator)*
Estimation *(Estimation)*

RÉFÉRENCE

Yates, F. (1949). *Sampling Methods for Censuses and Surveys.* Charles Griffin & Company Ltd.

ÉCHANTILLONNAGE SYSTÉMATIQUE

Systematic sampling

L'échantillonnage systématique est un **échantillonnage** aléatoire. Les individus sont prélevés dans la **population** à des intervalles fixes en termes de temps, d'espace ou d'ordre d'occurrence, le premier individu étant tiré au hasard.

HISTORIQUE

Voir **échantillonnage**.

ASPECTS MATHÉMATIQUES

Soit une **population** composée de N individus. Un **échantillon** systématique est formé d'individus dont les numéros constituent une progression arithmétique.

On choisit aléatoirement un premier nombre b compris entre 1 et r, où $r = \frac{N}{n}$, n étant la taille de l'échantillon et r la raison de la progression arithmétique. Les individus ainsi prélevés porteront alors les numéros suivants :

$$b, b + r, b + 2r, \ldots, b + (n - 1)\, r.$$

DOMAINES ET LIMITATIONS

L'**échantillonnage** systématique n'est possible que si les individus sont classés dans un certain ordre. Le principal avantage de cette méthode réside dans sa simplicité et par conséquent son faible coût.

En revanche, elle présente le désavantage de ne pas tenir compte d'une éventuelle périodicité du caractère étudié. Il peut en résulter des erreurs graves surtout si la période est un sous-multiple de la raison de la progression arithmétique.

EXEMPLE

On considère une **population** de 600 étudiants classés par ordre alphabétique dans laquelle on désire prélever un **échantillon** de 30 individus selon la méthode de l'**échantillonnage** systématique.

La raison de la progression arithmétique est égale à :

$$r = \frac{N}{n} = \frac{600}{30} = 20.$$

On choisit aléatoirement un premier nombre entre 1 et 20, par exemple $b = 17$.

L'échantillon est alors composé des étudiants portant les numéros :

$$17,\ 37,\ 57,\ldots,\ 597$$

MOTS CLÉS

Échantillonnage *(Sampling)*
Échantillonnage par grappes *(Cluster sampling)*

RÉFÉRENCE

Deming, W.E. (1960). *Sample Design in Business Research*, John Wiley & Sons, New York.

ÉCONOMÉTRIE

Econometrics

L'économétrie concerne l'application des méthodes statistiques à des données économiques. Les économistes utilisent la **statistique** dans le but de tester leurs théories ou de faire des prévisions.

Puisqu'en économie les données ne sont pas expérimentales et qu'elles contiennent très souvent une part d'aléatoire, l'économétrie fait appel à des modèles stochastiques plutôt qu'à des modèles déterministes.

HISTORIQUE

L'utilité des mathématiques et des statistiques dans le domaine de l'économie n'est pas négligeable, on s'en

rendit compte rapidement. Dans la seconde partie du XVII^{e} siècle, Sir William Petty (1676) publia un article très remarquable pour l'époque introduisant les fondements méthodologiques de l'économétrie.

Selon W. Jaffé (1968), Léon Walras, professeur à l'Université de Lausanne, est reconnu comme étant à l'origine de l'équilibre général en économie, qui constitue la base théorique de l'économétrie moderne.

Le 29 décembre 1930, à Cleveland, Ohio, un groupe d'économistes, de statisticiens et de mathématiciens fondèrent une Société d'économétrie dans le but de promouvoir la recherche des théories mathématiques et statistiques dans les domaines économiques. Ils créèrent la revue bimensuelle *Econometrica* qui parut pour la première fois en janvier 1933.

Le développement des modèles d'équations simultanées que l'on a pu utiliser dans les séries chronologiques de prévisions économiques a été le déclic historique de l'économétrie en tant que domaine distinct. Il reste aujourd'hui une part importante de l'économétrie.

MOTS CLÉS

Régression linéaire multiple *(Multiple linear regression)*
Régression linéaire simple *(Simple linear regression)*
Série chronologique *(Time series)*

RÉFÉRENCES

Jaffé, W. (1968). Walras, Léon. *International Encyclopedia of the Social Sciences.* Vol. 16, pp. 447-453. Édité par David L. Sills. Macmillan and Free Press, New York.

Petty, W. (1676). Political Arithmetick. In William Petty, *The Economic Writings ...*, vol. 1, pp. 233-313. Kelley, New York.

Wonnacott, R.J. and Wonnacott, T.H. (1970). *Econometrics.* John Wiley & Sons, New York.

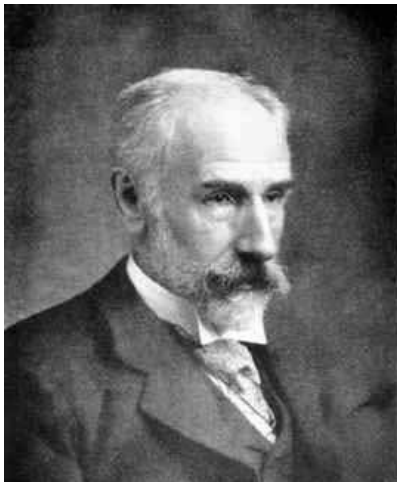
EDGEWORTH FRANCIS Y.

Edgeworth Francis Y.

Francis Ysidro Edgeworth est né en 1845 à Edgeworthtown (Irlande) et est mort en 1926. Ses contributions apparaissent dans des sujets aussi divers que les sciences morales, économiques, la probabilité et les statistiques. Son travail précoce et important sur les indices et la théorie de l'utilité furent suivis, plus tard, par un travail statistique qui est maintenant parfois entrevu comme « Bayesien ».

On lui doit les termes de « module » et de « fluctuation ».

Il a été le premier éditeur de l'*Economic Journal* dès 1881.



Quelques ouvrages et articles principaux de Francis Ysidro Edgeworth :

1881 *Mathematical Physics : an essay on the application of mathematics to the moral sciences.* Kegan Paul, London.

EFFET PRINCIPAL

Main effect

Dans une **expérience factorielle**, on appelle effet principal la contribution d'un **niveau** de **facteur** spécifique à la **variable** réponse.

ASPECTS MATHÉMATIQUES

On considère le cas simple d'un **modèle** additif d'un plan à deux facteurs (modèle linéaire qui prend en compte uniquement les effets principaux des facteurs et qui ignore les interactions entre les facteurs) :

$$Y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij},$$

$$i = 1, 2, \ldots, a \text{ et } j = 1, 2, \ldots, b,$$

où

- Y_{ij} représente la réponse recevant le **traitement** ij,
- μ est la **moyenne** commune à tous les traitements,
- α_i est l'effet du i-ème **niveau** du **facteur** A,
- β_j est l'effet du j-ème niveau du facteur B et
- ε_{ij} est l'**erreur** expérimentale de l'**observation** Y_{ij}.

Si $a = b = 2$, les effets principaux des facteurs A et B, notés a et b respectivement sont donnés par :

$$\begin{aligned} a &= \frac{(\alpha_1 - \alpha_0)(\beta_1 + \beta_0)}{2} \\ b &= \frac{(\alpha_1 + \alpha_0)(\beta_1 - \beta_0)}{2}. \end{aligned}$$

MOTS CLÉS

Analyse de variance *(Analysis of variance)*
Estimation *(Estimation)*
Expérience factorielle *(Factorial experiment)*
Facteur *(Factor)*
Interaction *(Interaction)*
Test d'hypothèse *(Hypothesis testing)*

RÉFÉRENCE

Box, G.E.P. and Hunter, W.G. and Hunter, J.S. (1978). *Statistics for experimenters. An introduction to design, data analysis, and model building.* John Wiley & Sons, New York.

EMPAN

Range

L'empan est la **mesure de dispersion** la plus simple à calculer. Il est défini comme la différence entre la valeur observée la plus grande et la valeur observée la plus petite d'un ensemble d'observations relatives à une **variable quantitative**.

L'empan est aussi parfois appelé *étendue*.

ASPECTS MATHÉMATIQUES

Soit un ensemble d'observations relatives à une **variable quantitative** X.

Si nous notons X_{max} la **valeur** de l'observation la plus grande dans un ensemble d'observations, et X_{min} la valeur la plus petite, l'empan est donné par :

$$\text{empan} = X_{max} - X_{min}.$$

Lorsque les observations sont groupées en classes, l'empan est égal à la différence entre les centres des deux classes extrêmes. Soit δ_1 le centre de la première classe et δ_k le centre de la dernière classe, l'empan est égal à :

$$\text{empan} = \delta_k - \delta_1.$$

DOMAINES ET LIMITATIONS

L'empan est une **mesure de dispersion** très peu utilisée car il ne fournit qu'une indication de l'étendue des valeurs, mais ne donne aucune information sur la distribution à l'intérieur de cette étendue.

Il présente cependant le grand avantage d'être simple à calculer. Toutefois, il faut faire très attention dans son interprétation.

L'empan est une mesure de dispersion qui ne tient compte que des deux valeurs extrêmes. Cette méthode peut donc mener à de fausses interprétations : si le nombre d'enfants dans sept familles est respectivement 0, 0, 1, 1, 2, 3 et 6, l'empan est 6, alors que toutes les familles ont entre 0 et 3 enfants, sauf une. Dans ce cas, la variabilité est relativement faible alors que l'empan est élevé.

L'utilisation de l'empan pour mesurer la dispersion (ou la variabilité) présente un autre inconvénient. En effet, celui-ci ne peut qu'augmenter lorsque le nombre d'observations augmente.

Ceci constitue un désavantage, car idéalement, une mesure de variabilité devrait être indépendante du nombre d'observations.

Il convient d'être très attentif à l'interprétation de l'empan en cas d'observations aberrantes. En effet, celles-ci ont une influence directe sur la valeur de l'empan.

EXEMPLE

Voici les âges de deux groupes de personnes suivant un cours du soir particulier :

1^{er} groupe :				
33	34	34	37	40
40	40	43	44	45
2^{e} groupe :				
22	25	31	40	40
41	44	45	50	52

La **moyenne arithmétique** $\bar{x}$ de ces deux ensembles d'observations est identique :

$$\bar{x} = \frac{390}{10} = 39,$$

cependant, la dispersion autour de la moyenne est bien différente.

L'empan pour le premier groupe est égal à 12 $(45 - 33)$, alors que l'empan pour le deuxième groupe est égal à $52 - 22 = 30$ ans.

Il faut toutefois être conscient que l'empan ne nous donne qu'une mesure indicative de l'étendue des valeurs mais ne nous indique pas quelle est la forme de la distribution à l'intérieur de cette étendue.

MOTS CLÉS

Intervalle interquartile *(Interquartile range)*
Mesure de dispersion *(Measure of dispersion)*

RÉFÉRENCE

Anderson, T.W. and Sclove, S.L. (1978). An Introduction to the Statistical Analysis of Data. Houghton Mifflin Company, Boston.

ENSEMBLE FONDAMENTAL

Sample space

L'ensemble fondamental d'une **expérience aléatoire** est l'ensemble de tous les résultats possibles de l'**expérience**, dénoté généralement par Ω.

HISTORIQUE

Voir **probabilité**.

ASPECTS MATHÉMATIQUES

Un ensemble fondamental peut être :

- fini (les résultats sont dénombrables) ;
- infini dénombrable (il est possible d'associer à chaque résultat possible un nombre entier naturel, de telle sorte que chaque résultat possible ait un nombre différent ; toutefois, le nombre de résultats possibles est infini) ;
- infini non dénombrable (il est impossible de dénombrer tous les résultats possibles).

L'ensemble fondamental Ω d'une **expérience aléatoire** est aussi appelé **événement certain**.

La **probabilité** de l'événement certain est égale à 1 :

$$P(\Omega) = 1.$$

EXEMPLE

Ensemble fondamental fini

Lancement d'une pièce de monnaie

$\Omega = \{Pile, Face\}$

Lancement d'un dé

$\Omega = \{1, 2, 3, 4, 5, 6\}$

Ensemble fondamental infini dénombrable

Lancement d'une pièce de monnaie autant de fois qu'il est nécessaire pour obtenir Face (F = Face et P = Pile)

$\Omega = \{F, PF, PPF, PPPF, PPPPF, \ldots\}$

Ensemble fondamental infini non dénombrable

Repérage des coordonnées de l'impact d'une balle de fusil, sur un panneau de 10 cm sur 10 cm. Dans ce cas, l'ensemble fondamental Ω est composé d'un ensemble infini non dénombrable de coordonnées puisque les coordonnées sont mesurées sur l'échelle des nombres réels (il est impossible de donner une liste exhaustive des résultats possibles).

MOTS CLÉS

Événement *(Event)*
Expérience aléatoire *(Random experiment)*
Probabilité *(Probability)*

RÉFÉRENCE

Bailey, D.E. (1971). Probability and statistics : models for research. John Wiley & Sons, New York.

ÉPIDÉMIOLOGIE

Epidemiology

L'épidémiologie, mot provenant des termes grecs *epidêmos* et *logos*, et signifiant « étude des épidémies », est l'étude de la fréquence et de la répartition des maladies, des paramètres et des facteurs de risque conditionnant les états de santé au sein d'une population, relativement au temps, à l'espace et à des groupes d'individus.

HISTORIQUE

L'origine de l'épidémiologie remonte au XVII[e] siècle, lorsque William Petty (1623-1687), médecin, économiste et scientifique anglais, récolta des informations chiffrées sur la population, consignées dans l'ouvrage *l'Arithmétique politique*. John Graunt (1620-1694) proposa la première analyse rigoureuse des causes de décès en 1662. **Adolphe Quetelet** (1796-1874), astronome et mathématicien belge, est considéré comme le fondateur de la statistique moderne des populations, discipline-mère de l'épidémiologie, de la statistique, de l'économétrie et des autres disciplines quantitatives décrivant les caractéristiques sociales ou biologiques de la vie humaine. Avec William Farr (1807-1883), l'épidémiologie s'affirma en tant que domaine à part entière de la statistique, étudiant les causes des décès et la façon dont celles-ci variaient avec l'âge, le sexe, la saison, le lieu de résidence ou la profession. Le médecin écossais James Lind (1716-1794), en prouvant que la consommation des agrumes permettait de guérir le scorbut, marqua une étape importante de l'histoire de l'épidémiologie. John Snow (1813-1858), médecin londonien, montra quant à lui que les épidémies se propageaient par contagion, mettant en évidence par une étude *in situ* que boire de l'eau de la Tamise, polluée par les rejets des égouts londoniens, contribuait à la dissémination du choléra.

ASPECTS MATHÉMATIQUES

Voir **cause et effet en épidémiologie, odds et odds ratio, risque relatif, risque attribuable, risque évitable, taux d'incidence, taux de prévalence.**

EXEMPLE

Prenons l'exemple de la mortalité entre deux régions fictives : supposons qu'il s'agisse d'analyser la mortalité par année dans deux régions A et B, peuplée chacune de 10 000 habitants. Les résultats sont indiqués dans le tableau ci-dessous :

	Région A	Région B
décès/habitants	500/10 000	1 000/10 000
mortalité globale	5 %	10 %
mortalité - de 60 ans	5 %	5 %
mortalité + de 60 ans	15 %	15 %
% - de 60 ans	100 %	50 %

De ces données, peut-on affirmer que les habitants de la région B subissent des facteurs de mortalité plus importants que les habitants de la région A ? Autrement dit, serait-il recommandable de quitter la région A pour B ? Si l'on décline les deux populations par classe d'âge, on remarque une identité des taux de mortalité, si bien qu'en réalité, la différence de mortalité globale entre les régions A et B n'apparaît ici comme étant seulement expliquée par la différence des structures par âge entre A et B. Il s'agit donc de ce que l'on nomme typiquement en statistique un « effet de structure ».

L'exemple précédent doit attirer l'attention sur la multitude de facteurs qui doivent être pris en compte lorsqu'on analyse des données. Parmi les facteurs principaux dont une étude épidémiologique doit tenir compte, on peut citer en particulier : l'âge (de manière absolue comme correspondant à un moment de la vie biologique et de manière relative comme représentatif de l'appartenance à une génération), le sexe, l'origine ethnique, etc.

MOTS CLÉS

Biostatistique *(Biostatistics)*
Cause et effet en épidémiologie *(Cause and effect in epidemiology)*
Odds et odds ratio *(Odds and odds ratio)*
Risque relatif *(Relative risk)*
Risque attribuable *(Attributable risk)*
Risque évitable *(Avoidable risk)*
Taux d'incidence *(Incidence rate)*
Taux de prévalence *(Prevalence rate)*

RÉFÉRENCES

Lilienfeld A.M. and D.E. Lilienfeld. (1980). *Foundations of epidemiology.* 2nd ed New York: Clarendon Press, Oxford.

MacMahon, B. and Pugh, T.F. (1970). *Epidemiology: Principles and Methods.* Boston: Little Brown.

Morabia A. (1996). *Épidémiologie causale.* Éditions Médecine et Hygiène, Genève.

Morabia A. (1996). *L'Épidémiologie clinique.* Presses universitaires de France, Que sais-je ? Paris.

ÉPIDÉMIOLOGIE CAUSALE

Causal Epidemiology

L'**épidémiologie causale** (ou analytique) vise à l'identification des relations de cause à effet en rapport avec la santé humaine.

Elle a pour objet l'étude des causes des maladies et cherche à mettre en évidence, soit un lien de causalité, soit une association, c'est-à-dire au minimum un lien statistique entre un facteur supposé explicatif et une maladie.

HISTORIQUE

Voir **épidémiologie.**

ASPECTS MATHÉMATIQUES

Voir **cause et effet en épidémiologie, odds et odds ratio, risque relatif, risque attribuable, risque évitable, taux d'incidence, taux de prévalence.**

DOMAINES ET LIMITATIONS

L'étude de la relation entre le tabac et les tumeurs bronchiques malignes ou entre le VIH et le SIDA sont des exemples de l'épidémiologie causale. La recherche d'une relation causale se caractérise souvent par sa complexité, nécessitant de nombreuses études et la synthèse de nombreuses informations issues d'expérimentations biologiques, animales, d'examens cliniques et d'enquêtes en population.

Si la cause ne peut toujours être clairement identifiée, la connaissance des facteurs de risque pour une maladie et, de ce fait, l'identification de populations ou de groupes à risque, permet d'envisager des interventions sanitaires et des mesures de prévention adaptées, ciblées.

EXEMPLE

Prenons l'exemple de la relation entre le fait de fumer et de développer un cancer du poumon. Considérons l'étude de 2 000 sujets : 1 000 fumeurs et 1 000 non-fumeurs en sachant que les structures par âge et les proportions hommes/femmes sont identiques dans les

deux groupes. Analysons le bilan des observations suivies au cours de plusieurs années. Les résultats sont présentés dans le tableau suivant :

	Fumeurs	Non-fumeurs
Nombre de cas de cancer du poumon (/1 000)	50	10
Proportion d'individus atteints de cancer du poumon	5 %	1 %

Si nous comparons la proportion d'individus atteints de cancer du poumon entre fumeurs et non-fumeurs, nous obtenons 5 %/1 % = 5, ce qui nous permet de conclure que les fumeurs ont 5 fois plus de risque de développer un cancer du poumon que les non-fumeurs. Géneralement, nous évaluons le caractère significatif de ce résultat en calculant la **valeur** p. Supposons qu'au moyen du **test du chi-carré**, une valeur p inférieure à $0,001$ soit obtenue. Il est ordinairement admis que si la valeur p devient inférieure à $0,05$, les résultats obtenus sont statistiquement significatifs.

Supposons que nous effectuions la même étude, mais que la taille de chacun des deux groupes est de 100 au lieu de 1 000 et que nous observions des résultats inchangés :

	Fumeurs	Non-fumeurs
Nombre de cas de cancer du poumon (/100)	5	1
Proportion d'individus atteints de cancer du poumon	5 %	1 %

Si les tests statistiques sont à nouveau effectués, la valeur p devient $0,212 > 0,05$. De cette étude nous ne pouvons tout simplement pas conclure qu'il existe une relation statistique significative entre les fumeurs et le cancer du poumon.

En effet, dans les études épidémiologiques, afin de pouvoir détecter un certain niveau de différence des résultats observés entre différentes populations, il est nécessaire de disposer d'un échantillon important. Ainsi, dans les cas où des études ne portent que sur un nombre d'individus, leurs conclusions restent fragiles.

MOTS CLÉS

Voir **épidémiologie.**

RÉFÉRENCES

Voir **épidémiologie.**

ÉPIDÉMIOLOGIE DESCRIPTIVE

Descriptive epidemiology

L'**épidémiologie descriptive** correspond à la description de la fréquence et de la répartition des maladies au sein d'une population.

À cet effet, elle est fondée sur le recueil d'informations sanitaires (tables de mortalité, registre de morbidité, maladies à déclaration obligatoire...) ou d'informations pouvant avoir un impact sur la santé des populations (données sur la pollution atmosphérique, comportements à risque...) et contribue à produire des indicateurs statistiques synthétiques permettant de forger une image des phénomènes de santé au sein d'une population.

HISTORIQUE

Voir **épidémiologie.**

ASPECTS MATHÉMATIQUES

Voir **cause et effet en épidémiologie, odds et odds ratio, risque relatif, risque attribuable, risque évitable, taux d'incidence, taux de prévalence.**

DOMAINES ET LIMITATIONS

Bien que l'épidémiologie descriptive n'apporte que des informations dites « élémentaires », son importance est primordiale :

- elle permet d'appréhender l'ampleur des phénomènes de santé (en appréciant leur **prévalence** et leur **incidence**) et de disposer d'une surveillance épidémiologique ;

- elle constitue une aide à la décision dans les domaines de la planification et de la gestion des organisations ou des programmes de santé ;

- elle peut être à l'origine d'hypothèses sur l'étiologie ou les facteurs de risque des maladies (exemple : répartition préférentielle des tumeurs malignes de la peau dans le sud de la France d'où la mise en évidence d'une association plausible tumeur maligne cutanée - rayons solaires).

L'épidémiologie descriptive ne peut établir de lien de cause à effet, elle n'est pas ou peu prédictive.

MOTS CLÉS

Voir **épidémiologie.**

RÉFÉRENCES

Voir **épidémiologie.**

ÉPREUVE DE BERNOULLI

Bernoulli trial

Les épreuves de Bernoulli sont des essais répétés d'une **expérience** qui obéissent aux conditions suivantes :

1. chaque épreuve aboutit soit à un succès soit à un échec ;

2. la **probabilité** de succès est la même pour chaque épreuve ; on note par p la probabilité de succès, et par $q = 1 - p$ la probabilité d'échec ;

3. les épreuves sont indépendantes.

HISTORIQUE

Les épreuves de Bernoulli tirent leur nom de celui du mathématicien suisse **Jakob Bernoulli** (1713).

Jakob Bernoulli (1654-1705) était le plus âgé de quatre frères et son père le destinait à la théologie. Après avoir terminé ses études de théologie à Bâle en 1676, il quitta cette ville pour y revenir en 1680 et s'intéressa dès lors aux mathématiques. Il obtint la chaire de mathématiques de l'Université de Bâle en 1687.

EXEMPLE

L'exemple le plus simple que l'on peut donner pour une épreuve de Bernoulli est le jet d'une pièce de monnaie. Si le fait d'obtenir « face » est considéré comme un succès (S) et le fait d'obtenir « pile » est considéré comme un échec (E), on a :

$$p = P(S) = \tfrac{1}{2} \quad \text{et}$$
$$q = P(E) = 1 - p = \tfrac{1}{2}.$$

MOTS CLÉS

Loi binomiale *(Binomial distribution)*
Loi de Bernoulli *(Bernoulli distribution)*

RÉFÉRENCES

Bernoulli, J. (1713). *Ars Conjectandi, Opus Posthumum. Accedit Tractatus de Seriebus infinitis, et Epistola Gallice scripta de ludo Pilæ recticularis.* Basileæ, Impensis Thurnisiorum, fratrum.

Ross, S.M. (1987). *Initiation aux probabilités.* Presses polytechniques romandes, Lausanne.

ÉQUATIONS NORMALES

Normal equations

Les équations normales sont les équations obtenues en annulant les dérivées partielles de la somme des carrés des erreurs (**moindres carrés**) et qui permettent d'estimer les paramètres d'une régression linéaire multiple.

HISTORIQUE

Voir **régression, analyse de**.

ASPECTS MATHÉMATIQUES

Considérons un **modèle** général de **régression linéaire multiple** :

$$Y_i = \beta_0 + \sum_{j=1}^{p-1} \beta_j \cdot X_{ji} + \varepsilon_i, \quad i = 1, \ldots, n$$

où

Y_i	est la **variable dépendante**,
X_{ji},	$j = 1, \ldots, p-1$, sont les variables indépendantes,
ε_i	est le terme d'**erreur** aléatoire non observable,
β_j,	$j = 0, \ldots, p-1$, sont les paramètres à estimer et
n	est le nombre d'observations.

Pour appliquer la méthode des **moindres carrés**, il faut minimiser la somme des erreurs au carré :

$$\begin{aligned} f(\beta_0, \beta_1, \ldots, \beta_{p-1}) &= \sum_{i=1}^{n} \varepsilon_i^2 \\ &= \sum_{i=1}^{n} \left(Y_i - \beta_0 - \sum_{j=1}^{p-1} \beta_j \cdot X_{ji} \right)^2. \end{aligned}$$

En annulant les p dérivées partielles par rapport aux paramètres à estimer, on obtient les p équations normales :

$$\begin{cases} \frac{\partial f(\beta_0, \beta_1, \ldots, \beta_{p-1})}{\partial \beta_0} = 0, \\ \frac{\partial f(\beta_0, \beta_1, \ldots, \beta_{p-1})}{\partial \beta_1} = 0, \\ \vdots \\ \frac{\partial f(\beta_0, \beta_1, \ldots, \beta_{p-1})}{\partial \beta_{p-1}} = 0. \end{cases}$$

Nous pouvons également exprimer ce même modèle matriciellement :

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

où

$\mathbf{Y}$	est le **vecteur** $(n \times 1)$ des observations relatives à la variable dépendante (n observations),
$\mathbf{X}$	est la **matrice** $(n \times p)$ du plan ayant trait aux variables indépendantes,
$\boldsymbol{\varepsilon}$	est le vecteur $(n \times 1)$ des erreurs, et
$\boldsymbol{\beta}$	est le vecteur $(p \times 1)$ des paramètres à estimer.

Par la méthode des **moindres carrés**, il faut minimiser :

$$\begin{aligned} \boldsymbol{\varepsilon}'\boldsymbol{\varepsilon} &= (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}')(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) \\ &= \mathbf{Y}'\mathbf{Y} - \boldsymbol{\beta}'\mathbf{X}'\mathbf{Y} - \mathbf{Y}'\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta} \\ &= \mathbf{Y}'\mathbf{Y} - 2\boldsymbol{\beta}'\mathbf{X}'\mathbf{Y} + \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta}. \end{aligned}$$

En annulant les dérivées par rapport à $\boldsymbol{\beta}$ (sous forme matricielle), correspondant aux dérivées partielles, mais écrites vectoriellement, nous trouvons les équations normales :

$$\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{Y}.$$

DOMAINES ET LIMITATIONS

Nous pouvons généraliser le concept d'équations normales au cas d'une **régression non linéaire** (avec r observations et $p+1$ paramètres à estimer).

Elles s'expriment alors par :

$$\begin{aligned} & \sum_{i=1}^{r} \frac{(Y_i - f(X_{1i}, \ldots, X_{ki}\,;\, \beta_0, \ldots, \beta_p))^2}{\partial \beta_j} \\ = & \sum_{i=1}^{r} (Y_i - f(X_{1i}, \ldots, X_{ki}\,;\, \beta_0, \ldots, \beta_p)) \\ & \cdot \left[\frac{\partial f(X_{1i}, \ldots, X_{ki}\,;\, \beta_0, \ldots, \beta_p)}{\partial \beta_j} \right] \\ = & \; 0 \end{aligned}$$

où $j = 0, \ldots, p$. Puisque f est non linéaire dans les paramètres à estimer, les équations normales sont aussi non linéaires, et donc peuvent s'avèrer très difficiles à résoudre. De plus, le système d'équations peut avoir plus d'une solution, ce qui correspond à la possibilité de plusieurs minima de f.

Normalement, nous essayons de développer un procédé itératif pour les résoudre, ou de se référer aux techniques expliquées dans la régression non linéaire.

EXEMPLE

Si le **modèle** contient par exemple trois paramètres β_0, β_1 et β_2, les équations normales sont alors les suivantes :

$$\begin{cases} \frac{\partial f(\beta_0, \beta_1, \beta_2)}{\partial \beta_0} &= 0 \\ \frac{\partial f(\beta_0, \beta_1, \beta_2)}{\partial \beta_1} &= 0 \\ \frac{\partial f(\beta_0, \beta_1, \beta_2)}{\partial \beta_1} &= 0. \end{cases}$$

C'est-à-dire :

$$\begin{cases} \beta_0 \cdot n + \beta_1 \cdot \sum_{i=1}^{n} X_{1i} + \beta_2 \cdot \sum_{i=1}^{n} X_{2i} = \sum_{i=1}^{n} Y_i \\ \beta_0 \cdot \sum_{i=1}^{n} X_{1i} + \beta_1 \cdot \sum_{i=1}^{n} X_{1i}^2 + \beta_2 \cdot \sum_{i=1}^{n} X_{1i} \cdot X_{2i} \\ \quad = \sum_{i=1}^{n} X_{1i} \cdot Y_i \\ \beta_0 \cdot \sum_{i=1}^{n} X_{2i} + \beta_1 \cdot \sum_{i=1}^{n} X_{2i} \cdot X_{1i} + \beta_2 \cdot \sum_{i=1}^{n} X_{2i}^2 \\ \quad = \sum_{i=1}^{n} X_{2i} \cdot Y_i. \end{cases}$$

Voir **régression linéaire simple**.

MOTS CLÉS

Erreur *(Error)*
Moindres carrés *(Least squares)*
Paramètre *(Parameter)*
Régression linéaire multiple *(Multiple linear regression)*
Régression linéaire simple *(Simple linear regression)*
Régression non linéaire *(Non-linear regression)*

RÉFÉRENCE

Legendre, A.M. (1805). *Nouvelles méthodes pour la détermination des orbites des comètes*, Courcier, Paris.

ERREUR

Error

La notion d'erreur implique l'existence d'une certaine représentation de la réalité. Par définition, l'erreur est la différence entre la **valeur** observée et la vraie valeur (ou valeur de référence) de la quantité concernée.

HISTORIQUE

Les lois des erreurs sont les distributions de **probabilité** qui décrivent les distributions des erreurs apparaissant lors de mesures répétées d'une même quantité sous les mêmes conditions. Elles ont été introduites dans la deuxième moitié du XVIII^{e} siècle pour montrer l'utilité de la **moyenne arithmétique** afin d'obtenir une bonne approximation de la **valeur** de référence de la quantité cherchée.

Dans une lettre au président de la Royal Society, Thomas Simpson (1756) suggéra que les chances respectives des différentes erreurs auxquelles est assujettie une **observation** peuvent être exprimées par une **loi de probabilité discrète**. Il introduisit alors les deux premières lois des erreurs. D'autres lois discrètes furent proposées et étudiées par Joseph-Louis Lagrange (1776).

T. Simpson (1757) proposa également des lois d'erreurs continues tout comme **Pierre Simon de Laplace** (1774 et 1781). Cependant, la plus importante loi des erreurs est due à **Carl Friedrich Gauss** (1809). Elle est plus connue sous le nom **loi normale** et sera reconnue universellement comme « la » loi des erreurs.

ASPECTS MATHÉMATIQUES

Dans le cadre plus général de la métrologie, on distingue les types d'erreurs suivants :

- l'erreur absolue, qui est la différence prise en valeur absolue de la **valeur** observée et la valeur de référence d'une certaine grandeur ;
- l'erreur relative, rapport de l'erreur absolue à la valeur de la grandeur elle-même. Elle caractérise l'exactitude d'une mesure physique ;
- l'erreur accidentelle n'est pas liée à l'outil de mesure employé mais à l'expérimentateur lui-même ;
- l'erreur expérimentale, erreur due à des variables non contrôlées ;
- l'erreur aléatoire, erreur résultant de la **combinaison**, due au hasard, de multiples erreurs partielles imputables soit à l'instrument soit à l'utilisateur. Les propriétés statistiques des erreurs aléatoires et leur **estimation** s'étudient à partir du calcul des probabilités ;
- l'erreur résiduelle, correspond à la différence entre la valeur estimée et la valeur observée. On utilise parfois le terme de **résidu** ;

- l'erreur systématique, erreur qui provient de causes permanentes (par exemple du réglage défectueux d'un appareil de mesure) et qui se produit toujours dans le même sens. Le **biais**, en statistique, en est un exemple particulier ;
- l'erreur d'arrondi, qui est l'erreur effectuée en remplaçant une valeur numérique par une valeur proche mais tronquée. Lors de calculs avec des chiffres à n décimales, l'erreur d'arrondi est comprise entre :

$$-\frac{1}{2}\cdot 10^{-n} \text{ et } \frac{1}{2}\cdot 10^{-n}.$$

  À l'occasion de longues procédures de calcul, les erreurs d'arrondis peuvent se cumuler et fournir des résultats très imprécis. Ce type d'erreur est très fréquent avec l'utilisation de l'ordinateur.

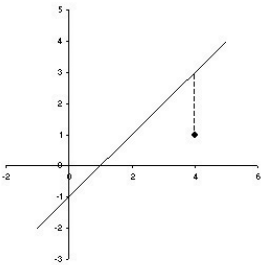
Toute erreur qui survient dans le cadre d'un problème **statistique** n'est en fait qu'une combinaison des différents types d'erreurs présentés ci-dessus.

En statistique et plus spécialement lors d'un **test d'hypothèse**, on rencontre également :

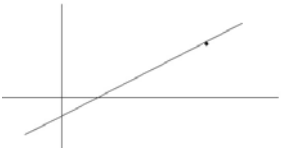
- l'**erreur de première espèce** correspondant à l'erreur commise lorsqu'on rejette l'**hypothèse nulle** alors que celle-ci est vraie ;
- l'**erreur de seconde espèce** ou erreur commise lorsqu'on ne rejette pas l'hypothèse nulle alors que celle-ci est fausse.

EXEMPLE

Pour erreurs relative et absolue



Pour erreur d'arrondi



MOTS CLÉS

Analyse des résidus *(Analysis of residuals)*
Biais *(Bias)*
Erreur de première espèce *(Type I error)*
Erreur de seconde espèce *(Type II error)*
Estimation *(Estimation)*
Résidu *(Residual)*
Test d'hypothèse *(Hypothesis testing)*

RÉFÉRENCES

Gauss, C.F. (1809). *Theoria motus corporum cœlestium.* Werke 7.

Lagrange, J.L. (1776). Mémoire sur l'utilité de la méthode de prendre le milieu entre les résultats de plusieurs observations ; dans lequel on examine les avantages de cette méthode par le calcul des probabilités; et où l'on résoud différents problèmes relatifs à cette matière. *Miscellanea Taurinensia* 5, pp. 167-232.

Laplace, P.S. de (1774). *Mémoire sur la probabilité des causes par les événements.* Mémoire de l'Académie royale des sciences présenté par divers savants, 6, pp. 621-656 ou Vol. 8, pp. 27-65. In P.S. de Laplace, *Œuvres complètes* (1891). Gauthier-Villars, Paris.

Laplace, P.S. de (1781). Mémoire sur les probabilités. *Mémoires de l'Académie royale des sciences de Paris*, 1778, pp. 227-332 ou Vol. 9 pp. 383-485 dans P.S. de Laplace, *Œuvres complètes* (1891). Gauthier-Villars, Paris.

Simpson, T. (1755). A Letter to the Right Honorable George Earl of Macclesfield, President of the Royal Society, on the Advantage of Taking the Mean

of a Number of Observations in Practical Astronomy. *Philosophical Transactions of the Royal Society of London*, 49, pp. 82-93.

Simpson, T. (1757). *Miscellaneous Tracts on Some Curious and Very Interesting Subjects in Mechanics, Physical-Astronomy and Speculative Mathematics.* J. Nourse, London.

ERREUR DE PREMIÈRE ESPÈCE

Type I error

Dans la réalisation d'un **test d'hypothèse**, nous appelons erreur de première espèce α la **probabilité** de rejeter l'**hypothèse nulle** H_0 alors que celle-ci est vraie :

$$\alpha = P(\text{rejeter } H_0 \mid H_0 \text{ est vraie})$$

L'erreur de première espèce α est égale au **seuil de signification** du test d'hypothèse.

HISTORIQUE

En 1928, **Jerzy Neyman** et **Egon Sharpe Pearson** furent les premiers auteurs à reconnaître que le choix rationnel d'un **test d'hypothèse** doit tenir compte non seulement de l'**hypothèse** qu'on souhaite vérifier, mais aussi de l'**hypothèse alternative**. Ils introduisirent la distinction entre l'**erreur** de première espèce et l'**erreur de seconde espèce** et de cette façon ils présentèrent leurs travaux relatifs au critère de vraisemblance comme une généralisation de la méthode de construction d'un test d'hypothèse.

DOMAINES ET LIMITATIONS

L'erreur de première espèce est une des deux erreurs auxquelles le statisticien est confronté dans la réalisation d'un **test d'hypothèse** : l'erreur qui peut survenir dans la prise de décision si l'**hypothèse nulle** est vraie.

Si l'hypothèse nulle est fausse, il est confronté à un autre type d'erreur, appelée **erreur de seconde espèce**, à savoir la **probabilité** d'accepter l'hypothèse nulle H_0 alors que celle-ci est fausse.

Nous pouvons représenter les différents types d'erreur par le tableau suivant :

Situation	Décision	
	accepter H_0	rejeter H_0
H_0 vraie	$1-\alpha$	α
H_0 fausse	β	$1-\beta$

α étant l'erreur de première espèce et β l'erreur de seconde espèce.

EXEMPLE

Nous illustrons par un exemple de la vie courante les différentes erreurs que nous pouvons commettre en prenant une décision :

Lorsque nous sortons de chez nous le matin, nous nous demandons quel temps il fera dans la journée. Si nous pensons qu'il risque de pleuvoir, nous prendrons notre parapluie. En revanche, si nous estimons qu'il y aura du soleil, nous ne nous équiperons pas contre le mauvais temps.

Nous sommes donc confrontés aux hypothèses suivantes :

hypothèse nulle H_0 : il va pleuvoir,
hypothèse alternative H_1 : il y aura du soleil.

Supposons que nous acceptions l'hypothèse de la pluie et que nous prenions donc notre parapluie.

S'il pleut véritablement, nous aurons pris la bonne décision, alors que s'il y a du soleil, nous aurons commis une erreur : celle d'avoir accepté une hypothèse fausse.

Dans le cas contraire, si nous refusons l'hypothèse de la pluie, nous aurons pris une bonne décision s'il y a du soleil, mais nous aurons commis une erreur s'il pleut : celle d'avoir rejeté une hypothèse vraie.

Nous pouvons représenter ces différents types d'erreur par le tableau suivant :

Situation	Décision	
	accepter H_0 (je prends le parapluie)	rejeter H_0 (je ne prends pas le parapluie)
H_0 vraie (il pleut)	bonne décision $1-\alpha$	erreur I α
H_0 fausse (il y a du soleil)	erreur II β	bonne décision $1-\beta$

La **probabilité** α de rejeter une hypothèse vraie s'appelle erreur de première espèce (erreur I dans le tableau).

La probabilité β d'accepter une hypothèse fausse s'appelle **erreur de seconde espèce** (erreur II dans le tableau).

MOTS CLÉS

Erreur de seconde espèce *(Type II error)*
Seuil de signification *(Significance level)*
Test d'hypothèse *(Hypothesis testing)*

RÉFÉRENCE

Neyman, J. and Pearson, E.S. (1928). On the use and interpretation of certain test criteria for purposes of statistical inference. I, II. *Biometrika*, 20A, pp. 175-240 and 263-294.

ERREUR DE SECONDE ESPÈCE

Type II error

Dans la réalisation d'un **test d'hypothèse**, on appelle **erreur** de seconde espèce β la **probabilité** d'accepter l'**hypothèse nulle** H_0 alors que celle-ci est fausse :

$$\beta = P(\text{accepter } H_0 \mid H_0 \text{ est fausse}).$$

HISTORIQUE

Voir **erreur de première espèce**.

DOMAINES ET LIMITATIONS

Si on note β l'**erreur** de seconde espèce, la **probabilité** de rejeter l'**hypothèse nulle** alors qu'elle est fausse est donnée par $1-\beta$:

$$1-\beta = P(\text{rejeter } H_0 \mid H_0 \text{ est fausse})$$

Cette probabilité $1-\beta$ est appelée la puissance du test.
Il faut donc toujours chercher à minimiser le risque d'erreur de seconde espèce, ce qui revient à chercher à augmenter la puissance du **test d'hypothèse**.

EXEMPLE

Voir **erreur de première espèce**.

MOTS CLÉS

Erreur de première espèce *(Type I error)*
Test d'hypothèse *(Hypothesis testing)*

ERREUR QUADRATIQUE MOYENNE

Mean squared error

Soit $\hat{\theta}$ un **estimateur** du **paramètre** θ. L'**erreur** quadratique en estimant θ par $\hat{\theta}$ est $\left(\theta - \hat{\theta}\right)^2$.

L'erreur quadratique moyenne (EQM) est l'**espérance mathématique** de cette quantité :

$$EQM(\hat{\theta}) = E\left[\left(\theta - \hat{\theta}\right)^2\right].$$

On peut montrer que l'**erreur** quadratique moyenne est égale à la somme de la **variance** de l'estimateur et du carré du **biais** :

$$EQM(\hat{\theta}) = Var\left(\hat{\theta}\right) + \left(E\left[\hat{\theta}\right] - \theta\right)^2.$$

Ainsi lorsque l'estimateur considéré est sans biais, l'erreur quadratique moyenne correspond exactement à la variance de l'estimateur.

HISTORIQUE

Au début du XIX^e siècle, **Pierre Simon de Laplace** (1820) et **Carl Friedrich Gauss** (1821) ont comparé l'**estimation** d'une quantité inconnue, basée sur des observations soumises à des erreurs aléatoires, avec un jeu de chance. Ils ont fait un parallèle entre l'erreur de la **valeur** estimée et la perte due à un tel jeu. C'est Gauss qui proposa comme mesure d'inexactitude le carré de l'erreur. Il reconnaissait que son choix était arbitraire mais il le justifiait par la simplicité mathématique de la fonction « élever au carré ».

ASPECTS MATHÉMATIQUES

Soient $X_1, X_2, \ldots, X_n$ un **échantillon** de taille n et $\hat{\theta}$ un **estimateur** du **paramètre** θ, l'écart (au carré) de l'**estimation** est donné par $\left(\theta - \hat{\theta}\right)^2$.

L'**erreur** quadratique moyenne est définie par l'**espérance mathématique** de ce dernier :

$$EQM\left(\hat{\theta}\right) = E\left[\left(\theta - \hat{\theta}\right)^2\right].$$

En anglais, on utilise la notation $MSE\left(\hat{\theta}\right)$ qui signifie *mean squared error*.

On peut montrer que :

$$EQM\left(\hat{\theta}\right) = Var\left(\hat{\theta}\right) + \left(E\left[\hat{\theta}\right] - \theta\right)^2.$$

c'est-à-dire que l'erreur quadratique moyenne est égale à la somme de la **variance** de l'estimateur et du carré du **biais**.

En effet :

$$\begin{aligned} EQM(\hat{\theta}) &= E\left[\left(\theta - \hat{\theta}\right)^2\right] \\ &= E\left[\left(\theta - E\left[\hat{\theta}\right] + E\left[\hat{\theta}\right] - \hat{\theta}\right)^2\right] \\ &= E\left[\left(\theta - E\left[\hat{\theta}\right]\right)^2 + \left(E\left[\hat{\theta}\right] - \hat{\theta}\right)^2 \right. \\ &\quad \left. +2 \cdot \left(\theta - E\left[\hat{\theta}\right]\right) \cdot \left(E\left[\hat{\theta}\right] - \hat{\theta}\right)\right] \\ &= E\left[\left(\theta - E\left[\hat{\theta}\right]\right)^2\right] + E\left[\left(E\left[\hat{\theta}\right] - \hat{\theta}\right)^2\right] \\ &\quad +2 \cdot \left(\theta - E\left[\hat{\theta}\right]\right) \cdot \left(E\left[\hat{\theta}\right] - E\left[\hat{\theta}\right]\right) \\ &= E\left[\left(E\left[\hat{\theta}\right] - \hat{\theta}\right)^2\right] + \left(\theta - E\left[\hat{\theta}\right]\right)^2 \\ &= Var\left(\hat{\theta}\right) + (biais)^2 \end{aligned}$$

où le biais est $E\left[\hat{\theta}\right] - \theta$.

Ainsi, lorsque l'estimateur $\hat{\theta}$ est sans biais, c'est-à-dire $E\left[\hat{\theta}\right] = \theta$, on obtient :

$$EQM\left(\hat{\theta}\right) = Var\left(\hat{\theta}\right).$$

L'efficacité d'un estimateur est inversement proportionnelle à son erreur quadratique moyenne.

On dit que l'erreur quadratique moyenne est consistante si

$$\lim_{n\to\infty} EQM\left(\hat{\theta}\right) = 0,$$

dans ce cas la variance et le biais de l'estimateur tendent vers 0 lorsque le nombre d'observations n tend vers l'infini.

DOMAINES ET LIMITATIONS

Certains auteurs préfèrent utiliser l'**écart moyen** en valeur absolue défini par $E\left[\left|\theta - \hat{\theta}\right|\right]$ où $\hat{\theta}$ est l'**estimateur** du **paramètre** θ. Le désavantage de ce dernier par rapport à l'**erreur** quadratique moyenne se trouve au niveau des propriétés mathématiques, car la fonction valeur absolue ne peut être dérivée partout.

EXEMPLE

On considère une **régression linéaire multiple** sous forme matricielle :

$$\mathbf{Y} = \mathbf{X} \cdot \boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

où

- $\mathbf{Y}$ est un **vecteur** d'ordre n (n observations de la **variable dépendante**),
- $\mathbf{X}$ est une **matrice** $n \times p$ donnée,
- $\boldsymbol{\beta}$ est le vecteur d'ordre p des coefficients à déterminer et

$\boldsymbol{\varepsilon}$ est le vecteur des erreurs que l'on suppose indépendantes et suivant chacune une **loi normale** de **moyenne** 0 et de **variance** σ^2.

Le vecteur $\hat{\boldsymbol{\beta}} = (\mathbf{X}' \cdot \mathbf{X})^{-1} \cdot \mathbf{X}' \cdot \mathbf{Y}$ donne un **estimateur** sans **biais** de $\boldsymbol{\beta}$.

On obtient alors :

$$EQM(\hat{\boldsymbol{\beta}}) = Var(\hat{\boldsymbol{\beta}}).$$

MOTS CLÉS

Biais *(Bias)*
Espérance mathématique *(Expected value)*
Estimateur *(Estimator)*
Variance *(Variance)*

RÉFÉRENCES

Gauss, C.F. (1821). In *Gauss's Work* (1803-1826) on the Theory of least Squares. Trans. H. F. Trotter. Statist. Techniques Res. Group. Tech. Rep. N° 5. Princeton University Press, Princeton NJ.

Laplace, P.S. de (1820). *Théorie analytique des probabilités.* 3e éd. Courcier, Paris.

ERREUR QUADRATIQUE MOYENNE TOTALE

Criterion of total mean squared error

L'erreur quadratique moyenne totale est un moyen de comparaison entre les estimations des paramètres d'un modèle, qu'il soit biaisé ou non.

ASPECTS MATHÉMATIQUES

Soit

$$\widehat{\boldsymbol{\beta}} = \left(\widehat{\beta}_1, \ldots, \widehat{\beta}_{p-1}\right)$$

un vecteur d'estimateurs des paramètres d'un modèle de régression. On définit l'erreur quadratique moyenne totale (en anglais, *total mean square error*, symbolisé par $EQMT$ en français et $TMSE$ en anglais) du vecteur $\widehat{\beta}$ d'estimateurs par la somme des erreurs quadratiques moyennes (EQM) de ses composantes.

On rappelle que

$$\begin{aligned} EQM\left(\widehat{\beta}_j\right) &= E\left(\left(\widehat{\beta}_j - \beta_j\right)^2\right) \\ &= V\left(\widehat{\beta}_j\right) + \left(E\left(\widehat{\beta}_j\right) - \beta_j\right)^2 \end{aligned}$$

avec $E(.)$ et $V(.)$, les notations usuelles pour l'**espérance** et la **variance**. On définit ainsi l'erreur quadratique moyenne totale par :

$$\begin{aligned} EQMT(\widehat{\beta}_j) &= \sum_{j=1}^{p-1} EQM\left(\widehat{\beta}_j\right) \\ &= \sum_{j=1}^{p-1} E\left(\left(\widehat{\beta}_j - \beta_j\right)^2\right) \\ &= \sum_{j=1}^{p-1} \left[Var\left(\widehat{\beta}_j\right) + \left(E\left(\widehat{\beta}_j\right) - \beta_j\right)^2\right] \\ &= (p-1)\sigma^2 \cdot Trace\,(V_j) \\ &\quad + \sum_{j=1}^{p-1} \left(E\left(\widehat{\beta}_j\right) - \beta_j\right)^2 ; \end{aligned}$$

où V_j désigne la **matrice de variances-covariances** de $\boldsymbol{\beta}$.

DOMAINES ET LIMITATIONS

Malheureusement, lorsqu'il s'agit de calculer l'erreur quadratique moyenne totale

$$EQMT\left(\widehat{\boldsymbol{\beta}}\right) = \sum_{j=1}^{p-1} E\left(\left(\widehat{\beta}_j - \beta_j\right)^2\right)$$

d'un vecteur d'estimateurs

$$\widehat{\boldsymbol{\beta}} = \left(\widehat{\beta}_1, \ldots, \widehat{\beta}_{p-1}\right)$$

des paramètres du modèle

$$Y_i = \beta_0 + \beta_1 X_{i1}^s + \ldots + \beta_{p-1} X_{ip-1}^s + \varepsilon_i,$$

il faut connaître les valeurs des paramètres β_j de ce modèle. Or ces paramètres sont évidemment inconnus pour un ensemble de données réelles. La notation X_j^s signifie que les données de la j-ème variable explicative ont été standardisées (voir **données standardisées**).

Pour parvenir tout de même à estimer ces $EQMT$, on définit un modèle artificiel jugé similaire au processus qui a généré les données réelles. En utilisant un générateur de nombres pseudo-aléatoires, on peut ensuite simuler un ensemble de données selon ce modèle artificiel, l'analyser selon plusieurs méthodes (régression simple, **régression ridge** etc.), et calculer ce que l'on appelle l'*erreur totale au carré* (en anglais, *total squared error*) symbolisé par ETC, des vecteurs d'estimateurs $\widehat{\beta}$ relatifs à chacune des méthodes :

$$ETC\left(\widehat{\boldsymbol{\beta}}\right) = \sum_{j=1}^{p-1}\left(\widehat{\beta}_j - \beta_j\right)^2.$$

On répète cette opération 100 fois, en s'assurant que les 100 ensembles de données sont pseudo-indépendants. La moyenne des 100 ETC obtenus de la sorte, pour chaque méthode, donne alors une bonne estimation de leurs $EQMT$. Si l'on est prêt à investir encore plus de temps, on peut répéter la même opération en choisissant pour approximer le processus sous-jacent aux données étudiées, une équation avec pour coefficients les estimateurs obtenus par la méthode de la sélection de variables ou obtenus par la méthode ridge. Notons que certains statisticiens préfèreront le modèle obtenu par la méthode de sélection de variables à cause de sa simplicité. En revanche, d'autres statisticiens choisiront plutôt la méthode ridge car elle prend en compte toute l'information disponible.

EXEMPLE

De manière générale on peut comparer dans le domaine d'une régression linéaire les méthodes suivantes : la **régression linéaire** par moindres carrés, la régression ridge ou encore la méthode de sélection des variables.

Dans l'exemple suivant, les observations sont relatives à treize fournées de ciment. Chaque fournée est composée de quatre ingrédients cités dans le tableau. Le but de l'expérience est de déterminer comment les quantités x_{i1}, x_{i2}, x_{i3} et x_{i4} de ces quatre ingrédients affectent la quantité y_i de chaleur émise par le durcissement du ciment.

Chaleur émise par le ciment

Fournée	Ingrédient				Chaleur
	1	2	3	4	
i	x_{i1}	x_{i2}	x_{i3}	x_{i4}	y_i
1	7	26	6	60	78,5
2	1	29	15	52	74,3
3	11	56	8	20	104,3
4	11	31	8	47	87,6
5	7	52	6	33	95,9
6	11	55	9	22	109,2
7	3	71	17	6	102,7
8	1	31	22	44	72,5
9	2	54	18	22	93,1
10	21	47	4	26	115,9
11	1	40	23	34	83,9
12	11	66	9	12	113,3
13	10	68	8	12	109,4

- y_i quantité de chaleur émise par le durcissement de la i-ème fournée (en joules) ;
- x_{i1} quantité de l'ingrédient 1 (aluminate de tricalcium) dans la i-ème fournée ;
- x_{i2} quantité de l'ingrédient 2 (silicate de tricalcium) dans la i-ème fournée ;
- x_{i3} quantité de l'ingrédient 3 (alumino-ferrite de tetracalcium) dans la i-ème fournée ;
- x_{i4} quantité de l'ingrédient 4 (silicate de dicalcium) dans la i-ème fournée.

Dans ce paragraphe, il sera introduit la comparaison entre les estimateurs par **moindres carrés** (MC), celui obtenu par la régression ridge (R) et enfin celui obtenu par la méthode de sélection des variables (SV) selon de critère de l'erreur quadratique moyenne totale $EQMT$. Les trois vecteurs d'estimation obtenus à partir de chaque méthode sont :

$$\begin{aligned}
\widehat{\mathbf{Y}}_{MC} &= 95,4 + 9,12\mathbf{X}_1{}^s + 7,94\mathbf{X}_2{}^s + 0,65\mathbf{X}_3{}^s \\
&\quad -2,41\mathbf{X}_4{}^s \\
\widehat{\mathbf{Y}}_{R} &= 95,4 + 7,64\mathbf{X}_1{}^s + 4,67\mathbf{X}_2{}^s - 0,91\mathbf{X}_3{}^s \\
&\quad -5,84\mathbf{X}_4{}^s \\
\widehat{\mathbf{Y}}_{SV} &= 95,4 + 8,64\mathbf{X}_1{}^s + 10,3\mathbf{X}_2{}^s.
\end{aligned}$$

On remarque que tous les estimations sont obtenues à partir des variables explicatives standardisées. On compare ces trois estimations selon le critère de

l'erreur quadratique moyenne totale $EQMT$ de ces trois vecteurs d'estimateurs, que l'on notera ici par :

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{MC} &= \left(\hat{\beta}_{MC_1}, \hat{\beta}_{MC_2}, \hat{\beta}_{MC_3}, \hat{\beta}_{MC_4}\right)' \\ \hat{\boldsymbol{\beta}}_{R} &= \left(\hat{\beta}_{R_1}, \hat{\beta}_{R_2}, \hat{\beta}_{R_3}, \hat{\beta}_{R_4}\right)' \\ \hat{\boldsymbol{\beta}}_{SV} &= \left(\hat{\beta}_{SV_1}, \hat{\beta}_{SV_2}, \hat{\beta}_{SV_3}, \hat{\beta}_{SV_4}\right)'\end{aligned}$$

correspondant respectivement à la méthode des moindres carrés, à la méthode ridge et à la méthode de la sélection de variables. Pour la méthode de sélection des variables, on considère comme étant égales à zéro, les estimations des coefficients pour les variables non sélectionnées dans le modèle. Dans notre cas, on a :

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{MC} &= (9,12\,;\,7,94\,;\,0,65\,;\,-2,41)' \\ \hat{\boldsymbol{\beta}}_{R} &= (7,64\,;\,4,67\,;\,-0,91\,;\,-5,84)' \\ \hat{\boldsymbol{\beta}}_{SV} &= (8,64\,;\,10,3\,;\,0,0)'.\end{aligned}$$

On a choisi pour approximer le processus sous-jacent aux données sur le ciment par l'équation suivante des moindres carrés :

$$\begin{aligned}Y_{iMC} = 95,4 &+ 9,12X^s_{i1} + 7,94X^s_{i2} \\ &+ 0,65X^s_{i3} - 2,41X^s_{i4} + \varepsilon_i.\end{aligned}$$

La procédure consiste à générer 100 fois 13 termes d'erreur aléatoires $\varepsilon_1, \ldots, \varepsilon_{13}$ selon une distribution normale de moyenne 0 et d'écart type 2,446 (on rappelle que 2,446 est l'estimateur des moindres carrés de σ pour les données sur le ciment). On calcule ensuite $Y_{1MC}, \ldots, Y_{13MC}$ en utilisant les valeurs des X^s_{ij} issues du tableau des données. On génère ainsi de la sorte 100 échantillons Y_{1MC}, ..., Y_{13MC} à partir de 100 échantillons ε_1, ..., ε_{13}.

Sur chacun de ces 100 échantillons Y_{1MC}, ..., Y_{13MC}, les trois méthodes sont appliquées (en utilisant toujours les mêmes valeurs pour les X^s_{ij}) fournissant ainsi 100 estimateurs $\hat{\boldsymbol{\beta}}_{\boldsymbol{MC}}$, $\hat{\boldsymbol{\beta}}_{\boldsymbol{R}}$ et $\hat{\boldsymbol{\beta}}_{\boldsymbol{SV}}$. On note que ces trois méthodes sont appliquées à ces 100 échantillons sans égard aux résultats des équations :

$$\begin{aligned}\widehat{Y}_{iMC} &= 95,4 + 9,12X^s_{i1} + 7,94X^s_{i2} + 0,65X^s_{i3} \\ &\quad -2,41X^s_{i4}, \\ \widehat{Y}_{iR} &= 95,4 + 7,64X^s_{i1} + 4,67X^s_{i2} - 0,91X^s_{i3} \\ &\quad -5,84X^s_{i4} \text{ et} \\ \widehat{Y}_{iSV} &= 95,4 + 8,64X^s_{i1} + 10,3X^s_{i2}\end{aligned}$$

obtenus pour l'échantillon original. En particulier, malgré le fait que la méthode de sélection des variables ait sélectionné les variables X_{i1} et X_{i2} dans $\widehat{Y}_{iSV} = 95,4 + 8,64X^s_{i1} + 10,3X^s_{i2}$, il est tout à fait possible que, pour l'un de ces 100 échantillons, la méthode sélectionne plutôt X_{i2} et X_{i3}, ou seulement X_{i3}, ou tout sous-ensemble des quatre variables disponibles. De même, malgré le fait que l'équation $\widehat{Y}_{iR} = 95,4 + 7,64X^s_{i1} + 4,67xX^s_{i2} - 0,91X^s_{i3} - 5,84X^s_{i4}$ ait été obtenue avec $k = 0,157$, la valeur de k est recalculée pour chacun de ces 100 échantillons (pour plus de détail se référer à l'exemple présenté pour la régression ridge).

À partir des 100 estimations $\hat{\beta}_{MC}$, $\hat{\beta}_{SV}$ et $\hat{\beta}_{R}$, on peut calculer 100 ETC pour chaque méthode, que l'on note respectivement par :

$$\begin{aligned}ETC_{MC} &= \left(\hat{\beta}_{MC_1} - 9,12\right)^2 + \left(\hat{\beta}_{MC_2} - 7,94\right)^2 \\ &\quad + \left(\hat{\beta}_{MC_3} - 0,65\right)^2 + \left(\hat{\beta}_{MC_4} + 2,41\right)^2 \\ ETC_{R} &= \left(\hat{\beta}_{R_1} - 9,12\right)^2 + \left(\hat{\beta}_{R_2} - 7,94\right)^2 \\ &\quad + \left(\hat{\beta}_{R_3} - 0,65\right)^2 + \left(\beta_{R_4} + 2,41\right)^2 \\ ETC_{SV} &= \left(\hat{\beta}_{SV_1} - 9,12\right)^2 + \left(\hat{\beta}_{SV_2} - 7,94\right)^2 \\ &\quad + \left(\hat{\beta}_{SV_3} - 0,65\right)^2 + \left(\hat{\beta}_{SV_4} + 2,41\right)^2.\end{aligned}$$

Les moyennes des 100 valeurs de ETC_{MC}, ETC_R et ETC_{SV} sont alors des estimations de $EQMT_{MC}$, $EQMT_{SV}$ et $EQMT_R$, c'est-à-dire des estimations des $EQMT$ pour les trois méthodes considérées.

Après avoir effectué une telle simulation, les estimations suivantes ont été obtenues :

$$\begin{aligned}EQMT_{MC} &= 270, \\ EQMT_{R} &= 75, \\ EQMT_{SV} &= 166.\end{aligned}$$

On a donc les différences :

$$\begin{aligned} EQMT_{MC} - EQMT_{SV} &= 104, \\ EQMT_{MC} - EQMT_{R} &= 195, \\ EQMT_{SV} - EQMT_{R} &= 91. \end{aligned}$$

Comme les écarts types des 100 différences observées sont respectivement 350, 290 et 280, on peut calculer des intervalles de confiance approximatifs de 95 % pour les différences entre les $EQMT$ de deux méthodes de la sorte :

$$\begin{aligned} EQMT_{MC} - EQMT_{SV} &= 104 \pm \frac{2 \cdot 350}{\sqrt{100}}, \\ EQMT_{MC} - EQMT_{R} &= 195 \pm \frac{2 \cdot 290}{\sqrt{100}}, \\ EQMT_{SV} - EQMT_{R} &= 91 \pm \frac{2 \cdot 280}{\sqrt{100}}. \end{aligned}$$

On obtient :

$$\begin{aligned} 34 &< EQMT_{MC} - EQMT_{SV} < 174, \\ 137 &< EQMT_{MC} - EQMT_{R} < 253, \\ 35 &< EQMT_{SV} - EQMT_{R} < 147. \end{aligned}$$

On peut donc conclure, du moins en ce qui concerne le modèle particulier utilisé pour générer les données simulées, et compte tenu de notre objectif d'obtenir de petits $EQMT$, que la méthode ridge est la meilleure des trois méthodes considérées, suivie en deuxième position par la procédure de sélection de variables.

MOTS CLÉS

Biais *(Bias)*
Données standardisées *(Standardised data)*
Erreur quadratique moyenne *(mean squared error)*
Espérance *(Expected value)*
Matrice de variance-covariance *(Variance-covariance matrix)*
Matrice H (*Hat matrix)*
Moindres carrés pondérés, méthode des *(Weighted leasted squares method)*
Régression ridge *(Ridge regression)*
Variance *(Variance)*

RÉFÉRENCES

Box. G.E.P et Draper, N.R. (1959). A basis for the selection of reponse surface design. *Journal of the American Statistical Association*, 54, pp. 622-654.

Dodge, Y. (1999). *Analyse de régression appliquée.* Dunod, Paris.

ERREUR TYPE

Standard error

L'**erreur type** est la racine carrée de la **variance** estimée d'une **statistique**, c'est-à-dire l'**écart type** de la **distribution d'échantillonnage** de cette statistique.

HISTORIQUE

La notion d'erreur type est attribuée à **Carl Friedrich Gauss** (1816), alors qu'il ne semblait pas connaître le concept de **distribution d'échantillonnage** d'une **statistique** utilisée pour estimer la **valeur** d'un **paramètre** de la **population**.

ASPECTS MATHÉMATIQUES

L'**erreur** type de la **moyenne**, ou **écart type** de la **distribution d'échantillonnage** des moyennes, noté $\sigma_{\bar{x}}$, se calcule en fonction de l'écart type de la **population**, et de la taille respective de la population finie et de l'**échantillon** :

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N}}$$

où

- σ est l'écart type de la population,
- N la taille de la population et
- n la taille de l'échantillon.

Si la taille N de la population est infinie, le facteur de correction :

$$\sqrt{\frac{N-n}{N}}$$

peut être omis puisque celui-ci tend vers 1 lorsque N tend vers l'infini.

Si la population est infinie, ou si l'échantillonnage est non exhaustif (avec remise), l'erreur type de la moyenne est donc égale à l'écart type de la population divisé par la racine carrée de la taille de l'échantillon :

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}.$$

Si l'écart type σ de la population est inconnu, il peut être estimé par l'écart type S de l'échantillon donné par la racine carrée de la **variance** S^2 :

$$S^2 = \frac{\sum_{j=1}^{n} (x_j - \bar{x})^2}{n-1}.$$

Les mêmes formules sont valables pour l'erreur type d'un **pourcentage**, ou écart type de la distribution d'échantillonnage des pourcentages, noté σ_p :

$$\sigma_p = \frac{\sigma}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N}}$$

dans le cas d'une population finie et

$$\sigma_p = \frac{\sigma}{\sqrt{n}}$$

dans le cas d'une population infinie, où σ est l'écart type de la population :

$$\begin{aligned} \sigma &= \sqrt{p \cdot (1-p)} \\ &= \sqrt{p \cdot q}, \end{aligned}$$

p étant la **probabilité** qu'un élément possède le caractère étudié, et $q = 1 - p$ la probabilité qu'il ne le possède pas.

L'erreur type de la différence entre deux quantités indépendantes est la racine carrée de la somme des erreurs types au carré des deux quantités.

Par exemple, l'erreur type de la différence entre deux moyennes est égale à :

$$\begin{aligned} \sigma_{\bar{x}_1 - \bar{x}_2} &= \sqrt{\sigma^2_{\bar{x}_1} + \sigma^2_{\bar{x}_2}} \\ &= \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \end{aligned}$$

où σ_1^2 et σ_2^2 sont les variances respectives des deux populations infinies à comparer et n_1, n_2 les tailles respectives des deux échantillons.

Si les variances des deux populations sont inconnues, nous pouvons les estimer par les variances calculées sur les deux échantillons, ce qui nous donne :

$$\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$$

où S_1^2 et S_2^2 sont les variances respectives des deux échantillons.

Si nous considérons que les variances des deux populations sont égales, nous pouvons estimer la valeur de cette variance par une **variance pondérée** S_p^2 calculée en fonction de S_1^2 et S_2^2 :

$$S_p^2 = \frac{(n_1 - 1) \cdot S_1^2 + (n_2 - 1) \cdot S_2^2}{n_1 + n_2 - 2},$$

ce qui nous donne :

$$\sigma_{\bar{x}_1 - \bar{x}_2} = S_p \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}.$$

L'erreur type de la différence entre deux pourcentages se calcule de la même manière que l'erreur type de la différence entre deux moyennes. Nous aurons donc :

$$\begin{aligned} \sigma_{p_1 - p_2} &= \sqrt{\sigma^2_{p_1} + \sigma^2_{p_2}} \\ &= \sqrt{\frac{p_1 \cdot (1-p_1)}{n_1} + \frac{p_2 \cdot (1-p_2)}{n_2}} \end{aligned}$$

où p_1 et p_2 sont les pourcentages pour des populations infinies calculés sur les deux échantillons de taille n_1 et n_2.

EXEMPLES

1. *Erreur type d'une moyenne*

Un fabricant veut tester la précision d'une nouvelle machine qui devrait lui fournir des boulons de 8 mm de diamètre. Sur un lot de $N = 10\,000$ pièces, il examine un **échantillon** de $n = 100$ pièces dont l'**écart type** S est de $1,2$ mm.

L'erreur type de la **moyenne** est égale à :

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N}}.$$

L'écart type σ de la **population** étant inconnu, nous pouvons l'estimer par l'écart type de l'échantillon S, ce qui nous donne :

$$\begin{aligned} \sigma_{\bar{x}} &= \frac{S}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N}} \\ &= \frac{1,2}{10} \cdot \sqrt{\frac{10\,000-100}{10\,000}} \\ &= 0,12 \cdot 0,995 = 0,119. \end{aligned}$$

Nous pouvons remarquer que le facteur

$$\sqrt{\frac{N-n}{N}} = 0,995$$

n'a que peu d'influence sur le résultat, la taille de la population étant suffisamment grande.

2. *Erreur type d'un pourcentage*

Un candidat aux élections réalise un sondage sur un échantillon de 200 personnes pour savoir s'il bénéficiera de plus de 50 % ($= \pi_0$) des voix, π_0 étant la valeur présumée du **paramètre** π (**pourcentage** de la population).

L'erreur type du pourcentage σ_p est égal à :

$$\sigma_p = \sqrt{\frac{\pi_0 \cdot (1-\pi_0)}{n}},$$

n étant la taille de l'échantillon (200 dans notre exemple). Nous pouvons considérer que la population est infinie, c'est pourquoi nous ne tenons pas compte ici du facteur correctif.

$$\sigma_p = \sqrt{\frac{0,5 \cdot (1-0,5)}{200}} = 0,0354.$$

3. *Erreur type de la différence entre deux quantités*

Une compagnie d'assurances a décidé d'équiper ses bureaux de micro-ordinateurs. Elle désire acheter ces micro-ordinateurs à deux fournisseurs différents pour autant qu'il n'y ait pas de différence significative de durabilité entre les deux marques. Elle teste donc un échantillon de 35 micro-ordinateurs de la marque 1 et 32 de la marque 2, en relevant le temps écoulé avant la première panne. Les données observées sont les suivantes :

Marque 1 : écart type $S_1 = 57,916$;
Marque 2 : écart type $S_2 = 57,247$.

Cas 1 *:*

Nous supposons que les **variances** des deux populations sont égales. L'écart type pondéré est égal à :

$$\begin{aligned} S_p &= \sqrt{\frac{(n_1-1) \cdot S_1^2 + (n_2-1) \cdot S_2^2}{n_1+n_2-2}} \\ &= \sqrt{\frac{34 \cdot 57,916^2 + 31 \cdot 57,247^2}{35+32-2}} \\ &= 57,5979. \end{aligned}$$

Connaissant l'écart type pondéré, nous pouvons calculer la **valeur** de l'écart type de la **distribution d'échantillonnage** par la formule suivante :

$$\begin{aligned} \sigma_{\bar{x}_1-\bar{x}_2} &= S_p \cdot \sqrt{\frac{1}{n_1}+\frac{1}{n_2}} \\ &= 57,5979 \cdot \sqrt{\frac{1}{35}+\frac{1}{32}} = 14,0875. \end{aligned}$$

Cas 2 *:*

Si nous considérons les variances des deux populations comme inégales, nous devons commencer par calculer l'erreur type de la moyenne pour chaque échantillon :

$$\sigma_{\bar{x}_1} = \frac{S_1}{\sqrt{n_1}} = \frac{57,916}{\sqrt{35}} = 9,79$$

$$\sigma_{\bar{x}_2} = \frac{S_2}{\sqrt{n_2}} = \frac{57,247}{\sqrt{32}} = 10,12$$

$$\begin{aligned} \sigma_{\bar{x}_1-\bar{x}_2} &= \sqrt{\sigma_{\bar{x}_1}^2 + \sigma_{\bar{x}_2}^2} \\ &= \sqrt{9,79^2 + 10,12^2} \\ &= 14,08. \end{aligned}$$

Nous pouvons constater dans cet exemple que les deux cas nous donnent une valeur de l'erreur type pratiquement identique. Ceci s'explique par la très faible différence entre les écarts types des deux échantillons S_1 et S_2.

MOTS CLÉS

Distribution d'échantillonnage *(Sampling distribution)*
Écart type *(Standard deviation)*
Test d'hypothèse *(Hypothesis testing)*

RÉFÉRENCES

Box, G.E.P. and Hunter, W.G. and Hunter, J.S. (1978). *Statistics for experimenters. An introduction to design, data analysis, and model building.* John Wiley & Sons, New York.

Cochran, W.G. (1963). *Sampling Techniques*, 2nd ed. John Wiley & Sons, New York.

Gauss, C.F. (1816). Bestimmung der Genauigkeit der Beobachtungen. Vol. 4, pp. 109-117. In *C.F. Gauss Werke* (publié en 1880). Dieterichsche Universitäts-Druckerei, Göttingen.

Gauss, C.F. (1957). *Gauss' Works (1803-1826) on the Theory of Least Squares.* Traduit par H.F. Trotter. Statistical Techniques Research Group, Technical Report, N° 5, Princeton University Press, Princeton NJ.

ESPÉRANCE MATHÉMATIQUE

Expected value

L'espérance mathématique d'une **variable aléatoire** est la **moyenne** pondérée des valeurs que la variable aléatoire peut prendre, les poids étant les probabilités avec lesquelles ces valeurs peuvent être prises.

HISTORIQUE

Le principe de l'espérance mathématique est apparu pour la première fois dans l'ouvrage intitulé *De Ratiociniis in Aleae Ludo* publié en 1657 par le scientifique hollandais Christiaan Huygens (1629-1695). Sa réflexion sur le sujet semble avoir eu passablement d'influence sur les travaux de Pascal et Fermat relatifs aux probabilités.

ASPECTS MATHÉMATIQUES

Selon que la **variable aléatoire** est discrète ou continue, on parlera de l'espérance mathématique d'une variable aléatoire discrète ou de l'espérance mathématique d'une variable aléatoire continue.

Soit une variable aléatoire discrète X dont la **fonction de probabilité** est $p(X)$, l'espérance mathématique de X, notée $E[X]$ ou μ, est définie par :

$$\mu = E[X] = \sum_{i=1}^{n} x_i p(x_i),$$

si X peut prendre n valeurs.

Si la variable aléatoire X est continue, l'espérance mathématique devient :

$$\mu = E[X] = \int_D x f(x)\, dx,$$

si X prend ses valeurs sur un **intervalle** D, où $f(x)$ est la **fonction de densité** de X.

Propriétés de l'espérance mathématique

Soient a et b deux constantes et X et Y deux variables aléatoires. Nous avons alors :

1. $E[aX + b] = aE[X] + b$;
2. $E[X + Y] = E[X] + E[Y]$;
3. $E[X - Y] = E[X] - E[Y]$;
4. Si X et Y sont indépendantes, alors :

$$E[X \cdot Y] = E[X] \cdot E[Y].$$

EXEMPLES

On présente deux exemples, l'un concernant une **variable aléatoire** discrète, l'autre une variable aléatoire continue.

On considère le jeu aléatoire suivant : on lance un dé plusieurs fois de suite. On suppose que l'on gagne 1 euro si le résultat obtenu est pair, 2 euros si le résultat est 1 ou 3 et que l'on perde 3 euros si le résultat est 5.

La variable aléatoire X considérée est le nombre d'euros gagnés ou perdus. Le tableau ci-dessous représente les différentes valeurs de X et leur **probabilité** respective :

X	-3	1	2
$P(X)$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

L'espérance mathématique est égale à :

$$\begin{aligned} E[X] &= \sum_{i=1}^{3} x_i p(x_i) \\ &= -3 \cdot \frac{1}{6} + 1 \cdot \frac{3}{6} + 2 \cdot \frac{2}{6} \\ &= \frac{2}{3}. \end{aligned}$$

En d'autres termes, le joueur gagne en moyenne $\frac{2}{3}$ d'euro par lancer.

Soit une variable aléatoire continue X dont la **fonction de densité** est donnée par :

$$f(x) = \begin{cases} 1 & \text{pour } 0 < x < 1, \\ 0 & \text{sinon.} \end{cases}$$

L'espérance mathématique de cette variable aléatoire X est égale à :

$$\begin{aligned} E[X] &= \int_0^1 x \cdot 1 dx \\ &= \left. \frac{x^2}{2} \right|_0^1 \\ &= \frac{1}{2}. \end{aligned}$$

MOTS CLÉS

Fonction de densité *(Density function)*
Fonction de probabilité *(Probability function)*
Moyenne arithmétique pondérée *(Weighted arithmetic mean)*

ESTIMATEUR

Estimator

On appelle estimateur toute fonction statistique d'un **échantillon** utilisée pour estimer un **paramètre** inconnu de la **population**.

Toute **valeur** de cet estimateur est une **estimation** du paramètre de la population.

HISTORIQUE

Voir **estimation**.

ASPECTS MATHÉMATIQUES

Soit θ un **paramètre** inconnu défini au sein d'une **population** et soit $(X_1, X_2, \ldots, X_n)$ un **échantillon** tiré de cette population. Pour estimer θ, on utilise une fonction statistique (ou simplement une **statistique**) de cet échantillon $g(X_1, \ldots, X_n)$.

De manière à distinguer la statistique $g(X_1, \ldots, X_n)$, qui est une **variable aléatoire**, de la **valeur** qu'elle prend dans un cas particulier, l'usage est de qualifier la statistique $g(X_1, \ldots, X_n)$ d'estimateur et la valeur, prise dans un cas particulier, d'**estimation**.

Un estimateur ne doit pas être confondu avec une méthode d'estimation.

On note généralement un paramètre d'une population par une lettre grecque et son estimation par cette même lettre accompagnée d'un accent circonflexe (ˆ) pour le distinguer du paramètre de la population. On trouve parfois aussi la lettre romaine correspondante.

DOMAINES ET LIMITATIONS

Lors du calcul d'un estimateur, comme par exemple celui de la **moyenne arithmétique**, on s'attend à ce que la **valeur** de cet estimateur soit d'autant plus proche de la valeur du vrai **paramètre**, c'est-à-dire la **moyenne** de la **population** dans ce cas, à mesure que la **taille d'échantillon** augmente.

Il est intuitivement clair qu'un bon estimateur doit avoir la propriété d'être proche du vrai paramètre. De ce point de vue, les estimateurs peuvent être qualifiés d'autant meilleurs qu'ils se concentrent plus autour des vrais paramètres.

Dans ce sens, les bons estimateurs sont fortement liés aux **mesures de dispersion** statistiques telle que la **variance**.

Pour être de bons estimateurs, ces derniers doivent posséder certaines qualités telles que celles décrites ci-dessous :

1. *Estimateur sans biais*

On dit d'un estimateur $\hat{\theta}$ d'un paramètre inconnu θ qu'il est sans biais si son **espérance mathématique** est égale à θ,

$$E\left[\hat{\theta}\right] = \theta.$$

La moyenne $\bar{x}$ et la variance S^2 d'un **échantillon** sont des estimateurs sans biais respectivement de la moyenne μ et de la variance σ^2 de la population, avec

$$\bar{x} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i \text{ et } S^2 = \frac{1}{n-1} \cdot \sum_{i=1}^{n} (x_i - \bar{x})^2 .$$

2. *Estimateur efficace*

On dit d'un estimateur $\hat{\theta}$ d'un paramètre θ qu'il est efficace si la variance de $\hat{\theta}$ est plus petite que celle de tout autre estimateur de θ.

Ainsi, de deux estimateurs sans biais de θ, l'un sera plus efficace que l'autre si sa variance est plus petite.

3. *Estimateur consistant*

On dit d'un estimateur $\hat{\theta}$ d'un paramètre θ qu'il est consistant si la **probabilité** qu'il diffère de θ décroît à mesure que la taille d'échantillon augmente.

Ainsi, un estimateur est consistant si, quelque soit un nombre $\varepsilon > 0$ arbitrairement petit, on a :

$$\lim_{n \to \infty} P\left(\left|\hat{\theta} - \theta\right| < \varepsilon\right) = 1.$$

EXEMPLE

Soit $(x_1, x_2, \ldots, x_n)$ un **échantillon** issu d'une **population** de **moyenne** μ et de **variance** σ^2.

L'estimateur $\bar{x}$ calculé sur l'échantillon et utilisé pour estimer le **paramètre** μ de la population est un estimateur sans biais. En effet, on peut montrer que :

$$\begin{aligned} E[\bar{x}] &= E\left[\frac{1}{n}\sum_{i=1}^{n} x_i\right] \\ &= \frac{1}{n} E\left[\sum_{i=1}^{n} x_i\right] \\ &= \frac{1}{n}\sum_{i=1}^{n} E[x_i] \\ &= \frac{1}{n}\sum_{i=1}^{n} \mu \\ &= \frac{1}{n} n \cdot \mu \\ &= \mu. \end{aligned}$$

Pour estimer la variance σ^2 de la population, on est tenté d'utiliser l'estimateur S^2 calculé sur l'échantillon $(x_1, x_2, \ldots, x_n)$ et qui est défini par :

$$S^2 = \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^2 .$$

Or cet estimateur est biaisé. En effet, on peut montrer que $E\left[S^2\right]$ est différent de σ^2.

On montre d'abord que :

$$\sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} (x_i - \mu)^2 - n(\bar{x} - \mu)^2$$

de la façon suivante :

$$\begin{aligned} &\sum_{i=1}^{n} (x_i - \bar{x})^2 = \\ &= \sum_{i=1}^{n} [(x_i - \mu) - (\bar{x} - \mu)]^2 \\ &= \sum_{i=1}^{n} [(x_i - \mu)^2 - 2(x_i - \mu)(\bar{x} - \mu) \end{aligned}$$

$$
\begin{aligned}
& +(\bar{x}-\mu)^2] \\
= & \sum_{i=1}^{n}(x_i-\mu)^2 - 2(\bar{x}-\mu)\cdot\sum_{i=1}^{n}(x_i-\mu) \\
& +n(\bar{x}-\mu)^2 \\
= & \sum_{i=1}^{n}(x_i-\mu)^2 - 2(\bar{x}-\mu)\left(\sum_{i=1}^{n}x_i - n\cdot\mu\right) \\
& +n(\bar{x}-\mu)^2 \\
= & \sum_{i=1}^{n}(x_i-\mu)^2 - 2(\bar{x}-\mu)(n\cdot\bar{x}-n\cdot\mu) \\
& +n(\bar{x}-\mu)^2 \\
= & \sum_{i=1}^{n}(x_i-\mu)^2 - 2n(\bar{x}-\mu)(\bar{x}-\mu) \\
& +n(\bar{x}-\mu)^2 \\
= & \sum_{i=1}^{n}(x_i-\mu)^2 - 2n(\bar{x}-\mu)^2 + n(\bar{x}-\mu)^2 \\
= & \sum_{i=1}^{n}(x_i-\mu)^2 - n(\bar{x}-\mu)^2 .
\end{aligned}
$$

On trouve alors :

$$
\begin{aligned}
E\left[S^2\right] &= E\left[\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2\right] \\
&= E\left[\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2 - (\bar{x}-\mu)^2\right] \\
&= \frac{1}{n}E\left[\sum_{i=1}^{n}(x_i-\mu)^2\right] - E\left[(\bar{x}-\mu)^2\right] \\
&= \frac{1}{n}\sum_{i=1}^{n}E\left[(x_i-\mu)^2\right] - E\left[(\bar{x}-\mu)^2\right] \\
&= Var(x_i) - Var(\bar{x}) \\
&= \sigma^2 - \frac{\sigma^2}{n} \\
&= \frac{n-1}{n}\sigma^2.
\end{aligned}
$$

Par conséquent, pour que l'estimateur de la variance σ^2 soit non biaisé, il ne faut pas prendre S^2 mais il faut le diviser par $\frac{n-1}{n}$:

$$
\begin{aligned}
S'^2 &= \frac{n}{n-1}S^2 \\
&= \frac{n}{n-1}\cdot\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2 \\
&= \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2 .
\end{aligned}
$$

MOTS CLÉS

Biais *(Bias)*
Échantillon *(Sample)*
Échantillonnage *(Sampling)*
Estimation *(Estimation)*

RÉFÉRENCES

Hogg, R.V. and Craig, A.T. (1965). *Introduction to mathematical statistics.* 2th ed. The Macmillan Company, New York.

Lehmann, E.L. (1983). *Theory of Point Estimation*, 2nd ed. John Wiley & Sons, New York.

ESTIMATION

Estimation

On appelle estimation le procédé par lequel on cherche à déterminer la **valeur** d'un **paramètre** d'une **population**. Pour estimer le paramètre, on forme un **échantillon** à partir de la population et on calcule la valeur de l'**estimateur** correspondant au paramètre inconnu.

L'estimation se divise en deux grandes **catégories** : l'**estimation ponctuelle** et l'estimation par **intervalle de confiance**.

HISTORIQUE

L'idée d'estimation remonte aux premiers écrits de la **statistique** mathématique, que l'on rencontre notamment chez **Jakob Bernoulli** (1713), **Pierre Simon de Laplace** (1774) et Daniel Bernoulli (1778).

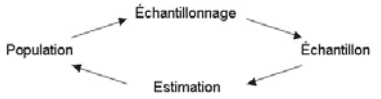
Le plus grand progrès dans la théorie de l'estimation, après l'introduction de la **méthode des moindres**

carrés, est peut-être la **méthode des moments** formulée par **Karl Pearson** (1894, 1898). Mais c'est à **Ronald Aylmer Fisher** que l'on doit les fondements de la théorie de l'estimation. En particulier, dans son premier texte de 1912, il introduisit la méthode du **maximum de vraisemblance**. C'est en 1922 avec le texte des fondements que R.A. Fisher pose pour la première fois et de manière claire ce qu'est l'estimation.

R.A. Fisher (1925) a également introduit un ensemble de définitions qui sont aujourd'hui adoptées pour décrire les estimateurs ; c'est à lui que l'on doit les termes tels que estimateurs biaisés, efficaces ou suffisants introduits dans la théorie de l'estimation.

DOMAINES ET LIMITATIONS

Le graphique suivant montre la relation entre échantillonnage et estimation. L'« échantillonnage » est le passage de la population à l'échantillon, et l'« estimation » est le passage inverse de l'échantillon à la population.



MOTS CLÉS

Échantillon *(Sample)*
Échantillonnage *(Sampling)*
Estimateur *(Estimator)*
Estimation L1 *(L1 estimation)*
Estimation ponctuelle *(Point estimation)*
Estimation robuste *(Robust estimation)*
Intervalle de confiance *(Confidence interval)*

RÉFÉRENCES

Armatte, M. (1988). La construction des notions d'estimation et de vraisemblance chez Ronald A. Fisher. In *Estimation et sondages.* J. Mairesse. *Economica*, Paris.

Bernoulli, D. (1778). The most probable choice between several discrepant observations, traduction et commentaires de E.S. Pearson et M.G. Kendall, (1970). *Studies in the History of Statistics and Probability.* Charles Griffin & Company Ltd.

Bernoulli, J. (1713). *Ars Conjectandi, Opus Posthumum. Accedit Tractatus de Seriebus infinitis, et Epistola Gallice scripta de ludo Pilæ recticularis.* Basileæ, Impensis Thurnisiorum, fratrum.

Fisher, R.A. (1912). On an absolute criterium for fitting frequency curves. *Messenger of Mathematics*, 41, pp. 155-160.

Fisher, R.A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London.* Series A, Vol. 222, pp. 309-368.

Fisher, R.A. (1925). Theory of statistical Estimation. *Proceedings of the Cambridge Philosophical Society*, 22, pp. 700-725.

Gauss, C.F. (1809). *Theoria motus corporum cœlestium.* Werke 7.

Laplace, P.S. de (1774). *Mémoire sur la probabilité des causes par les événements.* Mémoire de l'Académie royale des sciences présenté par divers savants, 6, pp. 621-656 ou Vol. 8, pp. 27-65. In P.S. de Laplace, *Œuvres complètes* (1891). Gauthier-Villars, Paris.

Laplace, P.S. de (1812). *Théorie analytique des probabilités.* Courcier, Paris.

Legendre, A.M. (1805). *Nouvelles méthodes pour la détermination des orbites des comètes*, Courcier, Paris.

Pearson, K. (1894). 1948. Contributions to the Mathematical Theory of Evolution. I. pp. 1-40 in Karl Pearson, *Karl Pearson's Early Statistical Papers.* Cambridge University Press, Cambridge. First published as: On the Dissection of Asymmetrical Frequency Curves, *Philosophical Transactions of the Royal Society of London.* Series A, Vol. 185, pp. 71-110.

Pearson, K. and Filon, L.N.G (1898). Mathematical Contributions to the Theory of Evolution. IV. On the Probable Errors of Frequency Constants and on the Influence of Random Selection on Variation and Correlation. *Philosophical Transactions of the Royal Society of London.* Series A, Vol. 191, pp. 229-311.

ESTIMATION L1

L1 estimation

La méthode d'estimation L1 est une alternative à celle des **moindres carrés** utilisée en régression linéaire pour l'**estimation** des paramètres d'un **modèle**. Au lieu de minimiser la somme des carrés des résidus, on rend minimale la somme des valeurs absolues de ces mêmes résidus.

Les estimateurs L1 présentent l'avantage de ne pas être aussi sensibles aux effets des points éloignés que ceux des moindres carrés.

HISTORIQUE

L'utilisation de l'estimation L1, sous forme très élémentaire, remonte à Galilée en 1632, qui l'utilisa pour la détermination de la **distance** de la terre à une étoile qui venait d'être découverte.

Roger Joseph Boscovich (1757) proposa deux critères pour l'ajustement d'une droite :

- les sommes respectives des résidus positifs et négatifs doivent être égales, la droite optimale passe par le centroïde des observations ;
- la somme des résidus en valeur absolue doit être minimale.

R.J. Boscovich donna une méthode géométrique pour résoudre les équations fournies par ses critères.

La procédure analytique pour l'**estimation** des paramètres pour les deux critères de R.J. Boscovich est due à **Pierre Simon de Laplace** (1793). **Francis Ysidro Edgeworth** (1887) supprima la première condition disant que la somme des résidus doit être nulle, et développa une méthode qui permet de traiter le cas de la **régression linéaire multiple**.

Durant la première moitié du XX^{e} siècle, les statisticiens ne s'intéressèrent guère à la régression linéaire basée sur l'estimation L1. Les raisons principales en étaient :

- les difficultés numériques, dues à l'absence de formule fermée ;
- l'absence de théorie asymptotique pour l'estimation L1 dans le **modèle** de régression ;
- l'estimation par la méthode des **moindres carrés** est satisfaisante pour de petits échantillons et nous ne sommes pas certains que l'estimation L1 soit meilleure.

À partir de 1970, le domaine de l'estimation L1 devint le pôle de recherche pour une multitude de statisticiens et de nombreux ouvrages furent publiés. Citons, en particulier, Basset et Koenker (1978) qui développèrent la théorie asymptotique des estimateurs L1.

Dès 1987, des congrès annuels consacrés à ces problèmes (International Conference on Statistical Data Analysis based on the L1-norm and Related Methods) ont eu lieu à Neuchâtel (Suisse). Le dernier en date s'est déroulé en 2002 : The Fourth International Conference on Statistical Data Analysis based on the L1-norm and Related Methods.

ASPECTS MATHÉMATIQUES

Dans le cas *d'un modèle à un paramètre*, la méthode d'**estimation** L1 consiste à estimer le **paramètre** β_0 dans le **modèle**

$$y_i = \beta_0 + \varepsilon_i$$

de manière à ce que la somme des erreurs en valeur absolue soit minimale :

$$\min_{\beta_0} \sum_{i=1}^{n} |\varepsilon_i| = \min_{\beta_0} \sum_{i=1}^{n} |y_i - \beta_0| \,.$$

Ce problème revient donc à trouver β_0 tel que la somme $\sum_{i=1}^{n} |y_i - \beta_0|$ soit minimale.

La valeur de β_0 qui rend cette somme minimale est la **médiane** de l'**échantillon** unidimensionnel $y_1, y_2, \ldots, y_n$.

Dans le cas d'un modèle à deux paramètres :

$$y_i = \beta_0 + \beta_1 \cdot x_i + \varepsilon_i,$$

les paramètres β_0 et β_1 estimés par la méthode d'**estimation** L1 sont généralement calculés par un **algorithme** itératif permettant de minimiser

$$\sum_{i=1}^{n} |y_i - \beta_0 - \beta_1 \cdot x_i| .$$

C'est finalement grâce aux techniques de **programmation linéaire** qu'il est aujourd'hui possible de calculer relativement facilement les estimateurs L1, en particulier dans le modèle de **régression linéaire multiple**.

DOMAINES ET LIMITATIONS

Le problème de l'estimation L1 réside dans l'absence de formule générale pour les estimations des paramètres à l'inverse de ce qui existe avec la méthode des **moindres carrés**.

Pour un ensemble de données unidimensionnelles, la **mesure de tendance centrale** déterminée par l'estimation L1 est la **médiane** qui est par définition la **valeur** qui minimise la somme des écarts pris en valeur absolue.

MOTS CLÉS

Écart moyen *(Mean deviation)*
Erreur *(Error)*
Estimateur *(Estimator)*
Estimation *(Estimation)*
Estimation robuste *(Estimation robuste)*
Moindres carrés *(Least squares)*
Paramètre *(Parameter)*
Résidu *(Residual)*

RÉFÉRENCES

Basset, G. and Koenker, R. (1978). Asymptotic Theory of Least absolute Error Regression. *Journal of the American Statistical Association*, 73, pp. 618-622.

Boscovich, R.J. (1757). *De Litteraria Expeditione per Pontificiam ditionem, et Synopsis amplioris Operis, ac habentur plura eius ex exemplaria etiam sensorum impressa.* Bononiensi Scientiarium et Artium Instituto Atque Academia Commentarii, Tomus IV, pp. 353-396.

Dodge, Y. éditeur (1987). *Statistical Data Analysis Based on the L1-norm and Related Methods.* Elsevier Science Publishers B.V.

Edgeworth F.Y. (1887). A New Method of Reducing Observations Relating to Several Quantities. *Philosophical Magazine* (5^{th} series), Vol. 24, pp. 222-223.

ESTIMATION PONCTUELLE

Point estimation

L'estimation ponctuelle d'un **paramètre** de la **population** permet d'obtenir une **valeur** unique calculée à partir d'un **échantillon**, valeur qui sera prise comme **estimation** de ce paramètre inconnu.

ASPECTS MATHÉMATIQUES

Soit une **population** dont nous étudions le caractère X ; supposons que la forme de la distribution de X soit connue, mais que le **paramètre** θ dont dépend cette distribution soit inconnu.

Pour estimer θ, nous tirons de la population un **échantillon** de taille n $(X_1, X_2, \ldots, X_n)$ et cherchons à construire une fonction $G(X_1, X_2, \ldots, X_n)$ qui, pour une réalisation particulière $(x_1, x_2, \ldots, x_n)$, fournit une **valeur** unique comme **estimation** de la valeur du paramètre θ.

DOMAINES ET LIMITATIONS

En général, l'estimation ponctuelle d'un **paramètre** de la **population** ne correspond pas exactement à la **valeur** de ce paramètre. En effet, il existe une **erreur** d'**échantillonnage** du fait qu'une partie de la population a été omise. Il est possible de mesurer cette erreur en calculant la **variance** ou l'**écart type** de l'**estimateur** utilisé pour évaluer le paramètre de la population.

EXEMPLE

Une entreprise fabriquant des ampoules électriques veut étudier la durée de vie moyenne de ses ampoules. Elle tire alors un **échantillon** de taille n (X_1, X_2, …, X_n) dans la production d'ampoules.

La fonction $G(X_1, X_2, \ldots, X_n)$ qu'elle utilise pour estimer le **paramètre** θ correspondant à la **moyenne** μ de la **population** est la suivante :

$$\bar{X} = \sum_{i=1}^{n} \frac{X_i}{n}.$$

Cet **estimateur** $\bar{X}$ de μ est une **variable aléatoire** et, pour une réalisation particulière $(x_1, x_2, \ldots, x_n)$ de l'échantillon, il prend une **valeur** précise notée $\bar{x}$. Ainsi, par exemple, pour $n = 5$, nous pouvons noter l'échantillon suivant (en heures) :

$$(812,\ 1\,067,\ 604,\ 918,\ 895)$$

L'estimateur $\bar{x}$ vaut alors :

$$\begin{aligned} \bar{x} &= \frac{812 + 1\,067 + 604 + 918 + 895}{5} \\ &= \frac{4\,296}{5} \\ &= 859,2. \end{aligned}$$

La valeur $859,2$ est une estimation ponctuelle de la durée de vie moyenne de ces ampoules.

MOTS CLÉS

Échantillon *(Sample)*
Échantillonnage *(Sampling)*
Erreur *(Error)*
Estimateur *(Estimator)*
Estimation *(Estimation)*

ESTIMATION ROBUSTE

Robust estimation

Une **estimation** est dite robuste si les estimateurs utilisés sont peu sensibles aux observations aberrantes. Ces observations aberrantes sont des erreurs provenant d'une mauvaise lecture, d'un mauvais enregistrement ou de toute autre cause due à l'environnement expérimental. La **moyenne arithmétique** et d'autres estimateurs des **moindres carrés** sont très sensibles à de telles valeurs aberrantes. C'est pourquoi il est souhaitable dans ces cas-là de faire appel à des estimateurs plus robustes.

HISTORIQUE

Le terme « robuste » ainsi que le domaine auquel il est lié a été introduit en 1953 par G.E.P Box. Ce terme ne fut reconnu comme champ d'investigation qu'au milieu des années soixante.

Il faut cependant noter que ce concept n'est pas nouveau puisqu'à la fin du XIX^{e} siècle, plusieurs scientifiques avaient déjà une idée relativement claire de ce concept. En fait, le premier travail mathématique sur l'estimation robuste semble remonter à 1818 avec **Pierre Simon de Laplace** dans son *Deuxième supplément à la théorie analytique des probabilités*. On y trouve en particulier la distribution de la **médiane**.

C'est surtout à partir de 1964, avec l'article de Peter J. Huber, que ce domaine est devenu une branche à part entière de la **statistique**.

ASPECTS MATHÉMATIQUES

Soit le **modèle** de régression linéaire multiple :

$$\mathbf{Y} = \mathbf{X} \cdot \boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

avec

- $\mathbf{Y}$ est le vecteur $(n \times 1)$ des observations relatives à la **variable dépendante** (n observations),
- $\boldsymbol{\beta}$ est le vecteur $(p \times 1)$ des paramètres à estimer,
- $\boldsymbol{\varepsilon}$ est le vecteur $(n \times 1)$ des erreurs.

Il existe plusieurs méthodes d'estimation robuste permettant d'estimer le **vecteur** de paramètres $\boldsymbol{\beta}$.

Les plus connues sont :

- L'**estimation L1** :

$$\text{minimiser} \sum_{i=1}^{n} |\varepsilon_i|.$$

- De manière plus générale, on définit le concept de M-estimateurs basés sur l'idée de remplacer le carré des erreurs ε_i^2 (méthode des **moindres carrés**) par une fonction $\delta(\varepsilon_i)$, où δ est une fonction symétrique possédant un minimum unique en zéro :

$$\text{minimiser} \sum_{i=1}^{n} \delta(\varepsilon_i).$$

En 1964, P.J. Huber proposa la fonction suivante :

$$\delta(x) = \begin{cases} \dfrac{x^2}{2} & \text{si } |x| \leq k \\ k \cdot |x| - \dfrac{k^2}{2} & \text{si } |x| > k \end{cases}$$

où k est une constante positive.

On remarque que pour $\delta(x) = |x|$, on obtient les estimateurs L1.

- La méthode de la moindre **médiane** des carrés :

$$\text{minimiser } Med_i \varepsilon_i^2.$$

MOTS CLÉS

Estimateur *(Estimator)*
Estimation *(Estimation)*
Estimation L1 *(L1 estimation)*
Maximum de vraisemblance *(Maximum likelihood)*
Observation aberrante *(Outlier)*

RÉFÉRENCES

Box, G.E.P. (1953). Non-normality and tests on variance. *Biometrika*, 40, pp. 318-335.

Huber, P.J. (1964). Robust estimation of a location parameter. *Annals of Mathematical Statistics*, 35, pp. 73-101.

Huber, P.J. (1981). *Robust Statistics.* John Wiley & Sons, New York.

Laplace, P.S. de (1818). *Deuxième supplément à la théorie analytique des probabilités.* Courcier, Paris, pp. 531-580 dans les *Œuvres complètes de Laplace*, vol. 7, Gauthier-Villars, Paris, 1886.

Rousseeuw, P.J. and Leroy, A.M. (1987). *Robust Regression & Outliers Detection.* John Wiley & Sons, New York.

ÉVÉNEMENT

Event

Un événement d'une **expérience aléatoire** est un sous-ensemble de l'ensemble des parties de l'**ensemble fondamental**. En d'autres termes, un événement est un ensemble de résultats possibles d'une expérience aléatoire.

ASPECTS MATHÉMATIQUES

Soit Ω l'**ensemble fondamental** d'une **expérience aléatoire**, tout sous-ensemble de l'ensemble des parties de Ω est appelé un événement.

Différents types d'événements

- *Événement impossible*

L'événement impossible correspond à l'ensemble vide. La **probabilité** de l'événement impossible est égale à 0 :

$$\text{si } B = \phi,\ P(B) = 0.$$

- *Événement certain*

L'événement certain est l'événement correspondant à l'ensemble fondamental. La probabilité de l'événement certain est égale à 1 :

$$P(\Omega) = 1.$$

- *Événement simple*

Un événement simple (ou un singleton) est un événement qui ne contient qu'un seul résultat.

Opérations sur les événements

- *Négation d'un événement*

Soit l'événement E, la négation de cet événement est un autre événement, appelé événement contraire de E, qui se réalise lorsque E ne se réalise

pas et vice-versa. L'événement contraire de E est noté par l'ensemble $\bar{E}$, **complémentaire** de E :

$$\begin{aligned}\overline{E} &= \Omega \backslash E \\ &= \{\omega \in \Omega\,;\, \omega \notin E\}\,.\end{aligned}$$

- *Intersection d'événements*

  Soient E et F deux événements, l'intersection de ces deux événements est un autre événement se réalisant lorsque les deux événements se réalisent simultanément. On parle de l'événement « E et F » que l'on note $E \cap F$:

  $$E \cap F = \{\omega \in \Omega\,;\, \omega \in E \text{ et } \omega \in F\}\,.$$

  De façon plus générale, si $E_1, E_2, \ldots, E_r$ sont r événements, on définit :

  $$\bigcap_{i=1}^{r} E_i = \{\omega \in \Omega\,;\, \omega \in E_i, \forall i = 1, \ldots, n\}$$

  comme étant l'intersection des r événements, qui se réalise lorsque chacun des r événements se réalise.

- *Réunion d'événements*

  Soit E et F deux événements, la réunion de ces deux événements est un autre événement qui se réalise lorsque au moins l'un des deux événements E ou F se réalise. On parle parfois aussi de l'événement « E ou F », et on le note $E \cup F$:

  $$E \cup F = \{\omega \in \Omega\,;\, \omega \in E \text{ ou } \omega \in F\}\,.$$

  De façon plus générale, si $E_1, E_2, \ldots, E_r$ sont r événements, on définit :

  $$\bigcup_{i=1}^{r} E_i = \{\omega \in \Omega\,;\, \exists i_0 \in \{1, \ldots, n\}\,;\, \omega \in E_{i_0}\}$$

  comme étant la réunion des r événements, qui se réalise lorsque au moins un des r événements se réalise.

Propriétés des opérations sur les événements

Soient E, F et G trois événements. L'intersection et la réunion d'événements satisfont aux propriétés suivantes :

— *commutativité* :

$$\begin{aligned}E \cap F &= F \cap E \\ E \cup F &= F \cup E\,;\end{aligned}$$

— *associativité* :

$$\begin{aligned}E \cap (F \cap G) &= (E \cap F) \cap G \\ E \cup (F \cup G) &= (E \cup F) \cup G\,;\end{aligned}$$

— *idempotence* :

$$\begin{aligned}E \cap E &= E \\ E \cup E &= E.\end{aligned}$$

La distributivité de l'intersection par rapport à la réunion se présente sous la forme :

$$E \cap (F \cup G) = (E \cap F) \cup (E \cap G).$$

La distributivité de la réunion par rapport à l'intersection se résume à :

$$E \cup (F \cap G) = (E \cup F) \cap (E \cup G).$$

Les principales relations faisant intervenir la négation sont :

$$\begin{aligned}\overline{E \cap F} &= \bar{E} \cup \bar{F}\,; \\ \overline{E \cup F} &= \bar{E} \cap \bar{F}.\end{aligned}$$

EXEMPLES

On considère l'**expérience aléatoire** qui consiste à lancer simultanément un dé jaune et un dé bleu.

L'**ensemble fondamental** de cette **expérience** est formé de l'ensemble des couples de résultats possibles des deux dés :

$$\begin{aligned}\Omega = \{&(1,1), (1,2), (1,3), \ldots, (2,1), (2,2), \\ &\ldots, (6,5), (6,6)\}\end{aligned}$$

Exemple :

$(1,2)$ = '1' sur le dé jaune et '2' sur le dé bleu.

Le nombre de couples, ou d'événements simples, qu'il est possible de former est donné par le nombre d'arrangements avec répétitions de 2 objets parmi 6 :

$$A_6^2 = 6^2 = 36.$$

En fonction de cet ensemble fondamental, on peut par exemple décrire les événements suivants :

1. la somme des points est égale à six
   $A = \{(1,5),(2,4),(3,3),(4,2),(5,1)\}$;
2. la somme des points est paire
   $B = \{(1,1),(1,3),(1,5),(2,2),(2,4),\ldots,(6,4),(6,6)\}$;
3. la somme des points est inférieure à 5
   $C = \{(1,1),(1,2),(1,3),(2,1),(2,2),(3,1)\}$;
4. la somme des points est impaire
   $D = \{(1,2),(1,4),(1,6),(2,1),(2,3),\ldots,(6,5)\} = \bar{B}$;
5. la somme des points est paire et inférieure à 5
   $E = \{(1,1),(1,3),(2,2),(3,1))\} = B \cap C$;
6. '2' est apparu sur le dé bleu
   $F = \{(1,2),(2,2),(3,2),(4,2),(5,2),(6,2)\}$;
7. '2' est apparu sur le dé bleu ou la somme des points est inférieure à 5
   $G = \{(1,2),(2,2),(3,2),(4,2),(5,2),(6,2),(1,1),(1,3),(2,1),(3,1)\} = F \cup C$;
8. A et D sont incompatibles ;
9. A implique B.

MOTS CLÉS

Compatibilité *(Compatibility)*
Ensemble fondamental *(Sample space)*
Expérience aléatoire *(Random experiment)*
Indépendance *(Independence)*
Probabilité *(Probability)*
Probabilité conditionnelle *(Conditional probability)*

ÉVÉNEMENT CERTAIN

Sure event

Voir **ensemble fondamental**.

EXACTITUDE

Accuracy

L'exactitude au sens général exprime la proximité d'une **valeur** ou d'une **statistique** par rapport à une valeur de référence. Dans un sens plus particulier, elle mesure la proximité qu'il y a entre l'**estimateur** T du **paramètre** θ inconnu et la vraie valeur de θ.

On peut mesurer l'exactitude d'un estimateur par l'**espérance mathématique** de l'écart au carré entre T et θ, c'est-à-dire par :

$$E\left[(T-\theta)^2\right].$$

L'exactitude est à ne pas confondre avec le terme « précision » qui indique le degré de justesse d'une mesure et qui est généralement donné par le nombre de décimales placées après la virgule.

MOTS CLÉS

Biais *(Bias)*
Estimateur *(Estimator)*
Paramètre *(Parameter)*
Statistique *(Statistics)*

EXPÉRIENCE

Experiment

Une expérience est une opération conduite sous des conditions contrôlées dans le but de découvrir un effet inconnu, de tester ou d'établir une **hypothèse** ou encore de démontrer une loi connue.

Le but de l'expérience est de clarifier la **relation** entre les conditions contrôlables et le résultat de l'expérience.

L'analyse est basée sur des **observations** qui ne sont pas uniquement affectées par les conditions contrôlées, mais également par des conditions incontrôlées et des erreurs de mesure.

DOMAINES ET LIMITATIONS

Lorsque les résultats possibles de l'expérience peuvent être décrits et qu'il est possible d'attribuer une **probabilité** de réalisation à chacun de ces résultats, l'expérience est alors dite **expérience aléatoire**. L'ensemble des résultats ou événements possibles d'une expérience est appelé **ensemble fondamental**.

On parle d'**expérience factorielle** lorsque l'expérimentateur organise une expérience avec deux ou plusieurs facteurs. Les facteurs sont les conditions contrôlables d'une expérience. Les erreurs expérimentales proviennent des conditions incontrôlées et des erreurs de mesure présentes dans toute expérience.

En fonction du but de l'expérience, du nombre de facteurs utilisés et des ressources disponibles, on établira un **plan d'expérience**.

EXEMPLES

Voir **expérience aléatoire** et **expérience factorielle**.

MOTS CLÉS

Expérience aléatoire *(Random experiment)*
Expérience factorielle *(Factorial experiment)*
Plan d'expérience *(Design of experiments)*
Test d'hypothèse *(Hypothesis testing)*

RÉFÉRENCES

Box, G.E.P. and Hunter, W.G. and Hunter, J.S. (1978). *Statistics for experimenters. An introduction to design, data analysis, and model building.* John Wiley & Sons, New York.

Rüegg, A. (1985). *Probabilités et statistique.* Presses Polytechniques Romandes, Lausanne, Suisse.

EXPÉRIENCE ALÉATOIRE

Random experiment

Une **expérience** est composée d'une action réalisée par une personne, suivie d'un résultat.

Lorsque cette action peut engendrer plusieurs résultats possibles, l'expérience est alors dite aléatoire.

On peut caractériser une expérience aléatoire par les faits suivants :

1. on peut décrire, avant l'expérience, l'ensemble de tous les résultats possibles (appelé **ensemble fondamental** de l'expérience aléatoire) ;
2. on ne peut pas prédire avec certitude le résultat ;
3. en revanche, on peut associer à chaque résultat possible une certaine **probabilité** de réalisation.

DOMAINES ET LIMITATIONS

Les phénomènes aléatoires auxquels s'applique le calcul des **probabilités** se rencontrent dans des domaines très variés, tels que les sciences économiques, les sciences sociales, la physique, la médecine, la psychologie, etc.

Lorsque l'homme participe à la réalisation d'un phénomène aléatoire, on parle alors d'expérience aléatoire.

EXEMPLES

Voici quelques exemples d'expériences aléatoires, caractérisées par leur action et par l'ensemble de leurs résultats possibles (leur **ensemble fondamental**).

Action	Ensemble fondamental
- Lancement d'une pièce de monnaie	$\Omega = \{pile, face\}$
- Lancement d'un dé	$\Omega = \{1, 2, 3, 4, 5, 6\}$
- Tirage d'une boule dans une urne contenant 1 boule verte, 1 bleue et 1 jaune	$\Omega = \{bleue, verte, jaune\}$
- Mesure des précipitations sur une période déterminée	Ω = ensemble des réels non négatifs (ensemble infini non dénombrable)

MOTS CLÉS

Ensemble fondamental *(Sample space)*
Événement *(Event)*
Probabilité *(Probability)*

EXPÉRIENCE FACTORIELLE

Factorial experiment

Une expérience factorielle est une **expérience** dans laquelle on étudie tous les traitements possibles formés à partir de deux facteurs ou plus, chaque facteur comportant deux niveaux ou plus, de telle sorte que les effets principaux et les interactions puissent être étudiés.

Le terme d'expérience factorielle signifie que les différents facteurs sont combinés de toutes les façons possibles, mais il ne décrit pas le **plan d'expérience** dans lequel ces combinaisons, ou une partie d'entre elles, seront étudiées.

ASPECTS MATHÉMATIQUES

Dans le contexte des expériences factorielles, la notion de **modèle** fait appel à une équation permettant de relier la **variable** réponse aux **variables indépendantes**.

Ainsi, si l'expérience factorielle implique deux facteurs, le modèle qui y est associé est le suivant :

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk},$$

i $= 1, 2, \ldots, a$ (niveaux du facteur A),
j $= 1, 2, \ldots, b$ (niveaux du facteur B),
k $= 1, 2, \ldots, n_{ij}$ (nombre d'observations recevant le **traitement** ij),

où

Y_{ijk} est la variable réponse,
μ est la **moyenne** générale commune à tous les traitements,
α_i l'effet du i-ème niveau du facteur A,
β_j l'effet du j-ème niveau du facteur B,
$(\alpha\beta)_{ij}$ l'effet de l'interaction entre α_i et β_j et
ε_{ijk} l'**erreur** expérimentale de l'observation Y_{ijk}.

En général, μ est une constante et ε_{ijk} est une **variable aléatoire** distribuée selon la **loi normale** de moyenne 0 et de **variance** σ^2.

Si on supprime le terme d'interaction $(\alpha\beta)_{ij}$, on obtient un modèle additif, c'est-à-dire que l'effet de chaque facteur est une constante qui s'ajoute à la moyenne générale μ et qui ne dépend pas du niveau de l'autre facteur. On parle alors de modèle sans interactions.

DOMAINES ET LIMITATIONS

Lorsque, dans une expérience factorielle, tous les facteurs ont le même nombre de niveaux, la définition de l'expérience est généralement donnée sous la forme du nombre de niveaux élevé à une puissance égale au nombre de facteurs. Ainsi, par exemple, pour une expérience impliquant trois facteurs, chacun à deux niveaux, on obtient une expérience factorielle 2^3 qui comporte huit traitements.

La notation la plus utilisée pour définir les combinaisons de traitements dans une expérience factorielle est la suivante : utiliser une lettre majuscule pour désigner le facteur et un indice numérique pour désigner son niveau.

Soit par exemple, une expérience factorielle comprenant trois facteurs : un facteur A avec deux niveaux, un facteur B avec deux niveaux et un facteur C avec trois niveaux ; les douze combinaisons seront :

$$\begin{array}{lll} A_1B_1C_1, & A_1B_1C_2, & A_1B_1C_3, \\ A_1B_2C_1, & A_1B_2C_2, & A_1B_2C_3, \\ A_2B_1C_1, & A_2B_1C_2, & A_2B_1C_3, \\ A_2B_2C_1, & A_2B_2C_2, & A_2B_2C_3. \end{array}$$

Parfois, seuls les indices énumérés dans le même ordre que les facteurs sont utilisés :

$$\begin{array}{lll} 111, & 112, & 113, \\ 121, & 122, & 123, \\ 211, & 212, & 213, \\ 221, & 222, & 223. \end{array}$$

Les expériences factorielles doubles, ou plus généralement multiples, sont importantes pour les raisons suivantes :

- Une expérience factorielle double est plus efficace par rapport aux ressources que ne le serait la combinaison de deux expériences à un seul facteur. La première prend moins de temps et nécessite moins d'unités expérimentales pour un niveau donné de précision.

- Une expérience factorielle double permet aux effets sur les réponses dues à l'augmentation du niveau de chaque facteur d'être estimés à chaque niveau de l'autre facteur. Ceci amène des conclusions valables sur une plus vaste étendue de conditions expérimentales qu'une suite de plans à un seul facteur.

- Finalement, la recherche simultanée sur deux facteurs est nécessaire quand des interactions entre les facteurs sont présentes, c'est-à-dire quand l'effet d'un facteur dépend du niveau de l'autre facteur.

MOTS CLÉS

Expérience *(Experiment)*
Plan d'expérience *(Design of experiments)*

RÉFÉRENCE

Dodge, Y. (1986). Principaux plans d'expériences. In *Méthodologies et techniques de plans d'expériences*, 28e cours de perfectionnement de l'Association vaudoise des chercheurs en physique, Saas-Fee, 2-8 mars 1986. Édité par Aeschlimann J., Bonjour C. et Stocker E.

EXPONENTIELLE, LOI

Exponential distribution

Voir **loi exponentielle**.

EXTRAPOLATION

Extrapolation

Une extrapolation au sens statistique est une **estimation** de la **variable dépendante** pour des valeurs des **variables indépendantes** qui se situent en dehors de l'ensemble des observations.

L'extrapolation est souvent utilisée dans les séries chronologiques quand le **modèle** déterminé par des valeurs $u_1, u_2, \ldots, u_n$ observées à des temps $t_1, t_2, \ldots, t_n$ (avec $t_1 < t_2 < \ldots < t_n$) sert à prédire la valeur au temps t_{n+1}.

Plus généralement, une fonction de **régression linéaire simple** fondée sur les observations (X_i, Y_i), $i = 1, 2, \ldots, n$, peut être utilisée pour estimer des valeurs de Y correspondant à des valeurs de X qui se situent en dehors de l'ensemble $X_1, X_2, \ldots, X_n$. On peut aussi faire des extrapolations sur un modèle de **régression linéaire multiple**.

DOMAINES ET LIMITATIONS

Il faut prendre de grandes précautions lors des extrapolations car elles sont basées sur l'**hypothèse** que le **modèle** ne change pas pour les valeurs utilisées pour l'extrapolation.

MOTS CLÉS

Prévision *(Forecasting)*
Régression linéaire multiple *(Multiple linear regression)*
Régression linéaire simple *(Simple linear regression)*
Série chronologique *(Time series)*


F

FACTEUR

Factor

On appelle facteur toute condition contrôlable d'une **expérience**. Chaque facteur peut prendre un nombre déterminé de valeurs, appelées les niveaux d'un facteur.

Les expériences faisant usage de facteurs sont appelées expériences factorielles.

EXEMPLE

Un sociologue souhaite connaître l'opinion des habitants d'une ville face au problème de la drogue. Il s'intéresse à l'influence du sexe et du revenu des citoyens exercée sur leur attitude.

Les niveaux du facteur sexe sont (H) pour les hommes et (F) pour les femmes ; les niveaux du facteur revenu correspondent à bas (B), moyen (M) et élevé (E), selon la spécification de l'expérimentateur.

L'espace factoriel G de cette **expérience** est composé des six combinaisons suivantes des niveaux des deux facteurs :

$$G = \{(H,B)\ ;\ (H,M)\ ;\ (H,E)\ ;\ (F,B)\ ;\ (F,M)\ ;\ (F,E)\}.$$

Dans cet exemple, le facteur sexe est qualitatif et prend des valeurs nominales ; alors que le facteur revenu, quantitatif, prend des valeurs ordinales.

MOTS CLÉS

Analyse de variance *(Analysis of variance)*
Expérience *(Experiment)*
Expérience factorielle *(Factorial experiment)*
Niveau *(Level)*
Plan d'expérience *(Design of experiments)*
Traitement *(Treatment)*

RÉFÉRENCE

Raktoe, B.L. Hedayat, A. and Federer, W.T. (1981). *Factorial Designs.* John Wiley & Sons, New York.

FISHER, INDICE DE

Fisher index

Voir **indice de Fisher**.

FISHER, LOI DE

Fisher distribution

Voir **loi de Fisher**.

FISHER, TABLE DE

Fisher table

Voir **table de Fisher**.

FISHER, TEST DE

Fisher test

Voir **test de Fisher**.

FISHER IRVING

Fisher Irving

Irving Fisher est né à New York, le 27 février 1867. Il a obtenu son doctorat à l'Université de Yale en 1892. Son approche mathématique de la théorie de la valeur et des prix dans sa thèse lui valut d'être considéré comme l'un des premiers économistes mathématiciens d'Amérique. Il a fondé l'Econometric Society en 1930 avec Ragnar Frisch et Charles F. Roos et en est devenu le premier président. I. Fisher a fait toute sa carrière à l'Université de Yale. Il a commencé y par enseigner les mathématiques, puis l'économie. Après un séjour en Europe, il a été nommé professeur de sciences sociales en 1898 et est mort à New York en 1947.

I. Fisher a été un auteur très prolifique, la liste de ses

publications comporte plus de 2 000 titres. Il s'est intéressé non seulement à la théorie économique mais également à la recherche statistique. En 1920, il a proposé des méthodes économétriques et statistiques pour calculer les indices. Son « indice idéal », que l'on appelle maintenant l'**indice de Fisher**, est une moyenne géométrique de l'**indice de Laspeyres** et de l'**indice de Paasche**.



Quelques ouvrages et articles principaux d'Irving Fisher :

1892 *Mathematical Investigations in the Theory of Value and Prices*, Connecticut Academy of Arts and Science, New Haven.

1906 *The Nature of Capital and Income.* The Macmillan Company, New York.

1907 *The Rate of Interest, Its nature, determination and relation to economic phenomena.* The Macmillan Company, New York.

1910 *Introduction to Economic Science.* The Macmillan Company, New York.

1911 *The Purchasing power of Money.* The Macmillan Company, New York.

1912 *Elementary principles of Economics.* The Macmillan Company, New York.

1921 The Best Form of Index Number, *American Statistical Association Quarterly*, 17, pp. 533-537.

1922 *The Making of Index Numbers,* Houghton Mifflin, Boston.

1926 A Statistical Relation Between Unemployment and Price Changes, *International Labour Review*, 13, pp. 785-792.

1927 A Statistical Method for Measuring "Marginal Utility" and Testing the Justice of a Progressive Income Tax, in *Economic Essays contributed in honor of John Bates Clark*, ed. J.H. Hollander. The Macmillan Company, New York.

1930 *The Theory of Interest as determined by impatience to spend income and opportunity to invest it.* The Macmillan Company, New York.

1932 *Booms and Depressions.* Adelphi, New York.

1933 The Debt-Deflation Theory of Great Depressions. *Econometric*, 1, pp. 337-357.

1935 *100 % Money.* Adelphi, New York.

MOTS CLÉS

Indice *(Index number)*
Indice de Fisher *(Fisher index)*

FISHER RONALD AYLMER

Fisher Ronald Aylmer

Né en 1890 à East Finchley, près de Londres, Ronald Aylmer Fisher fit des études de mathématiques à Harrow et au Caius College, Cambridge. Il commença sa carrière de statisticien en 1919 en travaillant à l'Institut de recherche agricole de Rothamsted (Rothamsted Experimental Station). La période qu'il y passa et qui dura jusqu'en 1933 fut très productive. En 1933, il fut nommé professeur d'eugénique à l'University College, puis, de 1943 à 1957, il occupa la chaire de génétique Arthur Balfour à

Cambridge. Une fois retraité, Fisher fit des travaux de recherche à la division de statistiques mathématiques de la Commonwealth Scientific and Industrial Research Organisation à Adélaïde, en Australie. C'est là qu'il mourut en 1962.

R.A. Fisher est reconnu comme l'un des fondateurs de la **statistique** moderne. Réputé dans le domaine de la génétique, il apporta également d'importantes contributions en théorie générale de l'**estimation** du **maximum de vraisemblance**. Ses travaux sur les plans d'expériences nécessaires pour l'étude des données des expériences agricoles, méritent également d'être soulignés. Ils sont à inscrire en parallèle à l'**analyse de variance**.



C'est avec R.A. Fisher qu'apparurent les principes de **randomisation**, de blocs randomisés, de plans carrés latins et d'**arrangement** factoriel.

Quelques ouvrages et articles principaux de Ronald Aylmer Fisher :

1912 On an absolute criterium for fitting frequency curves. *Messenger of Mathematics*, 41, pp. 155-160.

1918 The correlation between relatives on the supposition of Mendelian inheritance. *Trans. Roy. Soc. Edinburgh*, 52, pp. 399-433.

1920 A mathematical examination of methods of determining the accuracy of an observation by the mean error and by the mean square error. *Mon. Not. Rev. Astr. Soc.* 80, pp. 758-770.

1921 On the "probable error" of a coefficient of correlation deduced from a small sample. *Metron*, 1(4), pp. 1-32.

1922 On the mathematical foundation of theoretical statistics. *Philosophical Transactions of the Royal Society of London.* Series A, Vol. 222, pp. 309-368.

1925 Theory of statistical estimation. *Proceedings of the Cambridge Philosophical Society,* 22, pp. 700-725.

1925 *Statistical Methods for Research Workers.* Oliver & Boyd, Edinburgh.

1934 Probability, likelihood and quantity of information in the logic of uncertain inference. *Proceedings of the Royal Society of London.* Series A, 146, pp. 1-8.

1935 *The Design of Experiments.* Oliver & Boyd, Edinburgh.

1971-1974 *Collected Papers of R.A. Fisher.* Ed. Benett J.H., University of Adelaide.

MOTS CLÉS

Loi de Fisher *(Fisher distribution)*
Table de Fisher *(Fisher table)*
Test de Fisher *(Fisher test)*

RÉFÉRENCE

Fisher-Box, J. (1978). *R.A. Fisher, the life of a scientist.* John Wiley & Sons, New York.

FLUCTUATION CYCLIQUE

Cyclical fluctuation

Les fluctuations cycliques caractérisent les oscillations

qui se présentent sur une grande période autour de la droite (ou courbe) de **tendance séculaire** d'une **série chronologique**.

HISTORIQUE

Voir **série chronologique**.

ASPECTS MATHÉMATIQUES

Soit Y_t une **série chronologique** donnée à l'aide de ses composantes ; Y_t s'écrit :

$Y_t = T_t \cdot S_t \cdot C_t \cdot I_t$ (**modèle** multiplicatif) ou $Y_t = T_t + S_t + C_t + I_t$ (modèle additif),

où

Y_t est la **donnée** au temps t ;
T_t est la **tendance séculaire** au temps t ;
S_t est la **variation saisonnière** au temps t ;
C_t est la fluctuation cyclique au temps t et
I_t est la **variation irrégulière** au temps t.

Lors de l'étude d'une série chronologique, on commence toujours par déterminer la tendance séculaire T_t, puis les variations saisonnières S_t. Ensuite il est possible d'ajuster les données initiales de la série chronologique Y_t suivant ces deux composantes :

$$\frac{Y_t}{S_t \cdot T_t} = C_t \cdot I_t \text{ (modèle multiplicatif),}$$

$$Y_t - S_t - T_t = C_t + I_t \text{ (modèle additif).}$$

Pour évaluer les fluctuations cycliques, on effectue une **moyenne mobile** pondérée prise sur quelques mois seulement. L'utilisation des moyennes mobiles permet de lisser les variations irrégulières I_t en préservant les fluctuations cycliques C_t.

Le choix d'une moyenne mobile pondérée permet de donner plus de poids au valeurs centrales par rapport à celles des extrémités, afin de reproduire les fluctuations cycliques de manière plus fidèle. On choisit donc de grands poids pour les valeurs centrales et des petits aux extrémités, pour effectuer une pondération judicieuse.

Par exemple pour une moyenne mobile pondérée sur 5 mois, on peut choisir les poids $-0,1$, $0,3$, $0,6$, $0,3$, $-0,1$; comme leur somme est 1, il n'y a pas de division à effectuer.

Si on note X_t les valeurs de $C_t \cdot I_t$ (résultant de l'ajustement par rapport à la tendance séculaire et aux variations saisonnières), la valeur de la fluctuation cyclique pour le mois t est déterminée par l'expression :

$$\begin{aligned} C_t &= -0,1 \cdot X_{t-2} + 0,3 \cdot X_{t-1} + 0,6 \cdot X_t \\ &\quad +0,3 \cdot X_{t+1} - 0,1 \cdot X_{t+2}. \end{aligned}$$

DOMAINES ET LIMITATIONS

L'estimation des fluctuations cycliques permet :

- de déterminer les maxima ou les minima que peut atteindre une **série chronologique** ;
- d'effectuer une **prévision** à court ou à moyen terme ;
- d'identifier les composantes cycliques.

Les limitations et avantages de l'utilisation des **moyennes mobiles** pondérées pour l'évaluation des fluctuations cycliques sont les suivants :

- les moyennes mobiles pondérées ont l'avantage de lisser la courbe des mouvements cycliques tout en restant proche des mouvements initiaux car elles préservent plus fidèlement l'amplitude de ces cycles ;
- l'utilisation d'un nombre impair de mois pour faire la moyenne mobile permet de mieux centrer les valeurs obtenues ;
- la fluctuation cyclique d'une série chronologique est la plus difficile à étudier car les cycles varient en général en longueur et en amplitude. Elle est due à une multitude de facteurs, dont l'effet peut changer d'un cycle à l'autre. Aucun **modèle** mis au point pour expliquer et prévoir les fluctuations ne s'est avéré jusqu'à maintenant satisfaisant.

EXEMPLE

On effectue une **moyenne mobile** pondérée sur cinq mois des données ajustées suivant la **tendance**

séculaire et suivant les variations saisonnières figurant dans le tableau ci-dessous.

On choisit les poids $-0,1$, $0,3$, $0,6$, $0,3$, $-0,1$; comme leur somme est égale à 1, il n'y a pas de division à effectuer.

Si l'on pose X_i les valeurs de $C_i \cdot I_i$ (ajustées) :

$$\begin{aligned} C_i &= -0,1 \cdot X_{i-2} + 0,3 \cdot X_{i-1} + 0,6 \cdot X_i \\ &\quad +0,3 \cdot X_{i+1} - 0,1 \cdot X_{i+2}. \end{aligned}$$

Ce tableau indique la puissance électrique consommée chaque mois en millions de kilowatt heures pendant les années 1952 et 1953, pour l'éclairage des rues et des routes, les données ayant préalablement été ajustées suivant la tendance séculaire et les variations saisonnières.

Année Mois	Donnée X_i	Moyenne mobile pond. sur 5 mois C_i
1952 J	99,9	
F	100,4	
M	100,2	100,1
A	99,0	99,1
M	98,1	98,4
J	99,0	98,7
J	98,5	98,3
A	97,8	98,3
S	100,3	99,9
O	101,1	101,3
N	101,2	101,1
D	100,4	100,7
1953 J	100,6	100,3
F	100,1	100,4
M	100,5	100,1
A	99,2	99,5
M	98,9	98,7
J	98,2	98,2
J	98,4	98,5
A	99,7	99,5
S	100,4	100,5
O	101,0	101,0
N	101,1	
D	101,1	

Dans le problème actuel, il n'apparaît aucun effet cyclique important, ou bien, s'il en existe un, il est négligeable. En économie, on cherche souvent le début des cycles mais, il arrive qu'il ne fasse son apparition qu'après une période de vingt ans.

MOTS CLÉS

Moyenne mobile *(Moving average)*
Prévision *(Forecasting)*
Série chronologique *(Time series)*
Tendance séculaire *(Secular trend)*
Variation irrégulière *(Irregular variation)*
Variation saisonnière *(Seasonal variation)*

RÉFÉRENCES

Box, G.E.P. and Jenkins, G.M. (1970). *Time series analysis: Forecasting and control, Holden-Day Series in Time Series Analysis*, San Francisco.

Wold, H.O. editor (1965). *Bibliography on Time Series and Stochastic Processes.* Oliver & Boyd, Edinburgh, Edinburgh.

FONCTION DE DENSITÉ

Density function

La fonction de densité d'une **variable aléatoire** continue permet de déterminer la **probabilité** qu'une variable aléatoire X prenne une **valeur** dans un **intervalle** fixé.

HISTORIQUE

Voir **probabilité**.

ASPECTS MATHÉMATIQUES

Une variable X est une variable aléatoire continue s'il existe une fonction f non négative, définie sur les réels et vérifiant pour tout intervalle $[a, b]$ la propriété suivante :

$$P(a \leq X \leq b) = \int_a^b f(x)\, dx$$

où $P(a \leq X \leq b)$ est la **fonction de probabilité**. Ainsi, la **probabilité** que la **variable aléatoire** continue X prenne une **valeur** dans l'**intervalle** $[a, b]$ peut être obtenue en intégrant la fonction de probabilité sur $[a, b]$.

Il y a la condition suivante sur f :

$$\int_{\infty}^{\infty} f(x)\, dx = P(\infty \leq X \leq \infty) = 1.$$

On représente graphiquement la fonction de densité sur un système d'axes. Les différentes valeurs de la variable aléatoire X sont placées en abscisse, et celles prises par la fonction f en ordonnée.

Le graphe de la fonction f ne permet pas de déterminer la probabilité en un point, mais de visualiser la probabilité d'un intervalle par une surface.

DOMAINES ET LIMITATIONS

Si l'on pose $a = b$ dans $P(a \leq X \leq b) = \int_a^b f(x)\, dx$, il en résulte :

$$P(X = a) = \int_a^a f(x)\, dx = 0.$$

Cela signifie que la **probabilité** qu'une **variable aléatoire** continue prenne une **valeur** isolée fixée est toujours nulle.

Il est donc indifférent d'inclure ou d'exclure les bornes de l'**intervalle** dans le calcul de la probabilité associé à un intervalle.

EXEMPLE

Soit X une **variable aléatoire** continue dont la fonction de densité est :

$$f(x) = \begin{cases} \frac{3}{4}\left(-x^2 + 2x\right) & \text{si } 0 < x < 2 \\ 0 & \text{sinon.} \end{cases}$$

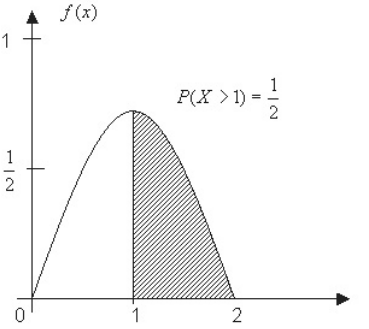
On désire calculer la **probabilité** que X soit supérieure à 1 :

$$P(X > 1) = \int_1^{\infty} f(x)\, dx$$

qui peut être divisé en :

$$\begin{aligned} P(X > 1) &= \int_1^2 \frac{3}{4}\left(-x^2 + 2x\right) dx + \int_2^{\infty} 0 dx \\ &= \frac{3}{4}\left(-\frac{x^3}{3} + x^2\Big|_1^2\right) \\ &= \frac{1}{2}. \end{aligned}$$

On peut représenter graphiquement la fonction de densité ci-dessus ainsi que $P(X > 1)$:



La surface hachurée représente la probabilité que $X > 1$.

MOTS CLÉS

Fonction de répartition continue *(Continuous distribution function)*
Loi de probabilité continue *(Continuous probability distribution)*
Probabilité *(Probability)*

FONCTION DE DENSITÉ CONJOINTE

Joint density function

La fonction de densité conjointe d'un **couple de variables aléatoires** continues est une fonction qui permet, par intégration, de calculer la **probabilité** associée à un **événement**.

HISTORIQUE

Voir **probabilité**.

ASPECTS MATHÉMATIQUES

Considérons la **fonction de répartition conjointe**

$$F(a, b) = P(X \leq a, Y \leq b), \text{ pour } -\infty < a, b < \infty.$$

Si les dérivées partielles sont définies, il suffit de dériver F pour obtenir la **fonction de densité** conjointe du **couple de variables aléatoires** continues :

$$f(a,b) = \frac{\partial^2 F(a,b)}{\partial a \partial b}.$$

MOTS CLÉS

Couple de variables aléatoires *(Pair of random variables)*
Fonction de densité *(Density function)*
Fonction de densité marginale *(Marginal density function)*
Fonction de répartition conjointe *(Joint distribution function)*

RÉFÉRENCES

Hogg, R.V. and Craig, A.T. (1965). *Introduction to mathematical statistics.* 2th ed. The Macmillan Company, New York.

Ross, S.M. (1987). *Initiation aux probabilités.* Presses polytechniques romandes, Lausanne.

FONCTION DE DENSITÉ MARGINALE

Marginal density function

Lorsque, dans le cas d'un **couple de variables aléatoires** (X,Y), la **variable aléatoire** X est considérée pour elle-même, sa **fonction de densité** est appelée fonction de densité marginale (de même pour la variable aléatoire Y).

HISTORIQUE

Voir **probabilité**.

ASPECTS MATHÉMATIQUES

Dans le cas d'un **couple de variables aléatoires** continues (X,Y) de **fonction de densité conjointe** $f(x,y)$, on obtient la **fonction de densité** marginale de X par intégration :

$$P(X \leq a, -\infty < Y < \infty) = \int_{-\infty}^{a} \int_{-\infty}^{\infty} f(x,y)\, dydx$$
$$= \int_{-\infty}^{a} f_X(x)\, dx$$

où $f_X(x) = \int_{-\infty}^{\infty} f(x,y)\, dy$ est la densité de X.

De même, on obtient la fonction de densité marginale de Y :

$$f_Y(y) = \int_{-\infty}^{\infty} f(x,y)\, dx.$$

MOTS CLÉS

Couple de variables aléatoires *(Pair of random variables)*
Fonction de densité *(Density function)*
Fonction de densité conjointe *(Joint density function)*

RÉFÉRENCES

Hogg, R.V. and Craig, A.T. (1965). *Introduction to mathematical statistics.* 2th ed. The Macmillan Company, New York.

Ross, S.M. (1987). *Initiation aux probabilités.* Presses polytechniques romandes, Lausanne.

FONCTION DE PROBABILITÉ

Probability function

La fonction de probabilité d'une **variable aléatoire** discrète est une fonction qui associe à chaque **valeur** de cette variable aléatoire sa **probabilité**.

HISTORIQUE

Voir **probabilité**.

ASPECTS MATHÉMATIQUES

La fonction $P(b) = P(X = b)$ où b varie selon les valeurs possibles de la **variable aléatoire** discrète X est appelée fonction de probabilité de X.

Comme $P(X = b)$ est toujours positif ou nul, la fonction de probabilité est elle aussi positive ou nulle.

On représente graphiquement la fonction de probabilité sur un système d'axes. Les différentes valeurs b de X sont placées en abscisse, les images $P(b)$ en ordonnée.

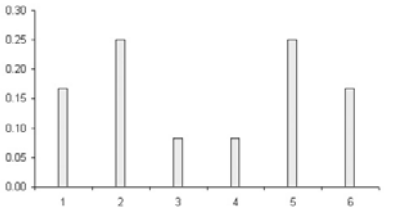
La **probabilité** $P(b)$ est représentée par des rectangles de largeur égale à l'unité et de hauteur égale à la probabilité de b.

Comme X doit prendre au moins une des valeurs b, la somme des $P(b)$ doit être égale à 1, c'est-à-dire que la somme des aires des rectangles doit être égale à 1.

EXEMPLE

On considère l'**expérience aléatoire** qui consiste à lancer un dé truqué. Soit la **variable aléatoire** X correspondant au nombre de points obtenus. La fonction de probabilité $P(X = b)$ est donnée par les probabilités associées à chaque **valeur** de X :

b	1	2	3	4	5	6
$P(b)$	$\frac{1}{6}$	$\frac{1}{4}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{4}$	$\frac{1}{6}$



MOTS CLÉS

Fonction de répartition discrète *(Discrete distribution function)*
Loi de probabilité discrète *(Discrete probability distribution)*
Probabilité *(Probability)*

FONCTION DE RÉPARTITION

Distribution function

La fonction de répartition d'une **variable aléatoire** est définie pour tout nombre réel comme la **probabilité** que la variable aléatoire prenne une **valeur** inférieure ou égale à ce nombre réel.

Selon que la variable aléatoire est discrète ou continue, on parle de **fonction de répartition discrète** ou de **fonction de répartition continue**.

HISTORIQUE

Voir **probabilité**.

ASPECTS MATHÉMATIQUES

On définit la fonction de répartition d'une **variable aléatoire** X au point b (avec $0 < b < \infty$) par :

$$F(b) = P(X \leq b).$$

MOTS CLÉS

Fonction de répartition continue *(Continuous distribution function)*
Fonction de répartition discrète *(Discrete distribution function)*
Probabilité *(Probability)*
Valeur *(Value)*
Variable aléatoire *(Random variable)*

FONCTION DE RÉPARTITION CONJOINTE

Joint distribution function

La fonction de répartition conjointe d'un **couple de variables aléatoires** est définie, pour tout couple de nombres réels, comme la **probabilité** que la première **variable aléatoire** prenne une **valeur** inférieure ou égale au premier nombre réel et que simultanément la deuxième variable aléatoire prenne une valeur inférieure ou égale au deuxième nombre réel.

ASPECTS MATHÉMATIQUES

Soient deux **variables aléatoires** X et Y.

La fonction de répartition conjointe est définie comme suit :

$$F(a,b) = P(X \leq a, Y \leq b), \quad -\infty < a, b < \infty.$$

MOTS CLÉS

Couple de variables aléatoires *(Pair of random variables)*
Fonction de répartition *(Distribution function)*
Fonction de répartition marginale *(Marginal distribution function)*

RÉFÉRENCES

Hogg, R.V. and Craig, A.T. (1965). *Introduction to mathematical statistics.* 2th ed. The Macmillan Company, New York.

Ross, S.M. (1987). *Initiation aux probabilités.* Presses polytechniques romandes, Lausanne.

FONCTION DE RÉPARTITION CONTINUE

Continuous distribution function

La **fonction de répartition** d'une **variable aléatoire** continue est définie pour tout nombre réel comme la **probabilité** que la variable aléatoire prenne une **valeur** inférieure ou égale à ce nombre réel.

HISTORIQUE

Voir **probabilité**.

ASPECTS MATHÉMATIQUES

On appelle **fonction de répartition** d'une **variable aléatoire** continue la fonction définie par :

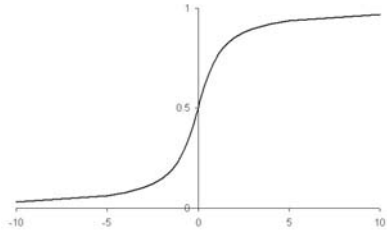
$$F(b) = P(X \leq b) = \int_{-\infty}^{b} f(x)\,dx.$$

Autrement dit, la **fonction de densité** f est la dérivée de la fonction de répartition continue.

Propriétés de la fonction de répartition continue

1. $F(x)$ est une fonction croissante puisque $f(x) \geq 0$ pour tout x ;
2. F prend ses valeurs dans l'**intervalle** $[0,1]$;
3. $\lim_{b \to -\infty} F(b) = 0$;
4. $\lim_{b \to \infty} F(b) = 1$;
5. $F(x)$ est une fonction continue et dérivable.

On peut représenter graphiquement la fonction de répartition sur un système d'axes. En abscisse nous mettons les différentes valeurs de la variable aléatoire X et en ordonnée les valeurs correspondantes de $F(x)$.



DOMAINES ET LIMITATIONS

La **probabilité** que la **variable aléatoire** continue X prenne une **valeur** de l'**intervalle** $]a,b]$, pour tout $a < b$, c'est-à-dire $P(a < X \leq b)$, est égale à $F(b) - F(a)$, où F est la fonction de répartition de la variable aléatoire X.

Démonstration : on peut écrire l'**événement** $\{X \leq b\}$ comme réunion des deux événements mutuellement exclusifs : $\{X \leq a\}$ et $\{a < X \leq b\}$;

$$\{X \leq b\} = \{X \leq a\} \cup \{a < X \leq b\}.$$

En prenant la probabilité de chaque côté de l'équation, on obtient :

$$\begin{aligned} P(X \leq b) &= P(\{X \leq a\} \cup \{a < X \leq b\}) \\ &= P(X \leq a) + P(a < X \leq b). \end{aligned}$$

La somme des probabilités résultant du fait que les deux événements sont exclusifs.

En soustrayant de chaque côté $P(X \leq a)$, nous avons :

$$P(a < X \leq b) = P(X \leq b) - P(X \leq a).$$

Finalement, par la définition de la **fonction de répartition**, on obtient :

$$P(a < X \leq b) = F(b) - F(a).$$

EXEMPLE

Soit une **variable aléatoire** continue X dont la **fonction de densité** est donnée par :

$$f(x) = \begin{cases} 1 & \text{si } 0 < x < 1 \\ 0 & \text{sinon.} \end{cases}$$

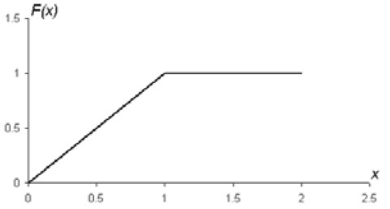
La **probabilité** que X prenne une **valeur** dans l'**intervalle** $[a, b]$, avec $0 < a$ et $b < 1$, est la suivante :

$$\begin{aligned} P(a \leq X \leq b) &= \int_a^b f(x)\,dx \\ &= \int_a^b 1 dx \\ &= x\,\Big|_a^b \\ &= b - a. \end{aligned}$$

La fonction de répartition est donc pour $0 < x < 1$:

$$\begin{aligned} F(x) &= P(X \leq x) \\ &= P(0 \leq X \leq x) \\ &= x. \end{aligned}$$

Cette fonction est représentée sur la figure suivante :



MOTS CLÉS

Expérience aléatoire *(Random experiment)*
Fonction de densité *(Density function)*
Probabilité *(Probability)*
Valeur *(Value)*
Variable aléatoire *(Random variable)*

FONCTION DE RÉPARTITION DISCRÈTE

Discrete distribution function

La fonction de répartition d'une **variable aléatoire** discrète est définie pour tout nombre réel comme la **probabilité** que la variable aléatoire prenne une **valeur** inférieure ou égale à ce nombre réel.

HISTORIQUE

Voir **probabilité**.

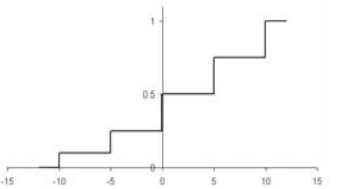
ASPECTS MATHÉMATIQUES

On appelle fonction de répartition d'une **variable aléatoire** discrète X la fonction F définie par :

$$F(b) = P(X \leq b).$$

La fonction de répartition correspond donc, étant donné un nombre réel b, à la **probabilité** que X soit inférieur ou égal à b.

On peut représenter graphiquement la fonction de répartition sur un système d'axes. En abscisse nous mettons les différentes valeurs de la variable aléatoire discrète X et en ordonnée les probabilités cumulées correspondant aux différentes valeurs de X, c'est-à-dire $F(x)$.



Dans le cas où les valeurs possibles de la variable aléatoire discrète sont b_1, b_2,... avec $b_1 < b_2 < \ldots$, la **fonction de répartition** discrète est une fonction en escaliers. $F(b_i)$ est constant sur l'**intervalle** $[b_i, b_{i+1}[$.

Propriétés de la fonction de répartition discrète

1. F est une fonction non décroissante, autrement dit si $a < b$, alors $F(a) \leq F(b)$;
2. F prend ses valeurs dans l'intervalle $[0, 1]$;
3. $\lim\limits_{b \to -\infty} F(b) = 0$;
4. $\lim\limits_{b \to \infty} F(b) = 1$.

EXEMPLE

On considère l'**expérience aléatoire** qui consiste à lancer deux dés simultanément. Soit la **variable aléatoire** discrète X correspondant à la somme des points obtenus.

On cherche la **probabilité** de l'**événement** $\{X \leq 7\}$ qui est par définition la **valeur** prise par la **fonction de répartition** pour la valeur $x = 7$.

La variable aléatoire discrète X prend ses valeurs dans l'ensemble $E = \{2, 3, \ldots, 12\}$. Les probabilités associées à chaque valeur de X sont données par le tableau suivant :

X	2	3	4	5	6
$P(X)$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$

X	7	8	9	10	11	12
$P(X)$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

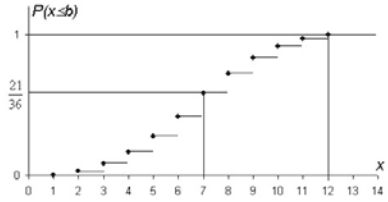
Pour construire la fonction de répartition, il faut calculer pour chaque valeur b de X la somme des probabilités pour toute valeur inférieure ou égale à b. On crée donc un nouveau tableau contenant les probabilités cumulées :

b	2	3	4	5	6	7
$P(X \leq b)$	$\frac{1}{36}$	$\frac{3}{36}$	$\frac{6}{36}$	$\frac{10}{36}$	$\frac{15}{36}$	$\frac{21}{36}$

b	8	9	10	11		
$P(X \leq b)$	$\frac{26}{36}$	$\frac{30}{36}$	$\frac{33}{36}$	$\frac{35}{36}$	$\frac{36}{36} = 1$	

La probabilité de l'événement $\{X \leq 7\}$ est donc égale à $\frac{21}{36}$.

On peut représenter la fonction de répartition de la variable aléatoire discrète X comme suit :



MOTS CLÉS

Expérience aléatoire *(Random experiment)*
Probabilité *(Probability)*
Valeur *(Value)*

FONCTION DE RÉPARTITION MARGINALE

Marginal distribution function

Lorsque, dans le cas d'un **couple de variables aléatoires** (X, Y), la **variable aléatoire** X est considérée pour elle-même, sa **fonction de répartition** est appelée fonction de répartition marginale (de même pour la variable aléatoire Y).

ASPECTS MATHÉMATIQUES

Soit un **couple de variables aléatoires** (X, Y) et sa **fonction de répartition conjointe** :

$$F(a, b) = P(X \leq a, Y \leq b).$$

La fonction de répartition marginale de X est égale à :

$$\begin{aligned} F_X(a) &= P(X \leq a) \\ &= P(X \leq a, Y < \infty) \\ &= P\left[\lim_{b\to\infty}(X \leq a, Y \leq b)\right] \\ &= \lim_{b\to\infty} P(X \leq a, Y \leq b) \\ &= \lim_{b\to\infty} F(a,b) \\ &= F(a,\infty). \end{aligned}$$

De façon analogue, la fonction de répartition marginale de Y est égale à :

$$F_Y(b) = F(\infty, b).$$

MOTS CLÉS

Couple de variables aléatoires *(Pair of random variables)*
Fonction de répartition *(Distribution function)*
Fonction de répartition conjointe *(Joint distribution function)*

FRACTION DE PLAN D'EXPÉRIENCE FACTORIELLE

Fractional factorial design

Une fraction de **plan d'expérience** factorielle est une **expérience factorielle** dans laquelle n'est réalisée qu'une fraction des combinaisons de niveaux de facteurs.

On utilise ce genre de plans lorsqu'une **expérience** comporte de nombreux facteurs et/ou de nombreux niveaux de facteurs. Cela permet de réduire le nombre d'unités expérimentales nécessaires.

HISTORIQUE

La notion de fraction de plan d'expérience factorielle a été introduite d'abord par Frank Yates en 1935 et c'est ensuite D.J. Finney qui l'a présentée de manière formelle.

EXEMPLE

Pour une **expérience factorielle** comprenant deux facteurs, chacun à trois niveaux, le nombre de combinaisons possibles est égal à 9. Si l'on supprime quatre observations, il reste une fraction de plan d'expérience factorielle que l'on peut, par exemple, schématiser ainsi :

		Facteur B		
		1	2	3
Facteur A	1	Y_{11}	Y_{12}	Y_{13}
	2	Y_{21}		
	3	Y_{31}		

où Y_{ij} représente l'observation faite au i-ème niveau du facteur A et au j-ème niveau du facteur B.

MOTS CLÉS

Expérience *(Experiment)*
Plan d'expérience *(Design of experiments)*

RÉFÉRENCES

Finney, D.J. (1945). The fractional replication of factorial arrangements. *Annals of Eugenics* (London), 12, pp. 291-301.

Yates, F. (1935). Complex Experiments. *Supplement to the Journal of the Royal Statistical Society*, 2, pp. 181-247.

FRÉQUENCE

Frequency

La fréquence ou fréquence absolue correspond au nombre d'apparitions d'une **observation** ou du résultat d'une **expérience**.

On distingue la fréquence absolue de la fréquence relative : la fréquence relative d'une observation est définie par le rapport du nombre d'apparitions au nombre total d'observations.

Pour certaines analyses, il peut être souhaitable de connaître le nombre d'observations dont la **valeur** est « inférieure ou égale », ou « supérieure » à une valeur donnée. On appelle fréquence cumulée le

nombre de valeurs inférieures ou égales à cette valeur fixée.

De même, on peut s'intéresser au **pourcentage** d'observations dont la valeur est « inférieure ou égale », ou « supérieure » à une valeur donnée. On appelle fréquence relative cumulée la proportion de valeurs inférieures ou égales à cette valeur.

EXEMPLE

On considère par exemple les seize observations suivantes qui représentent les tailles (en cm) d'un groupe d'individus :

174	169	172	174
171	179	174	176
177	161	174	172
177	168	171	172

La fréquence de l'observation 171 cm est de 2 ; tandis que la fréquence de l'observation 174 cm est de 4.

Dans cet exemple, la fréquence relative de l'**événement** 174 est de $\frac{4}{16} = 0,25$ (ce qui représente 25 % des observations), 4 étant la fréquence absolue et 16 le nombre d'observations.

MOTS CLÉS

Courbe de fréquences *(Frequency curve)*
Distribution de fréquences *(Frequency distribution)*
Tableau de fréquences *(Frequency table)*

FRÉQUENCES, POLYGONE DE

Frequency polygon

Voir **polygone de fréquences**.

FRÉQUENCES, TABLEAU DE

Frequency table

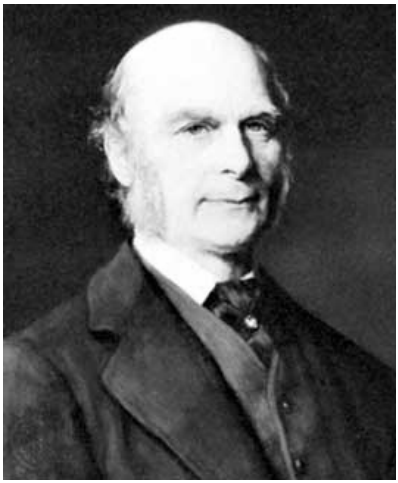
Voir **tableau de fréquences**.



G

GALTON FRANCIS

Galton Francis



Francis Galton né en 1822 en Angleterre, à proximité de Birmingham, est issu d'une famille d'intellectuels : Charles Darwin est son cousin et son grand-père était membre de la Royal Society.

Son intérêt pour la science se manifesta d'abord dans les domaines de la géographie et la météorologie. Élu membre de la Royal Society en 1860, il commença à s'intéresser à la génétique et aux méthodes statistiques dès 1864.

F. Galton fut un ami proche de **Karl Pearson** qu'il aida financièrement à fonder la revue *Biometrika*. Son Eugenics Record Office fusionna avec le laboratoire de biométrie de Karl Pearson de l'University College de Londres et prit bientôt le nom de Galton Laboratory.

Il mourut en 1911 et laissa plus de 300 publications dont 17 livres, notamment sur les méthodes statistiques relatives à l'**analyse de régression** et à la notion de **corrélation** qui lui est attribuée.

Quelques ouvrages et articles principaux de Francis Galton :

1869 *Hereditary Genius : An Inquiry into its Laws and Consequences.* The Macmillan Company, London.

1877 Typical Laws in Heredity. *Nature*, 15.

1889 *Natural Inheritance.* The Macmillan Company, London.

1907 *Probability, the Foundation of Eugenics.* Clarendon Press, Oxford.

1908 *Memories of my Life.* Methuen, London.

1914-1930 *The Life, Letters and Labours of Francis Galton.* Edited by K. Pearson. Cambridge University Press, Cambridge.

MOTS CLÉS

Analyse de régression *(Regression analysis)*
Coefficient de corrélation *(Correlation coefficient)*

GAMMA, LOI

Gamma distribution

Voir **loi gamma**.

GAUSS CARL FRIEDRICH

Gauss Carl Friedrich

Carl Friedrich Gauss naquit en 1777 à Brunswick en Allemagne. Il devint très rapidement un astronome et mathématicien renommé, si bien qu'actuellement il est toujours considéré comme l'un des plus grands mathématiciens de tous les temps, au même titre qu'Archimède et Newton.

Dès 1799, il avait acquis le titre de docteur. Il travailla quelque temps à l'Université de Helmsted

puis en 1807, il se rendit à Göttingen où il fut nommé directeur de laboratoire. Il passa sa vie à Göttingen où il mourut en 1855.



Ses contributions à la science et en particulier à la physique sont d'une très grande importance. En **statistique**, ses travaux concernent la théorie de l'**estimation** et on lui doit notamment la méthode des **moindres carrés** et le développement de la **loi normale** pour les problèmes d'erreurs de mesure.

Quelques ouvrages et articles principaux de Carl Friedrich Gauss :

1803-1809 *Disquisitiones de Elementis Ellipticis Pallidis.* Werke, 6, pp. 1-24.

1809 *Theoria Motus Corporum Coelestium.* Werke 7.

1816 *Bestimmung der Genauigkeit der Beobachtungen.* Werke 4, pp. 109-117. In C.F. Gauss Werke (publié en 1880). Dieterichsche Universitäts-Druckerei, Göttingen.

1821, 1823 et 1826 *Theoria Combinationis Observationum Erroribus Minimis Obnoxiæ*, partie 1, 2 et suppl. Werke 4, pp. 1-108.

1823 *Anwendungen der Wahrscheinlichkeitsrechnung auf eine Aufgabe der praktischen Geometrie.* Werke 9, pp. 231-237.

1855 *Méthode des moindres carrés : mémoires sur la combinaison des observations.* Traduction française de l'œuvre de C.F. Gauss par J. Bertrand (autorisée par C.F. Gauss lui-même). Mallet-Bachelier, Paris.

1957 *Gauss' Work (1803-1826). On The Theory of Least Squares.* Traduction anglaise par H.F. Trotter. Statistical Techniques Research Group, Technical Report, N° 5. Princeton University Press, Princeton NJ.

MOTS CLÉS

Théorème de Gauss-Markov *(Gauss-Markov theorem)*
Loi normale *(Normal distribution)*
Moivre Abraham de *(Moivre Abraham de)*

GAUSS-MARKOV, THÉORÈME DE

Gauss-Markov theorem

Voir **théorème de Gauss-Markov**.

GÉNÉRATION DE NOMBRES ALÉATOIRES

Generation of random numbers

Une génération de nombres aléatoires consiste en la réalisation d'une suite de nombres aléatoires indépendants les uns des autres. Ils sont utilisés pour effectuer des simulations et pour résoudre des problèmes analytiquement insolubles par la **méthode de Monte Carlo**.

HISTORIQUE

Les premiers travaux concernant les nombres aléatoires sont ceux de John von Neumann (1951) lorsqu'il introduit la méthode de *middle four square.*

Lehmer (1951) a introduit la méthode des congruences simples encore utilisée aujourd'hui.

Sowey a publié deux articles dans lesquels il établit la bibliographie de tout ce qui est paru sur les nombres aléatoires, le premier pour la période de 1927 à 1972 et le second pour celle allant de 1972 à 1976.

En 1996, Yadolah Dodge a proposé une méthode naturelle de génération de nombres aléatoires à partir des décimales de π.

ASPECTS MATHÉMATIQUES

Il existe différentes manières de générer des nombres aléatoires. La première consiste à tirer d'une urne contenant des billets numérotés, un billet à la fois et ensuite de le remettre dans l'urne.

Nous pouvons utiliser certains systèmes physiques pour générer des nombres aléatoires : par exemple le jeu de la roulette. Nous pouvons lister les nombres obtenus ci-dessus dans une table appelée table de nombres aléatoires (voir annexe A). Ces tables sont utiles si nous souhaitons effectuer plusieurs **simulations** pour des **modèles** différents et tester lequel est le meilleur. Dans ce cas, la même suite de nombres aléatoires peut être employée plusieurs fois.

Certains ordinateurs utilisent des circuits électroniques ou des moyens électro-mécaniques pour fournir des suites de nombres aléatoires. Toutefois, ces listes n'étant pas reproductibles, John von Neumann et D.H. Lehmer ont proposé des méthodes fournissant des suites de nombres qui, bien qu'entièrement déterministes, avaient l'apparence de nombres aléatoires. Une telle suite s'appelle suite de nombres pseudo-aléatoires.

Voici la méthode de Lehmer : ces nombres pseudo-aléatoires sont calculés à partir d'une formule algébrique déterminée. Par exemple la méthode de Lehmer permet de calculer la suite de nombres pseudo-aléatoires $x_0, x_1, \ldots, x_n, \ldots$ en calculant :

$$x_i = (a \cdot x_{i-1} + b) \pmod{m} \quad i = 1, 2, \ldots$$

avec $x_0 = b$ et a, b, m donnés.

Le choix de a, b, m influence la qualité de l'**échantillon** de nombres pseudo-aléatoires.

L'**indépendance** entre les nombres pseudo-aléatoires implique que la longueur du cycle, c'est-à-dire la quantité de nombres générés avant d'obtenir la même séquence, soit suffisamment longue. Pour obtenir le cycle le plus long possible, il faut que les critères suivants soient satisfaits :

— a ne doit pas diviser m ;

— x_0 ne doit pas diviser m ;

— m doit être grand.

DOMAINES ET LIMITATIONS

Dans le cas où la **simulation** requiert des observations provenant de distributions qui ne sont pas uniformes, il existe plusieurs techniques permettant de générer n'importe quelle loi à partir d'une **loi uniforme**. Certaines de ces méthodes sont applicables à de nombreuses distributions ; d'autres plus spécifiques sont utilisées pour une loi particulière.

EXEMPLE

Calculons avec la méthode de Lehmer les premiers nombres pseudo-aléatoires de la suite engendrée avec $a = 33$, $b = 7$, et $m = 1\,000$:

$$\begin{aligned} x_0 &= 7 \\ x_1 &= 7 \cdot 33 \pmod{1\,000} = 231 \\ x_2 &= 231 \cdot 33 \pmod{1\,000} = 623 \\ x_3 &= 623 \cdot 33 \pmod{1\,000} = 559 \\ &\text{etc.} \end{aligned}$$

Nous obtenons une suite de nombres r_i dont la distribution ne diffère pas significativement d'une distribution uniforme en divisant x_i par m :

$$r_i = \frac{x_i}{m} \; i = 1, 2, \ldots$$

Ainsi : $r_0 = 0,007$
$r_1 = 0,231$
$r_2 = 0,623$
$r_3 = 0,559$
etc.

MOTS CLÉS

Loi uniforme *(Uniform distribution)*
Méthode de Monte-Carlo *(Monte Carlo method)*
Nombre aléatoire *(Random number)*
Simulation *(Simulation)*

RÉFÉRENCES

Lehmer, D.H. (1951). Mathematical methods in large scale computing units. *Annals Comp. Lab. Harvard* 26, pp. 141-146.

Ripley, B.D. (1990). Thoughts on pseudorandom number generators. *Journal of Computational and Applied Mathematics*, 31, pp. 153-163.

Von Neumann, J. (1951). Various Techniques Used in Connection with Random Digits. *National Bureau of Standards symposium (NBS) Applied Mathematics Series* 12, pp. 36-38, National Bureau of Standards, Washington, D.C.

GÉOMÉTRIQUE, LOI

Geometric distribution

Voir **loi géométrique**.

GÉOSTATISTIQUE

Geostatistics

La géostatistique est une méthodologie de la **statistique spatiale** utilisée pour la prédiction optimale des parties inobservables d'un processus décrit par des données spatiales dont l'emplacement est connu. Les modèles hiérarchiques spatiaux suivent le principe suivant : modéliser localement, analyser globalement. À tous les niveaux de la hiérarchie, des modèles conditionnels simples sont construits (modélisation locale). Le résultat est un modèle conjoint qui peut être très complexe mais pour lequel l'analyse est possible (l'analyse globale). La géostatistique comprend l'ensemble de la théorie, des techniques et des applications statistiques appliquées pour analyser et prévoir les valeurs d'une variable distribuée dans l'espace et éventuellement dans le temps.

HISTORIQUE

Le terme de géostatistique a été utilisé pour la première fois par J.F. Hart (1954) dans un contexte géographique, afin de souligner l'utilisation des techniques statistiques particulières pour des observations ayant une distribution régionale. Les premiers concepts de la géostatistique ont été formulés par G. Matheron (1965) à l'École des mines de Paris afin d'estimer les réserves dans un minerai à partir des données (observations) spatialement distribuées.

DOMAINES ET LIMITATIONS

La géostatistique est utilisée dans un spectre de domaines très large : lors de prévisions (précipitations, porosité du sol, concentration en métaux lourds), de traitement des images satellitaires, etc.

MOTS CLÉS

Données spatiales *(Spatial data)*
Krigéage *(Krieginig)*
Statistique spatiale *(Spatial statistics)*

RÉFÉRENCES

Cressie, Noel A.C. (1991). *Statistics for Spatial Data.* John Wiley & Sons, New York.

Goovaerts, P. (1997). *Geostatistics for Natural Resources Evaluation.* Oxford : Oxford University Press.

Hart, J.F. (1954). Central tendency in areal distributions, *Economic Geography*, 30 pp. 48-59.

Matheron, G. (1965). *Les variables régionalisées et leur estimation.* Masson. Paris.

Oliver, M.A. and Webster, R. (1990). Kriging : a method of interpolation for geographical information system, *Int. J. Geographical Information Systems*, Vol. 4, No. 3, pp. 313-332.

Pannatier, Y. (1996). *Variowin 2.2. Software for Spatial Data Analysis in 2D.* Springer-Verlag, New York.

GINI CORRADO

Gini Corrado



Corrado Gini (1884-1965) fut licencié en droit en 1905 et durant ses études, il suivit des cours de mathématiques et de biologie. En 1909, il devint professeur temporaire de statistique à l'Université de Cagliari et accéda à la chaire de statistique de cette même université en 1920. En 1913, il enseigna à l'Université de Padua et, dès 1927, il tint la chaire de statistique de l'Université de Rome.

Entre 1926 et 1932, il fut président du Central Institute of Statistics (ISTAT).

Il fonda deux journaux, le premier, en 1920, *Metron*, était un journal international de statistique et le second, en 1934, *Genus*, était un journal du Comité italien pour l'étude des problèmes de population.

D'un point de vue strictement statistique, les contributions de Gini se rapportent principalement aux valeurs moyennes et à la variabilité ainsi qu'une association entre les variables aléatoires. Il a aussi contribué de manière originale à l'économie, la sociologie, la démographie et la biologie.

Quelques ouvrages et articles principaux de Corrado Gini :

1910 Indici di concentratione e di dependenza. *Atti della III Riunione della Societa Italiana per il Progresso delle Scienze*. In ref. 12, pp. 3-210.

1912 Variabilità e mutabilità. *Studi economico-giuridici*, Università di Cagliari, III 2a. In ref. 12, pp. 211-382.

GOSSET WILLIAM SEALY

Gosset William Sealy

William Sealy Gosset, dit « Student », naquit à Canterbury en 1876 ; il entreprit des études de mathématiques et de chimie au New College, d'Oxford. En 1899, il commença à travailler comme brasseur pour le compte des Brasseries Guinness à Dublin. Cette société, qui favorisait la recherche, mettait des laboratoires à la disposition de ses brasseurs et de ses chimistes et, en 1900, elle ouvrit le Guinness Research Laboratory dirigé par l'un des plus grands chimistes de l'époque, Horace Brown. Les travaux effectués dans ce laboratoire portaient sur la qualité et les coûts de diverses variétés d'orge et de houblon.

C'est dans cet environnement que se développera l'intérêt de W.S. Gosset pour les statistiques.

À Oxford, W.S. Gosset avait étudié les mathématiques, il fut donc sollicité par ses collègues de Guinness lorsque se posaient certains problèmes. C'est ainsi qu'il fut amené à étudier la théorie des erreurs et, dans cette optique, à consulter **Karl Pearson** qu'il rencontra en juillet 1905.

La taille toujours réduite des données expérimentales constituait la difficulté majeure à laquelle se heurtait W.S. Gosset. Ainsi se trouva-t-il confronté à la nécessité de trouver une méthode correcte de traitement des petits échantillons.

En 1906-1907, W.S. Gosset passa une année à collaborer avec Karl Pearson, dans le laboratoire que celui-ci dirigeait à l'Université de Londres. Il y travailla à résoudre ses questions concernant l'erreur probable de la **moyenne**.



En 1907, W.S. Gosset fut nommé responsable de la Brasserie expérimentale de Guinness, et utilisa notamment la **table de Student** qu'il avait définie dans une expérience visant à déterminer la meilleure variété d'orge. Guinness l'autorisa à publier ses articles sous un pseudonyme, qui pouvait être « Pupil » ou « Student », il choisit le second.

Il mourut en 1937, laissant d'importants écrits, tous publiés sous le nom de « Student ».

Les travaux de Gosset n'eurent qu'un retentissement très limité pendant plusieurs années et peu de scientifiques utilisaient sa table à l'exception des chercheurs des brasseries Guinness et ceux de la station expérimentale de Rothamsted où **Ronald Aylmer Fisher** devint statisticien.

Quelques ouvrages et articles principaux de William Sealy Gosset :

1907 On the error of counting with a haemacytometer. *Biometrika*, 5, pp. 351-360.

1908 The Probable Error of a Mean. *Biometrika*, 6, pp. 1-25

1925 New tables for testing the significance of observations. *Metron*, 5, pp. 105-120.

1942 *Student's collected papers* (ed. E.S. Pearson and J. Wishart, with a foreward by L. McMullen). Issued by the Biometrika Office, University College London and printed in Great Britain at the University Press, Cambridge. Reprinted 1958.

MOTS CLÉS

Loi de Student *(Student distribution)*
Table de Student *(Student table)*
Test de Student *(Student test)*

RÉFÉRENCES

Fisher-Box, J. (1987). Guinness, Gosset, Fisher, and small samples. *Statistical Science*, 2, pp. 45-52.

McMullen, L. and Pearson, E.S. (1970). *William Sealy Gosset, 1876-1937. Studies in the history of statistics and probability.* Vol. I. Edited by E.S. Pearson et M. Kendall. Charles Griffin & Company Ltd.

Plackett, R.L. and Barnard, G.A. (editors). (1990). *Student: A Statistical Biography of William Sealy Gosset.* Based on writings by E.S. Pearson. Clarendon Press, Oxford.

GRANDS NOMBRES, LOI DES

Law of large numbers

Voir **loi des grands nombres**.

GRAPHIQUE DE DISPERSION

Scatter plot

On obtient un graphique de dispersion en rapportant sur un graphe les données d'un **échantillon**.

En deux dimensions, un graphique de dispersion permet de représenter n paires d'**observations** (x_i, y_i), $i = 1, 2, \ldots, n$.

ASPECTS MATHÉMATIQUES

Un graphique de dispersion s'obtient en plaçant les couples de valeurs observées dans un système d'axes.

Pour deux **variables** X et Y, on commence par placer les couples de points correspondants (x_1, y_1), (x_2, y_2), $\ldots$, (x_n, y_n) dans un système d'axes rectangulaires.

On place généralement la **variable dépendante** Y sur l'axe vertical (l'ordonnée) et la **variable indépendante** X sur l'axe horizontal (l'abscisse).

DOMAINES ET LIMITATIONS

Le graphique de dispersion est un outil puissant et utile qui est souvent utilisé en **analyse de régression**. Chaque point représente alors un couple de **valeurs** observées de la **variable dépendante** et de la **variable indépendante**.

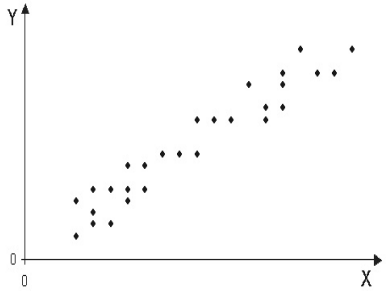
Il permet en particulier de déterminer de manière visuelle s'il existe une **relation** entre les deux variables avant de choisir un **modèle** approprié.

De plus, ces graphiques de dispersion sont très utiles dans l'**analyse des résidus** puisqu'ils permettent de vérifier l'adéquation du modèle.

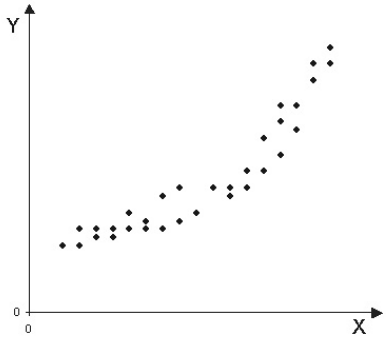
EXEMPLES

Voici deux exemples fictifs de graphiques de dispersion représentant les couples d'observations (x_i, y_i) relatives à deux variables quantitatives X et Y :

Dans cet exemple, la distribution des observations semble indiquer une **relation** linéaire entre les deux variables :



Le distribution des observations sur le graphique suivant semble indiquer une relation non linéaire entre X et Y :



MOTS CLÉS

Analyse des résidus *(Analysis of residuals)*
Régression, analyse de *(Regression analysis)*

GRAPHIQUE GAUSSO-ARITHMÉTIQUE

Normal probability plot

Un graphique gausso-arithmétique permet de vérifier si un ensemble de données est distribué selon une **loi normale**. Lorsque tel est le cas, il est possible d'en estimer la **moyenne** et l'**écart type** à partir du graphique.

On reporte les fréquences cumulées en ordonnée sur une échelle gaussienne, graduée en portant la **valeur** de la **fonction de répartition** $F(t)$ de la loi normale (centrée réduite) en face du point situé à la **distance** t de l'origine. En abscisse, on représente les observations sur une échelle arithmétique.

En pratique, on utilise les papiers gausso-arithmétiques vendus dans le commerce.

ASPECTS MATHÉMATIQUES

Sur un système d'axes, on fait le choix d'une origine et d'une unité de longueur d'après les observations à représenter. On gradue alors arithmétiquement l'abscisse comme sur du papier millimétré.

En ordonnée, on reporte les valeurs de la **fonction de répartition** $F(t)$ de la **loi normale** à la hauteur t d'une origine fictive placée à mi-hauteur de la feuille.

Ainsi, on place :

$F(t)$	t
$0,9772$	2
$0,8412$	1
$0,5000$	0
$0,1588$	-1
$0,0$	

et on ne garde que l'échelle située à gauche qui sera graduée de 0 à 1 (ou de 0 à 100 si l'on note en %).

Dans la pratique, on procède de la manière suivante :

Soit un ensemble de n données ponctuelles qu'on suppose classées dans l'ordre croissant. On appelle x_1, x_2, …, x_n cette suite, et pour chaque abscisse x_i $(i = 1, \ldots, n)$, on calcule l'ordonnée

$$100 \cdot \frac{i - \frac{3}{8}}{n + \frac{1}{4}}$$

qu'on représente alors sur le graphique gausso-arithmétique.

On peut également reporter

$$100 \cdot \frac{i - \frac{1}{2}}{n}$$

en fonction des x_i sur le graphique. Cette deuxième approximation de la **fréquence** cumulée s'explique de la façon suivante : si l'on divise la surface située sous la courbe normale (et d'aire égale à 1) en n parties égales, on pourrait supposer que chacune de nos n observations se trouve dans l'une de ces parties. Ainsi la i-ème observation x_i (dans l'ordre croissant), est située au milieu de la i-ème plage ; ce qui en terme de fréquence cumulée correspond à

$$\frac{i - \frac{1}{2}}{n}.$$

On peut alors tester la normalité de la distribution observée en examinant l'alignement des points : si les points sont alignés, alors la distribution est normale.

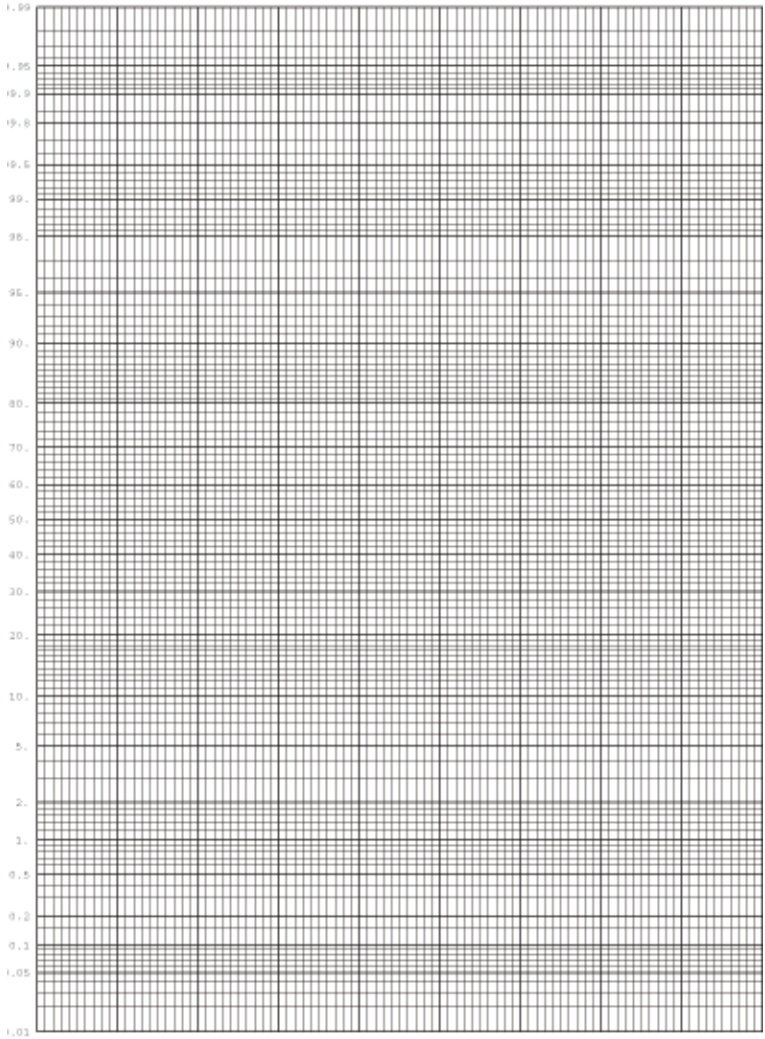
DOMAINES ET LIMITATIONS

En représentant les fréquences cumulées en fonction des valeurs de la **variable** statistique sur un graphique gausso-arithmétique, il est possible de vérifier si la distribution observée suit bien une **loi normale**. En effet, si tel est le cas, les points sont approximativement alignés.

Ainsi l'ajustement graphique d'une droite à l'ensemble des points permet :

- l'appréciation du caractère normal de la distribution observée de façon plus nette que sur un **histogramme** ;
- l'**estimation** graphique de la **moyenne** et de l'**écart type** de la distribution observée.

Papier gausso-arithmétique. Échelle verticale : 0,01 - 99,99 %



En effet, la droite des **moindres carrés** coupe l'ordonnée $0,50$ (ou 50 %) à l'estimation de la moyenne (qu'on peut lire sur l'abscisse).

L'écart type estimé est donné par l'inverse de la pente de la droite.

Pour déterminer cette dernière, il suffit de considérer deux points ; on choisit $(\hat{\mu}, 0,5)$ et le point d'ordonnée

$$0,9772 \left(= F(2) = F\left[\frac{X_{0,9772} - \hat{\mu}}{\hat{\sigma}}\right]\right)$$

dont l'abscisse $X_{0,9772}$ est lue sur la droite.

On obtient :

$$\begin{aligned} \frac{1}{\hat{\sigma}} &= \frac{F^{-1}(0,9772) - F^{-1}(0,5)}{X - \mu} \\ &= \text{pente de la droite des moindres carrés} \end{aligned}$$

d'où on tire finalement :

$$\hat{\sigma} = \frac{X_{0,9772} - \hat{\mu}}{2}.$$

EXEMPLE

Soient 10 pièces fabriquées par une machine X, tirées de façon aléatoire, aux diamètres suivants :

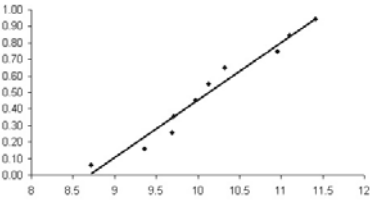
$$9,36\ ;\ 9,69\ ;\ 11,10\ ;\ 8,72\ ;\ 10,13\ ;$$
$$9,98\ ;\ 11,42\ ;\ 10,33\ ;\ 9,71\ ;\ 10,96.$$

Ces pièces, supposées distribuées selon une **loi normale**, sont alors classées dans l'ordre croissant et on calcule pour chaque diamètre

$$f_i = \frac{i - \frac{3}{8}}{10 + \frac{1}{4}},$$

i étant le numéro d'ordre de chacune des observations :

x_i	f_i
$8,72$	$5/82$
$9,36$	$13/82$
$9,69$	$21/82$
$9,71$	$29/82$
$9,98$	$37/82$
$10,13$	$45/82$
$10,33$	$53/82$
$10,96$	$61/82$
$11,10$	$69/82$
$11,42$	$77/82$



La **représentation graphique** de f_i par rapport à x_i sur un papier gausso-arithmétique fournit pour notre **échantillon** aléatoire :

— la **moyenne** estimée des diamètres : $\hat{\mu} = 10,2$;

— l'**écart type** estimé : $\hat{\sigma} = 0,9$.

MOTS CLÉS

Distribution de fréquences *(Frequency distribution)*
Fonction de répartition *(Distribution function)*
Fréquence *(Frequency)*
Loi normale *(Normal distribution)*

GRAPHIQUE LOGARITHMIQUE

Logarithmic plot

Un graphique logarithmique est une **représentation graphique** d'une **série chronologique**. Il est défini par deux échelles logarithmiques, en abscisse et en ordonnée.

En général, on utilise les papiers logarithmiques, dits « papiers log-log ».

ASPECTS MATHÉMATIQUES

Sur un système d'axes, on choisit une origine et on place une échelle logarithmique sur chacun des deux axes.

En général, la longueur du module, c'est-à-dire de l'unité de longueur sur l'échelle arithmétique correspondante, est la même sur chacun des deux axes, le nombre de modules pouvant être différent. L'utilisateur dispose d'une constante multiplicative sur chacun des axes pour en déterminer l'échelle.

DOMAINES ET LIMITATIONS

Sur un graphique logarithmique, les courbes représentant des fonctions puissances sont représentées par une droite.

En effet, considérons $y(x)$ la fonction puissance

$$y(x) = B \cdot x^{\alpha}$$

où B est une constante, et le **paramètre** α est appelé l'élasticité de la grandeur y par rapport à x :

$$\text{élasticité} = \frac{dy}{dx} \cdot \frac{x}{y}.$$

Cette fonction représente donc la variation d'un phénomène qui possède la propriété d'avoir, par rapport à x, une élasticité constante α. Cette élasticité est justement égale à la pente de la droite obtenue.

Cela se montre aisément :

Soit la fonction puissance :

$$y(x) = B \cdot x^{\alpha}$$

où B et α sont constantes, en prenant le logarithme,

$$\log y(x) = \log B + \alpha \cdot \log x,$$

$\log y(x)$ est donc bien linéaire par rapport à $\log x$.

Un graphique logarithmique permet :

- la lecture directe de la **variable** représentée ;
- la représentation de valeurs extrêmes qui « sortiraient » du graphe en cas d'utilisation d'un graphique arithmétique ;
- la détermination graphique de l'élasticité entre deux points : α étant la pente de la droite passant par les deux points ;
- la comparaison graphique de deux élasticités.

MOTS CLÉS

Représentation graphique *(Graphical representation)*
Série chronologique *(Time series)*
Test d'Anderson-Darling *(Anderson-Darling test)*

GRAPHIQUE QUANTITATIF

Quantitative graph

Les graphiques quantitatifs sont utilisés pour présenter et résumer des informations numériques provenant de l'étude d'une **variable qualitative catégorielle**.

Les types de graphiques quantitatifs les plus fréquemment utilisés sont :

— le **diagramme en barres** ;

— le **diagramme figuratif** ;

— le **diagramme circulaire**.

HISTORIQUE

Beniger et Robyn (1978) retracent le développement historique des diagrammes quantitatifs depuis le XVII^{e} siècle. Schmid et Schmid (1979) parlent des nombreuses formes de graphiques quantitatifs.

MOTS CLÉS

Diagramme circulaire *(Pie chart)*
Diagramme en barres *(Bar chart)*
Diagramme figuratif *(Pictogram)*

Représentation graphique *(Graphical representation)*

RÉFÉRENCES

Beniger, J.R. and Robyn, D.L. (1978). Quantitative Graphics in Statistics: A Brief History. *The American Statistician*, 32, pp. 1-11.

Schmid, C.F. and Schmid, S.E. (1979). *Handbook of Graphic Presentation.* 2nd ed. John Wiley & Sons, New York.

GRAPHIQUE SEMI-LOGARITHMIQUE

Semi-logarithmic plot

Un graphique semi-logarithmique est une **représentation graphique** de **série chronologique**.

Il est défini par une échelle arithmétique du temps t en abscisse et une échelle logarithmique en ordonnée ; c'est-à-dire qu'on reportera le logarithme de la **valeur** observée Y_t en ordonnée sur une échelle arithmétique.

En général, on utilise les papiers semi-logarithmiques imprimés.

ASPECTS MATHÉMATIQUES

Sur un système d'axes, on choisit une origine et une unité de longueur quelconques ; on gradue arithmétiquement l'abscisse comme sur du papier millimétré.

Sur l'axe des ordonnées, on peut modifier l'unité de longueur adoptée par le fabricant de papier semi-logarithmique qui correspond à la distance entre deux puissances successives de 10 sur l'échelle logarithmique. Cette distance qui n'est autre que l'unité de longueur sur l'échelle arithmétique correspondante est appelée le module.

On peut multiplier les valeurs imprimées en marge de l'échelle logarithmique par un même nombre ; cette constante multiplicative est choisie aussi simple que possible, de façon à ce que les valeurs obtenues soient d'une utilisation commode pour porter les points sur le graphique.

Les papiers semi-logarithmiques que l'on trouve habituellement dans le commerce ont 1, 2, 3 ou 4 modules.

DOMAINES ET LIMITATIONS

Le graphique semi-logarithmique présente les propriétés suivantes :

- les courbes exponentielles Y sont représentées par une droite. En effet, si $y(x)$ est une exponentielle

  $$y(x) = c \cdot a^x$$

  en prenant le logarithme

  $$\begin{aligned} \log y(x) &= \log c + x \cdot \log a \\ \text{ou} \quad \log y(x) &= m \cdot x + h \\ \text{où} \quad m &= \log a \ \text{ et } \ h = \log c, \end{aligned}$$

  on trouve bien une expression linéaire en x.

- Si $a = 1 + i$ où i est le taux d'intérêt annuel de Y et avec une somme d'argent de départ de Y_0 :

  $$Y = Y_0 \cdot (1 + i)^x$$

  nous donne la somme d'argent après x années.

  On peut également travailler en base népérienne (e), $y(x) = c \cdot \exp(r \cdot x)$ où $r = \ln a$ est appelé le taux de croissance instantané de la grandeur y. Dans ce cas précis, c'est le rapport de la variation instantanée $\frac{dy}{dx}$ à la **valeur** de y.

L'utilisation du graphique semi-logarithmique s'avère judicieuse :

- quand il existe de grandes différences de valeurs dans la **variable** pour éviter de sortir du graphe ;
- pour faire apparaître des variations relatives ;
- lorsque des cumuls de taux de croissance, qui feraient apparaître des exponentielles, doivent être représentés de façon linéaire.

MOTS CLÉS

Représentation graphique *(Graphical representation)*
Série chronologique *(Time series)*



H

HÁJEK JAROSLAV

Hájek Jaroslav



Né en 1926 et ingénieur statisticien de formation, Jaroslav Hájek obtient son doctorat en 1954. De 1954 à 1964, il travaille comme chercheur à l'Institut de mathématiques de l'Académie tchécoslovaque des sciences. Il rejoint ensuite la Charles University à Prague où il est professeur de 1966 jusqu'à sa mort prématurée en 1974. Les travaux principaux de J. Hájek portent sur la théorie des probabilités d'échantillonnage et sur la théorie des tests de rang. Il développa, en particulier, une théorie asymptotique de statistiques de rangs linéaires. Il fut le premier à appliquer le concept d'invariance dans la théorie asymptotique des tests de rang.

Quelques ouvrages et articles de Jaroslav Hájek :

1955 Some rank distributions and treir applications. *Cas. Pest. Mat.* 80, 17-31 (en tchèque) ; translation in Selected *Transl. Math. Statis. Probab.* 2 (1962), pp. 41-61.

1958 Some contributions to the theory of probability sampling. *Bulletin of the Institute of Intenational Statistics*, 36, pp. 127-134.

1961 Some extensions of the Wald-Wolfowitz-Noether theorem. *Annals of Mathematical Statistics*, 32, pp. 506-523.

1964 Asymtotic theory of rejective sampling with varying probabilities from a finite population. *Annals of Mathematical Statistics*, 35, pp. 1419-1523.

1965 Extension of the Kolmogorov-Smirnov test to regression alternatives. *Bernoulli-Bayes-Laplace : Proceedings of an International Seminar*, 1963 (J. Neyman et L. LeCam, eds.), pp. 45-60. Springer-Verlag, Berlin.

1981 *Sampling from a Finite Population.* (Ed. by V. Dupac) Marcel Dekker, New York.

HISTOGRAMME

Histogram

L'histogramme est une **représentation graphique** d'une distribution sur des données groupées consistant en un ensemble de rectangles. Il appartient à la famille des graphiques de fréquences.

Les données sont groupées sous la forme de plusieurs intervalles appelés intervalles de classe. À chaque intervalle de classe correspond une fréquence (ou une fréquence relative).

Un rectangle de l'histogramme représente un classe de données et sa base correspond à la longueur de l'intervalle et la surface du rectangle traduit la fréquence de chaque classe.

HISTORIQUE

Le terme d'histogramme a été utilisé pour la première fois par **Karl Pearson** en 1895.

Voir également **représentation graphique**.

ASPECTS MATHÉMATIQUES

La première étape de la construction d'un histogramme consiste à présenter les données sous la forme d'un **tableau de fréquences** (afin de les grouper en classes) : il faut d'abord procéder à la détermination des différents intervalles de classe et ensuite distribuer les fréquences à l'intérieur de ces classes. On obtient alors un tableau de fréquences que l'on peut représenter graphiquement de la manière suivante : on place sur l'**axe** horizontal les différentes classes, puis on calcule la hauteur que doit prendre chaque rectangle pour que la surface soit égale à la fréquence de la classe considérée.

DOMAINES ET LIMITATIONS

Les histogrammes sont utilisés pour condenser un ensemble de données sous une forme visuelle facile à comprendre. Ils permettent d'extraire certaines caractéristiques générales telles que des valeurs typiques, l'étendue ou la forme des données.

La forme d'un histogramme peut suggérer des modèles de **probabilité** (**loi normale, loi lognormale**, etc.).

À l'aide d'un histogramme on peut aussi détecter un comportement inattendu ou des valeurs anormales.

C'est en économie que ce type de **représentation graphique** est le plus utilisé, mais comme c'est une manière simple de visualiser un ensemble de données, on l'utilise dans tous les domaines.

On peut aussi construire un histogramme à partir d'une **fréquence** relative comme axe vertical ; la hauteur de chaque rectangle est égale à la fréquence relative de la classe correspondante divisée par la longueur de cet **intervalle**. Dans ce cas, l'aire totale sous l'histogramme est égale à 100 % (= 1).

EXEMPLE

Voici les données brutes des moyennes annuelles des précipitations dans 69 villes des États-Unis. À partir de ces données, nous allons former un **tableau de fréquences** et construire un histogramme.

Moyennes annuelles des précipitations dans 69 villes des États-Unis (en pouces)

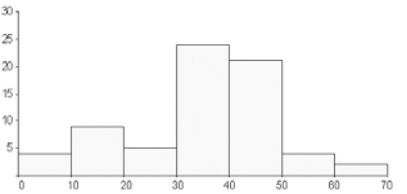
Mob.	67,0	Chic.	34,4	St.L.	35,9
Sacr.	17,2	Loui.	43,1	At.C.	45,5
Wash.	38,9	Detr.	31,0	Char.	42,7
Boise	11,5	K.C.	37,0	Col.	37,0
Wich.	30,6	Conc.	36,2	Prov.	42,8
Bost.	42,5	N.Y.	40,2	Dall.	35,9
Jack.	49,2	Clev.	35,0	Norf.	44,7
Reno	7,2	Pitt.	36,2	Chey.	14,6
Buff.	36,1	Nash.	46,0	L.R.	48,5
Cinc.	39,0	Burl.	32,5	Hart.	43,4
Phil.	39,9	Milw.	29,1	Atl.	48,3
Memp.	49,1	Phoen.	7,0	Indi.	38,7
S.L.	15,2	Denv.	13,0	Port.	40,8
Char.	40,8	Miami	59,8	Dul.	30,2
Jun.	54,7	Peor.	35.1	G.Fls	15,0
S.F.	20,7	N.Or.	56,8	Albq	7,8
Jack.	54,5	S.S.M.	31,7	Ral.	42,5
Spok.	17,4	L.A.	14,0	Omaha	30,2
Ok.C.	31,4	Wilm.	40,2	Alb.	33,4
Col.	46,4	Hon.	22,9	Bism.	16,2
ElPaso	7,8	D.Mn.	30,8	Port.	37,6
S-Tac	38,8	Balt.	41,8	S.Fls	24,7
S.J.	59,2	Mn-SP	25,9	Hstn.	48,2

Source : Statistical Abstract of the United States, 1975.

Ces données peuvent être représentées par le tableau de fréquences suivant :

Classes	Fréquences
0 – 10	4
10 – 20	9
20 – 30	5
30 – 40	24
40 – 50	21
50 – 60	4
60 – 70	2
Total	69

On construit alors l'histogramme :



Sur l'axe horizontal, on reproduit les classes et sur l'axe vertical les fréquences (les classes ayant la même longueur).

MOTS CLÉS

Diagramme stem and leaf *(Stem and leaf diagram)*
Distribution de fréquences *(Frequency distribution)*
Ogive *(Ogive)*
Polygone de fréquences *(Frequency polygon)*
Représentation graphique *(Graphical representation)*
Tableau de fréquences *(Frequency table)*

RÉFÉRENCES

Dodge, Y. (1986). Some Difficulties Involving Nonparametric Estimation of a Density Function. *Journal of Official Statistics*, vol. 2, n° 2, pp. 193-202.

Freedman, D., Pisani, R. and Purves, R. (1978). *Statistics*. Norton & Company, New York.

Pearson, K. (1895). 1948. Contributions to the Mathematical Theory of Evolution. II: Skew Variation in Homogeneous Material. pp. 41-112 in Karl Pearson, *Karl Pearson's Early Statistical Papers*. Cambridge University Press, Cambridge. First published in *Philosophical Transactions of the Royal Society of London*. Series A, Vol. 186, pp. 343-414.

HOMOGÉNÉITÉ, TEST D'

Homogeneity test

Voir **test d'homogénéité**.

HOTELLING HAROLD

Hotelling Harold

Harold Hotelling (1895-1973) est considéré comme un des plus grands pionniers dans le domaine des mathématiques économiques des années 1920 à 1930. Il introduisit le test multivarié T^2, l'analyse des composantes principales et les corrélations canoniques.

Il effectua ses études à l'Université de Washington, où il fut licencié ès lettres en journalisme en 1919, puis il entra à l'Université de Princeton pour faire son doctorat en mathématiques qu'il obtint en 1924. La même année, il commença d'enseigner à l'Université de Stanford. Ses premières applications des mathématiques aux sciences sociales concernèrent le journalisme et les sciences politiques, puis la population et les ravitaillements.



En 1931, il se déplaça à l'Université de Colombia où il participa activement à la création d'un département de statistiques. Durant la Deuxième Guerre mondiale, il appliqua les statistiques à l'armée.

En 1946, il fut recruté par l'Université de la Caroline du Nord à Chapel Hill afin d'y créer un département de statistiques.

Quelques ouvrages et articles principaux de Harold Hotelling :

1933 Analysis of a Complex of Statistical Variables with Principal Components. *Journal of Educational Psychology*, 24, pp. 417-441 and pp. 498-520.

1936 Relation Between Two Sets of Variates. *Biometrika*, 28, pp. 321-377.

HUBER PETER J.

Huber Peter J.



Peter J. Huber est né à Wohlen (Suisse) en 1934. Il a brillamment effectué ses études et son doctorat en mathématiques à l'École polytechnique fédérale de Zurich où il reçut la médaille d'argent pour la qualité scientifique de sa thèse. Il entama ensuite une carrière impressionnante. Tout d'abord professeur de statistique mathématique à l'École polytechnique fédérale de Zurich, il séjourna ensuite aux États-Unis dans les plus prestigieuses universités (Princeton, Yale, Berkeley) en tant que professeur invité. En 1977, il fut nommé professeur de mathématiques appliquées au Massachusetts Institute of Technology. Le professeur Huber est un statisticien mondialement connu. Il est membre de la prestigieuse American Academy of Arts and Sciences, de la Bernoulli Society et de la National Science Fundation aux États-Unis dont les membres étrangers sont extrêmement rares. Depuis son article « Robust estimation of location parameter » paru en 1964 il est considéré comme le fondateur de la statistique robuste.

Peter J. Huber a reçu le titre de docteur honoris causa de l'Université de Neuchâtel en 1994.

Quelques ouvrages et publications de Peter J. Huber :

1964 Robust estimation of a location parameter. *Annals of Mathematical Statistics*, 35, pp. 73-101.

1968 *Robust Statistical Procedures.* Cbms-Nsf Regional Conference Series in Applied Mathematics.

1981 *Robust Statistics.* John Wiley & Sons, New York.

1995 Robustness : Where are we now? *Student*, 1, pp. 75-86.

HYPERGÉOMÉTRIQUE, LOI

Hypergeometric distribution

Voir **loi hypergéométrique**.

HYPOTHÈSE

Hypothesis

Une hypothèse statistique est une assertion concernant la distribution d'une ou de plusieurs variables aléatoires. Elle peut notamment concerner les paramètres d'une distribution donnée ou encore la **loi de probabilité** de la **population** étudiée.

L'examen de la validité d'une hypothèse se fait par un **test d'hypothèse** à partir d'observations recueillies sur un **échantillon** de la population étudiée.

Dans le cas d'un test d'hypothèse sur la loi de probabilité suivie par la population étudiée, on appelle **hypothèse nulle** l'hypothèse selon laquelle

la population étudiée suit une loi de probabilité donnée. L'hypothèse selon laquelle la population ne suit pas la loi de probabilité donnée s'appelle **hypothèse alternative** (ou hypothèse contraire).

Dans le cas d'un test d'hypothèse sur les paramètres d'une distribution, on appelle hypothèse nulle l'hypothèse selon laquelle le paramètre étudié est égal à une **valeur** donnée. L'hypothèse selon laquelle le paramètre est différent de cette valeur s'appelle hypothèse alternative.

L'hypothèse nulle est généralement notée H_0 et l'hypothèse alternative H_1.

HISTORIQUE

Dans un **test d'hypothèse**, l'hypothèse que l'on désire tester porte le nom d'**hypothèse nulle**. C'est à **Ronald Aylmer Fisher** (1935) que l'on doit le terme « nulle ». En introduisant ce concept, il mentionna le problème bien connu de la dégustation de thé, dans lequel une dame prétendait être capable de reconnaître au goût si on avait versé en premier le thé ou le lait dans la tasse qu'elle buvait. L'hypothèse qu'il s'agissait de tester était que les appréciations données n'étaient aucunement influencées par l'ordre de préparation du thé.

On parlait d'hypothèse nulle principalement pour un **traitement** n'ayant pas d'effet ou lorsqu'il n'y avait pas de différence entre les effets des différents traitements.

On parle de nos jours surtout de l'hypothèse nulle pour mentionner l'hypothèse qu'on désire tester par opposition à l'**hypothèse alternative**.

Voir également **test d'hypothèses**.

EXEMPLE

De nombreux problèmes résultent de la répétition d'une **expérience** avec deux résultats possibles, par exemple le sexe d'un enfant à sa naissance.

On s'intéresse dans ce cas à la proportion de garçons et de filles dans une **population** donnée. Soit p la proportion de filles que l'on veut estimer à partir d'un **échantillon** observé. Pour déterminer si la proportion de garçons et de filles à la naissance est la même, on fera l'hypothèse statistique que $p = \frac{1}{2}$.

MOTS CLÉS

Analyse de variance *(Analysis of variance)*
Hypothèse alternative *(Alternative hypothesis)*
Hypothèse nulle *(Null hypothesis)*
Test d'hypothèse *(Hypothesis testing)*

RÉFÉRENCE

Fisher, R.A. (1935). *The Design of Experiments.* Oliver & Boyd, Edinburgh.

HYPOTHÈSE, TEST D'

Hypothesis testing

Voir **test d'hypothèse**.

HYPOTHÈSE ALTERNATIVE

Alternative hypothesis

Dans la réalisation d'un **test d'hypothèse**, l'hypothèse alternative est l'hypothèse qui sera acceptée si la règle de décision conduit à rejeter l'**hypothèse nulle** H_0.

L'hypothèse alternative est généralement notée par H_1.

HISTORIQUE

Voir **hypothèse** et **test d'hypothèse**.

ASPECTS MATHÉMATIQUES

Lors de la réalisation d'un **test d'hypothèse** concernant un **paramètre** d'une **population**, l'**hypothèse nulle** est présentée sous la forme suivante :

$$H_0 \quad : \quad \theta = \theta_0$$

où θ est le paramètre de la population à estimer et θ_0 la **valeur** présumée de ce paramètre.

L'hypothèse alternative peut alors prendre trois formes différentes :

1. H_1: $\theta > \theta_0$,
2. H_1: $\theta < \theta_0$,
3. H_1: $\theta \neq \theta_0$.

Dans les deux premiers cas, le test d'hypothèse est dit **test unilatéral**, alors que dans le troisième cas, il sera dit **test bilatéral**.

Dans la réalisation d'un test d'hypothèse concernant les paramètres de deux populations, l'hypothèse alternative peut aussi prendre trois formes différentes. Si l'hypothèse nulle pose l'égalité des deux paramètres θ_1 et θ_2 :

$$\begin{array}{rcl} H_0 & : & \theta_1 = \theta_2 \text{ ou} \\ H_0 & : & \theta_1 - \theta_2 = 0, \end{array}$$

l'hypothèse alternative pourra être :

- H_1: $\theta_1 > \theta_2$ ou H_1: $\theta_1 - \theta_2 > 0$,
- H_1: $\theta_1 < \theta_2$ ou H_1: $\theta_1 - \theta_2 < 0$,
- H_1: $\theta_1 \neq \theta_2$ ou H_1: $\theta_1 - \theta_2 \neq 0$.

Lors de la comparaison de plus de deux populations, l'hypothèse nulle suppose que les **valeurs** de tous les paramètres sont identiques. Si nous désirons comparer k populations, l'hypothèse nulle est donc la suivante :

$$H_0 \quad : \quad \theta_1 = \theta_2 = \ldots = \theta_k.$$

L'hypothèse alternative sera donc formulée comme suit :

H_1 : les valeurs de θ_i $(i = 1, \ldots, k)$ ne sont pas toutes identiques.

Cela signifie qu'il suffit que la valeur d'un paramètre soit différente des autres pour que l'hypothèse nulle soit rejetée au profit de l'hypothèse alternative.

EXEMPLES

Nous allons examiner les hypothèses alternatives sur trois exemples de tests d'hypothèse :

1. *Test d'hypothèse sur le pourcentage d'une population*

   Un candidat aux élections désire savoir s'il bénéficiera de plus de 50 % des voix.

   L'**hypothèse nulle** pour ce problème peut se poser ainsi :

   $$H_0 \quad : \quad \pi = 0,5,$$

   π étant le **pourcentage** de la **population** à estimer.

   Nous effectuons ici un **test unilatéral** à droite, ce qui nous permettra de répondre à la question que se pose ce candidat aux élections. L'hypothèse alternative sera donc la suivante :

   $$H_1 \quad : \quad \pi > 0,5.$$

2. *Test d'hypothèse sur la moyenne d'une population*

   Un fabricant de boulons veut tester la précision d'une nouvelle machine qui devrait lui fournir des boulons de 8 mm de diamètre.

   Nous pouvons poser l'hypothèse nulle suivante :

   $$H_0 \quad : \quad \mu = 8,$$

   μ étant la **moyenne** de la population à estimer.

   Nous allons effectuer un **test bilatéral** afin de déceler, d'une part, si le diamètre des pièces fabriquées est trop petit et, d'autre part, s'il est trop grand.

   L'hypothèse alternative peut donc être formulée de la manière suivante :

   $$H_1 \quad : \quad \mu \neq 8.$$

3. *Test d'hypothèse sur la comparaison des moyennes de deux populations*

   Une compagnie d'assurances a décidé d'équiper ses bureaux de micro-ordinateurs. Elle désire acheter ces micro-ordinateurs à deux fournisseurs différents pour autant qu'il n'y ait pas de différence significative de durabilité entre les deux marques. Elle teste donc le temps écoulé avant

la première panne sur un **échantillon** de micro-ordinateurs de chaque marque.

Selon l'hypothèse nulle, la **moyenne** du temps écoulé avant la première panne est la même pour chacune des deux marques :

$$H_0 \quad : \quad \mu_1 - \mu_2 = 0,$$

μ_1 et μ_2 étant les moyennes respectives des deux populations.

Comme nous ne savons pas *a priori* laquelle des moyennes sera la plus élevée, nous allons effectuer un test bilatéral. L'hypothèse alternative sera donc :

$$H_1 \quad : \quad \mu_1 - \mu_2 \neq 0.$$

MOTS CLÉS

Analyse de variance *(Analysis of variance)*
Hypothèse *(Hypothesis)*
Hypothèse nulle *(Null hypothesis)*
Test d'hypothèse *(Hypothesis testing)*

HYPOTHÈSE NULLE

Null hypothesis

Dans la réalisation d'un **test d'hypothèse**, l'hypothèse nulle est l'hypothèse que nous voulons tester. Elle est désignée par H_0.

L'**hypothèse** contraire est appelée **hypothèse alternative**. C'est l'hypothèse alternative qui sera acceptée si le test conduit à rejeter l'hypothèse nulle.

HISTORIQUE

Voir **hypothèse** et **test d'hypothèse**.

ASPECTS MATHÉMATIQUES

Lors de la réalisation d'un **test d'hypothèse** sur un **paramètre** d'une **population**, l'hypothèse nulle est généralement une supposition sur la **valeur** présumée de ce paramètre. L'hypothèse nulle se présente alors de la façon suivante :

$$H_0 \quad : \quad \theta = \theta_0 \ ,$$

θ étant le paramètre inconnu de la population et θ_0 la valeur présumée de ce paramètre.

Ce paramètre peut être par exemple la **moyenne** de la population.

Lorsque le test d'hypothèse a pour but de comparer deux populations, l'hypothèse nulle suppose généralement que les paramètres sont égaux :

$$H_0 \quad : \quad \theta_1 = \theta_2 \ ,$$

θ_1 étant le paramètre de la population dont est issu le premier échantillon et θ_2 celui dont est issu le deuxième échantillon.

EXEMPLES

Nous allons examiner les hypothèses nulles sur trois exemples de tests d'hypothèses :

1. *Test d'hypothèse sur le pourcentage d'un échantillon*

   Un candidat aux élections désire savoir s'il bénéficiera de plus de 50 % des voix.

   L'**hypothèse nulle** pour ce problème peut se poser ainsi :

   $$H_0 \quad : \quad \pi = 0,5,$$

   π étant le **pourcentage** de la **population** à estimer.

2. *Test d'hypothèse sur la moyenne d'une population*

   Un fabricant de boulons veut tester la précision d'une nouvelle machine qui devrait lui fournir des boulons de 8 mm de diamètre.

   Nous pouvons poser l'hypothèse nulle suivante :

   $$H_0 \quad : \quad \mu = 8,$$

   μ étant la **moyenne** de la population à estimer.

3. *Test d'hypothèse sur la comparaison des moyennes de deux populations*

   Une compagnie d'assurances a décidé d'équiper ses bureaux de micro-ordinateurs. Elle désire acheter ces micro-ordinateurs à deux fournisseurs différents pour autant qu'il n'y ait pas de différence significative de durabilité entre les deux marques. Elle teste donc le temps écoulé avant la première panne sur un échantillon de micro-ordinateurs de chaque marque.

   Selon l'hypothèse nulle, la moyenne du temps écoulé avant la première panne est la même pour chacune des deux marques :

   $$H_0 \quad : \quad \mu_1 - \mu_2 = 0,$$

   μ_1 etμ_2 étant les moyennes respectives des deux populations.

   Cette **hypothèse** peut également s'écrire

   $$H_0 \quad : \quad \mu_1 = \mu_2.$$

MOTS CLÉS

Analyse de variance *(Analysis of variance)*
Hypothèse *(Hypothesis)*
Hypothèse alternative *(Alternative hypothesis)*
Test d'hypothèse *(Hypothesis testing)*



I

INCIDENCE

Incidence

Voir **taux d'incidence**.

INDÉPENDANCE

Independence

La notion d'indépendance peut prendre deux sens : le premier concerne les événements d'une **expérience aléatoire**. On dit que deux événements sont indépendants si la réalisation du second événement ne dépend pas de la réalisation du premier.

L'indépendance de deux événements A et B peut être déterminée par un calcul de **probabilité** : si la probabilité de l'événement A lorsque B est réalisé est identique à la probabilité de A, alors A et B sont dits indépendants.

Le deuxième sens du mot indépendance concerne la **relation** qui peut exister entre des variables aléatoires. On peut tester l'indépendance de deux variables aléatoires par le test de Kendall et le test de Spearman.

HISTORIQUE

La notion d'indépendance a été implicitement utilisée bien avant qu'un ensemble formel d'axiomes des probabilités ait été établi. Selon Maistrov (1974), Cardano utilisait déjà la règle de la multiplication des probabilités ($P(A \cap B) = P(A) \cdot P(B)$). Maistrov mentionne également que les notions d'indépendance et de **dépendance** entre les événements étaient très familières à Blaise Pascal, Pierre de Fermat et Christiaan Huygens.

ASPECTS MATHÉMATIQUES

Mathématiquement, nous pouvons dire que l'**événement** A est indépendant de l'événement B si

$$P(A) = P(A \mid B)$$

où $P(A)$ est la **probabilité** de réalisation de l'événement A et $P(A|B)$ est la **probabilité conditionnelle** de l'événement A en fonction de la réalisation de B, c'est-à-dire la probabilité que A se réalise en sachant que B est réalisé.

En utilisant la définition de la probabilité conditionnelle, l'équation précédente peut s'écrire :

$$P(A) = \frac{P(A \cap B)}{P(B)}$$

ou encore

$$P(A) \cdot P(B) = P(A \cap B),$$

ce qui nous donne :

$$P(B) = \frac{P(A \cap B)}{P(A)} = P(B \mid A).$$

Par cette simple transformation, nous avons démontré que si A est indépendant de B, B est aussi indépendant de A. Nous dirons alors que les événements A et B sont indépendants.

DOMAINES ET LIMITATIONS

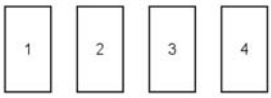
Si deux événements A et B sont indépendants, ils sont alors toujours compatibles. La **probabilité** que les deux événements A et B se réalisent simultanément peut être déterminée en multipliant les probabilités individuelles des deux événements :

$$P(A \cap B) = P(A) \cdot P(B).$$

Toutefois, si A et B sont dépendants, ils peuvent alors être compatibles ou incompatibles. Il n'existe pas de règle absolue.

EXEMPLES

- Considérons une boîte contenant quatre billets numérotés de 1 à 4 de laquelle nous allons effectuer deux tirages successifs sans remettre le billet tiré dans la boîte avant le second tirage.



L'**ensemble fondamental** peut être représenté par les 12 couples suivants :

$$\begin{aligned}\Omega &= \{(1,2),(1,3),(1,4),(2,1),(2,3),(2,4),\\ &\quad (3,1),(3,2),(3,4),(4,1),(4,2),(4,3)\}.\end{aligned}$$

Considérons les deux événements suivants :

$$\begin{aligned}A &= \text{« tirer le billet n}^{\text{o}}\text{ 3 au premier tirage »}\\ &= ((3,1),(3,2),(3,4)\}\end{aligned}$$

$$\begin{aligned}B &= \text{« tirer le billet n}^{\text{o}}\text{ 3 au second tirage »}\\ &= ((1,3),(2,3),(4,3)\}\end{aligned}$$

Ces deux événements sont-ils dépendants ou indépendants ?

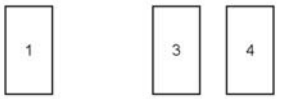
S'ils sont indépendants, la **probabilité** de l'événement B doit être identique quel que soit le résultat de l'événement A.

Si A est réalisé, il restera dans la boîte les billets suivants pour effectuer le second tirage :



La probabilité que B se réalise est alors nulle puisque le billet n$^{\text{o}}$ 3 n'est plus dans la boîte.

En revanche, si A n'est pas réalisé (par exemple, le billet n$^{\text{o}}$ 2 a été tiré au premier tirage), le second tirage se fera à partir de la boîte suivante :



La probabilité que B se réalise sera alors de $\frac{1}{3}$.

Nous pouvons donc conclure que l'événement B est dépendant de l'événement A, ou que les événements A et B sont dépendants.

Nous pouvons vérifier notre conclusion de la manière suivante : si A et B sont indépendants, l'équation suivante doit être vérifiée :

$$P(A \cap B) = P(A) \cdot P(B).$$

Dans notre exemple, $P(A) = \frac{3}{12}$ et $P(B) = \frac{3}{12}$.

L'intersection entre A et B est nulle. Nous avons donc :

$$P(A \cap B) = 0 \neq P(A) \cdot \frac{1}{16} = P(B).$$

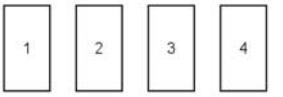
L'équation n'étant pas vérifiée, les événements A et B ne sont pas indépendants. Ils sont donc dépendants.

- Considérons à présent une expérience similaire, mais en remettant le billet tiré dans la boîte avant le second tirage.

L'ensemble fondamental de cette nouvelle **expérience aléatoire** est composé des 16 couples suivants :

$$\begin{aligned}\Omega &= \{(1,1),(1,2),(1,3),(1,4),(2,1),(2,2)\\ &\quad (2,3),(2,4),(3,1),(3,2),(3,3),(3,4),\\ &\quad (4,1),(4,2),(4,3),(4,4)\}.\end{aligned}$$

Dans ces conditions, le second tirage se fera à partir d'une boîte identique à celle du premier tirage :



Les événements A et B contiennent donc les résultats suivants :

$$\begin{aligned}A &= ((3,1),(3,2),(3,3),(3,4)\}\\ B &= \{(1,3),(2,3),(3,3),(4,3)\}\end{aligned}$$

La probabilité de A est donc $\frac{4}{16}$, ainsi que celle de B. L'intersection entre A et B est donnée par

l'événement $A \cap B = ((3,3)\}$; sa probabilité est donc de $\frac{1}{16}$. Nous avons donc :

$$P(A \cap B) = \frac{1}{16} = P(A) \cdot P(B)$$

et nous pouvons conclure que l'événement B est indépendant de l'événement A, ou que les événements A et B sont indépendants.

MOTS CLÉS

Coefficient de corrélation des rangs de Kendall *(Kendall rank correlation coefficient)*
Coefficient de corrélation des rangs de Spearman *(Spearman rank correlation coefficient)*
Compatibilité *(Compatibility)*
Dépendance *(Dependence)*
Événement *(Event)*
Probabilité *(Probability)*
Probabilité conditionnelle *(Conditional probability)*
Test d'indépendance *(Test of independence)*
Test d'indépendance du chi-carré *(Chi-square test of independence)*

RÉFÉRENCE

Maistrov, L.E. (1974). *Probability Theory - A History Sketch.* Translated and edited by Samuel Kotz. Academic Press, New York.

INDÉPENDANCE, TEST D'

Test of independence

Voir **test d'indépendance**.

INDICATEUR

Indicator

Un indicateur est une statistique (**statistique officielle**) dont l'objectif est de renseigner sur l'état, le comportement et les changements survenus au cours du temps, d'un phénomène économique et social. Un indicateur doit informer sur les variations de grandeur comme sur les modifications de nature des phénomènes observés et doit servir à instruire les décideurs.

Dans de bonnes conditions, un indicateur est le reflet de la politique sociale ou économique mise en place et met sur la voie des modifications à y apporter. Il sert de signal d'alarme, attire l'attention et invite à l'action (pilotage). Plusieurs indicateurs sont généralement nécessaire pour donner un reflet global des phénomènes et des politiques que l'on souhaite évaluer. Un ensemble d'indicateurs renseignant sur le même sujet, est appelé un système d'indicateurs.

Cependant, toutes les statistiques ne sont pas des indicateurs, les indicateurs doivent avoir les qualités suivantes :

1. Être quantifiables, ce qui ne veut pas dire strictement mesurables.
2. Être unidimensionnels (éviter le chevauchement).
3. Couvrir l'ensemble des objectifs prioritaires.
4. Évaluer directement les performances d'une politique.
5. Mettre sur la voie d'améliorations.
6. Porter sur les aspects majeurs du système et modifiables par des changements politiques.
7. Être vérifiables.
8. Être susceptibles d'être mis en relation avec d'autres indicateurs.
9. Être aussi économiques que possible, la collecte doit en effet être faite en un temps, à un prix et avec un niveau d'expertise raisonnables.
10. Être fournis en temps utile.
11. Être aisément compréhensibles par un large public.
12. La validité et la fiabilité des indicateurs doivent être reconnues par tous.

DOMAINES ET LIMITATIONS

La construction des indicateurs

Les indicateurs comme toutes les mesures relatives aux phénomènes sociaux ne peuvent pas prétendre à l'exactitude et à l'exhaustivité. Concevoir des indicateurs fait intervenir des considérations philosophiques (jugements de valeurs) et des connaissances techniques et scientifiques, mais ils sont basés également sur le bons sens et reflètent souvent des préocupations politiques.

La collecte et le choix des indicateurs

Pour collecter de bons indicateurs, un modèle cohérent de fonctionnement du système économique, social ou éducatif, est nécessaire. Le modèle général adopté par l'OCDE est un modèle systémique : dans un environnement donné, on étudie les ressources, le processus et les effets. Mais le choix des indicateurs réunis dans un système d'indicateurs dépend également des possibilités techniques de la récolte de données. Cela explique la grande disparité des indicateurs. Le nombre d'indicateurs nécessaire au bon pilotage d'un système économique ou social est difficile à évaluer. Trop nombreux, les indicateurs perdent de l'importance et risquent de ne plus jouer leur rôle de clignotant. Mais ils doivent ensemble représenter le système que l'on veut évaluer. Un trop petit nombre risque de nuire à la validité et à la représentatitvité.

L'interprétation des indicateurs

est un aspect aussi important du système que leur choix, leur construction et leur collecte. En effet, étant donné la signification que les indicateurs peuvent prendre pour les responsables politiques, pour le financement et le succès de certains projets, toutes les précautions doivent être prises pour assurer l'intégrité, la qualité et la neutralité des analyses interprétatives des indicateurs.

La validation des indicateurs

Avant d'utiliser ces statistiques comme indicateurs de la performance, qualité ou santé d'un système économique, social ou éducatif, il serait important de les valider. En effet, si leur valeur a été calculée de manière scientifique, avec des méthodes et des instruments mathématiques, leur influence sur le système que l'on veut décrire et sur lequel on veut agir grâce à eux, n'a pas toujours été prouvé scientifiquement. De plus, lorsqu'on utilise un ensemble d'indicateurs, on est loin de savoir comment ces indicateurs combinent leurs effets pour produire le résultat observé. Les méthodes de validation des indicateurs restent encore à développer.

EXEMPLES

Les indicateurs économiques

Produit intérieur brut (PIB), revenu national, indice des prix à la consommation, indice des prix à la production, taux de chômage standardisé, emploi total, taux d'intéret à long et à court terme, exportations et importations des biens manufacturés, balance des paiements...

Les indicateurs de l'enseignement

Les dépenses pour l'éducation, les actifs dans le système d'enseignement, l'espérance de scolarisation, les disparités géographiques dans l'accès au niveau du baccalauréat, le niveau général des conscrits, le niveau de formation des jeunes sortant de formation initiale, l'effet de possession d'un diplôme sur les chances de trouver un emploi, la scolarisation des enfants de deux à cinq ans, la dépense d'éducation pour la formation continue...

Les indicateurs de la science et de la technologie

Les dépenses intérieures brutes de recherche et développement (DIRD), les DIRD par habitant, les DIRD en pourcentage du PIB, le personnel total de RD, le total des diplômés universitaires, le pourcentage des DIRD financé par les entreprises, le pourcentage des DIRD financé par l'État, le pourcentage des DIRD financé par l'étranger, les demandes nationales de brevets...

MOT CLÉ

Statistique officielle *(Official statistics)*

RÉFÉRENCES

L'OCDE et les indicateurs internationaux de l'enseignement (1992). Centre de la Recherche et l'innovation dans l'enseignement, OCDE, Paris.

Principaux indicateurs de la science et de la technologie (1998). OCDE, Paris.

INDICE

Index number

Dans sa définition la plus générale, un indice est une valeur représentant la variation relative d'une **variable** entre deux périodes (ou deux situations) déterminées.

Il faut distinguer les indices élémentaires des indices synthétiques :

- les indices élémentaires décrivent le changement relatif d'une seule variable ;
- les indices synthétiques ou indices composés permettent de décrire par un nombre unique la comparaison de l'ensemble des valeurs que prennent plusieurs variables dans une certaine situation par rapport à l'ensemble des valeurs des mêmes variables dans une situation de référence.

La situation de référence définit la base de l'indice. Nous pourrons dire par exemple que sur une base de 100 en 1980, un indice quelconque a une valeur de 120 en 1982.

HISTORIQUE

Il faut remonter au début du XVIII^{e} siècle pour trouver les premières études utilisant la notion d'indice.

C'est en effet en 1707 que l'évêque anglican Fleetwood entreprit d'étudier l'évolution des prix entre les années 1440 et 1700. Cette étude dont les résultats figurent dans son ouvrage *Chronicon Preciosum* (1745) avait été menée dans un but précis : lors de la fondation d'un collège dans les années 1440-1460, une clause essentielle avait été établie, qui stipulait que tout membre admis devait jurer de quitter ce collège si sa fortune dépassait 5£ par an. Fleetwood désirait savoir si, compte tenu de l'évolution des prix, une telle promesse pouvait encore être tenue trois siècles plus tard.

Il prit en considération quatre produits courants : le blé, la viande, la boisson et les vêtements et en étudia l'évolution. Au terme de ses travaux, il parvint à la conclusion que 5 livres (£) en 1440-1460 avaient la même valeur que 30 livres (£) en 1700.

Un peu plus tard, en 1738, Dutot, quant à lui, s'intéressa à la diminution de la valeur de la monnaie dans un ouvrage intitulé *Réflexions politiques sur les finances et le commerce*. Dans ce but, il étudia les revenus respectifs de deux rois, Louis XII et Louis XV. Chacun d'eux avait perçu :

Louis XII : 7 650 000£ en 1515 ;
Louis XV : 100 000 000£ en 1735.

Pour déterminer lequel des deux souverains avait disposé du revenu le plus élevé, il releva les différents prix, aux dates considérées, d'un certain nombre de biens : un poulet, un lapin, un pigeon, la valeur d'une journée de travail, etc.

La mesure de Dutot fut la suivante :

$$I = \frac{\sum P_1}{\sum P_0}$$

où $\sum P_1$ et $\sum P_0$ sont les sommes des prix relevés aux deux dates respectives. Il est à noter que cet indice n'est pas pondéré. Il aboutit à la conclusion que la richesse de Louis XII était inférieure à celle de Louis XV.

La diminution de la valeur de la monnaie fit également l'objet d'une étude de Gian Rinaldo Carli. Ce professeur d'astronomie de Padoue s'intéressa en 1764 à l'évolution des prix depuis la découverte de l'Amérique.

Il prit pour références trois produits : les grains, le vin et l'huile, dont il étudia les prix sur la période comprise entre 1500 et 1750.

La mesure de Carli s'écrit comme suit :

$$I = \frac{1}{n} \cdot \sum \frac{P_1}{P_0}.$$

L'indice obtenu est une **moyenne arithmétique** des prix relatifs pour chaque produit observé.

1798 fut l'année de la parution d'un ouvrage intitulé *An account of some endeavours to ascertain a standard of weight and measure*, signé de Sir George Shuckburgh Evelyn. Ce précurseur en matière d'indices y définit un indice de prix couvrant les années 1050 à 1800 en introduisant les notions d'année de base (1550) et de prix relatifs. Il incluait également le prix de certains services dans ses observations.

C'est toutefois en Joseph Lowe qu'il faut voir, selon M.G. Kendall (1977), le véritable père des indices. L'ouvrage qu'il publia en 1822 sous le titre *The present state of England*, porte sur de nombreux problèmes relatifs à la construction des indices. En outre, il introduisit la notion de pondération des prix aux différentes dates (P_1 et P_0) par les quantités observées à la première date (Q_0). La mesure de Lowe est la suivante :

$$I = \frac{\sum (P_1 \cdot Q_0)}{\sum (P_0 \cdot Q_0)}.$$

Plus tard cet indice sera connu sous le nom d'**indice de Laspeyres**.

Dans la seconde moitié du XIX^{e} siècle, les statisticiens poursuivirent leurs efforts dans ce domaine.

En 1863, avec *A serious fall in the value of gold ascertained and its social effects set forth*, Jevons étudia le type de **moyenne** à employer lors de la construction d'un indice. Il y recommandait l'utilisation de la **moyenne géométrique**.

Entre 1864 et 1880, trois scientifiques allemands, Laspeyres (1864, 1871), Drobisch (1871) et Paasche (1874) travaillèrent sur l'évolution des prix de biens matériels relevés à Hambourg, sans toutefois tenir compte des services. Tous trois rejetèrent les méthodes de calcul non pondérées.

L'**indice de Paasche** est déterminé par la formule suivante :

$$I = \frac{\sum (P_1 \cdot Q_1)}{\sum (P_0 \cdot Q_1)}.$$

Palgrave (1886) proposa de pondérer les prix relatifs par la **valeur** totale du bien concerné. Cette méthode conduit à calculer l'indice comme suit :

$$I = \frac{\sum \left(P_1 \cdot Q_1 \cdot \frac{P_1}{P_0}\right)}{\sum (P_1 \cdot Q_1)}.$$

Ce furent enfin les travaux de statisticiens tels que **Irving Fisher** qui, en 1922, définit un indice qui porte son nom et qui constitue la moyenne géométrique de l'indice de Laspeyres et de l'indice de Paasche :

$$I = \sqrt{\frac{\sum (P_1 \cdot Q_0) \sum (P_1 \cdot Q_1)}{\sum (P_0 \cdot Q_0) \sum (P_0 \cdot Q_1)}},$$

tandis qu'à la même époque, Marshall et **Francis Ysidro Edgeworth** proposèrent la formule suivante :

$$I = \frac{\sum (P_1 (Q_1 + Q_0))}{\sum (P_0 (Q_1 + Q_0))}.$$

Il est à noter que, par cette formule, on calcule la moyenne des poids.

Les études statistiques modernes font aujourd'hui encore très largement appel aux indices que nous venons de définir.

DOMAINES ET LIMITATIONS

Le calcul d'indices est une méthode usuelle de description des changements dans des **variables** économiques. Bien que les indices les plus courants concernent essentiellement la variable prix et la variable quantité, des indices peuvent aussi être calculés pour des valeurs, des qualités, etc.

Généralement associés au domaine des affaires et de l'économie, les indices sont aussi largement utilisés dans d'autres sphères d'activité.

MOTS CLÉS

Indice de Fisher *(Fisher index)*
Indice de Laspeyres *(Laspeyres index)*
Indice de Paasche *(Paasche index)*
Indice élémentaire *(Simple index number)*
Indice synthétique *(Composite index number)*

RÉFÉRENCES

Carli, G.R. (1764). Del valore, etc. *Opere scelte di Carli*, ed. Custodi. Vol I, pp. 299.

Drobisch, M.W. (1871). Über Mittelgrössen und die Anwendbarkeit derselben auf die Berechnung des Steigens und Sinkens des Geldwerts. *Berichte über die Verhandlungen der König. Sachs. Ges. Wiss.* Leipzig. Math-Phy. Klasse 23, pp. 25.

Dutot, C. (1738). *Réflexions politiques sur les finances, et le commerce.* La Haye.

Evelyn, Sir G.S. (1798). An account of some endeavours to ascertain a standard of weight and measure. *Phil. Trans.* 113.

Fleetwood, W. (1707). *Chronicon Preciosum, London.* 2nd ed., 1745.

Jevons, W.S. (1863). *A serious fall in the value of gold ascertained and its social effects set forth.* London.

Kendall, M.G. (1977). *The early history of index numbers. Studies in History of Statistics and Probability.* Vol. II. Edited by M. Kendall and R.L. Plackett. Charles Griffin & Company Ltd.

Laspeyres, E. (1864). Hamburger Warenpreise 1850-1863 und die kalifornisch-australischen Geldentdeckung seit 1848. *Jahrbücher für Nationalökonomie und Statistik*, 3, pp. 81-118, 209-236.

Laspeyres, E. (1871). Die Berechnung einer mittleren Waarenpreissteigerung. *Jahrbücher fur Nationalökonomie und Statistik*, 16, pp. 296-314.

Lowe, J. (1822). The present state of England in regard to Agriculture, Trade, and Finance. London.

Paasche, H. (1874). Über die Preisentwicklung der letzten Jahre nach den Hamburger Borsennotirungen. *Jahrbücher fur Nationalökonomie und Statistik*, 23, pp. 168-178.

Palgrave, R.H.I. (1886). *Currency and standard of value in England, France and India, etc.* Memorandum to the Royal Commission on Depression of Trade and Industry, 3rd Report, Appendix B.

INDICE DE FISHER

Fisher index

L'**indice** de Fisher est un **indice synthétique** permettant de mesurer les changements du coût de la vie. Il représente la **moyenne géométrique** de deux indices :

- l'**indice de Laspeyres**
- l'**indice de Paasche**.

Indice de Fisher =

$$\sqrt{(\text{Indice de Laspeyres})\cdot(\text{Indice de Paasche})}$$

HISTORIQUE

L'économiste et mathématicien américain Irving Fisher a établi en 1922 le modèle de l'indice portant son nom en vue de pallier certains inconvénients rencontrés par les indices de Laspeyres et de Paasche.

Comme l'**indice de Laspeyres** pondère les prix par les quantités de la période de base, il peut surestimer l'augmentation du coût de la vie (car les gens peuvent avoir réduit leur consommation des biens qui sont devenus proportionnellement plus chers que les autres).

À l'inverse, comme l'**indice de Paasche** utilise les quantités de la période courante, il peut sous-estimer l'augmentation du coût de la vie.

L'indice de Fisher, la **moyenne géométrique** des deux indices de Laspeyres et de Paasche, se veut donc un indice idéal représentant un compromis entre ces deux indices.

ASPECTS MATHÉMATIQUES

En reprenant la définition de l'**indice de Laspeyres** et de l'**indice de Paasche**, le calcul de l'indice de Fisher $F_{n/0}$ s'effectue de la manière suivante :

$$F_{n/0} = 100 \cdot \sqrt{\frac{\sum P_n Q_0 \cdot \sum P_n Q_n}{\sum P_0 Q_0 \cdot \sum P_0 Q_n}}$$

où P_n et Q_n représentent le prix et la quantité vendue d'un bien à la période courante n, P_0 et Q_0 le prix et la quantité vendue d'un bien à la période de référence 0. Les différentes sommes portent sur les biens considérés.

Remarquons que l'indice de Fisher est réversible, ce qui signifie que

$$F_{n/0} = \frac{1}{F_{0/n}}.$$

DOMAINES ET LIMITATIONS

Bien que l'**indice** de Fisher soit en quelque sorte un indice idéal, il constitue un compromis entre les indices de Paasche et de Laspeyres.

EXEMPLE

Considérons le tableau fictif suivant indiquant les prix et les quantités respectives de trois biens de consommation à l'année de référence 0 et à l'année courante n.

Biens	Quantités (milliers)		Prix (euros)	
	1970 (Q_0)	1988 (Q_n)	1970 (P_0)	1988 (P_n)
Lait	50,5	85,5	0,20	1,20
Pain	42,8	50,5	0,15	1,10
Beurre	15,5	40,5	0,50	2,00

D'après le tableau suivant, nous avons

$$\begin{aligned} \sum P_n Q_n &= 239,15, \\ \sum P_0 Q_n &= 44,925, \\ \sum P_n Q_0 &= 138,68 \quad \text{et} \\ \sum P_0 Q_0 &= 24,27. \end{aligned}$$

Biens	$\sum P_n Q_n$	$\sum P_0 Q_n$	$\sum P_n Q_0$	$\sum P_0 Q_0$
Lait	102,60	17,100	60,60	10,10
Pain	55,55	7,575	47,08	6,42
Beurre	81,00	20,250	31,00	7,75
Total	239,15	44,925	138,68	24,27

Nous pouvons alors trouver l'**indice de Paasche**

$$\begin{aligned} I_{n/0} &= \frac{\sum P_n \cdot Q_n}{\sum P_0 \cdot Q_n} \cdot 100 \\ &= \frac{239,15}{44,925} \cdot 100 = 532,3 \end{aligned}$$

et l'**indice de Laspeyres**

$$\begin{aligned} I_{n/0} &= \frac{\sum P_n \cdot Q_0}{\sum P_0 \cdot Q_0} \cdot 100 \\ &= \frac{138,68}{24,27} \cdot 100 = 571,4. \end{aligned}$$

L'indice de Fisher est alors :

$$\text{indice de Fisher} = \sqrt{532,3 \cdot 571,4} = 551,5.$$

Selon l'indice de Ficher, le prix des biens considérés a donc augmenté de 451,5 % (551,5−100) entre l'année de référence et l'année courante.

MOTS CLÉS

Indice *(Index number)*
Indice de Laspeyres *(Laspeyres index)*
Indice de Paasche *(Paasche index)*
Indice élémentaire *(Simple index number)*
Indice synthétique *(Composite index number)*

RÉFÉRENCE

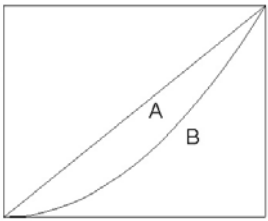
Fisher, I. (1922). *The Making of Index Numbers*, Houghton Mifflin, Boston.

INDICE DE GINI

Gini index

L'indice de Gini est l'indice d'inégalité le plus fréquemment utilisé ; il permet de mesurer le degré d'inégalité dans la répartition des revenus d'une population donnée. Graphiquement, l'indice de Gini est représenté par la surface entre la droite à 45° et la courbe de Lorenz (une représentation graphique des pourcentages cumulés du revenu total revenant aux différents pourcentages cumulés du nombre des bénéficiaires, les individus ou ménages les plus pauvres étant pris en premier). Cette surface s'exprime comme une fraction de l'aire totale sous la diagonale :

Ainsi, on a : $I_G = \dfrac{A}{A+B}$



Le coefficient de Gini prend la valeur 0 si la distribution des revenus est parfaitement égale et la valeur 1 si la distribution des revenus est totalement inégale ; ainsi $0 \leq I_G \leq 1$. En d'autres termes, on peut dire que plus la distribution des revenus est inégale, plus cet indice prend des valeurs proches de 1. Dans les pays développés, la valeur de l'indice de Gini se situe autour de $0,4$ (voir l'exemple).

HISTORIQUE

Le coefficient de Gini a été inventé en 1910 par le statisticien et démographe italien **Corrado Gini**.

ASPECTS MATHÉMATIQUES

L'aire sous la droite à 45° est égale à $\frac{1}{2}$, on peut donc écrire :

$$I_G = \frac{A}{A+B} = \frac{A}{\frac{1}{2}} = 2A = 2\left(\frac{1}{2} - B\right) \Rightarrow I_G = 1-2B \tag{1}$$

Pour calculer la valeur du coefficient de Gini, il faut donc connaître B (aire sous la courbe de Lorenz).

a) *Indice de Gini : cas discret*

Courbe de Lorenz discrète : pour une distribution de revenus rangés par ordre croissant $(x_1 \leq ... \leq x_n)$.

Pour calculer B, il faut diviser la surface sous la courbe de Lorenz en triangles et en trapèzes (la première surface est un triangle et l'autre est un trapèze).

La surface du triangle est définie ainsi :

$$\frac{1}{2n}\frac{x_1}{X}.$$

La surface du j-ème trapèze est égale à :

$$\frac{1}{2n}\left[\sum_{i=1}^{j-1}\frac{x_i}{X} + \sum_{i=1}^{j}\frac{x_i}{X}\right] = \frac{1}{2nX}\left[2\cdot\sum_{i=1}^{j-1}x_i + x_j\right],$$

avec $j = 2, \ldots, n$. On a donc

$$B = \frac{1}{2n}\cdot\frac{x_1}{X} + \sum_{j=2}^{n}\frac{1}{2nX}\left[2\cdot\sum_{i=1}^{j-1}x_i + x_j\right],$$

où

$x_1, \ldots, x_n$	revenus
X	revenu total
$\dfrac{x_1 + ... + x_n}{X}$	revenu total cumulé en %.

Après développement, on obtient :

$$B = \frac{1}{2nX}\cdot\left[-X + 2\cdot\sum_{i=1}^{n}(n-i+1)x_i\right] \tag{2}$$

D'après l'équation (1), on obtient l'indice de Gini :

$$I_G \;=\; 1 - 2B$$

$$= 1 - \frac{2}{2nX}\left[-X + 2 \cdot \sum_{i=1}^{n}(n-i+1)x_i\right]$$

$$= 1 + \frac{1}{n} - 2 \cdot \frac{\sum_{i=1}^{n}(n-i+1)\,x_i}{nX}$$

$$= 1 + \frac{1}{n} - \frac{2}{n^2\lambda(x)}\sum_{i=1}^{n} i x_{n-i+1},$$

où $\lambda(x) = \frac{1}{n}\sum_{i=1}^{n} x_i$ = revenu moyen de la distribution.

N.B. Cette formule est illustrée dans l'exemple. Le coefficient de Gini peut être reformulé de plusieurs manières. Soit, par exemple :

$$I_G = \sum_{i=1}^{n}\sum_{j=1}^{n}\frac{|x_i - x_j|}{2n^2\lambda(x)}.$$

b) *Indice de Gini : cas continu*

D'après l'équation (1), il suffit de calculer B. Dans ce cas, B correspond à l'intégrale de 0 à 1 de la fonction de Lorenz. On a :

$$I_G = 1 - 2B = 1 - 2\int_0^1 L(P)\,dP$$

où $P = F(y) = \int_0^y f(x)\,dx$ = pourcentage des individus qui ont un revenu inférieur à y.

$$I_G = \int_0^1 \left(1 - (1-P)^{\frac{\alpha-1}{\alpha}}\right) dP = 1 - \frac{\alpha}{2\alpha - 1}$$

$$= 1 - 2\left(1 - \frac{\alpha}{2\alpha-1}\right)$$

$$= \frac{1}{2\alpha - 1} \quad : \text{ loi de Pareto}$$

$$= \frac{1}{3}\frac{(x_{max} - x_0)}{(x_{max} + x_0)} : \text{ loi uniforme.}$$

DOMAINES ET LIMITATIONS

Le coefficient de Gini mesure l'ampleur de l'inégalité dans une répartition. Il est le plus souvent utilisé pour calculer les disparités en termes de revenu. Il peut aussi être utilisé pour mesurer les disparités cantonales du chômage.

Cet indice s'est heurté à un grand nombre de critiques. On dit qu'il n'est pas très sensible, qu'il est difficilement calculable avec précision et qu'il n'a aucun contenu normatif.

EXEMPLE

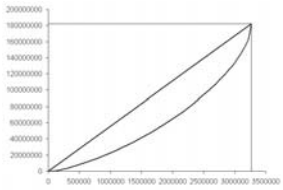
Pour calculer le coefficient de Gini, prenons le cas discret.

On a : $I_G = 1 - 2B$

La surface B est égale à la somme des trapèzes et du premier triangle découpés sous la courbe de Lorenz.

Les données pour 1993-1994 sont les suivantes :

Population	Revenu	Surface
0	0	
94 569	173 100, 4	0, 000013733
37 287	527 244, 8	0, 000027323
186 540	3 242 874, 9	0, 000726701
203 530	4 603 346, 0	0, 00213263
266 762	7 329 989, 6	0, 005465843
290 524	9 428 392, 6	0, 010037284
297 924	11 153 963, 1	0, 015437321
284 686	12 067 808, 2	0, 02029755
251 155	11 898 201, 1	0, 022956602
216 164	11 317 647, 7	0, 023968454
182 268	10 455 257, 4	0, 023539381
153 903	9 596 182, 1	0, 022465371
127 101	8 560 949, 7	0, 020488918
104 788	7 581 237, 4	0, 018311098
85 609	6 621 873, 3	0, 01597976
70 108	5 773 641, 3	0, 013815408
57 483	5 021 336, 9	0, 011848125
47 385	4 376 126, 7	0, 010140353
39 392	3 835 258, 6	0, 008701224
103 116	11 213 361, 3	0, 024078927
71 262	9 463 751, 6	0, 017876779
46 387	7 917 846, 5	0, 012313061
50 603	20 181 113, 8	0, 0114625049



Ainsi, la surface B est égale à :

$$\begin{aligned} B &= \frac{1}{2nX} \cdot \left[-X + 2 \cdot \sum_{i=1}^{n} (n - i + 1) \cdot x_i \right] \\ &= 0,315246894. \end{aligned}$$

On a donc :

$$I_G = 1 - 2B = 1 - 2 \cdot 0,315246894 = 0,369506212.$$

MOTS CLÉS

Indice élémentaire *(Simple index number)*
Loi uniforme *(Uniform distribution)*

RÉFÉRENCES

Douglas Greenwald ed. (1984). *Encyclopédie économique*, Economica, Paris, pp.164-165.

Gini, C. (1910). Indici di concentratione e di dependenza. *Atti della III Riunione della Societa Italiana per il Progresso delle Scienze.* In ref. 12, pp. 3-210.

INDICE DE LASPEYRES

Laspeyres index

L'**indice** de Laspeyres est un **indice synthétique** de prix construit par la méthode de la somme pondérée.

Cet indice représente le ratio de la somme des prix de la période actuelle n par la somme des prix de la période de référence 0, ces sommes étant pondérées par les quantités respectives de la période de référence.

L'indice mesure donc le changement relatif de prix des biens, les quantités respectives étant considérées comme inchangées.

L'indice de Laspeyres se différencie de l'**indice de Paasche** par le choix de la pondération : dans l'indice de Paasche, la pondération se fait par les quantités vendues de la période courante Q_n, alors que dans l'indice de Laspeyres, elle se fait par les quantités vendues de la période de référence Q_0.

HISTORIQUE

Étienne Laspeyres, économiste et statisticien allemand d'origine française, a repris l'indice développé par Lowe qui deviendra connu sous le nom d'indice de Laspeyres au milieu du XIX^{e} siècle. Durant les années 1864-1871, E. Laspeyres travailla sur des biens dont les prix étaient observés à Hambourg (il avait exclu les services).

ASPECTS MATHÉMATIQUES

L'**indice** de Laspeyres se calcule de la façon suivante :

$$I_{n/0} = \frac{\sum P_n \cdot Q_0}{\sum P_0 \cdot Q_0},$$

où P_n et Q_n sont les prix et les quantités vendues à la période courante, P_0 et Q_0 les prix et les quantités vendues à la période de référence 0 et exprimées en base 100 :

$$I_{n/0} = \frac{\sum P_n \cdot Q_0}{\sum P_0 \cdot Q_0} \cdot 100.$$

Les sommes portent sur les biens considérés.

Le modèle de Laspeyres peut aussi être appliqué pour calculer un indice de quantité (appelé aussi indice de volume). Dans ce cas, ce sont les prix qui sont constants et les quantités variables :

$$I_{n/0} = \frac{\sum Q_n \cdot P_0}{\sum Q_0 \cdot P_0} \cdot 100.$$

EXEMPLE

Considérons le tableau fictif suivant indiquant les prix respectifs de trois biens de consommation à l'année de référence 0 et à l'année courante n, de même que les quantités vendues à l'année de référence :

Biens	Quantités vendues en 1970 (Q_0) (milliers)	Prix (euros) 1970 (P_0)	1988 (P_n)
Lait	50, 5	0, 20	1, 20
Pain	42, 8	0, 15	1, 10
Beurre	15, 5	0, 50	2, 00

D'après le tableau suivant, nous avons :

$$\sum P_n Q_0 = 138,68 \text{ et } \sum P_0 Q_0 = 24,27.$$

Biens	$\sum P_n Q_0$	$\sum P_0 Q_0$
Lait	60, 60	10, 10
Pain	47, 08	6, 42
Beurre	31, 00	7, 75
Total	138, 68	24, 27

Nous pouvons alors trouver l'indice de Laspeyres :

$$\begin{aligned} I_{n/0} &= \frac{\sum P_n \cdot Q_0}{\sum P_0 \cdot Q_0} \cdot 100 \\ &= \frac{138,68}{24,27} \cdot 100 = 571,4. \end{aligned}$$

En d'autres termes, selon l'indice de Laspeyres, l'indice des prix de ces biens a augmenté de 471, 4 % (571, 4 − 100) pendant la période considérée.

MOTS CLÉS

Indice *(Index number)*
Indice de Fisher *(Fisher index)*
Indice de Paasche *(Paasche index)*
Indice élémentaire *(Simple index number)*
Indice synthétique *(Composite index number)*

RÉFÉRENCES

Laspeyres, E. (1864). Hamburger Warenpreise 1850-1863 und die kalifornisch-australischen Geldentdeckung seit 1848. *Jahrbücher für Nationalökonomie und Statistik*, 3, pp. 81-118, 209-236.

Laspeyres, E. (1871). Die Berechnung einer mittleren Waarenpreissteigerung. *Jahrbücher fur Nationalökonomie und Statistik*, 16, pp. 296-314.

INDICE DE PAASCHE

Paasche index

L'**indice** de Paasche est un **indice synthétique** de prix construit par la méthode de la somme pondérée.

Cet indice représente le ratio de la somme des prix de la période actuelle n par la somme des prix de la période de référence 0, ces sommes étant pondérées par les quantités respectives de la période actuelle.

L'indice de Paasche ne se différencie de l'**indice de Laspeyres** que par le choix de la pondération. En effet, dans l'indice de Laspeyres, la pondération se fait par les quantités de la période de référence Q_0 plutôt que par les quantités de la période courante Q_n.

HISTORIQUE

Statisticien allemand, Hermann Paasche a développé au milieu du XIX^{e} siècle la formule d'**indice** qui porte son nom. H. Paasche (1874) a travaillé sur des prix relevés à Hambourg. Il était opposé aux méthodes non pondérées.

ASPECTS MATHÉMATIQUES

L'**indice** de Paasche se calcule de la façon suivante :

$$I_{n/0} = \frac{\sum P_n \cdot Q_n}{\sum P_0 \cdot Q_n}$$

où P_n et Q_n sont les prix et les quantités vendues à la période courante, P_0 et Q_0 les prix et les quantités vendues à la période de référence. Les sommes portent sur les biens considérés et exprimés dans la base 100 :

$$I_{n/0} = \frac{\sum P_n \cdot Q_n}{\sum P_0 \cdot Q_n} \cdot 100.$$

Le modèle de Paasche peut aussi être appliqué pour calculer un indice de quantité (appelé aussi indice de volume). Dans ce cas, ce sont les prix qui sont constants et les quantités variables :

$$I_{n/0} = \frac{\sum Q_n \cdot P_n}{\sum Q_0 \cdot P_n} \cdot 100.$$

EXEMPLE

Considérons le tableau fictif suivant indiquant les prix respectifs de trois biens de consommation à l'année de référence 0 et à l'année courante n, de même que les quantités vendues à l'année courante :

	Quantités vendues en 1988 (Q_n)	Prix (euros)	
		1970	1988
Biens	(milliers)	(P_0)	(P_n)
Lait	85,5	0,20	1,20
Pain	50,5	0,15	1,10
Beurre	40,5	0,50	2,00

D'après le tableau suivant, nous avons :

$$\sum P_n Q_n = 239,15 \text{ et } \sum P_0 Q_n = 44,925.$$

Biens	$\sum P_n Q_n$	$\sum P_0 Q_n$
Lait	102,60	17,100
Pain	55,55	7,575
Beurre	81,00	20,250
Total	239,15	44,925

Nous pouvons alors trouver l'indice de Paasche :

$$\begin{aligned} I_{n/0} &= \frac{\sum P_n \cdot Q_n}{\sum P_0 \cdot Q_n} \cdot 100 \\ &= \frac{239,15}{44,925} \cdot 100 = 532,3. \end{aligned}$$

En d'autres termes, selon l'indice de Paasche, l'indice des prix de ces biens a augmenté de 432,3 % (532,3 − 100) pendant la période considérée.

MOTS CLÉS

Indice *(Index number)*
Indice de Fisher *(Fisher index)*
Indice de Laspeyres *(Laspeyres index)*
Indice élémentaire *(Simple index number)*
Indice synthétique *(Composite index number)*

RÉFÉRENCE

Paasche, H. (1874). Über die Preisentwicklung der letzten Jahre nach den Hamburger Borsennotirungen. *Jahrbücher fur Nationalökonomie und Statistik*, 23, pp. 168-178.

INDICE ÉLÉMENTAIRE

Simple index number

On appelle **indice** élémentaire le rapport de deux **valeurs** représentant la même **variable**, mesurées dans deux situations différentes, ou à deux périodes différentes.

Par exemple, un indice élémentaire de prix nous donnera la variation relative du prix entre la période courante et une période de référence, ou période de base.

Les indices élémentaires les plus couramment rencontrés sont les indices de prix, de quantité et de valeur.

HISTORIQUE

Voir **indice**.

ASPECTS MATHÉMATIQUES

L'**indice** $I_{n/0}$ représentatif d'une **variable** G dans la situation n par rapport à la même variable dans la situation 0 (situation de base) est défini par :

$$I_{n/0} = \frac{G_n}{G_0},$$

où G_n est la **valeur** de la variable G dans la situation n et G_0 est la valeur de la variable G dans la situation 0.

Généralement, un indice élémentaire est exprimé en base 100 dans la situation de référence 0 :

$$I_{n/0} = \frac{G_n}{G_0} \cdot 100.$$

Propriétés des indices élémentaires

- *Identité* : si les deux situations (ou les deux périodes) comparées sont identiques, la valeur de l'indice est égale à 1 (ou 100) :

$$I_{n/n} = I_{0/0} = 1$$

- Un indice élémentaire est *réversible* :

$$I_{n/0} = \frac{1}{I_{0/n}}$$

- Un indice élémentaire est *transférable* : considérons deux indices $I_{1/0}$ et $I_{2/0}$ représentant deux situations 1 et 2 par rapport à la même situation de base 0. Si nous considérons l'indice $I_{2/1}$, nous pouvons dire que :

$$I_{2/0} = I_{2/1} \cdot I_{1/0}.$$

Remarquons que la transférabilité (appelée aussi circularité) implique la réversibilité :
si $I_{0/1} \cdot I_{1/0} = I_{0/0} = 1$, alors $I_{0/1} = \dfrac{1}{I_{1/0}}$.

DOMAINES ET LIMITATIONS

Dans la vie économique, il est très rare que des **indices** élémentaires soient exploitables pour eux-mêmes, l'information qu'ils contiennent étant relativement limitée.

Toutefois, ils sont d'une importance fondamentale pour la construction des indices synthétiques.

EXEMPLE

Considérons le tableau suivant représentant les prix d'un bien quelconque et les quantités vendues à deux périodes différentes :

Prix		Quantités	
Période 1	Période 2	Période 1	Période 2
$P_0 = 50$	$P_n = 70$	$Q_0 = 25$	$Q_n = 20$

Si P_n est le prix de la période qui nous intéresse et P_0 est le prix de la période de base, nous avons :

$$I_{n/0} = \frac{P_n}{P_0} \cdot 100 = \frac{70}{50} \cdot 100 = 140$$

ce qui signifie que sur une base de 100 en période 1, l'indice du prix du bien est de 140 à la période 2. Le prix du bien a donc augmenté de 40 % (140 − 100) entre la période de référence et la période actuelle.

Nous pouvons de même calculer un indice élémentaire de quantité :

$$I_{n/0} = \frac{Q_n}{Q_0} \cdot 100 = \frac{20}{25} \cdot 100 = 80.$$

La quantité vendue a donc diminué de 20 % (100 − 80) entre la période de référence et la période actuelle.

Sur la base de ces chiffres, il nous est aussi possible de calculer un indice de valeur, la valeur étant définie comme le prix multiplié par la quantité :

$$I_{n/0} = \frac{P_n \cdot Q_n}{P_0 \cdot Q_0} = \frac{70 \cdot 20}{50 \cdot 25} \cdot 100 = 112.$$

La valeur du bien considéré a donc augmenté de 12 % (112 − 100) entre les deux périodes considérées.

MOTS CLÉS

Indice *(Index number)*
Indice de Fisher *(Fisher index)*
Indice de Laspeyres *(Laspeyres index)*
Indice de Paasche *(Paasche index)*
Indice synthétique *(Composite index number)*

INDICE SAISONNIER

Seasonal index

Considérons une **série chronologique** Y_t, dont les composantes sont :

- la **tendance séculaire** T_t ;
- les **fluctuations cycliques** C_t ;
- les **variations saisonnières** S_t ;
- les **variations irrégulières** I_t.

On sait que l'influence des variations saisonnières doit être neutre sur l'année et que les variations saisonnières S_t se répètent théoriquement de manière identique de période en période (ce qui n'est pas vérifié sur un cas réel).

Pour satisfaire aux exigences du **modèle** théorique et pour pouvoir étudier la série chronologique

réelle, il faut estimer, au lieu des S_t observées, des variations périodiques identiques chaque année (mois par mois ou trimestre par trimestre) qu'on appelle indices saisonniers.

On les note S_j, où j varie comme suit :

$j = 1$ à 12 pour les mois (sur n années), ou
$j = 1$ à 4 pour les trimestres (sur n années).

On remarque que sur n années, il n'existe que 12 indices saisonniers en ce qui concerne les mois et 4 en ce qui concerne les trimestres.

ASPECTS MATHÉMATIQUES

Il existe plusieurs méthodes de calcul des indices saisonniers :

1. *Méthode du pourcentage à la moyenne*

   Les données correspondant à chaque mois sont exprimées en **pourcentage** de la **moyenne**. On calcule alors avec la moyenne (sans tenir compte des valeurs extrêmes) ou la **médiane** les pourcentages de chaque mois pour des années différentes. En prenant la moyenne pour chaque mois, on obtient alors 12 pourcentages qui donnent des indices. Si leur moyenne n'est pas égale à 100 %, il faut faire un ajustement en multipliant les indices par un facteur convenable.

   Soit une **série chronologique** mensuelle sur n années, dénotée par $Y_{i,j}$, où $j = 1, \ldots, 2$ pour les différents mois et $i = 1, \ldots, n$ indiquant l'année.

   On regroupe d'abord les données dans un tableau du type :

	Mois		
Année	janvier	...	décembre
	1	...	12
1	$Y_{1,1}$	...	$Y_{1,12}$
...	...	...	...
n	$Y_{n,1}$	...	$Y_{n,12}$

   Puis on calcule les moyennes annuelles pour les n années :

$$M_i = \frac{\bar{Y}_{i,.}}{12}$$

   pour $i = 1, \ldots, n$.

   On calcule alors pour chaque donnée :

$$Q_{i,j} = \frac{Y_{i,j}}{M_i}$$

   pour $i = 1, \ldots, n$ et $j = 1, \ldots, 12$ que l'on reporte dans un tableau :

	Mois		
Année	janvier	...	décembre
	1	...	12
1	$Q_{1,1}$	...	$Q_{1,12}$
...	...	...	...
n	$Q_{n,1}$	...	$Q_{n,12}$

   Dans chaque colonne on effectue la **moyenne arithmétique**, notée $\bar{Q}_{i,.}$ et on calcule la somme de ces douze moyennes ; si cette dernière n'est pas égale à 1 200, on introduit un facteur de correction :

$$k = \frac{1\,200}{\sum_{i=1}^{12} \bar{Q}_{i,.}}.$$

   Finalement, les indices saisonniers sont donnés pour chaque mois par la moyenne du mois multipliée par k.

2. *Méthode du rapport à la tendance*

   Les données de chaque mois sont exprimées en pourcentage des valeurs mensuelles de la **tendance séculaire**. L'indice s'obtient à partir d'une **moyenne** appropriée des pourcentages des mois correspondants. Il faut ajuster les résultats si la moyenne n'est pas égale à 100 % pour obtenir les indices saisonniers cherchés.

3. *Méthode du rapport à la moyenne mobile*

   On calcule une **moyenne mobile** sur 12 mois. Comme les résultats correspondants s'étendent sur plusieurs mois, au lieu d'être au milieu du mois comme les données d'origine, on calcule une moyenne mobile sur deux mois de la moyenne mobile sur 12 mois. C'est ce qu'on appelle une moyenne mobile centrée sur 12 mois.

   Ensuite on exprime les données d'origine de chaque mois en pourcentage de la moyenne

mobile centrée sur 12 mois correspondante. On prend la moyenne des pourcentages correspondant à chaque mois, ce qui donne des indices. Il faut ensuite ajuster à 100 %, pour obtenir les indices saisonniers souhaités.

4. *Méthode des chaînes relatives*

   Les données de chaque mois sont exprimées en pourcentage des données du mois précédent (d'où l'expression chaîne relative). On calcule alors une moyenne appropriée des chaînes pour les mois correspondants. On peut obtenir les pourcentages relatifs de chaque mois par rapport à ceux du mois de janvier (= 100 %). On trouve ensuite que le mois de janvier suivant aura ordinairement un pourcentage associé soit inférieur soit supérieur à 100 %. Il y a donc croissance ou décroissance de la tendance. Les indices finalement obtenus, ajustés pour que leur moyenne soit égale à 100 %, donnent les indices saisonniers désirés.

EXEMPLE

On considère la **série chronologique** d'un phénomène quelconque ; il s'agit de données trimestrielles prises sur une durée de six ans résumées dans le tableau suivant :

Année	1^{er} trim.	2^{e} trim.	3^{e} trim.	4^{e} trim.
1	19, 65	16, 35	21, 30	14, 90
2	28, 15	25, 00	29, 85	23, 40
3	36, 75	33, 60	38, 55	32, 10
4	45, 30	42, 25	47, 00	40, 65
5	54, 15	51, 00	55, 75	49, 50
6	62, 80	59, 55	64, 40	58, 05

On détermine les indices saisonniers à l'aide de la méthode du **pourcentage** à la **moyenne** :

On calcule d'abord les moyennes arithmétiques trimestrielles pour les six années :

Année	Somme	Moyenne
1	72, 2	18, 05
2	106, 4	26, 6
3	141, 0	35, 25
4	175, 2	43, 8
5	210, 4	52, 6
6	244, 8	61, 2

On divise alors les lignes du tableau initial par la moyenne trimestrielle de l'année correspondante et on multiplie par 100.

Ensuite, puisque les résultats montrent d'assez grands écarts pour les valeurs obtenues pour un même trimestre, il est plus judicieux de prendre la **médiane** au lieu de la moyenne arithmétique pour obtenir des chiffres plus significatifs.

Année	1^{er} trim.	2^{e} trim.	3^{e} trim.	4^{e} trim.
1	108, 86	90, 58	118, 01	82, 55
2	105, 83	93, 98	112, 22	87, 97
3	104, 26	95, 32	109, 36	91, 06
4	103, 42	96, 46	107, 31	92, 81
5	102, 95	96, 96	105, 99	94, 11
6	102, 61	97, 30	105, 23	94, 85
Médiane	103, 84	95, 89	108, 33	91, 94

Le pourcentage médian pour chaque trimestre est donné dans la dernière ligne du tableau ci-dessus. La somme de ces pourcentages étant exactement de 400 %, un ajustement n'est pas nécessaire, et les chiffres figurant sur la dernière ligne représentent les indices saisonniers demandés.

On a vu qu'il existe trois autres méthodes pour déterminer les indices saisonniers, ces dernières fournissant les valeurs suivantes pour ce même problème :

	Trimestre			
Méthode	1	2	3	4
Pourcentage à la moyenne	103, 84	95, 89	108, 33	91, 94
Rapport à la tendance	112, 62	98, 08	104, 85	84, 45
Rapport à la moy. mobile	111, 56	98, 87	106, 14	83, 44
Chaînes	112, 49	98, 24	104, 99	84, 28

On constate que les résultats sont en accord les uns avec les autres, malgré la diversité des méthodes utilisées pour les déterminer.

MOTS CLÉS

Fluctuation cyclique *(Cyclical fluctuation)*
Série chronologique *(Time series)*

Tendance séculaire *(Secular trend)*
Variation irrégulière *(Irregular variation)*
Variation saisonnière *(Seasonal variation)*

INDICE SYNTHÉTIQUE

Composite index number

Les indices synthétiques permettent de mesurer par un nombre unique les variations relatives d'un groupe de variables dans deux situations différentes.

L'indice des prix à la consommation, l'indice des prix de gros, l'indice de l'emploi, l'indice boursier Dow-Jones, etc., sont des indices synthétiques.

Le but des indices synthétiques est de résumer dans un seul indice l'ensemble des indices élémentaires que nous pouvons calculer à partir d'une grandeur complexe (grandeur formée par un ensemble de grandeurs simples).

Les indices synthétiques les plus courants sont :

- l'**indice de Laspeyres** ;
- l'**indice de Paasche** ;
- l'**indice de Fisher**.

HISTORIQUE

Voir **indice**.

ASPECTS MATHÉMATIQUES

Il existe plusieurs méthodes de construction d'indices synthétiques.

Pour illustrer ces méthodes, prenons le cas de la détermination d'un **indice** de prix à la période courante n par rapport à une période de référence 0.

1. *Indice des moyennes arithmétiques* (méthode de la somme)

$$I_{n/0} = \frac{\sum P_n}{\sum P_0} \cdot 100,$$

où $\sum P_n$ est la somme des prix des biens à la période courante, et $\sum P_0$ est la somme des prix des mêmes biens à la période de référence.

2. *Moyenne arithmétique des indices élémentaires*

$$I_{n/0} = \frac{1}{N} \cdot \sum \left(\frac{P_n}{P_0} \right) \cdot 100,$$

où N est le nombre de biens considérés, et $\frac{P_n}{P_0}$ l'**indice élémentaire** de chaque bien.

Par ces deux méthodes, chaque bien a la même importance, ce qui ne correspond pas forcément à la réalité.

3. *Indice des moyennes arithmétiques pondérées* (méthode de la somme pondérée)

La formule générale d'un indice calculé par la méthode de la somme pondérée est la suivante :

$$I_{n/0} = \frac{\sum P_n \cdot Q}{\sum P_0 \cdot Q} \cdot 100.$$

Le problème réside dans le choix de la quantité Q pour chaque bien considéré. Celle-ci doit être la même au numérateur et au dénominateur (dans le cas du calcul d'un indice de prix).

Dans l'**indice de Laspeyres**, c'est la quantité de l'année de référence qui est utilisée. En revanche, dans l'**indice de Paasche**, c'est la quantité de l'année courante. D'autres statisticiens proposent d'utiliser la quantité à une année donnée.

EXEMPLE

Considérons le tableau fictif suivant indiquant les prix de trois biens de consommation à l'année de référence 0 et à l'année courante n.

	Prix (euros)	
Biens	1970 (P_0)	1988 (P_n)
Lait	0,20	1,20
Pain	0,15	1,10
Beurre	0,50	2,00

Sur la base des chiffres ci-dessus, nous allons examiner les trois principales méthodes de construction des indices synthétiques.

1. *Indice des moyennes arithmétiques* (méthode de la somme)

$$\begin{aligned} I_{n/0} &= \frac{\sum P_n}{\sum P_0} \cdot 100 \\ &= \frac{4,30}{0,85} \cdot 100 = 505,9. \end{aligned}$$

   Selon ce mode de calcul, l'**indice** des prix a augmenté de $405,9$ % $(505,9-100)$ entre l'année de référence et l'année courante.

2. *Moyenne arithmétique des indices élémentaires*

$$\begin{aligned} I_{n/0} &= \frac{1}{N} \cdot \sum \left(\frac{P_n}{P_0} \right) \cdot 100 \\ &= \frac{1}{3} \cdot \left(\frac{1,20}{0,20} + \frac{1,10}{0,15} + \frac{2,00}{0,50} \right) \cdot 100 \\ &= 577,8. \end{aligned}$$

   Ce mode de calcul nous donne un résultat quelque peu différent du premier puisque nous obtenons une augmentation de $477,8$ % $(577,8-100)$ de l'indice des prix entre l'année de référence et l'année courante.

3. *Indice des moyennes arithmétiques pondérées* (méthode de la somme pondérée)

$$I_{n/0} = \frac{\sum P_n \cdot Q}{\sum P_0 \cdot Q} \cdot 100.$$

   Des applications de cette méthode sont données par l'**indice de Laspeyres** et l'**indice de Paasche**.

MOTS CLÉS

Indice *(Index number)*
Indice de Fisher *(Fisher index)*
Indice de Laspeyres *(Laspeyres index)*
Indice de Paasche *(Paasche index)*
Indice élémentaire *(Simple index number)*

INÉGALITÉ DE TCHEBYCHEV

Chebyshev's inequality

Voir **loi des grands nombres**.

INERTIE

Inertia

L'inertie est une notion très utilisée en mécanique. On dira qu'un corps a d'autant plus d'inertie autour d'un axe qu'il faut lui appliquer d'énergie pour le mettre en rotation autour de cet axe. Dans l'inertie intervient la masse du corps et la **distance** (au carré) de ce corps à l'axe. On reconnaît la définition même de la **variance**, si l'on remplace « masse » par « **fréquence** » et « distance » par « écart ».

ASPECTS MATHÉMATIQUES

En supposant que l'on ait un nuage de n points centré dans un espace de dimension p, on peut construire la **matrice d'inertie** V à l'aide des coordonnées et des fréquences des points.

Les valeurs propres $k_1, \ldots, k_p$ (écrites dans un ordre décroissant) de la **matrice** V sont les inerties expliquées par les p axes factoriels du nuage de points. Leur somme, égale à la **trace** de V, est égale à l'inertie totale du nuage.

On dit que le premier axe factoriel (ou axe d'inertie) explique :

$$100 \cdot \frac{k_1}{\sum_{i=1}^{p} k_i} (\text{ en \%}) \text{ d'inertie.}$$

De même, les deux premiers axes d'inertie expliquent :

$$100 \cdot \frac{k_1 + k_2}{\sum_{i=1}^{p} k_i} (\text{ en \%}) \text{ d'inertie.}$$

DOMAINES ET LIMITATIONS

La notion d'inertie ou plus spécifiquement la **matrice d'inertie** d'un nuage de points est utilisée entre autres pour effectuer une **analyse factorielle des correspondances** entre les points considérés.

MOTS CLÉS

Analyse factorielle des correspondances *(Correspondence analysis)*
Axe factoriel *(Factorial axis)*
Distance *(Distance)*
Fréquence *(Frequency)*
Matrice d'inertie *(Inertia matrix)*
Valeur propre *(Eigenvalue)*

INFÉRENCE

Inference

L'inférence est un raisonnement par induction que l'on fait à partir d'informations récoltées sur un **échantillon** afin de les généraliser à la **population** associée à cet échantillon.

HISTORIQUE

Voir **statistique inférentielle**.

MOT CLÉ

Statistique inférentielle *(Inferential statistics)*

INTERACTION

Interaction

Le terme interaction est utilisé dans le domaine des plans d'expérience, plus précisément dans une **expérience factorielle** où un certain nombre de facteurs peuvent être étudiés simultanément. Il apparaît lorsque l'effet sur la **variable dépendante** d'un changement de **niveau** d'un facteur dépend des niveaux des autres facteurs.

HISTORIQUE

Le concept d'interaction dans le domaine des plans d'expérience est dû à **Ronald Aylmer Fisher** (1925 et 1935).

ASPECTS MATHÉMATIQUES

Le **modèle** d'un plan à classification double avec interaction se présente ainsi :

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk},$$

$i = 1, 2, \ldots, a$ (**niveau** du **facteur**) A,
$j = 1, 2, \ldots, b$ (niveau du facteur) B et
$k = 1, 2, \ldots, n_{ij}$ (nombre d'observation recevant le traitement),

où

Y_{ijk}	représente la k-ème **observation** recevant le **traitement** ij,
μ	est la **moyenne** générale commune à tous les traitements,
α_i	est l'effet du i-ème niveau du facteur A,
β_j	est l'effet du j-ème niveau du facteur B,
$(\alpha\beta)_{ij}$	est l'effet de l'interaction entre α_i et β_j et
ε_{ijk}	est l'**erreur** expérimentale de l'observation Y_{ijk}.

DOMAINES ET LIMITATIONS

Les interactions entre deux **facteurs** sont appelées interactions de premier ordre, celles concernant trois facteurs sont appelées interactions de second ordre et ainsi de suite.

EXEMPLES

Voir **analyse de variance pour deux facteurs**.

MOTS CLÉS

Analyse de variance *(Analysis of variance)*
Analyse de variance pour deux facteurs *(Two-way analysis of variance)*
Expérience *(Experiment)*
Loi de Fisher *(Fisher distribution)*
Plan d'expérience *(Design of experiments)*
Table de Fisher *(Fisher table)*
Test de Fisher *(Fisher test)*

RÉFÉRENCES

Cox, D.R. (1984). Interaction, *International Statistical Review* 52, pp. 1-25.

Fisher, R.A. (1925). *Statistical Methods for Research Workers.* Oliver & Boyd, Edinburgh.

Fisher, R.A. (1935). *The Design of Experiments.* Oliver & Boyd, Edinburgh.

INTERVALLE

Interval

Un intervalle est un ensemble de nombres réels déterminé par deux bornes, a et b. Ces dernières sont appelées les valeurs limites de l'intervalle.

On distingue l'intervalle fermé, ouvert, semi-ouvert ou semi-fermé.

- *Intervalle fermé :*
  noté $[a, b]$, représente l'ensemble des x tels que
  $$a \leq x \leq b.$$
- *Intervalle ouvert :*
  noté $]a, b[$, représente l'ensemble des x tels que
  $$a < x < b.$$
- *Intervalle semi-ouvert à gauche ou semi-fermé à droite :*
  noté $]a, b]$, représente l'ensemble des x tels que :
  $$a < x \leq b.$$
- *Intervalle semi-ouvert à droite ou semi-fermé à gauche :*
  noté $[a, b[$, représente l'ensemble des x tels que
  $$a \leq x < b.$$

MOTS CLÉS

Distribution de fréquences *(Frequency distribution)*
Histogramme *(Histogram)*
Ogive *(Ogive)*

INTERVALLE DE CONFIANCE

Confidence interval

On appelle intervalle de confiance tout **intervalle** construit autour d'un **estimateur** ayant une certaine **probabilité** de contenir la **valeur** du **paramètre** correspondant de la **population**.

HISTORIQUE

Selon A. Desrosières (1988), A.L. Bowley fut l'un des premiers à s'intéresser à la notion d'intervalle de confiance. C'est en 1906 que A.L. Bowley présenta à la Royal Statistical Society ses premiers calculs d'intervalle de confiance.

ASPECTS MATHÉMATIQUES

Pour construire un intervalle de confiance qui contient la **valeur** du **paramètre** θ avec une **probabilité** donnée, on doit résoudre une équation de la forme :

$$P\left(L_i \leq \theta \leq L_s\right) = 1 - \alpha$$

où

θ	est le paramètre à estimer,
L_i	la limite inférieure de l'**intervalle**,
L_s	la limite supérieure de l'intervalle et
$1 - \alpha$	la probabilité donnée que l'on appelle **niveau de confiance** de l'intervalle.

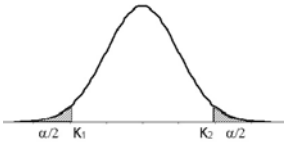
La probabilité α mesure le risque d'**erreur** de l'intervalle, c'est-à-dire la probabilité que l'intervalle ne contienne pas la vraie valeur du paramètre θ.

Pour résoudre cette équation, il faut commencer par définir une fonction $f(t, \theta)$, où t est un **estimateur** de θ, dont on connaît la **loi de probabilité**.

Définir cet intervalle pour $f\left(t, \theta\right)$ revient à écrire l'équation :

$$P\left(k_1 \leq f\left(t, \theta\right) \leq k_2\right) = 1 - \alpha$$

où les constantes k_1 et k_2 sont données par la loi de probabilité de la fonction $f(t,\theta)$. Généralement, le risque d'erreur α est divisé en deux parties égales à $\frac{\alpha}{2}$ et il est réparti de chaque côté de la distribution de $f(t,\theta)$. Si, par exemple, la fonction $f(t,\theta)$ suit une **loi normale** centrée réduite, les constantes k_1 et k_2 seront symétriques et pourront être représentées par $-z_{\frac{\alpha}{2}}$ et $+z_{\frac{\alpha}{2}}$ comme le montre la figure ci-dessous.



Une fois les constantes k_1 et k_2 trouvées, il faut isoler le paramètre θ dans l'équation ci-dessus. On obtient ainsi un intervalle de confiance pour θ au niveau de confiance $1-\alpha$.

DOMAINES ET LIMITATIONS

Il faut être très prudent dans l'interprétation de l'intervalle de confiance. Si, à un **niveau de confiance** de 95 %, nous trouvons un intervalle de confiance pour une **moyenne** μ dont les bornes inférieures et supérieures sont respectivement k_1 et k_2, nous pouvons conclure, par exemple :

« sur la base de l'**échantillon** étudié, nous pouvons affirmer qu'il est probable que la moyenne de la **population** se trouve dans l'**intervalle** que nous avons établi. »

En revanche, il ne serait pas exact de conclure qu'« il y a 95 % de chances pour que la moyenne de la population se trouve dans l'intervalle ». En effet, μ étant une constante ainsi que les bornes k_1 et k_2, l'intervalle contient ou ne contient pas μ. Cependant, si le statisticien a la possibilité de répéter plusieurs fois l'**expérience** qui consiste à tirer un échantillon de la population, 95 % des intervalles obtenus contiendront la vraie **valeur** de μ.

EXEMPLE

Une entreprise fabriquant des ampoules électriques veut étudier la durée de vie moyenne de ses ampoules. La distribution de la **variable aléatoire** X qui représente la durée de vie en heures est la **loi normale** de **moyenne** μ et d'**écart type** $\sigma = 30$.

Pour estimer μ, l'entreprise fait brûler $n = 25$ ampoules.

Elle obtient un temps moyen de durée $\bar{x} = 860$ heures. Elle veut construire un intervalle de confiance autour de l'**estimateur** $\bar{x}$ à un **niveau de confiance** égal à $0,95$. Il faut d'abord trouver une fonction $f(t,\theta) = f(\bar{x},\mu)$ dont on connaît la distribution. Ici, on prend :

$$f(t,\theta) = f(\bar{x},\mu) = \frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}}$$

qui suit une loi normale centrée réduite. L'équation $P(k_1 \leq f(t,\theta) \leq k_2) = 1-\alpha$ devient alors :

$$P\left(-z_{0,025} \leq \frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}} \leq z_{0,025}\right) = 0,95$$

puisque le risque d'**erreur** α a été divisé en deux parties égales à $\frac{\alpha}{2} = 0,025$.

La table de la loi normale centrée réduite, ou **table normale**, nous donne $z_{0,025} = 1,96$. On a donc :

$$P\left(-1,96 \leq \frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}} \leq 1,96\right) = 0,95.$$

Pour obtenir l'intervalle de confiance pour μ au niveau de confiance $0,95$, il faut isoler μ dans l'équation ci-dessus :

$$\begin{aligned} P\left(-1,96 \leq \frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}} \leq 1,96\right) &= 0,95 \\ P\left(-1,96\frac{\sigma}{\sqrt{n}} \leq \bar{x}-\mu \leq 1,96\frac{\sigma}{\sqrt{n}}\right) &= 0,95 \\ P\left(\bar{x}-1,96\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x}+1,96\frac{\sigma}{\sqrt{n}}\right) &= 0,95. \end{aligned}$$

En remplaçant $\bar{x}$, σ et n par leur **valeur** respective, on obtient :

$$P\left(860 - 1,96\frac{30}{\sqrt{25}} \leq \mu \leq 860 + 1,96\frac{30}{\sqrt{25}}\right) = 0,95$$

$$P\left(848,24 \leq \mu \leq 871,76\right) = 0,95.$$

L'intervalle de confiance pour μ au niveau 0,95 est donc :

$$[848,24\,;\,871,76]\,.$$

Cela signifie qu'on peut affirmer avec une **probabilité** de 95 % que cet **intervalle** contient la vraie **valeur** du **paramètre** μ qui correspond à la durée de vie moyenne des ampoules.

MOTS CLÉS

Estimateur *(Estimator)*
Estimation *(Estimation)*
Niveau de confiance *(Confidence level)*

RÉFÉRENCES

Bowley, A.L. (1906). Presidential address to the economic section of the British Association. *Journal of the Royal Statistical Society*, 69, pp. 540-558.

Desrosières, A. (1988). La partie pour le tout : comment généraliser ? La préhistoire de la contrainte de représentativité. *Estimation et sondages, cinq contributions à l'histoire de la statistique.* Economica, Paris.

INTERVALLE INTERQUARTILE

Interquartile range

L'intervalle interquartile est une **mesure de dispersion** correspondant à la différence entre le premier et le troisième **quartile**.

Il correspond donc à l'**intervalle** comprenant 50 % des observations les plus au centre de la distribution.

ASPECTS MATHÉMATIQUES

Soient Q_1 et Q_3 les premier et troisième **quartiles** d'une distribution. L'intervalle interquartile se calcule de la manière suivante :

$$\text{intervalle interquartile } = Q_3 - Q_1.$$

DOMAINES ET LIMITATIONS

L'intervalle interquartile est une mesure de variabilité qui ne dépend pas du nombre d'observations. De plus, et contrairement à l'**empan**, cette mesure est nettement moins sensible aux observations aberrantes.

EXEMPLE

Considérons l'ensemble suivant de 10 observations :

$$0\ 1\ 1\ 2\ 2\ 2\ 3\ 3\ 4\ 5$$

Le premier **quartile** Q_1 est égal à 1 et le troisième quartile Q_3 est égal à 3. L'intervalle interquartile est donc égal à :

$$Q_3 - Q_1 = 3 - 1 = 2.$$

Nous pouvons donc conclure que 50 % des observations se situent dans un **intervalle** de longueur 2, soit l'intervalle allant de 1 à 3.

MOTS CLÉS

Empan *(Range)*
Mesure de dispersion *(Measure of dispersion)*
Quartile *(Quartile)*

INVERSE GÉNÉRALISÉ

Generalized inverse

L'inverse généralisé est le correspondant de l'inverse d'une **matrice** carrée non singulière, mais pour une matrice de dimension et de rang quelconques.

L'inverse généralisé est utilisé dans la résolution des systèmes d'équations linéaires.

HISTORIQUE

Il semble que ce soit Erik Ivar Fredholm (1903) qui le premier ait considéré le concept d'inverse généralisé. Par la suite, Eliakim Hastings Moore a défini un inverse généralisé unique dans son livre *General Analysis* (1935) publié après sa mort. Mais son travail a été inutilisé jusque vers les années 50 quand un regain d'intérêt dû à l'application de l'inverse généralisé à des problèmes liés aux **moindres carrés** se manifesta.

En 1955, Penrose a montré que l'inverse généralisé de Moore est l'unique **matrice**, notée **G**, qui puisse satisfaire les 4 équations suivantes :

1. $\mathbf{A} = \mathbf{A} \cdot \mathbf{G} \cdot \mathbf{A}$
2. $\mathbf{G} = \mathbf{G} \cdot \mathbf{A} \cdot \mathbf{G}$
3. $(\mathbf{A} \cdot \mathbf{G})' = \mathbf{A} \cdot \mathbf{G}$
4. $(\mathbf{G} \cdot \mathbf{A})' = \mathbf{G} \cdot \mathbf{A}$

Cet inverse généralisé unique est actuellement connu sous le nom d'inverse de Moore-Penrose et désigné par $\mathbf{A}^+$.

ASPECTS MATHÉMATIQUES

La **matrice G** de dimension $n \times m$ est l'inverse généralisé de la matrice **A** de dimension $m \times n$ si

$$\mathbf{A} = \mathbf{A} \cdot \mathbf{G} \cdot \mathbf{A}.$$

La matrice **G** est unique si et seulement si $m = n$ et **A** est non singulière.

La notation la plus courante de l'inverse généralisé d'une matrice **A** est $\mathbf{A}^-$.

MOTS CLÉS

Matrice inverse *(Inverse matrix)*
Moindres carrés *(Least squares)*

RÉFÉRENCES

Bjerhammar, A. (1951). Rectangular reciprocal matrices with special reference to geodetic calculations. *Bull. Geod.*, 20, pp. 188-220.

Dodge, Y. and Majumdar, D. (1979). An Algorithm for Finding Least Square Generalized Inverses for Classification Models with Arbitrary Patterns. *Journal of Statistical Computation and Simulation*, Vol. 9, pp. 1-17.

Fredholm, J. (1903). Sur une classe d'équations fonctionnelles. *Acta Mathematica*, 27, pp. 365-390.

Moore, E.H. (1935). *General Analysis*, Part I. Memoirs of the American Philosophical Society, Philadelphia, pp. 147-209.

Penrose, R. (1955). A generalized inverse for matrices. *Proceedings of the Cambridge Philosophical Society*, 51, pp. 406-413.

Rao, C.R. (1962). A note on generalized inverse of a matrix with applications to problems in mathematical statistics. *Journal of the Royal Statistical Society.* Series B, 24, pp. 152-158.

Rao, C.R. (1967). Calculus of generalized inverse of matrices. I : General theory, *Sankhya*, A29, pp. 317-342.

Rao, C.R. and Mitra, S.K. (1971). *Generalized Inverses of Matrices and its Applications.* John Wiley & Sons, New York.

INVERSION

Inversion

Considérons une **permutation** des entiers 1, 2, ..., n. Lorsque, dans une paire d'éléments (non nécessairement adjacents) de la permutation, le premier élément de la paire est plus grand que le second, on dit qu'il y a inversion. Ainsi, on peut compter dans une permutation le nombre d'inversions qui n'est autre que le nombre de couples dans lesquels le grand entier précède le petit.

ASPECTS MATHÉMATIQUES

Soit une **permutation** des entiers $1, 2, \ldots, n$ donnée par $\sigma_1, \sigma_2, \ldots, \sigma_n$; on détermine le nombre d'inver-

sions en comptant le nombre de couples (σ_i, σ_j) tels que $\sigma_i > \sigma_j$.

Pour être certain de ne pas en oublier, on commence par prendre σ_1 et on compte les σ_i, $i > 1$, plus petits que σ_1 ; ensuite on prend σ_2 et on compte les éléments qui lui sont inférieurs et qui sont situés à sa droite. On procède ainsi jusqu'à σ_{n-1} ; on obtient ainsi le nombre d'inversions de la permutation considérée.

DOMAINES ET LIMITATIONS

Le nombre d'inversions est utilisé pour le calcul du **déterminant** d'une **matrice** carrée. On utilise pour cela la notation a_{ij} pour l'élément de la matrice situé à l'intersection de la i-ème ligne et de la j-ème colonne. En effectuant le produit $a_{1i} \cdot a_{1j} \cdot \ldots \cdot a_{1r}$, par exemple, la parité du nombre d'inversions de la permutation $i\ j \ldots r$ déterminera le signe attribué au produit pour le calcul du déterminant.

Pour une **permutation** des éléments $1, 2, \ldots, n$; le nombre d'inversions est toujours un entier compris entre 0 et $\frac{n \cdot (n-1)}{2}$.

EXEMPLES

Prenons les entiers 1, 2, 3, 4, 5 ; une **permutation** de ces cinq entiers est donnée, par exemple, par :

$$1\ 4\ 5\ 2\ 3$$

On peut alors déterminer le nombre d'inversions dans cette permutation en écrivant les couples dans lesquels il y a inversion :

$$(4, 2)$$
$$(4, 3)$$
$$(5, 2)$$
$$(5, 3)$$

D'où le nombre d'inversions est égal à 4.

Pour la permutation 1 2 3 4 5 (permutation identité), on trouve qu'il n'y a aucune inversion.

Pour la permutation 5 4 3 2 1, on obtient 10 inversions, car toute paire est une inversion et donc le nombre d'inversions est égal au nombre de **combinaisons** de 2 éléments choisis sans répétition sur 5, à savoir :

$$C_5^2 = \frac{5!}{2! \cdot 3!} = 10.$$

MOTS CLÉS

Déterminant *(Determinant)*
Permutation *(Permutation)*



J

JACKKNIFE, MÉTHODE DU

Jackknife method

Voir **méthode du jackknife**.

JEFFREYS HAROLD

Jeffreys Harold



Harold Jeffreys est né à Fatfield (Angleterre) en 1891. Il étudia les mathématiques, la physique, la chimie et la géologie au Armstrong College à Newcastle d'où il fut diplômé en 1910. Ensuite, il rejoignit le St. John's College à Cambridge où il devint chargé de cours en 1914 et où il resta durant soixante-quinze ans. Il mourut en 1989.

Ouvrages principaux d'Harold Jeffreys :

1939 *Theory of Probability.* Oxford : Oxford University Press.

1946 (and Jeffreys, B.S.) *Methods of mathematical physics*, Cambridge University Press, Cambridge. The Macmillan Company, New York.



K

KOLMOGOROV ANDREÏ NICOLAÏÉVITCH

Kolmogorov Andrei Nikolaevich



Né à Tambov (Russie) en 1903, Andreï Nicolaïévitch Kolmogorov est un des grands fondateurs des probabilités modernes. En 1920, il entre à l'Université d'État de Moscou et fait ses études de mathématiques, histoire et métallurgie. En 1925, il publie son premier article en probabilités sur les inégalités des sommes partielles des variables aléatoires qui devient la référence principale dans le domaine des processus stochastiques. Il obtient son doctorat en 1929 et publie 18 articles qui portent sur la loi des grands nombres ainsi que sur la logique intuitive. Il est nommé professeur ordinaire à l'Université de Moscou en 1931. En 1933, il publie sa monographie sur la théorie des probabilités, *Grundbegriffe der Wahrscheinlichkeitsrechnung*, où il procède de manière très rigoureuse débutant par la base axiomatique des probabilités d'une manière comparable de celle d'Euclide sur la géométrie.

En 1939 il est élu membre de l'Académie des sciences de l'URSS. Il a reçu le Prix Lénine en 1965, l'Ordre de Lénine dans six occasions distinctes ainsi que le Prix Lobachevsky en 1987. Il a été élu membre de plusieurs autres académies étrangères parmi lesquelles l'Académie roumaine des sciences (1956), la Royal Statistical Society de Londres (1956), l'Académie Leopoldina d'Allemagne (1959), l'American Academy of Arts and Sciences (1959), ainsi que de la London Mathematical Society (1959), l'American Philosophical Society (1961), l'Institut indien de statistique (1962), l'Académie hollandaise des sciences (1963), la Royal Society de Londres (1964), l'Académie nationale des États-Unis (1967) et l'Académie française des sciences (1968).

Il décède à Moscou en 1987.

Quelques ouvrages et articles principaux d'Andreï Nicolaïévitch Kolmogorov :

1933 *Grundbegriffe der Wahrscheinlichkeitsrechnung.* Springer-Verlag, Berlin.

1933 Sulla determinazione empirica di una lege di distributione. *Giornale dell'Instituto Italiano degli Attuari*, 4, pp. 83-91 (6.1).

1941 Local structure of turbulence in incompressible fluids with very high Reynolds number. *Dan SSSR*, 30, 229.

1941 Dissipation of energy in locally isotropic turbulence. *Dokl. Akad. Nauk. SSSR*, 32, pp. 16-18.

1958 (and Uspenskii, V.A.) Kopredeleniyu algoritma. (On the Definition of an Algorithm). *Uspekhi Matematicheskikh Nauk* 13(4), pp. 3-28, *American Mathematical Society Translations Series* 2 29, pp. 217-245, 1963.

1961 (and Fomin, S.V.) *Measure, Lebesgue integrals and Hilbert space.* Natascha Artin Brunswick and Alan Jeffrey. Academic Press, New York.

1963 On the Representation of Continuous Functions of Many Variables by Superposition of Continuous Functions of One Variable and Addition. *Doklady Akademii Nauk SSR*, Vol. 114, pp. 953-956 ; 1957. English translation. *Mathematical Society Transaction*, Vol. 28, pp. 55-59.

1965 Three approaches to the quantitative definition of information. *Problems of Information Transmission*, vol. 1, pp. 1-17, translation of *Problemy Peredachi Informatsii* 1, n° 1, pp. 3-11 (1965).

1987 (and Uspenskii, V.A.) Algorithms and Randomness. *Theoria Veroyatnostey i ee Primeneniya (Theory of Probability and its Applications),* 3(32), pp. 389-412.

MOT CLÉ

Test de Kolmogorov-Smirnov *(Kolmogorov-Smirnov test)*

KOLMOGOROV-SMIRNOV, TEST DE

Kolmogorov-Smirnov test

Voir **test de Kolmogorov-Smirnov**.

KRUSKAL-WALLIS, TABLE DE

Kruskal-Wallis table

Voir **table de Kruskal-Wallis**.

KRUSKAL-WALLIS, TEST DE

Kruskal-Wallis test

Voir **test de Kruskal-Wallis**.



L

LAPLACE, LOI DE

Laplace distribution

Voir **loi de Laplace**.

LAPLACE PIERRE SIMON DE

Laplace Pierre Simon de



Né en France en 1749, dans une famille bourgeoise, le marquis Pierre Simon de Laplace fut l'un des pionniers des statistiques. Intéressé par les mathématiques, l'astronomie théorique, les probabilités et les statistiques, il fit paraître ses premières publications au début des années 1770.

Membre de l'Académie des sciences en 1785, il dirigea également le Bureau des longitudes.

En 1799, il fut nommé ministre de l'Intérieur par Bonaparte qui, en 1806, lui conféra le titre de comte.

Élu à l'Académie française en 1816, il poursuivit son œuvre scientifique jusqu'à sa mort, à Paris, en 1827.

Quelques ouvrages et articles principaux de Pierre Simon de Laplace :

1774 *Mémoire sur la probabilité des causes par les événements.* Mémoire de l'Académie royale des sciences présenté par divers savants, 6, pp. 621-656 ou Vol. 8, pp. 27-65 dans P.S. de Laplace, *Œuvres complètes* (1891). Gauthier-Villars, Paris.

1781 *Mémoire sur les probabilités.* Mémoire de l'Académie royale des sciences de Paris, 1778, pp. 227-332 ou Vol. 9 pp. 383-485 dans P.S. de Laplace, *Œuvres complètes* (1891). Gauthier-Villars, Paris.

1805 *Traité de mécanique céleste.* Vol. 1-4. Duprot. Paris. et vol. 5. Bachelier, Imprimeur-Libraire pour les Mathématiques, Paris.

1812 *Théorie analytique des probabilités.* Courcier, Paris.

1878-1912 *Œuvres complètes de Laplace.* 14 volumes. Gauthier-Villars, Paris.

LEGENDRE ADRIEN-MARIE

Legendre Adrien-Marie

Né en 1752, Adrien-Marie Legendre succéda à **Pierre Simon de Laplace** comme professeur de mathématiques, d'abord à l'École militaire, puis à l'École normale de Paris. Il s'intéressa à la théorie et à la pratique de l'astronomie et de la géodésie. Dès 1792, il fit partie de la commission française chargée de mesurer la longueur d'un quadrant de méridien du pôle Nord à l'Équateur en passant par Paris. Il mourut en 1833.

Dans l'histoire de la **statistique**, son nom reste

attaché à la publication de son ouvrage de 1805, dans lequel figure un appendice sur la méthode des **moindres carrés**. Cette publication fut à l'origine de la controverse qui l'opposa à **Carl Friedrich Gauss** qui lui contestait la découverte de cette méthode.



Ouvrage principal d'Adrien-Marie Legendre :

1805 *Nouvelles méthodes pour la détermination des orbites des comètes.* Courcier, Paris.

RÉFÉRENCE

Plackett, R.L. (1977). *Studies in the history of probability and statistics. The discovery of the method of least squares.* Vol. II. Edited by M. Kendall and R.L. Plackett. Charles Griffin & Company Ltd.

LOGICIEL STATISTIQUE

Statistical software

Un logiciel statistique est un ensemble de programmes informatiques et de procédés relatifs au traitement de données statistiques.

Librairies

Une librairie est une série de programmes ou sous-routines qui sont installés sous un même système d'exploitation (Dos, Windows ou Unix) et pouvant être appelés en tapant des commandes individuelles.

IMSL (*international mathematics and statistics library*) : IMSL est un ensemble d'environ 540 sous-routines écrites en FORTRAN concernant spécifiquement les domaines des mathématiques et des statistiques.

NAG (*numerical algorithms group*) : NAG est une librairie d'algorithmes écrits à l'aide de trois langages (ALGOL 60, ALGOL 68 et FORTRAN ANSI). On y trouve plus de 600 sous-routines avancées en ce qui concerne la **simulation**, l'analyse de régression, les techniques d'**optimisation**, Glim (*General LInear Models*), GenStat, les séries chronologiques et les représentations graphiques. De plus, on peut traiter d'une manière simple les réponses obtenues par sondage et appliquer des analyses multivariées ainsi que des tests non paramétriques.

Logiciels

Un logiciel est un ensemble complexe d'algorithmes présentant une structure commune des données et nécessitant un minimum de programmation ultérieure de la part de l'utilisateur.

BMDP : à la base destiné au domaine bio-médical, BMDP est un logiciel complet, formé de sous-routines écrites en FORTRAN, qui permet de traiter des enquêtes effectuées par **sondage**, de générer des représentations graphiques, d'appliquer des techniques multivariées, des tests non paramétriques ou d'effectuer une **analyse de régression**, aussi bien linéaire que non linéaire, sans oublier l'étude de séries chronologiques. Chaque sous-programme est basé sur les algorithmes de calcul les plus compétitifs rigoureusement testés dans les vingt ans d'existence du logiciel.

DataDesk : DataDesk est un logiciel interactif pour l'analyse et le traitement statistique des

données, très visuel, présentant des outils permettant l'établissement des liaisons entre les données, des tendances, des sous-groupes et des valeurs aberrantes.

EViews : logiciel essentiellement destiné aux économistes, EViews est l'outil de prédilection pour l'analyse, la prévision et la modélisation économétrique. Cependant, EViews peut être utilisé pour l'analyse statistique générale, pour l'estimation des séries chronologique, pour des simulations à large échelle, des graphiques de présentation ou simplement pour la gestion des données.

Excel : les utilitaires d'analyse Microsoft Excel permettent à l'utilisateur de représenter et d'analyser les données à l'aide des méthodes statistiques de base : statistique descriptive, analyse de variance, régression et corrélation, tests d'hypothèse.

Gauss : programme statistique orienté matrice, Gauss est un puissant outil statistique, économétrique et graphique.

JMP : logiciel statistique destiné à l'analyse des plans d'expérience en particulier, JMP contient aussi des outils statistiques de base exhaustifs.

Lisrel : logiciel utilisé dans le domaine des sciences sociales et plus spécifiquement pour l'analyse factorielle et la modélisation, Lisrel est particulièrement utile pour la modélisation présentant des variables latentes.

Maple : Maple est un environnement mathématique complet destiné à la manipulation des expressions algébriques, à la résolution d'équations et intégrales et qui est muni de très puissants outils graphiques en deux et trois dimensions. L'utilisateur peut également programmer dans Maple.

Mathematica : le module statistique de Mathematica permet d'effectuer des statistiques descriptives univariées et multivariées, du lissage des données, des tests d'hypothèse classiques, l'estimation d'intervalles de confiance, ainsi que la régression linéaire et non linéaire.

Matlab : Matlab est un environnement interactif qui peut être utilisé pour l'analyse scientifique et statistique des données. Les objets de base de Matlab sont des matrices. L'utilisateur peut entreprendre des analyses numériques, des traitements de signaux, des traitements d'images, ainsi que des statistiques sur les matrices. L'utilisateur peut programmer des fonctions à l'aide de C et de FORTRAN.

MIM : MIM est un programme de modélisation graphique pour des données discrètes et continues. Parmi les familles de graphiques disponibles dans ce logiciel, on peut citer : les modèles log-linéaires, les modèles graphiques gaussiens, les modèles Manova standards et de multiples autres modèles utiles pour l'**analyse multivariée**.

Minitab : Minitab est certainement le système le plus simple à employer. Pouvant être utilisé en mode interactif, il permet d'effectuer des analyses de types variés : **statistique descriptive**, **analyse de variance**, **analyse de régression**, tests non paramétriques, **génération de nombres aléatoires** ou étude de séries chronologiques, entre autres.

P-STAT : le logiciel P-STAT a été créé en 1960 pour le traitement des données pharmacologiques. Aujourd'hui destiné aux chercheurs des sciences sociales, il offre notamment la possibilité de faire de l'analyse de sondages. La puissance du logiciel P-STAT réside dans l'existence de Macros et de programmes écrits dans le langage PPL (*P-STAT Programming Language*) et utiles dans les applications statistiques liées au monde des affaires. Il offre la possibilité de traiter de larges sets de données. Il traite des plans d'expérience, des représentations graphiques, des tests non paramétriques et de l'**analyse de données**.

R : langage de programmation similaire à S, R est également un logiciel gratuit largement utilisé dans les milieux académiques. Il présente une implémentation différente de S, cependant une large palette de commandes dans S fonctionne dans R. R contient une large variété de commandes statistiques (modélisation linéaire et non

linéaire, exécution des tests classiques, analyse de séries temporelles, classification, segmentation, ...) ainsi que des techniques graphiques.

SAS (*Statistical Analysis System*) : SAS est un ensemble de logiciels incluant un langage SAS de base ainsi que des logiciels spécifiques. Sans doute le plus complet, il permet d'analyser des données provenant de domaines très divers : la recherche, l'ingénierie, la médecine et les applications en affaires. Il inclut une importante quantité d'outils statistiques : analyse de variance, régression, analyse de données catégoriques, analyse multivariée, analyse de survie, arbres de décision, analyse non paramétrique et *data mining*.

S-plus : logiciel écrit dans un langage de programmation puissant et flexible orienté objet appelé S, il permet le calcul matriciel, les calculs mathématiques et statistiques. Il inclut des outils d'analyse exploratoire visuelle et permet également l'analyse approfondie des données ainsi que leur modélisation. Pour les géostatisticiens, S-plus permet de rattacher un module « S+SpatialStats » contenant différents outils de **statistique spatiale**.

Spad-T : c'est un produit purement basé sur des techniques statistiques d'analyse multivariée. On réalise des tables de contingence entre variables traitées à l'aide des méthodes multifactorielles donnant lieu à des représentations graphiques.

SPSS : apparu sur le marché pour la première fois en 1968, le logiciel SPSS (*Statistical Package for the Social Sciences*) est un des plus utilisés actuellement par les statisticiens des domaines les plus divers. Il offre des moyens interactifs pour le traitement de grandes bases de données. On y trouve aussi quelques algorithmes de représentations graphiques.

Stata : c'est un logiciel statistique permettant la réalisation d'analyses statistiques de base allant de la modélisation linéaire aux modèles linéaires généralisés, traitant des données binaires, des méthodes non paramétriques, de l'analyse multivariée, de l'analyse des séries temporelles ainsi que des méthodes graphiques d'exploration des données. De plus, le logiciel permet de programmer des commandes spécifiques à l'utilisateur. Stata contient, entre autres, des outils de statistique destinés aux épidémiologistes.

Statistica : arrivé sur le marché en 1993, Statistica est un logiciel extrêmement complet, qui contient de puissants outils d'exploration de larges jeux de données. Le logiciel de base inclut les statistiques de base, l'analyse de variance, les statistiques non paramétriques, les modèles linéaires et non linéaires, les techniques d'analyse exploratoire multivariée, les réseaux neuronaux, les calculs de la taille des échantillons, des outils d'analyse des plans d'expérience, des outils de contrôle de qualité, etc.

StatView : logiciel statistique adressé aux scientifiques et chercheurs du domaine médical, StatView permet de faire de la statistique descriptive, des tests d'hypothèse paramétriques et non paramétriques, le calcul de la corrélation et de la covariance, de la régression linéaire et non linéaire, de la régression logistique, des tables de contingence, de l'analyse factorielle, de l'analyse de survie, du contrôle de qualité, de l'analyse de Pareto. Une bonne interface graphique est disponible.

Systat : puissant logiciel graphique, Systat permet la visualisation des données ainsi que l'utilisation de statistiques de base afin de trouver la modélisation adéquate.

Solas : logiciel de traitement des données manquantes, Solas intègre les algorithmes d'imputation multiple et l'analyse combinée des données ainsi obtenues.

Vista : destiné autant à ceux qui s'initient à l'étude statistique qu'au professeurs, Vista peut être utilisé comme logiciel d'apprentisage dans les domaines de la statistique univariée, multivariée, graphique et calculatoire.

WinBugs : c'est un programme de modélisation bayesienne de données basé sur des techniques de simulation Monte-Carlo. Il inglobe des représentations des modèles bayesiens avec des commandes pour réaliser des simulations. De plus, il existe une interface permettant à tout moment le contrôle de l'analyse. Des outils graphiques permettent de suivre la convergence de la simulation.

XploRe : XploRe est un environnement statistique très vaste destiné à l'analyse exploratoire des données. Étant programmé dans un langage orienté matrice, il contient une large gamme d'opérations statistiques et possède une puissante interface graphique interactive. L'utilisateur a la possibilité d'écrire ses propres macros et le programme supporte des jeux de données très importants.

HISTORIQUE

Les premiers efforts entrepris par les statisticiens pour simplifier leurs calculs statistiques ont été la construction des tables statistiques. Avec l'avénement des ordinateurs, au milieu des années 1940, le codage des routines statistiques prend forme. Cependant, le langage de programmation de l'époque était peu connu par les chercheurs et il a fallu attendre le début des années 1960, lorsque le langage de programmation FORTRAN vit le jour, pour que les logiciels statistiques puissent se dévélopper. Parmi les premières routines statistiques introduites par IBM, la plus connue est le générateur de nombres pseudo-aléatoires RANDU. Le développement systématique des logiciels statistiques commence en 1960 par la création des programmes BDM et SPSS suivis par SAS.

RÉFÉRENCES

Asimov, D. (1985). The grand tour : A tool for viewing multidimensional data. *SIAM Journal on Scientific and Statisting Computing*, 6(1), pp. 128-143.

Becker, R.A., Chambers, J.M. and Wilks, A.R. (1988). *The New S Language.* Waldsworth and Brooks/Cole, Pacific Grove, CA.

Chambers, J.M and Hastie, T.K (eds.) (1992). *Statistical models in S.* Waldsworth and Brooks/ Cole, Pacific Grove, CA.

Dixon, W.J. chief editor (1981). *BMDP Statistical Software Communications.* The University of California Press. Berkeley, C.A.

Dodge, Y. and Hand, D. (1991). *What Should Future Statistical Software Look like ?* CSDA.

Dodge, Y. and Whittaker, D.R. editors (1992). *Computational statistics.* Volumes 1 and 2. COMPSTAT, Neuchâtel, Switzerland, August 24-28. (English) Heidelberg : Physica-Verlag.

Edwards, D. (1995). *Introduction to Graphical Modelling.* Springer-Verlag, New York.

Francis, I. (1981). *Statistical software: A comparative Review.* North Holland, New York.

Hayes, A.R. (1982). *Statistical Software: A survey and Critique of its Development.* Office of Naval Research, Arlington, VA.

International Mathematical and Statistical Libraries, Inc. (1981). IMSL Library Information. IMSLK, Houston, TX.

Ripley, B.D. and Venables, W.N. (2000) *S programming.* Springer-Verlag, New York.

Statistical Sciences (1993). *Statistical Analysis in S-Plus*, Version 3.1. Seattle : StatSci, a division of MathSoft, Inc.

Venables, W.N. and Ripley, B.D. (1994). *Modern Applied Statistics with S-PLUS.* Springer-Verlag, New York.

Wegman, E.J. and Hayes, A.R. (1988). Statistical Software. *Encyclopedia of Statistical Sciences*, 8, pp. 667-674.

LOGNORMALE, LOI

Lognormal distribution

Voir **loi lognormale**.

LOI BÊTA

Beta distribution

Une **variable aléatoire** X suit une loi bêta de paramètres α et β si sa **fonction de densité** est de la forme :

$$f(x) = \frac{(x-a)^{\alpha-1}(b-x)^{\beta-1}(b-a)^{-(\alpha+\beta-1)}}{B(\alpha,\beta)},$$

$$a \leq x \leq b,$$
$$\alpha > 0 \text{ et } \beta > 0.$$

où :

$$\begin{aligned} B(\alpha,\beta) &= \int_0^1 t^{\alpha-1}(1-t)^{\beta-1}\,dt \\ &= \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} \end{aligned}$$

et Γ est la fonction gamma (voir **loi gamma**).



Loi bêta $\alpha = 2$, $\beta = 3$, $a = 3$, $b = 7$.

La loi bêta est une **loi de probabilité continue**.

ASPECTS MATHÉMATIQUES

En faisant le changement de **variable** $Y = \frac{X-a}{b-a}$, on obtient la **fonction de densité** de la loi bêta standard :

$$f(y) = \begin{cases} \frac{1}{B(\alpha,\beta)} y^{\alpha-1} \cdot (1-y)^{\beta-1} & \text{si } 0 < y < 1, \\ 0 & \text{sinon.} \end{cases}$$

Soit X une **variable aléatoire** suivant une loi bêta standard. L'**espérance mathématique** et la **variance** de X sont respectivement données par :

$$E[X] = \frac{\alpha}{\alpha+\beta}$$

$$Var(X) = \frac{\alpha \cdot \beta}{(\alpha+\beta)^2 \cdot (\alpha+\beta+1)}.$$

DOMAINES ET LIMITATIONS

La loi bêta est une des lois les plus fréquemment utilisées pour ajuster des distributions empiriques dont l'**empan** (ou **intervalle** de variation) $[a,b]$ est connu.

Citons quelques cas particuliers de loi bêta, se rapportant à d'autres lois de probabilité continues :

- si X_1 et X_2 sont deux variables aléatoires indépendantes distribuées selon une **loi gamma** de paramètres $(\alpha_1, 1)$ et $(\alpha_2, 1)$ respectivement, la variable aléatoire

  $$\frac{X_1}{X_1 + X_2}$$

  est distribuée selon une loi bêta de paramètres (α_1, α_2).

- la loi bêta devient une **loi uniforme** quand

  $$\alpha = \beta = 1.$$

- quand les paramètres α et β tendent vers l'infini, la loi bêta tend vers la **loi normale** standard.

MOTS CLÉS

Loi de probabilité continue *(Continuous probability distribution)*

Loi gamma *(Gamma distribution)*
Loi normale *(Normal distribution)*
Loi uniforme *(Uniform distribution)*

LOI BINOMIALE

Binomial distribution

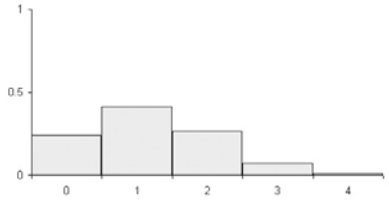
Une **variable aléatoire** X suit une loi binomiale de paramètres n et p si sa **fonction de probabilité** est de la forme :

$$P(X = x) = C_n^x \cdot p^x \cdot q^{n-x}, \quad x = 0, 1, 2, \ldots, n,$$

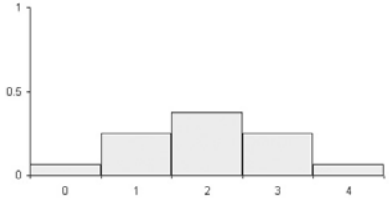
où

p	est la **probabilité** de succès,
q	le probabilité d'échec,
C_n^x	le nombre de combinaisons de x objets parmi n.

Ainsi si un événement est formé de x « succès » et $(n-x)$ « échecs », où p est la probabilité de « succès » et $q = 1-p$ la probabilité d'« échec », la loi ci-dessus permet de calculer la probabilité d'obtenir x « succès » parmi n épreuves indépendantes.



Loi binomiale, $p = 0,3$, $q = 0,7$.



Loi binomiale, $p = 0,5$, $q = 0,5$.

La loi binomiale de paramètres n et p, notée $B(n,p)$, est une **loi de probabilité discrète**.

HISTORIQUE

La loi binomiale compte parmi les plus anciennes lois de probabilités. Cette distribution fut découverte par **Jakob Bernoulli** dans son ouvrage intitulé *Ars Conjectandi* (1713). Ce dernier est divisé en quatre parties : dans la première, l'auteur commente le traité de Huygens ; la deuxième partie est consacrée à la théorie des permutations et combinaisons, tandis que la troisième est relative aux solutions de divers problèmes sur les jeux de hasard ; enfin, dans la quatrième, on trouve des propositions pour appliquer la théorie des probabilités aux questions d'intérêts moraux et de la science économique.

ASPECTS MATHÉMATIQUES

Si $X_1, X_2, \ldots, X_n$, sont n **variables aléatoires** indépendantes suivant une **loi de Bernoulli** de **paramètre** p, alors la **variable** aléatoire

$$X = X_1 + X_2 + \ldots + X_n$$

suit une loi binomiale $B(n,p)$.

Pour calculer l'**espérance mathématique** de X, on utilise la propriété suivante, où Y et Z sont deux variables aléatoires :

$$E[Y + Z] = E[Y] + E[Z].$$

Nous avons donc :

$$\begin{aligned} E[X] &= E[X_1 + X_2 + \ldots + X_n] \\ &= E[X_1] + E[X_2] + \ldots + E[X_n] \\ &= p + p + \ldots + p \\ &= np. \end{aligned}$$

Pour calculer la **variance** de X, on utilise la propriété suivante, où Y et Z sont deux variables indépendantes :

$$Var(Y + Z) = Var(Y) + Var(Z).$$

Nous avons donc :

$$\begin{aligned} Var(X) &= Var(X_1 + X_2 + \ldots + X_n) \\ &= Var(X_1) + Var(X_2) + \ldots + Var(X_n) \\ &= pq + pq + \ldots + pq \\ &= npq. \end{aligned}$$

Des tables de la loi binomiale (de la **fonction de probabilité** et de la **fonction de répartition**) ont été publiées par Rao *et al.* (1985), ainsi que par le National Bureau of Standards (1950). Des tables étendues de la fonction de répartition se trouvent dans les *Annals of the Computation Laboratory* (1955) (voir **table binomiale**).

EXEMPLE

On lance dix fois une pièce de monnaie. Soit la **variable aléatoire** X représentant le nombre de « piles » apparues.

Nous avons donc :

nombre d'épreuves :	$n = 10$
probabilité d'un succès :	$p = \frac{1}{2}$ (obtenir « pile »)
probabilité d'un échec :	$q = \frac{1}{2}$ (obtenir « face »)

La probabilité d'obtenir x fois « pile » parmi les dix épreuves est donnée par :

$$P(X = x) = C_{10}^{x} \cdot \left(\frac{1}{2}\right)^{x} \cdot \left(\frac{1}{2}\right)^{10-x}.$$

La probabilité d'obtenir exactement huit fois « pile » est donc égale à :

$$\begin{aligned} P(X = 8) &= C_{10}^{8} \cdot p^{8} \cdot q^{10-8} \\ &= \frac{10!}{8!\,(10-8)!} \cdot \left(\frac{1}{2}\right)^{8} \cdot \left(\frac{1}{2}\right)^{2} \\ &= 0,0439. \end{aligned}$$

La variable aléatoire X suit une loi binomiale $B\left(10, \frac{1}{2}\right)$.

MOTS CLÉS

Loi binomiale négative *(Negative binomial distribution)*
Loi de Bernoulli *(Bernoulli distribution)*
Loi de probabilité discrète *(Discrete probability distribution)*
Table binomiale *(Binomial table)*
Test binomial *(Binomial test)*

RÉFÉRENCES

Annals of the Computation Laboratory. (1955). *Tables of the Cumulative Binomial Probability Distribution.* Vol. 35, Harvard University.

Bernoulli, J. (1713). *Ars Conjectandi, Opus Posthumum. Accedit Tractatus de Seriebus infinitis, et Epistola Gallice scripta de ludo Pilæ recticularis.* Basileæ, Impensis Thurnisiorum, fratrum.

National Bureau of Standards. (1950). Tables of the Binomial Probability Distribution. *Applied Mathematics Series* 6.

Rao, C.R., Mitra, S.K., Matthai, A. and Ramamurthy, K.G. (1966). *Formulæ and Tables for Statistical Work.* Statistical Publishing Company, Calcutta, pp. 34-37.

LOI BINOMIALE NÉGATIVE

Negative binomial distribution

Une **variable aléatoire** X suit une loi binomiale négative de paramètres k et r si sa **fonction de probabilité** est de la forme :

$$P(X = k) = C_{r+k-1}^{k} \cdot p^{r} \cdot q^{k},$$

où

p	est la **probabilité** de succès,
q	la probabilité d'échec,
C_n^x	le nombre de combinaisons de x objets parmi n.

En se plaçant dans la situation où l'on répète de façon indépendante une **expérience aléatoire** « succès-échec », on note p, la **probabilité** de succès et $q = 1 - p$ la probabilité d'échec.

On décide de répéter l'**expérience** autant de fois qu'il faut pour obtenir r succès. Le nombre d'échecs obtenus avant de réaliser ce but est une variable aléatoire suivant la loi binomiale négative décrite ci-dessus.

La loi binomiale négative est une **loi de probabilité discrète**.

HISTORIQUE

Le premier à avoir traité la loi binomiale négative fut Blaise Pascal (1679). Puis Pierre Rémond de Montmort (1714) appliqua la loi binomiale négative pour représenter le nombre de fois qu'il faut lancer une pièce de monnaie pour obtenir un certain nombre de « faces ».

Student (1907) a utilisé la loi binomiale négative comme une alternative à la **loi de Poisson**.

Greenwood et **George Udny Yule** (1920) ainsi qu'Eggenberger et George Polya (1923) ont trouvé des applications à la loi binomiale négative. Dès lors, il y eut un nombre croissant d'applications de cette loi et parallèlement se sont développées les techniques statistiques basées sur cette distribution.

ASPECTS MATHÉMATIQUES

Si $X_1, X_2, \ldots, X_n$ sont n variables aléatoires indépendantes suivant une **loi géométrique**, alors la variable aléatoire :

$$X = X_1 + X_2 + \ldots + X_r$$

suit une loi binomiale négative.

Pour calculer l'**espérance mathématique** de X, on utilise la propriété suivante, où Y et Z sont des variables aléatoires :

$$E[Y + Z] = E[Y] + E[Z].$$

Nous avons donc :

$$\begin{aligned} E[X] &= E[X_1 + X_2 + \ldots + X_r] \\ &= E[X_1] + E[X_2] + \ldots + E[X_r] \\ &= \frac{q}{p} + \frac{q}{p} + \ldots + \frac{q}{p} \\ &= r \cdot \frac{q}{p}. \end{aligned}$$

Pour calculer la **variance** de X, on utilise la propriété suivante, où Y et Z sont des variables indépendantes :

$$Var(Y + Z) = Var(Y) + Var(Z).$$

Nous avons donc :

$$\begin{aligned} Var(X) &= Var(X_1 + X_2 + \ldots + X_r) \\ &= Var(X_1) + Var(X_2) + \ldots + Var(X_r) \\ &= \frac{q}{p^2} + \frac{q}{p^2} + \ldots + \frac{q}{p^2} \\ &= r \cdot \frac{q}{p^2}. \end{aligned}$$

DOMAINES ET LIMITATIONS

Parmi les domaines spécifiques auxquels la loi binomiale négative a été appliquée, on peut citer les statistiques sur les accidents, les sciences biologiques, l'écologie, l'étude de marché, la recherche médicale et la psychologie.

EXEMPLE

On lance une pièce de monnaie et on s'intéresse à la **probabilité** d'obtenir une troisième « face » au sixième lancer.

Nous avons alors :

- nombre de succès : $r = 3$
- nombre d'échecs : $k = 6 - r = 3$
- probabilité d'un succès : $p = \frac{1}{2}$ (obtenir « face »)
- probabilité d'un échec : $q = \frac{1}{2}$ (obtenir « pile »)

La probabilité d'obtenir k « piles » avant la troisième « face » est donnée par :

$$P(X = k) = C_{3+k-1}^{k} \cdot \left(\frac{1}{2}\right)^3 \cdot \left(\frac{1}{2}\right)^k.$$

La probabilité d'avoir une troisième « face » au sixième lancer, c'est-à-dire la probabilité d'obtenir trois fois « pile » avant d'obtenir la troisième « face » est donc égale à :

$$\begin{aligned} P(X = 3) &= C_{3+3-1}^{3} \cdot \frac{1}{2^3} \cdot \frac{1}{2^3} \\ &= C_5^3 \cdot \frac{1}{2^3} \cdot \frac{1}{2^3} \\ &= \frac{5!}{3!\,(5-3)!} \cdot \frac{1}{2^3} \cdot \frac{1}{2^3} \\ &= 0,1563. \end{aligned}$$

MOTS CLÉS

Loi binomiale *(Binomial distribution)*
Loi de Bernoulli *(Bernoulli distribution)*
Loi de Poisson *(Poisson distribution)*
Loi de probabilité discrète *(Discrete probability distribution)*

RÉFÉRENCES

Eggenberger, F. und Polya, G. (1923). Über die Statistik verketteter Vorgänge. *Zeitschrift für angewandte Mathematische Mechanik*, 3, pp. 279-289.

Gosset, S.W. « **Student** » (1907). On the error of counting with a haemacytometer. *Biometrika* 5, pp. 351-360.

Greenwood, M. and Yule, G.U. (1920). An Inquiry into the Nature of Frequency Distributions Representative of Multiple Happenings with Particular Reference to the Occurrence of Multiple Attacks of Disease or of Repeated Accidents. *Journal of the Royal Statistical Society.* Series A, 83, pp. 255-279.

Montmort, P.R. de (1713). *Essai d'analyse sur les jeux de hasard.* Seconde édition. Quillau, Paris.

Pascal, B. (1679). *Varia Opera Mathematica.* D. Petri de Fermat. Tolosae.

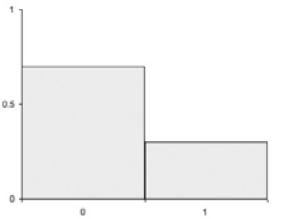
LOI DE BERNOULLI

Bernoulli distribution

Une **variable aléatoire** X suit une loi de Bernoulli de **paramètre** p si sa **fonction de probabilité** est de la forme :

$$P(X=x) = \begin{cases} p & \text{pour } x = 1 \\ q = 1-p & \text{pour } x = 0. \end{cases}$$

où p et q représentent respectivement les probabilités de « succès » et d'« échec » symbolisés par les valeurs 1 et 0.



Loi de Bernoulli, $p = 0,3$*,* $q = 0,7$*.*

La loi de Bernoulli est une **loi de probabilité discrète**.

HISTORIQUE

Voir **loi binomiale**.

ASPECTS MATHÉMATIQUES

L'**espérance mathématique** de la loi de Bernoulli est par définition :

$$\begin{aligned} E[X] &= \sum_{x=0}^{1} x \cdot P(X=x) \\ &= 1 \cdot p + 0 \cdot q \\ &= p. \end{aligned}$$

La **variance** de la loi de Bernoulli est par définition :

$$\begin{aligned} Var(X) &= E\left[X^2\right] - (E[X])^2 \\ &= 1^2 \cdot p + 0^2 \cdot q - p^2 \\ &= p - p^2 \\ &= pq. \end{aligned}$$

DOMAINES ET LIMITATIONS

La loi de Bernoulli est utilisée lorsqu'une **expérience aléatoire** n'a que deux résultats possibles : le « succès » et l'« échec ». On symbolise généralement ces deux résultats par 0 et 1.

MOTS CLÉS

Loi binomiale *(Binomial distribution)*
Loi de probabilité discrète *(Discrete probability distribution)*

LOI DE CAUCHY

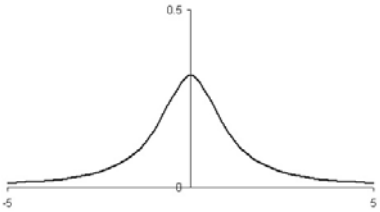
Cauchy distribution

Une **variable aléatoire** X suit une loi de Cauchy si sa **fonction de densité** est de la forme :

$$f(x) = \frac{1}{\pi\theta} \cdot \left[1 + \left(\frac{x-\alpha}{\theta}\right)^2\right]^{-1}, \quad \theta > 0.$$

Les paramètres α et θ sont des paramètres de tendance centrale et de dispersion respectivement.

La distribution de Cauchy est symétrique en $x = \alpha$ qui représente la **médiane**. Le premier et le troisième **quartile** sont donnés par $\alpha \pm \theta$.



Loi de Cauchy, $\theta = 1$, $\alpha = 0$

La loi de Cauchy est une **loi de probabilité continue**.

ASPECTS MATHÉMATIQUES

L'**espérance mathématique** $E[X]$ et la **variance** $Var(X)$ n'existent pas.

Si $\alpha = 0$ et $\theta = 1$, la loi de Cauchy est identique à la **loi de Student** avec 1 **degré de liberté**.

DOMAINES ET LIMITATIONS

La loi de Cauchy peut, par exemple, être utilisée pour décrire la distribution des points d'impact de particules émises en faisceau, à partir d'un point source.

MOTS CLÉS

Loi de probabilité continue *(Continuous probability distribution)*
Loi de Student *(Student distribution)*

RÉFÉRENCE

Cauchy, A.L. (1853). Sur les résultats moyens d'observations de même nature, et sur les résultats les plus probables. *Comptes rendus des séances de l'Académie des sciences*, Paris, 37, pp. 198-206.

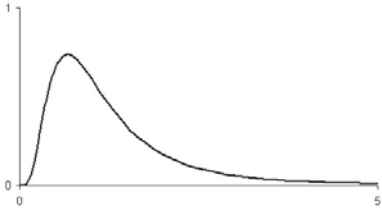
LOI DE FISHER

Fisher distribution

La **variable aléatoire** X suit une loi de Fisher si sa **fonction de densité** est de la forme :

$$f(x) = \frac{\Gamma\left(\frac{m+n}{2}\right)}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)} \left(\frac{m}{n}\right)^{\frac{m}{2}} \cdot \frac{x^{\frac{m-2}{2}}}{\left(1+\frac{m}{n}x\right)^{\frac{m+n}{2}}}$$

où $x \geq 0$ et Γ est la fonction gamma (voir **loi gamma**) et m, n sont les **degrés de liberté** $(m, n = 1, 2, \ldots)$.



Loi de Fisher, $m = 12$, $n = 8$

La loi de Fisher est une **loi de probabilité continue**.

HISTORIQUE

La loi de Fisher a été découverte par **Ronald Aylmer Fisher** (1925). Le symbole F, utilisé pour dénoter la loi de Fisher, a été introduit par **George Waddel Snedecor** (1934), en hommage à son auteur.

ASPECTS MATHÉMATIQUES

Si U et V sont deux variables aléatoires indépendantes suivant chacune une **loi du chi-carré**, avec respectivement m et n degrés de libertés, alors la variable aléatoire :

$$F = \frac{U/m}{V/n}$$

suit une loi de Fisher avec m et n degrés de liberté.

L'**espérance mathématique** d'une variable aléatoire F suivant une loi de Fisher est donnée par :

$$E[F] = \frac{n}{n-2} \text{ pour } n > 2$$

et la **variance** est égale à :

$$Var(F) = \frac{2n^2(m+n-2)}{m(n-2)^2(n-4)} \quad n > 4.$$

La loi de Fisher avec 1 et v degrés de libertés est identique au carré de la **loi de Student** avec v degrés de libertés.

DOMAINES ET LIMITATIONS

L'importance de la loi de Fisher dans la théorie statistique est notamment liée à son application à la distribution de ratios d'estimateurs indépendants de la **variance**. L'application la plus courante de cette loi se trouve dans les tests standards associés à l'**analyse de variance** et à l'**analyse de régression**.

MOTS CLÉS

Analyse de variance *(Analysis of variance)*
Loi de probabilité continue *(Continuous probability distribution)*
Loi de Student *(Student distribution)*
Loi du chi-carré *(Chi-square distribution)*
Table de Fisher *(Fisher table)*
Test de Fisher *(Fisher test)*

RÉFÉRENCES

Fisher, R.A. (1925). *Statistical Methods for Research Workers.* Oliver & Boyd, Edinburgh.

Snedecor, G.W. (1934). *Calculation and interpretation of analysis of variance and covariance.* Collegiate Press, Inc. Ames, Iowa.

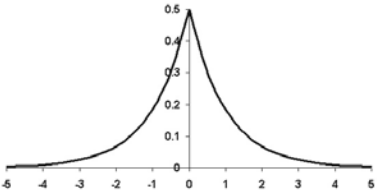
LOI DE LAPLACE

Laplace distribution

La **variable aléatoire** X suit une loi de Laplace si sa **fonction de densité** est de la forme :

$$f(x) = \frac{1}{2\Phi} \cdot \exp\left[\left(-\frac{|x-\theta|}{\Phi}\right)\right], \quad \Phi > 0$$

où Φ est un **paramètre** de dispersion et θ est l'**espérance mathématique**.



Loi de Laplace, $\theta = 0$, $\Phi = 1$

La loi de Laplace est symétrique autour de son espérance mathématique θ, qui correspond également au **mode** et à la **médiane** de la distribution. La loi de Laplace est une **loi de probabilité continue**.

HISTORIQUE

Cette **loi de probabilité continue** porte le nom de son auteur. En 1774, **Pierre Simon de Laplace** écrivit un article fondamental sur les distributions symétriques afin d'en décrire les erreurs de mesure.

ASPECTS MATHÉMATIQUES

L'**espérance mathématique** d'une **variable aléatoire** X suivant une loi de Laplace est donnée par :

$$E[X] = \theta$$

et la **variance** est égale à :

$$Var(X) = 2 \cdot \Phi^2.$$

La forme standard de la **fonction de densité** de la loi de Laplace s'écrit :

$$f(x) = \frac{1}{\sigma\sqrt{2}} \cdot \exp\left[\left(-\sqrt{2} \cdot \frac{|x-\mu|}{\sigma}\right)\right], \quad \sigma > 0$$

$$\text{où } \mu = \theta \text{ et } \sigma = \sqrt{2} \cdot \Phi.$$

Elle permet d'écrire la **fonction de répartition** de la loi de Laplace donnée par :

$$F(x) = \begin{cases} \frac{1}{2} \exp\left[\left(\frac{x-\theta}{\Phi}\right)\right] & \text{si } x \leq \theta \\ 1 - \frac{1}{2} \exp\left[\left(-\frac{x-\theta}{\Phi}\right)\right] & \text{si } x \geq \theta. \end{cases}$$

MOT CLÉ

Loi de probabilité continue *(Continuous probability distribution)*

RÉFÉRENCE

Laplace, P.S. de (1774). *Mémoire sur la probabilité des causes par les événements.* Mémoire de l'Académie royale des sciences présenté par divers savants, 6, pp. 621-656 ou Vol. 8, pp. 27-65. In P.S. de Laplace, *Œuvres complètes* (1891). Gauthier-Villars, Paris.

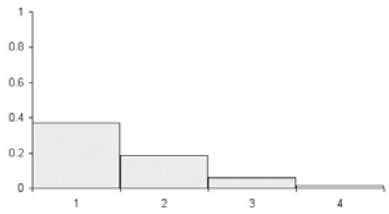
LOI DE POISSON

Poisson distribution

La loi de Poisson est une **loi de probabilité discrète**. Elle convient particulièrement pour des phénomènes de comptage dans une unité de temps ou d'espace.

La **variable aléatoire** X correspondant au nombre de phénomènes observés par unité de temps ou d'espace, suit une loi de Poisson de paramètre θ, notée $P(\theta)$, si sa **fonction de probabilité** est :

$$P(X = x) = \frac{\exp(-\theta) \cdot \theta^x}{x!}.$$



Loi de Poisson, $\theta = 1$

HISTORIQUE

C'est en 1837 que **Siméon Denis Poisson** a publié la distribution qui porte son nom. Il a approché cette distribution en s'intéressant aux limites de la **loi binomiale**.

ASPECTS MATHÉMATIQUES

L'**espérance mathématique** de la loi de Poisson est par définition :

$$\begin{aligned} E[X] &= \sum_{x=0}^{\infty} x \cdot P(X = x) \\ &= \sum_{x=0}^{\infty} x \cdot \frac{\exp(-\theta)\,\theta^x}{x!} \\ &= \sum_{x=1}^{\infty} x \cdot \frac{\exp(-\theta)\,\theta^x}{x!} \\ &= \theta \cdot \exp(-\theta) \cdot \sum_{x=1}^{\infty} \frac{\theta^{x-1}}{(x-1)!} \\ &= \theta \cdot \exp(-\theta) \cdot \exp(\theta) \\ &= \theta. \end{aligned}$$

La **variance** de X est égale à :

$$\begin{aligned} Var(X) &= E\left[X^2\right] - (E[X])^2 \\ &= E[X(X-1) + X] - (E[X])^2 \\ &= E[X(X-1)] + E[X] - (E[X])^2. \end{aligned}$$

Comme

$$E[X(X-1)] = \sum_{x=0}^{\infty} x(x-1) \frac{\exp(-\theta) \cdot \theta^x}{x!}$$

$$
\begin{aligned}
&= \sum_{x=2}^{\infty} x\,(x-1)\,\frac{\exp(-\theta)\cdot\theta^x}{x!} \\
&= \exp(-\theta)\cdot\theta^2\cdot\sum_{x=2}^{\infty}\frac{\theta^{x-2}}{(x-2)!} \\
&= \exp(-\theta)\cdot\theta^2\cdot\exp(\theta) \\
&= \theta^2,
\end{aligned}
$$

on obtient :

$$
\begin{aligned}
Var(X) &= \theta^2+\theta-\theta^2 \\
&= \theta.
\end{aligned}
$$

DOMAINES ET LIMITATIONS

La loi de Poisson est utilisée dans les cas suivants :

- nombre de particules émises par une substance radioactive ;
- nombre d'appels téléphoniques enregistrés par un central ;
- nombre d'accidents survenus à un assuré ;
- nombre d'arrivées à un guichet ;
- nombre de bactéries dans une préparation microscopique ;
- nombre de plantes ou d'animaux dans une surface déterminée dans la nature par l'observateur ;
- etc.

Lorsque θ tend vers l'infini, on peut faire une approximation de la loi de Poisson par la **loi normale** de **moyenne** θ et de **variance** θ.

EXEMPLE

Une secrétaire fait en moyenne deux fautes par page. Quelle est la **probabilité** qu'il y ait trois fautes sur une page ?

Le nombre de fautes par page, X, suit une loi de Poisson de **paramètre** $\theta = 2$ et sa **fonction de probabilité** est alors :

$$P(X=x) = \frac{\exp(-\theta)\cdot\theta^x}{x!} = \frac{\exp(-2)\cdot 2^x}{x!}.$$

La probabilité qu'il y ait trois fautes sur une page est égale à :

$$
\begin{aligned}
P(X=3) &= \frac{\exp(-2)\cdot 2^3}{3!} \\
&= 0,1804.
\end{aligned}
$$

MOTS CLÉS

Loi binomiale *(Binomial distribution)*
Loi de probabilité discrète *(Discrete probability distribution)*
Loi normale *(Normal distribution)*

RÉFÉRENCE

Poisson, S.D. (1837). *Recherches sur la probabilité des jugements en matière criminelle et en matière civile. Procédés des règles générales du calcul des probabilités.* Bachelier, Imprimeur-Libraire pour les Mathématiques, Paris.

LOI DE PROBABILITÉ

Probability distribution

Une loi de probabilité est un **modèle** représentant au mieux une **distribution de fréquences** d'une **variable aléatoire**.

Dans le cas d'une **variable** continue, cette distribution est continue.

Dans le cas d'une variable discrète, on obtient la loi de probabilité (ou distribution de probabilité) d'une variable aléatoire en associant sa **probabilité** à chacune des valeurs possibles de cette variable.

HISTORIQUE

C'est au XVII^{e} siècle que commença l'étude systématique des problèmes liés aux phénomènes aléatoires. L'éminent physicien Galilée essaya déjà d'étudier les erreurs des mesures physiques, considérant celles-ci comme aléatoires et estimant leur **probabilité**. À cette époque apparut aussi la théorie des assurances, basée sur l'analyse des lois régissant des phénomènes aléatoires tels que la morbidité, la mortalité, les accidents, etc.

Cependant, les statisticiens de l'époque étudièrent tout d'abord des phénomènes plus simples, tels que les jeux de hasard. Ceux-ci encore aujourd'hui fournissent des modèles particulièrement simples et clairs de phénomènes aléatoires, permettant d'observer et d'étudier les lois spécifiques qui les régissent ; de plus, la possibilité de répéter de nombreuses fois une même **expérience** assure une vérification expérimentale de ces lois.

MOTS CLÉS

Fonction de densité *(Density function)*
Fonction de probabilité *(Probability function)*
Fonction de répartition *(Distribution function)*
Loi de probabilité continue *(Continuous probability distribution)*
Loi de probabilité discrète *(Discrete probability distribution)*
Probabilité *(Probability)*
Variable aléatoire *(Random variable)*

RÉFÉRENCES

Johnson, N.L. and Kotz, S. (1969). *Distributions in Statistics: discrete distributions.* John Wiley & Sons, New York.

Johnson, N.L. and Kotz, S. (1970). *Distributions in Statistics: continuous univariate distributions.* Vol. 1, 2. John Wiley & Sons, New York.

Rothschild, V. and Logothetis, N. (1986). *Probability Distributions.* John Wiley & Sons, New York.

LOI DE PROBABILITÉ CONTINUE

Continuous probability distribution

À toute **variable** peut être associée une **distribution de fréquences**. Dans le cas d'une variable continue, cette distribution est continue.

Une loi de probabilité continue, aussi appelée distribution continue, est un **modèle** représentant au mieux une distribution de fréquences d'une variable continue.

ASPECTS MATHÉMATIQUES

La **loi de probabilité** d'une **variable aléatoire** continue X est donnée par sa **fonction de densité** $f(x)$ ou sa **fonction de répartition** $F(x)$.

Elle peut généralement être caractérisée par son **espérance mathématique** :

$$E[X] = \int_D x \cdot f(x)\, dx = \mu$$

et sa **variance**

$$\begin{aligned} Var(X) &= \int_D (x-\mu)^2 \cdot f(x)\, dx \\ &= E\left[(X-\mu)^2\right] \\ &= E\left[X^2\right] - E[X]^2, \end{aligned}$$

où D représente l'**intervalle** sur lequel X prend ses valeurs.

Une propriété essentielle d'une variable aléatoire continue réside dans le fait que la **probabilité** qu'elle prenne une valeur numérique spécifique est nulle, alors que la probabilité qu'elle prenne une valeur sur un intervalle (fini ou infini) est en général non nulle.

DOMAINES ET LIMITATIONS

La loi de probabilité continue la plus connue est la **loi normale**.

Les lois de probabilité continues, ou distributions continues, sont souvent utilisées comme approximation des lois de probabilité discrètes, autant pour la construction du modèle que pour l'application des techniques statistiques.

MOTS CLÉS

Espérance mathématique *(Expected value)*
Fonction de densité *(Density function)*
Fonction de répartition continue *(Continuous distribution function)*
Loi bêta *(Beta distribution)*
Loi de Cauchy *(Cauchy distribution)*
Loi de Fisher *(Fisher distribution)*
Loi de Laplace *(Laplace distribution)*

Loi de probabilité *(Probability distribution)*
Loi de probabilité discrète *(Discrete probability distribution)*
Loi de Student *(Student distribution)*
Loi du chi-carré *(Chi-square distribution)*
Loi exponentielle *(Exponential distribution)*
Loi gamma *(Gamma distribution)*
Loi lognormale *(Lognormal distribution)*
Loi normale *(Normal distribution)*
Loi uniforme *(Uniform distribution)*
Probabilité *(Probability)*
Variable aléatoire *(Random variable)*
Variance d'une variable aléatoire *(Variance of a random variable)*

RÉFÉRENCE

Johnson, N.L. and Kotz, S. (1970). *Distributions in Statistics: continuous univariate distributions.* Vol. 1, 2. John Wiley & Sons, New York.

LOI DE PROBABILITÉ DISCRÈTE

Discrete probability distribution

Si on associe sa **probabilité** à chacune des valeurs possibles d'une **variable aléatoire** discrète, on obtient la loi de probabilité discrète de cette variable aléatoire.

ASPECTS MATHÉMATIQUES

La loi de probabilité d'une **variable aléatoire** discrète X est donnée par sa **fonction de probabilité** $P(x)$ ou sa **fonction de répartition** $F(x)$.

Elle peut généralement être caractérisée par son **espérance mathématique** :

$$E[X] = \sum_D x \cdot P(X = x)$$

et sa **variance** :

$$Var(X) = \sum_D (x - E[X])^2 \cdot P(X = x)$$

où D représente l'**intervalle** sur lequel X peut prendre ses **valeurs**.

Les lois de probabilité discrète les plus utilisées sont la **loi de Bernoulli**, la **loi binomiale**, la **loi binomiale négative**, la **loi géométrique**, la **loi multinomiale**, la **loi hypergéométrique** et la **loi de Poisson**.

MOTS CLÉS

Espérance mathématique *(Expected value)*
Fonction de probabilité *(Probability function)*
Fonction de répartition discrète *(Discrete distribution function)*
Loi binomiale *(Binomial distribution)*
Loi binomiale négative *(Negative binomial distribution)*
Loi de Bernoulli *(Bernoulli distribution)*
Loi de Poisson *(Poisson distribution)*
Loi de probabilité *(Probability distribution)*
Loi de probabilité continue *(Continuous probability distribution)*
Loi géométrique *(Geometric distribution)*
Loi hypergéométrique *(Hypergeometric distribution)*
Loi multinomiale *(Multinomial distribution)*
Probabilité *(Probability)*
Variable aléatoire *(Random variable)*
Variance d'une variable aléatoire *(Variance of a random variable)*

RÉFÉRENCE

Johnson, N.L. and Kotz, S. (1969). *Distributions in Statistics: discrete distributions.* John Wiley & Sons, New York.

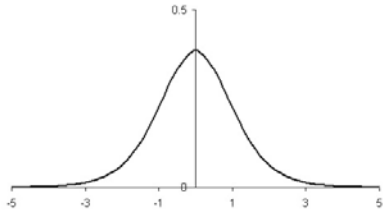
LOI DE STUDENT

Student distribution

La **variable aléatoire** T suit une loi de Student si sa **fonction de densité** est de la forme :

$$f(t) = \frac{\Gamma\left(\frac{v+1}{2}\right)}{\sqrt{v\pi} \cdot \Gamma\left(\frac{v}{2}\right)} \left(1 + \frac{t^2}{v}\right)^{-\frac{v+1}{2}}$$

où Γ est la fonction gamma (voir **loi gamma**) et v le nombre de degrés de liberté.



Loi de Student, $\nu = 7$

La loi de Student est une **loi de probabilité continue**.

HISTORIQUE

La loi de Student est due à **William Sealy Gosset**, dit « Student », qui publia en 1908 un article dans lequel il décrivit la **fonction de densité** de la différence entre la **moyenne** d'un **échantillon** et la moyenne de la **population**, divisée par l'**écart type** de l'échantillon. Il fournit également dans le même article une première table de la **fonction de répartition** correspondante.

Il poursuivit ses travaux et publia en 1925 un article dans lequel il proposa de nouvelles tables plus étendues et plus précises.

Ronald Aylmer Fisher s'intéressa aux travaux de Gosset. Il lui écrivit en 1912 pour lui proposer une démonstration géométrique de la loi de Student, et pour introduire la notion de degrés de liberté. Il publia notamment en 1925 un article dans lequel il définit le ratio t :

$$t = \frac{Z}{\sqrt{\frac{X}{v}}}$$

où X et Z sont deux variables aléatoires indépendantes, X est distribuée selon une **loi du chi-carré** avec v degrés de liberté et Z est distribuée selon la **loi normale** centrée réduite. Ce ratio est appelé t de Student.

ASPECTS MATHÉMATIQUES

Si X est une **variable aléatoire** qui suit une **loi du chi-carré** avec v degrés de liberté, et Z est une variable aléatoire distribuée selon la **loi normale** centrée réduite, alors la variable aléatoire :

$$T = \frac{Z}{\sqrt{\frac{X}{v}}},$$

suit une loi de Student avec v degrés de liberté, si X et Z sont indépendantes.

La variable aléatoire T :

$$T = \frac{\sqrt{n}\,(\bar{x} - \mu)}{S},$$

où

$\bar{x}$ est la **moyenne** d'un **échantillon**,
μ la moyenne de la **population**, et
S l'**écart type** de l'échantillon,

$$S = \sqrt{\frac{1}{n-1} \cdot \sum_{i=1}^{n} (x_i - \bar{x})^2},$$

suit donc une loi de Student avec $n - 1$ degrés de liberté.

L'**espérance mathématique** d'une variable aléatoire T suivant une loi de Student est donnée par :

$$E[T] = 0, \quad v > 1$$

et la **variance** est égale à :

$$Var(T) = \frac{v}{v-2}, \quad v > 2,$$

v étant le nombre de degrés de liberté.

Le loi de Student est liée à d'autres lois de probabilité continues :

- lorsque le nombre de degrés de liberté est suffisamment grand, la loi de Student s'approche d'une loi normale ;
- quand le nombre de degrés de liberté v est égal à 1, la loi de Student est identique à une **loi de Cauchy** avec $\alpha = 0$ et $\theta = 1$;

- si la variable aléatoire X suit une loi de Student avec v degrés de liberté, la variable aléatoire X^2 suit une **loi de Fisher** avec 1 et v degrés de liberté.

DOMAINES ET LIMITATIONS

La loi de Student est une distribution symétrique de **moyenne** égale à 0 dont l'**écart type** est d'autant plus petit que le nombre de degrés de liberté augmente.

La loi de Student est utilisée en **statistique inférentielle** en relation avec le ratio t de Student dans les tests d'hypothèse et dans la construction d'intervalles de confiance pour la **moyenne** de la **population**.

Dans les tests d'**analyse de variance**, la loi de Student peut être utilisée lorsque la somme de carrés entre les groupes (ou la somme des carrés des facteurs) à comparer n'a qu'un degré de liberté.

La **loi normale** peut être utilisée comme approximation de la loi de Student quand le nombre d'observations n est grand ($n > 30$ selon la plupart des auteurs).

MOTS CLÉS

Loi de Cauchy *(Cauchy distribution)*
Loi de Fisher *(Fisher distribution)*
Loi de probabilité continue *(Continuous probability distribution)*
Loi normale *(Normal distribution)*
Table de Student *(Student table)*
Test de Student *(Student test)*

RÉFÉRENCES

Fisher, R.A. (1925). Applications of "Student's" Distribution. *Metron*, 5, pp. 90-104.

Gosset, S.W. « Student » (1908). The Probable Error of a Mean. *Biometrika*, 6, pp. 1-25.

Gosset, S.W. « Student » (1925). New tables for testing the significance of observations. *Metron*, 5, pp. 105-120.

LOI DES GRANDS NOMBRES

Law of large numbers

La loi des grands nombres établit que la **moyenne** d'une somme de variables aléatoires indépendantes et identiquement distribuées converge vers l'**espérance mathématique** de cette distribution commune lorsque la **taille d'échantillon** tend vers l'infini.

HISTORIQUE

La loi des grands nombres fut établie pour la première fois par **Jakob Bernoulli** dans son ouvrage intitulé *Ars Conjectandi*, publié en 1713, dans le cadre d'une définition empirique de la **probabilité**. Il exprima que la **fréquence** relative d'un **événement** converge, lors de la répétition d'épreuves identiques (**loi de Bernoulli**), vers un nombre qui constitue sa probabilité.

L'inégalité de Bienaymé-Tchebychev a été découverte initialement par **Irénée-Jules Bienaymé** (1853), avant que **Pafnouti Lvovitch Tchebychev** (1867) ne la découvre de manière tout à fait indépendante quelques années plus tard.

La version générale de la loi des grands nombres est attribuée au mathématicien russe Aleksandr Yakovlevich Khintchin (1894-1959).

ASPECTS MATHÉMATIQUES

Présentons d'abord l'inégalité de Bienaymé-Tchebychev :

Soit X une **variable aléatoire** d'**espérance mathématique** μ et de **variance** σ^2 finies. Pour tout $\varepsilon > 0$ réel, nous avons :

$$P\left(|X - \mu| \geq \varepsilon\right) \leq \frac{\sigma^2}{\varepsilon^2}$$

ce qui signifie que la **probabilité** qu'une variable aléatoire X diffère de plus de ϵ de son espérance mathématique ne peut pas excéder $\frac{\sigma^2}{\varepsilon^2}$.

L'importance de cette inégalité réside en ce qu'elle permet de borner la **valeur** de certaines probabilités là où seule l'espérance mathématique de la distribution est connue, plus éventuellement sa variance. Il est évident que si la distribution elle-même est connue, nous ne recourrons pas à des bornes, puisque la valeur exacte de ces probabilités est calculable.

Par cette inégalité, nous pouvons calculer la convergence de la **moyenne** observée d'une variable aléatoire vers son espérance mathématique.

Soient n variables aléatoires indépendantes et identiquement distribuées $X_1, X_2, \ldots, X_n$ suivant une **loi de probabilité** quelconque d'espérance mathématique μ et de variance σ^2, alors pour tout $\varepsilon > 0$, la loi des grands nombres est donnée par :

$$P\left(\left|\frac{X_1 + X_2 + \ldots + X_n}{n} - \mu\right| > \varepsilon\right) \overset{n\to\infty}{\longrightarrow} 0.$$

DOMAINES ET LIMITATIONS

La loi des grands nombres laisse entendre qu'il subsiste malgré tout une **probabilité**, même si elle est très faible, d'écart substantiel entre la **moyenne** empirique d'une suite d'événements et son **espérance mathématique**. C'est pour cette raison que nous parlons de loi faible des grands nombres.

En fait, nous pourrions démontrer qu'il n'en est rien, que la convergence est quasi certaine. Ce résultat plus fort que le précédent constitue la loi forte des grands nombres et s'écrit :

$$P\left(\lim_{n\to\infty} \bar{X} = \mu\right) = 1.$$

EXEMPLE

La loi des grands nombres nous dit qu'il est toujours possible de trouver une **valeur** pour n telle que la **probabilité** que $\bar{X}$ soit inclus dans un **intervalle** de $\mu \pm \varepsilon$ soit aussi grande que nous le désirons.

Prenons comme exemple une **loi de probabilité** ayant comme **variance** $\sigma^2 = 1$. En choisissant un intervalle $\varepsilon = 0,5$ et une probabilité de $0,05$, nous pouvons poser :

$$P\left(\left|\bar{X} - \mu\right| \geq 0,5\right) = 0,05.$$

Par l'inégalité de Tchebychev, nous savons que

$$P\left(\left|\bar{X} - \mu\right| \geq \varepsilon\right) \leq \frac{\sigma^2}{n\varepsilon^2}.$$

Nous pouvons donc poser l'inégalité suivante :

$$0,05 \leq \frac{1}{n \cdot 0,25}.$$

D'où :

$$\begin{aligned} 0,05 \cdot 0,25 &\leq 1/n \\ n &\geq 80 \end{aligned}$$

MOTS CLÉS

Théorème central limite *(Central limit theorem)*
Théorème de convergence *(Convergence theorem)*

RÉFÉRENCES

Bernoulli, J. (1713). *Ars Conjectandi, Opus Posthumum. Accedit Tractatus de Seriebus infinitis, et Epistola Gallice scripta de ludo Pilæ recticularis.* Basileæ, Impensis Thurnisiorum, fratrum.

Bienaymé, I.J. (1853). Considérations à l'appui de la découverte de Laplace sur la loi de probabilité dans la méthode des moindres carrés. *Comptes rendus de l'Académie des sciences*, Paris 37, pp. 5-13 ; réédité en 1867 dans le journal de Liouville précédant la preuve des inégalités de Bienaymé-Tchebyshev, *Journal de mathématiques pures et appliquées* 12, pp. 158-176.

Tchebychev, P.L. (1867). Des valeurs moyennes. *Journal de mathématiques pures et appliquées* (2) 12, pp. 177-184. Publié simultanément en russe, dans *Mat. Sbornik*, (2) 2, pp. 1-9.

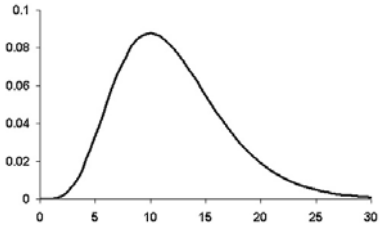
LOI DU CHI-CARRÉ

Chi-square distribution

Une **variable aléatoire** X suit une loi du chi-carré à n degrés de liberté si sa **fonction de densité** est :

$$f(x) = \frac{x^{\frac{n}{2}-1} \exp\left(-\frac{x}{2}\right)}{2^{\frac{n}{2}} \Gamma\left(\frac{n}{2}\right)}, \quad x \geq 0$$

où Γ est la fonction gamma (voir **loi gamma**).



Loi du χ^2, $\nu = 12$

La loi du chi-carré est une **loi de probabilité continue**.

HISTORIQUE

La loi de distribution du chi-carré est due, selon O.B. Sheynin (1977), à Abbe en 1863. John Frederick William Herschell semble être arrivé à cette distribution pour deux degrés de liberté en 1869, James Clerk Maxwell quelques années auparavant (1860) pour trois degrés de liberté, et Ludwig Boltzman en découvrit le cas général en 1881.

Parallèlement à ces travaux, il faut noter que selon H.O. Lancaster (1966), **Irénée-Jules Bienaymé** a obtenu en 1838 la loi du chi-carré en tant que limite de la **variable aléatoire** discrète :

$$\sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i},$$

si $(N_1, N_2, \ldots, N_k)$ suivent une distribution multinomiale conjointe de paramètres $n, p_1, p_2, \ldots, p_k$.

Ellis a démontré en 1844 que la somme de k variables aléatoires distribuées selon une loi du chi-carré avec 2 degrés de liberté suit une loi du chi-carré avec $2k$ degrés de liberté. Le résultat général a été démontré en 1852 par I.-J. Bienaymé.

Les travaux de **Karl Pearson** sont également très importants dans ce domaine. C'est en 1900 qu'il fait intervenir la loi du chi-carré comme approximation pour la **statistique** du chi-carré utilisée dans différents tests basés sur des tables de contingence.

ASPECTS MATHÉMATIQUES

La loi du chi-carré apparaît dans la théorie associée aux variables aléatoires distribuées selon une **loi normale**, en tant que distribution de la somme des carrés de variables aléatoires normales centrées réduites (de **moyenne** égale à 0 et de **variance** égale à 1).

Soient $Z_1, Z_2, \ldots, Z_n$, n variables aléatoires normales centrées réduites indépendantes. La somme de leurs carrés :

$$X = Z_1^2 + Z_2^2 + \ldots + Z_n^2 = \sum_{i=1}^{n} Z_i^2$$

est une variable aléatoire distribuée selon une loi du chi-carré avec n degrés de liberté.

L'**espérance mathématique** de la loi du chi-carré est donnée par :

$$E[X] = n.$$

La variance est égale à :

$$Var(X) = 2n.$$

La loi du chi-carré est liée à d'autres lois de probabilité continues :

- la loi du chi-carré est un cas particulier de la **loi gamma** ;
- si deux variables aléatoires X_1 et X_2 suivent une loi du chi-carré avec respectivement n_1 et n_2 degrés de liberté, la variable aléatoire

$$Y = \frac{X_1/n_1}{X_2/n_2}$$

suit une **loi de Fisher** avec n_1 et n_2 degrés de liberté ;

- lorsque le nombre de degrés de liberté n tend vers l'infini, la loi du chi-carré tend (relativement lentement) vers une loi normale.

DOMAINES ET LIMITATIONS

La loi du chi-carré est utilisée dans de nombreux tests d'hypothèse, le plus important étant le **test d'adéquation du chi-carré** (en anglais, « goodness of fit »), qui est fondé sur la comparaison entre les fréquences observées et les fréquences hypothétiques de classes spécifiques.

Elle est aussi utilisée pour la comparaison entre la **variance** observée et la variance hypothétique d'échantillons normalement distribués, ainsi que pour tester l'**indépendance** de deux variables.

MOTS CLÉS

Loi de Fisher *(Fisher distribution)*
Loi de probabilité continue *(Continuous probability distribution)*
Loi gamma *(Gamma distribution)*
Loi normale *(Normal distribution)*
Table du chi-carré *(Chi-square table)*
Test d'adéquation du chi-carré *(Chi-square goodness of fit test)*
Test du chi-carré *(Chi-square test)*

RÉFÉRENCES

Lancaster, H.O. (1966). Forerunners of the Pearson chi-square. *Australian Journal of Statistics*, 8, pp. 117-126.

Pearson, K. (1900). 1948. On the Criterion, That a Given System of Deviations from the Probable in the Case of a Correlated System of Variables is such that it can be reasonably supposed to have arisen from random Sampling. pp. 339-357 in Karl Pearson, *Karl Pearson's Early Statistical Papers.* Cambridge Univ. Press. First published in the *Philosophical Magazine*, 5th Series, 50, pp. 157-175.

Sheynin, O.B. (1977). On the history of some statistical laws of distribution. *Studies in the History of Statistics and Probability.* Vol. II. Edited by M. Kendall & R.L. Plackett. Charles Griffin & Company Ltd.

LOI EXPONENTIELLE

Exponential distribution

Une **variable aléatoire** X suit une loi exponentielle de **paramètre** θ ($\theta > 0$) si sa **fonction de densité** est donnée par :

$$f(x) = \begin{cases} \theta \cdot \exp(-\theta x) & \text{si } x \geq 0, \\ 0 & \text{sinon.} \end{cases}$$



Loi exponentielle, $\theta = 1$, $\sigma = 2$

La loi exponentielle est appelée aussi exponentielle négative.

La loi exponentielle est une **loi de probabilité continue**.

ASPECTS MATHÉMATIQUES

La **fonction de répartition** d'une **variable aléatoire** X suivant une loi exponentielle de **paramètre** θ est la suivante :

$$\begin{aligned} F(x) &= \int_0^x \theta \cdot \exp(-\theta t)\, dt \\ &= 1 - \exp(-\theta x). \end{aligned}$$

L'**espérance mathématique** est donnée par :

$$E[X] = \frac{1}{\theta}.$$

Comme nous avons

$$E\left[X^2\right] = \frac{2}{\theta^2},$$

la **variance** est égale à :

$$Var\,(X) = E\left[X^2\right] - (E\,[X])^2 = \frac{2}{\theta^2} - \frac{1}{\theta^2} = \frac{1}{\theta^2}.$$

La loi exponentielle est la **loi de probabilité continue** analogue à la **loi géométrique**.

Elle est un cas particulier de la **loi gamma**, quand le paramètre α de la loi gamma est égal à 1.

Elle devient une **loi du chi-carré** avec deux degrés de liberté quand le paramètre θ de la loi exponentielle est égal à $\frac{1}{2}$.

DOMAINES ET LIMITATIONS

La loi exponentielle est notamment utilisée pour décrire des **événements** aléatoires dans le temps. Par exemple, la durée de vie est une des caractéristiques très fréquemment représentées par une **variable aléatoire** exponentielle.

MOTS CLÉS

Loi de probabilité continue *(Continuous probability distribution)*
Loi du chi-carré *(Chi-square distribution)*
Loi gamma *(Gamma distribution)*
Loi géométrique *(Geometric distribution)*

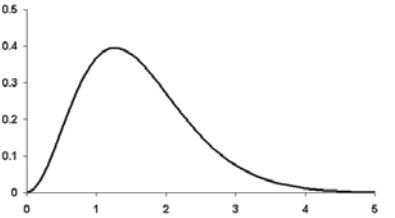
LOI GAMMA

Gamma distribution

Une **variable aléatoire** X suit une loi gamma de **paramètre** α si sa **fonction de densité** est donnée par

$$f\,(x) = \frac{\beta^\alpha}{\Gamma\,(\alpha)} x^{\alpha-1} \cdot \exp\,(-\beta x)\,,$$

avec $x > 0, \alpha > 0, \beta > 0$, où $\Gamma(\alpha)$ est la fonction gamma.



Loi gamma, $\alpha = 2$, $\beta = 1$

La forme standard de la loi gamma est obtenue en posant $\beta = 1$, ce qui donne

$$f\,(x) = \frac{1}{\Gamma\,(\alpha)} x^{\alpha-1} \cdot \exp\,(-x)\,, \quad \text{si } x > 0,\ \alpha > 0.$$

La fonction gamma (Γ) apparaît fréquemment en théorie statistique. Elle est définie de la manière suivante :

$$\Gamma\,(\alpha) = \int_0^\infty t^{\alpha-1} \exp\,(-t)\,dt.$$

Nous ne nous intéressons en **statistique** qu'à des valeurs de $\alpha > 0$.

En intégrant par parties, nous obtenons :

$$\Gamma(\alpha + 1) = \alpha\Gamma\,(\alpha)\,.$$

Comme

$$\Gamma\,(1) = 1$$

nous avons :

$$\Gamma\,(\alpha + 1) = \alpha! \quad \text{pour tout } \alpha \text{ entier positif.}$$

HISTORIQUE

Selon Lancaster (1966), cette **loi de probabilité continue** est attribuée à **Pierre Simon de Laplace** (1836).

ASPECTS MATHÉMATIQUES

L'**espérance mathématique** de la loi gamma est donnée par :

$$E\,[X] = \frac{1}{\beta} \cdot \frac{\Gamma\,(\alpha)}{\Gamma\,(\alpha+1)} = \frac{\alpha}{\beta}.$$

Comme

$$E\left[X^2\right] = \frac{1}{\beta^2} \cdot \frac{\Gamma(\alpha+2)}{\Gamma(\alpha)} = \frac{\alpha(\alpha+1)}{\beta^2},$$

la **variance** est égale à :

$$\begin{aligned} Var(X) &= E\left[X^2\right] - (E[X])^2 \\ &= \frac{\alpha(\alpha+1) - \alpha^2}{\beta^2} \\ &= \frac{\alpha}{\beta^2}. \end{aligned}$$

La **loi du chi-carré** est un cas particulier de la loi gamma, quand $\alpha = \frac{n}{2}$ et $\beta = \frac{1}{2}$, n étant le nombre de **degrés de liberté** de la loi du chi-carré.

La loi gamma de **paramètre** $\alpha = 1$ donne la **loi exponentielle** :

$$f(x) = \beta \cdot \exp(-\beta x).$$

MOTS CLÉS

Loi de probabilité continue *(Continuous probability distribution)*
Loi du chi-carré *(Chi-square distribution)*
Loi exponentielle *(Exponential distribution)*

RÉFÉRENCES

Lancaster, H.O. (1966). Forerunners of the Pearson chi-square. *Australian Journal of Statistics*, 8, pp. 117-126.

Laplace, P.S. de (1836). *Théorie analytique des probabilités*, supplément à la 3e éd. Courcier, Paris.

LOI GÉOMÉTRIQUE

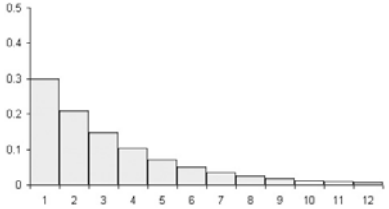
Geometric distribution

On considère un processus de la **loi de Bernoulli** avec une **probabilité** p de succès et une probabilité q d'échec.

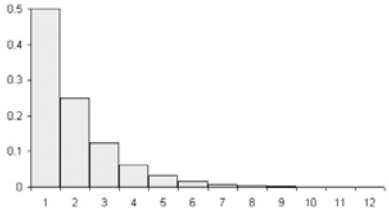
La loi géométrique peut aussi être exprimée sous cette forme :

$$P(X = k) = q^{k-1} \cdot p, \quad k = 1, 2, \ldots$$

où la **variable aléatoire** X représente le nombre d'épreuves nécessaires pour parvenir au premier succès (y compris celui-ci).



Loi géométrique, $p = 0,3$, $q = 0,7$.



Loi géométrique, $p = 0,5$, $q = 0,5$.

La loi géométrique est une **loi de probabilité discrète**.

On appelle cette loi « géométrique » parce que les termes successifs de la **fonction de probabilité** ci-dessus forment une progression géométrique de raison égale à q.

La loi géométrique est un cas particulier de la **loi binomiale négative** où $r = 1$, c'est-à-dire que le processus continue jusqu'au premier succès.

ASPECTS MATHÉMATIQUES

L'**espérance mathématique** de la loi géométrique est par définition égale à :

$$\begin{aligned} E[X] &= \sum_{x=1}^{\infty} x \cdot P(X = x) \\ &= \sum_{x=1}^{\infty} x \cdot q^{x-1} \cdot p \\ &= \frac{1}{p}. \end{aligned}$$

En effet :

$$\begin{aligned} E[X] &= \sum_{x=1}^{\infty} x \cdot q^{x-1} \cdot p \\ &= p \cdot \sum_{x=1}^{\infty} x \cdot q^{x-1} \\ &= 0 + p\left(1 + 2 \cdot q + 3 \cdot q^2 + \ldots\right). \end{aligned}$$

Or la somme $1 + 2 \cdot q + 3 \cdot q^2 + \ldots = \frac{1}{(1-q)^2}$, puisque la dérivée de la série géométrique suivante donne :

$$\begin{aligned} q + q^2 + q^3 + \ldots &= \frac{q}{1-q} \\ \left(q + q^2 + q^3 + \ldots\right)' &= \left(\frac{q}{1-q}\right)' \\ 1 + 2 \cdot q + 3 \cdot q^2 + \ldots &= \frac{1}{(1-q)^2}. \end{aligned}$$

La **variance** de la loi géométrique est par définition égale à :

$$\begin{aligned} Var(X) &= E\left[X^2\right] - (E[X])^2 \\ &= \sum_{x=1}^{\infty} x^2 \cdot P(X = x) - \left(\frac{1}{p}\right)^2 \\ &= \sum_{x=1}^{\infty} x^2 \cdot q^{x-1} \cdot p - \frac{1}{p^2} \\ &= p \cdot q\left(1 + 2^2 \cdot q + 3^2 q^2 + \ldots\right) \\ &\quad - \left(\frac{q}{p}\right)^2 \\ &= \frac{q}{p^2}. \end{aligned}$$

En effet, la somme $1 + 2^2 \cdot q + 3^2 q^2 + \ldots = \frac{1+q}{p^2}$, puisque :

$$\begin{aligned} \left(q\left(1 + 2 \cdot q + 3 \cdot q^2 + \ldots\right)\right)' &= \left(\frac{q}{(1-q)^2}\right)' \\ 1 + 2^2 \cdot q + 3^2 \cdot q^2 + \ldots &= \frac{1+q}{(1-q)^2} = \frac{1+q}{p^2}. \end{aligned}$$

DOMAINES ET LIMITATIONS

La loi géométrique est utilisée relativement souvent dans les modèles météorologiques. Elle est aussi utilisée dans les processus stochastiques et la théorie des files d'attente.

La **variable aléatoire** X correspondant au nombre d'échecs k qui se produit lorsque le processus est poursuivi jusqu'au premier succès, suit une loi géométrique de **paramètre** p et on a :

$$P(X = x) = p \cdot q^k.$$

EXEMPLE

On lance un dé non truqué et on veut connaître la **probabilité** qu'un « six » apparaisse pour la première fois au quatrième lancer.

On a donc :

- $p = \frac{1}{6}$ (probabilité qu'un « six » apparaisse),
- $q = \frac{5}{6}$ (probabilité qu'un « six » n'apparaisse pas),
- $k = 3$ (nombre d'échecs : puisque le quatrième lancer est un succès).

La probabilité cherchée est la suivante :

$$\begin{aligned} P(X = 3) &= \frac{1}{6} \cdot \left(\frac{5}{6}\right)^3 \\ &= 0,0965. \end{aligned}$$

MOTS CLÉS

Loi binomiale négative *(Negative binomial distribution)*
Loi de Bernoulli *(Bernoulli distribution)*
Loi de probabilité discrète *(Discrete probability distribution)*

LOI HYPERGÉOMÉTRIQUE

Hypergeometric distribution

La loi hypergéométrique décrit les probabilités des résultats de tirages sans remise.

Elle est utilisée lorsqu'une **expérience aléatoire** n'a que deux résultats possibles : le « succès » et l'« échec ».

Supposons un ensemble de N événements à l'intérieur duquel il y a M « succès » et $N-M$ « échecs ». La **variable aléatoire** X correspondant au nombre de « succès » obtenus dans n tirages sans remise suit une loi hypergéométrique de paramètres N, M et n et notée $H(n, M, n)$.

La loi hypergéométrique est une **loi de probabilité discrète**.

Le nombre de possibilités de faire n tirages parmi N événements est égal à :

$$C_N^n = \left(\begin{array}{c} N \\ n \end{array} \right) = \frac{n!}{N! \cdot (n-N)!}.$$

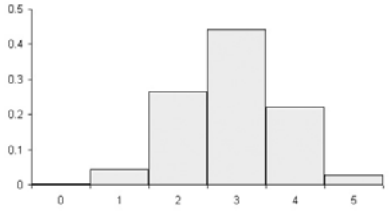
Le nombre d'événements élémentaires est dépendant de X et vaut :

$$C_M^x \cdot C_{N-M}^{n-x},$$

ce qui nous donne la **fonction de probabilité** suivante :

$$P(X = x) = \frac{C_M^x \cdot C_{N-M}^{n-x}}{C_N^n}, \quad \text{pour } x = 0, 1, \ldots, n$$

(convention : $C_v^u = 0$ si $u > v$).



Loi hypergéométrique, $N = 12$, $M = 7$, $n = 5$

ASPECTS MATHÉMATIQUES

Soit la **variable aléatoire** $X = X_1 + X_2 + \ldots + X_n$ où :

$$X_i = \left\{ \begin{array}{ll} 1 & \text{si le } i\text{-ème tirage est un succès} \\ 0 & \text{si le } i\text{-ème tirage est un échec.} \end{array} \right.$$

Nous avons la **loi de probabilité** pour X_i :

X_i	0	1
$P(X_i)$	$\frac{N-M}{N}$	$\frac{M}{N}$.

L'**espérance mathématique** de X_i est donc égale à :

$$\begin{aligned} E[X_i] &= \sum_{j=1}^{2} x_j P(X_i = x_j) \\ &= 0 \cdot \frac{N-M}{N} + 1 \cdot \frac{M}{N} \\ &= \frac{M}{N}. \end{aligned}$$

En utilisant le fait que $X = X_1 + X_2 + \ldots + X_n$, nous avons :

$$E[X] = \sum_{i=1}^{n} E[X_i] = \sum_{i=1}^{n} \frac{M}{N} = n\frac{M}{N}.$$

La **variance** de X_i est par définition :

$$\begin{aligned} Var(X_i) &= E\left[X_i^2\right] - (E[X_i])^2 \\ &= \frac{M}{N} - \left(\frac{M}{N}\right)^2 \\ &= \frac{M(n-M)}{N^2}. \end{aligned}$$

Comme les X_i, $i = 1, 2, \ldots, n$, sont des variables aléatoires dépendantes, il faut tenir compte de la **covariance** pour calculer la variance de X.

La **probabilité** que X_i et X_j $(i \neq j)$ soient tous les deux des succès est égale à :

$$P(X_i = 1, X_j = 1) = \frac{M(m-1)}{N(N-1)}.$$

Si nous posons $V = X_i \cdot X_j$, les valeurs de V et les probabilités associées sont :

V	0	1
$P(V)$	$1 - \frac{M(m-1)}{N(n-1)}$	$\frac{M(m-1)}{N(N-1)}$

L'espérance mathématique de V est donc :

$$E[V] = 0 \cdot \left(1 - \frac{M(m-1)}{N(N-1)}\right) + 1 \cdot \frac{M(m-1)}{N(N-1)}.$$

La **covariance** de X_i et X_j est par définition égale à :

$$Cov(X_i, X_j) = E[X_i \cdot X_j] - E[X_i] \cdot E[X_j]$$

$$= \frac{M(m-1)}{N(n-1)} - \left(\frac{M}{N}\right)^2$$
$$= -\frac{M(n-M)}{N^2(n-1)}$$
$$= -\frac{1}{N-1} Var(x_i).$$

Nous pouvons maintenant calculer la variance de X :

$$\begin{aligned} Var(X) &= \sum_{i=1}^{n} Var(X_i) + 2\sum_{j=1}^{n}\sum_{i<j} Cov(X_i, X_j) \\ &= \sum_{i=1}^{n} Var(X_i) + n(n-1) Cov(X_i, X_j) \\ &= n\left[Var(X_i) - \frac{n-1}{N-1} Var(X_i)\right] \\ &= nVar(X_i)\frac{N-n}{N-1} \\ &= n\frac{M(n-M)}{N^2}\frac{N-n}{N-1} \\ &= \frac{N-n}{N-1} n \frac{M}{N}\left(1-\frac{M}{N}\right). \end{aligned}$$

DOMAINES ET LIMITATIONS

La loi hypergéométrique est souvent utilisée pour le contrôle de qualité.

Supposons qu'une production de taille N soit soumise à une vérification. Nous tirons d'abord un **échantillon** de taille n puis notons le nombre de produits défectueux. Il est alors possible de donner, par **inférence**, une information sur la grandeur probable du nombre de produits défectueux dans la production totale.

EXEMPLE

Une boîte contient 30 fusibles dont 12 sont défectueux. En prenant au hasard 5 fusibles, la **probabilité** qu'aucun d'entre eux ne soit défectueux est égale à :

$$\begin{aligned} P(X=0) &= \frac{C_M^x C_{N-M}^{n-x}}{C_N^n} \\ &= \frac{C_{12}^0 C_{18}^5}{C_{30}^5} \\ &= 0,0601. \end{aligned}$$

MOTS CLÉS

Loi binomiale *(Binomial distribution)*
Loi de Bernoulli *(Bernoulli distribution)*
Loi de probabilité discrète *(Discrete probability distribution)*

LOI LOGNORMALE

Lognormal distribution

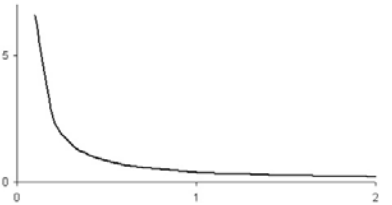
Une **variable aléatoire** Y suit une loi lognormale si les valeurs de Y sont fonction de celles de X selon l'équation :

$$y = \exp(x)$$

où la variable aléatoire X suit une **loi normale**.

La **fonction de densité** de la loi lognormale est donnée par :

$$f(y) = \frac{1}{\sigma y\sqrt{2\pi}} \exp\left(\left[\frac{-(\log y - \mu)^2}{2\sigma^2}\right]\right).$$



Loi lognormale, $\mu = 0$, $\sigma = 1$

La loi lognormale est une **loi de probabilité continue**.

HISTORIQUE

La loi lognormale est due aux travaux de **Francis Galton** (1879) et D. McAlister (1879) qui obtinrent des expressions pour la **moyenne**, la **médiane**, le **mode**, la **variance** et certains quantiles de la distribution résultante. Francis Galton était parti de n variables aléatoires indépendantes positives et en avait construit le produit à l'aide du logarithme ; il passa du produit à une somme de nouvelles variables aléatoires.

J.C. Kapteyn (1903) reconsidéra la construction d'une variable aléatoire suivant une loi lognormale et en compagnie de M.J. van Uven (1916), ils développèrent une méthode graphique pour l'estimation de paramètres. Dès lors, de nombreuses publications ont été faites relativement à la loi lognormale ; une liste exhaustive est citée par N.L. Johnson et S. Kotz (1970).

ASPECTS MATHÉMATIQUES

L'**espérance mathématique** de la loi lognormale est donnée par :

$$E[Y] = \exp\left(\mu + \frac{1}{2\sigma^2}\right).$$

La **variance** est égale à :

$$Var(X) = \exp(2\mu) \cdot \left[\exp\left(2\sigma^2\right) - \exp\left(\sigma^2\right)\right].$$

Si la **variable aléatoire** Y suit une loi lognormale, la variable aléatoire :

$$X = \ln Y$$

suit une **loi normale**.

DOMAINES ET LIMITATIONS

La loi lognormale est très largement appliquée dans les travaux statistiques pratiques. Par exemple, les dosages critiques dans des applications pharmaceutiques, la durée d'une consultation chez le médecin, la répartition des forces économiques suivent une loi lognormale.

La loi lognormale peut servir d'approximation de la **loi normale**. Plus σ est petit et plus la loi lognormale s'approche de la loi normale.

MOTS CLÉS

Loi de probabilité continue *(Continuous probability distribution)*
Loi normale *(Normal distribution)*

RÉFÉRENCES

Galton, F. (1879). The geometric mean in vital and social statistics. *Proceedings of the Royal Society of London.* 29, pp. 365-367.

Johnson, N.L. and Kotz, S. (1970). *Distributions in Statistics: continuous univariate distributions.* Vol. 1, 2. John Wiley & Sons, New York.

Kapteyn, J. (1903). *Skew frequency curves in biology and statistics.* Astronomical Laboratory Noordhoff, Groningen.

Kapteyn, J. and van Uven, M.J. (1916). *Skew frequency curves in biology and statistics.* Hoitsema Brothers, Groningen.

McAlister, D. (1879). The law of the geometric mean. *Proceedings of the Royal Society of London.* 29, pp. 367-376.

LOI MULTINOMIALE
Multinomial distribution

Considérons une **expérience aléatoire** de n épreuves indépendantes. Supposons que chaque épreuve ne puisse donner qu'un seul des événements E_i, avec une **probabilité** p_i.

La **variable aléatoire** X_i représente le nombre d'apparitions de l'événement E_i.

La loi multinomiale de paramètres n et p_i est définie par la **fonction de densité conjointe** suivante :

$$P(X_1 = x_1, X_2 = x_2, \ldots, X_s = x_s) = \frac{n!}{\prod_{i=1}^{s} x_i!} \prod_{i=1}^{s} p_i^{x_i},$$

avec les entiers $x_i \geq 0$ tels que $\sum_{i=1}^{s} x_i = n$.

En effet, la probabilité d'obtenir la séquence

$$\underbrace{E_1, \ldots, E_1}_{x_1}; \underbrace{E_2, \ldots, E_2}_{x_2}; \ldots; \underbrace{E_s, \ldots, E_s}_{x_s}$$

est égale à $\prod_{i=1}^{s} p_i^{x_i}$.

Le nombre de possibilités existant est le nombre de permutations de n objets dont x_1 sont des

événements E_1, x_2 sont des événements $E_2, \ldots, x_s$ sont des événements E_s, c'est-à-dire :

$$\frac{n!}{\prod_{i=1}^{s} x_i!}$$

La loi multinomiale est une **loi de probabilité discrète**.

ASPECTS MATHÉMATIQUES

Pour la loi multinomiale

$$\sum_{i=1}^{s} p_i = 1 \quad \text{et} \quad \sum_{i=1}^{s} x_i = n,$$

avec $0 < p_i < 1$.

L'**espérance mathématique** se calcule pour chaque **variable aléatoire** X_i.

Nous avons donc :

$$E[X_i] = n \cdot p_i.$$

En effet, chaque variable aléatoire peut être considérée individuellement comme suivant une **loi binomiale**.

Nous pouvons faire la même remarque pour la **variance**. Elle est alors égale à :

$$Var(X_i) = n \cdot p_i \cdot q_i.$$

DOMAINES ET LIMITATIONS

La loi multinomiale a de nombreuses applications dans l'analyse des données statistiques. En particulier, elle est utilisée dans des situations où les données doivent être classées en plusieurs catégories d'événements (par exemple dans l'analyse des données catégoriques).

EXEMPLE

Considérons l'**expérience aléatoire** suivante.

Une urne contient neuf boules, à savoir deux boules blanches, trois boules rouges et quatre boules noires. Une boule est choisie au hasard ; nous notons sa couleur et la remettons dans l'urne.

Nous pouvons décrire cette expérience aléatoire de la manière suivante.

Nous avons trois événements possibles :

E_1 : obtenir une boule blanche,
E_2 : obtenir une boule rouge,
E_3 : obtenir une boule noire ;

et trois variables aléatoires :

X_1 : nombre de boules blanches obtenues,
X_2 : nombre de boules rouges obtenues,
X_3 : nombre de boules noires obtenues.

Les probabilités associées à ces trois événements sont :

$$\begin{aligned} p_1 &= \frac{2}{9} \\ p_2 &= \frac{3}{9} \\ p_3 &= \frac{4}{9}. \end{aligned}$$

Nous allons calculer la probabilité que sur cinq boules choisies ($n = 5$), une soit blanche, deux soient rouges et deux soient noires.

Le nombre de possibilités est égal à :

$$\frac{n!}{\prod_{i=1}^{3} x_i!} = \frac{5!}{1!2!2!} = 30.$$

La probabilité d'obtenir la séquence (b, r, r, n, n) où b, r et n sont respectivement des boules blanches, rouges et noires vaut :

$$\prod_{i=1}^{3} p_i^{x_i} = \frac{2}{9} \cdot \left(\frac{3}{9}\right)^2 \cdot \left(\frac{4}{9}\right)^2 = 0,0049.$$

La probabilité que sur 5 boules, on ait une blanche, deux rouges et deux noires est donc égale à :

$$\begin{aligned} &P(X_1 = x_1, X_2 = x_2, X_3 = x_3) \\ &= \frac{5!}{\prod_{i=1}^{3} x_i!} \cdot \left(\frac{2}{9}\right)^{x_1} \cdot \left(\frac{3}{9}\right)^{x_2} \cdot \left(\frac{4}{9}\right)^{x_3} \end{aligned}$$

$P(X_1 = 1, X_2 = 2, X_3 = 2) = 30 \cdot 0,0049 = 0,147.$

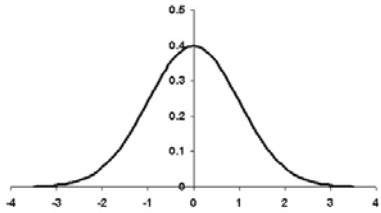
MOTS CLÉS

Loi binomiale *(Binomial distribution)*
Loi de probabilité discrète *(Discrete probability distribution)*

LOI NORMALE
Normal distribution

La **variable aléatoire** X est distribuée selon une loi normale si elle a une **fonction de densité** de la forme :

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right), \quad (\sigma > 0).$$



Loi normale, $\mu = 0$, $\sigma = 1$

Nous dirons que X suit une loi normale de moyenne μ et de variance σ^2. La loi normale est une **loi de probabilité continue**.

HISTORIQUE

La loi normale est souvent attribuée à **Pierre Simon de Laplace** et **Carl Friedrich Gauss** dont elle porte également le nom. Toutefois, son origine remonte aux travaux de **Jakob Bernoulli** qui dans son œuvre *Ars Conjectandi* (1713) a fourni les premiers éléments de base à la **loi des grands nombres**.

Abraham de Moivre fut le premier qui, en 1733, obtint la loi normale, comme approximation à la **loi binomiale**. Cet écrit était en latin ; il en publia une version anglaise en 1738. A. de Moivre parla de ce qu'il avait trouvé comme étant une « courbe » ; il découvrit cette courbe alors qu'il devait calculer les probabilités de gain pour différents jeux de hasard.

P.S. de Laplace, à la suite d'A. de Moivre, étudia cette loi et obtint un résultat plus formel et général que l'approximation d'A. de Moivre. Il obtint en 1774 la distribution normale comme approximation de la **loi hypergéométrique**.

C.F. Gauss s'intéressa à cette distribution par le biais des problèmes de mesure en astronomie. Ses travaux en 1809 et 1816 établissent des techniques basées sur la distribution normale qui deviendront des méthodes standards utilisées durant le XIX^{e} siècle.

Remarquons que bien que la première approximation de cette loi de distribution soit due à A. de Moivre, Galilée avait déjà trouvé que les erreurs d'observations étaient distribuées de façon symétrique et tendaient à se grouper autour de leur vraie **valeur**.

La littérature propose de nombreuses dénominations de la loi normale. **Adolphe Quetelet** parlait de la « courbe des possibilités » ou « loi des possibilités ».

Notons également que **Francis Galton** parlait de « loi de **fréquence** des erreurs » ou de « loi de **déviation** d'après une **moyenne** ». S.M. Stigler (1980) présente une discussion plus détaillée sur les différents noms de cette courbe.

ASPECTS MATHÉMATIQUES

L'**espérance mathématique** de la loi normale est donnée par :

$$E[X] = \mu.$$

La **variance** est égale à :

$$Var(X) = \sigma^2.$$

Si la **moyenne** μ est égale à 0 et la variance σ^2 est égale à 1, nous obtenons alors une distribution normale centrée réduite (ou loi normale standard), dont la **fonction de densité** est donnée par :

$$f(x) = \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-x^2}{2}\right).$$

Si une **variable aléatoire** X suit une loi normale de moyenne μ et de variance σ^2, la variable aléatoire

$$Z = \frac{X - \mu}{\sigma}$$

suit une loi normale centrée réduite (de moyenne 0 et de variance 1).

DOMAINES ET LIMITATIONS

La loi normale est la plus connue des lois de probabilité continues. Elle joue un rôle central dans la théorie des probabilités et dans ses applications statistiques.

Beaucoup de mesures, telles que la taille ou le poids d'individus, le diamètre d'une pièce de machine, les résultats à un test du QI suivent « approximativement » une loi normale.

La loi normale est fréquemment utilisée comme approximation, soit lorsque la normalité est attribuée à une distribution dans la construction d'un **modèle**, soit lorsqu'une distribution connue est remplacée par une distribution normale, de même **espérance mathématique** et de même **variance**. Elle est notamment utilisée pour l'approximation de la **loi du chi-carré**, de la **loi de Student**, de même que de lois de probabilité discrètes telle que la **loi binomiale** et la **loi de Poisson**.

La loi normale est également un élément fondamental de la théorie d'**échantillonnage**, où son rôle est prépondérant notamment dans l'étude de la corrélation, l'**analyse de régression**, l'**analyse de variance** et l'**analyse de covariance**.

MOTS CLÉS

Loi de probabilité continue *(Continuous probability distribution)*
Loi lognormale *(Lognormal distribution)*
Table normale *(Normal table)*

RÉFÉRENCES

Bernoulli, J. (1713). *Ars Conjectandi, Opus Posthumum. Accedit Tractatus de Seriebus infinitis, et Epistola Gallice scripta de ludo Pilæ recticularis.* Basileæ, Impensis Thurnisiorum, fratrum.

Gauss, C.F. (1809). *Theoria motus corporum cœlestium.* Werke 7.

Gauss, C.F. (1816). Bestimmung der Genauigkeit der Beobachtungen. Vol. 4, pp. 109-117. In *C.F. Gauss Werke* (publié en 1880). Dieterichsche Universitäts-Druckerei, Göttingen.

Laplace, P.S. de (1774). *Mémoire sur la probabilité des causes par les événements.* Mémoire de l'Académie royale des sciences présenté par divers savants, 6, pp. 621-656 ou Vol. 8, pp. 27-65. In P.S. de Laplace, *Œuvres complètes* (1891). Gauthier-Villars, Paris.

Moivre, A. de (1718). *The Doctrine of Chances: or, A Method of Calculating the Probability of Events in Play.* W. Pearson, London.

Moivre, A. de (1733). Approximatio ad summam terminorum binomii $(a + b)^n$, in Seriem Expansi, Supplementum II to *Miscellanæ Analytica*, pp. 1-7. Photographically reprinted in *A rare pamphlet of Moivre and some of his discoveries.* Publié par Archibald R.C. (1926). *Isis* 8, pp. 671-683.

Stigler, S. (1980). Stigler's law of eponymy, *Transactions of the New York Academy of Sciences*, 2nd series, 39, pp. 147- 157.

LOI UNIFORME
Uniform distribution

Une loi uniforme est utilisée pour décrire une **population** uniformément distribuée sur un **intervalle**.

Une **variable aléatoire** X est dite uniformément distribuée sur l'**intervalle** $[a, b]$ si sa **fonction de densité** est donnée par :

$$f(x) = \begin{cases} \dfrac{1}{b-a} & \text{si } a \leq x \leq b \\ 0 & \text{sinon.} \end{cases}$$

En raison de la forme rectangulaire de sa fonction de densité, la loi uniforme est aussi parfois appelée loi rectangulaire.



Loi uniforme, $a = 1$, $b = 3$

La loi uniforme est une **loi de probabilité continue**.

ASPECTS MATHÉMATIQUES

La **fonction de répartition** d'une loi uniforme est la suivante :

$$F(x) = \begin{cases} 0 & \text{si } x \leq a, \\ \dfrac{x-a}{b-a} & \text{si } a < x < b, \\ 1 & \text{sinon.} \end{cases}$$

L'**espérance mathématique** d'une **variable aléatoire** uniformément distribuée sur un **intervalle** $[a, b]$ est égale à :

$$E[X] = \int_a^b \frac{x}{b-a} dx = \frac{b^2 - a^2}{2(b-a)} = \frac{a+b}{2}.$$

Ce résultat est intuitivement très aisé à comprendre puisque la **fonction de densité** est symétrique autour du point milieu de l'intervalle $[a, b]$.

Nous pouvons calculer

$$E\left[X^2\right] = \int_a^b x^2 f(x)\, dx = \frac{b^2 + ab + a^2}{3}.$$

La **variance** est donc égale à :

$$Var(X) = E\left[X^2\right] - (E[X])^2 = \frac{(b-a)^2}{12}.$$

La loi uniforme est un cas particulier de la **loi bêta** quand $\alpha = \beta = 1$.

DOMAINES ET LIMITATIONS

Les applications de la loi uniforme sont nombreuses dans la construction de modèles pour des phénomènes physiques, biologiques et sociaux.

La loi uniforme est souvent utilisée comme modèle pour les erreurs d'arrondis lors de mesures. Par exemple, si nous mesurons la longueur d'objets au centimètre, une longueur de 35 cm peut représenter tous les objets dont la longueur est comprise entre $34,5$ et $35,5$ cm. Les erreurs d'arrondis suivent alors une loi uniforme sur l'intervalle $[-0,5\,;\, 0,5]$.

La loi uniforme est souvent utilisée pour générer des nombres aléatoires de n'importe quelle **loi de probabilité** discrète ou continue. La méthode utilisée pour générer une loi continue est la suivante :

Soit X une **variable aléatoire** continue et $F(x)$ sa **fonction de répartition**. Supposons que F est continue et strictement croissante. Alors $U = F(x)$ est une variable aléatoire uniforme et continue. U prend ses valeurs dans l'intervalle $[0, 1]$.

Nous avons alors $X = F^{-1}(u)$, où F^{-1} est la fonction inverse de la fonction de répartition.

Par conséquent, si $u_1, u_2, \ldots, u_n$ est une suite de nombres aléatoires

$$\begin{aligned} x_1 &= F^{-1}(u_1), \\ x_2 &= F^{-1}(u_2), \\ &\ldots, \\ x_n &= F^{-1}(u_n) \end{aligned}$$

est une suite de nombres aléatoires issus de la loi de la variable aléatoire X.

EXEMPLE

Si la **variable aléatoire** X suit une **loi exponentielle** négative de moyenne égale à 1, alors sa **fonction de répartition** est

$$F(x) = 1 - \exp(-x).$$

Comme $F(x) = u$, nous avons $u = 1 - \exp(-x)$.

Pour trouver x, nous cherchons F^{-1} :

$$\begin{aligned} u &= 1 - \exp(-x) \\ u - 1 &= -\exp(-x) \\ 1 - u &= \exp(-x) \end{aligned}$$

$$\begin{aligned} \ln(1-u) &= -x \\ -\ln(1-u) &= x. \end{aligned}$$

Remarquons que si u est une variable aléatoire uniforme sur $[0,1]$, $(1-u)$ est aussi une variable aléatoire uniforme sur $[0,1]$.

En appliquant ce qui précède à une suite de nombres aléatoires, nous trouvons un **échantillon** issu de X :

U	X
$0,14$	$1,97$
$0,97$	$0,03$
$0,53$	$0,63$
$0,73$	$0,31$
$\vdots$	$\vdots$

MOTS CLÉS

Loi bêta *(Beta distribution)*
Loi de probabilité continue *(Continuous probability distribution)*
Loi uniforme discrète *(Discrete uniform distribution)*

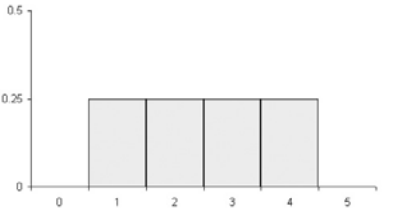
LOI UNIFORME DISCRÈTE

Discrete uniform distribution

La loi uniforme discrète est une **loi de probabilité discrète**. La **loi de probabilité continue** correspondante est la **loi uniforme** (continue).

Soient n événements ayant chacun la même **probabilité** $P(X=x) = \frac{1}{n}$, la **variable aléatoire** X suit une loi uniforme discrète et sa **fonction de probabilité** est :

$$P(X=x) = \begin{cases} \frac{1}{n} & \text{si } x = 1,2,\ldots,n, \\ 0 & \text{sinon.} \end{cases}$$



Loi uniforme discrète, $n = 4$

ASPECTS MATHÉMATIQUES

L'**espérance mathématique** de la loi uniforme discrète est par définition :

$$\begin{aligned} E[X] &= \sum_{x=1}^{n} x \cdot P(X=x) \\ &= \frac{1}{n} \cdot \sum_{x=1}^{n} x \\ &= \frac{1}{n} \cdot (1+2+\ldots+n) \\ &= \frac{1}{n} \cdot \frac{n^2+n}{2} \\ &= \frac{n+1}{2}. \end{aligned}$$

La **variance** de la loi uniforme discrète est égale à :

$$Var(x) = E\left[X^2\right] - (E[X])^2.$$

Comme

$$\begin{aligned} E\left[X^2\right] &= \sum_{x=1}^{n} x^2 \cdot P(X=x) \\ &= \frac{1}{n} \cdot \sum_{x=1}^{n} x^2 \\ &= \frac{1}{n} \cdot \left(1^2+2^2+\ldots+n^2\right) \\ &= \frac{1}{n} \cdot \frac{n(n+1)(2n+1)}{6} \\ &= \frac{2n^2+3n+1}{6} \end{aligned}$$

et

$$(E[X])^2 = \frac{(n+1)^2}{2^2},$$

nous avons :

$$\begin{aligned} Var(X) &= \frac{2n^2+3n+1}{6} - \frac{(n+1)^2}{2^2} \\ &= \frac{n^2-1}{12}. \end{aligned}$$

DOMAINES ET LIMITATIONS

La loi uniforme discrète est souvent utilisée pour générer des nombres aléatoires de n'importe quelle **loi de probabilité** discrète ou continue.

EXEMPLE

Considérons la **variable aléatoire** X donnant le nombre de points obtenus dans un jet de dé. Si le dé n'est pas truqué, la **probabilité** d'obtenir un 1 est égale à $\frac{1}{6}$. La probabilité d'obtenir un 2 est aussi de $\frac{1}{6}$, etc., de sorte que nous avons :

$$\begin{aligned} P(X=1) &= \frac{1}{6} \\ P(X=2) &= \frac{1}{6} \\ &\dots \\ P(X=6) &= \frac{1}{6}. \end{aligned}$$

Le nombre d'événements possibles est $n = 6$. Nous avons donc bien :

$$P(X=x) = \begin{cases} \frac{1}{6} & \text{pour } x = 1, 2, \dots, 6, \\ 0 & \text{sinon.} \end{cases}$$

C'est-à-dire que la variable aléatoire X suit une loi uniforme discrète.

MOTS CLÉS

Loi de probabilité discrète *(Discrete probability distribution)*
Loi uniforme *(Uniform distribution)*



M

MAHALANOBIS PRASANTA CHANDRA

Mahalanobis Prasanta Chandra



Prasanta Chandra Mahalanobis (1893-1972) effectua sa scolarité à Calcutta jusqu'en 1908. En 1912, il obtint sa licence de physique au Presidency College de Calcutta. En 1915, il alla en Angleterre afin de faire un *Tripos* (licence obtenue à Cambridge) en mathématique et physique au King's College à Cambridge. Avant d'entamer ses recherches, il retourna à Calcutta pour quelques vacances, mais ne revint pas en Angleterre car la guerre éclata.

En tant que statisticien, P.C. Mahalanobis ne concevait pas les statistiques pour elles-mêmes, mais plutôt comme un outil pour la compréhension et l'interprétation des données scientifiques et d'ingénieurie ainsi que les décisions pour le bien-être de la société. Il les utilisa pour étudier tout d'abord l'anthropologie (vers 1917) puis les inondations.

En 1930, il publia un article sur la statistique D^2 intitulé : *On Tests and Measures of Groups Divergence.*

Son plus grand apport aux statistiques est sans aucun doute l'enquête par sondage à grande échelle. Il y apporta plusieurs contributions telles que le choix optimal d'échantillonnage utilisant la variance et des fonctions de coût. Il fut nommé président de l'United Nations Subcommission on Statistical Sampling en 1947 et conserva ce poste jusqu'en 1951.

On peut encore citer son amitié avec **Ronald Aylmer Fisher** qui, depuis 1926, dura jusqu'à la mort de ce dernier.

Quelques ouvrages et articles principaux de Prasanta Chandra Mahalanobis :

1930 On Tests and Measures of Groups Divergence. Part I. Theoretical formulæ. *Journal of the Asiatic Society of Bengal*, 26, pp. 541-588.

1940 A sample survey of the acreage under jute in Bengal, *Sankhya* 4, pp. 511-530.

1944 On Large-scale Sample Surveys. *Philosophical Transactions of the Royal Society of London.* Series B, Vol. 231, pp. 329-451.

1963 *The Approach of Operational Research to Planning in India.* Asia Publishing House, Bombay.

MOT CLÉ

Distance de Mahalanobis *(Mahalanobis distance)*

MANN-WITHNEY, TEST DE

Mann-Whitney test

Voir **test de Mann-Whitney**.

MARKOV ANDREÏ ANDREÏEVITCH

Markov Andrei Andreevich

Markov naquit en 1856 à Ryazan et mourut en 1922 à Saint-Pétersbourg.

On doit à cet élève de **Pafnouti Lvovitch Tchebychev** de très importants travaux en calcul des probabilités. Il donna, notamment, une preuve rigoureuse du **théorème central limite**. À travers ses travaux sur les chaînes de Markov, le concept de dépendance markovienne s'infiltra dans la théorie moderne et les applications de processus aléatoires. La globalité de son œuvre influença le développement des probabilités et des statistiques internationales.



Il étudia les mathématiques à l'Université de Saint-Pétersbourg, où il obtint sa licence, sa maîtrise et son doctorat en 1878. Ensuite, il y enseigna, reprenant le cours de probabilités de P.L. Tchebychev au moment où celui-ci quitta l'université. En 1886, il fut élu membre de l'École de mathématiques fondée par P.L. Tchebychev à l'Académie des sciences de Saint-Pétersbourg. Il en devint un membre à part entière en 1886 et se retira de l'université en 1905, où il continuait d'enseigner.

Markov et Liapunov furent des élèves célèbres de Tchebychev et tentèrent sans cesse d'établir les probabilités comme une science mathématique exacte et pratique.

La première apparition des chaînes de Markov dans l'œuvre de Markov a eu lieu dans *Izvestiia*, bulletin de la Société physico-mathématique de l'Université de Kazan.

Markov compléta la preuve du théorème central limite que P.L. Tchebychev avait commencée mais non terminée. Il en fit une approche via la méthode des moments et la preuve fut publiée dans la troisième édition de *Ischislenie Veroiatnostei* (*Calcul des probabilités*).

Ouvrage principal d'Andreï Andreïevitch Markov :

1913 *Ischislenie Veroiatnostei*, 3e édition, Gosizdat, Moscow.

MOTS CLÉS

Théorème central limite *(Central limit theorem)*
Théorème de Gauss-Markov *(Gauss-Markov theorem)*

MATRICE

Matrix

On appelle matrice $(m \times n)$ un tableau rectangulaire de nombres à m lignes et n colonnes. Les éléments figurant dans le tableau sont appelés les coefficients ou composantes de la matrice.

ASPECTS MATHÉMATIQUES

On peut représenter une matrice de la manière suivante :

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{bmatrix}.$$

Un coefficient de $\mathbf{A}$ est caractérisé par sa **valeur** et sa position. Donc a_{ij} représente l'élément de la matrice

$\mathbf{A}$ à la i-ème ligne et j-ème colonne. Une matrice de m lignes et de n colonnes est dite d'ordre $(m \times n)$. Une matrice de n lignes et de n colonnes est dite carrée d'ordre $(n \times n)$. On trouve différents types de matrice :

- *Matrice nulle*

  C'est la matrice dont tous les éléments sont nuls. Elle sera notée $\mathbf{0}$.

- **Matrice identité** *d'ordre n*

  C'est une matrice carrée d'ordre n, dont tous les éléments diagonaux (ceux qui ont le même indice pour la ligne que pour la colonne, les a_{ii}) valent 1 et des 0 partout ailleurs. On la désigne par $\mathbf{I}$, ou parfois par $\mathbf{I_n}$ où n est l'ordre de la matrice identité.

$$\mathbf{I_n} = \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{bmatrix}.$$

- *Matrice symétrique*

  Une matrice carrée $\mathbf{A}$ est symétrique si $\mathbf{A} = \mathbf{A}'$ où $\mathbf{A}'$ est la **transposée** de la matrice $\mathbf{A}$.

$$\begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{12} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{1n} & a_{2n} & \dots & a_{nn} \end{bmatrix}.$$

- *Matrice anti-symétrique*

  Une matrice carrée $\mathbf{A}$ est anti-symétrique si $\mathbf{A} = -\mathbf{A}'$ où $\mathbf{A}'$ est la transposée de la matrice $\mathbf{A}$; il s'ensuit que les éléments de la diagonale principale a_{ii} sont nuls.

$$\begin{bmatrix} 0 & a_{12} & \dots & a_{1n} \\ -a_{12} & 0 & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ -a_{1n} & -a_{2n} & \dots & 0 \end{bmatrix}.$$

- *Matrice scalaire*

  Une matrice carrée $\mathbf{A}$ est dite matrice scalaire si :

$$a_{ij} = \beta \text{ pour } i = j$$
$$a_{ij} = 0 \text{ pour } i \neq j$$

  où β est un scalaire :

$$\begin{bmatrix} \beta & 0 & \dots & 0 \\ 0 & \beta & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \beta \end{bmatrix}.$$

- *Matrice diagonale*

  Une matrice carrée $\mathbf{A}$ est dite matrice diagonale si :

$$a_{ij} = 0 \quad \text{pour } i \neq j$$

  et où les a_{ii} sont des nombres quelconques :

$$\begin{bmatrix} a_{11} & 0 & \dots & 0 \\ 0 & a_{22} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & a_{nn} \end{bmatrix}.$$

- *Matrice triangulaire*

  Une matrice carrée $\mathbf{A}$ est dite triangulaire supérieure si $a_{ij} = 0$ pour $i > j$:

$$\begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ 0 & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & a_{nn} \end{bmatrix},$$

  ou inférieure si $a_{ij} = 0$ pour $i < j$:

$$\begin{bmatrix} a_{11} & 0 & \dots & 0 \\ a_{12} & a_{22} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ a_{1n} & a_{2n} & \dots & a_{nn} \end{bmatrix}.$$

- **Matrice idempotente**

  Une matrice carrée $\mathbf{A}$ est idempotente si elle possède la propriété

$$\mathbf{A} \cdot \mathbf{A} = \mathbf{A}.$$

  Si une matrice est idempotente, son **déterminant** est égal à 1 ou 0.

  La seule matrice idempotente non singulière (c'est-à-dire dont le déterminant est non nul) est la matrice identité.

Opérations sur les matrices

- *Addition de matrices*

  Si $\mathbf{A} = (a_{ij})$ et $\mathbf{B} = (b_{ij})$ sont deux matrices d'ordre $(m \times n)$, leur somme $\mathbf{A} + \mathbf{B}$ est définie par la matrice : $\mathbf{C} = (c_{ij})$ d'ordre $(m \times n)$ dont chaque élément est la somme des éléments correspondants de $\mathbf{A}$ et $\mathbf{B}$:

  $$\mathbf{C} = (c_{ij}) = (a_{ij} + b_{ij}) .$$

- *Multiplication d'une matrice par un scalaire*

  Si α est un scalaire et $\mathbf{A}$ une matrice, le produit $\alpha\mathbf{A}$ est la matrice de même ordre que $\mathbf{A}$ obtenue en multipliant chaque élément a_{ij} de $\mathbf{A}$ par le scalaire α :

  $$\alpha\mathbf{A} = \mathbf{A}\alpha = (\alpha \cdot a_{ij}) .$$

- *Multiplication de matrices*

  Soient $\mathbf{A}$ et $\mathbf{B}$ deux matrices, le produit $\mathbf{AB}$ est défini si et seulement si le nombre de colonnes de $\mathbf{A}$ est égal au nombre de lignes de $\mathbf{B}$. Si $\mathbf{A}$ est d'ordre $(m \times n)$ et $\mathbf{B}$ d'ordre $(n \times q)$, alors le produit $\mathbf{AB}$ est défini par la matrice $\mathbf{C}$ d'ordre $(m \times q)$ dont les éléments sont obtenus par :

  $$c_{ij} = \sum_{k=1}^{n} a_{ik} \cdot b_{kj} \quad \begin{array}{l} \text{pour } i = 1, \ldots, m \\ \text{et } j = 1, \ldots, q. \end{array}$$

  Cela signifie qu'on multiplie les éléments de la i-ème ligne de $\mathbf{A}$ par les éléments correspondants de la j-ème colonne de $\mathbf{B}$ et qu'on additionne les résultats.

- *Puissance entière de matrices carrées*

  Soit $\mathbf{A}$ une matrice carrée, si p est un entier positif, on définit la puissance p-ème de $\mathbf{A}$ par le produit de la matrice $\mathbf{A}$ par elle-même et ceci p fois :

  $$\mathbf{A}^p = \mathbf{A} \cdot \mathbf{A} \cdot \ldots \cdot \mathbf{A} \quad p \text{ fois.}$$

  Par convention, pour $p = 0$, on pose : $\mathbf{A}^0 = \mathbf{I}$ où $\mathbf{I}$ est la matrice identité de même ordre que $\mathbf{A}$.

- *Racine carrée d'une matrice diagonale :*

  Soit $\mathbf{A}$ une matrice diagonale à coefficients non négatifs, on note $\sqrt{\mathbf{A}}$ la racine carrée de $\mathbf{A}$, définie par la matrice diagonale de même ordre que $\mathbf{A}$, dans laquelle on a remplacé chaque élément par sa racine carrée :

  $$\sqrt{\mathbf{A}} = \begin{bmatrix} \sqrt{a_{11}} & 0 & \ldots & 0 \\ 0 & \sqrt{a_{22}} & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \ldots & 0 & \sqrt{a_{nn}} \end{bmatrix} .$$

DOMAINES ET LIMITATIONS

L'utilisation des matrices, ou plus généralement du calcul matriciel, est très courante en analyse de régression, en **analyse de variance**, ainsi que dans de nombreux autres domaines de la **statistique**. Elle permet d'écrire de façon plus concise un système d'équations et allège passablement les notations. Pour ces mêmes raisons, les matrices apparaissent également en **analyse de données** où l'on rencontre en général de grands tableaux de données.

EXEMPLE

Soient $\mathbf{A}$ et $\mathbf{B}$ deux matrices carrées d'ordre 2 :

$$\mathbf{A} = \begin{bmatrix} 1 & 2 \\ 1 & 3 \end{bmatrix} \text{ et } \mathbf{B} = \begin{bmatrix} 6 & 2 \\ -2 & 1 \end{bmatrix} .$$

La somme des deux matrices est définie par :

$$\begin{aligned} \mathbf{A} + \mathbf{B} &= \begin{bmatrix} 1 & 2 \\ 1 & 3 \end{bmatrix} + \begin{bmatrix} 6 & 2 \\ -2 & 1 \end{bmatrix} \\ &= \begin{bmatrix} 1+6 & 2+2 \\ 1-2 & 3+1 \end{bmatrix} = \begin{bmatrix} 7 & 4 \\ -1 & 4 \end{bmatrix} . \end{aligned}$$

La multiplication de la matrice $\mathbf{A}$ par 3 donne :

$$\begin{aligned} 3 \cdot \mathbf{A} &= 3 \cdot \begin{bmatrix} 1 & 2 \\ 1 & 3 \end{bmatrix} = \begin{bmatrix} 3 \cdot 1 & 3 \cdot 2 \\ 3 \cdot 1 & 3 \cdot 3 \end{bmatrix} \\ &= \begin{bmatrix} 3 & 6 \\ 3 & 9 \end{bmatrix} . \end{aligned}$$

Le produit matriciel de $\mathbf{A}$ par $\mathbf{B}$ vaut :

$$\mathbf{A} \cdot \mathbf{B} = \begin{bmatrix} 1 & 2 \\ 1 & 3 \end{bmatrix} \cdot \begin{bmatrix} 6 & 2 \\ -2 & 1 \end{bmatrix}$$

$$= \begin{bmatrix} 1\cdot 6+2\cdot(-2) & 1\cdot 2+2\cdot 1 \\ 1\cdot 6+3\cdot(-2) & 1\cdot 2+3\cdot 1 \end{bmatrix}$$
$$\mathbf{A}\cdot\mathbf{B} = \begin{bmatrix} 2 & 4 \\ 0 & 5 \end{bmatrix}.$$

Remarque : $\mathbf{A}\cdot\mathbf{B}$ et $\mathbf{B}\cdot\mathbf{A}$ ne sont en général pas égaux, ce qui illustre la non-commutativité de la multiplication de matrices.

Le carré de la matrice $\mathbf{A}$ s'obtient en multipliant $\mathbf{A}$ par $\mathbf{A}$, ainsi :

$$\mathbf{A}^2 = \mathbf{A}\cdot\mathbf{A} = \begin{bmatrix} 1 & 2 \\ 1 & 3 \end{bmatrix}\cdot\begin{bmatrix} 1 & 2 \\ 1 & 3 \end{bmatrix}$$
$$= \begin{bmatrix} 1^2+2\cdot 1 & 1\cdot 2+2\cdot 3 \\ 1^2+3\cdot 1 & 1\cdot 2+3\cdot 3 \end{bmatrix}$$
$$\mathbf{A}^2 = \begin{bmatrix} 3 & 8 \\ 4 & 11 \end{bmatrix}.$$

Soit $\mathbf{D}$ la matrice diagonale d'ordre 2 suivante :

$$\mathbf{D} = \begin{bmatrix} 9 & 0 \\ 0 & 2 \end{bmatrix}.$$

On définit la racine carrée de $\mathbf{D}$, notée $\sqrt{\mathbf{D}}$ par :

$$\sqrt{\mathbf{D}} = \begin{bmatrix} \sqrt{9} & 0 \\ 0 & \sqrt{2} \end{bmatrix} = \begin{bmatrix} 3 & 0 \\ 0 & \sqrt{2} \end{bmatrix}.$$

MOTS CLÉS

Analyse de données *(Data analysis)*
Analyse factorielle des correspondances *(Correspondence analysis)*
Déterminant *(Determinant)*
Matrice d'inertie *(Inertia matrix)*
Matrice H *(Hat matrix)*
Matrice inverse *(Inverse matrix)*
Produit scalaire *(Scalar product)*
Régression, analyse de *(Regression analysis)*
Régression linéaire multiple *(Multiple linear regression)*
Trace *(Trace)*
Transposée *(Transpose)*

RÉFÉRENCES

Balestra, P. (1972). *Calcul matriciel pour économistes.* Éditions Castella, Albeuve, Suisse.

Graybill, F.A. (1976). *Theory and Applications of the Linear Model.* Duxbury Press, North Scituate, Massachusetts; Waldsworth and Brooks/Cole, Pacific Grove, CA.

Searle, S.R. (1982). *Matrix Algebra Useful for Statistics.* John Wiley & Sons, New York.

MATRICE DE VARIANCE-COVARIANCE

Variance-covariance matrix

La matrice $\mathbf{V}$ de variance-covariance entre n variables $X_1, \ldots, X_n$, est une matrice d'ordre $(n \times n)$ où la diagonale contient les **variances** de chaque variable X_i et, en dehors des diagonales, les **covariances** entre les couples de variables (X_i, X_j) pour tout $i \neq j$. Pour n variables, on calcule n variances et $n(n-1)$ covariances. Il en résulte que la matrice $\mathbf{V}$ est une matrice diagonale de dimension $n \times n$. Dans la littérature, on rencontre souvent cette matrice sous deux autres noms : matrice de dispersion (*dispersion matrix*) ou matrice de covariances (*covariances matrix*).

Si les variables sont standardisées (voir **données standardisées**), les variances des variables sont égales à 1 et les covariances deviennent des corrélations et ainsi la matrice devient une matrice de corrélation.

HISTORIQUE

Voir **variance** et **covariance**.

ASPECTS MATHÉMATIQUES

Soit $X_1, \ldots, X_n$ n variables aléatoires.

La matrice de variance-covariance est une matrice d'ordre $(n \times n)$, symétrique, qui contient sur la diagonale les variances de chaque variable et dans les termes extérieurs les covariances entre les variables. La matrice $\mathbf{V}$ de variance-covariance est la matrice suivante :

$$\begin{bmatrix} Var(X_1) & Cov(X_1,X_2) & \cdots & Cov(X_1,X_n) \\ Cov(X_2,X_1) & Var(X_2) & \cdots & Cov(X_2,X_n) \\ \vdots & \vdots & \ddots & \vdots \\ Cov(X_n,X_1) & Cov(X_n,X_2) & \cdots & Var(X_n) \end{bmatrix}$$

avec

$Var(X_i) = E\left[X_i^2\right] - (E[X_i])^2$ et
$Cov(X_i,X_j) = E[(X_i - E[X_i])(X_j - E[X_j])]$.

EXEMPLE

Soit les trois variables suivantes X_1, X_2 et X_3 résumées dans le tableau suivant :

X_1	X_2	X_3
1	2	3
2	3	4
1	2	3
5	4	3
4	4	4

Dans ce cas, $n = 3$. Chaque variable aléatoire à cinq termes, que l'on note $N = 5$.

On désire calculer la matrice de variance-covariance $\mathbf{V}$ de ces variables. On passe successivement par les étapes suivantes :

On estime l'espérance de chaque variable par sa moyenne empirique $\overline{x}_i$. On obtient les résultats suivants :

$$\begin{aligned} \overline{X}_1 &= 2,6 \\ \overline{X}_2 &= 3 \\ \overline{X}_3 &= 3,4. \end{aligned}$$

La variance pour une variable X_j est donnée par :

$$\frac{\sum_{i=1}^{N} (x_{ij} - \overline{x}_j)^2}{4}.$$

La covariance entre deux variables X_r et X_s se calcule de la manière suivante :

$$\frac{\sum_{i=1}^{N} (x_{ir} - \overline{x}_r)(x_{is} - \overline{x}_s)}{4}.$$

Ensuite il faut procéder à la soustraction de la moyenne $\overline{X}_i$ de chaque variable X_i. On obtient une nouvelle variable $Y_i = X_i - \overline{X}_i$.

Y_1	Y_2	Y_3
$-1,6$	-1	$-0,4$
$-0,6$	0	0,6
$-1,6$	-1	$-0,4$
2,4	1	$-0,4$
1,4	1	0,6

On remarque que la moyenne de chaque nouvelle variable Y_i est égale à 0. On met ces résultats sous forme matricielle :

$$\mathbf{Y} = \begin{bmatrix} -1,6 & -1 & -0,4 \\ -0,6 & 0 & 0,6 \\ -1,6 & -1 & -0,4 \\ 2,4 & 1 & -0,4 \\ 1,4 & 1 & 0,6 \end{bmatrix}.$$

À partir de cette matrice, on calcule la matrice $\mathbf{Y}' \cdot \mathbf{Y}$, matrice symétrique, de la forme suivante :

$$\begin{bmatrix} \sum_{i=1}^{N} y_{i1}^2 & \sum_{i=1}^{N} y_{i2} \cdot y_{i1} & \sum_{i=1}^{N} y_{i3} \cdot y_{i1} \\ \sum_{i=1}^{N} y_{i2} \cdot y_{i1} & \sum_{i=1}^{N} y_{i2}^2 & \sum_{i=1}^{N} y_{i3} \cdot y_{i2} \\ \sum_{i=1}^{N} y_{i3} \cdot y_{i1} & \sum_{i=1}^{N} y_{i3} \cdot y_{i2} & \sum_{i=1}^{N} y_{i3}^2 \end{bmatrix}.$$

Cette matrice contient les valeurs de la somme des produits des carrés de chaque variable Y_i (sur la diagonale) et le produit croisé de chaque paire de variables (Y_i, Y_j) sur le reste de la matrice. Pour les variables originales X_i, la matrice ci-dessous contient sur la diagonale la somme des carrés des déviations par rapport à la moyenne pour chaque variable et sur les termes extérieurs les produits croisés de ces déviations.

Numériquement on obtient :

$$\begin{bmatrix} 13,2 & 7 & 0,8 \\ 7 & 4 & 1 \\ 0,8 & 1 & 1,2 \end{bmatrix}.$$

La moyenne de la somme des carrés des déviations par rapport à la moyenne d'une variable aléatoire

X_i à N termes est connue sous le nom de **variance**. La moyenne des produits croisés de ces déviations s'appelle la **covariance**. Ainsi, pour obtenir la matrice de variance-covariance $\mathbf{V}$ des variables X_1, ..., X_n il faut diviser les éléments de la matrice $\mathbf{Y}' \cdot \mathbf{Y}$ par $N-1$, ce qui revient mathématiquement à :

$$\begin{bmatrix} \frac{\sum_{i=1}^{N} y_{i1}^2}{N-1} & \frac{\sum_{i=1}^{N} y_{i2} \cdot y_{i1}}{N-1} & \frac{\sum_{i=1}^{N} y_{i3} \cdot y_{i1}}{N-1} \\ \frac{\sum_{i=1}^{N} y_{i2} \cdot y_{i1}}{N-1} & \frac{\sum_{i=1}^{N} y_{i2}^2}{N-1} & \frac{\sum_{i=1}^{N} y_{i3} \cdot y_{i2}}{N-1} \\ \frac{\sum_{i=1}^{N} y_{i3} \cdot y_{i1}}{N-1} & \frac{\sum_{i=1}^{N} y_{i3} \cdot y_{i2}}{N-1} & \frac{\sum_{i=1}^{N} y_{i3}^2}{N-1} \end{bmatrix}.$$

Le fait de diviser par $N-1$ et non pas par N pour moyenner est lié au fait que l'on travaille généralement avec des échantillons de variables aléatoires dont on ne connaît pas la vraie population et pour lesquels on fait une estimation non biaisée de la variance et de la covariance. Numériquement, on obtient :

$$\mathbf{V} = \begin{bmatrix} 3,3 & 1,75 & 1,75 \\ 1,75 & 1 & 0,2 \\ 0,2 & 0,25 & 0,3 \end{bmatrix}.$$

MOTS CLÉS

Coefficient de corrélation *(Correlation coefficient)*
Covariance *(Covariance)*
Données standardisées *(Standardised Data)*
Variance *(Variance)*

MATRICE D'INERTIE

Inertia matrix

Considérons un nuage de points dans un espace à p dimensions, nous supposons que ce nuage est centré à l'origine du système de référence ; si tel n'est pas le cas, il suffit d'effectuer un changement de variables en soustrayant à chaque point le centre du nuage. Désignons par $\mathbf{A}$ la **matrice** des coordonnées multipliées par la racine carrée de la **fréquence** associée à chaque point, cette dernière sera, pour n points distincts, de dimension $(n \times p)$. Supposons encore que p est inférieur ou égal à n, sinon nous utilisons la **transposée** de la matrice $\mathbf{A}$ comme nouvelle matrice de coordonnées.

Nous obtenons la matrice d'inertie $\mathbf{V}$ en effectuant le produit matriciel de la transposée de $\mathbf{A}$, $\mathbf{A}'$ par la matrice $\mathbf{A}$ elle-même. $\mathbf{V}$ est alors une matrice carrée d'ordre p symétrique.

La **trace** de cette matrice $\mathbf{V}$, aussi appelée **matrice de variance-covariance**, n'est autre que l'**inertie** totale (ou variance totale).

ASPECTS MATHÉMATIQUES

Soient n points $\mathbf{X}_i$ distincts à p composantes

$$\mathbf{X}_i = (x_{i1}, x_{i2}, \ldots, x_{ip}) \quad i = 1, 2, \ldots, n$$

chaque point $\mathbf{X}_i$ a une **fréquence** f_i $(i = 1, 2, \ldots, n)$.

Si les points ne sont pas centrés autour de l'origine, il faut faire le changement de variables correspondant au déplacement de l'origine au centre du nuage de points :

$$\mathbf{X}'_i = \mathbf{X}_i - \mathbf{Y}$$

où $\mathbf{Y} = (y_1, y_2, \ldots, y_p)$ avec

$$y_j = \frac{1}{\sum_{i=1}^{n} f_i} \cdot \left(\sum_{i=1}^{n} f_i \cdot x_{ij} \right).$$

Le terme général de la matrice d'inertie $\mathbf{V}$ est donc :

$$v_{ij} = \sum_{k=1}^{n} f_k \cdot (x_{ki} - y_i) \cdot (x_{kj} - y_j).$$

Matriciellement, $\mathbf{V} = \mathbf{A}' \cdot \mathbf{A}$ où les coefficients de la matrice $\mathbf{A}$ sont :

$$a_{ki} = \sqrt{f_k} \cdot (x_{ki} - y_i).$$

L'**inertie** totale du nuage de points s'obtient en calculant la **trace** de la matrice d'inertie $\mathbf{V}$.

Elle vaut :

$$I_{tot} = tr(\mathbf{V}) = \sum_{i=1}^{p} v_{ii} = \sum_{i=1}^{p} \sum_{k=1}^{n} f_k \cdot (x_{ki} - y_i)^2.$$

DOMAINES ET LIMITATIONS

La matrice d'inertie d'un nuage de points est utilisée pour effectuer une **analyse factorielle des correspondances** entre les points considérés.

EXEMPLE

Considérons les points suivants dans l'espace à 3 dimensions :

$$\begin{array}{c}(2\,;\ \ 0\,;\ \ 0,5)\\(3\,;\ \ 1\,;\ \ \ 2)\\(-2\,;-1\,;\ \ -1)\\(-1\,;-1\,;-0,5)\\(-1\,;-1\,;-0,5)\\(-1\,;-1\,;-0,5)\end{array}$$

Il y a donc quatre points distincts de fréquences respectivement égales à 1, 1, 1 et 3.

Le centre du nuage formé par ces points est donné par :

$$\begin{aligned} y_1 &= \frac{2+3-2+3\cdot(-1)}{6} = 0 \\ y_2 &= \frac{1-1+3\cdot(-1)}{6} = -0,5 \\ y_3 &= \frac{0,5+2-1+3\cdot(-0,5)}{6} = 0. \end{aligned}$$

Les nouvelles coordonnées de nos 4 points distincts seront données par :

$$(2\,;\,0,5\,;\,0,5)\ (3\,;\,1,5\,;\,2)\ (-2\,;\,-0,5\,;\,-1)\ (-1\,;\,-0,5\,;\,-0,5) \text{ de fréquence 3.}$$

D'où $\mathbf{A}$ la matrice des coordonnées multipliées par la racine carrée de la fréquence :

$$\mathbf{A} = \begin{bmatrix} 2 & 0,5 & 0,5 \\ 3 & 1,5 & 2 \\ -2 & -0,5 & -1 \\ -\sqrt{3} & -\dfrac{\sqrt{3}}{2} & -\dfrac{\sqrt{3}}{2} \end{bmatrix}$$

et $\mathbf{V}$ la matrice d'inertie :

$$\mathbf{V} = \mathbf{A}' \cdot \mathbf{A} = \begin{bmatrix} 20 & 8 & 10,5 \\ 8 & 3,5 & 4,5 \\ 10,5 & 4,5 & 6 \end{bmatrix}.$$

L'**inertie** du nuage de points vaut :

$$I_{tot} = tr\,(\mathbf{V}) = 20 + 3,5 + 6 = 29,5.$$

MOTS CLÉS

Analyse factorielle des correspondances *(Correspondence analysis)*
Fréquence *(Frequency)*
Matrice *(Matrix)*
Trace *(Trace)*
Transposée *(Transpose)*
Variance *(Variance)*

RÉFÉRENCE

Lagarde, J. de (1983). *Initiation à l'analyse des données.* Dunod, Paris.

MATRICE H

Hat matrix

La matrice **H** est une **matrice** utilisée en **analyse de régression** et en **analyse de variance**. Elle est définie comme la matrice qui permet de passer des valeurs de la **variable** observée aux estimations obtenues par les **moindres carrés**.

Ainsi, lors d'une régression linéaire sous forme matricielle, si $\hat{\mathbf{Y}}$ est le **vecteur** formé par les estimations calculées à partir des paramètres des moindres carrés et si $\mathbf{Y}$ est le vecteur des observations relatives à la **variable dépendante**, la matrice **H** permet d'obtenir $\hat{\mathbf{Y}}$ par multiplication matricielle de $\mathbf{Y}$.

HISTORIQUE

La matrice **H** a été introduite par **John Wilder Tukey** en 1972. L'article de D.C. Hoaglin et R.E. Welsch (1978) donne les propriétés de la matrice **H** ainsi que plusieurs exemples d'application.

ASPECTS MATHÉMATIQUES

Considérons le **modèle** général d'une régression linéaire sous forme matricielle :

$$\boldsymbol{Y} = \boldsymbol{X} \cdot \boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

où

- $\mathbf{Y}$ est le **vecteur** $(n \times 1)$ des observations relatives à la **variable dépendante** (n observations),
- $\mathbf{X}$ est la **matrice** $(n \times p)$ ayant trait aux variables indépendantes (p variables indépendantes),
- $\boldsymbol{\varepsilon}$ est le vecteur $(n \times 1)$ des **erreurs** et
- $\boldsymbol{\beta}$ est le vecteur $(p \times 1)$ des paramètres à estimer.

L'estimation $\hat{\boldsymbol{\beta}}$ du vecteur $\boldsymbol{\beta}$ est donnée par :

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X'} \cdot \boldsymbol{X})^{-1} \cdot \boldsymbol{X'} \cdot \boldsymbol{Y}$$

et nous pouvons alors calculer les valeurs estimées $\hat{\mathbf{Y}}$ de $\mathbf{Y}$ connaissant $\hat{\boldsymbol{\beta}}$. En effet :

$$\hat{\boldsymbol{Y}} = \boldsymbol{X} \cdot \hat{\boldsymbol{\beta}} = \boldsymbol{X} \cdot (\boldsymbol{X'} \cdot \boldsymbol{X})^{-1} \cdot \boldsymbol{X'} \cdot \boldsymbol{Y} \ ;$$

la matrice $\mathbf{H}$ est donc définie par :

$$\mathbf{H} = \mathbf{X} \cdot (\mathbf{X}' \cdot \mathbf{X})^{-1} \cdot \mathbf{X}'.$$

En particulier, l'élément diagonal h_{ii} sera défini par :

$$h_{ii} = \mathbf{x}_i \cdot (\mathbf{X}' \cdot \mathbf{X})^{-1} \cdot \mathbf{x}_i',$$

où $\mathbf{x}_i$ est la i-ème ligne de $\mathbf{X}$.

DOMAINES ET LIMITATIONS

La matrice $\mathbf{H}$ permettant d'obtenir les n estimations de la **variable dépendante** à l'aide des n observations relatives à cette dernière, est une **matrice idempotente** carrée d'ordre n symétrique. L'élément (i, j) de cette **matrice** (se trouvant à l'intersection de la i-ème ligne et de la j-ème colonne pour i et j allant de 1 à n) mesure l'influence de la j-ème observation sur la i-ème **valeur** estimée. En particulier, les éléments diagonaux évaluent l'effet de l'observation sur l'estimation correspondante de la variable dépendante. La valeur de chaque élément diagonal de la matrice $\mathbf{H}$ est compris entre 0 et 1 (bornes comprises).

En écrivant $\mathbf{H} = (h_{ij})$ pour $i, j = 1, \ldots, n$, nous avons la relation :

$$h_{ii} = h_{ii}^2 + \sum_{\substack{i=1 \\ i \neq j}}^{n} \sum_{j=1}^{n} h_{ij}^2$$

démontrée en utilisant la propriété d'idempotence de $\mathbf{H}$, à savoir que $\mathbf{H} = \mathbf{H}^2$.

Ainsi $tr(\mathbf{H}) = \sum_{i=1}^{n} h_{ii} = p =$ nombre de paramètres à estimer.

La matrice $\mathbf{H}$ est utilisée pour déterminer les points leviers lors d'une analyse de régression.

EXEMPLE

Considérons le tableau suivant où Y est une **variable dépendante** expliquée par la **variable indépendante** X :

X	Y
50	6
52	8
55	9
75	7
57	8
58	10

Le **modèle** de **régression linéaire simple** sous forme matricielle s'écrit :

$$\boldsymbol{Y} = \boldsymbol{X} \cdot \boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

où

$$\mathbf{Y} = \begin{bmatrix} 6 \\ 8 \\ 9 \\ 7 \\ 8 \\ 10 \end{bmatrix} \quad \mathbf{X} = \begin{bmatrix} 1 & 50 \\ 1 & 52 \\ 1 & 55 \\ 1 & 75 \\ 1 & 57 \\ 1 & 58 \end{bmatrix},$$

$\boldsymbol{\varepsilon}$ est le **vecteur** (6×1) des erreurs, et $\boldsymbol{\beta}$ est le vecteur (2×1) des paramètres.

Nous trouvons la **matrice** $\mathbf{H}$ en appliquant le résultat :

$$\mathbf{H} = \mathbf{X} \cdot (\mathbf{X}' \cdot \mathbf{X})^{-1} \cdot \mathbf{X}'.$$

En effectuant les calculs matriciels par étape, nous obtenons :

$$\begin{aligned} (\mathbf{X}' \cdot \mathbf{X})^{-1} &= \frac{1}{2\,393} \cdot \begin{bmatrix} 20\,467 & -347 \\ -347 & 6 \end{bmatrix} \\ &= \begin{bmatrix} 8,5528 & -0,1450 \\ -0,1450 & 0,0025 \end{bmatrix} \end{aligned}$$

et finalement :

$$\mathbf{H} = \mathbf{X} \cdot (\mathbf{X}' \cdot \mathbf{X})^{-1} \cdot \mathbf{X}'$$

$$= \begin{bmatrix} 0,32 & 0,28 & 0,22 & -0,17 & 0,18 & 0,16 \\ 0,28 & 0,25 & 0,21 & -0,08 & 0,18 & 0,16 \\ 0,22 & 0,21 & 0,19 & 0,04 & 0,17 & 0,17 \\ -0,17 & -0,08 & 0,04 & 0,90 & 0,13 & 0,17 \\ 0,18 & 0,18 & 0,17 & 0,13 & 0,17 & 0,17 \\ 0,16 & 0,16 & 0,17 & 0,17 & 0,17 & 0,17 \end{bmatrix}.$$

Nous remarquons, par exemple, que le poids de y_1 dans l'**estimation** de $\hat{y}_1$ est égal à $0,32$.

Nous vérifions alors que la **trace** de $\mathbf{H}$ vaut bien 2, c'est-à-dire est égale au nombre de paramètres du **modèle**.

$$\begin{aligned} tr(\mathbf{H}) &= 0,32 + 0,25 + 0,19 + 0,90 + 0,17 + 0,17 \\ &= 2. \end{aligned}$$

MOTS CLÉS

Matrice *(Matrix)*
Point levier *(Leverage point)*
Régression, analyse de *(Regression analysis)*
Régression linéaire multiple *(Multiple linear regression)*
Régression linéaire simple *(Simple linear regression)*

RÉFÉRENCES

Belsley, D.A., Kuh, E. and Welsch, R.E. (1980). *Regression diagnostics.* John Wiley & Sons, New York, pp. 16-19.

Hoaglin, D.C. and Welsch, R.E. (1978). The hat matrix in regression and ANOVA. *The American Statistician*, 32, pp. 17-22 and corriged 32, p. 146.

Tukey, J.W. (1972). Some graphical and Semigraphical Displays. In *Statistical Papers in Honor of George W. Snedecor*, ed. T.A. Bancroft, Arnes: The Iowa State University Press. Iowa, U.S.A, pp. 293-316.

MATRICE IDEMPOTENTE

Idempotent matrix

Voir **matrice**.

MATRICE IDENTITÉ

Identity matrix

Voir **matrice**.

MATRICE INVERSE

Inverse matrix

Pour une **matrice** carrée $\mathbf{A}$ d'ordre n, on appelle matrice inverse de $\mathbf{A}$, la matrice carrée d'ordre n telle que le résultat du produit matriciel de $\mathbf{A}$ par son inverse (tout comme celui de l'inverse multipliant $\mathbf{A}$) soit égal à la **matrice identité** d'ordre n. Pour une matrice carrée quelconque, la matrice inverse, si elle existe, est unique. Pour qu'une matrice carrée $\mathbf{A}$ soit inversible, il faut que son déterminant soit différent de zéro.

ASPECTS MATHÉMATIQUES

Soit $\mathbf{A}$ une **matrice** carrée d'ordre n de déterminant non nul, on appelle matrice inverse de $\mathbf{A}$, l'unique matrice carrée d'ordre n, notée $\mathbf{A}^{-1}$, telle que :

$$\mathbf{A} \cdot \mathbf{A}^{-1} = \mathbf{A}^{-1} \cdot \mathbf{A} = \mathbf{I_n}$$

où $\mathbf{I_n}$ est la matrice identité d'ordre n. On dit alors que c'est une matrice régulière. Dans le cas où $\det(\mathbf{A}) = 0$, c'est-à-dire que la matrice $\mathbf{A}$ n'est pas inversible, on dit qu'elle est singulière.

La condition que le déterminant de $\mathbf{A}$ soit non nul s'explique par les propriétés du déterminant :

$$1 = \det(\mathbf{I_n}) = \det\left(\mathbf{A} \cdot \mathbf{A}^{-1}\right) = \det(\mathbf{A}) \cdot \det\left(\mathbf{A}^{-1}\right),$$

c'est-à-dire que $\det\left(\mathbf{A}^{-1}\right) = \frac{1}{\det(\mathbf{A})}$.

En outre, l'expression de $\mathbf{A}^{-1}$, à partir de $\mathbf{A}$, est obtenue à l'aide de $\frac{1}{\det(\mathbf{A})}$ (qui doit donc être défini), elle vaut :

$$\mathbf{A}^{-1} = \frac{1}{\det(\mathbf{A})} \cdot \left[(-1)^{i+j} \cdot D_{ij}\right]$$

où D_{ij} est le déterminant d'ordre $n-1$ que l'on obtient en supprimant dans $\mathbf{A}'$ (la **transposée** de $\mathbf{A}$) la

ligne i et la colonne j. D_{ij} est appelé le cofacteur de l'élément a_{ij}.

DOMAINES ET LIMITATIONS

La matrice inverse s'utilise principalement lors d'une régression linéaire sous forme matricielle afin de déterminer les paramètres à estimer.

En général, on peut dire que la notion de matrice inverse permet de résoudre un système d'équations linéaires.

EXEMPLE

Traitons le cas général d'une matrice carrée d'ordre 2 :

Soit $\mathbf{A} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$. Le **déterminant** de $\mathbf{A}$ est donné par :

$$\text{dét}\,(\mathbf{A}) = a \cdot d - b \cdot c.$$

Si ce dernier est non nul, on définit :

$$\mathbf{A}^{-1} = \frac{1}{a \cdot d - b \cdot c} \cdot \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}.$$

On peut alors vérifier que $\mathbf{A} \cdot \mathbf{A}^{-1} = \mathbf{A}^{-1} \cdot \mathbf{A} = \mathbf{I_2}$ par calcul direct.
Soit $\mathbf{B}$ une matrice carrée d'ordre 3 :

$$\mathbf{B} = \begin{bmatrix} 3 & 2 & 1 \\ 4 & -1 & 0 \\ 2 & -2 & 0 \end{bmatrix},$$

proposons-nous de déterminer la matrice inverse, $\mathbf{B}^{-1}$, de $\mathbf{B}$ si elle existe.

Calculons donc en premier le déterminant de $\mathbf{B}$ en développant selon la dernière colonne :

$$\begin{aligned} \text{dét}\,(\mathbf{B}) &= 1 \cdot [4 \cdot (-2) - (-1) \cdot 2] \\ \text{dét}\,(\mathbf{B}) &= -8 + 2 = -6. \end{aligned}$$

Ce dernier étant non nul, on peut calculer $\mathbf{B}^{-1}$, en déterminant les D_{ij}, c'est-à-dire les déterminants d'ordre 2 obtenus en biffant la i-ème ligne et la j-ème colonne de la **transposée** $\mathbf{B}'$ de $\mathbf{B}$:

$$\mathbf{B}' = \begin{bmatrix} 3 & 4 & 2 \\ 2 & -1 & -2 \\ 1 & 0 & 0 \end{bmatrix}$$

$$\begin{aligned} D_{11} &= -1 \cdot 0 - (-2) \cdot 0 = 0, \\ D_{12} &= 2 \cdot 0 - (-2) \cdot 1 = 2, \\ D_{13} &= 2 \cdot 0 - (-1) \cdot 1 = 1, \\ &\vdots \\ D_{33} &= 3 \cdot (-1) - 4 \cdot 2 = -11. \end{aligned}$$

On obtient alors la matrice inverse $\mathbf{B}^{-1}$ en adjugeant le signe correct à l'élément D_{ij} (c'est-à-dire $(-)$ si la somme de i et j est impaire et $(+)$ si elle est paire).

$$\begin{aligned} \mathbf{B}^{-1} &= \frac{1}{-6} \cdot \begin{bmatrix} 0 & -2 & 1 \\ 0 & -2 & 4 \\ -6 & 10 & -11 \end{bmatrix} \\ &= \frac{1}{6} \cdot \begin{bmatrix} 0 & 2 & -1 \\ 0 & 2 & -4 \\ 6 & -10 & 11 \end{bmatrix}. \end{aligned}$$

MOTS CLÉS

Déterminant *(Determinant)*
Inverse généralisé *(Generalized inverse)*
Matrice *(Matrix)*
Régression linéaire multiple *(Multiple linear regression)*
Transposée *(Transpose)*

MAXIMUM DE VRAISEMBLANCE

Maximum likelihood

Le terme maximum de vraisemblance se réfère à une méthode d'**estimation** de paramètres d'une **population** à partir d'un **échantillon** aléatoire.

Elle s'applique quand on connaît la forme générale de distribution de la population mais quand un ou plusieurs paramètres de cette distribution sont inconnus.

La méthode consiste à choisir comme estimateurs des paramètres inconnus, les valeurs qui maximisent la **probabilité** d'avoir obtenu l'échantillon observé.

HISTORIQUE

C'est généralement à **Ronald Aylmer Fisher** qu'est attribuée la méthode du maximum de vraisemblance. Cependant, pour retrouver la trace de cette méthode, il faut remonter jusqu'au XVIII^{e} siècle avec les écrits de J.-H. Lambert et D. Bernoulli.

R.A. Fisher introduisit la méthode du maximum de vraisemblance dans sa première publication statistique en 1912.

ASPECTS MATHÉMATIQUES

Soit un **échantillon** de n variables aléatoires X_1, X_2, ..., X_n, où chaque variable aléatoire est distribuée selon une **loi de probabilité** ayant une **fonction de densité** $f(x, \theta)$, avec $\theta \in \Omega$, où Ω est l'espace des paramètres.

La **fonction de densité conjointe** de X_1, X_2, ..., X_n est $f(x_1, \theta) \cdot f(x_2, \theta) \cdot \ldots \cdot f(x_n, \theta)$.

Pour x fixé, cette fonction est une fonction de θ et est appelée fonction de vraisemblance L :

$$L(\theta\,;\, x_1, x_2, \ldots, x_n) = f(x_1, \theta) \cdot f(x_2, \theta) \cdot \ldots \cdot f(x_n, \theta)\,.$$

On appelle **estimateur** du maximum de vraisemblance tout estimateur $\hat{\theta}$ qui maximise la fonction de vraisemblance.

On note que cette fonction L peut être maximisée en annulant les premières dérivées partielles par rapport à θ et en résolvant les équations ainsi obtenues.

De plus, chacune des fonctions L et $\ln L$ étant maximales pour la même valeur de θ, il est souvent plus facile de résoudre :

$$\frac{\partial \ln L(\theta\,;\, x_1, x_2, \ldots, x_n)}{\partial \theta} = 0.$$

Finalement, il faut mentionner que les estimateurs du maximum de vraisemblance ne sont pas forcément uniques et que cette méthode ne fournit pas forcément des estimateurs non biaisés.

EXEMPLE

Soit X_1, X_2, ..., X_n un **échantillon** aléatoire provenant d'une distribution normale de **moyenne** μ et de **variance** σ^2. La méthode d'**estimation** du maximum de vraisemblance consiste à choisir comme estimateurs des paramètres inconnus μ et σ^2, les valeurs qui maximisent la **probabilité** d'avoir obtenu l'échantillon observé.

Cela consiste donc à maximiser la fonction de vraisemblance :

$$L\left(\mu, \sigma^2\,;\, x_1 \ldots, x_n\right) = \frac{1}{\left(\sigma\sqrt{2\pi}\right)^n} \exp\left(\frac{-\sum_{i=1}^{n} (x_i - \mu)^2}{2\sigma^2}\right)$$

ou, ce qui revient au même, son logarithme naturel :

$$\begin{aligned}
\ln L &= \ln\left[\frac{1}{\left(\sigma\sqrt{2\pi}\right)^n} \exp\left(\frac{-\sum_{i=1}^{n} (x_i - \mu)^2}{2\sigma^2}\right)\right] \\
&= \ln\left(\sigma\sqrt{2\pi}\right)^{-n} - \frac{\sum_{i=1}^{n} (x_i - \mu)^2}{2\sigma^2} \\
&= \ln(2\pi)^{-\frac{n}{2}} + ln(\sigma)^{-n} - \frac{\sum_{i=1}^{n} (x_i - \mu)^2}{2\sigma^2} \\
&= -\frac{n}{2} \cdot \ln(2\pi) - n \cdot \ln(\sigma) - \frac{\sum_{i=1}^{n} (x_i - \mu)^2}{2\sigma^2}.
\end{aligned}$$

On annule à présent les dérivées partielles $\dfrac{\partial \ln L}{\partial \mu}$ et $\dfrac{\partial \ln L}{\partial \sigma^2}$:

$$\frac{\partial \ln L}{\partial \mu} = \frac{1}{\sigma^2} \cdot \sum_{i=1}^{n} (x_i - \mu) = 0 \qquad (1)$$

$$\frac{\partial \ln L}{\partial \sigma^2} = -\frac{n}{\sigma} + \frac{1}{\sigma \cdot \sigma^2} \cdot \sum_{i=1}^{n} (x_i - \mu)^2 = 0 \qquad (2)$$

De l'équation (1), on déduit :

$$\begin{aligned}
\frac{1}{\sigma^2} \cdot \sum_{i=1}^{n} (x_i - \mu) &= 0 \\
\sum_{i=1}^{n} (x_i - \mu) &= 0 \\
\sum_{i=1}^{n} x_i - n \cdot \mu &= 0 \\
\hat{\mu} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i &= \bar{x}
\end{aligned}$$

Ainsi, pour la **loi normale**, la **moyenne arithmétique** est l'estimateur du maximum de vraisemblance.

De l'équation (2), on obtient :

$$\begin{aligned}
-\frac{n}{\sigma} + \frac{1}{\sigma \cdot \sigma^2} \cdot \sum_{i=1}^{n} (x_i - \mu)^2 &= 0 \\
\frac{1}{\sigma \cdot \sigma^2} \cdot \sum_{i=1}^{n} (x_i - \mu)^2 &= \frac{n}{\sigma} \\
\hat{\sigma}^2 &= \frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - \mu)^2 .
\end{aligned}$$

On trouve dans ce cas comme estimateur du maximum de vraisemblance de la variance, la variance empirique de l'échantillon.
Dans ce cas, on trouve la variance empirique de l'échantillon comme estimateur du maximum de vraisemblance de la variance.

On note que l'estimation du maximum de vraisemblance pour la **moyenne** est sans **biais**, alors que celui de la variance est biaisé (division par n au lieu de $n-1$).

On trouve donc que, pour la loi normale, la moyenne et la variance de l'échantillon sont les estimateurs du maximum de vraisemblance de la moyenne et de la variance de la **population**.

Cette correspondance n'est pas nécessairement valable pour d'autres distributions : on considère par exemple, le cas de la **loi uniforme** sur l'intervalle $[a, b]$, où a et b sont inconnus et où le paramètre à estimer est la moyenne de la distribution $\mu = (a+b)/2$. On peut vérifier que l'estimateur du maximum de vraisemblance de μ est égal à :

$$\hat{\mu} = \frac{x_1 + x_n}{2}$$

où x_1 et x_n sont respectivement l'**observation** la plus petite et l'observation la plus grande de l'échantillon $(x_1, x_2, \ldots, x_n)$. Il s'agit du point milieu de l'intervalle dont les bornes sont x_1 et x_n respectivement. On constate que, dans ce cas, l'estimateur du maximum de vraisemblance est différent de la moyenne de l'échantillon :

$$\hat{\mu} \neq \bar{x}.$$

MOTS CLÉS

Estimateur *(Estimator)*
Estimation *(Estimation)*
Estimation robuste *(Robust estimation)*
Test du rapport de vraisemblance *(Likelihood ratio test)*

RÉFÉRENCE

Fisher, R.A. (1912). On an absolute criterium for fitting frequency curves. *Messenger of Mathematics*, 41, pp. 155-160.

MÉDIANE

Median

La médiane est une **mesure de tendance centrale** définie comme la **valeur** qui se trouve au centre d'un ensemble d'observations lorsque celles-ci sont rangées par ordre croissant ou décroissant.

Soit une **variable aléatoire** quantitative X. La médiane M_d est définie telle que la **probabilité** qu'une observation soit inférieure ou égale à la médiane soit inférieure ou égale à $\frac{1}{2}$:

$$P(X \leq M_d) \leq \frac{1}{2} \text{ et } P(X \geq M_d) \leq \frac{1}{2}.$$

Nous trouvons donc 50 % des observations de chaque côté de la médiane :



Si nous avons un nombre impair d'observations, la médiane correspond donc à la valeur de l'observation du milieu.

Si, en revanche, nous avons un nombre pair d'observations, il n'existe pas une observation unique du milieu. La médiane sera donnée par la **moyenne arithmétique** des valeurs des deux observations du milieu.

HISTORIQUE

En 1748, Leonhard Euler et Johann Tobias Mayer ont proposé, d'une manière indépendante, une méthode qui consiste à diviser les observations d'un ensemble de données en deux parties égales.

ASPECTS MATHÉMATIQUES

Lorsque les observations nous sont données de façon individuelle, le processus de calcul de la médiane est simple :

1. Classer les n observations par ordre de grandeur.
2. Si le nombre d'observations est impair :
   - repérer l'observation du milieu $\frac{n+1}{2}$;
   - la médiane est égale à la **valeur** de l'observation du milieu.
3. Si le nombre d'observations est pair :
   - repérer les 2 observations du milieu $\frac{n}{2}$ et $\frac{n}{2}+1$;
   - la médiane est égale à la **moyenne arithmétique** des valeurs de ces deux observations.

Lorsque nous ne possédons pas les observations individuelles mais que celles-ci sont groupées en classes, la médiane se détermine de la manière suivante :

1. Déterminer la classe médiane (celle qui comprend la $\frac{n}{2}$-ème observation).
2. Calculer la médiane par la formule suivante :

$$M_d = L + \left[\frac{\frac{n}{2} - \sum f_{inf}}{f_{M_d}}\right] \cdot c$$

où

L	=	borne inférieure de la classe médiane,
n	=	nombre total d'observations,
$\sum f_{inf}$	=	somme des fréquences inférieures à la classe médiane,
f_{M_d}	=	fréquence de la classe médiane et
c	=	longueur de l'**intervalle** de la classe médiane.

Cette formule repose sur l'**hypothèse** selon laquelle les observations sont uniformément réparties à l'intérieur de chaque classe. Elle suppose en outre que les bornes inférieure et supérieure de la classe médiane sont définies et connues.

Propriété

La somme des écarts en valeur absolue entre chaque observation x_i d'un ensemble de données et une valeur α est minimale lorsque α est égale à la médiane :

$$\underset{\alpha}{\text{minimiser}} \sum_{i=1}^{n} |x_i - \alpha| \Rightarrow \alpha = M_d.$$

DOMAINES ET LIMITATIONS

Dans le calcul de la **moyenne arithmétique**, nous tenons compte de la **valeur** de chaque **observation**. Ainsi, une **observation aberrante** influence fortement la moyenne arithmétique.

En revanche, dans le calcul de la médiane, nous ne tenons compte que du rang des observations. Les observations aberrantes n'influencent donc pas la médiane.

Lorsque nous sommes en présence d'une **distribution de fréquences** fortement asymétrique, nous avons donc intérêt à choisir la médiane comme **mesure de tendance centrale** plutôt que la moyenne arithmétique, en vue de neutraliser l'effet des valeurs extrêmes.

EXEMPLE

La médiane de l'ensemble des cinq nombres 42, 25, 75, 40 et 20 est 40, car 40 est le nombre qui se trouve

au milieu lorsque ces nombres sont rangés dans un ordre croissant :

$$20 \quad 25 \quad \underbrace{40}_{médiane} \quad 42 \quad 75$$

Si nous rajoutons le nombre 62 à cet ensemble, le nombre d'observations est alors pair, et la médiane est égale à la **moyenne arithmétique** des deux observations du milieu :

$$20 \quad 25 \quad \underbrace{40 \quad 42}_{\frac{40+42}{2}=41=médiane} \quad 62 \quad 75$$

Considérons à présent la médiane de données groupées en classes. Prenons comme exemple le **tableau de fréquences** suivant qui représente les revenus journaliers de cinquante épiceries :

Classes (revenus en euros)	Fréquences f_i (nombre d'épiceries)	Fréquences cumulées
$500-550$	3	3
$550-600$	12	15
$600-650$	17	32
$650-700$	8	40
$700-750$	6	46
$750-800$	4	50
Total	50	

La classe médiane est la classe $600-650$ puisque celle-ci contient l'observation du milieu (la 25^{e} observation).

En supposant que les observations soient réparties uniformément dans chaque classe, la médiane est égale à :

$$M_d = 600 + \frac{\frac{50}{2} - 15}{17} \cdot (650 - 600) = 629,41.$$

Ce résultat signifie que 50 % des épiceries réalisent des revenus journaliers supérieurs à $629,41$ francs, et les autres 50 % des revenus inférieurs à $629,41$ francs.

MOTS CLÉS

Mesure de tendance centrale *(Measure of central tendency)*
Mode *(Mode)*
Moyenne *(Mean)*
Quantile *(Quantile)*

RÉFÉRENCES

Boscovich, R.J. (1757). *De Litteraria Expeditione per Pontificiam ditionem, et Synopsis amplioris Operis, ac habentur plura eius ex exemplaria etiam sensorum impressa.* Bononiensi Scientiarium et Artium Instituto Atque Academia Commentarii, Tomus IV, pp. 353-396.

Euler, L. (1749). *Recherches sur la question des inégalités du mouvement de Saturne et de Jupiter, pièce ayant remporté le prix de l'année 1748, par l'Académie royale des sciences de Paris.* Réédité en 1960 in Leonhardi Euleri, *Opera Omnia*, 2^{e} série, 25, pp. 47-157. Turici, Bâle.

Mayer, T. (1750). Abhandlung über die Umwalzung des Monds um seine Axe und die scheinbare Bewegung der Mondflecken. *Kosmographische Nachrichten und Sammlungen auf das Jahr 1748*, 1, pp. 52-183.

MESURE D'APLATISSEMENT

Measure of kurtosis

Les mesures d'aplatissement font partie des mesures de forme et caractérisent un aspect de la forme d'une distribution donnée. Plus précisément, elles caractérisent le degré d'aplatissement de la distribution par rapport à l'aplatissement de la distribution normale. En effet, certaines distributions se rapprochent de la **loi normale** sans toutefois y être totalement identiques. Il est alors utile de pouvoir tester si la forme de la distribution présente une déviation par rapport à l'aplatissement de la distribution normale.

On parle de distribution platicurtique si la courbe est plus aplatie que la courbe normale et de distribution leptocurtique si la courbe est plus pointue que la courbe normale.

Pour tester l'aplatissement d'une courbe, on utilise le **coefficient β_2 de Pearson**.

MOTS CLÉS

Coefficient β_2 de Pearson *(Coefficient of kurtosis)*
Mesure d'asymétrie *(Measure of skewness)*
Mesure de forme *(Measure of shape)*

MESURE D'ASYMÉTRIE

Measure of skewness

Dans une distribution symétrique, la **médiane**, la **moyenne arithmétique** et le **mode** se confondent en un même point central.

Cet équilibre n'existe plus lorsqu'il y a asymétrie. Dans ce cas, le mode est séparé de la moyenne arithmétique, et la médiane se trouve entre les deux. Par conséquent, il est nécessaire de développer des mesures d'asymétrie pour appréhender le degré de déviation de la forme de la distribution par rapport à une distribution symétrique. Les principales mesures d'asymétrie sont :

— le **coefficient de Yule et Kendall** ;

— le **coefficient β_1 de Pearson**.

MOTS CLÉS

Coefficient β_1 de Pearson *(Coefficient of skewness)*
Coefficient de Yule et Kendall *(Yule and Kendall coefficient)*
Mesure d'aplatissement *(Measure of kurtosis)*
Mesure de forme *(Measure of shape)*

MESURE DE DISPERSION

Measure of dispersion

Une mesure de dispersion permet de décrire un ensemble de données concernant une **variable** particulière, en fournissant une indication sur la variabilité des **valeurs** au sein de l'ensemble des données.

La mesure de dispersion complète la description fournie par la **mesure de tendance centrale** d'une distribution.

Si nous observons différentes distributions, nous pouvons constater que, pour certaines, toutes les données sont groupées à une **distance** plus ou moins faible de la valeur centrale, alors que, pour d'autres, l'étalement des données est nettement plus grand.

Nous pouvons classer les mesures de dispersion en deux groupes :

1. Mesures définies par la distance entre deux valeurs représentatives de la distribution :
   - l'**empan**, appelé aussi étendue ou **intervalle** de variation ;
   - l'**intervalle interquartile**.
2. Mesures calculées en fonction des déviations de chaque donnée par rapport à une valeur centrale :
   - l'**écart géométrique** ;
   - l'**écart médian** ;
   - l'**écart moyen** ;
   - l'**écart type**.

Parmi les mesures de dispersion, la plus importante et la plus utilisée est l'écart type.

DOMAINES ET LIMITATIONS

Au-delà de l'aspect descriptif permettant de caractériser une distribution ou comparer plusieurs distributions, la connaissance de la dispersion ou de la variabilité d'une distribution peut avoir un intérêt pratique considérable. Qu'arriverait-il par exemple si nos décisions n'étaient basées que sur la **moyenne** ?

- Nos routes seraient construites pour absorber uniquement le trafic moyen et les embouteillages des retours de vacances seraient incommensurables.
- Les immeubles seraient construits pour résister à la force moyenne du vent avec les conséquences qui en résulteraient en cas de forte tempête.

Ces exemples montrent bien que, dans certains cas, disposer d'une **mesure de tendance centrale** n'est

pas suffisant pour pouvoir valablement prendre une décision.

Par ailleurs, la mesure de dispersion apporte elle-même une indication sur le caractère de la mesure de tendance centrale calculée. En effet, plus la variabilité d'un ensemble de données est petite, plus la valeur de la mesure de tendance centrale sera représentative de l'ensemble des données.

MOTS CLÉS

Coefficient de variation *(Coefficient of variation)*
Écart moyen *(Mean deviation)*
Écart type *(Standard deviation)*
Empan*(Range)*
Intervalle interquartile *(Interquartile range)*
Mesure de tendance centrale *(Measure of central tendency)*
Variance *(Variance)*

MESURE DE DISSIMILARITÉ

Measure of dissimilarity

On appelle mesure de dissimilarité une mesure de ressemblance, qui, à un couple d'objets de la **population** étudiée, fait correspondre un nombre réel positif (ou nul).

Si l'on suppose, par exemple, que chaque objet est caractérisé par n variables, une mesure de dissimilarité entre deux objets consisterait à donner le nombre de points différents que présentent deux objets considérés. Ce nombre est un entier compris entre 0 et n, et il est égal à n lorsque les deux objets n'ont aucun point commun.

ASPECTS MATHÉMATIQUES

Soit Ω une **population** sur laquelle on aimerait définir une mesure de dissimilarité s'.

On appelle mesure de dissimilarité s' toute application de $\Omega \times \Omega$ dans les nombres réels positifs (ou nul), satisfaisant :

- pour tous w, w' dans Ω, $s'(w, w') = s'(w', w)$, c'est-à-dire que s' est symétrique ;
- pour tout w dans Ω, $s'(w, w) = 0$, c'est-à-dire que la mesure de dissimilarité entre un objet et lui-même est nulle.

DOMAINES ET LIMITATIONS

Comme une mesure de dissimilarité s' sur une **population** Ω finie est une application bornée, c'est-à-dire :

$$\max \{s'(w, w') \; ; \; w, w' \in \Omega\} = m < \infty,$$

on peut définir une **mesure de similarité** sur la population Ω par :

$$s(w, w') = m - s'(w, w').$$

Ces deux notions, à savoir mesure de similarité et mesure de dissimilarité, peuvent être considérées comme complémentaires.

On utilise les mesures de dissimilarité pour résoudre des problèmes de **classification**, au lieu d'une **distance** qui en est un cas particulier et qui possède la propriété dite de l'inégalité triangulaire.

EXEMPLE

On considère la **population** Ω formée de trois fleurs

$$\Omega = \{w_1, w_2, w_3\}$$

sur laquelle ont été mesurées les 3 variables suivantes :

— v_1 : couleur de la fleur ;

— v_2 : nombre de pétales ;

— v_3 : nombre de feuilles ;

où

v_1 se résume ici à une **variable dichotomique**, car on ne considère que des fleurs rouges (R) ou jaunes (J) et
v_2 et v_3 sont des variables aléatoires discrètes prenant leurs valeurs dans les nombres cardinaux.

On a observé :

		Variables		
		v_1	v_2	v_3
Ω	w_1	R	5	2
	w_2	J	4	2
	w_3	R	5	1

On détermine pour chaque couple de fleurs le nombre de points qui les distinguent. On constate alors qu'il s'agit d'une mesure de dissimilarité sur notre ensemble de trois fleurs.

$$s'(w_1, w_1) = s'(w_2, w_2) = s'(w_3, w_3) = 0$$

car une fleur ne se distingue pas d'elle-même ;

$$s'(w_1, w_2) = s'(w_2, w_1) = 2$$

car la couleur et le nombre de pétales sont différents chez w_1 et w_2 ;

$$s'(w_1, w_3) = s'(w_3, w_1) = 1$$

car seul le nombre de feuilles permet de distinguer w_1 de w_3 ;

$$s'(w_2, w_3) = s'(w_3, w_2) = 3$$

car w_2 et w_3 diffèrent en chacune des 3 caractéristiques étudiées.

À l'aide de toutes ces valeurs possibles, on peut déduire que s' satisfait les deux propriétés d'une mesure de dissimilarité.

Il s'agit du complémentaire du nombre de points communs pour un couple de fleurs, qui est une **mesure de similarité**.

MOTS CLÉS

Classification *(Classification)*
Distance *(Distance)*
Mesure de similarité *(Measure of similarity)*
Population *(Population)*

RÉFÉRENCES

Celeux, G., Diday, E., Govaert, G., Lechevallier, Y. et Ralambondrainy, H. (1989). *Classification automatique des données - aspects statistiques et informatiques.* Dunod, Paris.

Hand, D.J. (1981). *Discrimination and classification.* John Wiley & Sons, New York.

MESURE DE FORME

Measure of shape

On peut chercher à caractériser la forme d'une **distribution de fréquences** au moyen de coefficients appropriés. Certaines distributions de fréquences se rapprochent de la **loi normale** (« courbe en cloche »). Ces distributions peuvent aussi présenter une asymétrie ou un aplatissement par rapport à la courbe normale. Les deux mesures de forme sont :

— les **mesures d'asymétrie** ;

— les **mesures d'aplatissement**.

HISTORIQUE

Karl Pearson (1894,1895) a été le premier à tester les différences entre certaines distributions et la **loi normale**.

Il a démontré que les écarts par rapport à la courbe normale peuvent être caractérisés par les moments d'ordre 3 et 4 d'une distribution.

Avant 1890, Jørgen Pedersen Gram et Thorvald Nicolai Thiele, au Danemark, ont développé une théorie sur la symétrie des courbes de fréquences.

Après que K. Pearson ait publié son système élaboré et extrêmement intéressant (1894 et 1895), de nombreux articles furent publiés sur ce sujet.

MOTS CLÉS

Mesure d'aplatissement *(Measure of kurtosis)*
Mesure d'asymétrie *(Measure of skewness)*
Mesure de dispersion *(Measure of dispersion)*
Mesure de position *(Measure of location)*

RÉFÉRENCES

Pearson, K. (1894). 1948. Contributions to the Mathematical Theory of Evolution. I. pp. 1-40 in Karl Pearson, *Karl Pearson's Early Statistical Papers.*

Cambridge University Press, Cambridge. First published as: On the Dissection of Asymmetrical Frequency Curves, *Philosophical Transactions of the Royal Society of London.* Series A, Vol. 185, pp. 71-110.

Pearson, K. (1895). 1948. Contributions to the Mathematical Theory of Evolution. II: Skew Variation in Homogeneous Material. pp. 41-112 in Karl Pearson, *Karl Pearson's Early Statistical Papers.* Cambridge University Press, Cambridge. First published in *Philosophical Transactions of the Royal Society of London.* Series A, Vol. 186, pp. 343-414.

MESURE DE POSITION

Measure of location

Une mesure de position (ou de location) est une mesure qui se propose de synthétiser un ensemble de données statistiques par une seule **valeur** fixe, en faisant ressortir un point particulier de la **variable** étudiée.

Les mesures de position les plus fréquentes sont :

- les mesures de tendance centrale (**moyenne**, **mode**, **médiane**) qui tendent à déterminer le centre d'un ensemble de données ;
- les **quantiles** (**quartile**, **décile**, **centile**) qui ne tendent pas à déterminer le centre, mais une position particulière d'un sous-ensemble de données.

DOMAINES ET LIMITATIONS

Les mesures de position, de même que les mesures de dispersion et les mesures de forme, sont notamment utilisées en vue de comparer plusieurs ensembles de données ou plusieurs distributions de fréquences.

MOTS CLÉS

Mesure de tendance centrale *(Measure of central tendency)*
Quantile *(Quantile)*

MESURE DE SIMILARITÉ

Measure of similarity

On appelle mesure de similarité une mesure de ressemblance, qui, à un couple d'objets de la **population** étudiée, fait correspondre un nombre réel positif (ou nul).

Si l'on suppose, par exemple, que chaque objet est caractérisé par n variables, une mesure de similarité entre deux objets consisterait à donner le nombre de points communs aux deux objets considérés. Ce nombre est un entier compris entre 0 et n, et il est égal à n lorsque les deux objets sont identiques (ou confondus).

ASPECTS MATHÉMATIQUES

Soit Ω une **population** sur laquelle on aimerait définir une mesure de similarité s.

On appelle mesure de similarité s toute application de $\Omega \times \Omega$ dans les nombres réels positifs (ou nul), satisfaisant :

- pour tous w, w' dans Ω, $s(w, w') = s(w', w)$, c'est-à-dire que s est symétrique ;
- pour tous w, w' dans Ω, avec $w \neq w'$

$$s(w, w) = s(w', w') > s(w, w'),$$

c'est-à-dire qu'un objet est plus « semblable » à lui-même qu' à tout autre objet distinct.

DOMAINES ET LIMITATIONS

Comme une mesure de similarité s sur une **population** Ω est une application bornée, c'est-à-dire :

$$\max\{s(w, w') \; ; \; w, w' \in \Omega\} = s(w, w) < \infty,$$

où w est un objet quelconque de Ω, si on pose $m = s(w, w)$, on peut définir une **mesure de dissimilarité** sur la population Ω par :

$$s'(w, w') = m - s(w, w').$$

Ces deux notions, à savoir mesure de similarité et mesure de dissimilarité, peuvent être considérées comme complémentaires.

On utilise les mesures de similarité, de même que les mesures de dissimilarité pour des problèmes de **classification**.

EXEMPLE

On considérons la **population** Ω formée de trois fleurs

$$\Omega = \{w_1, w_2, w_3\}$$

sur laquelle ont été mesurées les 3 variables suivantes :

— v_1 : couleur de la fleur ;

— v_2 : nombre de pétales ;

— v_3 : nombre de feuilles ;

où

v_1 se résume ici à une **variable dichotomique**, car on ne considère que des fleurs rouges (R) ou jaunes (J) et
v_2 et v_3 sont des variables aléatoires discrètes prenant leurs valeurs dans les nombres cardinaux.

On a observé :

		Variables		
		v_1	v_2	v_3
Ω	w_1	R	5	2
	w_2	J	4	2
	w_3	R	5	1

On détermine pour chaque couple de fleurs le nombre de points qu'elles ont en commun. On voit alors qu'il s'agit d'une mesure de similarité sur notre ensemble de trois fleurs.

$$s(w_1, w_1) = s(w_2, w_2) = s(w_3, w_3) = 3$$

qui correspond au nombre de variables ;

$$s(w_1, w_2) = s(w_2, w_1) = 1$$

car seul le nombre de feuilles est identique pour les deux fleurs w_1 et w_2 ;

$$s(w_1, w_3) = s(w_3, w_1) = 2$$

car les fleurs w_1 et w_3 ont la même couleur et le même nombre de pétales ;

$$s(w_2, w_3) = s(w_3, w_2) = 0$$

car w_2 et w_3 n'ont aucun point commun.

À l'aide de toutes ces valeurs possibles, on peut déduire que s satisfait les deux propriétés d'une mesure de similarité.

Il s'agit du complémentaire de la **mesure de dissimilarité** donnée par le nombre de points permettant de distinguer deux fleurs.

MOTS CLÉS

Distance *(Distance)*
Mesure de dissimilarité *(Measure of dissimilarity)*
Population *(Population)*

RÉFÉRENCES

Voir **mesure de dissimilarité**.

MESURE DE TENDANCE CENTRALE

Measure of central tendency

Une mesure de tendance centrale permet de résumer un ensemble de données relatives à une **variable quantitative**.

Plus précisément, elle permet de déterminer une **valeur** fixe, appelée valeur centrale, autour de laquelle l'ensemble des données a tendance à se rassembler.

Les principales mesures de tendance centrale sont :

— la **moyenne arithmétique** ;

— la **médiane** ;

— le **mode**.

ASPECTS MATHÉMATIQUES

La **moyenne** μ, de même que la **médiane** M_d et le **mode** M_o sont des mesures du centre d'une distribution.

Dans une distribution symétrique, la moyenne, la médiane et le mode coïncident en une **valeur** commune. Cette valeur donne le centre de symétrie de la distribution. Pour une distribution symétrique et unimodale, nous avons donc :

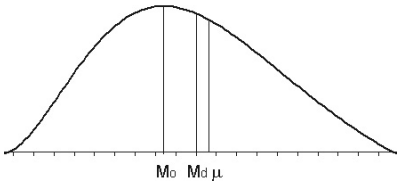
$$\mu = M_d = M_o.$$

Dans un tel cas, une **estimation** du centre de la distribution peut être donnée par la moyenne, la médiane ou le mode.

Dans une distribution asymétrique, les différentes mesures de tendance centrale ont des valeurs distinctes. (Leur comparaison peut être utilisée comme **mesure d'asymétrie**.)

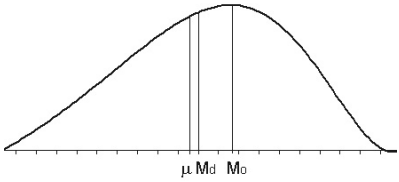
Pour une distribution unimodale et étirée à droite, nous avons généralement :

$$\mu \geq M_d \geq M_o.$$



Cette relation est inversée lorsque la distribution est unimodale et étirée à gauche :

$$\mu \leq M_d \leq M_o.$$



Pour une distribution unimodale, modérément asymétrique, il a été montré empiriquement que le mode M_o, la médiane M_d et la moyenne μ satisfont souvent à la relation suivante, de manière approximative :

$$M_o - \mu = 3 \cdot (M_d - \mu)\,.$$

DOMAINES ET LIMITATIONS

Dans la pratique, le choix de la mesure permettant de caractériser au mieux le centre d'un ensemble d'observations dépend des traits spécifiques de l'étude statistique et de l'information que l'on veut obtenir.

Comparons la **moyenne arithmétique**, le **mode** et la **médiane** en fonction de différents critères :

1. La *moyenne arithmétique* :
   - elle dépend de la **valeur** de toutes les observations ;
   - elle est simple à interpréter ;
   - c'est la mesure la plus familière et la plus utilisée ;
   - elle est fréquemment utilisée comme **estimateur** de la **moyenne** de la **population** ;
   - par contre, sa valeur peut être faussée par des observations aberrantes ;
   - la somme des écarts au carré entre chaque observation x_i d'un ensemble de données et une valeur α est minimale lorsque α est égale à la moyenne arithmétique :
   $$\min_{\alpha} \sum_{i=1}^{n} (x_i - \alpha)^2 \Rightarrow \alpha = \text{moy. arithm.}$$
2. La *médiane* :
   - elle est facile à déterminer puisqu'elle n'exige qu'un classement des données ;
   - elle est facile à comprendre, mais moins utilisée que la moyenne arithmétique ;
   - elle n'est pas influencée par des observations aberrantes, ce qui lui donne un avantage sur la moyenne arithmétique, si la série comporte effectivement des observations aberrantes ;

- elle est utilisée comme **estimateur** de la valeur centrale d'une distribution notamment lorsque celle-ci est asymétrique ou qu'elle comporte des observations aberrantes ;
- la somme des écarts en valeur absolue entre chaque observation x_i d'un ensemble de données et une valeur α est minimale lorsque α est égale à la médiane :

$$\min_{\alpha} \sum_{i=1}^{n} |x_i - \alpha| \Rightarrow \alpha = M_d.$$

3. Le *mode* :

- il a un intérêt pratique puisqu'il est la valeur la plus représentée dans un ensemble ;
- c'est toutefois une mesure peu utilisée ;
- il peut y avoir plusieurs (ou aucun) modes dans un ensemble de données ;
- sa valeur n'est pas trop influencée par des observations aberrantes ;
- sa valeur est fortement influencée par les fluctuations d'**échantillonnage**. Elle peut donc varier fortement d'un **échantillon** à l'autre.

MOTS CLÉS

Espérance mathématique *(Expected value)*
Médiane *(Median)*
Mesure d'asymétrie *(Measure of skewness)*
Mode *(Mode)*
Moyenne *(Mean)*
Moyenne arithmétique *(Arithmetic mean)*
Moyenne arithmétique pondérée *(Weighted arithmetic mean)*

RÉFÉRENCE

Stuart, A., Ord, J.K. (1943). *Kendall's advanced theory of statistics. Vol. I. Distribution theory.* John Wiley & Sons, New York.

MÉTHODE DE MONTE CARLO

Monte Carlo method

Toute technique numérique de résolution des problèmes mathématiques qui utilise des nombres aléatoires ou des nombres pseudo-aléatoires est appellée méthode de Monte Carlo.

HISTORIQUE

Le nom de Monte Carlo provient de la ville du même nom dans la principauté de Monaco, célèbre pour son casino. En effet, la roulette est l'un des mécanismes les plus simples pour la **génération de nombres aléatoires**.

Selon I.M. Sobol (1975), la méthode de Monte Carlo est due aux mathématiciens américains John von Neumann (1951), Stanislaw Marcin Ulam et N. Metropolis (1949). Elle a été développée vers 1949 ; cependant c'est avec l'avènement des ordinateurs qu'elle est devenue réalisable.

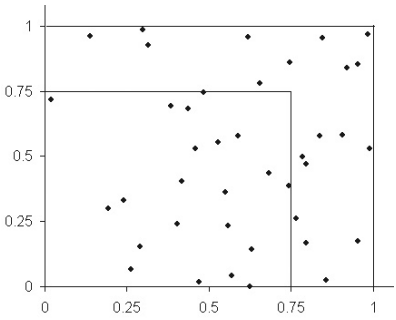
EXEMPLE

Supposons que nous devions calculer la surface d'une figure S qui se situe à l'intérieur d'un carré de côté égal à 1. Nous allons générer N points aléatoires à l'intérieur du carré. Pour ce faire, nous reportons le carré sur un système d'axes perpendiculaires dont l'origine est l'angle inférieur gauche du carré. Cela signifie que tous les points que l'on va générer à l'intérieur du carré auront des coordonnées comprises entre 0 et 1. Il suffit donc de prendre des variables aléatoires uniformes pour obtenir un point. En effet, le premier **nombre aléatoire** sera l'abscisse et le second l'ordonnée.

Dans l'exemple suivant, choisissons de disposer 40 points aléatoires, nous aurons alors besoin de deux échantillons de 40 nombres aléatoires. La surface S qu'on aimerait estimer est le carré de $0,75$ de côté, autrement dit la **valeur** exacte vaut $0,5625$.

Après avoir représenté les 40 points aléatoires

dans le carré unitaire, il suffit de compter le nombre de points qui se trouvent dans S (il y en a 21 dans le cas précis). L'**estimation** de la surface cherchée s'obtient par le ratio $\frac{21}{40} = 0,525$.



Il est possible d'améliorer cette estimation en répétant l'**expérience** plusieurs fois et en prenant la **moyenne** des surfaces obtenues.

MOTS CLÉS

Génération de nombres aléatoires *(Generation of random numbers)*
Méthode du jackknife *(Jackknife method)*
Nombre aléatoire *(Random number)*
Simulation *(Simulation)*

RÉFÉRENCES

Metropolis, N. and Ulam, S. (1949). The Monte Carlo method. *Journal of the American Statistical Association*, 44, pp. 335-341.

Sobol, I.M. (1975). *The Monte Carlo Method.* Mir Publishers, Moscow.

Von Neumann, J. (1951). Various Techniques Used in Connection with Random Digits. *National Bureau of Standards symposium (NBS) Applied Mathematics Series* 12, pp. 36-38, National Bureau of Standards, Washington, D.C.

MÉTHODE DES MOINDRES CARRÉS

Least squares method

Le principe des moindres carrés se réfère à une méthode d'**estimation** des paramètres d'un **modèle** stochastique.

La méthode des moindres carrés consiste à minimiser la somme des carrés des écarts entre les valeurs observées de la **variable dépendante** du modèle et les valeurs estimées par le modèle.

HISTORIQUE

La méthode des moindres carrés a été introduite par **Adrien-Marie Legendre** (1805).

ASPECTS MATHÉMATIQUES

Considérons le **modèle** général de **régression linéaire multiple** :

$$Y_i = \beta_0 + \sum_{j=1}^{p-1} \beta_j \cdot X_{ji} + \varepsilon_i,$$

$$i = 1, \ldots, n$$

où

Y_i	est la **variable dépendante**,
X_{ji},	$j = 1, \ldots, p-1$, sont les variables indépendantes,
β_j,	$j = 0, \ldots, p-1$, sont les paramètres à estimer et
ε_i	est le terme d'**erreur** aléatoire non observable.

L'**estimation** des paramètres de la fonction par la méthode des moindres carrés

$$f(\beta_0, \ldots, \beta_{p-1}) = \sum_{i=1}^{n} \varepsilon_i^2$$

consiste à minimiser la somme des carrés des erreurs :

$$\text{minimum } f(\beta_0, \ldots, \beta_{p-1}).$$

La somme des carrés des erreurs étant égale à :

$$\begin{aligned} f(\beta_0, \ldots, \beta_{p-1}) &= \sum_{i=1}^{n} \epsilon_i^2 \\ &= \sum_{i=1}^{n} \left(Y_i - \beta_0 - \sum_{j=1}^{p-1} \beta_j \cdot X_{ji} \right)^2, \end{aligned}$$

nous trouvons les paramètres estimés de la fonction en annulant les dérivées partielles de cette fonction par rapport à chaque paramètre :

$$\begin{aligned} \frac{\partial f(\beta_0, \ldots, \beta_{p-1})}{\beta_0} &= 0, \\ \frac{\partial f(\beta_0, \ldots, \beta_{p-1})}{\beta_1} &= 0, \\ &\vdots \\ \frac{\partial f(\beta_0, \ldots, \beta_{p-1})}{\beta_{p-1}} &= 0. \end{aligned}$$

Supposons que $p = 2$ et que nous disposions de n observations :

$$(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n).$$

L'équation liant les Y_i et les X_i peut être définie par le modèle linéaire

$$Y_i = \beta_0 + \beta_1 \cdot X_i + \varepsilon_i, \quad i = 1, \ldots, n.$$

La somme des carrés des écarts à la droite estimée de régression est :

$$f(\beta_0, \beta_1) = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 \cdot X_i)^2.$$

Mathématiquement, nous déterminons β_0 et β_1 en prenant les dérivées partielles de f d'abord par rapport à β_0, ensuite par rapport à β_1 et en les posant égales à zéro.

Dérivée partielle par rapport à β_0 :

$$\frac{\partial f}{\partial \beta_0} = -2 \cdot \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 \cdot X_i).$$

Dérivée partielle par rapport à β_1 :

$$\frac{\partial f}{\partial \beta_1} = -2 \cdot \sum_{i=1}^{n} X_i \cdot (y_i - \beta_0 - \beta_1 \cdot X_i).$$

Ainsi les **valeurs** estimées de β_0 et β_1, notées $\hat{\beta}_0$ et $\hat{\beta}_1$, sont données par les solutions des équations :

$$\begin{aligned} \sum_{i=1}^{n} \left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 \cdot X_i \right) &= 0, \\ \sum_{i=1}^{n} X_i \cdot \left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 \cdot X_i \right) &= 0. \end{aligned}$$

En développant ces deux équations, nous obtenons :

$$\sum_{i=1}^{n} Y_i = n \cdot \hat{\beta}_0 + \hat{\beta}_1 \cdot \sum_{i=1}^{n} X_i$$

et

$$\sum_{i=1}^{n} (X_i \cdot Y_i) = \hat{\beta}_0 \cdot \sum_{i=1}^{n} X_i + \hat{\beta}_1 \cdot \sum_{i=1}^{n} X_i^2$$

que l'on appelle **équations normales** de la droite des moindres carrés.

Sachant que :

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} \text{ et } \bar{Y} = \frac{\sum_{i=1}^{n} Y_i}{n},$$

la résolution des équations normales de la droite des moindres carrés nous donne les estimations des paramètres suivantes :

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \cdot \bar{X}$$

$$\begin{aligned} \hat{\beta}_1 &= \frac{n \cdot \sum_{i=1}^{n} (X_i \cdot Y_i) - (\sum_{i=1}^{n} X_i) \cdot (\sum_{i=1}^{n} Y_i)}{n \cdot \sum_{i=1}^{n} X_i^2 - (\sum_{i=1}^{n} X_i)^2} \\ &= \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (X_i - \bar{X})^2}. \end{aligned}$$

Les estimations par les moindres carrés de la variable dépendante sont donc :

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 \cdot X_i, \quad i = 1, \ldots, n$$

ou

$$\hat{Y}_i = \bar{Y} + \hat{\beta}_1 \cdot (X_i - \bar{X}), \quad i = 1, \ldots, n.$$

D'après cette équation, la droite des moindres carrés passe par le point $(\bar{X}, \bar{Y})$ appelé barycentre ou centre de gravité du nuage de points.

Nous pouvons également exprimer le modèle de régression linéaire multiple sous forme matricielle :

$$\mathbf{Y} = \mathbf{X} \cdot \boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

où

- $\mathbf{Y}$ est le **vecteur** $(n \times 1)$ des observations relatives à la variable dépendante (n observations),
- $\mathbf{X}$ est la **matrice** $(n \times p)$ du plan ayant trait aux variables indépendantes,
- $\boldsymbol{\varepsilon}$ est le vecteur $(n \times 1)$ des erreurs et
- $\boldsymbol{\beta}$ est le vecteur $(p \times 1)$ des paramètres à estimer.

L'application de la méthode des moindres carrés consiste à minimiser :

$$\begin{aligned} \boldsymbol{\varepsilon}' \cdot \boldsymbol{\varepsilon} &= (\mathbf{Y} - \mathbf{X} \cdot \boldsymbol{\beta})' \cdot (\mathbf{Y} - \mathbf{X} \cdot \boldsymbol{\beta}) \\ &= \mathbf{Y}' \cdot \mathbf{Y} - \boldsymbol{\beta}' \cdot \mathbf{X}' \cdot \mathbf{Y} - \mathbf{Y}' \cdot \mathbf{X} \cdot \boldsymbol{\beta} \\ &\quad + \boldsymbol{\beta}' \cdot \mathbf{X}' \cdot \mathbf{X} \cdot \boldsymbol{\beta} \\ &= \mathbf{Y}' \cdot \mathbf{Y} - 2 \cdot \boldsymbol{\beta}' \cdot \mathbf{X}' \cdot \mathbf{Y} + \boldsymbol{\beta}' \cdot \mathbf{X}' \cdot \mathbf{X} \cdot \boldsymbol{\beta} \end{aligned}$$

Les équations normales sont données matriciellement par :

$$\mathbf{X}' \cdot \mathbf{X} \cdot \hat{\boldsymbol{\beta}} = \mathbf{X}' \cdot \mathbf{Y}.$$

Si $\mathbf{X}' \cdot \mathbf{X}$ est inversible, les paramètres $\boldsymbol{\beta}$ sont donc estimés par :

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}' \cdot \mathbf{X})^{-1} \cdot \mathbf{X}' \cdot \mathbf{Y}.$$

EXEMPLE

Voir **régression linéaire simple**.

MOTS CLÉS

Estimation *(Estimation)*
Régression, analyse de *(Regression analysis)*
Régression linéaire multiple *(Multiple linear regression)*
Régression linéaire simple *(Simple linear regression)*

RÉFÉRENCE

Legendre, A.M. (1805). *Nouvelles méthodes pour la détermination des orbites des comètes*, Courcier, Paris.

MÉTHODE DES MOINDRES CARRÉS PONDÉRÉS

Weighted least squares method

La méthode des moindres carrés pondérés est utilisé lorsque la variance des erreurs n'est pas constante, c'est-à-dire lorsque l'hypothèse suivante de la méthode des moindres carrés classique est violée : la variance des erreurs est constante (égale à la valeur inconnue σ^2) quelle que soit l'observation i (c'est-à-dire quelles que soient les valeurs des x_{ij} concernées). Ainsi, à la place d'avoir, pour tout $i = 1, \ldots, n$, $Var(\varepsilon_i) = \sigma^2$ on a

$$Var(\varepsilon_i) = \sigma^2 w_i$$

où les *poids* (en anglais *weights*) $w_i > 0$ peuvent être différents pour chaque $i = 1, \ldots, n$.

ASPECTS MATHÉMATIQUES

Matriciellement, on a le modèle

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

où

- $\mathbf{Y}$ est le vecteur $(n \times 1)$ des observations relatives à la **variable dépendante** (n observations),
- $\boldsymbol{\beta}$ est le vecteur $(p \times 1)$ des paramètres à estimer,
- $\boldsymbol{\varepsilon}$ est le vecteur $(n \times 1)$ des erreurs,

et $\mathbf{X} = \begin{pmatrix} 1 & X_{11} & \ldots & X_{1(p-1)} \\ \vdots & \vdots & & \vdots \\ 1 & X_{n1} & \ldots & X_{n(p-1)} \end{pmatrix}$ est la **matrice** $(n \times p)$ du plan ayant trait aux variables indépendantes. De plus, on a :

$$Var(\boldsymbol{\varepsilon}) = \sigma^2 \mathbf{V}$$

où

$$\mathbf{V} = \begin{pmatrix} 1/w_1 & 0 & \cdots & 0 \\ 0 & 1/w_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1/w_n \end{pmatrix}.$$

En posant

$$\mathbf{Y}_w = \mathbf{WY}$$

$$\begin{aligned}\mathbf{X}_w &= \mathbf{W}\mathbf{X}\\ \boldsymbol{\varepsilon}_w &= \mathbf{W}\boldsymbol{\varepsilon},\end{aligned}$$

avec

$$\mathbf{W}=\begin{pmatrix}\sqrt{w_1} & 0 & \cdots & 0\\ 0 & \sqrt{w_2} & \cdots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \cdots & \sqrt{w_n}\end{pmatrix}$$

tel que $\mathbf{W}'\mathbf{W}=\mathbf{V}^{-1}$, on obtient le modèle équivalent

$$\mathbf{Y}_w = \mathbf{X}_w\boldsymbol{\beta}+\boldsymbol{\varepsilon}_w$$

où

$$\begin{aligned}Var\left(\boldsymbol{\varepsilon}_w\right) = Var\left(\mathbf{W}\boldsymbol{\varepsilon}\right) = \mathbf{W}Var\left(\boldsymbol{\varepsilon}\right)\mathbf{W}' =\\ = \sigma^2\cdot\mathbf{W}\mathbf{V}\mathbf{W} = \sigma^2\mathbf{I}_n\end{aligned}$$

où I_n est la **matrice identité** de dimension n. Comme pour ce nouveau modèle la variance des erreurs est constante, la méthode des moindres carrés est utilisée. Le vecteur des estimateurs est :

$$\begin{aligned}\hat{\boldsymbol{\beta}}_w &= \left(\mathbf{X}_w'\mathbf{X}_w\right)^{-1}\mathbf{X}_w'\mathbf{Y}_w\\ &= \left(\mathbf{X}'\mathbf{W}'\mathbf{W}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{W}'\mathbf{W}\mathbf{Y}\\ &= \left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{Y}\end{aligned}$$

Le vecteur des valeurs estimées pour $\mathbf{Y}_w = \mathbf{W}\mathbf{Y}$ devient

$$\begin{aligned}\widehat{\mathbf{Y}}_w &= \mathbf{X}_w\hat{\boldsymbol{\beta}}_w\\ &= \mathbf{W}\mathbf{X}\left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{Y}.\end{aligned}$$

On peut alors obtenir un vecteur de valeurs estimées pour $\mathbf{Y}=\mathbf{W}^{-1}\mathbf{Y}_w$ en posant

$$\begin{aligned}\widehat{\mathbf{Y}} &= \mathbf{W}^{-1}\widehat{\mathbf{Y}}_w\\ &= \mathbf{W}^{-1}\mathbf{W}\mathbf{X}\left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{Y}\\ &= \mathbf{X}\hat{\boldsymbol{\beta}}_w.\end{aligned}$$

La variance σ^2 est estimée par

$$\begin{aligned}s_w^2 &= \frac{\left(\mathbf{Y}_w-\widehat{\mathbf{Y}}_w\right)'\left(\mathbf{Y}_w-\widehat{\mathbf{Y}}_w\right)}{n-p}\\ &= \frac{\left(\mathbf{Y}-\widehat{\mathbf{Y}}\right)'\mathbf{W}'\mathbf{W}\left(\mathbf{Y}-\widehat{\mathbf{Y}}\right)}{n-p}\\ &= \frac{\sum_{i=1}^{n}w_i\left(y_i-\widehat{y}_i\right)^2}{n-p}.\end{aligned}$$

Dans le cas d'une régression simple, à partir d'un échantillon de n observations (x_i,y_i) et de poids w_i on a :

$$\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}=\begin{pmatrix}\sum_{i=1}^{n}w_i & \sum_{i=1}^{n}w_ix_i\\ \sum_{i=1}^{n}w_ix_i & \sum_{i=1}^{n}w_ix_i^2\end{pmatrix},$$

$$\mathbf{X}\mathbf{V}^{-1}\mathbf{Y}=\begin{pmatrix}\sum_{i=1}^{n}w_iy_i\\ \sum_{i=1}^{n}w_ix_iy_i\end{pmatrix}$$

et ainsi, si on note $\hat{\boldsymbol{\beta}}_w=\begin{pmatrix}\widehat{\beta}_{0w}\\ \widehat{\beta}_{1w}\end{pmatrix}$, on trouve :

$$\begin{aligned}\widehat{\beta}_{0w} &= \frac{\sum_{i=1}^{n}w_iy_i\sum_{i=1}^{n}w_ix_i^2-\sum_{i=1}^{n}w_ix_i\sum_{i=1}^{n}w_ix_iy_i}{\sum_{i=1}^{n}w_i\sum_{i=1}^{n}w_ix_i^2-\left(\sum_{i=1}^{n}w_ix_i\right)^2}\\ \widehat{\beta}_{1w} &= \frac{\sum_{i=1}^{n}w_i\sum_{i=1}^{n}w_ix_iy_i-\sum_{i=1}^{n}w_ix_i\sum_{i=1}^{n}w_iy_i}{\sum_{i=1}^{n}w_i\sum_{i=1}^{n}w_ix_i^2-\left(\sum_{i=1}^{n}w_ix_i\right)^2}.\end{aligned}$$

En considérant les moyennes pondérées

$$\begin{aligned}\overline{x}_w &= \frac{\sum_{i=1}^{n}w_ix_i}{\sum_{i=1}^{n}w_i}\\ \overline{y}_w &= \frac{\sum_{i=1}^{n}w_iy_i}{\sum_{i=1}^{n}w_i},\end{aligned}$$

on a les estimateurs

$$\begin{aligned}\widehat{\beta}_{1w} &= \frac{\sum_{i=1}^{n}w_ix_iy_i-\sum_{i=1}^{n}w_i\overline{x}_w\overline{y}_w}{\sum_{i=1}^{n}w_ix_i^2-\sum_{i=1}^{n}w_i\overline{x}_w^2} =\\ &= \frac{\sum_{i=1}^{n}w_i\left(x_i-\overline{x}_w\right)\left(y_i-\overline{y}_w\right)}{\sum_{i=1}^{n}w_i\left(x_i-\overline{x}_w\right)^2}\\ \widehat{\beta}_{0w} &= \overline{y}_w-\widehat{\beta}_{1w}\overline{x}_w\end{aligned}$$

qui ressemblent fortement aux estimateurs des moindres carrés, sauf que l'on donne plus de poids aux observations avec un large w_i (c'est-à-dire aux observations pour lesquelles les erreurs ont une variance plus faible dans le modèle initial). Les valeurs estimées sont alors données par

$$\widehat{y}_i = \widehat{\beta}_{0w} + \widehat{\beta}_{1w} x_i = \overline{y}_w + \widehat{\beta}_{1w} (x_i - \overline{x}_w)$$

et la variance σ^2 est estimée par

$$s_w^2 = \frac{\sum_{i=1}^{n} w_i (y_i - \widehat{y}_i)^2}{n-2}$$

$$= \frac{\sum_{i=1}^{n} w_i (y_i - \overline{y}_w)^2 - \widehat{\beta}_{1w}^2 \sum_{i=1}^{n} w_i (x_i - \overline{x}_w)^2}{n-2}.$$

Des exemples de choix de poids w_i en régression simple sont donnés par $w_i = 1/x_i$ dans les cas où la variance de ε_i est proportionnelle à x_i, ou alors $w_i = 1/x_i^2$ dans le cas où la variance de ε_i est proportionnelle à x_i^2.

DOMAINES ET LIMITATIONS

Le plus grand inconvénient de la méthode des moindres carrés pondérés, que beaucoup de personnes choisissent d'ignorer, est probablement le fait que la théorie sur laquelle cette méthode est fondée se base sur l'hypothèse que les poids sont connus exactement. Ceci n'est presque jamais le cas dans la pratique et donc des poids estimés sont utilisés à la place. En général, il est difficile d'évaluer l'effet de l'utilisation des poids estimés, mais l'expérience indique que les résultats de la plupart des analyses de régression ne sont pas très sensibles aux poids utilisés. Les avantages d'une analyse pondérée sont souvent obtenus à large échelle, bien que pas entièrement, avec des poids approximatifs. Cependant il est important de rester averti de ce problème, et d'utiliser seulement des poids qui peuvent être estimés avec précision relativement l'un à l'autre. La régression des moindres carrés pondérés, comme d'ailleurs les autres méthodes des moindres carrés, est également sensible aux valeurs aberrantes.

EXEMPLE

Soit un ensemble de données comprenant 10 observations avec comme variable explicative X, la suite composée des valeurs entières allant de 1 à 10. La variable Y est générée en utilisant le modèle

$$y_i = 3 + 2x_i + \varepsilon_i$$

où les ε_i sont normalement distribués avec $E(\varepsilon_i) = 0$ et $Var(\varepsilon_i) = (0, 2x_i)^2$. Ainsi, la variance de l'erreur de la première observation correspond à $[(1)(0,2)]^2 = 0,04$ alors que celle de la dixième observation correspond à $[(10)(0,2)]^2 = 4$. Les données ainsi générées sont présentées dans le tableau suivant :

Valeurs x_i et y_i générées par le modèle $y_i = 3 + 2x_i + \varepsilon_i$

i	x_i	y_i
1	1	4,90
2	2	6,55
3	3	8,67
4	4	12,59
5	5	17,38
6	6	13,81
7	7	14,60
8	8	32,46
9	9	18,73
10	10	20,27

On effectue tout d'abord une régression non pondérée. On obtient l'équation de régression :

$$\widehat{y}_i = 3,49 + 2,09 x_i.$$

$\widehat{\beta}_0 = 3,49$ avec $t_c = 0,99$
$\widehat{\beta}_1 = 2,09$ avec $t_c = 3,69$

Analyse de variance

Source de variation	Degré de liberté	Somme des carrés	Moyennes des carrés	F
Régression	1	360,68	360,68	13,59
Résiduelle	8	212,39	26,55	
Totale	9	573,07		

Le R^2 de cette régression est 62,94 %. À première vue, ces résultats paraissent bons. Le coefficient estimé associé à la variable explicative est clairement significatif. Ce résultat n'est guère surprenant car le

fait que la variance des erreurs n'est pas constante n'a que rarement un effet sur les coefficients estimés. Toutefois, le carré moyen des résidus est difficilement explicable compte tenu du fait qu'il n'y a pas une variance unique à estimer. De plus, les erreurs types et les largeurs des intervalles de confiance à 95 % pour les moyennes conditionnelles estimées sont relativement constantes pour toutes les observations. On devrait pourtant s'attendre à ce que les observations avec peu de variance soient estimées de façon plus précise. Ce fait implique que l'analyse non pondérée ne prend pas en considération les inégalités observées entre les variances des erreurs.

On effectue à présent une régression pondérée en utilisant comme pondération les valeurs $1/x^2$. Ces pondérations sont connues puisqu'elles doivent être proportionnelles aux vraies variances, en l'occurence égales à $(0,2x)^2$. Les résultats sont les suivants :

$$\widehat{y}_i = 2,53 + 2,28x_i$$

$\widehat{\beta}_0 = 2,53$ avec $t_c = 3,07$
$\widehat{\beta}_1 = 2,28$ avec $t_c = 7,04$

Analyse de variance

Source de variation	Degré de liberté	Somme des carrés	Moyennes des carrés	F
Régression	1	23, 29	23, 29	49, 6
Résiduelle	8	3, 75	0, 47	
Totale	9	27, 05		

Le R^2 de cette régression est 86, 11 %. Sur cette base, on peut formuler les commentaires suivants :

- toutes les quantités reliées à la somme des carrés de la variable dépendante sont affectées par les pondérations et ne sont pas comparables à celles obtenues par la régression non pondérée ;
- le test de Fisher et le R^2 peuvent être comparés entre les deux modèles. Dans notre exemple, les deux valeurs sont supérieures à celle de la régression non pondérée, mais cela n'est pas forcément le cas ;
- les coefficients estimés sont relativement proches de ceux de la régression non pondérée. C'est souvent le cas mais pas forcément une généralité ;
- les erreurs types et les intervalles de confiance pour les moyennes conditionnelles reflètent le fait que la précision des estimations décroît avec les grandes valeurs de x. Ce résultat est la raison primordiale d'utiliser la régression pondérée.

MOTS CLÉS

Régression, analyse de *(Regression analysis)*
Moindres carrés *(Least squares)*
Régression linéaire généralisée *(Generalized linear regression)*

RÉFÉRENCE

Seber, G.A.F. (1977) *Linear Regression Analysis,* John Wiley & Sons, New York.

MÉTHODE DES MOMENTS

Method of moments

La méthode des moments sert à estimer les paramètres d'une distribution à partir d'un **échantillon**.

L'idée est la suivante : si le nombre de paramètres à estimer est égal à k, on pose les k premiers moments de la loi (qui s'expriment en fonction des paramètres inconnus) égaux aux moments empiriques correspondants, c'est-à-dire aux estimateurs des moments d'ordre k calculés sur l'échantillon que l'on considère. Il s'agit donc de résoudre un système de k équations avec k inconnues.

HISTORIQUE

À la fin du XIX^{e} siècle, **Karl Pearson** a utilisé la méthode des moments pour estimer les paramètres d'un modèle normal mixte (avec paramètres p_1, p_2, μ_1, μ_2, σ_1^2, σ_2^2) donné par

$$P(x) = \sum_{i=1}^{2} p_i \cdot g_i(x),$$

où les g_i sont normalement distribués de moyenne μ_i

et variance σ_i^2. Ensuite, la méthode a été développée et étudiée plus en détail par Tchouproff, Thorvald Nicolaï Thiele, **Ronald Aylmer Fisher** et Karl Pearson parmi d'autres.

ASPECTS MATHÉMATIQUES

Soient $\theta_1, \ldots, \theta_k$ les k paramètres à estimer de la distribution d'une variable aléatoire X.

En ce qui concerne les moments empiriques, que l'on note avec m_k, on renvoie le lecteur au mot **moment** sous la section DOMAINES ET LIMITATIONS.

On note $\mu_j((\theta_1, \ldots, \theta_k)) = E\left[X^j\right]$ le moment (initial) d'ordre j de X exprimé en fonction des paramètres à estimer.

La méthode des moments consiste dans la résolution du système d'équations suivant :

$$\mu_j(\theta_1, \ldots, \theta_k) = m_j \text{ pour } j = 1, \ldots, k.$$

Pour distinguer entre les vrais paramètres et les estimateurs, on note ces derniers par $\hat{\theta}_1, \ldots, \hat{\theta}_k$.

EXEMPLE

1. Soit X une variable aléatoire suivant une **loi exponentielle** de paramètre inconnu θ. On dispose de n observations $x_1, \ldots, x_n$ de X et on désire estimer θ.

   Comme le moment initial d'ordre 1 est la moyenne de la loi, pour estimer θ il faut égaliser la moyenne théorique et la moyenne empirique. On sait (voir loi exponentielle) que la moyenne (théorique) est égale $\frac{1}{\theta}$.

   L'estimateur voulu $\hat{\theta}$ satisfait $\frac{1}{\theta} = \frac{1}{n}\sum_{i=1}^{n} x_i$.

2. Selon la méthode des moments, la **moyenne** μ et la **variance** σ^2 d'une variable aléatoire X suivant une **loi normale** sont estimées par la moyenne empirique $\bar{x}$ et par la variance suivante :

   $$\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

   d'un échantillon $\{x_1, x_2, \ldots, x_n\}$ de X.

DOMAINES ET LIMITATIONS

Il faut observer tout d'abord que le système d'équations donné auparavant peut ne pas avoir des solutions, par exemple si les paramètres à estimer sont soumis à des contraintes particulières, ou bien en avoir plus d'une. En tout cas, la méthode ne garantit pas des estimateurs particulièrement efficaces. L'estimateur trouvé pour la variance d'une loi normale est, par exemple, biaisé.

MOTS CLÉS

Distance du chi-carré *(Chi-square distance)*
Échantillon *(Sample)*
Estimateur *(Estimator)*
Maximum de vraisemblance *(Maximum likelihood)*
Moindres carrés *(Least squares)*
Moment *(Moment)*
Paramètre *(Parameter)*

RÉFÉRENCES

Fisher, R.A. (1929). Moments and product moments of sampling distributions. *Proceedings London Mathematical Society* 30, pp. 199-238.

Fisher, R.A. (1930). The Moments of the Distribution for Normal Samples of Measures of Departure from Normality. *Proceedings of the Royal Society of London.* Series A, 130, pp. 16-28.

Kendall, M.G. and Stuard, A. (1977). *The advanced Theory of Statistics*, Vol. 1 Distribution theory. 4th ed. Charles Griffin & Company Ltd.

Pearson, K. (1894). 1948. Contributions to the Mathematical Theory of Evolution. I. pp. 1-40 in Karl Pearson, *Karl Pearson's Early Statistical Papers.* Cambridge University Press, Cambridge. First published as: On the Dissection of Asymmetrical Frequency Curves, *Philosophical Transactions of the Royal Society of London.* Series A, Vol. 185, pp. 71-110.

Thiele, T.N. (1903). Theory of Observations. C. and E. Layton, London; *Annals of Mathematical Statistics*, 2, pp. 165-308.

MÉTHODE DU CHAÎNAGE COMPLET

Complete linkage method

La méthode du chaînage complet est une méthode de **classification** hiérarchique telle que :

- la **distance** utilisée pour construire le **tableau de distances** est la distance euclidienne ;
- la distance entre deux classes est donnée par la distance euclidienne entre leurs deux éléments les plus éloignés (distance maximum).

ASPECTS MATHÉMATIQUES

Voir **classification automatique**.

MOTS CLÉS

Classification *(Classification)*
Classification automatique *(Cluster analysis)*
Dendrogramme *(Dendrogram)*
Distance *(Distance)*
Méthode du chaînage simple *(Single link method)*
Tableau de distances *(Distance table)*

MÉTHODE DU CHAÎNAGE SIMPLE

Single link method

La méthode du chaînage simple est une méthode de **classification** hiérarchique telle que :

- la **distance** utilisée pour construire le **tableau de distances** est la distance euclidienne ;
- la distance entre deux classes est donnée par la distance euclidienne entre leurs deux éléments les plus proches (distance minimum).

ASPECTS MATHÉMATIQUES

Voir **classification automatique**.

MOTS CLÉS

Classification *(Classification)*
Classification automatique *(Cluster analysis)*
Dendrogramme *(Dendrogram)*
Distance *(Distance)*
Méthode du chaînage complet *(Complete linkage method)*
Tableau de distances *(Distance table)*

MÉTHODE DU JACKKNIFE

Jackknife method

Lors d'estimations ponctuelles ou d'estimations par **intervalle de confiance**, il est souvent difficile de déterminer le **biais** ou l'**erreur type** de l'**estimateur** utilisé. Les raisons se résument généralement en deux points :

- le manque d'information à propos de la forme de la distribution théorique des données ;
- la complexité du **modèle** théorique, les hypothèses considérées étant trop fortes.

La méthode du jackknife permet d'obtenir un biais réduit, et des estimations numériques de l'erreur type ainsi qu'un intervalle de confiance. Elle s'effectue en générant des sous-échantillons artificiels à partir des données de l'**échantillon** complet.

HISTORIQUE

La méthode du jackknife a été développée par Quenouille (1949, 1956) et **John Wilder Tukey** (1958). Elle est plus largement utilisée actuellement grâce aux ordinateurs qui peuvent générer un grand nombre de données en un temps très court.

La dénomination jackknife est due à J.W. Tukey (1958). Le nom de cette méthode est issu d'un couteau scout à plusieurs lames prêt à être utilisé dans un grand nombre de cas.

ASPECTS MATHÉMATIQUES

Soient x_1, ..., x_n, n observations indépendantes d'une **variable aléatoire**. On souhaite estimer un **paramètre** θ de la **population**. La technique du

jackknife est la suivante : l'**estimateur** $\hat{\theta}$ de θ est calculé à partir de l'**échantillon** initial. Ce dernier est partagé en J groupes avec $m = \frac{n}{J}$ observations dans chacun d'eux (pour simplifier le problème, on suppose que n est un multiple de J).

L'estimateur de θ par la méthode du jackknife est la **moyenne** des pseudo-valeurs :

$$\hat{\bar{\theta}} = \frac{1}{J} \cdot \sum_{j=1}^{J} \hat{\theta}_j$$

où $\hat{\theta}_j$ est appelée la *pseudo-valeur* et définie par :

$$\hat{\theta}_j = J \cdot \hat{\theta} - (J-1) \cdot \hat{\theta}_{(j)}$$

et les estimateurs $\hat{\theta}_{(j)}$ $(j = 1, \ldots, J)$ sont calculés de la même façon que $\hat{\theta}$, mais en enlevant le j-ème groupe.

L'**estimation** de sa **variance** est définie par :

$$S^2 = Var(\hat{\bar{\theta}}) = \sum_{j=1}^{J} \frac{(\hat{\theta}_j - \hat{\bar{\theta}})^2}{(J-1) \cdot J}.$$

L'estimateur du **biais** est :

$$B(\hat{\bar{\theta}}) = (J-1) \cdot (\hat{\theta}_{(.)} - \hat{\theta})$$

où $\hat{\theta}_{(.)}$ est défini comme suit :

$$\hat{\theta}_{(.)} = \frac{1}{J} \cdot \sum_{j=1}^{J} \hat{\theta}_{(j)}.$$

L'**intervalle de confiance** au **seuil de signification** α est alors donné par :

$$[\hat{\bar{\theta}} - \frac{S}{\sqrt{n}} \cdot t_{n-1,1-\frac{\alpha}{2}} \; ; \; \hat{\bar{\theta}} + \frac{S}{\sqrt{n}} \cdot t_{n-1,1-\frac{\alpha}{2}}]$$

Une des propriétés importantes du jackknife est le fait que l'on enlève du biais un terme d'ordre $\frac{1}{n}$.

On suppose que $a_1, a_2, \ldots$ soient des fonctions de θ, on peut poser comme **espérance mathématique** de $\hat{\theta}$:

$$\begin{aligned} E[\hat{\theta}] &= \theta + \frac{a_1}{n} + \frac{a_2}{n^2} + \ldots \\ &= \theta + \frac{a_1}{m \cdot J} + \frac{a_2}{(m \cdot J)^2} + \ldots \end{aligned}$$

et

$$\begin{aligned} E[\hat{\theta}_j] &= J \cdot E[\hat{\theta}] - (J-1) \cdot E[\hat{\theta}_{(j)}] \\ &= J \cdot \left[\theta + \frac{a_1}{m \cdot J} + \frac{a_2}{(m \cdot J)^2} + \ldots\right] \\ &\quad -(J-1) \cdot \left[\theta + \frac{a_1}{m \cdot (J-1)} + \ldots\right] \\ &= J \cdot \theta + \frac{a_1}{m} + \frac{a_2}{m^2 \cdot J} + \ldots \\ &\quad -J \cdot \theta + \theta - \frac{a_1}{m} - \frac{a_2}{m^2 \cdot (J-1)} - \ldots \\ &= \theta - a_2 \cdot \frac{1}{m^2 \cdot J \cdot (J-1)} + \ldots \end{aligned}$$

DOMAINES ET LIMITATIONS

L'idée de la méthode du jackknife, relativement ancienne, est devenue actuelle grâce à l'informatique. Elle génère de manière artificielle des données pour déterminer l'**intervalle de confiance** à un **seuil de signification** choisi d'une **statistique** inconnue.

Dans les cas pratiques, les groupes ne sont composés que d'une seule **observation** $(m = 1)$. Les estimateurs $\hat{\theta}_{(j)}$ sont calculés en enlevant la j-ème observation.

EXEMPLE

Une grande chaîne de magasins souhaite déterminer l'évolution de son chiffre d'affaires entre 1950 et 1960.

Pour obtenir une **estimation** correcte, on a tiré un **échantillon** de huit points de vente dont on connaît les chiffres d'affaires respectifs.

Les données sont résumées dans le tableau suivant :

Point de vente	Chiffre d'affaires (milliards de $)	
	1950	1960
1	6,03	7,59
2	12,42	15,65
3	11,89	14,95
4	3,12	5,03
5	8,17	11,08
6	13,34	18,48
7	5,30	7,98
8	4,23	5,61

L'estimation doit reposer sur les observations de l'échantillon. Le chiffre d'affaires total des 8 points de ventes est passé de $64,5$ en 1950 à $86,37$ milliards de $ en 1960.

Ainsi, l'estimation du facteur d'accroissement vaut :

$$\hat{\theta} = \frac{86,37}{64,5} = 1,34.$$

On souhaite calculer le **biais** de cette estimation ainsi que son **écart type**.

On calcule les $\hat{\theta}_{(j)}$ pour j allant de 1 à 8, c'est-à-dire les facteurs d'accroissement calculés sans les données du j-ème point de vente. On obtient :

$$\hat{\theta}_{(1)} = \frac{86,37 - 7,59}{64,5 - 6,03} = 1,347,$$

$$\begin{array}{ll} \hat{\theta}_{(2)} = 1,358 & \hat{\theta}_{(6)} = 1,327 \\ \hat{\theta}_{(3)} = 1,358 & \hat{\theta}_{(7)} = 1,324 \\ \hat{\theta}_{(4)} = 1,325 & \hat{\theta}_{(8)} = 1,340 \\ \hat{\theta}_{(5)} = 1,337 & \end{array}$$

On peut alors calculer $\hat{\theta}_{(.)}$:

$$\hat{\theta}_{(.)} = \frac{1}{J} \cdot \sum_{j=1}^{J} \hat{\theta}_{(j)} = \frac{1}{8} \cdot (10,716) = 1,339.$$

On calcule les pseudo-valeurs $\hat{\theta}_j$. Rappelons que $\hat{\theta}_j$ se calcule comme suit :

$$\hat{\theta}_j = J \cdot \hat{\theta} - (J-1) \cdot \hat{\theta}_{(j)}.$$

On trouve :

$$\begin{array}{ll} \hat{\theta}_1 = 1,281 & \hat{\theta}_5 = 1,356 \\ \hat{\theta}_2 = 1,207 & \hat{\theta}_6 = 1,423 \\ \hat{\theta}_3 = 1,210 & \hat{\theta}_7 = 1,443 \\ \hat{\theta}_4 = 1,436 & \hat{\theta}_8 = 1,333. \end{array}$$

On calcule l'**estimateur** de θ par le jackknife :

$$\hat{\tilde{\theta}} = \frac{1}{J} \cdot \sum_{j=1}^{J} \hat{\theta}_j = \frac{1}{8} \cdot (10,689) = 1,336.$$

On détermine alors le biais :

$$\begin{aligned} B(\hat{\tilde{\theta}}) &= (J-1) \cdot (\hat{\theta}_{(.)} - \hat{\theta}) \\ &= (8-1) \cdot (1,339 - 1,34) = -0,0035. \end{aligned}$$

L'**erreur type** est donnée par :

$$\begin{aligned} S &= \sqrt{Var(\hat{\tilde{\theta}})} \\ &= \sqrt{\sum_{j=1}^{J} \frac{(\hat{\theta}_j - \hat{\tilde{\theta}})^2}{(J-1) \cdot J}} \\ &= \sqrt{\frac{0,0649}{7 \cdot 8}} = 0,034. \end{aligned}$$

On peut donc calculer l'**intervalle de confiance** du facteur d'accroissement entre 1950 et 1960 du chiffre d'affaires de la chaîne. Pour un **seuil de signification** α de 5 %, la **valeur** critique de la **table de Student** pour 7 degrés de liberté est de :

$$t_{7,0,975} = 2,365.$$

L'intervalle de confiance pour l'estimation du facteur d'accroissement est donc :

$$[\hat{\tilde{\theta}} - \frac{S}{\sqrt{n}} \cdot t_{7,0,975} \,;\, \hat{\tilde{\theta}} + \frac{S}{\sqrt{n}} \cdot t_{7,0,975}]$$

$$[1,34 - \frac{0,034}{\sqrt{8}} \cdot 2,365 \,;\, 1,34 + \frac{0,034}{\sqrt{8}} \cdot 2,365]$$

c'est-à-dire :

$$[1,312 \,;\, 1,368].$$

On peut donc affirmer que l'accroissement moyen du chiffre d'affaires est compris approximativement entre 31 et 37 % pendant la décennie.

MOTS CLÉS

Biais *(Bias)*
Erreur type *(Standard error)*
Estimateur *(Estimator)*
Estimation *(Estimation)*
Estimation ponctuelle *(Point estimation)*
Intervalle de confiance *(Confidence interval)*

RÉFÉRENCES

Efron, B. (1982). *The Jackknife, the Bootstrap and the other Resampling Plans.* Society for Industrial and Applied Mathematics. Philadelphia.

Quenouille, M.H. (1949). Approximate tests of correlation in time series. *Journal of the Royal Statistical Society.* Series B, 11, pp. 68-84.

Quenouille, M.H. (1956). Notes on bias in estimation. *Biometrika*, 43, pp. 353-360.

Tukey, J.W. (1958). Bias and confidence in not quite large samples. *Annals of Mathematical Statistics*, 29, pp. 614.

MODE

Mode

Le mode est une **mesure de tendance centrale** dont la définition la plus générale est la suivante : le mode d'un ensemble d'observations est la **valeur** rencontrée le plus fréquemment, c'est-à-dire celle qui a la plus grande **fréquence**.

Toutefois, nous rencontrons aussi la définition suivante : le mode d'un ensemble d'observations est une valeur dont la fréquence est supérieure ou égale aux fréquences des valeurs adjacentes.

D'après cette dernière définition, une distribution peut avoir un mode unique (dite distribution unimodale) ou plusieurs modes (dite distribution plurimodale, bimodale, trimodale, etc.).

Cette définition peut être formulée d'une manière théorique de la façon suivante :

Un minimum local est appelé anti-mode. Une distribution possédant un anti-mode est en forme de « U ».

ASPECTS MATHÉMATIQUES

Considérons une **variable aléatoire** X dont la **fonction de densité** est $f(x)$. M_o est un mode de la distribution si $f(x)$ a un maximum local en M_o.

Empiriquement, la détermination du mode d'un ensemble d'observations présentées individuellement s'effectue de la manière suivante :

1. Établir la **distribution de fréquences** à partir de l'ensemble des observations.

2. Le mode ou les modes sont les valeurs dont la **fréquence** est supérieure ou égale aux fréquences des valeurs adjacentes.

Pour des distributions de valeurs groupées en classes de même amplitude ($=$ *maximum* $-$ *minimum*) et dont les bornes sont connues, la procédure de détermination du mode est la suivante :

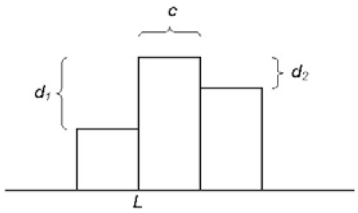
1. Déterminer la (ou les) classe(s) modale(s) ; les classes dont la fréquence est supérieure ou égale aux fréquences des classes adjacentes. (Si plusieurs classes adjacentes ont la même fréquence, nous regrouperons ces classes pour ne calculer qu'un seul mode sur cet ensemble de classes).

2. Calculer le mode M_o en tenant compte des fréquences des classes adjacentes :

$$M_o = L + \left(\frac{d_1}{d_1 + d_2}\right) \cdot c$$

où

L	limite inférieure de la classe modale
d_1	différence entre la fréquence de la classe modale et la fréquence de la classe précédente
d_2	différence entre la fréquence de la classe modale et la fréquence de la classe suivante
c	longueur de la classe modale, commune à toutes les classes.



Si la longueur des classes n'est pas identique, le mode doit être calculé en modifiant le découpage des classes pour obtenir, si possible, des classes de longueur égale.

DOMAINES ET LIMITATIONS

Pour une **variable** discrète, le mode présente l'avantage d'être facile à déterminer et à interpréter. Dans ce cas, le mode est surtout utilisé pour se faire rapidement une idée de la tendance centrale d'une série. Il est notamment intéressant de connaître sa valeur lorsque la distribution est asymétrique.

Pour une variable continue, la détermination du mode n'est pas toujours précise puisqu'elle dépend du découpage en classes.

Généralement, le mode n'est un bon **indicateur** du centre des données que s'il n'y a qu'une **valeur** dominante dans l'ensemble des données.

S'il y a plusieurs valeurs dominantes, la distribution est dite plurimodale. Dans un tel cas, les modes ne sont plus des mesures de tendance centrale. Une distribution bimodale indique généralement que la **population** considérée est en réalité hétérogène et composée de deux sous-populations n'ayant pas la même valeur centrale.

EXEMPLE

Soit l'ensemble de nombres suivants :

1 2 2 3 3 3 4 4 4 5 5 5 5 6 6 7 7 8 9

Le **tableau de fréquences** dérivé de cet ensemble est :

Valeur	1	2	3	4	5	6	7	8	9
Fréquence	1	2	3	3	4	2	2	1	1

Le mode de cet ensemble de nombres est 5 puisqu'il apparaît avec la plus grande **fréquence** (4).

Considérons à présent le tableau de fréquences suivant qui représente les revenus journaliers de cinquante épiceries, groupées par tranche ou classe de revenu :

Classes (revenus en euros)	Fréquences f_i (nombre d'épiceries)
$500-550$	3
$550-600$	12
$600-650$	17
$650-700$	8
$700-750$	6
$750-800$	4
Total	50

Nous constatons immédiatement que la classe modale est $600-650$ puisque c'est celle qui possède la fréquence la plus élevée.

Le mode se calcule de la manière suivante :

$$\begin{aligned} M_o &= L_1 + \left(\frac{d_1}{d_1+d_2}\right)\cdot c \\ &= 600 + \left(\frac{5}{5+9}\right)\cdot 50 = 617,86. \end{aligned}$$

Reprenons à présent le même exemple avec un regroupement en classes différent :

Classes (revenus en euros)	Fréquences f_i (nombre d'épiceries)
$500-600$	15
$600-650$	17
$650-700$	8
$700-800$	10
Total	50

Les classes ne sont pas de même longueur. Nous allons donc effectuer un nouveau découpage en vue d'obtenir des classes d'égale longueur.

Cas 1. Prenons par exemple comme référence la classe $600-650$ dont la longueur est 50 et supposons que les revenus sont répartis d'une façon uniforme dans chaque tranche.

Le tableau résultant de ce nouveau découpage est le suivant :

Classes (revenus en euros)	Fréquences f_i (nombre d'épiceries)
$500-550$	$7,5$
$550-600$	$7,5$
$600-650$	17
$650-700$	8
$700-750$	5
$750-800$	5
Total	50

La classe modale est la classe $600-650$, et la valeur du mode est :

$$M_o = 600 + \left(\frac{9,5}{9,5+9}\right) \cdot 50 = 625,67.$$

Cas 2. Nous aurions aussi pu nous baser sur la longueur de la première classe, ce qui nous donne le tableau suivant :

Classes (revenus en euros)	Fréquences f_i (nombre d'épiceries)
$500-600$	15
$600-700$	25
$700-800$	10
Total	50

La classe modale est la classe $600-700$, et la valeur du mode est alors :

$$M_o = 600 + \left(\frac{10}{10+15}\right) \cdot 100 = 640.$$

Ces résultats montrent bien qu'il faut être prudent dans l'utilisation du mode puisque le résultat obtenu est différent selon le découpage en classes.

MOTS CLÉS

Médiane *(Median)*
Mesure d'asymétrie *(Measure of skewness)*
Mesure de tendance centrale *(Measure of central tendency)*
Moyenne *(Mean)*

MODÈLE

Model

Un modèle est une représentation théorique d'une situation réelle. Il est constitué de symboles mathématiques, on parle alors de modèle mathématique.

Un modèle mathématique, basé sur un certain nombre d'observations et d'hypothèses, cherche à fournir la meilleure description possible du phénomène étudié.

Un modèle mathématique contient essentiellement deux types d'éléments : des variables, directement ou indirectement observables, concernant le phénomène à étudier, et des paramètres, quantités fixes généralement inconnues qui lient les variables entre elles au sein du modèle.

HISTORIQUE

La notion de « modèle » est apparue au XVIII[e] siècle, à travers trois problèmes majeurs de l'époque :

1. on souhaitait rendre compte de l'inégalité observée entre les mouvements respectifs de Jupiter et de Saturne ;

2. on cherchait à déterminer et décrire mathématiquement les mouvements de la Lune ;

3. on voulait déterminer la forme de la Terre.

Ces trois problèmes faisaient appel à des observations liées à l'astronomie et à certaines connaissances de la théorie de la gravitation. Ils préoccupaient alors de nombreux scientifiques, notamment Leonhard Euler (1749) qui traita le premier problème et Tobias Mayer (1750) intéressé par les mouvements de la Lune. Le troisième problème fut traité par **Roger Joseph Boscovich** et Christopher Maire (1755).

Ces trois ouvrages constituent les premiers travaux relatifs à la détermination d'un modèle lié chacun à une étude approfondie en astronomie.

ASPECTS MATHÉMATIQUES

Il existe plusieurs types de modèles. Parmi eux, le modèle déterministe est un modèle dans lequel les liens entre les variables décrits par le modèle sont considérés comme exacts. En revanche, le modèle stochastique (ou statistique) suppose que les liens ne sont pas exacts, mais établis par un processus aléatoire.

Un modèle statistique est déterminé principalement par la nature des variables indépendantes (exogènes ou explicatives). On parle de modèle continu lorsque les variables indépendantes X_j sont des variables quantitatives continues. On cherche alors à exprimer la **variable dépendante** Y (appelée également variable expliquée, endogène ou réponse) par une fonction du type :

$$Y = f\left(X_1, \ldots, X_p \,|\, \beta_0, \beta_1, \ldots, \beta_{p-1}\right) + \varepsilon$$

où les β_j sont des paramètres (ou coefficients) à déterminer et ε est un terme d'**erreur** aléatoire non observable. Parmi ce type de modèles, on peut distinguer des modèles linéaires et des modèles non linéaires, la linéarité étant relative aux coefficients. Un modèle linéaire peut s'écrire sous la forme d'un polynôme. Par exemple,

$$Y = \beta_0 + \beta_1 \cdot X + \beta_2 \cdot X^2 + \varepsilon$$

est un modèle linéaire dit « de deuxième ordre » (l'ordre provenant de la plus grande puissance de la variable X). Le modèle :

$$\log V = \beta_0 + \frac{\beta_1}{W}$$

est également un modèle linéaire par rapport aux paramètres β_0 et β_1. Ce modèle devient également linéaire par rapport aux variables en posant :

$$Y = \log V \text{ et } X = \frac{1}{W}.$$

En effet, cela donne :

$$Y = \beta_0 + \beta_1 \cdot X.$$

Par contre, le modèle

$$Y = \frac{\beta_0}{X + \beta_1} + \varepsilon$$

est un modèle non linéaire par rapport au paramètre β_1.

Il existe plusieurs méthodes pour estimer les paramètres d'un modèle. Parmi ces méthodes, on peut citer la méthode des **moindres carrés** et l'**estimation L1**.

Outre les modèles continus, on parle de modèles discrets lorsque la **population** étudiée peut être divisée en sous-populations sur la base de variables qualitatives catégorielles (ou facteurs) qui peuvent chacune prendre un certain nombre de modalités (ou **niveaux**). Le type de modèle dépend alors du type de **plan d'expérience** établi. L'objectif d'un tel plan est de déterminer comment les conditions expérimentales affectent les valeurs observées de la variable réponse. On peut distinguer différents types de modèles discrets :

- Un modèle à effets fixes (ou modèle I) est un modèle linéaire pour lequel les facteurs sont fixés, c'est- à-dire qu'ils prennent un nombre fini de niveaux qui sont tous présents dans le plan d'expérience établi. Dans ce cas, le modèle peut s'écrire :

  valeur observée = (fonction linéaire de paramètres inconnus) + erreur,

  les erreurs étant des variables aléatoires indépendantes d'**espérance mathématique** égale à zéro.

- Un modèle à effets variables (ou modèle II) est un modèle linéaire pour lequel les niveaux étudiés pour chaque facteur ont été choisis aléatoirement parmi un nombre, fini ou non, de niveaux possibles. Dans ce cas, on peut écrire :

  valeur observée = constante + (fonction linéaire de variables aléatoires) + erreur,

  les variables aléatoires étant indépendantes et ayant une espérance mathématique égale à zéro.

- Un modèle à effets mixtes (ou modèle III) est un modèle linéaire qui comprend à la fois des facteurs fixés et des facteurs aléatoires. On peut alors écrire :

valeur observée = (fonction linéaire de paramètres inconnus) + (fonction linéaire de variables aléatoires) + erreur.

Lorsque le plan d'expériences ne comprend qu'un seul facteur, noté α, le modèle peut s'écrire :

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij},$$

$$i = 1, \ldots, I \text{ et } j = 1, \ldots, n_i$$

où

- I est le nombre de niveaux du facteur α,
- n_i est le nombre d'observations à chaque niveau,
- μ est la **moyenne** de la **population**,
- α_i sont les effets associés à chaque niveau du facteur α et
- ε_{ij} sont des erreurs aléatoires indépendantes, distribuées selon une **loi normale** d'espérance mathématique égale à zéro et de **variance** homogène σ^2.

Un tel modèle peut être de type I ou de type II, en fonction de la nature du facteur α : il est de type I si les niveaux du facteur α sont fixes et s'ils apparaissent tous dans le plan. Il est de type II si α est une variable aléatoire indépendante de ϵ, distribuée selon une loi quelconque d'espérance mathématique égale à zéro et de variance σ^2_α (donc si les niveaux de α sont choisis aléatoirement parmi un nombre fini ou infini de niveaux possibles).

Lorsque le plan d'expérience comprend plusieurs facteurs, par exemple :

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk},$$

$$i = 1, \ldots, I \ ; \ j = 1, \ldots, J \text{ et } k = 1, \ldots, n_{ij},$$

le modèle peut alors être de type I (les niveaux de tous les facteurs sont fixés), II (tous les facteurs sont des variables aléatoires soumises aux mêmes restrictions que dans le modèle précédent) ou III (certains facteurs ont des niveaux fixés et d'autres variables).

Le terme $(\alpha\beta)_{ij}$ représente ici l'effet de l'**interaction** entre les I niveaux du facteur α et les J niveaux du facteur β.
Un modèle de type I peut être analysé par une **analyse de variance**.

MOTS CLÉS

Analyse de variance *(Analysis of variance)*
Paramètre *(Parameter)*
Plan d'expérience *(Design of experiments)*
Régression, analyse de *(Regression analysis)*
Variable *(Variable)*
Variable aléatoire *(Random variable)*
Variable dépendante *(Dependent variable)*
Variable indépendante *(Independent variable)*

RÉFÉRENCES

Anderson, V.L. and McLean, R.A. (1974). *Design of Experiments: A realistic approach.* Marcel Dekker, New York.

Boscovich, R.J. et Maire, C. (1755). *De Litteraria Expeditione per Pontificiam ditionem ad dimetiendas duas Meridiani gradus.* Palladis, Rome.

Euler, L. (1749). *Recherches sur la question des inégalités du mouvement de Saturne et de Jupiter, pièce ayant remporté le prix de l'année 1748, par l'Académie royale des sciences de Paris.* Réédité en 1960 in Leonhardi Euleri, *Opera Omnia*, 2e série, 25, pp. 47-157. Turici, Bâle.

Mayer, T. (1750). Abhandlung über die Umwalzung des Monds um seine Axe und die scheinbare Bewegung der Mondflecken. *Kosmographische Nachrichten und Sammlungen auf das Jahr 1748*, 1, pp. 52-183.

MODÈLES ARMA

ARMA Models

Les modèles *ARMA* (de l'anglais *autoregressive moving average models*) sont des modèles censés recouvrir une large évolution possible des **séries chronologiques**. La partie autorégressive d'un processus, notée *AR*, est constituée par une combinaison

linéaire finie des valeurs passées du processus. La partie moyenne mobile, MA, est constituée d'une combinaison linéaire finie en t des valeurs passées d'un **bruit blanc**.

1. Modèle AR (autorégressif)
Dans le processus autorégressif d'ordre p, l'observation présente y_t est générée par une moyenne pondérée des observations passées jusqu'à la p-ème période sous la forme suivante :

$$\begin{aligned} AR(1) &: y_t = \theta_1 y_{t-1} + \varepsilon_t \\ AR(2) &: y_t = \theta_1 y_{t-1} + \theta_2 y_{t-2} + \varepsilon_t \\ &\dots \\ AR(p) &: y_t = \theta_1 y_{t-1} + \theta_2 y_{t-2} + \ldots + \theta_p y_{t-p} + \varepsilon_t \end{aligned}$$

où $\theta_1, \theta_2, \ldots, \theta_p$ sont des paramètres à estimer positifs ou négatifs et ε_t est le facteur d'erreur distribué selon une loi normale.

2. Modèle MA (*moving average* : moyenne mobile)
Dans le processus de moyenne mobile d'ordre q, chaque observation y_t est générée par une **moyenne arithmétique pondérée** d'aléas jusqu'à la q-ème période.

$$\begin{aligned} MA(1) &: y_t = \varepsilon_t - \alpha_1 \varepsilon_{t-1} \\ MA(2) &: y_t = \varepsilon_t - \alpha_1 \varepsilon_{t-1} - \alpha_2 \varepsilon_{t-2} \\ &\dots \\ MA(p) &: y_t = \varepsilon_t - \alpha_1 \varepsilon_{t-1} - \alpha_2 \varepsilon_{t-2} - \ldots - \alpha_q \varepsilon_{t-q} \end{aligned}$$

où $\alpha_1, \alpha_2, \ldots, \alpha_q$ sont des paramètres positifs ou négatifs et ε_t est un aléa gaussien.

Le modèle MA est représentatif pour une série chronologique fluctuant autour de sa moyenne de manière aléatoire, d'où le terme de moyenne mobile car celle-ci lisse la série et soustrait le bruit blanc généré par l'élément de l'aléa.

3. Modèle $ARMA$ (*autoregressive moving average models*)
Les modèles $ARMA$ sont représentatifs d'un processus généré par une combinaison des valeurs passées et des erreurs passées. Ils sont définis par l'équation suivante :

$$ARMA(p,q) \quad : y_t = \theta_1 y_{t-1} + \theta_2 y_{t-2} + \ldots + \theta_p y_{t-p} + \varepsilon_t - \alpha_1 \varepsilon_{t-1} - \alpha_2 \varepsilon_{t-2} - \ldots - \alpha_q \varepsilon_{t-q}$$

avec $\theta_p \neq 0, \alpha_q \neq 0$ et $(\varepsilon_t, t \in Z)$ un bruit blanc faible.

MOTS CLÉS

Moyenne arithmétique pondérée *(Weighted arithmetic mean)*
Série chronologique *(Time series)*

MOINDRES CARRÉS

Least squares

Voir **méthode des moindres carrés**.

MOINDRES CARRÉS PONDÉRÉS, MÉTHODE DES

Weighted least squares method

Voir **méthode des moindres carrés pondérés**.

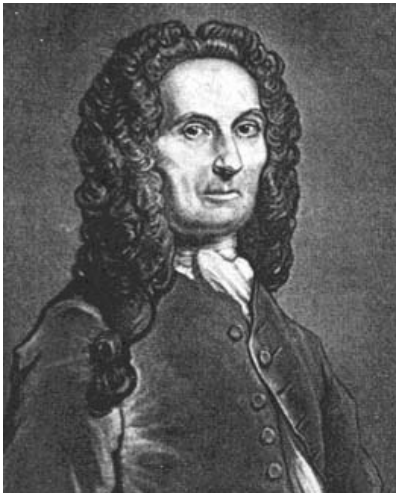
MOIVRE ABRAHAM DE

Moivre Abraham de

Abraham de Moivre (1667-1754) étudia d'abord à Sedan et à Saumur avant de poursuivre à la Sorbonne des études de mathématiques et de physique. Issu d'une famille protestante, il dut quitter la France en 1685, à la révocation de l'Édit de Nantes, et s'exila à Londres. Après des débuts difficiles, il put se consacrer aux mathématiques. Il se lia avec le scientifique Halley, puis avec Newton et fut élu membre de la Royal Society de Londres en 1697.

Ses ouvrages les plus importants furent *The Doctrine of Chances* paru en 1718, consacré à des questions qui toutes jouent un rôle fondamental dans la théorie des probabilités moderne, et *Miscellanea Analytica de Seriebus et Quadraturis* où pour la première fois apparaît la **fonction de densité** de la **loi normale** dans le cadre de calculs de probabilités pour des jeux de hasard. **Carl Friedrich Gauss**,

quant à lui, reprendra cette même fonction de densité pour des problèmes de mesures en astronomie. Il contribua aussi aux sciences actuarielles et à l'analyse mathématique.



Quelques ouvrages principaux d'Abraham de Moivre :

1718 *The Doctrine of Chances or A Method of Calculating the Probability of Events in Play.* W. Pearson, London.

1725 *Annuities upon Lives.* W. Pearson, London.

1730 *Miscellanea Analytica de Seriebus et Quadraturis.* J. Tonson and J. Watts, London.

MOTS CLÉS

Loi normale *(Normal distribution)*
Théorème de Moivre-Laplace *(Moivre-Laplace theorem)*

MOIVRE-LAPLACE, THÉORÈME DE

Moivre-Laplace theorem

Voir **théorème de Moivre-Laplace**.

MOMENT

Moment

On appelle moment d'ordre k de la **variable aléatoire** X par rapport à une **valeur** quelconque x_0, la différence moyenne entre la variable aléatoire et x_0, élevée à la puissance k.

Si $x_0 = 0$, on parle de moment (ou moment initial) d'ordre k.

Si $x_0 = \mu$, où μ est l'**espérance mathématique** de la variable aléatoire X, on parle de moment centré d'ordre k.

HISTORIQUE

Le terme « moment » provient de la statique où l'on parle du moment pour des forces.

Des formules corrigées pour le calcul des moments lors de données groupées ont été adaptées par W.F. Sheppard (1886).

ASPECTS MATHÉMATIQUES

Le moment initial d'ordre k d'une **variable aléatoire** discrète X, est exprimé mathématiquement par :

$$E\left[X^k\right] = \sum_{i=1}^{n} x_i^k P\left(x_i\right)$$

s'il existe, c'est-à-dire si $E\left[X^k\right] < \infty$; où $P\left(x\right)$ est la **fonction de probabilité** de la variable aléatoire discrète X prenant les n valeurs $x_1, \ldots, x_n$.

Dans le cas d'une variable aléatoire continue, si le moment existe, il est défini par :

$$E\left[X^k\right] = \int_{-\infty}^{\infty} x^k f\left(x\right) dx$$

où $f\left(x\right)$ est la **fonction de densité** de la variable aléatoire continue.

Pour $k = 1$, le moment initial se confond avec l'**espérance mathématique** de la variable aléatoire X (si cette espérance existe) :

$$(X \text{ discrète}) \; E[X] = \sum_{i=1}^{n} x_i P(x_i)$$

$$(X \text{ continue}) \; [X] = \int_{-\infty}^{\infty} x f(x)\, dx.$$

Le moment centré d'ordre k d'une variable aléatoire discrète X, s'il existe, est exprimé par :

$$E\left[(X - E[X])^k\right] = \sum_{i=1}^{n} (x_i - E[X])^k P(x_i).$$

De même, dans le cas d'une variable aléatoire continue, on a :

$$E\left[(X - E[X])^k\right] = \int_{-\infty}^{\infty} (x - E[X])^k f(x)\, dx$$

si l'intégrale est bien définie.

Il est évident que pour toute variable aléatoire X ayant une espérance mathématique finie, le moment centré d'ordre 1 est nul :

$$\begin{aligned} E[X - \mu] &= \sum_{i=1}^{n} (x_i - \mu) P(x_i) \\ &= \sum_{i=1}^{n} x_i P(x_i) - \mu \sum_{i=1}^{n} P(x_i) \\ &= \mu - \mu \\ &= 0. \end{aligned}$$

Par ailleurs, s'il existe, le moment centré d'ordre 2 se confond avec la **variance** de la variable aléatoire :

$$\begin{aligned} (X \text{ discrète}) \; E\left[(X-\mu)^2\right] &= \sum_{i=1}^{n} (x_i - \mu)^2 P(x_i) \\ &= Var(X) \end{aligned}$$

$$\begin{aligned} (X \text{ continue}) \; E\left[(X-\mu)^2\right] &= \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\, dx \\ &= Var(X). \end{aligned}$$

DOMAINES ET LIMITATIONS

L'**estimateur** du moment d'ordre k calculé sur un **échantillon** $(x_1, x_2, \ldots, x_n)$ est noté par m_k. Il est égal à :

$$m_k = \frac{\sum_{i=1}^{n} (x_i - x_0)^k}{n}$$

où n est le nombre total d'observations.

Si $x_0 = 0$, on parle de moment (ou moment initial) d'ordre k.

Si $x_0 = \bar{x}$, où $\bar{x}$ est la **moyenne arithmétique**, on parle de moment centré d'ordre k.

Pour le cas où une **variable aléatoire** X prend ses valeurs x_i avec les fréquences f_i, $i = 1, 2, \ldots, h$, le moment d'ordre k est donné par :

$$m_k = \frac{\sum_{i=1}^{h} f_i (x_i - x_0)^k}{\sum_{i=1}^{h} f_i}.$$

MOTS CLÉS

Espérance mathématique *(Expected value)*
Fonction de densité *(Density function)*
Fonction de probabilité *(Probability function)*
Moyenne arithmétique *(Arithmetic mean)*
Moyenne arithmétique pondérée *(Weighted arithmetic mean)*
Variable aléatoire *(Random variable)*
Variance *(Variance)*

RÉFÉRENCE

Sheppard, W.F. (1898). On the calculation of the most probable values of frequency constants for data arranged according to equidistant divisions of a scale. *London Mathematical Society Proceedings* 29, pp. 353-380.

MOUVEMENT EXTRA-SAISONNIER

Extra-seasonal movement

Dans la décomposition d'une **série chronologique**, on désigne conjointement sous le terme de mouvement extra-saisonnier ou parfois de mouvement conjoncturel, la **tendance séculaire** et les fluctuations cycliques qu'on ne cherche souvent pas à dissocier.

HISTORIQUE

Voir **série chronologique**.

ASPECTS MATHÉMATIQUES

D'une façon générale, en relation avec une **série chronologique** dont Y_t représente l'**observation** faite au temps t, on distingue :

— le mouvement extra-saisonnier au temps t, noté M_t ;

— la **variation saisonnière** au temps t, notée S_t ;

— la **variation irrégulière** au temps t, notée I_t.

Ces diverses composantes s'articulent suivant différents schémas, par exemple :

— le **modèle** additif

$$Y_t = M_t + S_t + I_t\,;$$

— le modèle multiplicatif de la forme

$$Y_t = M_t \cdot S_t \cdot I_t$$

ou de la forme $Y_t = M_t \cdot (1 + S_t) \cdot (1 + I_t)$.

À l'aide du procédé de moyennes mobiles appliqué à la série chronologique Y_t, on peut estimer les différentes composantes M_t, S_t et I_t de cette dernière.

On peut également utiliser la méthode des moyennes mobiles simplement pour :

- éliminer les variations saisonnières S_t de la série chronologique Y_t ;
- lisser les variations irrégulières I_t ;
- réduire la série chronologique Y_t à son mouvement extra-saisonnier.

DOMAINES ET LIMITATIONS

L'**estimation** du mouvement extra-saisonnier permet d'effectuer une **prévision** à court terme, pour laquelle la connaissance du mouvement extra-saisonnier est suffisante.

Cependant, en macro-économie, on prend soin de distinguer :

- la **tendance séculaire** qui montre l'évolution sur une période de l'ordre de plusieurs dizaines d'années ;
- les fluctuations cycliques, dont la période est de l'ordre de la dizaine d'années.

Lorsque le graphique de la série chronologique étudiée présente une tendance séculaire exponentielle ou plus généralement avec une forte courbure, la méthode des moyennes mobiles ne donne pas des résultats très précis pour l'estimation du mouvement extra-saisonnier.

MOTS CLÉS

Fluctuation cyclique *(Cyclical fluctuation)*
Moyenne mobile *(Moving average)*
Série chronologique *(Time series)*
Tendance séculaire *(Secular trend)*
Variation irrégulière *(Irregular variation)*
Variation saisonnière *(Seasonal variation)*

RÉFÉRENCES

Bowerman, B.L. and O'Connel, R.T. (1979). *Time Series and Forecasting: An Applied Approach.* Duxbury Press, Wadsworth, Belmont, California.

Harnett, D.L. and Murphy, J.L. (1975). *Introductory Statistical Analysis.* Addison-Wesley, Reading, Mass.

MOYENNE

Mean

La moyenne d'une **variable aléatoire** est une **mesure de tendance centrale** de cette **variable**. Elle est également appelée **espérance mathématique**.

Dans la pratique, le terme moyenne est très souvent utilisé dans le sens de **moyenne arithmétique**.

HISTORIQUE

Voir **moyenne arithmétique**.

ASPECTS MATHÉMATIQUES

Soit X une **variable aléatoire** continue dont la **fonction de densité** est $f(x)$. La moyenne μ de X, si elle existe, est donnée par :

$$\mu = \int x \cdot f(x)\, dx.$$

La moyenne d'une variable aléatoire discrète X de **fonction de probabilité** $P(X = x)$, si elle existe, est donnée par :

$$\mu = \sum x \cdot P(X = x).$$

DOMAINES ET LIMITATIONS

Empiriquement, la détermination de la **moyenne arithmétique** d'un ensemble d'observations nous fournit une mesure de la **valeur** centrale de cet ensemble. Cette mesure nous donne une **estimation** de la moyenne μ de la **variable aléatoire** X dont sont issues les observations.

MOTS CLÉS

Espérance mathématique *(Expected value)*
Mesure d'aplatissement *(Measure of kurtosis)*
Mesure d'asymétrie *(Measure of skewness)*
Mesure de tendance centrale *(Measure of central tendency)*
Moyenne arithmétique *(Arithmetic mean)*
Moyenne arithmétique pondérée *(Weighted arithmetic mean)*

MOYENNE ARITHMÉTIQUE

Arithmetic mean

La **moyenne** arithmétique est une **mesure de tendance centrale** qui permet de caractériser le centre de la **distribution de fréquences** d'une **variable quantitative** en considérant toutes les observations et en leur attribuant le même poids (par opposition à la **moyenne arithmétique pondérée**).

Elle est calculée par la somme des observations divisée par leur nombre.

HISTORIQUE

La moyenne arithmétique est l'une des plus anciennes méthodes employées pour combiner des observations afin d'obtenir une valeur approximative unique. Son utilisation semble en effet remonter au temps des astronomes babyloniens du III^{e} siècle avant J.-C. La science de l'astronomie utilisa la moyenne arithmétique pour déterminer la position du soleil, de la lune et des planètes. Selon R.L. Plackett (1958), c'est avec l'astronome grec Hipparchus que la moyenne arithmétique se généralise.

C'est en 1755 que Thomas Simpson propose officiellement l'utilisation de la moyenne arithmétique par une lettre au président de la Royal Society.

ASPECTS MATHÉMATIQUES

Soit $x_1, x_2, \ldots, x_n$ un ensemble de n quantités, ou de n observations relatives à une **variable quantitative** X.

La moyenne arithmétique $\bar{x}$ de $x_1, x_2, \ldots, x_n$ est la somme de ces observations divisée par le nombre n d'observations :

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}.$$

Lorsque les observations sont ordonnées sous forme d'une **distribution de fréquences**, le calcul de la

moyenne arithmétique se fait de la façon suivante :

$$\bar{x} = \frac{\sum_{i=1}^{k} x_i \cdot f_i}{\sum_{i=1}^{k} f_i}$$

où les x_i sont les différentes valeurs de la **variable**, les f_i sont les fréquences associées à ces valeurs, k est le nombre de valeurs différentes, et la somme des fréquences est égale au nombre d'observations :

$$\sum_{i=1}^{k} f_i = n.$$

Pour calculer la **moyenne** d'une distribution de fréquences où les valeurs de la variable quantitative X sont groupées en classes, on considère que toutes les observations appartenant à une certaine classe prennent la valeur du centre de la classe, en supposant que les observations sont uniformément réparties à l'intérieur des classes. (Si cette **hypothèse** n'est pas vérifiée, la valeur de la moyenne arithmétique obtenue ne sera qu'approximative.)

On a donc :

$$\bar{x} = \frac{\sum_{i=1}^{k} x_i \cdot f_i}{\sum_{i=1}^{k} f_i}$$

où les x_i sont ici considérés comme les centres des classes, les f_i sont les fréquences associées à chaque classe, et k est le nombre de classes.

Propriétés de la moyenne arithmétique

- la somme algébrique des écarts entre chacune des valeurs d'un ensemble et la moyenne arithmétique de cet ensemble est égale à 0 :

$$\sum_{i=1}^{n} (x_i - \bar{x}) = 0$$

- la somme des carrés des écarts de chaque valeur à un nombre donné « a » est minimale lorsque « a » est la moyenne arithmétique :

$$\sum_{i=1}^{n} (x_i - a)^2 \geq \sum_{i=1}^{n} (x_i - \bar{x})^2 .$$

Démonstration :

On peut écrire :

$$x_i - a = (x_i - \bar{x}) + (\bar{x} - a) .$$

En élevant les deux membres de l'égalité au carré et en en faisant la somme, on obtient après simplification :

$$\sum_{i=1}^{n} (x_i - a)^2 = \sum_{i=1}^{n} (x_i - \bar{x})^2 + n \cdot (\bar{x} - a)^2 .$$

Comme $n \cdot (\bar{x} - a)^2$ est non négatif, on a démontré que :

$$\sum_{i=1}^{n} (x_i - a)^2 \geq \sum_{i=1}^{n} (x_i - \bar{x})^2$$

- la moyenne arithmétique $\bar{x}$ d'un **échantillon** $(x_1, \ldots, x_n)$ est habituellement considérée comme un **estimateur** de la **moyenne** μ de la **population** à partir de laquelle l'échantillon a été tiré.

- on suppose que les x_i soient des variables aléatoires indépendantes et distribuées selon la même **fonction de répartition** de moyenne μ et de **variance** σ^2. On peut alors vérifier que :

  1. $E[\bar{x}] = \mu$,
  2. $Var(\bar{x}) = \frac{\sigma^2}{n}$,

  à condition que ces moments existent.

  L'**espérance mathématique** de $\bar{x}$ étant égale à μ, la moyenne arithmétique est un estimateur sans **biais** de la moyenne de la population.

- si les x_i résultent d'un **échantillonnage aléatoire** sans remise sur une population finie de moyenne μ, l'identité

$$E[\bar{x}] = \mu$$

continue à être valable, mais en revanche, la variance de $\bar{x}$ doit être ajustée par un facteur

dépendant de la taille N de la population et de la taille n de l'échantillon :

$$Var\left(\bar{x}\right) = \frac{\sigma^2}{n} \cdot \left[\frac{N-n}{N-1}\right],$$

où σ^2 est la variance de la population.

Relations entre la moyenne arithmétique et d'autres mesures de tendance centrale

- La moyenne arithmétique est liée aux deux autres principales mesures de tendance centrale, soit le **mode** M_o et la **médiane** M_d.

  Si la distribution est symétrique et unimodale :

  $$\bar{x} = M_d = M_o.$$

  Si la distribution est unimodale, il est généralement vrai que :

  $\bar{x} \geq M_d \geq M_o$ si la distribution est étirée à droite ;

  $\bar{x} \leq M_d \leq M_o$ si la distribution est étirée à gauche.

  Pour une distribution unimodale, modérément asymétrique, ces trois mesures de tendance centrale satisfont souvent à la relation suivante, de manière approximative :

  $$\left(\bar{x} - M_o\right) = 3 \cdot \left(\bar{x} - M_d\right).$$

- De même, pour une distribution unimodale, si l'on considère un ensemble de quantités positives, la **moyenne géométrique** G est toujours inférieure ou égale à la **moyenne arithmétique** $\bar{x}$, et toujours supérieure ou égale à la **moyenne harmonique** H. On a donc :

  $$H \leq G \leq \bar{x}.$$

  Ces trois moyennes ne sont identiques que si toutes les quantités sont égales.

DOMAINES ET LIMITATIONS

La moyenne arithmétique est une mesure simple de la **valeur** centrale d'un ensemble d'observations quantitatives. Toutefois, elle amène parfois à une interprétation faussée de la réalité.

Si les salaires mensuels de 5 personnes sont de 3 000, 3 200, 2 900, 3 500 et 6 500, on obtient une moyenne arithmétique du salaire égale à $\frac{19\,100}{5} = 3\,820$. Cette **moyenne** donne bien une idée de l'ordre de grandeur du salaire dans la mesure où elle se situe quelque part entre le salaire le plus bas et le salaire le plus haut. Cependant, elle ne donne pas une bonne représentation des cinq salaires puisque 80 % d'entre eux sont inférieurs à la moyenne.

Il faut donc être très attentif à la forme de la distribution et à la fiabilité des observations avant d'utiliser la moyenne arithmétique comme **mesure de tendance centrale** d'un ensemble de valeurs. Si l'on remarque des observations aberrantes dans la distribution, la moyenne arithmétique risque de donner une valeur peu représentative de la tendance centrale.

Si certaines observations sont considérées moins fiables que d'autres, il pourrait être avantageux d'attribuer à celles-ci moins d'importance, soit en calculant une **moyenne arithmétique pondérée**, soit en utilisant la **médiane** qui n'est pas trop influencée par les observations aberrantes.

EXEMPLES

Exemple 1 : dans l'entreprise A, les 9 employés reçoivent respectivement le salaire mensuel suivant :

3 000 3 200 2 900 3 440 5 050 4 150 3 150 3 300 5 200

La moyenne arithmétique du salaire mensuel des employés est égale à :

$$\begin{aligned}\bar{x} &= \frac{(3\,000 + 3\,200 + \cdots + 3\,300 + 5\,200)}{9} \\ &= \frac{33\,390}{9} = 3\,710.\end{aligned}$$

On examine à présent un cas où les valeurs données sont présentées sous forme d'une **distribution de fréquences**.

Le **tableau de fréquences** suivant donne le nombre de jours de maladie de 50 employés sur une période d'une année :

x_i : Jours de maladie	f_i : Nombre d'employés
0	7
1	12
2	19
3	8
4	4
Total	50

On cherche à calculer la durée moyenne de maladie des employés.

Le nombre total de jours de maladie des 50 employés est égal à la somme du produit de chaque valeur x_i par sa **fréquence** f_i respective :

$$\begin{aligned}\sum_{i=1}^{5} x_i \cdot f_i &= 0 \cdot 7 + 1 \cdot 12 + 2 \cdot 19 + 3 \cdot 8 + 4 \cdot 4 \\ &= 90.\end{aligned}$$

Le nombre total d'employés est égal à :

$$\sum_{i=1}^{5} f_i = 7 + 12 + 19 + 8 + 4 = 50.$$

La moyenne arithmétique du nombre de jours de maladie par employé est donc :

$$\bar{x} = \frac{\sum_{i=1}^{5} x_i \cdot f_i}{\sum_{i=1}^{5} f_i} = \frac{90}{50} = 1,8$$

ce qui signifie qu'en moyenne, les 50 employés concernés sont malades $1,8$ jour par an.

Exemple 2 : dans l'exemple suivant, les données sont groupées en classes.

On veut calculer la moyenne arithmétique des profits journaliers de 50 épiceries dont la distribution de fréquences est donnée par le tableau suivant :

Classes (profits en euros)	Points milieux x_i	Fréquences f_i (nombre d'épiceries)	$x_i \cdot f_i$
$500 - 550$	525	3	1 575
$550 - 600$	575	12	6 900
$600 - 650$	625	17	10 625
$650 - 700$	675	8	5 400
$700 - 750$	725	6	4 350
$750 - 800$	775	4	3 100
Total		50	31 950

La moyenne arithmétique des profits est égale à :

$$\bar{x} = \frac{\sum_{i=1}^{6} x_i \cdot f_i}{\sum_{i=1}^{6} f_i} = \frac{31\,950}{50} = 639$$

ce qui signifie qu'en **moyenne**, les 50 épiceries concernées ont un profit journalier de 639 euros.

MOTS CLÉS

Mesure de tendance centrale *(Measure of central tendency)*
Moyenne *(Mean)*
Moyenne arithmétique pondérée *(Weighted arithmetic mean)*
Moyenne géométrique *(Geometric mean)*
Moyenne harmonique *(Harmonic mean)*

RÉFÉRENCES

Plackett, R.L. (1958). Studies in the history of probability and statistics. VII: The Principle of the Arithmetic Mean. *Biometrika*, 45, pp. 130-135.

Simpson, T. (1755). A Letter to the Right Honorable George Earl of Macclesfield, President of the Royal Society, on the Advantage of Taking the Mean of a Number of Observations in Practical Astronomy. *Philosophical Transactions of the Royal Society of London*, 49, pp. 82-93.

Simpson, T. (1757). An attempt to show the advantage arising by taking the mean of a number of observations in practical astronomy. *Miscellaneous Tracts on some curious and very interesting Subjects in Mechanics, Physical-Astronomy, and Speculative Mathematics*, Nourse, London, pp. 64-75.

MOYENNE ARITHMÉTIQUE PONDÉRÉE

Weighted arithmetic mean

La **moyenne arithmétique** pondérée est une **mesure de tendance centrale** d'un ensemble d'observations quantitatives utilisée lorsque les observations n'ont pas toutes une importance identique.

Nous devons donc attribuer un poids à chaque observation en fonction de son importance relative par rapport aux autres observations.

La moyenne arithmétique pondérée est égale à la somme des observations multipliées par leur poids, divisée par la somme des poids.

HISTORIQUE

La moyenne arithmétique pondérée a été introduite par Roger Cotes en 1712. Son ouvrage majeur a été publié en 1722, six ans après sa mort.

ASPECTS MATHÉMATIQUES

Soit $x_1, x_2, \ldots, x_n$ un ensemble de n quantités, ou de n observations relatives à une **variable quantitative** X auxquelles nous attribuons les poids $w_1, w_2, \ldots, w_n$.

La moyenne arithmétique pondérée est égale à :

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i \cdot w_i}{\sum_{i=1}^{n} w_i}.$$

DOMAINES ET LIMITATIONS

La moyenne arithmétique pondérée est couramment utilisée dans le domaine de l'économie, notamment dans le calcul des indices de prix à la consommation, ou de prix de production, etc.

EXEMPLE

Supposons qu'un cours de gestion d'entreprise soit composé de trois exigences ayant des importances relatives différentes :

Projet individuel	30 %
Examen intermédiaire	20 %
Examen final	50 %

Chaque étudiant est noté sur une échelle entre 1 et 10. Un étudiant obtient la note 8 pour son projet individuel, 9 pour l'examen intermédiaire, et 4 pour l'examen final.

Sa note finale se calcule par la moyenne arithmétique pondérée :

$$\bar{x} = \frac{(8 \cdot 30) + (9 \cdot 20) + (4 \cdot 50)}{30 + 20 + 50} = 6,2.$$

Dans cet exemple, les poids correspondent à l'importance relative des différentes exigences du cours.

MOTS CLÉS

Mesure de tendance centrale *(Measure of central tendency)*
Moyenne *(Mean)*
Moyenne arithmétique *(Arithmetic mean)*

RÉFÉRENCE

Cotes, R. (1722). Aestimatio Errorum in Mixta Mathesi, per variationes partium Trianguli plani et sphaerici. *Opera Miscellania*, ed. Robert Smith, Cambridge.

MOYENNE GÉOMÉTRIQUE

Geometric mean

La **moyenne** géométrique est définie comme la n-ème racine du produit de n nombres non négatifs.

HISTORIQUE

Selon J-J. Droesbeke et Ph. Tassi (1990), la moyenne géométrique et la **moyenne harmonique** ont été introduites par William Stanley Jevons en 1874.

ASPECTS MATHÉMATIQUES

Soit x_1, x_2, ..., x_n, un ensemble de n quantités non négatives, ou de n observations relatives à une **variable quantitative** X. La moyenne géométrique G de cet ensemble est égale à :

$$G = \sqrt[n]{x_1 \cdot x_2 \cdot \ldots \cdot x_n}.$$

La moyenne géométrique est parfois exprimée sous sa forme logarithmique :

$$\ln(G) = \frac{\sum_{i=1}^{n} \ln(x_i)}{n}$$

ou

$$G = \exp\left[\frac{1}{n} \cdot \sum_{i=1}^{n} \ln(x_i)\right].$$

Nous remarquons donc que le logarithme de la moyenne géométrique d'un ensemble de nombres positifs est la **moyenne arithmétique** des logarithmes de ces nombres (ou la **moyenne arithmétique pondérée** dans le cas d'observations groupées).

DOMAINES ET LIMITATIONS

Dans la pratique, la moyenne géométrique est utilisée essentiellement pour calculer la **moyenne** de ratios, ou plus particulièrement la moyenne d'indices.

Comme la **moyenne arithmétique**, la moyenne géométrique prend en compte chaque **observation** individuellement. Toutefois, elle diminue l'effet des très grands nombres. C'est une raison pour laquelle elle est parfois préférée à la moyenne arithmétique.

Un des aspects importants de la moyenne géométrique est qu'elle ne s'applique qu'aux nombres positifs.

EXEMPLE

Voici un exemple de l'utilisation de la moyenne géométrique : en 1987, 1988 et 1989, un commerçant a réalisé respectivement un profit de 10 000, 20 000 et 160 000 euros. Sur la base de ces informations, on désire déterminer le taux moyen de croissance du profit.

De 1987 à 1988, le profit a doublé et de 1988 à 1989 il a été multiplié par huit. Si nous calculions simplement la **moyenne arithmétique** de ces deux nombres, nous obtiendrions

$$x = \frac{2+8}{2} = 5,$$

et pourrions conclure qu'en **moyenne**, les profits ont été multipliés par 5 chaque année. Ce résultat présenté ainsi n'est évidemment pas correct puisque si les profits en 1987 étaient de 10 000 euros et qu'ils étaient multipliés par 5 chaque année, nous aurions 50 000 euros de profits en 1988 et 250 000 euros en 1989. Or ces résultats sont manifestement trop élevés.

Par contre, si nous calculons la moyenne en utilisant la moyenne géométrique de ces deux taux de croissance plutôt que leur moyenne arithmétique, nous obtenons :

$$G = \sqrt{2 \cdot 8} = \sqrt{16} = 4$$

et nous pouvons dire correctement qu'en moyenne, les profits ont quadruplé chaque année. En effet, en appliquant ce taux de croissance moyen au profit de 10 000 euros en 1987, nous obtenons 40 000 euros en 1988 et 160 000 euros en 1989. Bien que le profit pour 1988 soit toujours trop élevé (mais moins que dans le résultat précédent), le profit pour 1989 est à présent correct.

Pour illustrer l'utilisation de la formule de la moyenne géométrique sous sa forme logarithmique, trouvons la moyenne des indices suivants :

112 99 105 96 85 100

Nous calculons (avec une table des logarithmes)

$$\begin{array}{lcr} \log(112) & = & 2,0492 \\ \log(99) & = & 1,9956 \\ \log(105) & = & 2,0212 \\ \log(96) & = & 1,9823 \\ \log(85) & = & 1,9294 \\ \log(100) & = & 2,0000 \\ \hline & & 11,9777 \end{array}$$

Nous obtenons donc :

$$\log(G) = \frac{11,9777}{6} = 1,9963,$$

d'où $G = 99,15$.

MOTS CLÉS

Mesure de tendance centrale *(Measure of central tendency)*
Moyenne *(Mean)*
Moyenne arithmétique *(Arithmetic mean)*
Moyenne harmonique *(Harmonic mean)*

RÉFÉRENCES

Droesbeke, J.J. et Tassi, Ph. (1990). *Histoire de la statistique*. Presses universitaires de France, Que sais-je ? Paris.

Jevons, W.S. (1874). *The principles of Science: A Treatise on logic and scientific methods*, 2 Vols. The Macmillan Company, London.

MOYENNE HARMONIQUE

Harmonic mean

La **moyenne** harmonique de n nombres est définie comme n divisé par la somme des inverses de chaque nombre.

HISTORIQUE

Voir **moyenne géométrique**.

La relation entre la moyenne harmonique, la moyenne géométrique et la **moyenne arithmétique** est présentée par D.S. Mitrinovic (1970).

ASPECTS MATHÉMATIQUES

Soient $x_1, x_2, \ldots, x_n$, n quantités non nulles, ou n observations relatives à une **variable quantitative** X. La moyenne harmonique H de ces n quantités se calcule comme suit :

$$H = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}.$$

Si $\{x_i\}_{i=1,\ldots,n}$ représente une série finie de nombres positifs, on note que :

$$\min_i x_i \leq H \leq G \leq \bar{x} \leq \max_i x_i.$$

DOMAINES ET LIMITATIONS

La moyenne harmonique est très peu utilisée en **statistique**. Toutefois, elle peut se révéler appropriée dans certains cas. Par exemple :

- si des investissements à des taux différents amènent chacun le même revenu, le taux unique auquel le capital total doit être investi pour produire le revenu total est égal à la moyenne harmonique des différents taux individuels ;

- si un montant identique est investi dans plusieurs biens de prix différents, le prix moyen des biens est égal à la moyenne harmonique des différents prix ;

- une propriété de la moyenne harmonique est d'être très peu influencée par des observations aberrantes lorsque celles-ci sont beaucoup plus grandes que l'ensemble des autres données. Par exemple, si l'on considère les valeurs suivantes : 1, 2, 3, 4, 5 et 100, la moyenne harmonique est égale à $2,62$ tandis que la **moyenne arithmétique** vaut dans ce cas $19,17$. Par contre, elle est beaucoup plus sensible aux valeurs aberrantes lorsque celles-ci sont beaucoup plus petites que l'ensemble des autres données. C'est ainsi qu'avec les observations 1, 6, 6, 6, 6, 6, on trouve $H = 3,27$ alors que la moyenne arithmétique vaut ici $5,17$.

EXEMPLES

Exemple 1 : trois investissements produisant chacun le même revenu ont les taux de rendement suivants : 5 %, 10 % et 15 %.

La moyenne harmonique donne le taux auquel il faudrait placer le capital pour produire le revenu total des trois investissements :

$$H = \frac{3}{\left[\frac{1}{5} + \frac{1}{10} + \frac{1}{15}\right]} = \frac{3}{\frac{11}{30}} = 8,18\,\%.$$

On remarque que ce résultat est différent de la **moyenne arithmétique** qui serait de 10 % $(5 + 10 + 15/3)$.

Exemple 2 : un représentant achète pour un montant de 200 euros trois cafés de qualité différente au prix respectif de 3, 2 et 1, 5 euros le kg.

Le prix moyen du kg de café est alors donné en divisant le coût total par la quantité totale achetée :

$$\begin{aligned}\text{prix moyen} &= \frac{\text{coût total}}{\text{quantité totale}} \\ &= \frac{3 \cdot 200}{66,66 + 100 + 133,33} \\ &= 2.- \,.\end{aligned}$$

Ceci correspond à la moyenne harmonique des prix des trois qualités de café différentes :

$$\text{prix moyen} = \frac{3}{\left[\frac{1}{3} + \frac{1}{2} + \frac{1}{1,5}\right]} = \frac{6}{3} = 2.- \,.$$

MOTS CLÉS

Mesure de tendance centrale *(Measure of central tendency)*
Moyenne *(Mean)*
Moyenne arithmétique *(Arithmetic mean)*
Moyenne géométrique *(Geometric mean)*

RÉFÉRENCE

Mitrinovic, D.S. (1970). *Analytic Inequalities.* In cooperation with P.M. Vasic. Springer-Verlag, New York.

MOYENNE MOBILE

Moving average

On appelle moyenne mobile d'ordre k la **moyenne arithmétique** calculée sur k valeurs successives d'une **série chronologique**. La suite de ces moyennes arithmétiques donne alors la série des moyennes mobiles d'ordre k. Le processus qui consiste à remplacer la série initiale par cette nouvelle série est appelé lissage des séries chronologiques.

Quand la série chronologique consiste en données annuelles ou mensuelles, les moyennes prennent respectivement les noms de moyennes mobiles sur k années ou k mois.

Lors de données mensuelles, par exemple, en calculant une moyenne mobile sur 12 mois, les résultats obtenus correspondent au milieu de la période considérée, à savoir à la fin du sixième mois (ou au premier jour du septième mois), au lieu d'être affectée au milieu d'un mois comme les données d'origine. On corrige ceci en effectuant une moyenne mobile centrée sur 12 mois.

En général, une moyenne mobile centrée d'ordre k est calculée en effectuant la moyenne mobile d'ordre 2 de la moyenne mobile d'ordre k de la série chronologique initiale.

HISTORIQUE

Voir **série chronologique**.

ASPECTS MATHÉMATIQUES

Étant donné la **série chronologique** : Y_1, Y_2, Y_3, ..., Y_N, on définit la série des moyennes mobiles d'ordre k par la suite des moyennes arithmétiques

$$\frac{Y_1 + Y_2 + \cdots + Y_k}{k}, \frac{Y_2 + Y_3 + \cdots + Y_{k+1}}{k}, \ldots$$
$$\frac{Y_{N-k+1} + Y_{N-k+2} + \cdots + Y_N}{k}.$$

Les sommes aux numérateurs sont appelées sommes mobiles d'ordre k.

On peut utiliser des moyennes arithmétiques pondérées, avec des poids spécifiés à l'avance ; la suite ainsi obtenue est appelée série des moyennes mobiles pondérées d'ordre k.

DOMAINES ET LIMITATIONS

Les moyennes mobiles sont utilisées dans l'étude des séries chronologiques, pour l'**estimation** de la **tendance séculaire**, des variations saisonnières et des fluctuations cycliques.

On prendra soin que chaque moyenne successive soit affectée à l'une des dates d'**observation** pour la **série chronologique**, ce qui est le cas lorsque l'on fait une moyenne mobile d'ordre impair. Cependant

lorsque l'ordre est pair, la moyenne mobile est affectée au centre de l'**intervalle** séparant deux dates consécutives d'observation.

Propriétés

L'opération « moyenne mobile » appliquée à une série chronologique permet :

- d'éliminer les variations saisonnières dans la mesure où celles-ci sont rigoureusement périodiques ;
- de lisser les variations irrégulières ;
- de conserver approximativement le **mouvement extra-saisonnier**.

Citons les inconvénients de la méthode des moyennes mobiles :

- les données de début et de fin de série sont « perdues » ;
- les moyennes mobiles peuvent engendrer des cycles ou d'autres mouvements qui n'étaient pas présents dans les données d'origine ;
- les moyennes mobiles sont fortement affectées par les valeurs aberrantes ou accidentelles ;
- lorsque le graphique de la série chronologique étudiée présente une tendance séculaire exponentielle ou plus généralement une forte courbure, la méthode des moyennes mobiles ne donne pas des résultats très précis pour l'estimation du mouvement extra-saisonnier.

EXEMPLE

Étant donné la **série chronologique** 7, 2, 3, 4, 5, 0, 1, on obtient la série des moyennes mobiles d'ordre 3 :

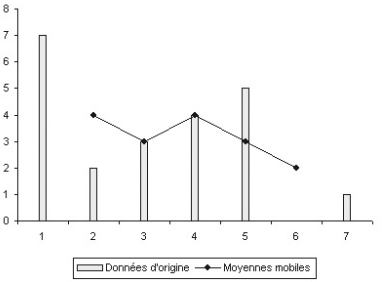
$$\frac{7+2+3}{3}, \frac{2+3+4}{3}, \frac{3+4+5}{3}, \frac{4+5+0}{3}, \frac{5+0+1}{3}$$

c'est-à-dire 4, 3, 4, 3, 2.

Dans une série de moyennes mobiles, il est commode de localiser chaque nombre à sa position relative par rapport aux données d'origine. Dans cet exemple, on écrit :

Données d'origine	7, 2, 3, 4, 5, 0, 1
Moyennes mobiles d'ordre 3	4, 3, 4, 3, 2

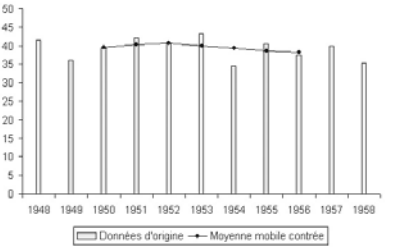
Chaque nombre de la série des moyennes mobiles étant la **moyenne arithmétique** des trois nombres immédiatement situés au-dessus de lui.



On se propose de calculer la série des moyennes mobiles centrées sur 4 ans de la production mensuelle moyenne de houille grasse d'un pays.

Dans une première étape, on effectue les sommes mobiles sur 4 ans, en additionnant les données de 4 années successives.

Ensuite, en divisant chacune des sommes mobiles par 4, on obtient les moyennes mobiles sur 4 ans ; ces dernières correspondent à la date du 1er janvier de la troisième année considérée dans les différentes sommes. Finalement, on obtient les moyennes mobiles centrées sur 4 ans, en calculant la moyenne arithmétique de deux moyennes mobiles sur 4 années successives.



Production mensuelle moyenne de houille grasse d'un pays, donnée en millions de tonnes.

Année	Donnée	Somme mobile sur 4 ans	Moyenne mobile sur 4 ans	Moyenne mobile centrée sur 4 ans
1948	41,4			
1949	36,0			
		158,8	39,7	
1950	39,3			39,6
		158,0	39,5	
1951	42,1			40,4
		165,2	41,3	
1952	40,6			40,7
		160,4	40,1	
1953	43,2			39,9
		158,8	39,7	
1954	34,5			39,3
		155,6	38,9	
1955	40,5			38,5
		152,4	38,1	
1956	37,4			38,2
		153,2	38,3	
1957	40,0			
1958	35,3			

MOTS CLÉS

Moyenne arithmétique *(Arithmetic mean)*
Moyenne arithmétique pondérée *(Weighted arithmetic mean)*
Série chronologique *(Time series)*

MULTICOLINÉARITÉ

Multicollinearity

Multicolinéarité est le terme utilisé pour décrire la situation où une **variable** est presque égale à une combinaison linéaire d'autres variables. La situation de multicolinéarité se présente en particulier dans les analyses de **régression linéaire multiple** quand il y a une forte **corrélation** entre les variables indépendantes.

L'existence de relations linéaires ou presque linéaires entre les variables indépendantes entraîne une grande imprécision dans la signification des coefficients de **régression**. À la limite, si ces liaisons étaient rigoureusement linéaires, les coefficients seraient alors indéterminés.

ASPECTS MATHÉMATIQUES

Voir **régression ridge**.

MOTS CLÉS

Coefficient de corrélation *(Correlation coefficient)*
Régression linéaire multiple *(Multiple linear regression)*

MULTINOMIALE, LOI

Multinomial distribution

Voir **loi multinomiale**.

MULTIPLICATEUR DE LAGRANGE

Lagrange multiplier

La méthode des multiplicateurs de Lagrange est une méthode d'**optimisation** classique permettant de déterminer l'extremum local d'une fonction sous certaines contraintes.

ASPECTS MATHÉMATIQUES

Soit $f(x, y)$ la fonction objectif qui doit être maximisée ou minimisée et soit la contrainte $g(x, y) = 0$. On forme une fonction auxiliaire appelée un lagrangien :

$$F(x, y, \theta) = f(x, y) - \theta g(x, y)$$

où θ est une **variable** inconnue appelée le multiplicateur de Lagrange.

On cherche les minima et maxima de la fonction F en annulant les dérivées partielles par rapport à x, y et θ.

$$\begin{aligned}\frac{\partial F}{\partial x} = \frac{\partial f}{\partial x} - \theta\frac{\partial g}{\partial x} &= 0\\ \frac{\partial F}{\partial y} = \frac{\partial f}{\partial y} - \theta\frac{\partial g}{\partial y} &= 0\\ \frac{\partial F}{\partial \theta} = g(x,y) &= 0.\end{aligned}$$

Ces trois équations peuvent être résolues pour x, y et θ. La solution donne les points en lesquels la fonction sous contrainte peut présenter un minimum ou un maximum. Il faut encore voir de quelle nature sont ces points, c'est-à-dire s'il s'agit d'un maximum, d'un minimum ou d'un col de la fonction.

On a un maximum si

$$\begin{aligned}\left[\frac{\partial^2 F}{\partial x^2}\right]\cdot\left[\frac{\partial^2 F}{\partial y^2}\right] - \left[\frac{\partial^2 F}{\partial x \partial y}\right]^2 &> 0,\\ \frac{\partial^2 F}{\partial x^2} &< 0\\ \text{et } \frac{\partial^2 F}{\partial y^2} &< 0.\end{aligned}$$

On a un minimum si

$$\begin{aligned}\left[\frac{\partial^2 F}{\partial x^2}\right]\cdot\left[\frac{\partial^2 F}{\partial y^2}\right] - \left[\frac{\partial^2 F}{\partial x \partial y}\right]^2 &> 0,\\ \frac{\partial^2 F}{\partial x^2} &> 0\\ \text{et } \frac{\partial^2 F}{\partial y^2} &> 0.\end{aligned}$$

Si

$$\left[\frac{\partial^2 F}{\partial x^2}\right]\cdot\left[\frac{\partial^2 F}{\partial y^2}\right] - \left[\frac{\partial^2 F}{\partial x \partial y}\right]^2 \leq 0,$$

le test échoue et pour déterminer la nature de la solution, il faut examiner la fonction au voisinage de (x, y).

La méthode des multiplicateurs de Lagrange peut être étendue à une fonction de n variables, $f(x_1, x_2, \ldots, x_n)$ assujettie à k contraintes $g_j(x_1, x_2, \ldots, x_n) = 0$, $j = 1, 2, \ldots, k$ où $k \leq n$.

D'une façon analogue au cas de deux variables, on définit le lagrangien :

$$F = f(x_1, x_2, \ldots, x_n) - \sum_{j=1}^{k} \theta_j \cdot g_j(x_1, x_2, \ldots, x_n)$$

où les paramètres $\theta_1, \ldots, \theta_k$ sont appelés les multiplicateurs de Lagrange. On établit alors les dérivées partielles, qui aboutissent à $n + k$ équations à $n + k$ inconnues donnant les coordonnées de l'extremum de la fonction, si ce dernier existe.

DOMAINES ET LIMITATIONS

La méthode des multiplicateurs de Lagrange ne s'applique pas quand les fonctions ne sont pas différentiables ou quand les contraintes se présentent sous forme d'inégalités.

Dans ces cas, il faut utiliser d'autres méthodes d'**optimisation** choisies parmi les méthodes de la **programmation linéaire**.

EXEMPLE

Soient la fonction objectif :

$$f(x, y) = 5x^2 + 6y^2 - x \cdot y$$

et la contrainte :

$$g(x, y) = x + 2y - 24 = 0.$$

On forme le lagrangien :

$$\begin{aligned}F(x, y, \theta) &= f(x, y) - \theta g(x, y)\\ &= 5x^2 + 6y^2 - x \cdot y - \theta(x + 2y - 24)\end{aligned}$$

et on annule les dérivées partielles :

$$\frac{\partial F}{\partial x} = 10x - y - \theta = 0 \quad (1)$$

$$\frac{\partial F}{\partial y} = 12y - x - 2\theta = 0 \quad (2)$$

$$\frac{\partial F}{\partial \theta} = x + 2y - 24 = 0. \quad (3)$$

En éliminant θ de (1) et (2) :

$$\begin{array}{rcl} 20x - 2y - 2\theta &=& 0\\ \underline{-x + 12y - 2\theta} &=& \underline{0}\\ 21x - 14y &=& 0\\ 3x - 2y &=& 0\end{array}$$

et en remplaçant $2y$ par $3x$ dans (3) :

$$0 = x + 2y - 24 = x + 3x - 24 = 4x - 24,$$

on obtient $x = 6$.

En remplaçant cette **valeur** de x dans (3), on trouve $y = 9$. Le point critique a donc les coordonnées $(6, 9)$. On calcule les dérivées de second ordre pour vérifier s'il s'agit d'un maximum, d'un minimum ou d'un col. On obtient :

$$\begin{aligned} \frac{\partial^2 F}{\partial x^2} &= 10 \\ \frac{\partial^2 F}{\partial y^2} &= 12 \\ \frac{\partial^2 F}{\partial x \partial y} &= -1 \end{aligned}$$

$$\begin{aligned} \left[\frac{\partial^2 F}{\partial x^2}\right] \cdot \left[\frac{\partial^2 F}{\partial y^2}\right] - \left[\frac{\partial^2 F}{\partial x \partial y}\right]^2 &= 10 \cdot 12 - (-1)^2 \\ &= 119 > 0 \\ \frac{\partial^2 F}{\partial x^2} &> 0 \\ \frac{\partial^2 F}{\partial y^2} &> 0. \end{aligned}$$

On constate donc que le point $(6, 9)$ correspond à un minimum pour la fonction objectif f sous la contrainte. Ce minimum vaut $f(6, 9) = 180 + 486 - 54 = 612$.

MOTS CLÉS

Optimisation *(Optimization)*
Programmation linéaire *(Linear programming)*

RÉFÉRENCE

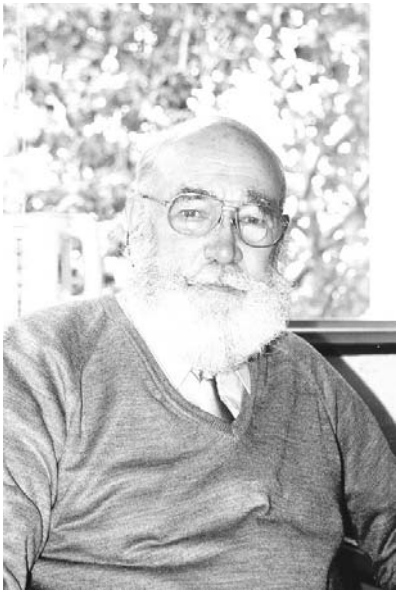
Lagrange, J.L. (1776). Mémoire sur l'utilité de la méthode de prendre le milieu entre les résultats de plusieurs observations ; dans lequel on examine les avantages de cette méthode par le calcul des probabilités; et où l'on résoud différents problèmes relatifs à cette matière. *Miscellanea Taurinensia* 5, pp. 167-232.



N

NELDER JOHN A.

Nelder John A.



John Nelder est né en 1924 à Dulverton (Somerset, Angleterre). Après l'obtention de son diplôme en statistique mathématique à Cambridge, il est nommé président de la section statistique de la National Vegetable Research Station à Wellesbourne, poste qu'il détient de 1951 à 1968. Il obtient le titre de docteur ès sciences à Birmingham et est ensuite élu président du département de statistique à la Rothamsted Experimental Station d'Harpenden de 1968 à 1984. Actuellement, il est professeur invité à l'Imperial College of Science, Technology and Medicine de Londres. Il a été élu membre de la Royal Society en 1984 et il a été président de deux sociétés prestigieuses : l'International Biometric Society et la Royal Statistical Society.

John Nelder est l'auteur des systèmes statistiques computationnelles Genstat et GLIM, actuellement répandus dans plus de 70 pays. Auteur de plus de 140 articles publiées dans des revues statistiques et biologiques, il est également l'auteur de *Computers in Biology* et co-auteur avec Peter McCullagh d'un livre traitant des modèles linéaires généralisés. John Nelder a développé l'idée des *Generally Balanced Designs* qui uniformise la plupart des plans d'expérience classiques.

Quelques ouvrages et articles principaux de John A. Nelder :

1965 (and Mead, R.) A simplex method for function minimization.*Computational Journal*, 7, pp. 303-333.

1971 *Computers in Biology.* London and Winchester : Wykeham Publications (London) Ltd., pp. 149.

1972 (and Wedderburn, R.W.M.) Generalized Linear Models. *Journal of the Royal Statistical Society.* Series A, 135, pp. 370-384.

1974 Genstat : a statistical system. In *COMPSTAT : Proceedings in Computational Statistics.* Vienna : Physica-Verlag.

1983 (and McCullagh, P.) *Generalized Linear Models.* Chapman & Hall, London, pp. 261.

MOT CLÉ

Logiciel statistique *(Statistical softwar)*

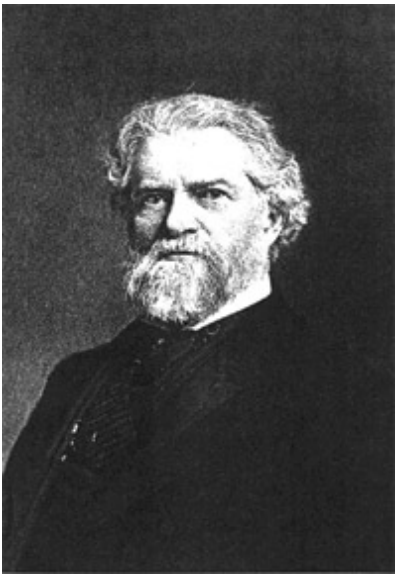
NEWCOMB SIMON

Newcomb Simon

Simon Newcomb est né en 1835 à Nova Scotia (Canada). Il est mort en 1909. Cet astronome d'origine canadienne contribua au traitement des observations aberrantes en statistique, à l'application de la théorie probabiliste aux données ainsi qu'au développement de ce qu'on appelle aujourd'hui les méthodes robustes de statistique.

Jusqu'à l'âge de 16 ans, il étudia essentiellement en consultant les livres que son père obtenait pour lui et,

à partir de ce moment, il commença un apprentissage chez un docteur. En 1857, il fut engagé comme calculateur à l'American Ephemeris and Nautical Almanac à Cambridge dans le Massachussetts. En même temps, il fit des études à Harvard et obtint sa licence ès sciences en 1858. En 1861, il devint professeur de mathématiques à l'Observatoire naval.



La collection de ses *Notes sur la théorie des probabilités* (*Mathematics Monthly*, 1859-61) constitue un ouvrage qui est aujourd'hui encore vu comme étant moderne.

La contribution la plus notable de Newcomb aux statistiques est son approche du traitement des aberrations dans les données astronomiques ; constatant que la distribution normale ne convenait pas, il inventa la distribution normale contaminée.

Quelques ouvrages et articles principaux de Simon Newcomb :

1859-61 Notes on the Theory of Probabilities. *Mathematics Monthly*, 1, pp. 136-139, 233-235, 331-355 and 349-350 ; 2, pp. 134-140 and 272-275 ; 3, pp. 68, 119-125 and 341-349).

NEYMAN JERZY

Neyman Jerzy



Jerzy Neyman (1894-1981) fut un des fondateurs de la **statistique** moderne, il posa notamment avec **Egon Sharpe Pearson** les bases de la théorie des tests d'hypothèse.

Né en Russie, Jerzy Neyman fit des études de physique et de mathématiques à l'Université de Karkov. En 1921, il revint en Pologne, le pays d'origine de sa famille, où il travailla quelque temps comme statisticien au National Institute of Agriculture de Bydgoszcz.

En 1924, il fit un séjour à Londres où il put étudier à l'University College sous la direction de **Karl Pearson**. Il eut à cette époque l'occasion de rencontrer Egon Sharpe Pearson, **William Sealy Gosset** et **Ronald Aylmer Fisher**.

En 1937, à l'issue d'un voyage aux États-Unis où il donna plusieurs conférences, il accepta le poste de professeur à l'Université de Berkeley. Il y créa le département de statistique et termina dans cette université sa brillante carrière.

Quelques articles principaux de J. Neyman :

1928 (and Pearson, E.S.) On the use and interpretation of certain test criteria for purposes of statistical inference. I, II. *Biometrika*, 20A, pp. 175-240 and 263-295.

1933 (and Pearson, E.S.) Testing of statistical hypotheses in relation to probabilities a priori. *Proceedings of the Cambridge Philosophical Society*, 29, pp. 492-510.

1934 On the two different aspects of the representative method : the method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, 97, pp. 558-606, discussion pp. 607-625.

1938 Contribution to the theory of sampling human populations. *Journal of the American Statistical Association*, 33, pp. 101-116.

MOT CLÉ

Test d'hypothèse *(Hypothesis testing)*

NIVEAU D'UN FACTEUR

Factor level

On appelle niveaux d'un **facteur** les différentes valeurs que peut prendre celui-ci. Ainsi un **plan d'expérience** détermine parmi toutes les combinaisons possibles des niveaux des différents facteurs lesquelles sont soumises à l'**expérience**.

MOTS CLÉS

Expérience *(Experiment)*
Facteur *(Factor)*
Plan d'expérience *(Design of experiments)*
Traitement *(Treatment)*

NIVEAU DE CONFIANCE

Confidence level

On appelle niveau de confiance la **probabilité** que l'**intervalle de confiance** construit autour d'un **estimateur** contienne la vraie **valeur** du **paramètre** correspondant de la **population**.

On désigne le niveau de confiance par $(1 - \alpha)$ où α correspond au risque d'**erreur**, c'est-à-dire à la probabilité que l'intervalle de confiance ne contienne pas la vraie valeur du paramètre.

HISTORIQUE

Le premier exemple d'un intervalle de confiance apparaît dans le travail de Laplace (1812). Selon A. Desrosières (1988), A.L. Bowley est l'un des premiers à s'intéresser à la notion d'intervalle de confiance.

Voir **test d'hypothèse**.

ASPECTS MATHÉMATIQUES

Soit θ un paramètre à estimer de la population et T son estimateur à partir d'un échantillon aléatoire. On évalue la précision de T comme estimateur de θ en construisant un **intervalle de confiance** autour de l'estimateur, qui souvent s'interprète comme une marge d'erreur. Pour construire cet intervalle de confiance, on procède, en terme général, de la manière suivante.

À partir de la loi de distribution de l'estimateur T, on détermine un intervalle calculé sur la base de l'échantillon tel que la probabilité qu'il englobe la

vraie valeur du paramètre recherché soit importante. Soit $(T-\varepsilon, T+\varepsilon)$ cet intervalle et $(1-\alpha)$ la probabilité d'appartenance, on peut dire que la marge d'erreur e est liée à α par la probabilité :

$$P(T-\varepsilon \leq \theta \leq T+\varepsilon) = 1-\alpha.$$

Le niveau de probabilité associé à un intervalle d'estimation est appelé niveau de confiance ou degré de confiance.

L'intervalle, $T-\varepsilon \leq \theta \leq T+\varepsilon$, est appelé intervalle de confiance de l'estimateur de θ au niveau de confiance $1-\alpha$. Prenons comme exemple $\alpha = 5$ %, l'intervalle de confiance du paramètre θ à un seuil de probabilité de 95 %. Ceci veut dire qu'en utilisant T comme estimateur de θ, en moyenne, sur 100 échantillonnages, 95 fois l'intervalle construit de la façon indiquée comprendra la vraie valeur de l'estimateur et 5 fois il ne l'incluera pas.

La quantité ε de l'intervalle de confiance mesure la moitié de l'étendue de l'intervalle. Elle indique donc, dans un certain sens, la marge d'erreur de l'estimateur. Un estimateur est d'autant plus efficace que, pour un niveau de confiance $1-\alpha$ donné, il conduit à un intervalle de confiance plus petit.

DOMAINES ET LIMITATIONS

Les valeurs de niveau de confiance les plus utilisées dans la construction d'intervalles de confiance sont 90 %, 95 % et 99 %. Cependant, si nécessaire, on peut choisir d'autres valeurs.

On est tenté de choisir le niveau de confiance le plus élevé afin d'augmenter la **probabilité** que l'intervalle de confiance contienne la vraie valeur du **paramètre**. Cependant plus le niveau de confiance est élevé, plus l'**intervalle** est large. Par conséquent, ce que l'on gagne en sûreté, on le perd en précision. Il faut donc trouver un compromis.

EXEMPLE

Une entreprise fabriquant des ampoules électriques veut étudier la durée de vie moyenne de ses ampoules. La distribution de la **variable aléatoire** X qui représente la durée de vie en heures est une **loi normale** de **moyenne** μ et d'**écart type** $\sigma = 30$.

Pour estimer μ, l'entreprise fait brûler un **échantillon** aléatoire de $n = 25$ ampoules.

Elle obtient un temps moyen de durée $\bar{x} = 860$ heures. Elle veut construire un **intervalle de confiance** pour μ autour de l'**estimateur** $\bar{x}$ à un niveau de confiance égal à $0,95$.

Quand l'écart type σ de la population est connu, la valeur de ε est égale à $z_{\alpha/2} \cdot \sigma_{\bar{X}}$. La valeur de $z_{\alpha/2}$ se lit dans la table de Gauss en fonction de la probabilité attribuée au paramètre α. On en déduit donc l'intervalle de confiance de l'estimateur de μ, au seuil de probabilité $1-\alpha$:

$$\bar{X} - z_{\alpha/2}\sigma_{\bar{X}} \leq \mu \leq \bar{X} + z_{\alpha/2}\sigma_{\bar{X}}.$$

Partant de l'**hypothèse** que la durée de vie X des ampoules suit une loi normale de moyenne μ et d'écart type $\sigma = 30$, on déduit que l'expression :

$$\frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}}$$

suit une loi normale centrée réduite.

D'où on obtient :

$$P\left[-z_{0,025} \leq \frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}} \leq z_{0,025}\right] = 0,95$$

puisque le risque d'**erreur** α a été divisé en deux parties égales à $\frac{\alpha}{2} = 0,025$.

La table de la loi normale centrée réduite, ou **table normale**, donne $z_{0,025} = 1,96$. On a donc :

$$P\left(-1,96 \leq \frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}} \leq 1,96\right) = 0,95.$$

Pour obtenir l'intervalle de confiance pour μ, il faut isoler μ dans l'équation ci-dessus par les transformations suivantes :

$$P\left(-1,96\frac{\sigma}{\sqrt{n}} \leq \bar{x}-\mu \leq 1,96\frac{\sigma}{\sqrt{n}}\right) = 0,95$$

$$P\left(\bar{x}-1,96\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x}+1,96\frac{\sigma}{\sqrt{n}}\right) = 0,95.$$

En remplaçant $\bar{x}$, σ et n par leur **valeur** respective, on évalue l'intervalle de confiance voulu :

$$860 - 1,96 \cdot \frac{30}{\sqrt{25}} \leq \mu \leq 860 + 1,96 \cdot \frac{30}{\sqrt{25}}$$

$$848,24 \leq \mu \leq 871,76.$$

Ceci signifie qu'on peut affirmer avec un niveau de confiance de 95 % que cet **intervalle** contient la vraie valeur du **paramètre** μ qui correspond à la durée de vie moyenne des ampoules.

MOTS CLÉS

Estimateur *(Estimator)*
Estimation *(Estimation)*
Intervalle de confiance *(Confidence interval)*
Test d'hypothèse *(Hypothesis testing)*

NOMBRE ALÉATOIRE

Random number

On appelle nombre aléatoire la réalisation d'une **variable aléatoire** distribuée uniformément dans l'intervalle $[0, 1]$. Un ensemble de nombres aléatoires est donc un **échantillon** issu de la **loi uniforme**.

DOMAINES ET LIMITATIONS

Si les conditions de définition ne sont pas vérifiées rigoureusement, c'est-à-dire si la distribution dont on tire les nombres aléatoires n'est pas tout à fait une **loi uniforme** et que les tirages successifs des nombres ne sont pas tout à fait indépendants, on dit que l'ensemble des nombres aléatoires obtenus forme une suite de nombres pseudo-aléatoires.

EXEMPLE

Une urne contient 10 boules numérotées de 0 à 9. Si l'on désire obtenir un nombre aléatoire, compris entre 0 et 1 avec m décimales, il suffit de tirer avec remise m boules dans le but de former un nombre de m chiffres. Puis, en considérant ces tirages comme faisant partie d'un seul nombre, il s'agit de le diviser par 10^m.

À titre d'exemple, on suppose que l'on ait tiré successivement les cinq boules suivantes numérotées : 8, 7, 5, 4 et 7. Le nombre aléatoire correspondant est $0,87547$.

MOTS CLÉS

Génération de nombres aléatoires *(Generation of random numbers)*
Loi uniforme *(Uniform distribution)*

NON PARAMÉTRIQUE, TEST

Nonparametric test

Voir **test non paramétrique**.

NORMALE, LOI

Normal distribution

Voir **loi normale**.

NORMALE, TABLE

Normal table

Voir **table normale** et annexe G.

NORMALISATION

Normalisation

La normalisation est le passage d'une **variable aléatoire** qui suit une **loi normale** à une variable aléatoire qui suit une loi normale centrée réduite. Elle permet de calculer et de comparer des valeurs appartenant à des courbes normales de **moyenne** et de **variance** différentes sur la base d'une loi normale de référence qui est la loi normale centrée réduite.

ASPECTS MATHÉMATIQUES

Soit la **variable aléatoire** X suivant la **loi normale** $\mathcal{N}(\mu, \sigma^2)$, sa normalisation nous donne une variable aléatoire :

$$Z = \frac{X - \mu}{\sigma}$$

qui suit la loi normale centrée réduite $\mathcal{N}(0, 1)$.

Chaque **valeur** de la distribution $\mathcal{N}(\mu, \sigma^2)$ peut être transformée en **variable** centrée réduite Z, chaque Z représentant un écart à la **moyenne** exprimé en unité d'**écart type**.

EXEMPLE

Les élèves d'une école professionnelle ont subi deux épreuves. Chaque épreuve a été notée sur une échelle de 1 à 60 et les scores sont considérés comme étant des réalisations de deux variables aléatoires de **loi normale**. Essayons de comparer les scores obtenus par un élève à ces deux épreuves.

Voici les moyennes et les écarts types de chaque épreuve calculés sur l'ensemble des élèves :

Épreuve 1 : $\mu_1 = 35$; $\sigma_1 = 4$

Épreuve 2 : $\mu_2 = 45$; $\sigma_2 = 1,5$.

L'élève Marc a obtenu les résultats suivants :

Épreuve 1 : $X_1 = 41$

Épreuve 2 : $X_2 = 48$.

La question est de savoir dans quel test l'élève Marc a le mieux réussi, comparativement à l'ensemble des élèves de l'école.

Nous ne pouvons pas comparer directement les résultats obtenus dans les deux épreuves puisque ces résultats appartiennent à des distributions de moyenne et d'écart type différents.

Si nous examinons simplement la différence de chaque note à la moyenne de sa distribution, nous obtenons :

Épreuve 1 : $X_1 - \mu_1 = 41 - 35 = 6$

Épreuve 2 : $X_2 - \mu_2 = 48 - 45 = 3$.

Nous constatons que Marc a obtenu 6 points de plus que la moyenne dans l'épreuve 1 alors qu'il n'a obtenu que 3 points de plus que la moyenne dans l'épreuve 2. Une conclusion hâtive nous ferait dire que Marc, comparativement à l'ensemble des élèves, a mieux réussi l'épreuve 1 que l'épreuve 2.

Toutefois cette conclusion ne tient compte que de la différence de chaque résultat à la moyenne. Elle néglige la dispersion des notes des élèves autour de chaque moyenne. Nous allons donc diviser la différence à la moyenne par l'écart type pour rendre les résultats comparables :

$$\text{Épreuve 1 : } Z_1 = \frac{X_1 - \mu_1}{\sigma_1} = \frac{6}{4} = 1,5$$

$$\text{Épreuve 2 : } Z_2 = \frac{X_2 - \mu_2}{\sigma_2} = \frac{3}{1,5} = 2.$$

Par ce calcul, nous avons normalisé les scores X_1 et X_2. Nous pouvons conclure que la **valeur** normalisée étant plus élevée pour l'épreuve 2 ($Z_2 = 2$) que pour l'épreuve 1 ($Z_1 = 1,5$), l'élève Marc a, comparativement aux autres élèves, mieux réussi l'épreuve 2 que l'épreuve 1.

MOT CLÉ

Loi normale *(Normal distribution)*

NORME D'UN VECTEUR

Norm of a vector

La norme d'un **vecteur** indique la longueur de ce vecteur définie à partir du **produit scalaire** du vecteur par lui-même. La norme d'un vecteur s'obtient en prenant la racine carrée de ce produit scalaire.

Un vecteur de norme 1 est appelé vecteur unitaire.

ASPECTS MATHÉMATIQUES

Pour un **vecteur x** donné,

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

on définit le **produit scalaire** de $\mathbf{x}$ par lui-même par :

$$\mathbf{x}' \cdot \mathbf{x} = \sum_{i=1}^{n} x_i^2.$$

La norme du vecteur $\mathbf{x}$ vaut alors :

$$\| \mathbf{x} \| = \sqrt{\mathbf{x}' \cdot \mathbf{x}} = \sqrt{\sum_{i=1}^{n} x_i^2}.$$

DOMAINES ET LIMITATIONS

Comme la norme d'un **vecteur** se calcule à l'aide du **produit scalaire**, les propriétés de la norme découlent de celles du produit scalaire ; ainsi :

- la norme d'un vecteur $\mathbf{x}$ est strictement positive si le vecteur est non nul, et nulle sinon,

$$\begin{aligned} \| \mathbf{x} \| &> 0 \quad \text{si } \mathbf{x} \neq 0 \\ \text{et } \| \mathbf{x} \| &= 0 \quad \text{si et seulement si } \mathbf{x} = 0 \text{ ;} \end{aligned}$$

- la norme du résultat d'une multiplication d'un vecteur $\mathbf{x}$ par un scalaire est égale au produit de la norme de $\mathbf{x}$ par la valeur absolue du scalaire,

$$\| k \cdot \mathbf{x} \| = |k| \cdot \| \mathbf{x} \|,$$

- pour deux vecteurs $\mathbf{x}$ et $\mathbf{y}$, on a :

$$\| \mathbf{x} + \mathbf{y} \| \leq \| \mathbf{x} \| + \| \mathbf{y} \|,$$

  c'est-à-dire que la norme d'une somme de deux vecteurs est inférieure ou égale à la somme des normes de ces vecteurs ;

- pour deux vecteurs $\mathbf{x}$ et $\mathbf{y}$, la valeur absolue du produit scalaire de $\mathbf{x}$ par $\mathbf{y}$ est inférieure ou égale au produit des normes,

$$|\mathbf{x}' \cdot \mathbf{y}| \leq \| \mathbf{x} \| \cdot \| \mathbf{y} \| .$$

  Ce résultat est appelé l'inégalité de Cauchy-Schwarz.

La norme d'un vecteur s'utilise aussi pour obtenir un vecteur unitaire ayant la même direction qu'un vecteur donné. Il suffit dans ce cas de diviser le vecteur initial par sa norme :

$$\mathbf{y} = \frac{\mathbf{x}}{\| \mathbf{x} \|}.$$

EXEMPLE

On considère le **vecteur** suivant défini dans l'espace euclidien à trois dimensions :

$$\mathbf{x} = \begin{bmatrix} 12 \\ 15 \\ 16 \end{bmatrix}$$

Le **produit scalaire** de ce vecteur par lui-même vaut :

$$\begin{aligned} \mathbf{x}' \cdot \mathbf{x} &= 12^2 + 15^2 + 16^2 \\ &= 144 + 225 + 256 = 625 \end{aligned}$$

d'où l'on obtient la norme de $\mathbf{x}$:

$$\| \mathbf{x} \| = \sqrt{\mathbf{x}' \cdot \mathbf{x}} = \sqrt{625} = 25.$$

Le vecteur unitaire ayant la direction de $\mathbf{x}$ est obtenu en divisant chacune des composantes de $\mathbf{x}$ par 25 :

$$\mathbf{y} = \begin{bmatrix} 0,48 \\ 0,6 \\ 0,64 \end{bmatrix}$$

et on peut vérifier que $\| \mathbf{y} \| = 1$. En effet,

$$\begin{aligned} \| \mathbf{y} \|^2 &= 0,48^2 + 0,6^2 + 0,64^2 \\ &= 0,2304 + 0,36 + 0,4096 = 1. \end{aligned}$$

MOTS CLÉS

Estimation L1 *(L1 estimation)*
Moindres carrés *(Least squares)*
Produit scalaire *(Scalar product)*
Vecteur *(Vector)*

RÉFÉRENCE

Dodge, Y. (2002). *Mathématiques de base pour économistes.* Springer-Verlag France, Paris.


O

OBSERVATION

Observation

Une observation est le résultat d'une étude scientifique rassemblant des informations sur une unité statistique appartenant à un **échantillon** ou à une **population** donnés.

MOTS CLÉS

Donnée *(Data)*
Échantillon *(Sample)*
Observation aberrante *(Outlier)*
Population *(Population)*

OBSERVATION ABERRANTE

Outlier

Une **observation** est dite aberrante quand sa **valeur** s'écarte fortement des valeurs des autres observations d'un ensemble.

Il est à noter qu'une observation ne peut être jugée aberrante qu'en fonction de la forme supposée de la distribution sous-jacente de l'ensemble des observations. Si la forme supposée est modifiée, l'observation peut alors devenir concordante avec le nouveau **modèle**.

HISTORIQUE

Le problème des observations aberrantes a été considéré par plusieurs scientifiques depuis 1750. On peut citer notamment les travaux de **Roger Joseph Boscovich**, **Pierre Simon de Laplace** et **Adrien-Marie Legendre**.

Selon Stephen M. Stigler (1973), A.-M. Legendre proposa, en 1805, le rejet des observations aberrantes ; et en 1852, Benjamin Peirce établit un premier critère pour déterminer les observations aberrantes. Des critiques émanèrent, notamment de la part de George Biddell Airy (1856), face à l'utilisation d'un tel critère.

DOMAINES ET LIMITATIONS

Dans le cadre d'une étude statistique, nous pouvons aborder la question des observations aberrantes de deux façons différentes :

- la première consiste à s'accommoder de ce type d'observation. Le but est alors de faire des inférences, à partir de l'**échantillon**, qui soient le moins possible affectées par les valeurs de chaque observation ;

- la seconde vise à éliminer les observations aberrantes. Après les avoir identifiées, on les retire de l'échantillon avant d'effectuer l'analyse statistique.

Le problème des observations aberrantes a conduit à rechercher de nouvelles méthodes statistiques dont font partie les méthodes statistiques dites robustes (voir **estimation robuste**).

MOTS CLÉS

Analyse des résidus *(Analysis of residuals)*
Estimation L1 *(L1 estimation)*
Estimation robuste *(Robust estimation)*
Matrice H *(Hat matrix)*
Observation *(Observation)*
Point levier *(Leverage point)*
Valeur *(Value)*

RÉFÉRENCES

Airy, G.B. (1856). Letter from Professor Airy, Astronomer Royal, to the editor. *Astronomical Journal*, 4, pp. 137-138.

Barnett, V. and Lewis, T. (1984). *Outliers in statistical Data*. John Wiley & Sons, New York.

Peirce, B. (1852). Criterion for the Rejection of Doubtful Observations. *Astronomical Journal*, 2, pp. 161-163.

Stigler, S. (1973). Simon Newcomb, Percy Daniell, and the History of Robust Estimation 1885-1920. *Journal of the American Statistical Association*, 68, pp. pp. 872-879.

OBSERVATION MANQUANTE

Missing data

On parle d'observations manquantes lorsque dans une étude statistique certains résultats font défaut. C'est le cas si, par accident, les données ne sont pas disponibles (récolte détruite par les intempéries, **erreur** dans l'enregistrement, mort d'un animal en cours d'**expérience**, etc.) mais cela peut aussi être structurel ou volontaire pour diverses raisons.

HISTORIQUE

Il semble que Allan et Wishart (1930) ont été les premiers à considérer le problème des observations manquantes dans l'analyse des plans d'expérience. Leurs travaux ont été suivis par ceux de Frank Yates (1933).

Allan et Wishart ainsi que F. Yates ont essayé de trouver une **estimation** des valeurs manquantes par la méthode d'itération. Dans la situation où plusieurs observations sont manquantes, la méthode proposée ne donne toutefois pas une solution satisfaisante.

M.S. Bartlett (1937) a utilisé l'**analyse de covariance** pour résoudre le même problème.

Depuis lors, beaucoup d'autres méthodes ont été proposées, parmi lesquelles on peut citer la méthode de Rubin (1972). Birkes, Dodge et Seely (1976) présentent une méthode exacte pour résoudre le problème des observations manquantes, en utilisant la **matrice** d'**incidence**. L'approche matricielle a été proposée par Dodge et Majumdar (1979).

De plus amples détails concernant l'aspect historique des observations manquantes depuis 1933 à 1985 se trouvent dans Dodge (1985).

ASPECTS MATHÉMATIQUES

On peut pallier le problème des observations manquantes par la répétition de l'étude sous les mêmes conditions et obtenir ainsi de nouvelles valeurs pour les observations manquantes. Toutefois, une telle solution, bien qu'idéale, n'est pas toujours possible économiquement ou matériellement.

Dans le cas des plans d'expérience, la solution la plus fréquemment utilisée consiste à appliquer l'analyse pour un plan d'expérience complet, en prenant des valeurs calculées à la place des observations manquantes. Toutefois, dans ce cas, l'analyse est approximative.

Il existe d'autres approches pour résoudre le problème des observations manquantes dans les plans d'expérience ; ainsi l'approche par l'**inverse généralisé** ou par la **matrice** d'indicence indique quelles observations sont manquantes dans le plan.

MOTS CLÉS

Donnée *(Data)*
Expérience *(Experiment)*
Observation *(Observation)*
Plan d'expérience *(Design of experiments)*

RÉFÉRENCES

Allan, F.E. and Wishart, J. (1930). A method of estimating the yield of a missing plot in field experimental work. *Journal of Agricultural Science*, 20, pp. 399-406.

Bartlett, M.S. (1937). Some Examples of Statistical Methods of Research in Agriculture and Applied Biology. *Journal of the Royal Statistical Society*, (Suppl.), 4, pp. 137-183.

Birkes, D., Dodge, Y. and Seely, J. (1976). Spanning sets for estimable contrasts in classification models. *Ann. Stat.* 4, pp. 86-107.

Dodge, Y. (1985). *Analysis of experiments with missing data.* John Wiley & Sons, New York.

Dodge, Y. and Majumdar, D. (1979). An Algorithm for Finding Least Square Generalized Inverses for Classification Models with Arbitrary Patterns. *Journal of Statistical Computation and Simulation*, Vol. 9, pp. 1-17.

Rubin, D.B. (1972). A non-iterative algorithm for least squares estimation of missing values in any analysis of variance design. *Applied Statistics*, 21, 136-141.

Yates, F. (1933). The Analysis of Replicated Experiments when the Field Results are Incomplete. *Empire Journal of Experimental Agriculture*, 1, pp. 129-142.

ODDS ET ODDS RATIO

Odds and odds ratio

Un odds est défini comme le rapport entre la probabilité d'un événement et la probabilité de l'événement complémentaire, c'est-à-dire le rapport entre la probabilité de survenue d'un événement (dans le cas particulier de l'épidémiologie : une maladie) et la probabilité qu'il ne survienne pas.

L'odds ratio est, au même titre que le **risque relatif**, une mesure de l'effet relatif d'une cause. Plus précisément, c'est le rapport des odds calculés au sein de populations exposées diversement à un facteur de risque.

L'acception française généralement admise du terme odds ratio est : « rapport des cotes ». Cependant, c'est la terminologie anglaise qui s'impose le plus largement.

ASPECTS MATHÉMATIQUES

Un odds attaché à l'événement « contracter une maladie » est, formellement, le rapport suivant, où p désigne la probabilité d'occurence de la maladie :

$$odds = \frac{p}{1-p}.$$

Soit un facteur de risque, auquel un groupe E est exposé et un groupe NE n'est pas exposé. L'odds ratio correspondant, abrégé OR, est le rapport entre l'odds du groupe E et l'odds du groupe NE :

$$OR = \frac{\text{odds du groupe E}}{\text{odds du groupe NE}}.$$

DOMAINES ET LIMITATIONS

Prenons l'exemple du calcul de l'odds ratio à partir des données du tableau suivant, comparant deux groupes d'individus atteints d'une maladie potentiellement grave et dont l'un (groupe « traité ») reçoit un traitement médical, l'autre non (groupe « de référence ») :

Groupe	Décès	Survie	Odds	OR
Traité	5 %	95 %	0,053	0,47
Référence	10 %	90 %	0,111	1,0*

* Par définition l'OR vaut 1 pour le groupe de référence.

Si l'on appelle ici « **risque** » le risque attaché à l'événement « décès », l'OR est une bonne approximation du risque relatif, à la condition que les risques de décès soient proches ou bien, s'ils diffèrent notablement, qu'ils soient faibles. On considère généralement un risque inférieur à 10 %, à cet égard, comme étant faible. L'odds ratio n'en demeure pas moins qu'une approximation du risque relatif, car en fait il y a un **biais** de l'OR qui intervient pour des risques relatifs sensiblement différents de l'unité. En effet :

$$OR = \frac{\dfrac{\text{risque dans le groupe traité}}{\text{taux de survie dans le groupe traité}}}{\dfrac{\text{risque dans le groupe de référence}}{\text{taux de survie dans le groupe de référence}}}$$

$$= \frac{\text{risque dans le groupe traité}}{\text{risque dans le groupe de référence}} \times \frac{\text{taux de survie dans le groupe de référence}}{\text{taux de survie dans le groupe traité}}$$

$$= \text{risque relatif} \times \text{biais de l'odds ratio}$$

avec :

$$\text{biais de l'OR} = \frac{\text{taux de survie dans le groupe de référence}}{\text{taux de survie dans le groupe traité}}.$$

Les estimations du risque relatif et de l'odds ratio sont utilisées pour mesurer la force de l'association causale entre l'exposition à un facteur de risque et une maladie. Les deux rapports peuvent prendre des valeurs comprises entre 0 et l'infini et qui sont justiciables d'une même interprétation générale. On peut distinguer deux cas nettement différents :

1. une valeur de l'OR supérieure à 1 indique un risque ou un odds de la maladie plus grand lorsque le groupe étudié est exposé à un facteur de risque spécifique. On parle alors d'une association positive ;
2. une valeur inférieure à un implique un risque réduit à l'exposition du facteur risque. On parle alors d'une association négative.

Intervalle de confiance à 95 % d'un odds ratio

La technique et la notation utilisées pour l'intervalle de confiance à 95 % de l'odds ratio sont les mêmes que celles du risque relatif.

$$exp\left\{(lnOR) \pm 1,96\sqrt{\frac{1}{a}-\frac{1}{n}+\frac{1}{c}-\frac{1}{m}}\right\}.$$

Il est requis que a,b,c et d soient grands. Si l'intervalle de confiance inclut la valeur 1,0, l'odds ratio n'est pas statistiquement significatif.

EXEMPLE

Soit l'étude de la relation entre le facteur de risque « tabagisme » et le cancer du sein. L'odds du cancer du sein est le rapport entre la probabilité de développer ce cancer et la probabilité de ne pas avoir cette maladie, au cours d'une période donnée. Il y a un odds de cancer du sein pour chaque groupe défini par son niveau d'exposition au tabagisme. Le tableau suivant illustre cet exemple :

Groupe	Cancer (/100 000/ 2 ans)	Pas de cancer (/100 000 /2 ans)	Odds du cancer du sein (/2 ans)	OR
Non-exposées	114	99 886	0,00114	1,0
Fumeuses passives	252	99 748	0,00252	2,2
Fumeuses actives	276	99 724	0,00276	2,4

Dans cet exemple, l'odds du cancer du sein des non-exposées est de 0,00114 (114 : 99 886) ; celui des fumeuses passives est de 0,00252 (252 : 99 748) et celui des fumeuses actives de 0,00276 (276 : 99 274). en prenant pour groupe de référence celui des non-exposées, l'odds ratio du cancer du sein est, pour les fumeuses actives :

$$\frac{\frac{276}{99\,724}}{\frac{114}{99\,886}} = 2,4 = \frac{276}{114}\cdot\frac{99\,886}{99\,724}.$$

Le risque du cancer du sein sur deux ans étant très faible, la survie des fumeuses actives comme celle des non-exposées est proche de 100 %, et par conséquent le rapport des survies (le biais de l'odds ratio) est pratiquement égal à 1,0, ce qui s'interprète comme une quasi-absence de biais de l'OR. Dans cet exemple, l'odds ratio de cancer du sein est identique au risque relatif de cancer du sein que cela soit pour les fumeuses passives ou pour les fumeuses actives. Le cancer du sein étant une maladie rare, l'odds ratio est une bonne approximation du risque relatif.

MOTS CLÉS

Cause et effet en épidémiologie *(Cause and effect in epidemiology)*
Incidence *(Incidence)*
Prévalence *(Prevalence)*
Odds et odds ratio *(Odds and odds ratio)*
Risque *(Risk)*
Risque relatif *(Relative risk)*
Risque attribuable *(Attributable risk)*
Risque évitable *(Avoidable risk)*
Taux d'incidence *(Incidence rate)*
Taux de prévalence*(Prevalence rate)*

RÉFÉRENCES

Lilienfeld A.M. and D.E. Lilienfeld. (1980). *Foundations of epidemiology.* 2nd ed New York: Clarendon Press, Oxford.

MacMahon, B. and Pugh, T.F. (1970). *Epidemiology: Principles and Methods.* Boston: Little Brown.

Morabia A. (1996). *Épidémiologie causale.* Éditions Médecine et Hygiène, Genève.

Morabia A. (1996). *L'Épidémiologie clinique.* Presses universitaires de France, Que sais-je ? Paris.

OGIVE

Ogive

Une ogive est une **représentation graphique** d'une **distribution de fréquences** cumulées. C'est une forme de graphique de fréquences aussi appelée **polygone de fréquences** cumulées.

Elle sert à donner le nombre (ou la proportion) d'observations inférieures ou égales à une **valeur** particulière.

ASPECTS MATHÉMATIQUES

L'ogive se construit sur un système d'axes perpendiculaires. On place sur l'axe horizontal les limites des intervalles de classe préalablement déterminées. Sur chacune de ces valeurs limites, on détermine en ordonnée une hauteur égale à la **fréquence** cumulée correspondant à cette valeur. En joignant par des segments de droite les points successifs ainsi déterminés, on obtient une ligne brisée appelée ogive.

On a l'habitude de placer la fréquence cumulée sur l'axe vertical gauche et le **pourcentage** de fréquences cumulé sur l'axe vertical droit.

DOMAINES ET LIMITATIONS

L'ogive est spécialement utile pour estimer les centiles dans la distribution. Par exemple, on peut connaître le point central, de sorte que 50 % des observations soient en-dessous de ce point et 50 % au-dessus.

Pour ce faire, on tire une ligne à partir du point des 50 % sur l'axe des pourcentages jusqu'à ce qu'elle intercepte la courbe. Ensuite on projette verticalement l'intersection sur l'axe horizontal. Cette dernière intersection nous donne la **valeur** recherchée.

Le **polygone de fréquences** ainsi que l'ogive sont utilisés pour comparer deux ensembles statistiques dont les effectifs pourraient être très différents.

EXEMPLE

À partir d'un tableau représentant la distribution du chiffre d'affaires d'une journée de 75 épiceries, nous allons construire une ogive.

Distribution du chiffre d'affaires d'une journée de 75 épiceries (en dollars) :

Chiffre d'affaires (intervalles de classes)	Fréquence	Fréquence cumulée	Pourcentage de fréquence cumulé
$215 - 235$	4	4	$5,3$
$235 - 255$	6	10	$13,3$
$255 - 275$	13	23	$30,7$
$275 - 295$	22	45	$60,0$
$295 - 315$	15	60	$80,0$
$315 - 335$	6	66	$88,0$
$335 - 355$	5	71	$94,7$
$355 - 375$	4	75	$100,0$
Total	75		

Dans cet exemple, on obtient l'**estimation** que 50 % des épiceries ont un chiffre d'affaires inférieur ou égal à environ 289 dollars.

MOTS CLÉS

Distribution de fréquences *(Frequency distribution)*
Polygone de fréquences *(Frequency polygon)*
Représentation graphique *(Graphical representation)*

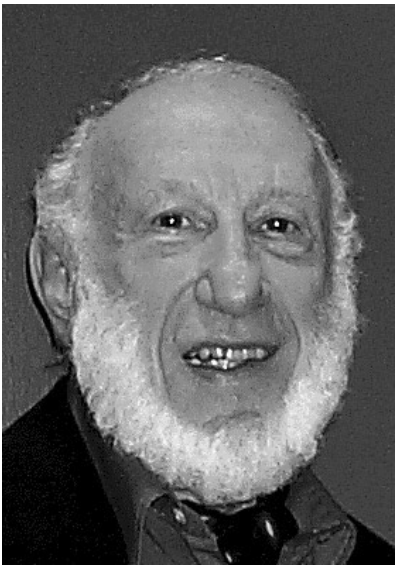
OLKIN INGRAM

Olkin Ingram

Ingram Olkin est né en 1924 à Waterbury (Connecticut) aux États-Unis. Diplômé en 1947 du College of the City à New York, il continue ses études à l'Université de Columbia où il obtient un diplôme postgrade en statistique (1949) puis le titre de docteur en mathématiques statistiques à l'Université de Caroline du Nord en 1951. Il enseigne ensuite à l'Université d'État de Michigan et à l'Université du Minnesota avant d'être nommé professeur de statistique à l'Université de Stanford, poste qu'il occupe actuellement.

Ingram Olkin a été éditeur des *Annals of Statistics*, éditeur associé de *Psychometrika*, du *Journal of Educational Statistics*, du *Journal of The American Statistical Association* et du *Linear Algebra and its Application* pour ne citer qu'une partie des journaux mathématiques et statistiques. Ses recherches se sont focalisées sur l'**analyse multivariée**, les inégalités, la théorie et l'application de la méta-analyse dans les sciences sociales et la biologie.

Il est co-auteur de nombreux ouvrages parmi lesquels : *Inequalities Theory of Majorization and its Applications* (1979), *Selecting and Ordering Populations* (1980), *Probability Models and Applications* (1980) et plus récemment *Statistical Methods in Meta-Analysis* (1985).



Quelques ouvrages et articles d'Ingram Olkin :

1979 (and Marshall, A.) *Inequalities : theory of majorization and its applications.* Academic Press, New York.

1984 (and Marsaglia, G) Generating Correlation Matrices. *SIAM Journal of Scientific and Statistical Computing* 5, pp. 470-475.

1985 (and Hedges, L.V.) *Statistical Methods for Meta-Analysis.* Academic Press, New York.

1994 (with Gleser, L. J. and Derman, C.) *Probability Models and Applications.* The Macmillan Company, New York.

1995 (and Moher, D.) Meta-analysis of randomised controlled trials. A concern for standards. *Journal of the American Medical Association*, 274, pp. 1962-1963.

1996 (with Begg C., Cho M., Eastwood S., Horton R., Moher D., Pitkin R., Rennie D., Schulz KF., Simel D. & Stroup DF.) Improving the quality of reporting of randomized controlled trials. The CONSORT statement. *Journal of the American Medical Association*, 276(8), pp. 637-639.

OPTIMISATION

Optimization

Un grand nombre de problèmes statistiques sont résolus par des techniques d'optimisation. On utilise des techniques d'optimisation dans des procédures statistiques telles que les **moindres carrés**, le **maximum de vraisemblance**, l'**estimation L1**, etc.

Dans plusieurs autres domaines de la **statistique** tels que l'analyse de régression, les tests d'hypothèses, les plans d'expérience, etc., l'optimisation joue un rôle caché, mais important.

Les méthodes d'optimisation peuvent être classées de la façon suivante :

1. méthodes d'optimisation classique (calcul différentiel, multiplicateurs de Lagrange) ;

2. méthodes de programmation mathématique (**programmation linéaire**, programmation non linéaire, programmation dynamique).

HISTORIQUE

Des problèmes d'optimisation étaient déjà formulés par Euclide, mais ce n'est qu'avec le développement du calcul différentiel et du calcul des variations aux XVII^{e} et XVIII^{e} siècles que furent forgés les outils mathématiques propres à résoudre de tels problèmes. Ces méthodes furent employées pour la première fois dans la résolution de certains problèmes d'optimisation en géométrie et en physique.

Des problèmes de ce genre ont été formulés et étudiés par des mathématiciens connus tels que Leonhard Euler, **Jakob Bernoulli**, Carl Gustav Jacob Jacobi et Joseph-Louis Lagrange auquel on doit les multiplicateurs de Lagrange.

La découverte d'importantes applications des techniques d'optimisation dans les problèmes militaires durant la Deuxième Guerre mondiale et les progrès de la technologie (ordinateurs) ont permis de résoudre des problèmes d'optimisation de plus en plus complexes dans des domaines de plus en plus variés.

MOTS CLÉS

Multiplicateur de Lagrange *(Lagrange multiplier)*
Programmation linéaire *(Linear programming)*
Recherche opérationnelle *(Operations research)*

RÉFÉRENCE

Arthanari, T.S. and Dodge, Y. (1981). *Mathematical Programming in Statistics.* John Wiley & Sons, New York.



P

PANEL

Panel

Le panel est un type de **sondage** que l'on répète périodiquement. Il est basé sur un **échantillon** permanent (ou presque) d'individus, de ménages, de magasins, etc., que l'on interroge régulièrement sur leur comportement ou leur opinion. Le panel offre donc l'avantage de suivre dans le temps le comportement individuel des personnes ou entités interrogées et de mesurer leur évolution. Les informations recueillies dans un panel sont souvent plus riches que dans un simple sondage, tout en ayant un coût de recrutement limité. On adapte le taux de renouvellement de l'échantillon en fonction des objectifs de l'enquête. Lorsque l'objectif du panel est de suivre l'évolution des individus au cours du temps (exemple : suivi du parcours professionnel), l'échantillon reste le même, et l'on parle d'enquête longitudinale. Lorsque l'objectif est de fournir des estimations des caractéristiques de la **population** à différentes périodes, l'échantillon est renouvelé partiellement de période en période, on parle de panel tournant. Ainsi, les données de panel constituent une source d'information très importante car elles contiennent une dimension individuelle et temporelle à la fois.

DOMAINES ET LIMITATIONS

Dans la pratique, il est rare d'avoir des panels au sens strict du terme. Il faut tenir compte des problèmes suivants propres au panel : difficulté de recrutement de certaines catégories de panelistes, lassitude des panelistes, principale responsable des non-réponses, et évolution de la **population** entraînant une déformation de l'**échantillon** par rapport à la population qu'il est censé représenter. On remédie souvent dans la pratique à ces problèmes en préférant au panel strict la technique du panel tournant.

Il existe également des erreurs de mesures spécifiques aux panels ; ce sont :

L'effet téléscope : erreur de date que commettent généralement les nouveaux panelistes lorsqu'ils déclarent la date d'un événement sur lequel ils sont interrogés.

L'effet de panel ou *biais de conditionnement* : changement de comportement des panelistes qui peut se produire au cours du temps en raison des interviews répétées. Les panelistes sont amenés ainsi à s'interroger sur leur comportement et finalement à le changer.

EXEMPLES

Enquête sur la santé et les soins médicaux : enquête réalisée par l'INSEE auprès de 10 000 ménages environ. Chaque ménage est suivi pendant trois mois au cours desquels, toutes les trois semaines, il est interrogé par un enquêteur sur sa consommation médicale.

Panel des consommateurs de la Sofres : le panel Métascope de la Sofres est basé sur un **échantillon** de 20 000 ménages. Les panelistes reçoivent chaque mois par la Poste un questionnaire auto-administré sur leurs achats divers. Ces questionnaires sont utilisés essentiellement par les entreprises qui désirent connaître les caractéristiques de leur clientèle.

Panel d'audience : le panel Médiamat est basé sur un échantillon de 2 300 foyers équipés d'audimètres à boutons-pressoirs. Il permet de mesurer l'audience des chaînes de télévision.

MOTS CLÉS

Biais *(Bias)*
Échantillon *(Sample)*
Population *(Population)*
Sondage *(Survey)*

RÉFÉRENCES

Duncan, G.J. (1992). Household Panel Studies : Prospects and Problems. *Social Science Methodology*, Trento, Italy.

Kasprzyk, D., Duncan, G. *et al.*, Eds. (1989). *Panel Surveys.* New York, Chichester, Brisbane, Toronto, Singapore, John Wiley & Sons, New York.

Rose, D. (1995). Household Panel Studies : an Overview. *Innovation* 8(1) : 7-24.

PARAMÈTRE

Parameter

Un paramètre caractérise un aspect quantitatif d'une **population**.

DOMAINES ET LIMITATIONS

Les paramètres de la **population** sont souvent inconnus. On peut toutefois estimer un paramètre par une **statistique** calculée à partir d'un **échantillon**, en utilisant une méthode d'**estimation**.

EXEMPLES

Un paramètre est généralement désigné par une lettre grecque :

μ : **moyenne** d'une **population**,
σ : **écart type** d'une population,
π : **pourcentage** relatif à une population.

MOTS CLÉS

Estimateur *(Estimator)*
Estimation *(Estimation)*
Maximum de vraisemblance *(Maximum likelihood)*
Moindres carrés *(Least squares)*
Moment *(Moment)*
Statistique *(Statistics)*

RÉFÉRENCE

Lehmann, E.L. (1983). *Theory of Point Estimation*, 2nd ed. John Wiley & Sons, New York.

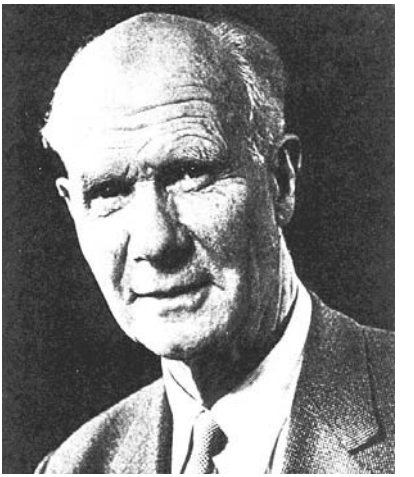
PARAMÉTRIQUE, TEST

Parametric test

Voir **test paramétrique**.

PEARSON EGON SHARPE

Pearson Egon Sharpe



Egon Sharpe Pearson (1895-1980), fils du statisticien **Karl Pearson**, fit des études de mathématiques au Trinity College à Cambridge. En 1921, il entra au département de **statistique** de l'University College à Londres. Il rencontra **Jerzy Neyman** avec lequel il commença à travailler durant le séjour de ce dernier à Londres en 1924-1925 et collabora également avec **William Sealy Gosset**. Lorsque son père quitta l'University College, en 1933, il prit en charge le département de statistique appliquée et, en 1936, devint directeur de la revue *Biometrika* après la mort de son père, fondateur de la revue.

Outre les travaux qu'il mena en collaboration avec J. Neyman et qui aboutirent au développement de la théorie des tests d'hypothèse, E.S. Pearson se consacra également aux problèmes de contrôle de qualité et de **recherche opérationnelle**.

Quelques articles de Egon Sharpe Pearson :

1928 (and Neyman, J.) On the use and interpretation of certain test criteria for purposes of statistical inference. *Biometrika*, 20A, pp. 175-240 and 263-295.

1931 The test of significance for the correlation coefficient. *Journal of the American Statistical Association*, 26, pp. 128-134.

1932 The percentage limits for the distribution of range in samples from a normal population ($n \leq 100$). *Biometrika*, 24, pp. 404-417.

1933 Neyman, J. and Pearson, E.S. On the testing of statistical hypotheses in relation to probability a priori. *Proceedings of the Cambridge Philosophical Society*, 29, pp. 492-510.

1935 The Application of Statistical Methods to industrial Standardization and Quality Control. *Brit. Stant.* 600. British Standards Institution, London.

MOT CLÉ

Test d'hypothèse *(Hypothesis testing)*

PEARSON KARL

Pearson Karl

Né à Londres, Karl Pearson (1857-1936) est connu pour ses nombreuses contributions à la **statistique**. Après des études au Kings College de Cambridge, il fut nommé dès 1885 à la chaire de mathématiques appliquées de l'University College de Londres.

Toute sa carrière se déroula dans cette université dont il devint titulaire de la chaire d'eugénique en 1911.

En 1901, avec l'aide de **Francis Galton**, il fonda la revue *Biometrika* dont il assuma la direction jusqu'à sa mort.

En 1906, il accueillit pour un an dans son laboratoire **William Sealy Gosset**, avec lequel il s'employa à résoudre les problèmes posés par les échantillons de petites tailles.

En 1933, il prit sa retraite et son département à l'Université fut divisé en deux : le département d'eugénique fut confié à **Ronald Aylmer Fisher** et celui de **statistique** à son propre fils **Egon Sharpe Pearson**.



Quelques articles de Karl Pearson :

1894 1948. Contributions to the Mathematical Theory of Evolution. I. pp. 1-40 in Karl Pearson, *Karl Pearson's Early Statistical Papers.* Cambridge University Press, Cambridge. First published as : On the Dissection of Asymmetrical Frequency Curves, *Philosophical Transactions of the Royal Society of London.* Series A, Vol. 185, pp. 71-110.

1895 1948. Contributions to the Mathematical Theory of Evolution. II. Skew Variation in Homogeneous Material. pp. 41-112 in Karl Pearson, *Karl Pearson's Early Statistical Papers.* Cambridge University Press, Cambridge. First published in *Philosophical Transactions of the Royal Society of London.* Series A, Vol. 186, pp. 343-414.

1896 1948. Mathematical Contributions to the Theory of Evolution III. Regression, Heredity and Panmixia. pp. 113-178 in Karl Pearson, *Karl Pearson's Early Statistical Papers.* Cambridge University Press, Cambridge. First published in *Philosophical Transactions of the Royal Society of London.* Series A, Vol. 187, pp. 253-318 and in the *Philosophical Magazine*, 5$^{\text{th}}$ Series, 50, pp. 157-175.

MOTS CLÉS

Loi du chi-carré *(Chi-square distribution)*
Coefficient de corrélation *(Correlation coefficient)*

PERMUTATION

Permutation

Le terme « permutation » se réfère à une notion d'**analyse combinatoire**.

Une permutation est un rangement ou un classement ordonné de n objets.

En théorie des probabilités, il faut faire la différence entre le cas où les n objets peuvent être distingués et celui où ils ne peuvent l'être que partiellement.

HISTORIQUE

Voir **analyse combinatoire**.

ASPECTS MATHÉMATIQUES

1. *Nombre de permutations d'objets distincts*

   Le nombre de permutations possibles de n objets est égal à n factorielle :

   $$\begin{aligned} \text{permutations} &= n! \\ &= n \cdot (n-1) \cdot \ldots \cdot 3 \cdot 2 \cdot 1 \end{aligned}$$

2. *Nombre de permutations d'objets partiellement distincts.*

   Le nombre de permutations possibles de n objets parmi lesquels n_1, n_2, ..., n_r ne sont pas distinguables entre eux est égal à :

   $$\text{nombre de permutations} = \frac{n!}{n_1! \cdot n_2! \cdot \ldots \cdot n_r!}$$

   où $n_1 + n_2 + \ldots + n_r = n$.

EXEMPLE

Si nous disposons des trois objets A, B et C, les permutations possibles sont les suivantes :

A B C
A C B
B A C
B C A
C A B
C B A

Nous avons donc 6 permutations possibles. Sans les énumérer, nous pouvons trouver le nombre de permutations par la formule suivante :

$$\text{nombre de permutations} = n!$$

n étant le nombre d'objets. Ce qui nous donne pour notre exemple :

$$\text{nombre de permutations} = 3! = 3 \cdot 2 \cdot 1 = 6.$$

Imaginons qu'il nous faille composer des signaux en alignant des drapeaux. Si nous disposons de cinq drapeaux de différentes couleurs, le nombre de signaux différents que nous pourrons composer sera :

$$5! = 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 = 120.$$

En revanche, si parmi ces cinq drapeaux, il y en a deux rouges, deux blancs et un noir, nous pourrons composer :

$$\frac{5!}{2! \cdot 2! \cdot 1!} = 30$$

signaux différents.

MOTS CLÉS

Analyse combinatoire *(Combinatory analysis)*
Arrangement *(Arrangement)*
Combinaison *(Combination)*

PITMAN EDWIN JAMES GEORGE

Pitman Edwin James George



Edwin James George Pitman est né à Melbourne (Victoria) en 1987. Il effectua sa scolarité à la Kensington State School et au South Melbourne College. Lors de sa dernière année d'étude, il obtint une bourse pour étudier les mathématiques à Wyselaskie et à Dixson, ainsi qu'au Ormond College. Il obtint une licence ès lettres (1921), une licence ès sciences (1922) et une maîtrise ès lettres (1923). Par la suite, il fut nommé professeur suppléant de mathématiques au Canterbury College, université de Nouvelle-Zélande (1922-23). Il retourna en Australie où il obtint un poste de professeur assistant aux Trinity and Ormond Colleges ainsi qu'un poste de « part-time lecturer » en physique à l'Université de Melbourne (1924-25). En 1926, Edwin Pitman fut nommé professeur de mathématiques à l'Université de Tasmanie, où il resta jusqu'à sa retraite en 1962. Il mourut en 1993, à Kingston (Tasmanie).

Une contribution majeure de Pitman à la théorie des probabilités concerne l'étude du comportement de la fonction caractéristique au voisinage de zéro.

Quelques ouvrages et articles principaux d'Edwin James George Pitman :

1937 Signifiance tests which may applied to samples from any populations. Supplement, *Journal of the Royal Statistical Society.* Series B, 4, pp. 119-130.

1937 Signifiance tests which may be applied to samples from any populations. II. The correlation coefficient test. Supplement, *Journal of the Royal Statistical Society.* Series B, 4, pp. 225-232.

1938 Signifiance tests which may be apply to samples from any populations. III. The analysis of variance test. *Biometrika*, 29, pp. 322-335.

1939 The estimation of the location and scale parameters. *Biometrika*, 30, pp. 391-421.

1939 Tests of hypotheses concerning location and scale parameters. *Biometrika*, 31, pp. 200-215.

PLAN CARRÉ GRÉCO-LATIN

Graeco-latin squares design

Un plan carré gréco-latin est un **plan d'expérience** dans lequel les unités expérimentales sont regroupées de trois façons différentes. Nous l'obtenons en superposant deux plans carrés latins de même dimension.

Si chaque lettre latine coïncide exactement une fois avec chaque lettre grecque, les deux plans carrés latins sont dits orthogonaux. Ensemble ils forment un plan carré gréco-latin.

Chaque **traitement** (lettre latine) apparaît alors une fois dans chaque ligne, une fois dans chaque colonne et une fois avec chaque lettre grecque.

HISTORIQUE

La construction d'un carré gréco-latin, appelé également carré eulérien, remonte à Leonhard Euler (1782).

Le livre de **Ronald Aylmer Fisher** et Frank Yates (1963) donne une table carrée gréco-latine d'ordre 3 jusqu'à 12 (non compris l'ordre 6). Pour tout complément d'information, on peut se référer à l'ouvrage de Dénes et Keedwell (1974). Yadolah Dodge et Shah (1977) traitent le cas de l'**estimation** lorsque nous sommes en présence d'observations manquantes dans un plan expérience gréco-latin.

DOMAINES ET LIMITATIONS

Les plans carrés gréco-latins sont utilisés pour réduire les effets de trois sources d'erreurs systématiques.

Il existe des plans carrés gréco-latins pour toutes les dimensions n sauf pour $n = 1, 2$ et 6.

EXEMPLE

Nous désirons tester trois marques d'essence différentes. Dans ce but, nous disposons de trois conducteurs et de trois véhicules. Comme la capacité du circuit d'essais n'est pas suffisante, il faut répartir les expériences sur trois jours. Il est clair que tant les conducteurs que les véhicules ou les jours peuvent être des sources d'erreurs systématiques. Par conséquent, afin d'éliminer ces erreurs, nous allons construire un plan carré gréco-latin. Ainsi chaque marque d'essence (**traitement**) va être testée une fois avec chaque conducteur, une fois avec chaque véhicule et une fois chaque jour.

Dans le plan carré gréco-latin ci-dessous, nous représentons les différentes marques d'essence par A, B, C et les différents jours par α, β, γ.

		Conducteurs		
		1	2	3
Véhicules	1	$A\alpha$	$B\beta$	$C\gamma$
	2	$B\gamma$	$C\alpha$	$A\beta$
	3	$C\beta$	$A\gamma$	$B\alpha$

Une **analyse de variance** permet de dire si, après élimination des effets des lignes (véhicules), des colonnes (conducteurs) et des lettres grecques (jours), il existe une différence significative entre les marques d'essence.

MOTS CLÉS

Analyse de variance *(Analysis of variance)*
Plan carré latin *(Latin squares design)*
Plan d'expérience *(Design of experiments)*

RÉFÉRENCES

Dénes, J. and Keedwell, A.D. (1974). *Latin squares and their applications.* Academic Press, New York.

Dodge, Y. and Shah, K.R. (1977). Estimation of parameters in Latin squares and Graeco-latin squares with missing observations. *Commun. Statist., Theory Methods*, A6(15), pp. 1465-1472.

Euler, L. (1782). Recherches sur une nouvelle espèce de carrés magiques. *Verh. Zeeuw. Gen. Weten. Vlissengen*, 9, pp. 85-239.

Fisher, R.A. and Yates, F. (1963). *Statistical Tables for Biological, Agricultural and Medical Research*, 6th ed. Hafner Press (Macmillan), New York.

PLAN CARRÉ LATIN

Latin squares design

Un plan carré latin est un **plan d'expérience** dans lequel les unités expérimentales sont regroupées de deux manières différentes. C'est un arrangement de t traitements, chacun répété t fois, de telle façon que chaque traitement apparaisse exactement une fois dans chaque ligne et dans chaque colonne. Nous avons donc un plan carré. Nous désignons les différents traitements par des lettres latines, d'où le nom de plan carré latin.

Ce genre de plans est utilisé pour réduire les effets de deux sources d'erreurs systématiques.

HISTORIQUE

Selon Preece (1983), l'histoire des carrés latins remonte à 1624. En **statistique**, et plus précisément dans le domaine des plans d'expérience, **Ronald Aylmer Fisher** (1925) introduisit les carrés latins lors d'expériences dans le domaine forestier.

EXEMPLE

Un agriculteur possède du terrain sur lequel il veut faire pousser du maïs. Il souhaite faire des essais car il a à sa disposition une variété de maïs et quatre sortes d'engrais (**traitements**). De plus, il sait que son terrain présente des hétérogénéités quant à son ensoleillement et à son humidité. Il y aura donc deux sources d'erreurs systématiques si nous admettons que le soleil et l'humidité influencent la croissance du maïs.

L'agriculteur fait alors un quadrillage de son terrain et attribue à chaque parcelle un des traitements (engrais) suivants, de sorte que chaque traitement apparaisse une fois dans chaque ligne et dans chaque colonne :

A : engrais 1
B : engrais 2
C : engrais 3
D : engrais 4.

De cette manière il peut en grande partie éliminer les effets de l'environnement expérimental (soleil et humidité).

La figure suivante montre le plan carré latin 4×4 où les lettres latines A, B, C, D représentent les quatre traitements ci-dessus :

A	B	C	D
B	C	D	A
C	D	A	B
D	A	B	C

Une **analyse de variance** va pouvoir nous dire si, après élimination des effets des lignes (soleil) et des colonnes (humidité), il existe une différence significative entre les traitements.

MOTS CLÉS

Analyse de variance *(Analysis of variance)*
Plan d'expérience *(Design of experiments)*

RÉFÉRENCES

Fisher, R.A. (1925). *Statistical Methods for Research Workers.* Oliver & Boyd, Edinburgh.

Preece, D.A. (1983). Latin squares, Latin cubes, Latin rectangles, etc. In Kotz S. and Johnson N.L. (editors). *Encyclopedia of Statistical Sciences.* Vol 4. John Wiley & Sons, New York.

PLAN COMPLÈTEMENT RANDOMISÉ

Completely randomized design

Un plan complètement randomisé est un **plan d'expérience** dans lequel les traitements sont attribués d'une manière complètement aléatoire aux unités expérimentales.

Il est utilisé quand les unités expérimentales sont supposées « uniformes », c'est-à-dire quand il n'y a pas de **facteur** incontrôlable dans l'**expérience**.

HISTORIQUE

Voir **plan d'expérience**.

EXEMPLE

On veut tester cinq sortes de médicaments à base d'aspirine. Dans ce but, on distribue aléatoirement les cinq sortes d'aspirine à quarante patients. Si l'on note par A, B, C, D, E les cinq médicaments, le résultat de la distribution aléatoire est le suivant :

— A est attribué à dix personnes ;

— B est attribué à douze personnes ;

— C est attribué à quatre personnes ;

— D est attribué à sept personnes ;

— E est attribué à sept personnes.

On a alors un plan complètement randomisé où les traitements (médicaments) sont distribués d'une manière aléatoire aux unités expérimentales (patients), chaque patient recevant un seul traitement. En effet, on suppose dans cet exemple que les patients sont « uniformes », c'est-à-dire qu'ils ne présentent aucune différence. De plus, on suppose qu'il n'y a pas de **facteur** incontrôlable qui intervienne au cours du traitement.

Dans cet exemple, le plan complètement randomisé est une **expérience factorielle** n'impliquant qu'un facteur ; en effet, l'aspirine est le seul facteur présent dans l'**expérience**. Les cinq sortes d'aspirine sont les différents niveaux du facteur.

MOTS CLÉS

Expérience *(Experiment)*
Plan d'expérience *(Design of experiments)*

PLAN D'EXPÉRIENCE

Design of experiments

Un plan d'expérience est en quelque sorte une programmation de l'**expérience** proprement dite. Comme chaque **facteur** engagé dans l'expérience peut prendre un nombre déterminé de valeurs différentes (appelées niveaux), le plan d'expérience va spécifier le choix des niveaux du ou des facteurs ou les combinaisons de facteurs à introduire dans l'expérience.

HISTORIQUE

Les plans d'expérience ont été utilisés pour la première fois dans les années 1920, essentiellement dans le domaine de l'agriculture. Sir **Ronald Aylmer Fisher** a été le premier à utiliser les statistiques mathématiques dans les plans d'expérience. Un document, rédigé en termes non mathématiques, décrivant les principes régissant les plans d'expérience est donné par R.A. Fisher (1926).

W.T. Federer et L.N. Balaam (1973) proposent une bibliographie très détaillée de la littérature relative aux plans d'expérience antérieure à 1969. Celle-ci regroupe une liste de plus de 8 000 références.

DOMAINES ET LIMITATIONS

Le but d'un plan d'expérience est de fournir les méthodes les plus efficaces et les plus économiques permettant d'obtenir des conclusions solides et adéquates.

Les plans les plus fréquemment utilisés sont le **plan complètement randomisé**, le **plan randomisé à blocs** et le **plan carré latin**, mais d'autres plans sont également utilisés.

Chaque type de plan fait l'objet d'une analyse mathématique différente. En effet, chaque sorte de plan possède un **modèle** mathématique différent. Parmi ces analyses, on peut citer l'**analyse de variance**, l'**analyse de covariance** ou l'**analyse de régression**.

MOTS CLÉS

Analyse de régression *(Regression analysis)*
Analyse de variance *(Analysis of variance)*
Expérience *(Experiment)*
Modèle *(Model)*
Plan optimal *(Optimal design)*

RÉFÉRENCES

Federer, W.T. and Balaam, L.N. (1973). *Bibliography on Experiment and Treatment Design Pre-1968.* Hafner, New York.

Fisher, R.A. (1926). The Arrangement of Field Experiments. *Journal of the Ministry of Agriculture*, Vol. XXXIII, pp. 503-513.

PLAN OPTIMAL

Optimal design

Un plan optimal est un **plan d'expérience** qui satisfait un certain critère d'optimalité (par exemple la minimisation des erreurs associées aux estimations que l'on doit faire).

En effet, les traitements appliqués aux diverses unités expérimentales étant fixés par l'expérimentateur, on peut alors essayer d'améliorer la qualité de

l'analyse des résultats. On pourra ainsi augmenter la précision des estimateurs, réduire l'étendue des intervalles de confiance ou accroître la puissance des tests d'hypothèses, en choisissant le plan optimal à l'intérieur de la **catégorie** à laquelle il appartient.

HISTORIQUE

La théorie des plans optimaux fut développée par Jack Kiefer (1924-1981). L'un des points principaux de son développement repose sur l'efficacité statistique d'un plan. Dans son article de 1959, il présente les principaux critères pour obtenir un plan optimal. L'ensemble des articles de Kiefer dans le domaine des plans d'expérience a été publié par Springer-Verlag avec la collaboration de l'Institute of Mathematical Statistics.

DOMAINES ET LIMITATIONS

Dans la mesure où l'on ne peut pas minimiser toutes les **erreurs** simultanément, les plans optimaux seront déterminés par le choix du critère d'optimalité. Ces différents critères sont exposés dans l'article de Kiefer (1959).

MOT CLÉ

Plan d'expérience *(Design of experiments)*

RÉFÉRENCES

Brown, L.D., Olkin, I., Sacks, J. and Wynn, H.P. (1985). *Jack Carl Kiefer, Collected Papers III : Design of Experiments.* Springer-Verlag, New York.

Fedorov, V.V. (1972). *Theory of Optimal Experiments.* Academic Press, New York.

Kiefer, J. (1959). Optimum experimental designs. *Journal of the Royal Statistical Society.* Series B, 21, pp. 272-319.

PLAN RANDOMISÉ À BLOCS

Randomized block design

Un plan randomisé à blocs est un **plan d'expérience** dans lequel les unités expérimentales sont regroupées dans des ensembles appelés blocs, de telle sorte qu'elles soient aussi semblables que possible à l'intérieur de chaque bloc. Les traitements sont affectés aléatoirement aux unités expérimentales à l'intérieur de chaque bloc. Si tous les traitements apparaissent au moins une fois dans chaque bloc, nous avons un plan complet randomisé à blocs. Sinon c'est un plan incomplet randomisé à blocs.

Ce genre de plans est utilisé pour réduire les effets d'une source d'**erreur** systématique. Dans la mesure où seules les différences entre les traitements intéressent l'expérimentateur, les effets des variations entre les blocs doivent être éliminés.

HISTORIQUE

Voir **plan d'expérience**.

EXEMPLE

Un agriculteur possède cinq parcelles de terrain sur lesquelles il veut faire pousser du maïs. Il souhaite faire des essais car il a à sa disposition deux variétés de maïs et deux sortes d'engrais. De plus, il sait que son terrain présente des hétérogénéités quant à son ensoleillement. Il y aura donc une source d'erreur si l'on admet que le soleil favorise la croissance du maïs.

L'agriculteur va alors diviser le terrain en cinq blocs et attribuer aléatoirement à chaque bloc les quatre traitements suivants :

A : maïs 1 avec engrais 1
B : maïs 1 avec engrais 2
C : maïs 2 avec engrais 1
D : maïs 2 avec engrais 2

Ainsi, il élimine la source d'**erreur** systématique due au **facteur** incontrôlable qu'est le soleil. En effet, en procédant d'une manière aléatoire à l'intérieur de chaque bloc, il ne favorise ni une variété de maïs par rapport à l'autre, ni une sorte d'engrais par rapport à l'autre. La figure suivante montre l'arrangement aléatoire des quatre traitements dans cinq blocs :

bloc 1	D	C	A	B
bloc 2	C	B	A	D
bloc 3	A	B	C	D
bloc 4	A	C	B	D
bloc 5	D	C	A	B

Une **analyse de variance** permet de conclure si, après élimination des effets des blocs, il existe une différence significative entre les traitements.

MOTS CLÉS

Analyse de variance *(Analysis of variance)*
Plan d'expérience *(Design of experiments)*

POINT LEVIER

Leverage point

Lors d'une **analyse de régression**, on appelle point levier une **observation** i telle que l'**estimation** de la i-ème **variable** réponse $\hat{Y}_i$ est fortement influencée par la **valeur** de la **variable indépendante** X_i correspondante. En d'autres termes, la notion de point levier est équivalente à celle de levier en physique, c'est-à-dire que si la valeur de l'observation est modifiée, même légèrement (par exemple par une **erreur** de mesure), l'estimation de la **variable dépendante** en est très affectée.

ASPECTS MATHÉMATIQUES

Les éléments diagonaux de la **matrice H** de l'**analyse de régression** décrivent l'influence de chaque **observation** sur l'**estimation** correspondante de la **variable dépendante**. On associe donc généralement les éléments diagonaux h_{ii} de la matrice **H** à la notion de point levier. Comme la **trace** de la matrice **H** est égale au nombre p de paramètres à estimer, on s'attend à avoir des h_{ii} de l'ordre de $\frac{p}{n}$. Les h_{ii} seraient de l'ordre de p/n si toutes les observations avaient la même influence sur les estimateurs. Si pour une observation on a donc un h_{ii} beaucoup plus grand que p/n, elle sera alors considérée comme un point levier. Dans la pratique, une observation i est considérée comme un point levier si :

$$h_{ii} > \frac{2p}{n}$$

On considère le **modèle** général d'une **régression** linéaire sous forme matricielle :

$$\mathbf{Y} = \mathbf{X} \cdot \boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

où

- $\mathbf{Y}$ est le **vecteur** $(n \times 1)$ des observations relatives à la variable dépendante (n observations),
- $\mathbf{X}$ est la **matrice** $(n \times p)$ ayant trait aux variables indépendantes (p variables indépendantes),
- $\boldsymbol{\varepsilon}$ est le vecteur $(n \times 1)$ des erreurs et
- $\boldsymbol{\beta}$ est le vecteur $(p \times 1)$ des paramètres à estimer.

L'**estimation** $\hat{\boldsymbol{\beta}}$ du **vecteur** $\boldsymbol{\beta}$ est donnée par :

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}.$$

La matrice **H** est définie par :

$$\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'.$$

En particulier, l'élément diagonal h_{ii} sera défini par :

$$h_{ii} = \mathbf{x}_i(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i',$$

où $\mathbf{x}_i$ est la i-ème ligne de $\mathbf{X}$.

On appelle alors point levier toute observation i telle que :

$$h_{ii} = \mathbf{x}_i(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i' > 2 \cdot \frac{p}{n}.$$

DOMAINES ET LIMITATIONS

Huber (1981) parle de point levier lorsque l'élément diagonal de la **matrice** $\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ correspondant est supérieur à $0,2$, pour que la limite critique ne dépende plus du nombre de paramètres p ni du nombre d'observations n.

EXEMPLE

On considère le tableau suivant où Y est une **variable dépendante** expliquée par la **variable indépendante** X :

X	Y
50	6
52	8
55	9
75	7
57	8
58	10

Le **modèle** de **régression linéaire simple** sous forme matricielle s'écrit :

$$\mathbf{Y} = \mathbf{X} \cdot \boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

où

$$\mathbf{Y} = \begin{bmatrix} 6 \\ 8 \\ 9 \\ 7 \\ 8 \\ 10 \end{bmatrix} \quad \mathbf{X} = \begin{bmatrix} 1 & 50 \\ 1 & 52 \\ 1 & 55 \\ 1 & 75 \\ 1 & 57 \\ 1 & 58 \end{bmatrix},$$

$\boldsymbol{\varepsilon}$ est le **vecteur** (6×1) des erreurs, et $\boldsymbol{\beta}$ est le vecteur (2×1) des paramètres.

On trouve la **matrice H** en appliquant le résultat :

$$\mathbf{H} = \mathbf{X} \left(\mathbf{X}'\mathbf{X}\right)^{-1} \mathbf{X}'.$$

En effectuant les calculs, on trouve :

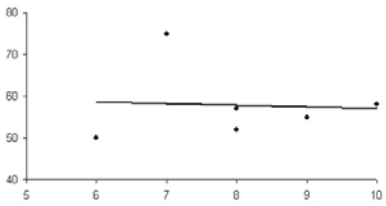
$$\mathbf{H} = \begin{bmatrix} 0,32 & 0,28 & 0,22 & -0,17 & 0,18 & 0,16 \\ 0,28 & 0,25 & 0,21 & -0,08 & 0,18 & 0,16 \\ 0,22 & 0,21 & 0,19 & 0,04 & 0,17 & 0,17 \\ -0,17 & -0,08 & 0,04 & 0,90 & 0,13 & 0,17 \\ 0,18 & 0,18 & 0,17 & 0,13 & 0,17 & 0,17 \\ 0,16 & 0,16 & 0,17 & 0,17 & 0,17 & 0,17 \end{bmatrix}.$$

On écrit la diagonale de la matrice **H** dans le vecteur **h** :

$$\mathbf{h} = \begin{bmatrix} 0,32 \\ 0,25 \\ 0,19 \\ 0,90 \\ 0,17 \\ 0,17 \end{bmatrix}.$$

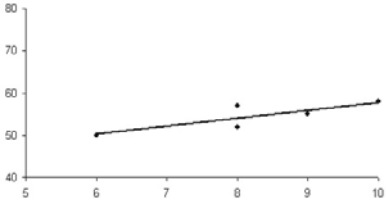
Dans ce problème il y a 2 paramètres ($p = 2$) et 6 observations ($n = 6$), c'est-à-dire que la **valeur critique** pour déterminer les points levier est donnée par $\frac{4}{6} = 0,67$.

En comparant les composantes de **h** à cette **valeur**, on constate qu'excepté la quatrième, elles sont toutes inférieures à cette valeur critique, c'est-à-dire que la quatrième observation est un point levier pour le problème étudié. Le graphique sur lequel on a représenté en abscisse la variable X et en ordonnée la variable Y montre bien que le point $(75, 7)$ est un point levier :



On peut constater sur ce graphique que la droite de régression est fortement influencée par ce point levier et ne représente donc pas correctement l'ensemble des données.

Sur le graphique suivant, on a supprimé ce point de l'ensemble des données. On peut constater que la nouvelle droite de régression donne à présent une représentation beaucoup plus correcte de l'ensemble des points.



MOTS CLÉS

Analyse de régression *(Regression analysis)*
Matrice *(Matrix)*
Matrice H *(Hat matrix)*
Régression linéaire multiple *(Multiple linear regression)*
Régression linéaire simple *(Simple linear regression)*

RÉFÉRENCES

Huber, P.J. (1981). *Robust Statistics.* John Wiley & Sons, New York.

Mosteller, F. and Tukey, J.W. (1977). *Data Analysis and Regression : A Second Course in Statistics.* Addison-Wesley, Reading, Mass.

POISSON, LOI DE

Poisson distribution

Voir **loi de Poisson**.

POISSON SIMÉON-DENIS

Poisson Siméon-Denis



Siméon-Denis Poisson (1781-1840), bien qu'issu d'une modeste famille de province, put poursuivre ses études à l'École polytechnique à Paris où il devint plus tard professeur.

À partir de 1815, il enseigna également à la Sorbonne et fut élu à l'Académie des sciences la même année.

Outre l'enseignement, il fut intéressé par la recherche dans différents domaines : mécanique, physique, et théorie des probabilités.

En 1837, il publia un ouvrage intitulé « *Recherches sur la probabilité des jugements en matière criminelle et en matière civile, précédées des règles générales du calcul des probabilités* ». Dans plusieurs de ses écrits, il posa les bases d'une méthode statistique pour les sciences sociales.

Quelques articles et ouvrages de Siméon-Denis Poisson :

1824 Sur la probabilité des résultats moyens des observations. *Connaissance des temps pour l'an 1827*, pp. 273-302.

1829 Suite du mémoire sur la probabilité du résultat moyen des observations, inséré dans *Connaissance des temps de l'année 1827. Connaissance des temps pour l'an 1832*, pp. 3-22.

1836a Note sur la loi des grands nombres. *Comptes rendus hebdomadaires des séances de l'Académie des sciences* 2 : pp. 377-382.

1836b Note sur le calcul des probabilités. *Comptes rendus hebdomadaires des séances de l'Académie des sciences* 2 : pp. 395-400.

1837 *Recherches sur la probabilité des jugements en matière criminelle et en matière civile, précédées des règles générales du calcul des probabilités.* Bachelier, Imprimeur-Libraire pour les Mathématiques, Paris.

MOT CLÉ

Loi de Poisson *(Poisson distribution)*

POLYGONE DE FRÉQUENCES

Frequency polygon

Un polygone de fréquences est une **représentation graphique** d'une **distribution de fréquences**, s'alignant à l'**histogramme** et appartenant à la catégorie des graphiques de fréquences.

Il se présente sous la forme d'une droite segmentée reliant les points centraux des sommets de chaque rectangle d'un histogramme.

ASPECTS MATHÉMATIQUES

On construit un polygone de fréquences en plaçant en abscisse les points milieux des différentes classes d'une **distribution de fréquences**. Sur chacune de ces valeurs ou chacun de ces points milieux, on porte en ordonnée une hauteur égale à la **fréquence** de la classe correspondante. On relie par des segments de droite les points ainsi déterminés.

Cela revient à relier les points centraux des sommets de chaque rectangle d'un histogramme.

Il est de coutume de fermer le polygone de fréquences aux deux extrémités en plaçant des points bien choisis sur l'axe horizontal et en reliant ces points aux premier et dernier rectangles.

DOMAINES ET LIMITATIONS

Le polygone de fréquences est principalement construit à partir de données groupées en intervalles. Dans ces cas, il est préférable que ces intervalles soient de même longueur. Si l'on fait subir au polygone de fréquences une sorte de « polissage », on obtient une courbe lisse que l'on appelle **courbe de fréquences**. Ce polissage consiste à construire des polygones de fréquences de plus en plus fins en subdivisant les intervalles dans des classes de plus en plus petites.

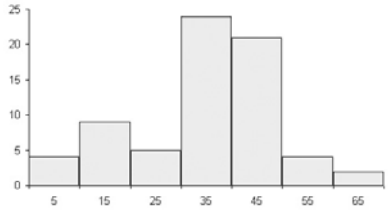
Il est évident qu'on peut effectuer ce type de polissage si le nombre d'unités dans l'ensemble statistique est suffisamment grand, tel qu'il y a assez d'observations dans les classes réduites.

EXEMPLE

Le **tableau de fréquences** ci-dessous présente les moyennes annuelles des précipitations dans 69 villes des États-Unis :

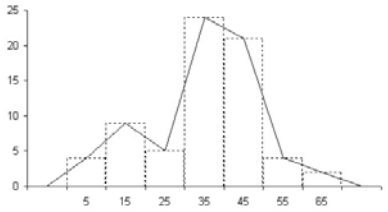
Classes (précipi--tations en pouces)	Fréquences
$0-10$	4
$10-20$	9
$20-30$	5
$30-40$	24
$40-50$	21
$50-60$	4
$60-70$	2
Total	69

Le polygone de fréquences se construit à partir de l'**histogramme** qui représente les différentes classes sur l'axe horizontal, et les fréquences sur l'axe vertical :



Pour construire le polygone de fréquences, on place un premier point sur l'axe horizontal, à une **distance** égale à la borne inférieure de la première classe moins un demi-**intervalle** de classe.

On relie ensuite tous les sommets de l'histogramme en leur milieu. Le dernier sommet est relié à un point situé sur l'axe horizontal, à une distance égale à la borne supérieure de la dernière classe plus un demi-intervalle de classe. Le polygone de fréquences est alors le suivant :



MOTS CLÉS

Distribution de fréquences *(Frequency distribution)*
Histogramme *(Histogram)*
Ogive *(Ogive)*
Représentation graphique *(Graphical representation)*
Tableau de fréquences *(Frequency table)*

POPULATION

Population

Une population est définie comme un ensemble d'unités statistiques de même nature sur lequel on recherche des informations quantifiables. La population constitue l'univers de référence lors de l'étude d'un problème statistique donné.

EXEMPLES

L'ensemble des personnes d'un pays, l'ensemble des arbres d'une forêt, l'ensemble de la production d'une usine, ou encore l'ensemble des prix d'articles de consommation forment chacun une population.

MOT CLÉ

Échantillon *(Sample)*

POPULATION-CIBLE

Target population

Le concept de population-cible est défini comme une **population** à laquelle on désire appliquer les résultats d'une enquête.

Ce terme est à différencier de celui de population qui désigne, quant à lui, la population effectivement disponible pour effectuer l'enquête.

EXEMPLE

On désire faire une enquête sur les étrangers résidant dans la ville X. La **population** est représentée par la liste des étrangers enregistrés dans la ville.

La population-cible, quant à elle, est définie par l'ensemble des étrangers habitant la ville X, y compris ceux ne figurant pas sur la liste (non annoncés, clandestins, etc.).

MOT CLÉ

Population *(Population)*

POURCENTAGE

Percentage

La notion de pourcentage est une mesure permettant de décrire la proportion des individus ou unités statistiques possédant une certaine caractéristique dans un ensemble, évaluée en général sur la centaine.

ASPECTS MATHÉMATIQUES

Le pourcentage est généralement dénoté par p lorsqu'il est mesuré sur un **échantillon** et par π lorsqu'il se réfère à la **population**.

Soit un échantillon de taille n. Si k individus possèdent la caractéristique étudiée, le pourcentage d'individus possédant cette caractéristique est donné par la **statistique** p :

$$p = \frac{k}{n} \cdot 100\,\%.$$

Le calcul du même pourcentage sur une population de taille N nous donne le **paramètre** π :

$$\pi = \frac{k}{N} \cdot 100\,\%.$$

PRÉVALENCE

Prevalence

Voir **taux de prévalence**.

PRÉVISION

Forecasting

La prévision est un ensemble d'études mettant en œuvre les données de la **statistique**, les théories économiques et les conditions non économiques visant à obtenir, par le raisonnement, une opinion sur les événements futurs.

On recourt à de nombreuses méthodes pour prévoir l'avenir. Elles reposent toutes sur un nombre restreint d'hypothèses de base.

Une de ces hypothèses affirme que les grandes tendances observées historiquement se maintiendront dans le futur. Une autre consiste à dire que les fluctuations mesurables d'une **variable** se reproduiront à intervalles réguliers, de telle sorte que ces variations deviendront prévisibles.

L'analyse des séries chronologiques dans un processus de prévision s'appuient sur ces deux hypothèses :

1. *La prévision à long terme*
   Lorsqu'il s'agit de faire des prévisions à long terme (plus de cinq années), l'analyse et, par la suite, l'extrapolation de la **tendance séculaire** s'avèrent importantes.

   Les prévisions à long terme basées sur la tendance séculaire ne prennent généralement pas en considération les fluctuations cycliques. De plus, les variations saisonnières n'exercent pas d'effet sur les données annuelles et n'entrent donc pas en considération pour une prévision à long terme.

2. *La prévision à moyen terme*
   Pour tenir compte de l'effet probable des fluctuations cycliques dans le mécanisme de prévision à moyen terme (une à cinq ans), on multiplie la **valeur** de la projection de la tendance séculaire par la fluctuation cyclique pour obtenir la valeur prévue. Il devient délicat de présumer, lorsqu'on traite de la composante cyclique, que les structures qui se dégagent des données se maintiendront et que les variations observées se reproduiront avec régularité.

   En fait, les cycles successifs ont tendance à varier énormément, tant au niveau de leur périodicité qu'à celui de leur importance ou du modèle qu'ils suivent.

3. *La prévision à court terme*
   L'**indice saisonnier** devra être pris en considération dans la prévision à court terme (quelques mois), en plus des valeurs de la projection de la tendance séculaire et des fluctuations cycliques.

   En général, on n'essaie pas de prévoir les variations irrégulières, ni dans les prévisions à court terme, ni dans les prévisions portant sur des périodes plus longues.

ASPECTS MATHÉMATIQUES

Soit Y une **série chronologique** ; on peut la décrire à l'aide de ses composantes, à savoir :

— la **tendance séculaire** T ;

— les **variations saisonnières** S ;

— les **fluctuations cycliques** C ;

— les **variations irrégulières** I.

On distingue :

— le **modèle** multiplicatif : $Y = T \cdot S \cdot C \cdot I$;

— le modèle additif : $Y = T + S + C + I$.

Si on note, pour t supérieur au temps d'**observation** :

Y_t = prévision de la **valeur** Y au temps t,
T_t = projection de la tendance séculaire au temps t,
C_t = fluctuation cyclique prévue au temps t,
S_t = variation saisonnière prévue au temps t,

on a :

1. *pour une prévision à long terme*, la valeur de la série chronologique sera estimée par :

$$Y_t = T_t \ ;$$

2. *pour une prévision à moyen terme*, la valeur de la série chronologique sera estimée par :
   $Y_t = T_t \cdot C_t$ (modèle multiplicatif)
   ou $Y_t = T_t + C_t$ (modèle additif) ;
3. *pour une prévision à court terme*, la valeur de la série chronologique sera estimée par :
   $Y_t = T_t \cdot C_t \cdot S_t$ (pour le modèle multiplicatif)
   ou $Y_t = T_t + C_t + S_t$ (pour le modèle additif).

DOMAINES ET LIMITATIONS

La planification et la prise de décisions sont deux activités qui s'appuient sur la vision de ce que l'avenir réserve. Les administrateurs doivent donc faire des prévisions.

Ce n'est pas parce qu'une prévision obtenue par l'analyse de séries chronologiques s'est révélée fausse qu'il faut immédiatement conclure que l'utilisation des techniques présentées est à proscrire. En fait, bien que fausse, cette prévision était peut-être plus proche de la réalité que ne l'aurait été toute autre prévision s'appuyant uniquement sur l'intuition. Malgré ses limites et les problèmes qu'elle soulève, l'analyse des séries chronologiques s'avère un outil très précieux comme point de départ d'un processus de prévision.

EXEMPLES

La notion de prévision dépend du sujet d'étude ; en effet,

- en météorologie :
  - — une prévision à court terme est de l'ordre de quelques heures ;
  - — une prévision à moyen terme est de l'ordre de quelques jours ;
- en économie :
  - — une prévision à court terme est de l'ordre de quelques mois ;
  - — une prévision à moyen terme est de l'ordre de un à cinq ans ;
  - — une prévision à long terme est de l'ordre de plus de cinq ans.

MOTS CLÉS

Fluctuation cyclique *(Cyclical fluctuation)*
Indice saisonnier *(Seasonal index)*
Série chronologique *(Time series)*
Tendance séculaire *(Secular trend)*
Variation irrégulière *(Irregular variation)*
Variation saisonnière *(Seasonal variation)*

PROBABILITÉ

Probability

On peut définir la probabilité d'un **événement** par deux approches : l'approche par les fréquences relatives, et l'approche axiomatique.

Dans la première approche, on suppose qu'une **expérience aléatoire** est répétée plusieurs fois sous les mêmes conditions. Pour chaque événement A défini dans l'**ensemble fondamental** Ω, on définit n_A comme le nombre de fois où l'événement A survient lors des n premières répétitions de l'**expérience**. Dans ce cas, la probabilité de l'événement A, notée $P(A)$, est définie par :

$$P(A) = \lim_{n\to\infty} \frac{n_A}{n},$$

c'est-à-dire que $P(A)$ est défini comme la limite du rapport du nombre de fois où l'événement A survient par rapport au nombre total de répétitions.

Dans la seconde approche, pour chaque événement A, on admet qu'il existe une probabilité de A, $P(A)$, satisfaisant les trois axiomes suivants :

1. $0 \leq P(A) \leq 1$;
2. $P(\Omega) = 1$;
3. pour chaque séquence d'événements mutuellement exclusifs $A_1, A_2, \ldots$ (c'est-à-dire d'événements pour lesquels $A_i \cap A_j = \phi$ si $i \neq j$) :

$$P\left[\bigcup_{i=1}^{\infty} A_i\right] = \sum_{i=1}^{\infty} P(A_i).$$

HISTORIQUE

Les premiers jeux de hasard marquent le début de l'histoire des probabilités et on peut affirmer qu'ils datent de l'époque de l'*Homo sapiens*.

L'origine du mot « hasard » toutefois offre moins de certitude. Selon Maurice George Kendall (1956), il aurait été rapporté en Europe au moment de la Troisième Croisade et serait dérivé du mot arabe « al zhar » désignant un dé.

J.P. Benzécri (1982), quant à lui, fait référence à un chroniqueur du XII^{e} siècle, Franc d'Orient, nommé Guillaume de Tyr, selon lequel le mot « hasard » serait lié au nom d'un château en Syrie. Ce château aurait fait l'objet d'un siège pendant lequel les assaillants auraient inventé un jeu de dés donnant son nom à l'ensemble de ces jeux.

On trouve la trace des jeux de hasard dans un grand nombre de civilisations anciennes. Ainsi les peintures et objets retrouvés dans les tombes égyptiennes de la Première Dynastie (3500 avant J.-C.) montrent-ils que ces jeux étaient déjà pratiqués à l'époque.

Pour Hérodote, historien grec, c'est au cours du siège de Troie que Palamède aurait inventé les jeux de hasard ; en effet, pendant un siège qui dura dix ans, il fallait trouver un moyen de distraire les soldats.

Le jeu d'astragales de chien fait partie de ces premiers jeux. L'astragale est un petit os du pied, symétrique par rapport à un axe vertical. Très prisé par les Grecs, puis plus tard par les Romains, il pourrait être, selon M.G. Kendall (1956), à l'origine du jeu de dés. En effet, dans la Rome antique et en Égypte, l'astragale fut imité par de petits morceaux de pierre taillée. Il aurait ensuite été façonné jusqu'à présenter la forme d'un cube.

De très anciens dés furent retrouvés en Égypte, mais le plus ancien, fabriqué en terre cuite, fut découvert au nord de l'Irak. Il date du troisième millénaire avant J.-C. Un autre, en terre également, et datant de la même époque, fut mis à jour en Inde. Il existe donc de nombreux indices prouvant la vogue de jeux de dés dans les temps anciens. Personne toutefois n'a songé à l'époque à établir la propriété équiprobable d'obtenir n'importe quelle face.

Pour F.N. David (1955) deux hypothèses peuvent expliquer cette lacune : la première réside dans le fait que les dés étaient généralement pipés, et la seconde dans le caractère religieux attaché au jet de dés. Celui-ci constituait en effet un des moyens d'interroger les dieux : l'interprétation du résultat du lancer de dés permettait de connaître la réponse de la divinité aux questions posées.

Parmi les jeux de hasard, les cartes sont également d'origine très ancienne. On sait qu'elles étaient utilisées en Chine, en Inde, en Arabie et en Égypte. En Occident, on en trouve les premières traces à Venise en 1377, à Nuremberg en 1380 et à Paris en 1397. Le jeu du tarot serait d'ailleurs le plus ancien.

Jusqu'au XV^{e} siècle, selon M.G. Kendall (1956), l'Église et les souverains luttèrent contre la pratique des jeux de dés. On peut citer à ce propos les lois promulguées sous Louis IX, Edward III et Henri VIII. Les jeux de cartes furent ajoutés à la liste de ces pratiques illégales, on en retrouve la mention sur une ordonnance de prohibition émise à Paris en 1397. Malgré ces interdits, les jeux de hasard furent présents sans interruption, des Romains à la Renaissance, dans toutes les classes sociales.

Il n'est donc pas étonnant de constater qu'ils sont à l'origine des travaux des mathématiciens sur les probabilités.

En Occident, les premiers écrits sur ce sujet viennent d'Italie. Ils sont dus à Cardano avec son traité *Liber de Ludo Aleæ* publié en 1663, après la mort de son auteur et à Galilée avec son *Sopra le Scoperte de i Dadi*. Dans les travaux de Galilée on trouve le principe de l'équi-probabilité des jeux de dés.

Dans le reste de l'Europe, ces problèmes soulevaient également l'intérêt des mathématiciens. Huygens, originaire des Pays-Bas, publia en 1657 un ouvrage intitulé *De Ratiociniis in Aleæ Ludo* qui selon P.

Holgate (1984) étendait à d'autres domaines l'intérêt des probabilités ; la théorie des jeux se serait en effet développée parallèlement à celle relative à la dynamique des corps. Les travaux de Huygens influenceront considérablement ceux de deux autres mathématiciens célèbres : **Jakob Bernoulli** et **Abraham de Moivre**.

Toutefois c'est en France, avec Blaise Pascal (1623-1662) et Pierre de Fermat (1601-1665), que la théorie des probabilités va véritablement prendre forme. Selon I. Todhunter (1949), B. Pascal fut sollicité par un joueur réputé, Antoine Gombauld, chevalier de Méré, pour résoudre le problème suivant : quelles sont les chances de succès de deux adversaires, sachant qu'à un certain stade du jeu l'un a gagné $n < m$ parties et l'autre $p < m$, le premier qui gagne m parties devant remporter toute la mise.

B. Pascal prit contact avec P. de Fermat qui trouva une solution. B. Pascal, quant à lui, découvrit la formule de récurrence lui permettant d'aboutir à un résultat identique.

En Europe, il fallut donc attendre quelque quatorze siècles pour qu'apparaissent les premières découvertes sur les probabilités. Cependant, selon J.P. Benzécri (1982), à l'époque où se produisaient en Occident la chute de l'Empire romain et l'avènement de l'ère chrétienne, l'Orient connaissait une intense activité scientifique et artistique.

Dans ces circonstances, des érudits orientaux, tels Omar Khayyâm, n'auraient-ils pas découvert ces règles de probabilités ?

Il est un aspect des probabilités dont on sait qu'il a particulièrement intéressé les mathématiciens : il s'agit des problèmes d'**analyse combinatoire**.

ASPECTS MATHÉMATIQUES

Bases axiomatiques des probabilités

On considère une **expérience aléatoire** dont l'**ensemble fondamental** Ω contient n éléments :

$$\Omega = \{x_1, x_2, \ldots, x_n\}.$$

On peut associer à chaque **événement** simple x_i une probabilité $P(x_i)$ ayant les propriétés suivantes :

1. les probabilités $P(x_i)$ sont non négatives :
$$P(x_i) \geq 0 \quad \text{pour } i = 1, 2, \ldots, n,$$
2. la somme des probabilités pour i allant de 1 à n est égale à 1 :
$$\sum_{i=1}^{n} P(x_i) = 1,$$
3. la probabilité P est une fonction qui assigne un nombre compris entre 0 et 1 à tout événement simple d'une expérience aléatoire :
$$P \ : \ \Omega \to [0, 1].$$

Propriétés des probabilités

1. La probabilité de l'événement certain est la plus grande probabilité qui peut être associée à un événement :
$$P(\Omega) = 1.$$
2. La probabilité de l'événement impossible est égale à 0 :
$$\text{si } A = \phi, \quad P(A) = 0.$$
3. Soit $\bar{A}$ le **complémentaire** de A dans Ω, la probabilité de $\bar{A}$ est égale à 1 moins la probabilité de A :
$$P\left(\bar{A}\right) = 1 - P(A).$$
4. Soient A et B deux événements incompatibles ($A \cap B = \phi$), la probabilité de $A \cup B$ est égale à la somme des probabilités de A et de B :
$$P(A \cup B) = P(A) + P(B).$$
5. Soient A et B deux événements quelconques, la probabilité de $A \cup B$ est égale à :
$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

MOTS CLÉS

Ensemble fondamental *(Sample space)*
Événement *(Event)*

Expérience aléatoire *(Random experiment)*
Loi de probabilité continue *(Continuous probability distribution)*
Loi de probabilité discrète *(Discrete probability distribution)*
Probabilité conditionnelle *(Conditional probability)*

RÉFÉRENCES

Benzécri, J.-P. (1982). *Histoire et préhistoire de l'analyse des données.* Dunod, Paris.

David, F.N. (1955). Dicing and gaming (a note on the history of probability). *Biometrika*, 42, pp. 1-15.

Holgate, P. (1984). The influence of Huygens'work in Dynamics on his Contribution to Probability. *International Statistical Review.* Vol. 52, pp. 137-140.

Kendall, M.G. (1956). Studies in the History of Probability and Statistics: II. The Beginnings of a Probability Calculus. *Biometrika*, 43, pp. 1-14.

Stigler, S. (1986). *The History of Statistics, the Measurement of Uncertainty before 1900.* The Belknap Press of Harvard University Press. London, England.

Todhunter, I. (1949). *A history of the mathematical theory of probability.* Chelsea Publishing Company, New York.

PROBABILITÉ, LOI DE

Probability distribution

Voir **loi de probabilité**.

PROBABILITÉ CONDITIONNELLE

Conditional probability

La probabilité conditionnelle de la réalisation de l'**événement** A en fonction de l'événement B représente la **probabilité** que A se réalise en sachant que B est réalisé.

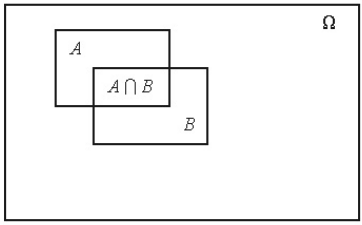
Cette probabilité conditionnelle se note $P(A|B)$ et se lit « probabilité de A conditionnée par B ».

HISTORIQUE

La notion d'**indépendance** domine les probabilités jusqu'au début du XX^{e} siècle. **Andreï Nicolaïévitch Kolmogorov** (1933) a introduit la notion de probabilité conditionnelle ; cette notion joue un rôle essentiel dans la théorie et dans l'application des probabilités et des statistiques.

ASPECTS MATHÉMATIQUES

Soit une **expérience aléatoire** dont nous connaissons l'**ensemble fondamental** Ω. Considérons deux événements A et B de cet ensemble. La **probabilité** de A, $P(A)$ est donnée en fonction de l'ensemble des événements possibles de l'**expérience** (Ω).



Considérons à présent que nous connaissons une information supplémentaire concernant l'expérience : l'événement B est réalisé. La probabilité de réalisation de l'événement A ne sera plus donnée en fonction de Ω, mais en fonction de l'ensemble B. Cette probabilité de A conditionnée par B se calcule de la façon suivante :

$$P(A|B) = \frac{P(A \cap B)}{P(B)}.$$

Si A et B sont deux événements incompatibles, l'intersection entre A et B est l'ensemble vide. Nous aurons donc $P(A|B) = P(B|A) = 0$.

DOMAINES ET LIMITATIONS

La notion de probabilité conditionnelle est l'une des plus importantes en théorie des probabilités. Cette importance ressort particulièrement des points suivants :

1. Nous nous intéressons souvent à calculer la probabilité d'un **événement** lorsqu'une partie de l'information concernant le résultat est déjà connue. Dans ce cas, la probabilité à calculer est une probabilité conditionnelle.

2. Même lorsqu'aucune information partielle concernant le résultat n'est connue, les probabilités conditionnelles peuvent être utiles pour permettre le calcul des probabilités cherchées.

EXEMPLE

Considérons un lot de 100 voitures réparties selon deux critères, le confort et la vitesse. Nous ferons les distinctions suivantes :

$$\text{une voiture peut être} \begin{cases} \text{rapide} \\ \text{lente} \end{cases}$$

$$\text{une voiture peut être} \begin{cases} \text{confortable} \\ \text{inconfortable} \end{cases}$$

La répartition des 100 voitures selon ces critères est donnée dans le tableau ci-dessous :

	rapide	lente	total
confortable	40	10	50
inconfortable	20	30	50
total	60	40	100

Soient les deux événements

$A =$ « choisir une voiture rapide »
et $B =$ « choisir une voiture confortable ».

Les probabilités de ces deux événements sont :

$$P(A) = 0,6$$
$$P(B) = 0,5$$

La probabilité de choisir une voiture rapide est donc de $0,6$ et celle de choisir une voiture confortable de $0,5$.

Imaginons à présent que nous connaissions une information supplémentaire : la voiture qui a été choisie est rapide. Quelle est alors la probabilité que cette voiture soit aussi confortable ?

Nous allons donc calculer la probabilité de B à condition que A soit réalisé, ou la probabilité conditionnelle de B en fonction de A.

Nous pouvons trouver dans le tableau que

$$P(A \cap B) = 0,4.$$

$$\Rightarrow P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{0,4}{0,6} = 0,667.$$

La probabilité que la voiture soit confortable alors que nous savons qu'elle est rapide est donc de $\frac{2}{3}$.

MOTS CLÉS

Ensemble fondamental *(Sample space)*
Événement *(Event)*
Expérience aléatoire *(Random experiment)*
Probabilité *(Probability)*

RÉFÉRENCE

Kolmogorov, A.N. (1933). *Grundbegriffe der Wahrscheinlichkeitsrechnung.* Springer-Verlag, Berlin (1956). *Foundations of the Theory of Probability.* Chelsea Publishing Company, New York.

PROBABILITÉ CONTINUE, LOI DE

Continuous probability distribution

Voir **loi de probabilité continue**.

PROBABILITÉ DISCRÈTE, LOI DE

Discrete probability distribution

Voir **loi de probabilité discrète**.

PROCESSUS STOCHASTIQUE

Stochastic Process

On appelle processus stochastique une famille de variables aléatoires. En pratique, ils servent à modéliser un grand nombre de phénomènes temporels où intervient le hasard.

On distingue plusieurs types de processus stochastiquey à l'aide de certaines propriétés mathématiques, les plus connus sont :

- le processus de Markov (ou chaînes de Markov) ;
- la martingale ;
- le processus stationnaire ;
- le processus à incréments indépendants.

HISTORIQUE

L'origine des processus stochastiques remonte aux progrès faits au début du XX^{e} siècle dans certaines branches appliquées, telles que la mécanique statistique (par Gibbs, Boltzmann, Poincaré, Smoluchowski et Langevin). Les bases théoriques ont été formulées plus tard par Doob, Kolmogorov et d'autres (1930-1940). C'est durant cette période que le mot « stochastique », qui provient du grec *stokhastikos* « conjectural », a commencé à être employé. D'autres progrés ont été faits aussi à partir du mouvement brownien en physique (par Einstein, Lévy et Wiener).

Le processus de Markov prend son nom d'**Andreï Andreïevitch Markov** (1856-1922) qui a posé les bases de la théorie des processus de Markov à temps fini, généralisée par le mathématicien russe **Andreï Nicolaïévitch Kolmogorov** (1936).

DOMAINES ET LIMITATIONS

Théoriquement, il n'y a pas de différences entre l'étude des processus stochastiques et la statistique en soi. En pratique, l'étude des processus stochastiques est confrontée plutôt avec la structure et les propriétés des modèles qu'avec l'inférence à partir des données réelles, domaine propre à la statistique.

ASPECTS MATHÉMATIQUES

Un processus stochastique est une famille $(X_t)_{t\in I}$ (où I est un ensemble discret ou continu) de variables aléatoires définies sur le même espace de probabilité. Les processus *à temps discrétisé* sont représentés par les familles suivantes $(X_t)_{t\in\mathbb{N}}$ (c'est-à-dire avec $I=\mathbb{N}$) et les processus *à temps continu* par les familles $(X_t)_{t\in\mathbb{R}_+}$ (c'est-à-dire avec $I=\mathbb{R}_+=[0,\infty)$).

Le terme « processus stochastique » est employé si l'ensemble I est infini. En physique, les éléments de I répresentent le temps.

Pour un **événement** w appartenant à l'**ensemble fondamental** Ω, on définit la *trajectoire* du processus étudié comme la famille suivante :

$$(X_t(w))_{t\in I}$$

donc comme la « série » (pour un processus à temps discrétisé) des valeurs des variables aléatoires prises pour un événement choisi et fixe.

Les objets principaux d'étude d'un processus aléatoire sont les proprietés des variables aléatoires et des trajectoires.

On caractérise les différents types de processus par rapport aux conditions probabilistes auxquelles ils sont soumis.

1. *Processus de Markov*

   Un processus de Markov est un processus aléatoire $(X_t)_{t\in\mathbb{N}}$ à temps discrétisé pour lequel la probabilité d'un événement (au temps $t+1$) ne dépend que du passé immédiat (temps t). Cela s'exprime mathématiquement par la condition suivante :

   pour tout $t\in\mathbb{N}$ et pour tout i_0, i_1, …, $i_{t+1}\in\mathbb{R}_+$ on a :
   $P(X_{t+1}=i_{t+1}|X_t=i_t,X_{t-1}=i_{t-1},\ldots,X_0=i_0)=P(X_{t+1}=i_{t+1}|X_t=i_t)$.

2. *Martingale*

   Une martingale est un processus aléatoire $(X_t)_{t\in\mathbb{N}}$ à temps discrétisé qui vérifie la condition suivante :

   pour tout $t \in \mathbb{N}$:

   $$E(X_{t+1}\,|X_t, X_{t-1}, \ldots, X_0\,) = X_t.$$

3. *Processus stationnaire*

   Un processus stationnaire (au sens strict) est un processus aléatoire $(X_t)_{t\in\mathbb{N}}$ pour lequel les distributions conjointes du processus dépendent seulement de la différence temporelle (entre les variables considérées dans la distribution conjointe) et pas du temps « absolu ». Mathématiquement :

   pour tout $t_1, t_2, \ldots, t_k \quad \in I(k \geq 2)$ et pour tout $a \in I$, $X_{t_1}, X_{t_2}, \ldots, X_{t_k}$ ont la même distribution conjointe que $X_{t_1+a}, X_{t_2+a}, \ldots, X_{t_k+a}$.

4. *Processus à incréments indépendants*

   Un processus à incréments indépendants est un processus aléatoire $(X_t)_{t\in I}$ pour lequel :

   pour tout $t_1 < t_2 < \ldots < t_k \in I$: $X_{t_2} - X_{t_1}$, $X_{t_3} - X_{t_2}$, …, $X_{t_k} - X_{t_{k-1}}$ sont des variables aléatoires réciproquement indépendantes .

EXEMPLE

Exemple 1 : on donne ici un exemple qui montre le lien mathématiquement étroit entre la statistique et les processus stochastiques.

On considère les variables aléatoires X_i = âge de l'individu i d'une population (potentiellement infinie). Un problème typique de statistique est d'estimer la moyenne $E\,(X)$ avec les hypothèses suivantes :

- les variables aléatoires X_i sont indépendantes ;
- les X_i sont identiquement distribuées.

Exemple 2 : marche aléatoire. Soient Z_n, $n \geq 2$, des variables aléatoires indépendantes et identiquement distribuées prenant la valeur $+1$ avec probabilité p et la valeur -1 avec probabilité $q = 1 - p$.

On définit le processus stochastique de marche aléatoire comme suit :

$$X_0 = 0, X_t = \sum_{n=1}^{t} Z_n, \quad \text{si } t > 1.$$

Un échantillon d'une marche aléatoire se présente typiquement comme la séquence suivante :

$$(0, 1, 2, 1, 0, -1, 0, -1, -2, -3, -2, \ldots)\,.$$

Il est facile de voir que le processus stochastique défini est un processus de Markov (ce fait est une conséquence directe de l'indépendance des X_t) avec les probabilités suivantes de transition :

$$\begin{aligned} P\,(X_{t+1} &= i+1\,|X_t = i\,) = p, \\ P\,(X_{t+1} &= i-1\,|X_t = i\,) = q. \end{aligned}$$

On a donc :

$$E(X_{t+1}\,|X_t\,) = X_t + (+1)\cdot p + (-1)\cdot q = X_t + (p-q).$$

La marche aléatoire est donc une martingale si et seulement si $p = q = 0,5$.

MOTS CLÉS

Variable aléatoire *(Random variable)*
Convergence *(Convergence)*
Série chronologique *(Time series)*
Statistique *(Statistics)*

RÉFÉRENCES

Bartlett, M.S. (1960). *An introduction to Stochastic Processes*, Cambridge University Press, Cambridge.

Billingsley, P. (1961). *Statistical inference for Markov Processes.* University of Chicago Press, Chicago, IL.

Cox, D.R. and Miller, H.D. (1965). *Theory of stochastic processes.* John Wiley & Sons, New York ; Methuen, London.

Doob, J.L. (1942). What is a stochastic process ? *The American Mathematical Monthly*, 49, pp. 648-653.

Doob, J.L. (1953). *Stochastic Processes.* John Wiley & Sons, New York.

Kolmogorov, A.N. (1936). *Math. Sbornik. N.S.*, 1, pp. 607-610.

PRODUIT SCALAIRE

Scalar product

On définit le produit scalaire de deux vecteurs $\mathbf{x}$ et $\mathbf{y}$ de même taille par le produit matriciel de la **transposée** de l'un par l'autre. Autrement dit, le produit scalaire de $\mathbf{x}$ et $\mathbf{y}$, noté $\mathbf{x}' \cdot \mathbf{y}$, est la somme des produits composante par composante des deux vecteurs.

Géométriquement, le produit scalaire de deux vecteurs de taille n correspond, dans l'espace à n dimensions, au nombre obtenu en multipliant la **norme** de l'un par celle de la projection de l'autre sur ce dernier. Cette définition équivaut à dire que le produit scalaire de deux vecteurs s'obtient en effectuant le produit des normes de ces deux vecteurs par le cosinus de l'angle déterminé par les deux vecteurs.

ASPECTS MATHÉMATIQUES

Soient $\mathbf{x}$ et $\mathbf{y}$ deux vecteurs de taille n, donnés par leurs composantes :

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \text{ et } \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

le produit scalaire de $\mathbf{x}$ et $\mathbf{y}$, noté $\mathbf{x}' \cdot \mathbf{y}$, est défini par :

$$\begin{aligned} \mathbf{x}' \cdot \mathbf{y} &= [x_1 \quad x_2 \quad \dots x_n] \cdot \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \\ \mathbf{x}' \cdot \mathbf{y} &= x_1 \cdot y_1 + x_2 \cdot y_2 + \dots + x_n \cdot y_n. \\ \mathbf{x}' \cdot \mathbf{y} &= \sum_{i=1}^{n} x_i \cdot y_i \,. \end{aligned}$$

Le produit scalaire d'un vecteur $\mathbf{x}$ par lui-même correspond à sa **norme** (ou longueur) au carré, notée $\| \mathbf{x} \|^2$.

$$\| \mathbf{x} \|^2 = \mathbf{x}' \cdot \mathbf{x} = \sum_{i=1}^{n} x_i^2.$$

À l'aide de la norme, on peut calculer le produit scalaire

$$\mathbf{x}' \cdot \mathbf{y} = \| \mathbf{x} \| \cdot \| \mathbf{y} \| \cdot \cos \alpha$$

où α est l'angle compris entre le vecteur $\mathbf{x}$ et le vecteur $\mathbf{y}$.

Le produit scalaire de deux vecteurs jouit des propriétés suivantes :

- il est commutatif, c'est-à-dire pour deux vecteurs de même dimension $\mathbf{x}$ et $\mathbf{y}$, $\mathbf{x}' \cdot \mathbf{y} = \mathbf{y}' \cdot \mathbf{x}$. Autrement dit, cela n'a aucune importance si l'on projette $\mathbf{x}$ sur $\mathbf{y}$ ou le contraire ;
- il peut prendre des valeurs négatives si, par exemple, la projection d'un vecteur et celui sur lequel on a fait la projection ont des sens opposés.

EXEMPLE

On considère deux vecteurs dans l'espace à trois dimensions, donnés par leurs composantes :

$$\mathbf{x} = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} \text{ et } \mathbf{y} = \begin{bmatrix} 0 \\ 2 \\ -1 \end{bmatrix}.$$

Le produit scalaire de $\mathbf{x}$ et $\mathbf{y}$, noté $\mathbf{x}' \cdot \mathbf{y}$, est défini par :

$$\begin{aligned} \mathbf{x}' \cdot \mathbf{y} &= \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}' \cdot \begin{bmatrix} 0 \\ 2 \\ -1 \end{bmatrix} = [1\ 2\ 3] \cdot \begin{bmatrix} 0 \\ 2 \\ -1 \end{bmatrix} \\ \mathbf{x}' \cdot \mathbf{y} &= 1 \cdot 0 + 2 \cdot 2 + 3 \cdot (-1) = 4 - 3. \\ \mathbf{x}' \cdot \mathbf{y} &= 1. \end{aligned}$$

La **norme** de $\mathbf{x}$, notée $\| \mathbf{x} \|$, est donnée par la racine carrée du produit scalaire par lui-même.

Ainsi :

$$\begin{aligned} \mathbf{x}' \cdot \mathbf{x} &= \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}' \cdot \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} = [1\ 2\ 3] \cdot \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} \\ \mathbf{x}' \cdot \mathbf{x} &= 1^2 + 2^2 + 3^2 = 1 + 4 + 9. \\ \mathbf{x}' \cdot \mathbf{x} &= 14. \end{aligned}$$

$$\text{d'où } \| \mathbf{x} \| = \sqrt{14} \approx 3,74.$$

Le vecteur $\mathbf{x}$ a donc une longueur de $3,74$ dans l'espace à trois dimensions.

De même, on peut calculer la norme du vecteur $\mathbf{y}$. Elle est donnée par :

$$\| \mathbf{y} \|^2 = \mathbf{y}' \cdot \mathbf{y} = \sum_{i=1}^{3} y_i^2 = 2^2 + (-1)^2 = 5.$$

et vaut $\| \mathbf{y} \| = \sqrt{5}$.
On peut alors déterminer l'angle α situé entre les deux vecteurs, à l'aide de la relation :

$$\cos \alpha = \frac{\mathbf{x}' \cdot \mathbf{y}}{\| \mathbf{x} \| \cdot \| \mathbf{y} \|} = \frac{1}{\sqrt{14} \cdot \sqrt{5}} \approx 0,1195.$$

Ce qui donne un angle approximatif de $83,1°$ entre les vecteurs $\mathbf{x}$ et $\mathbf{y}$.

MOTS CLÉS

Matrice *(Matrix)*
Norme *(Norm)*
Transposée *(Transpose)*
Vecteur *(Vector)*

PROGRAMMATION LINÉAIRE

Linear programming

Les problèmes de programmation mathématique concernent l'allocation optimale de ressources limitées pour atteindre des objectifs fixés.

On parle de programmation linéaire lorsque le **modèle** mathématique peut être posé en utilisant des relations linéaires.

Un problème de programmation linéaire se présente sous la forme d'un ensemble d'équations (ou d'inégalités) linéaires appelées contraintes et d'une fonction linéaire qui exprime l'objectif du problème appelée fonction objectif.

HISTORIQUE

Le problème de la programmation est né dans le domaine de l'économie dans les années 1930. On peut citer notamment les travaux remarquables dans le domaine des modèles linéaires réalisés par Johannes Von Neumann (1935-1936). En 1947, George B. Dantzig donne la méthode du simplexe pour la solution des problèmes de programmation linéaire.

ASPECTS MATHÉMATIQUES

Un problème de programmation linéaire peut être défini de la façon suivante :

Trouver un ensemble de valeurs pour les variables $x_1, x_2, \ldots, x_n$ qui maximise ou minimise la fonction objectif linéaire :

$$z = c_1 \cdot x_1 + c_2 \cdot x_2 + \ldots + c_n \cdot x_n$$

soumise aux contraintes linéaires :

$$\begin{aligned} a_{11} \cdot x_1 + a_{12} \cdot x_2 + \ldots + a_{1n} \cdot x_n &\left\{ \begin{matrix} \leq \\ = \\ \geq \end{matrix} \right\} b_1 \\ a_{21} \cdot x_1 + a_{22} \cdot x_2 + \ldots + a_{2n} \cdot x_n &\left\{ \begin{matrix} \leq \\ = \\ \geq \end{matrix} \right\} b_2 \\ &\vdots \\ a_{m1} \cdot x_1 + a_{m2} \cdot x_2 + \ldots + a_{mn} \cdot x_n &\left\{ \begin{matrix} \leq \\ = \\ \geq \end{matrix} \right\} b_m \end{aligned}$$

et aux contraintes de non-négativité :

$$x_j \geq 0,\ j = 1, 2, \ldots, n$$

où a_{ij}, b_i et c_j sont des constantes connues.

En d'autres termes, cela signifie que, dans un problème de programmation linéaire, on désire trouver une solution non négative qui satisfasse les contraintes et optimise la fonction objectif.

On peut écrire un problème de programmation linéaire sous forme matricielle :

Maximiser (ou minimiser) $\mathbf{z} = \mathbf{c}' \cdot \mathbf{x}$ sous les contraintes $\mathbf{A} \cdot \mathbf{x} \left\{ \begin{array}{c} \leq \\ = \\ \geq \end{array} \right\} \mathbf{b}$ et $\mathbf{x} \geq \mathbf{0}$

où

- $\mathbf{c}'$ est un vecteur-ligne de dimension $(1 \times n)$,
- $\mathbf{x}$ est un vecteur-colonne de dimension $(n \times 1)$,
- $\mathbf{A}$ est une **matrice** de dimension $(m \times n)$,
- $\mathbf{b}$ est un vecteur-colonne de dimension $(m \times 1)$ et
- $\mathbf{0}$ est le vecteur nul à n composantes.

L'outil principal pour résoudre des problèmes de programmation linéaire est la méthode du simplexe. Celle-ci consiste en un ensemble de règles à suivre pour obtenir la solution d'un problème donné. C'est une méthode itérative qui donne la solution exacte en un nombre fini d'étapes.

DOMAINES ET LIMITATIONS

Les problèmes de programmation linéaire sont des problèmes dans lesquels sont impliqués des hommes, des services, des machines et des matières diverses. La programmation linéaire cherche à déterminer l'allocation optimale de ces ressources pour produire un ou plusieurs produits, tout en maximisant ou minimisant des fonctions telles que le profit ou le coût.

C'est pour cette raison que les coefficients de la fonction objectif sont souvent appelés les « prix » associés aux variables.

EXEMPLE

Une entreprise fabrique des chaises et des tables à l'aide de deux machines A et B. Chaque produit passe obligatoirement par les deux machines. Pour produire une chaise, il faut deux heures de machine A et une heure de machine B. Pour produire une table, il faut une heure de machine A et deux heures de machine B. L'entreprise réalise un profit de 3 euros sur chaque chaise et de 4 euros sur chaque table. Les deux machines A et B sont disponibles 12 heures par jour au maximum.

À partir de ces données, il s'agit de trouver la production d'ensemble pour un jour assurant le maximum de profit. Autrement dit, le fabricant veut connaître le programme de production optimal, c'est-à-dire les quantités de chaises et de tables qu'il doit fabriquer par jour pour obtenir le bénéfice maximum.

On note par x_1 le nombre de chaises et par x_2 le nombre de tables qui doivent être produites par jour. Le bénéfice journalier est donc donné par la fonction :

$$z = 3 \cdot x_1 + 4 \cdot x_2.$$

Cette fonction objectif doit donc être maximisée.

On sait que le produit x_1 requiert deux heures de machine A et le produit x_2 requiert une heure de machine A. Par conséquent, le total en heures par jour pour la machine A est de :

$$2 \cdot x_1 + 1 \cdot x_2$$

puisqu'on fabrique x_1 chaises et x_2 tables. Mais le temps total de fonctionnement de la machine A ne peut excéder 12 heures. Mathématiquement, cela se traduit par :

$$2x_1 + x_2 \leq 12.$$

On fait le même raisonnement pour la machine B et on trouve :

$$x_1 + 2x_2 \leq 12.$$

On a donc deux contraintes. De plus, on ne peut pas produire des quantités négatives ; par conséquent, s'ajoutent deux contraintes de non-négativité :

$$\begin{aligned} x_1 &\geq 0 \\ \text{et } x_2 &\geq 0. \end{aligned}$$

On veut trouver des valeurs pour les variables x_1 et x_2 qui satisfassent les contraintes et maximisent

le profit. On peut finalement représenter le **modèle** mathématique de la façon suivante :

$$\begin{aligned} \max z &= 3x_1 + 4x_2 \\ \text{s.c. } 2x_1 + x_2 &\leq 12 \\ x_1 + 2x_2 &\leq 12 \\ \text{et } x_1, x_2 &\geq 0 \end{aligned}$$

(s.c. signifie sous les contraintes).

Le profit maximal est obtenu pour le couple $(x_1, x_2) = (4, 4)$ il vaut 28 euros et est fourni par la production de 4 chaises et 4 tables. Dans le cas précis, on peut trouver la réponse au problème en calculant le profit pour chacun des couples possibles (leur nombre étant inférieur à 36, $x_1 \leq 6$ et $x_2 \leq 6$ se déduisant des contraintes).

MOTS CLÉS

Optimisation *(Optimization)*
Recherche opérationnelle *(Operations research)*

RÉFÉRENCES

Arthanari, T.S. and Dodge, Y. (1981). *Mathematical Programming in Statistics.* John Wiley & Sons, New York.

Dantzig, G.B. (1966). *Applications et prolongements de la programmation linéaire.* Dunod, Paris.

Dodge, Y. (1987). *Programmation linéaire.* EDES, Neuchâtel.

Hadley, G. (1962). *Linear programming.* Addison-Wesley, Reading, Mass.

Von Neumann, J. (1937). Über ein ökonomisches Gleichungssystem und eine Verallgemeinerung des Brouwerschen Fixpunktsatzes. *Ergebn. math. Kolloqu. Wien* 8, 73-83.


Q

Q-Q PLOT (QUANTILE TO QUANTILE PLOT)

Q-Q plot (Quantile to Quantile Plot)

Le Q-Q plot (ou *Quantile to Quantile Plot*) est un graphique dont l'objectif est de tester la conformité entre la distribution empirique d'une variable et une distribution théorique donnée.

Une des méthodes utilisées pour vérifier la normalité des erreurs d'un modèle de régression est d'effectuer le Q-Q plot des résidus. Si les points sont alignés sur la droite $x = y$, alors les données sont distribuées selon la loi normale.

HISTORIQUE

La méthode de **représentation graphique** Q-Q plot est apparue au début des années 1960. Depuis lors, elle est devenue un outil incontournable lors de l'analyse des données et/ou des résidus, car c'est une méthode très suggestive et d'interprétation aisée.

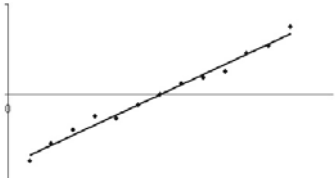
ASPECTS MATHÉMATIQUES

Pour construire un Q-Q plot on suit deux étapes. La première étape consiste dans le classement des données x_1, x_2, …, x_n dans l'ordre croissant :

$$x_{[1]} \leq x_{[2]} \leq \ldots \leq x_{[n]}.$$

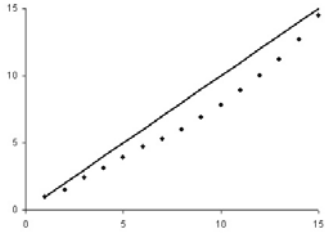
Pour la deuxième étape, on associe à chaque donnée $x_{[i]}$ le $i/(n+1)$-**quantile** q_i d'une loi normale centrée réduite. On représente alors sur un graphique, en ordonnée les données ordonnées x_i et en abcisse les quantiles q_i.

Si deux variables ont la même distribution, alors la représentation graphique entre les quantiles de la première variable par rapport aux quantiles de la deuxième distribution sera une ligne droite de pente 1. Donc si les données x_i sont normalement distribuées, les points sur le graphique doivent être à peu près alignés sur la droite d'équation $x_i = q_i$.

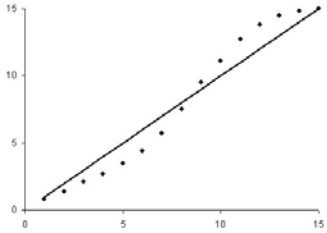


DOMAINES ET LIMITATIONS

Le Q-Q plot est utilisé pour vérifier si les données suivent une distribution particulière, ou si deux jeux de données ont la même distribution. Si les distributions sont les mêmes, le graphique montre une ligne droite. Plus le résultat obtenu s'écarte de la diagonale à 45°, plus la distribution empirique s'écarte de la distribution théorique. Les points extrêmes ont une variabilité plus grande que ceux se trouvant au centre de la distribution.



Ainsi, un graphique en forme de « U » signifie qu'une distribution est asymétrique (*skewed*) par rapport à l'autre.



Si le graphique est en forme de « S », alors une des distributions présente une plus grande influence des valeurs extrêmes que l'autre distribution (*long tail*).

Une utilisation fréquente des Q-Q plots est lorsqu'on désire connaître le comportement des résidus, lors d'une **régression linéaire simple** ou **multiple**, plus spécifiquement lorsqu'on désire savoir si leur distribution suit une **loi normale**.

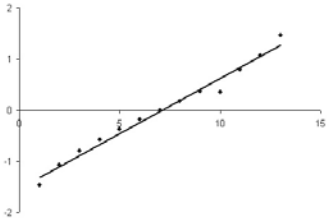
EXEMPLE

Le tableau ci-dessous contient les résidus (rangés par ordre croissant) d'une régression faite à partir de 13 observations. On leur associe les quantiles q_i comme décrit ci-dessus.

Résidus et quantiles

i	e_i	q_i
1	$-1,62$	$-1,47$
2	$-0,66$	$-1,07$
3	$-0,62$	$-0,79$
4	$-0,37$	$-0,57$
5	$-0,17$	$-0,37$
6	$-0,12$	$-0,18$
7	$0,04$	$0,00$
8	$0,18$	$0,18$
9	$0,34$	$0,37$
10	$0,42$	$0,57$
11	$0,64$	$0,79$
12	$1,33$	$1,07$
13	$2,54$	$1,47$

Le Q-Q plot obtenu est représenté par la figure ci-dessous. On voit qu'à part deux points qui constituent peut-être des valeurs aberrantes, les points sont bien alignés.



MOTS CLÉS

Analyse de données *(Data analysis)*
Analyse exploratoire des données *(Exploratory data analysis)*
Analyse des résidus *(Analysis of residuals)*
Quantile *(Quantile)*
Loi de probabilité *(Probability distribution)*
Régression linéaire multiple *(Multiple linear regression)*
Régression linéaire simple *(Simple linear regression)*
Résidu *(Residual)*

RÉFÉRENCES

Chambers, J.M., Cleveland, W.S., Kleiner, B. and Tukey, P.A. (1983). *Graphical Methods for Data Analysis.* Wadsworth, Belmont, California.

Hoaglin, D.C., Mosteller, F. and Tukey, J.W., editors (1983). *Understanding Robust and Exploratory Data Analysis.* John Wiley & Sons, New York.

Wilk, M.B., Ganadesikan, R. (1968). Probability Plotting Methods for the Analysis of Data. *Biometrika*, 55, pp. 1-17.

QUANTILE

Quantile

Les quantiles sont des mesures de position (ou **mesure de tendance centrale**) qui ne tentent pas nécessairement de déterminer le centre d'une distribution d'observations, mais de décrire une position particulière.

Cette notion est une extension du concept de la **médiane** (qui divise une distribution d'observations en deux parties). Les quantiles les plus fréquemment utilisés sont :

- les quartiles qui divisent un ensemble d'observations en quatre parties ;
- les déciles qui divisent un ensemble d'observations en dix parties ;

- les centiles qui divisent un ensemble d'observations en cent parties.

DOMAINES ET LIMITATIONS

Le calcul des quantiles n'a de sens que pour une **variable quantitative** pouvant prendre des valeurs sur un **intervalle** déterminé.

Le concept de quantile indique la division d'une distribution d'observations en un nombre quelconque de parties. Remarquons que plus le nombre d'observations est élevé, et plus nous pouvons diviser finement la distribution.

Les quartiles peuvent généralement être utilisés pour toute distribution.

Le calcul des déciles et *a fortiori* celui des centiles nécessitent un nombre d'observations relativement grand pour obtenir une interprétation valable.

MOTS CLÉS

Centile *(Percentile)*
Décile *(Decile)*
Médiane *(Median)*
Mesure de position *(Measure of location)*
Quartile *(Quartile)*

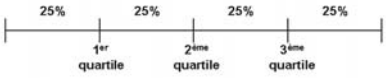
QUARTILE

Quartile

Les quartiles sont des mesures de position d'une distribution d'observations.

Nous appelons quartiles les quantiles qui partagent une distribution en quatre parties.

Nous aurons donc trois quartiles pour une distribution donnée. Entre chaque quartile se trouvent 25 % des observations :



Notons que le 2e quartile est égal à la **médiane**.

ASPECTS MATHÉMATIQUES

Le processus de calcul du quartile est similaire à celui de la **médiane**.

Lorsque nous possédons toutes les observations brutes, le processus de calcul des quartiles est le suivant :

1. Les n observations doivent être organisées sous forme d'une **distribution de fréquences**.
2. Les quartiles correspondent aux observations pour lesquelles la **fréquence** relative cumulée dépasse respectivement 25 %, 50 % et 75 %.

   Certains auteurs proposent la formule suivante qui permet de déterminer avec précision la **valeur** des différents quartiles :

   Calcul du j-ème quartile :

   Soit i la partie entière de $\frac{j\cdot(n+1)}{4}$ et k la partie fractionnelle de $\frac{j\cdot(n+1)}{4}$.

   Soient x_i et x_{i+1} les valeurs des observations classées respectivement en i-ème et $(i+1)$-ème position (lorsque les observations sont classées par ordre croissant).

   Le j-ème quartile est égal à :

   $$Q_j = x_i + k \cdot (x_{i+1} - x_i).$$

Lorsque nous possédons des observations groupées en classes, les quartiles se déterminent de la manière suivante :

1. Déterminer la classe dans laquelle se trouve le quartile :
   - 1er *quartile* : classe pour laquelle la fréquence relative cumulée dépasse 25 %.
   - 2e *quartile* : classe pour laquelle la fréquence relative cumulée dépasse 50 %.
   - 3e *quartile* : classe pour laquelle la fréquence relative cumulée dépasse 75 %.

2. Calculer la valeur du quartile en fonction de l'**hypothèse** selon laquelle les observations sont distribuées uniformément dans chaque classe, le j-ème quartile Q_j est :

$$Q_j = L_j + \left[\frac{n \cdot \frac{j}{4} - \sum f_{inf}}{f_j}\right] \cdot c_j$$

où

L_j = borne inférieure de la classe du quartile Q_j,
n = nombre total d'observations,
$\sum f_{inf}$ = somme des fréquences inférieures à la classe du quartile,
f_j = fréquence de la classe du quartile Q_j et
c_j = dimension de l'**intervalle** de la classe du quartile Q_j.

DOMAINES ET LIMITATIONS

Le calcul des quartiles n'a de sens que pour une **variable quantitative** pouvant prendre des valeurs sur un **intervalle** déterminé.

Il permet d'obtenir des informations relatives aux intervalles dans lesquels se situent les quarts successifs de l'ensemble des observations.

La notion de quartile est similaire à la notion de **médiane**. Elle est aussi basée sur le rang des observations plutôt que sur leur valeur. Une **observation aberrante** n'aura donc que peu d'influence sur la valeur des quartiles.

EXEMPLES

1. Prenons tout d'abord un exemple où nous avons un ensemble de dix observations ($n = 10$) :

   1 2 4 4 5 5 5 6 7 9

   Bien que le calcul des quartiles ne soit pas en principe applicable pour un si petit nombre d'observations (il faut être très prudent dans leur interprétation), nous allons étudier ce cas en vue de comprendre les principes de calcul.

   Le premier quartile Q_1 se trouve à la position $\frac{n+1}{4} = 2,75$. Q_1 est donc entre la 2e et la 3e observation (que nous appellerons x_2 et x_3), aux trois quarts de la **distance** entre ces deux observations. Nous pouvons donc calculer Q_1 de la manière suivante :

$$\begin{aligned} Q_1 &= x_2 + 0,75 \cdot (x_3 - x_2) \\ &= 2 + 0,75 \cdot (4 - 2) \\ &= 3,5. \end{aligned}$$

   Le deuxième quartile Q_2 (qui est égal à la **médiane**), se trouve à la position $2 . \frac{n+1}{4}$ (ou $\frac{n+1}{2}$), ce qui est égal pour notre exemple à 5, 5. Q_2 est donc égal à :

$$\begin{aligned} Q_2 &= x_5 + 0,5 \cdot (x_6 - x_5) \\ &= 5 + 0,5 \cdot (5 - 5) \\ &= 5. \end{aligned}$$

   Le troisième quartile Q_3 se trouve à la position $3 \cdot \frac{n+1}{4} = 8,25$. Q_3 est donc égal à :

$$\begin{aligned} Q_3 &= x_8 + 0,25 \cdot (x_9 - x_8) \\ &= 6 + 0,25 \cdot (7 - 6) \\ &= 6,25. \end{aligned}$$

   Nous pouvons donc dire que les valeurs 3, 5, 5 et 6, 25 partagent l'ensemble des observations en 4 parties sensiblement égales.

2. Pour un deuxième exemple, considérons le **tableau de fréquences** suivant représentant le nombre d'enfants par famille sur un ensemble de 200 familles :

Valeur (nombre d'enfants)	Fréquence (nombre de familles)	Fréquence relative	Fréquence relative cumulée
0	6	0, 03	0, 03
1	38	0, 19	0, 22
2	50	0, 25	0, 47
3	54	0, 27	0, 74
4	42	0, 21	0, 95
5	8	0, 04	0, 99
6	2	0, 01	1, 00
Total	200	1, 00	

Le 1er quartile est égal à l'observation dont la **fréquence** relative cumulée dépasse 25 %, ce qui correspond à 2 enfants (puisque la fréquence relative cumulée pour 2 enfants va de 22 à 47 %, ce qui inclut 25 %).

Le 2e quartile est égal à 3 enfants puisque la fréquence relative cumulée pour 3 enfants va de 47 à 74 %, ce qui inclut 50 %.

Le 3e quartile est égal à 4 enfants puisque la fréquence relative cumulée pour 4 enfants va de 74 à 95 %, ce qui inclut 75 %.

Les quartiles 2, 3 et 4 divisent les 200 familles en quarts. Nous pouvons donc attribuer 50 des 200 familles au premier quart avec 0, 1 ou 2 enfants, 50 au deuxième quart avec 2 ou 3 enfants, 50 au troisième quart avec 3 ou 4 enfants et 50 au quatrième quart avec 4, 5 ou 6 enfants.

3. Considérons à présent un exemple de calcul des quartiles à partir de la **distribution de fréquences** d'une **variable** continue où les observations sont groupées en classes.

   Le tableau de fréquences suivant représente les profits (en milliers d'euros) de cent épiceries :

Profit (milliers d'euros)	Fréquence	Fréquence cumulée	Fréquence relative cumulée
$100-200$	10	10	$0,1$
$200-300$	20	30	$0,3$
$300-400$	40	70	$0,7$
$400-500$	30	100	$1,0$
Total	100		

   La classe comprenant le 1er quartile est la classe $200-300$ (celle dont la fréquence relative cumulée comprend 25 %).

   En considérant que les observations sont distribuées de manière uniforme dans chaque classe, nous obtenons pour le premier quartile la **valeur** suivante :

$$\begin{aligned} 1^{\text{er}} \text{ quartile} &= 200 + \left[\frac{\left(100 \cdot \frac{1}{4}\right) - 10}{20}\right] \cdot 100 \\ &= 275. \end{aligned}$$

   La classe comprenant le 2e quartile est la classe $300-400$. La **valeur** du 2e quartile est égale à

$$\begin{aligned} 2^{\text{e}} \text{ quartile} &= 300 + \left[\frac{\left(100 \cdot \frac{2}{4}\right) - 30}{40}\right] \cdot 100 \\ &= 350. \end{aligned}$$

   La classe comprenant le 3e quartile est la classe $400-500$. La valeur du 3e quartile est égale à

$$\begin{aligned} 3^{\text{e}} \text{ quartile} &= 400 + \left[\frac{\left(100 \cdot \frac{3}{4}\right) - 70}{30}\right] \cdot 100 \\ &= 416,66. \end{aligned}$$

   Nous pouvons donc conclure que 25 des 100 épiceries ont un profit compris entre 100 et 275 milliers d'euros, 25 ont un profit compris entre 275 et 350 milliers d'euros, 25 ont un profit compris entre 350 et 416,66 milliers d'euros, et 25 ont un profit compris entre 416,66 et 500 milliers d'euros.

MOTS CLÉS

Centile *(Percentile)*
Décile *(Decile)*
Médiane *(Median)*
Mesure de position *(Measure of location)*
Quantile *(Quantile)*

QUETELET ADOLPHE

Quetelet Adolphe

Adolphe Quetelet est né en 1796 à Ghent (Belgique). En 1823, le ministre de l'Éducation publique lui demanda de superviser la création d'un observatoire pour la ville de Bruxelles. Pour ce faire, il fut envoyé à Paris afin d'apprendre l'astronomie et de préparer les plans pour la construction et l'équipement de l'observatoire. En France, Alexis Bouvard s'est chargé de sa formation. Il rencontra,

entre autres, **Siméon-Denis Poisson** et **Pierre Simon de Laplace** qui le convainquirent de l'importance du calcul probabiliste. Durant les quatre années qui suivirent, il voyagea beaucoup afin de visiter différents observatoires et de récolter des instruments. Il écrivit des ouvrages de vulgarisation scientifique sur la science : *Astronomie élémentaire*, *Astronomie populaire*, *Positions de physique* et *Instructions populaires sur le calcul des probabilités* en 1828. De plus, il fonda avec Garnier (son directeur de thèse) une revue périodique : *Correspondance mathématique et physique*.



En 1828, il fut nommé astronome de l'Observatoire de Bruxelles, puis rapidement directeur de l'Observatoire, poste qu'il occupa durant les vingt-cinq années qui suivirent.

En 1836, il devint le tuteur des princes Ernest et Albert de Saxe-Cobourg et Gotha, ce qui lui permit de satisfaire son intérêt pour le calcul des probabilités. Ses leçons furent publiées en 1846 sous le titre : *Lettres à S.A.R le Duc régnant de Saxe-Cobourg et Gotha sur la théorie des probabilités appliquées aux sciences morales et politiques.*

Il mourut en 1784.

Ouvrages principaux d'Adolphe Quetelet :

1826 *Astronomie élémentaire.* Malther, Paris.

1828 *Positions de physique.* Bruxelles.

1828 *Instructions populaires sur le calcul des probabilités.*

1846 *Lettres à S.A.R le Duc régnant de Saxe-Cobourg et Gotha sur la théorie des probabilités, appliquées aux sciences morales et politiques*, Bruxelles.



R

RANDOMISATION

Randomization

La randomisation est un procédé objectif par lequel les traitements sont aléatoirement affectés à des unités expérimentales. Elle est utilisée pour éviter des affectations subjectives des traitements. Elle permet d'éliminer une **erreur** systématique produite par des facteurs incontrôlables qui persiste même dans des expériences répétées. La randomisation peut être utilisée pour dissimuler les affectations de traitements aux personnes impliquées dans l'expérience afin d'empêcher les biais personnels qui altèrent les données. Finalement, la randomisation est nécessaire si l'on veut estimer l'erreur expérimentale.

HISTORIQUE

En septembre 1919, **Ronald Aylmer Fisher** commença sa carrière à la Rothamsted Experimental Station, où l'on s'occupait de recherche agronomique depuis 1843. Il a commencé avec des données historiques et, ensuite, il a considéré des données obtenues lors d'essais agricoles, pour lesquels il développa l'**analyse de variance**. Ces différents essais le conduisirent à développer les plans d'expérience. Il développa des plans d'une grande efficacité, en posant notamment les principes de randomisation, de carré latin, de blocs randomisés et d'arrangements factoriels.

Ces méthodes sont réunies dans l'ouvrage *Statistical Methods for Research Workers* (1925). Son article de 1926 sur l'arrangement des expériences agricoles se transforma en un livre : *The Design of Experiments* (1935).

MOTS CLÉS

Expérience *(Experiment)*
Plan complètement randomisé *(Completely randomized design)*
Plan d'expérience *(Design of experiments)*
Traitement *(Treatment)*
Unité expérimentale *(Experimental unit)*

RÉFÉRENCES

Fisher, R.A. (1925). *Statistical Methods for Research Workers.* Oliver & Boyd, Edinburgh.

Fisher, R.A. (1926). The Arrangement of Field Experiments. *Journal of the Ministry of Agriculture*, Vol. XXXIII, pp. 503-513.

Fisher, R.A. (1935). *The Design of Experiments.* Oliver & Boyd, Edinburgh.

Fisher, R.A. (1990). *Statistical Methods, Experimental Design and Scientific Inference.* A re-issue of *Statistical Methods for Research Workers*, *The Design of Experiments*, and *Statistical Methods and Scientific Inference.* Ed. Bennett, J.H. and Yates, F. Oxford : Oxford University Press.

RANG D'UNE MATRICE

Rank of a matrix

Le rang d'une **matrice** est le nombre maximum de lignes ou de colonnes linéairement indépendantes de la matrice.

On dit que deux ou plusieurs lignes ou colonnes sont linéairement indépendantes, si la seule combinaison linéaire de ces dernières égale à 0 est la combinaison linéaire triviale avec tous les coefficients nuls.

ASPECTS MATHÉMATIQUES

Si **A** est une **matrice** de dimension $m \times n$, on définit son rang, noté r_A, comme le nombre maximum de lignes ou de colonnes linéairement indépendantes. En particulier :

$$r_A \leq \min(m, n).$$

Si $r_A = \min(m, n)$, on dit que la matrice **A** est de rang plein.

Si le rang d'une matrice carrée d'ordre n vaut n alors le **déterminant** de cette matrice est non nul, en d'autres termes, cette matrice sera inversible.

EXEMPLE

Soit la matrice $\mathbf{A}$ de dimension 2×3 suivante :

$$\mathbf{A} = \begin{bmatrix} -7 & 3 & 2 \\ 1 & 0 & 4 \end{bmatrix}.$$

Son rang est inférieur ou égal à $\min(2, 3)$. Le nombre de lignes ou de colonnes linéairement indépendantes est égal à 2. Par conséquent, $r_A = 2$.

MOTS CLÉS

Déterminant *(Determinant)*
Matrice *(Matrix)*

RAO CALYAMPUDI RADHAKRISHNA

Rao Calyampudi Radhakrishna



Radhakrishna Rao est né en 1920 à Hadagali, Karnata (Inde). Il reçut le titre de docteur en 1948 à l'Université de Cambridge. Il est docteur honoris causa de plusieurs universités, notamment de l'Université de Neuchâtel (Suisse) depuis 1989.

Sa contribution au développement de la théorie statistique et de ses applications se situe dans le prolongement des œuvres de **Ronald Aylmer Fisher**, Jerzy Neyman et des autres sommités de la statistique moderne. Parmi ses contributions les plus importantes, on cite : le théorème de Fisher-Rao, le théorème de Rao-Blackwell, l'inégalité de Cramer-Rao ainsi que le test U de Rao. Il est le co-auteur de 12 livres et plus de 250 articles scientifiques. Le livre *Statistics and Truth : Putting Chance to Work* traite des fondements logiques et des applications de la statistique et il a été traduit dans plusieurs langues.

Quelques ouvrages et publications de Calyampudi Radhakrishna Rao :

1973 *Linear Statistical Inference and Its Applications.* 2nd ed., John Wiley & Sons, New York.

1989 (and Shanbhag, D.N.) Further extensions of the Choquet-Deny and Deny theorems with applications in characterization theory. *Quarterly Journal of Mathematics*, Oxford 40, pp. 333-350.

1992 (and Zhao, L.C.) Linear representation of R-estimates in linear models. *Canadian Journal of Statistics* 220, pp. 359-368.

1997 *Statistics and Truth : Putting Chance to Work.* 2nd ed., World Scientific Pub Co.

1999 (and Toutenburg, H.) *Linear Models. Least Squares and Alternatives.* 2nd ed. Springer Series in Statistics.

MOTS CLÉS
Inverse généralisé *(Generalized inverse)*
Théorème de Gauss-Markov *(Gauss-Markov theorem)*

RAPPORT DE VRAISEMBLANCE, TEST DU

Likelihood ratio test

Voir **test du rapport de vraisemblance**.

RÉ-ÉCHANTILLONNAGE

Resampling

Lors de l'utilisation d'un **échantillon** pour estimer une **statistique** (concernant la **population** de laquelle on tire l'échantillon), on peut réutiliser le même échantillon pour améliorer ou approfondir l'analyse statistique effectuée, en particulier en tirant des échantillons secondaires à partir de l'échantillon original. Voilà l'idée principale du ré-échantillonnage, qui implique aussi l'idée de **simulation**.

On recherche des informations concernant la fiabilité (**biais, écart type**, intervalles de confiance) d'un **paramètre** estimé ou la véracité d'une **hypothèse** faite sur la population donnée.

Parmi les techniques qui impliquent un ré-échantillonnage, on peut citer le **bootstrap**, le jackknife et la re-randomisation. Leur intérêt principal se trouve dans leur caractère potentiellement non paramétrique, qui se revèle important lorsque :

- on ne connaît pas du tout la distribution sous-jacente aux données ou
- les données proviennent d'une **expérience** non linéaire et difficile à modéliser.

Par contre, un problème lié à ce genre de procédures est que, si l'échantillon n'est pas représentatif, on risque de faire des inférences absolument fausses (puisqu'en quelque sorte on multiplie l'échantillon donné).

HISTORIQUE

L'explosion de ce genre de techniques se fit en coïncidence avec celle des ordinateurs dans les années 70, même si les idées sous-jacentes remontent au début du vingtième siècle. En effet, on peut voir l'estimation de la **variance** d'un **échantillon** par rapport à la **moyenne** comme application de ce principe.

Le jackknife fut introduit dans les années 50 comme moyen pour réduire le **biais** d'une **estimation**. Dans les années 60 et 70, le ré-échantillonnage fut beaucoup utilisé dans l'**échantillonnage** comme méthode d'estimation de la **variance**.

Le **bootstrap** a été découvert par B. Efron à la fin des années 70 et a été ensuite développé par de nombreux chercheurs dans différents domaines.

ASPECTS MATHÉMATIQUES

Voir **bootstrap**, **méthode du jackknife**.

MOTS CLÉS

Bootstrap *(Bootstrap)*
Méthode de Monte-Carlo *(Monte Carlo method)*
Méthode du jackknife *(Jackknife methode)*
Randomisation *(Randomization)*

RÉFÉRENCES

Edington, E.S. (1980). *Randomisation Tests.* Marcel Dekker, New York.

Efron, B. (1982). *The Jackknife, the bootstrap, and other resampling plans*, CBMS Monograph. N° 38. SIAM, Philadelphia.

Fisher, R.A. (1935). *The Design of Experiments.* Oliver & Boyd, Edinburgh.

Pitman, E.J.G. (1937). *Significance tests which may be applied to samples from any populations.* Supplement, *Journal of the Royal Statistical Society*, 4, pp. 119-130.

RECENSEMENT

Census

Le recensement est l'opération consistant à observer la totalité des individus de la **population**. Si le mot recensement provient du recensement de la population, c'est-à-dire le dénombrement des individus formant la population, son sens s'est étendu à toutes les enquêtes (dites exhaustives) où l'on cherche à obtenir des informations en observant tous les objets d'une population. Il est clair que lorsque la taille de la population est élevée, ces enquêtes sont

très coûteuses. C'est pourquoi le recensement est rarement utilisé ; on lui préfère le **sondage** qui consiste à n'observer qu'une partie restreinte de la population appelée **échantillon**.

HISTORIQUE

Le début des recensements se situe à l'époque des grandes civilisations de l'Antiquité, dont l'étendue et la complexité des empires nécessitaient une administration rigoureuse.

Parmi les plus anciennes, on peut citer celle de Sumer. Dès cette époque, c'est-à-dire entre 5000 et 2000 avant J.-C., on procédait au recensement en inscrivant sur des tablettes d'argile, en caractères cunéiformes, la liste des hommes et des biens.

Il en fut de même en Mésopotamie, vers 3000 avant J.-C., et en Égypte où, dès la première dynastie, on effectua des recensements. Ceux-ci furent pratiqués pendant les différentes dynasties qui se succédèrent, avec des objectifs tant militaires que fiscaux. Sous Amasis II, chacun devait, sous peine de mort, déclarer sa profession et ses sources de revenus.

La situation en Israël était plus complexe : les recensements y étaient tantôt prescrits, tantôt interdits car l'Ancien Testament liait au recensement un symbole de malédiction. Ce souvenir pesa d'ailleurs longtemps sur la civilisation chrétienne puisqu'au Moyen-Âge, saint Augustin et saint Ambroise condamnaient encore de telles opérations.

En Chine, en revanche, le recensement était pratiqué depuis très longtemps, sous des formes et dans des buts différents. J. Hecht (1987) en cite les principaux :

1. *Époque des Han (200 ans avant J.-C. à 200 ans après J.-C.) :* les recensements de la **population** sont liés au système de la conscription.

2. *Époque des Trois Royaumes aux Cinq Dynasties (221-959 après J.-C.) :* ils sont liés aux systèmes de distribution des terres.

3. *Époque des Sung et Yuan (960-1368 après J.-C.) :* ils sont effectués selon des objectifs fiscaux.

4. *Époque Ming (1368-1644 après J.-C.) :* des registres, dits « registres jaunes » sont établis au cours de recensements décennaux. Ils indiquent pour chaque personne le nom, la profession, le sexe et l'âge.

5. *Époque Ching (dès 1644 après J.-C.) :* les recensements sont effectués pour surveiller les déplacements de la population.

   À partir de 1755, ce fut l'instauration du système du « pao-chia ». Outre la tenue de registres, il consistait à apposer sur chaque maison les indications concernant ses occupants : nombre, sexe, âge, profession et montant de l'imposition.

Le Japon connut également, tout au long du Moyen-Âge, différents types de recensement. Le premier aurait été effectué sous le règne de l'empereur Soujin (86 avant J.-C.).

Enfin en Inde, un traité de science politique et d'économie intitulé *Arthasàstra* (traité du profit) donne des renseignements sur l'utilité de la tenue de registres extrêmement détaillés. Conçu par Kautilya, alors ministre du roi Candragupta (313-289 avant J.-C.), il préconisait un État centralisateur, qui dirigerait et contrôlerait toutes les activités de ses citoyens.

Une autre civilisation très importante, celle des Incas, connaissait également les recensements. On y pratiquait un système perfectionné de statistiques appelé « quipos ». Le quipo constituait à la fois l'instrument et le registre des renseignements saisis. Formé d'une série de cordelettes dont les couleurs, les combinaisons et les nœuds revêtaient des significations précises, le quipo était confié à des gardiens initiés qui rassemblaient la totalité des statistiques.

En Europe, Grecs et Romains pratiquaient les recensements depuis très longtemps. Aristote rapporte d'ailleurs qu'en Grèce, on offrait à la déesse Athéna une mesure de froment par naissance et une mesure d'orge par décès. À Rome, le premier recensement aurait été réalisé à l'instigation du roi Servius Tullius (578-534 avant J.-C.) pour classer les citoyens en fonction de leurs revenus et lever des impôts en conséquence. On peut à ce propos noter

l'origine du mot recensement (en anglais *census*) dérivé du latin *cens*. En 89 avant J.-C., une loi obligeait les citoyens à se soumettre au recensement sous peine de se voir retirer leurs biens ou leurs droits civiques.

Plus tard, selon les pays, la pratique des recensements s'instaura à des périodes et à des échelles différentes.

En 786, Charlemagne fit dénombrer ses sujets âgés de plus de douze ans ; au XII^{e} siècle, l'Italie commença également à compter ses habitants ; au XV^{e} siècle, ce furent de nombreuses villes qui procédèrent au recensement de leurs habitants : Nuremberg en 1449, ou Strasbourg en 1470. Toutefois ce fut au XVI^{e} siècle que se constituèrent en France de véritables registres d'État civil.

Le XVII^{e} siècle vit se développer trois courants différents : l'école allemande qui se vouait à la **statistique descriptive**, l'école française attachée à l'idéologie et à la méthodologie du recensement et l'école anglaise qui sera le précurseur de la statistique moderne.

Cependant, dans l'histoire des recensements en Europe, il est un pays qui occupa une place particulière. En effet, très en avance sur leurs contemporains, les Suédois mirent en place dès 1665 des registres tenus par les pasteurs chargés de noter le nombre de leurs paroissiens ; en 1668, un décret rendit ces dénombrements obligatoires et, de religieux, les registres devinrent administratifs. 1736 vit la parution d'un autre décret, stipulant que les gouverneurs de chaque province devaient rendre compte à chaque assemblée parlementaire des variations de la population dans leur province. La statistique suédoise fut officiellement créée le 3 février 1748 par le « Tabellverket » (Administration des tableaux). Le premier résumé de toutes les provinces en 1749 peut être considéré comme le premier recensement en Europe et le 11 novembre 1756 marqua la création du « Tabellkommissionen », premier organisme officiel de statistique. Dès 1762, des tableaux de chiffres figuraient dans le recueil de l'Académie des sciences.

Les recensements furent d'abord organisés chaque année (1749-1751), puis tous les trois ans (1754-1772) et enfin tous les cinq ans (dès 1775).

La fin du XVIII^{e} siècle voyait donc en Occident l'instauration officielle du recensement avec en France, en 1791, la promulgation d'une loi qui en instaurait le principe et aux États-Unis, en 1787, son inscription dans la Charte constitutionnelle.

Voir aussi **statistique officielle**.

MOTS CLÉS

Collecte des données *(Data collection)*
Démographie *(Demography)*
Échantillon *(Sample)*
Population *(Population)*
Sondage *(Survey)*
Statistique officielle *(Official Statistics)*

RÉFÉRENCE

Hecht, J. (1978). L'idée du dénombrement jusqu'à la Révolution. *Pour une histoire de la statistique*, tome 1, pp. 21-82 . Economica/INSEE.

RECHERCHE OPÉRATIONNELLE

Operations research

La recherche opérationnelle est une branche des mathématiques appliquées qui utilise des méthodes scientifiques dans le but de fournir une base nécessaire à la prise de décisions.

Elle est généralement appliquée à des problèmes complexes d'organisation de personnes et d'équipement en vue de trouver la meilleure solution pour atteindre un objectif fixé. Parmi ces méthodes scientifiques, on trouve la **simulation**, la **programmation linéaire**, la programmation mathématique (non linéaire), la théorie des jeux, etc.

HISTORIQUE

L'origine de la recherche opérationnelle est à chercher dans les applications militaires du Royaume-Uni durant les années 30.

Cette application militaire s'est ensuite largement étendue et, depuis la Deuxième Guerre mondiale, la recherche opérationnelle a vu son usage s'élargir à des problèmes publics non militaires et même à des problèmes d'entreprises privées.

L'impact sur la recherche opérationnelle du développement rapide des ordinateurs s'est révélé très important.

EXEMPLES

L'**optimisation**, la **programmation linéaire** et la **simulation** sont des exemples de recherche opérationnelle.

MOTS CLÉS

Optimisation *(Optimization)*
Programmation linéaire *(Linear programming)*
Simulation *(Simulation)*

RÉFÉRENCES

Bealle, E.M.L. (1988). *Introduction to Optimization.* John Wiley & Sons, New York.

Collatz, L. and Wetterling, W. (1975). *Optimization problems.* Springer-Verlag, New York.

RÉGION D'ACCEPTATION

Acceptance region

On appelle région d'acceptation l'**intervalle**, mesuré sur la **distribution d'échantillonnage** de la **statistique** étudiée, qui conduit à accepter l'**hypothèse nulle** H_0 d'un **test d'hypothèse**.

C'est la région complémentaire à la **région de rejet**.

La région d'acceptation est associée à une **probabilité** $1 - \alpha$, α étant le **seuil de signification** du test.

ASPECTS MATHÉMATIQUES

Voir **région de rejet**.

EXEMPLES

Voir **région de rejet**.

MOTS CLÉS

Région de rejet *(Rejection region)*
Seuil de signification *(Significance level)*
Test d'hypothèse *(Hypothesis testing)*
Valeur critique *(Critical value)*

RÉGION DE REJET

Rejection region

La région de rejet est l'**intervalle**, mesuré sur la **distribution d'échantillonnage** de la **statistique** étudiée, qui conduit à rejeter l'**hypothèse nulle** H_0 d'un **test d'hypothèse**.

La région de rejet est **complémentaire** à la **région d'acceptation**.

La région de rejet est associée à une **probabilité** α, appelée **seuil de signification** du test ou **erreur de première espèce**.

ASPECTS MATHÉMATIQUES

Considérons un **test d'hypothèse** sur un **paramètre** θ. Nous utilisons la **statistique** T en vue d'estimer ce paramètre.

Cas 1 : test unilatéral à droite

Les hypothèses sont les suivantes :

$$\begin{array}{lcl} H_0 & : & \theta = \theta_0 \\ H_1 & : & \theta > \theta_0 \end{array}$$

où θ_0 est la valeur présumée du paramètre θ.

Dans un **test unilatéral** à droite, la région de rejet correspond à un **intervalle** borné à gauche par la **valeur** de la **valeur critique** :

région de rejet= [valeur critique ; ∞[, et
région d'acceptation=]$-\infty$; valeur critique[,

la valeur critique étant égale à :

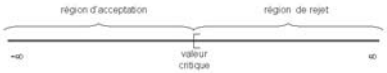
$$\mu_T + z_\alpha \cdot \sigma_T$$

où

μ_T	est la **moyenne** de la **distribution d'échantillonnage** de la statistique T,
σ_T	est l'**erreur type** de la statistique T et
α	est le **seuil de signification** du test.

La valeur de z_α peut être obtenue en consultant la **table statistique** relative à la distribution de T.

Si la valeur de T calculée sur l'**échantillon** est comprise dans la région de rejet, donc si T est supérieur ou égal à la valeur critique, l'**hypothèse nulle** H_0 doit être rejetée au profit de l'**hypothèse alternative** H_1. Dans le cas contraire, l'hypothèse nulle H_0 doit être acceptée.



Cas 2 : test unilatéral à gauche

Les hypothèses sont les suivantes :

$$\begin{array}{lcl} H_0 & : & \theta = \theta_0 \\ H_1 & : & \theta < \theta_0 \end{array}$$

Dans un test unilatéral à gauche, la région de rejet est bornée à droite par la valeur de la valeur critique :

$$\text{région de rejet} =]-\infty\,;\,\text{valeur critique}],$$

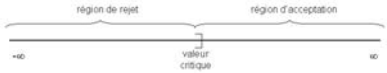
et la région d'acceptation est son complémentaire :

$$\text{région d'acceptation} =]\text{valeur critique}\,;\,\infty[,$$

la valeur critique étant égal à :

$$\mu_T - z_\alpha \cdot \sigma_T.$$

L'hypothèse nulle H_0 sera rejetée au profit de l'hypothèse alternative H_1 si la valeur calculée de T sur l'**échantillon** est comprise dans la région de rejet, donc si T est inférieur ou égal à la valeur critique. Dans le cas contraire, l'hypothèse nulle sera acceptée.



Cas 3 : test bilatéral

Les hypothèses sont les suivantes :

$$\begin{array}{lcl} H_0 & : & \theta = \theta_0 \\ H_1 & : & \theta \neq \theta_0 \end{array}$$

Dans un **test bilatéral**, la région de rejet est divisée en deux intervalles :

$$\begin{array}{rcl} \text{région de rejet} & = & \mathbb{R}\backslash]\text{valeur critique inférieure} \\ & & \text{valeur critique supérieure}[\end{array}$$

et la région d'acceptation est son complémentaire :

$$\begin{array}{rcl} \text{région d'acceptation} & = &]\text{valeur critique inférieure} \\ & & \text{valeur critique supérieure}[, \end{array}$$

où les valeurs critiques sont respectivement :

$$\mu_T - z_{\frac{\alpha}{2}} \cdot \sigma_T \text{ et } \mu_T + z_{\frac{\alpha}{2}} \cdot \sigma_T.$$

Si la valeur calculée de la **statistique** T est comprise dans la région de rejet, donc si T est inférieur ou égal à la valeur critique inférieure, ou supérieur ou égal à la valeur critique supérieure, l'hypothèse nulle H_0 sera rejetée au profit de l'hypothèse alternative H_1. Dans le cas contraire, l'hypothèse nulle sera acceptée.



La figure suivante illustre graphiquement les régions d'acceptation et de rejet pour les cas mentionnés ci-dessus.

DOMAINES ET LIMITATIONS

Dans un **test unilatéral**, la région de rejet est unique et se situe :

- à droite de la **distribution d'échantillonnage** pour un test unilatéral à droite ;
- à gauche de la distribution d'échantillonnage pour un test unilatéral à gauche.

Par contre, la région de rejet est divisée en deux parties dans un **test bilatéral**, se situant aux deux extrémités de la distribution d'échantillonnage, symétriquement par rapport à la **moyenne** de la distribution. À chacune de ces régions est alors associée une **probabilité** $\frac{\alpha}{2}$, α étant le **seuil de signification** du test.

EXEMPLE

Une compagnie fabrique des câbles d'acier. Elle veut vérifier sur un **échantillon** de taille $n = 100$ si le diamètre des câbles est conforme aux normes qui imposent un diamètre de $0,9$ cm.

L'**écart type** de la **population** est connu et vaut $0,05$ cm.

Le **test d'hypothèse** est dans ce cas un **test bilatéral**. Les hypothèses sont les suivantes :

$$\begin{array}{rrcl} \textbf{hypothèse nulle} & H_0 & : & \mu = 0,9 \\ \textbf{hypothèse alternative} & H_1 & : & \mu \neq 0,9 \end{array}$$

Avec un **seuil de signification** de $\alpha = 5$ %, la valeur $z_{\frac{\alpha}{2}}$ de la **table normale** est égale à $1,96$. Les valeurs critiques de l'hypothèse nulle sont :

$$\mu_{\bar{x}} \pm z_{\frac{\alpha}{2}} \cdot \sigma_{\bar{x}}$$

où $\mu_{\bar{x}}$ est la **moyenne** de la **distribution d'échantillonnage** des moyennes :

$$\mu_{\bar{x}} = \mu = 0,9,$$

et $\sigma_{\bar{x}}$ est l'**erreur type** de la moyenne :

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{0,05}{\sqrt{100}} = 0,005$$

$$\begin{array}{rcl} \Rightarrow \mu_{\bar{x}} \pm z_{\frac{\alpha}{2}} \sigma_{\bar{x}} & = & 0,9 \pm 1,96 \cdot 0,005 \\ & = & 0,9 \pm 0,0098 \end{array}$$

L'**intervalle** $]0,8902\,;\,0,9098[$ correspond à la **région d'acceptation** de l'hypothèse nulle.

La région de rejet est l'intervalle complémentaire à la région d'acceptation. Elle est donc divisée en deux parties :

$$]-\infty\,;\,0,8902] \text{ et } [0,9098\,;\,\infty[\,.$$

Si la moyenne échantillonnale obtenue est comprise dans ces intervalles, nous devrons rejeter l'hypothèse nulle au profit de l'hypothèse alternative et conclure qu'à un seuil de signification de 5 %, le diamètre des câbles n'est pas conforme aux normes.

MOTS CLÉS

Région d'acceptation *(Acceptance region)*
Seuil de signification *(Significance level)*
Test d'hypothèse *(Hypothesis testing)*
Valeur critique *(Critical value)*

RÉGRESSION, ANALYSE DE

Regression analysis

Voir **analyse de régression**.

RÉGRESSION LAD

Least Absolute Deviation regression

La méthode des moindres écarts en valeur absolue, dite la méthode LAD (*Least Absolute Deviation*), est l'une des principales alternatives à la **méthode des moindres carrés** lorsqu'il s'agit d'estimer les paramètres d'un modèle de régression. Le but de cette régression est de fournir un **estimateur robuste**.

HISTORIQUE

Elle a été introduite presque cinquante ans avant la **méthode des moindres carrés**, en 1757 par **Roger Joseph Boscovich**. Il utilisa cette procédure dans le but de concilier des mesures incohérentes dans le cadre de l'estimation de la forme de la Terre. **Pierre Simon de Laplace** adopta cette méthode trente ans plus tard, mais elle fut ensuite éclipsée par la méthode des moindres carrés développée par **Adrien-Marie Legendre** et **Carl Friedrich Gauss** et dont la popularité repose principalement sur la simplicité des calculs. Mais aujourd'hui, avec les progrès de l'informatique, la méthode LAD peut être utilisée presque aussi simplement.

ASPECTS MATHÉMATIQUES

Nous ne traitons ici que le cas de régression LAD simple.

Considérons le modèle de **régression** simple $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$. Le critère utilisé par la méthode LAD est de choisir $\widehat{\beta}_0$ et $\widehat{\beta}_1$ qui minimisent la somme des valeurs absolues des résidus du modèle, c'est-à-dire

$$\sum |e_i| = \sum \left| y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i \right|$$

où e_i est le i-ème résidu. La particularité de cette méthode est qu'il n'existe pas de formule explicite pour calculer $\widehat{\beta}_0$ et $\widehat{\beta}_1$, les estimateurs LAD peuvent toutefois être calculés en appliquant un **algorithme** itératif.

Algorithme LAD simple

Le point essentiel utilisé par cet algorithme est le fait que la droite de régression LAD (qui minimise la quantité $\sum |e_i|$) passe toujours par (au moins) deux points des données.

Pour résumer la méthode : choisissons un point des données, disons (x_1, y_1). Cherchons alors la meilleure droite qui passe par ce point (pour la méthode permettant de construire cette droite, voir ci-dessous). Cette droite passe au moins par un second point, disons par (x_2, y_2).

L'étape suivante est de chercher la meilleure droite, au sens de la somme $\sum |e_i|$, passant cette fois par (x_2, y_2). Elle passe alors au moins par un autre point, disons par (x_3, y_3).

Nous continuons l'itération en cherchant la meilleure droite passant par (x_3, y_3) et ainsi de suite. Au fur et à mesure des itérations, la quantité $\sum |e_i|$ diminue. Si, lors d'une itération, la droite obtenue est identique à celle de l'étape précédente, nous pouvons conclure qu'il s'agit de la meilleure droite au sens de cette méthode et ainsi que nous avons trouvé la droite LAD.

Si les cas de non-unicité de la droite LAD et de dégénérescence (lorsque la droite LAD passe par plus de deux points de données, voir DOMAINES ET LIMITATIONS) sont ignorés, la meilleure droite est unique et passe exactement par deux points lors de chaque itération et de ce fait l'algorithme converge.

Construction de la meilleure droite passant par (x_1, y_1) au sens de la somme $\sum |e_i|$

Comme cette droite passe par un autre point des données, il s'agit de trouver le point (x_k, y_k) pour lequel la droite

$$\widehat{y}(x) = y_1 + \frac{y_k - y_1}{x_k - x_1}(x - x_1),$$

c'est-à-dire la droite de pente

$$\widehat{\beta}_1 = \frac{y_k - y_1}{x_k - x_1}$$

et d'ordonnée à l'origine

$$\widehat{\beta}_0 = y_1 - \widehat{\beta}_1 x_1,$$

est la meilleure au sens $\sum |e_i|$. Pour se faire, renommons les $(n-1)$ points candidats (x_2, y_2), …, (x_n, y_n) de telle sorte que

$$\frac{y_2 - y_1}{x_2 - x_1} \leq \frac{y_3 - y_1}{x_3 - x_1} \leq \ldots \leq \frac{y_n - y_1}{x_n - x_1}.$$

Définissons ensuite $T = \sum_{i=1}^{n} |x_i - x_1|$. Le point (x_k, y_k) recherché est alors déterminé par l'indice k pour lequel

$$\begin{cases} |x_2 - x_1| + \ldots + |x_{k-1} - x_1| < T/2 \\ |x_2 - x_1| + \ldots + |x_{k-1} - x_1| + |x_k - x_1| > T/2. \end{cases}$$

Cette condition nous assure que le $\widehat{\beta}_1$ ainsi choisi minimise la quantité

$$\sum_{i=1}^{n} \left| (y_i - y_1) - \widehat{\beta}_1 (x_i - x_1) \right|$$

qui est l'analogue de la $\sum |e_i|$ pour les droites qui

passent par (x_1, y_1). À partir de là, l'ordonnée à l'origine $\widehat{\beta}_0$ est calculée de telle sorte que la droite passe par le point (x_1, y_1). Nous pouvons vérifier qu'elle passe également par (x_k, y_k). Il suffit alors de renommer le point (x_k, y_k) par (x_2, y_2) et de recommencer.

DOMAINES ET LIMITATIONS

Deux cas particuliers peuvent toutefois se présenter en utilisant la régression LAD : le cas de *non-unicité* et le cas de *dégénérescence*. La non-unicité signifie que la droite LAD n'est pas unique, c'est-à-dire qu'il y a au moins deux droites qui minimisent la somme des valeurs absolues des résidus, alors que la dégénérescence signifie que la droite LAD passe par plus de deux points des données.

EXEMPLE

Prenons l'exemple concernant 14 pays d'Amérique du Nord et d'Amérique centrale. Le tableau suivant contient les données s'y rapportant et indique, pour chaque pays, le taux de natalité (nombre de naissances par année pour 1 000 personnes) ainsi que le taux d'urbanisation (pourcentage de la population vivant dans des villes de plus de 100 000 habitants) en 1980.

Taux d'urbanisation et taux de natalité

Obs. i	Pays	Taux d'urbanisation x_i	Taux de natalité y_i
1	Canada	55,0	16,2
2	Costa Rica	27,3	30,5
3	Cuba	33,3	16,9
4	États-Unis	56,5	16,0
5	El Salvador	11,5	40,2
6	Guatemala	14,2	38,4
7	Haïti	13,9	41,3
8	Honduras	19,0	43,9
9	Jamaïque	33,1	28,3
10	Mexique	43,2	33,9
11	Nicaragua	28,5	44,2
12	Trinité-et-Tobago	6,8	24,6
13	Panama	37,7	28,0
14	Rép. Dominicaine	37,1	33,1

Source : Birkes & Dodge (1993)

Essayons d'estimer ici, en utilisant la régression LAD, le taux de natalité y_i en fonction du taux d'urbanisation x_i, c'est-à-dire par un modèle de régression simple

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i.$$

Afin d'appliquer l'algorithme LAD pour estimer β_0 et β_1, il s'agit dans un premier temps de chercher la meilleure droite passant par un point donné, le Canada (choisi arbitrairement).

Ainsi $(x_1, y_1) = (55,0\,;\,16,2)$. Calculons alors les pentes $\frac{y_i - 16,2}{x_i - 55,0}$ pour les 13 pays restants. Par exemple, pour le Costa-Rica, le résultat obtenu est $\frac{30,5-16,2}{27,3-55,0} = -0,5162$. Les 13 pays sont ainsi classés par ordre croissant de ces pentes, comme nous pouvons le constater dans le tableau ci-dessous. Ainsi le point (x_2, y_2) est le Mexique, le point (x_3, y_3) est le Nicaragua, et ainsi de suite. Calculons ensuite la quantité $T = \sum |x_i - 55,0| = 355,9$ et donc $\frac{T}{2} = 177,95$. Il s'agit alors de déterminer le pays k qui satisfait la condition

$$\begin{cases} |x_2 - x_1| + \ldots + |x_{k-1} - x_1| & < & T/2 \\ |x_2 - x_1| + \ldots + |x_{k-1} - x_1| + |x_k - x_1| & > & T/2 \end{cases}$$

Comme

$$\sum_{j=2}^{8} |x_j - x_1| = 172,5 < T/2$$

et

$$\sum_{j=2}^{9} |x_j - x_1| = 216,0 > T/2$$

nous en concluons que $k = 9$, et donc que la meilleure droite passant par le Canada (x_1, y_1) passe aussi par le Salvador (x_9, y_9). Ainsi

$$\begin{aligned} \widehat{\beta}_1 &= \tfrac{y_9 - y_1}{x_9 - x_1} = \tfrac{40,2-16,2}{11,5-55,0} = -0,552 \\ \widehat{\beta}_0 &= y_1 - \widehat{\beta}_1 x_1 = 16,2 - (-0,552) \cdot 55,0 = 46,54 \end{aligned}$$

et la droite recherchée est :

$$\widehat{y}(x) = 46,54 - 0,552x.$$

Calculs effectués pour trouver la régression LAD

Obs. i	Pays	$\frac{y_i-y_1}{x_i-x_1}$	$\lvert x_i - x_1\rvert$	$\sum_{j=2}^{i} \lvert x_j - x_1\rvert$
2	Mexique	$-1,5000$	$11,8$	$11,8$
3	Nicaragua	$-1,0566$	$26,5$	$38,3$
4	Rép. Dom.	$-0,9441$	$17,9$	$56,2$
5	Honduras	$-0,7694$	$36,0$	$92,2$
6	Panama	$-0,6821$	$17,3$	$109,5$
7	Haïti	$-0,6107$	$41,1$	$150,6$
8	Jamaïque	$-0,5525$	$21,9$	$172,5$
9	Salvador	$-0,5517$	$43,5$	$216,0$
10	Guatemala	$-0,5447$	$40,8$	$256,8$
11	Costa Rica	$-0,5162$	$27,7$	$284,5$
12	Trinité-et-Tobago	$-0,1743$	$48,2$	$332,7$
13	États-Unis	$-0,1333$	$1,5$	$334,2$
14	Cuba	$-0,0323$	$21,7$	$355,9$

Afin d'améliorer cette première estimation de la droite LAD, l'étape suivante consiste à trouver la meilleure droite qui passe par le Salvador. Renommons ce point $(x_1, y_1) = (11,5\,;\,40,2)$, classons les 13 autres pays par ordre croissant des pentes $\frac{y_i-11,5}{x_i-40,2}$ et, de manière analogue à ce qui est fait ci-dessus, nous obtenons que la meilleure droite qui passe par le Salvador passe aussi par les États-Unis. Alors :

$$\begin{aligned}\widehat{\beta}_1 &= \tfrac{40,2-16,0}{11,5-56,5} = -0,538\\ \widehat{\beta}_0 &= 40,2-(-0,538)\cdot 11,5 = 46,38\end{aligned}$$

et la droite recherchée est

$$\widehat{y}(x) = 46,38 - 0,538x.$$

La troisième étape consiste à trouver la meilleure droite qui passe par les États-Unis. Comme nous trouvons, de manière analogue à ce qui est fait ci-dessus, que cette droite passe également par le Salvador, notre algorithme s'arrête et la droite $\widehat{y}(x) = 46,38 - 0,538x$ est la droite LAD de notre problème.

En comparaison, notons que la droite calculée par la méthode des moindres carrés pour ce problème est donnée par

$$\widehat{y}(x) = 42,99 - 0,399x.$$

Nous constatons avec cette droite que Cuba, Nicaragua et Trinité-et-Tobago ont des résidus anormalement élevés. En supprimant ces trois points et en recalculant la droite des moindres carrés avec les 11 pays restants, nous obtenons :

$$\widehat{y}\,(x) = 48,52 - 0,528x$$

qui est très proche de la droite LAD.

MOTS CLÉS

Analyse de régression *(Regression analysis)*
Analyse de variance *(Analysis of variance)*
Estimation robuste *(Robust estimation)*
Moindres carrés *(Least squares)*
Régression linéaire multiple *(Muliple linear regression)*
Régression linéaire simple *(Simple linear regression)*
Résidus *(Residual)*

RÉFÉRENCE

Birkes, D. and Dodge, Y. (1993). *Alternative methods of Regression.* John Wiley & Sons, New York.

RÉGRESSION LINÉAIRE GÉNÉRALISÉE

Generalized Linear Regression

La régression linéaire généralisée permet de considérer des modèles de régression linéaire qui infirment certaines hypothèses sous-jacentes aux méthodes usuelles d'estimation de ces modèles et qui rendent souvent biaisé le résultat de l'application des **moindres carrées** (ordinaires).

En particulier on prend en considération :

- *l'hétéroscédasticité :* les erreurs observées dans un modèle (appelées aussi résidus) peuvent avoir des variances différentes parmi les observations (ou entre différents groupes d'observations) ;
- *l'autocorrélation :* il peut y avoir de la **corrélation** entre les erreurs de différentes observations.

ASPECTS MATHÉMATIQUES

On considère un modèle général de **régression linéaire multiple** :

$$Y_i = \beta_0 + \sum_{j=1}^{p-1} \beta_j X_{ij} + \varepsilon_i, \qquad i = 1, \ldots, n$$

où

Y_i	est la **variable dépendante**,
X_{ji},	$j = 1, \ldots, p-1$, sont les variables indépendantes,
β_j,	$j = 0, \ldots, p-1$, sont les paramètres à estimer et
ε_i	est le terme d'**erreur** aléatoire non observable.

Sous forme matricielle, il s'écrit :

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}.$$

où

$\mathbf{Y}$	est le vecteur $(n \times 1)$ des observations relatives à la **variable dépendante** (n observations),
$\boldsymbol{\beta}$	est le vecteur $(p \times 1)$ des paramètres à estimer,
$\boldsymbol{\varepsilon}$	est le vecteur $(n \times 1)$ des erreurs

et $\mathbf{X} = \begin{pmatrix} 1 & X_{11} & \ldots & X_{1(p-1)} \\ \vdots & \vdots & & \vdots \\ 1 & X_{n1} & \ldots & X_{n(p-1)} \end{pmatrix}$ est la **matrice** $(n \times p)$ ayant trait aux variables indépendantes.

Dans un modèle ordinaire de régression, les hypothèses sur les erreurs ε_i sont généralement les suivantes :

- leur **espérance** est égale à zéro (ce qui revient à dire que les observations ne sont pas biaisées) ;
- elles sont indépendantes et donc non corrélées ;
- leur **variance** $V(\varepsilon_i)$ est constante et égale à σ^2 (on suppose que les erreurs sont de même « grandeur »).

Si le modèle choisi est adéquat, la distribution des résidus doit confirmer ces hypothèses (voir **analyse de variance**).

Si ce n'est pas le cas ou si on a des informations supplémentaires qui indiquent que les erreurs sont corrélées, alors on peut se référer, pour une estimation plus précise des paramètres β_0, ..., β_{p-1}, à une régression linéaire généralisée. On introduit alors la **matrice** $\mathbf{V}(\boldsymbol{\varepsilon})$ des variance-covariance (de dimension $n \times n$) des erreurs, qui, à la ligne i et à la colonne j, est définie par :

- $V(\varepsilon_i)$ si $i = j$ (variance en ε_i),
- $Cov(\varepsilon_i, \varepsilon_j)$ (**covariance** entre ε_i et ε_j) si $i \neq j$.

Souvent cette matrice est à estimer, mais on considère d'abord le cas plus simple où $\mathbf{V}(\boldsymbol{\varepsilon})$ est connue. On met en évidence un terme commun de variance σ^2 : $\mathbf{V}(\boldsymbol{\varepsilon}) = \sigma^2 \mathbf{V}$.

Estimation du vecteur β lorsque V est connue

En transformant notre modèle du départ dans un modèle ordinaire de régression, il est possible d'estimer les paramètres du modèle généralisé par :

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-1} \mathbf{X}'\mathbf{V}^{-1}\mathbf{Y}.$$

On peut démontrer que :

1. l'estimation est non biaisée, c'est-à-dire $\mathbf{E}(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$,
2. l'égalité suivante est vérifiée :

$$\mathbf{V}\left(\hat{\boldsymbol{\beta}}\right) = \hat{\sigma}^2 \left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-1},$$

avec $\hat{\sigma} = \dfrac{1}{n-p}\, \boldsymbol{\varepsilon}'\mathbf{V}^{-1}\, \boldsymbol{\varepsilon}$.

3. $\hat{\boldsymbol{\beta}}$ est le meilleur (dans le sens qu'il a la variance plus petite) estimateur non biaisé de $\hat{\boldsymbol{\beta}}$; ce résultat est connu comme le théorème généralisé de Gauss-Markov.

Dans le cas où $\mathbf{V}(\boldsymbol{\varepsilon})$ est inconnue.

Estimation du vecteur β lorsque V est inconnue

Lorsque la matrice de variance-covariance n'est pas connue, le modèle est appelé de régression généralisée réalisable. Le premier pas consiste à exprimer $\mathbf{V}$ en fonction d'un paramètre θ, cela permet d'estimer $\mathbf{V}$ par $\hat{\mathbf{V}} = \mathbf{V}(\hat{\theta})$. L'estimation des paramètres est obtenue ensuite par substitution de $\mathbf{V}$ par $\hat{\mathbf{V}}$ dans la formule donnée auparavant :

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})^{-1}\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{Y}.$$

Estimation de V dans trois cas typiques

a) Hétéroscédasticité

On suppose que les observations sont de variance $\frac{\sigma^2}{w_i}$ où w_i est le poids et peut être différent pour chaque $i = 1, \ldots, n$. Alors :

$$\mathbf{V} = \sigma^2 \begin{pmatrix} \frac{1}{w_1} & 0 & \ldots & 0 \\ 0 & \frac{1}{w_2} & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \frac{1}{w_n} \end{pmatrix}.$$

En posant

$$\begin{aligned} \mathbf{Y}_w &= \mathbf{W}\mathbf{Y} \\ \mathbf{X}_w &= \mathbf{W}\mathbf{X} \\ \boldsymbol{\varepsilon}_w &= \mathbf{W}\boldsymbol{\varepsilon} \end{aligned}$$

où $\mathbf{W}$ est la matrice carrée d'ordre n suivante :

$$\mathbf{W} = \begin{pmatrix} \sqrt{w_1} & 0 & \ldots & 0 \\ 0 & \sqrt{w_2} & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \sqrt{w_n} \end{pmatrix}$$

telle que $\mathbf{W}'\mathbf{W} = \mathbf{V}^{-1}$, on obtient le modèle équivalent :

$$V(\boldsymbol{\varepsilon}_w) = V(\mathbf{W}\boldsymbol{\varepsilon}) = \mathbf{W}V(\boldsymbol{\varepsilon})\mathbf{W}' = \sigma^2\mathbf{W}\mathbf{V}\mathbf{W} = \sigma^2\mathbf{I}_n$$

où $\mathbf{I}_n$ est la matrice identité de dimension n. Comme pour ce nouveau modèle la variance des erreurs est constante, on peut appliquer la méthode des moindres carrés et on obtient le vecteur des estimateurs :

$$\begin{aligned} \hat{\boldsymbol{\beta}}_w &= (\mathbf{X}'_w\mathbf{X}_w)^{-1}\mathbf{X}'_w\mathbf{Y}_w \\ &= (\mathbf{X}'\mathbf{W}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}'\mathbf{W}\mathbf{Y} \\ &= \left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{Y} \end{aligned}$$

ainsi que le vecteur des valeurs estimées pour $\mathbf{Y} = \mathbf{W}^{-1}\mathbf{Y}_w$ en posant :

$$\begin{aligned} \hat{\mathbf{Y}} &= \mathbf{W}^{-1}\hat{\mathbf{Y}}_w \\ &= \mathbf{W}^{-1}\mathbf{W}\mathbf{X}\left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{Y} \\ &= \mathbf{X}\hat{\boldsymbol{\beta}}_w. \end{aligned}$$

b) Hétéroscédasticité par groupes

On suppose que les observations sont en trois groupes (de tailles n_1, n_2, n_3) et de variance respective σ_1^2, σ_2^2 et σ_3^2. Alors :

$$\mathbf{V} = \begin{pmatrix} \sigma_1^2\mathbf{I}_{n_1} & \mathbf{O}_{\mathbf{n_2}\times\mathbf{n_1}} & \mathbf{O}_{\mathbf{n_3}\times\mathbf{n_1}} \\ \mathbf{O}_{\mathbf{n_1}\times\mathbf{n_2}} & \sigma_2^2\mathbf{I}_{n_2} & \mathbf{O}_{\mathbf{n_3}\times\mathbf{n_2}} \\ \mathbf{O}_{\mathbf{n_1}\times\mathbf{n_3}} & \mathbf{O}_{\mathbf{n_2}\times\mathbf{n_3}} & \sigma_3^2\mathbf{I}_{n_3} \end{pmatrix}$$

où $\mathbf{I}_{n_j}$ désigne la **matrice identité** d'ordre n_j et $\mathbf{O}_{\mathbf{n_i}\times\mathbf{n_j}}$ la matrice nulle d'ordre $n_i \times n_j$.

Les variances σ_1^2, σ_2^2, σ_3^2 peuvent être estimées en effectuant des régressions ordinaires sur les différents groupes d'observations et en estimant les σ^2 par S^2 (voir **régression linéaire multiple**).

c) Autocorrélation d'ordre un

On suppose qu'il y a de la corrélation entre les erreurs des différentes observations. En particulier, on suppose que l'autocorrélation soit d'ordre un, c'est-à-dire que l'erreur ε_i à l'observation i dépend de l'erreur ε_{i-1} à l'observation précédente (ce qui est raisonable si, par exemple, les observations sont temporelles).

On a donc :

$$\varepsilon_i = \rho\varepsilon_{i-1} + u_i, \quad i = 2, \ldots, n$$

où les hypothèses habituelles sur les erreurs d'un modèle de régression s'appliquent ici aux u_i.

Dans ce cas, la matrice de variance-covariance des erreurs peut s'exprimer par :

$$\mathbf{V}(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{V}$$

$$= \sigma^2 \frac{1}{1-\rho^2} \begin{pmatrix} 1 & \rho & \rho^2 & \dots & \rho^{n-1} \\ \rho & 1 & \rho & \dots & \rho^{n-2} \\ \vdots & \rho & \ddots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & \rho \\ \rho^{n-1} & \rho^{n-2} & \dots & \rho & 1 \end{pmatrix}$$

On peut estimer $\mathbf{V}$ en remplaçant l'estimation suivante de ρ dans la matrice $\mathbf{V}$:

$$\hat{\rho} = \frac{\sum_{i=2}^{n} e_i e_{i-1}}{\sum_{i=2}^{n} e_{i-1}^2} \text{ (corrélation empirique)}$$

où les e_i sont les erreurs estimées (résidus) avec une régression ordinaire.

On peut tester l'autocorrélation d'ordre un par le test de Durbin-Watson sur les résidus, qui est décrit sous le mot **corrélation sérielle**.

EXEMPLE

Correction de l'hétéroscédasticité lorsque les données représentent une moyenne.

Afin de tester l'efficacité d'un nouveau fortifiant sur la croissance des poulets, un éleveur a retenu quarante poulets répartis en cinq groupes de taille inégale n_1, n_2, ..., n_5 et leur a administré des doses différentes de fortifiant. Les données sont résumées dans le tableau suivant :

Groupe i	Nombre de poulets n_i	Poids moyen Y_i	Dose moyenne X_i
1	12	1,7	5,8
2	8	1,9	6,4
3	6	1,2	4,8
4	9	2,3	6,9
5	5	1,8	6,2

Puisque les variables représentent des moyennes, la variance de l'erreur ε_i est égale à $\frac{\sigma^2}{n_i}$.

La matrice de variance-covariance des erreurs est donc de la forme :

$$\mathbf{V}(\boldsymbol{\varepsilon}) = \sigma^2 \begin{pmatrix} \frac{1}{12} & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{8} & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{6} & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{9} & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{5} \end{pmatrix} = \sigma^2\mathbf{V}.$$

La **matrice inverse** $\mathbf{V}^{-1}$ est alors la suivante :

$$\mathbf{V}^{-1} = \begin{pmatrix} 12 & 0 & 0 & 0 & 0 \\ 0 & 8 & 0 & 0 & 0 \\ 0 & 0 & 6 & 0 & 0 \\ 0 & 0 & 0 & 9 & 0 \\ 0 & 0 & 0 & 0 & 5 \end{pmatrix}$$

et l'estimation des paramètres β_0 et β_1 dans le modèle

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

se revèle être :

$$\begin{aligned} \hat{\beta}_0 &= -1,23 \\ \hat{\beta}_1 &= 0,502. \end{aligned}$$

Ce résultat revient à utiliser la **méthode des moindres carrés pondérée**.

MOTS CLÉS

Analyse de variance *(Analysis of variance)*
Modèle *(Model)*
Analyse de régression *(Regression analysis)*
Régression linéaire multiple *(Muliple linear regression)*
Régression non linéaire *(Non-linear regression)*
Théorème de Gauss-Markov *(Gauss-Markov theorem)*

RÉFÉRENCES

Bishop, Y.M.M., Fienberg, S.E., and Holland, P.W. (1975). *Discrete Multivariate Analysis : Theory and Practice.* MIT Press, Cambridge, Mass.

Graybill, F.A. (1976). *Theory and Applications of the Linear Model.* Duxbury Press, North Scituate,

Massachusetts; Waldsworth and Brooks/Cole, Pacific Grove, CA.

Rao, C.R. (1973). *Linear Statistical Inference and Its Applications*. 2nd ed., John Wiley & Sons, New York.

RÉGRESSION LINÉAIRE MULTIPLE

Multiple linear regression

Une **analyse de régression** où la **variable dépendante** Y dépend linéairement de plusieurs variables indépendantes X_1, X_2, ..., X_k est appelée régression linéaire multiple.

L'équation de régression linéaire multiple est de la forme :

$$Y = f\,(X_1, X_2, \ldots, X_k)$$

où $f\,(X_1, X_2, \ldots, X_k)$ est une fonction linéaire de X_1, X_2, ..., X_k.

HISTORIQUE

Voir **analyse de régression**.

ASPECTS MATHÉMATIQUES

Un **modèle** général de régression linéaire multiple comprenant $k = (p-1)$ variables indépendantes (et n observations) s'écrit :

$$Y_i = \beta_0 + \sum_{j=1}^{p-1} X_{ji}\beta_j + \varepsilon_i, \quad i = 1, \ldots, n$$

où

Y_i	est la **variable dépendante**,
X_{ji},	$j = 1, \ldots, p-1$, sont les variables indépendantes,
β_j,	$j = 0, \ldots, p-1$, sont les paramètres à estimer et
ε_i	est le terme d'**erreur** aléatoire non observable.

Sous forme matricielle, ce **modèle** s'écrit :

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

où

$\mathbf{Y}$	est le vecteur $(n \times 1)$ des observations relatives à la variable dépendante (n observations),
$\boldsymbol{\beta}$	est le vecteur $(p \times 1)$ des paramètres à estimer,
$\boldsymbol{\varepsilon}$	est le vecteur $(n \times 1)$ des erreurs

et $\mathbf{X} = \begin{pmatrix} 1 & X_{11} & \ldots & X_{1(p-1)} \\ \vdots & \vdots & & \vdots \\ 1 & X_{n1} & \ldots & X_{n(p-1)} \end{pmatrix}$ est la **matrice** $(n \times p)$ ayant trait aux variables indépendantes.

Estimation du vecteur $\boldsymbol{\beta}$

Partant du modèle :

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

nous obtenons l'estimation $\hat{\boldsymbol{\beta}}$ du vecteur $\boldsymbol{\beta}$ par la méthode des **moindres carrés** :

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}.$$

et nous calculons une valeur estimée $\widehat{\mathbf{Y}}$ pour $\mathbf{Y}$:

$$\widehat{\mathbf{Y}} = \mathbf{X} \cdot \hat{\boldsymbol{\beta}}.$$

À ce stade, nous pouvons calculer les résidus, notés dans le vecteur $\mathbf{e}$, que nous trouvons de la façon suivante :

$$\mathbf{e} = \mathbf{Y} - \widehat{\mathbf{Y}}.$$

Mesure de la fiabilité de l'estimation de Y

Pour savoir dans quelle mesure nous pouvons nous fier au modèle linéaire choisi, il est utile d'effectuer une **analyse de variance** et de tester les hypothèses sur le vecteur $\boldsymbol{\beta}$ du modèle de régression. Pour effectuer ces tests, nous devons faire les suppositions suivantes :

- Pour chaque **valeur** de X_{ji} et pour tout $i = 1, \ldots, n$ et $j = 1 \ldots, p-1$, $\mathbf{Y_i}$ est une **variable aléatoire** distribuée selon la **loi normale**.
- La **variance** de $\mathbf{Y}$ est la même pour tout X_{ji} ; elle est égale à σ^2 (inconnu).
- Les différentes observations sur $\mathbf{Y}$ sont indépendantes les unes des autres mais conditionnées par les valeurs de X_{ji}.

Analyse de variance sous forme matricielle

Sous forme matricielle, le tableau d'**analyse de variance** pour la régression s'écrit comme suit :

Analyse de variance

Source de variation	Degré de liberté	Somme des carrés	Moyenne des carrés
Régression	$p-1$	$\hat{\boldsymbol{\beta}}'\ \mathbf{X}'\mathbf{Y} - n\bar{\mathbf{Y}}^2$	$\dfrac{\hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{Y} - n\bar{\mathbf{Y}}^2}{p-1}$
Résiduelle	$n-p$	$\hat{\boldsymbol{\beta}}'\mathbf{Y} - \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{Y}$	$\dfrac{\mathbf{Y}'\mathbf{Y} - \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{Y}}{n-p}$
Total	$n-1$	$\mathbf{Y}'\mathbf{Y} - n\bar{\mathbf{Y}}^2$	

Si le **modèle** est correct,

$$S^2 = \frac{\mathbf{Y}'\mathbf{Y} - \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{Y}}{n-p}$$

est un **estimateur** sans biais de σ^2.

L'analyse de variance nous permet de tester l'**hypothèse nulle** :

$$H_0 \quad : \quad \beta_j = 0 \text{ pour } j = 1, \ldots, p-1$$

contre l'**hypothèse alternative** :

H_1 : au moins un des paramètres β_j, $j \neq 0$, est différent de zéro

en calculant la **statistique** :

$$F = \frac{MC_{reg}}{MC_{res}} = \frac{MC_{reg}}{S^2},$$

où

MC_{reg} est la moyenne des carrés de la régression,
MC_{res} est la moyenne des carrés des résidus et
MC_{tot} est la moyenne des carrés totale.

Cette statistique F doit être comparée avec la **valeur** $F_{\alpha,p-1,n-p}$ de la **table de Fisher**, où α est le **seuil de signification** du test :

$\Rightarrow$ si $F \leq F_{\alpha,p-1,n-p}$, nous acceptons H_0
si $F > F_{\alpha,p-1,n-p}$, nous rejetons H_0 au profit de H_1

Le coefficient de détermination R^2 se calcule de la manière suivante :

$$R^2 = \frac{SC_{reg}}{SC_{tot}} = \frac{\hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{Y} - n\bar{\mathbf{Y}}^2}{\mathbf{Y}'\mathbf{Y} - n\bar{\mathbf{Y}}^2}$$

où

SC_{reg} est la somme des carrés de la régression et
SC_{tot} est la somme des carrés totale.

DOMAINES ET LIMITATIONS

Voir **analyse de régression**.

EXEMPLES

Les données ci-dessous concernent 10 entreprises de l'industrie chimique. Nous cherchons à établir une relation entre la production, les heures de travail et le capital.

Production (100 tonnes)	Travail (heures)	Capital (machines heures)
60	1 100	300
120	1 200	400
190	1 430	420
250	1 500	400
300	1 520	510
360	1 620	590
380	1 800	600
430	1 820	630
440	1 800	610
490	1 750	630

Le **modèle** matriciel s'écrit

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

où

$$\mathbf{Y} = \begin{bmatrix} 60 \\ 120 \\ 190 \\ 250 \\ 300 \\ 360 \\ 380 \\ 430 \\ 440 \\ 490 \end{bmatrix} \qquad \mathbf{X} = \begin{bmatrix} 1 & 1\,100 & 300 \\ 1 & 1\,200 & 400 \\ 1 & 1\,430 & 420 \\ 1 & 1\,500 & 400 \\ 1 & 1\,520 & 510 \\ 1 & 1\,620 & 590 \\ 1 & 1\,800 & 600 \\ 1 & 1\,820 & 630 \\ 1 & 1\,800 & 610 \\ 1 & 1\,750 & 630 \end{bmatrix}$$

$$\boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \\ \varepsilon_6 \\ \varepsilon_7 \\ \varepsilon_8 \\ \varepsilon_9 \\ \varepsilon_{10} \end{bmatrix} \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix}.$$

Les équations qui forment le modèle sont alors les suivantes :

$$\begin{aligned} 60 &= \beta_0 + 1\,100\beta_1 + 300\beta_2 + \varepsilon_1 \\ 120 &= \beta_0 + 1\,200\beta_1 + 400\beta_2 + \varepsilon_2 \\ &\dots \\ 490 &= \beta_0 + 1\,750\beta_1 + 630\beta_2 + \varepsilon_{10} \end{aligned}$$

Nous calculons les estimateurs $\hat{\boldsymbol{\beta}}$ en utilisant le résultat :

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y},$$

avec

$$(\mathbf{X}'\mathbf{X}) = \begin{bmatrix} 10 & 15\,540 & 5\,090 \\ 15\,540 & 24\,734\,600 & 8\,168\,700 \\ 5\,090 & 8\,168\,700 & 2\,720\,500 \end{bmatrix}$$

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{bmatrix} 6,304288 & -0,007817 & 0,011677 \\ -0,007817 & 0,000015 & -0,000029 \\ 0,011677 & -0,000029 & 0,000066 \end{bmatrix}$$

et

$$\mathbf{X}'\mathbf{Y} = \begin{bmatrix} 3\,020 \\ 5\,012\,000 \\ 1\,687\,200 \end{bmatrix}.$$

Cela donne :

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} -439,269 \\ 0,283 \\ 0,591 \end{bmatrix} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix}.$$

Ainsi nous obtenons l'équation pour la droite estimée :

$$\hat{Y}_i = -439,269 + 0,283X_{i1} + 0,591X_{i2}.$$

Analyse de variance sous forme matricielle

Le calcul des différents éléments du tableau nous donne :

1. calcul des degrés de liberté :

$$\begin{aligned} dl_{reg} &= p - 1 = 3 - 1 = 2 \\ dl_{res} &= n - p = 10 - 3 = 7 \\ dl_{tot} &= n - 1 = 10 - 1 = 9 \end{aligned}$$

où

dl_{reg} est le degré de liberté de la régression,
dl_{res} est le degré de liberté des résidus et
dl_{tot} est le degré de liberté total.

2. calcul des sommes des carrés :

$$SC_{reg} = \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{Y} - n\bar{\mathbf{Y}}^2$$

$$= [-439,269\ 0,283\ 0,591] \cdot \begin{bmatrix} 3\,020 \\ 5\,012\,000 \\ 1\,687\,200 \end{bmatrix} - 912\,040$$

$$= 179\,063,39.$$

$$\begin{aligned} SC_{tot} &= \mathbf{Y}'\mathbf{Y} - n\bar{\mathbf{Y}}^2 \\ &= [60 \dots 490] \cdot \begin{bmatrix} 60 \\ \vdots \\ 490 \end{bmatrix} - 912\,040 \\ &= 187\,160. \end{aligned}$$

$$\begin{aligned} SC_{res} &= \mathbf{Y}'\mathbf{Y} - \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{Y}' \\ &= SC_{tot} - SC_{Reg} \\ &= 187\,160,00 - 179\,063,39 \\ &= 8\,096,61. \end{aligned}$$

Nous obtenons donc le tableau suivant :

Analyse de variance

Source de variation	Degré de liberté	Somme des carrés	Moyenne des carrés
Régression	2	179 063,39	89 531,70
Résiduelle	7	8 096,61	1 156,66 = S^2
Total	9	187 160,00	

Celle-ci nous permet de tester l'**hypothèse nulle** :

$$H_0 \quad : \quad \beta_1 = \beta_2 = 0.$$

La **valeur** calculée de F est :

$$\begin{aligned} F &= \frac{MC_{reg}}{MC_{res}} = \frac{MC_{reg}}{S^2} = \frac{89\,531,70}{1\,156,66} \\ &= 77,41. \end{aligned}$$

La valeur de F dans la **table de Fisher** avec un **seuil de signification** $\alpha = 0,05$ est

$$F_{0,05,2,7} = 4,74.$$

Par conséquent, l'**hypothèse nulle** H_0 est rejetée. L'**hypothèse alternative** peut être l'une des trois possibilités suivantes :

$$\begin{array}{lcl} H_1 & : & \beta_1 \neq 0 \text{ et } \beta_2 \neq 0 \\ H_2 & : & \beta_1 = 0 \text{ et } \beta_2 \neq 0 \\ H_3 & : & \beta_1 \neq 0 \text{ et } \beta_2 = 0. \end{array}$$

Nous pouvons alors calculer le **coefficient de détermination** R^2 :

$$R^2 = \frac{SC_{reg}}{SC_{tot}} = \frac{179\,063,39}{187\,160,00} \cong 0,96.$$

MOTS CLÉS

Analyse de variance *(Analysis of variance)*
Analyse des résidus *(Analysis of residuals)*
Coefficient de détermination *(Coefficient of determination)*
Corrélation *(Correlation)*
Colinéarité *(Collinearity)*
Équations normales *(Normal equations)*
Matrice H *(Hat matrix)*
Moindres carrés *(Least squares)*
Point levier *(Leverage point)*
Analyse de régression *(Regression analysis)*
Régression linéaire simple *(Simple linear regression)*
Résidu *(Residual)*

RÉGRESSION LINÉAIRE SIMPLE

Simple linear regression

La régression linéaire simple est une **analyse de régression** où la **variable dépendante** Y dépend linéairement d'une seule **variable indépendante** X.

La régression linéaire simple a pour objet non seulement d'estimer la fonction de régression relative au **modèle** choisi, mais également de tester la fiabilité des estimations obtenues.

HISTORIQUE

Voir **analyse de régression**.

ASPECTS MATHÉMATIQUES

Le **modèle** de régression linéaire simple est de la forme :

$$Y = \beta_0 + \beta_1 X + \varepsilon,$$

où

- Y est la **variable dépendante** (ou variable expliquée),
- X est la **variable indépendante** (ou variable explicative),
- ε est le terme d'**erreur** aléatoire non observable,
- β_0 et β_1 sont les paramètres à estimer.

Si nous avons un ensemble de n observations (X_1, Y_1), $\ldots$, (X_n, Y_n), où les Y_i dépendent linéairement des X_i correspondants, nous pouvons donc écrire :

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad i = 1, \ldots, n.$$

Le problème de l'**analyse de régression** consiste à estimer les paramètres β_0 et β_1, en choisissant les valeurs $\hat{\beta}_0$ et $\hat{\beta}_1$ telles que la **distance** entre Y_i et $(\beta_0 + \beta_1 \cdot X_i)$ soit minimale. Il faut donc que

$$\varepsilon_i = Y_i - \beta_0 - \beta_1 \cdot X_i$$

soit petit pour tout $i = 1, \ldots, n$. Pour ce faire, nous pouvons choisir parmi plusieurs critères comme :

1. $\min\limits_{\beta_0,\beta_1} \max\limits_{i} |\varepsilon_i|$
2. $\min\limits_{\beta_0,\beta_1} \sum\limits_{i=1}^{n} |\varepsilon_i|$
3. $\min\limits_{\beta_0,\beta_1} \sum\limits_{i=1}^{n} \varepsilon_i^2$

La méthode d'**estimation** des paramètres β_0 et β_1 la plus fréquemment utilisée est la troisième, appelée

méthode des moindres carrés. Celle-ci a pour objet de minimiser la somme des carrés des erreurs.

L'estimation des paramètres par la méthode des moindres carrés donne les estimateurs $\hat{\beta}_0$ et $\hat{\beta}_1$ suivants :

$$\begin{aligned}\hat{\beta}_1 &= \frac{\sum_{i=1}^{n}\left(X_i-\bar{X}\right)\left(\hat{Y}_i-\bar{Y}\right)}{\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2}\\ \hat{\beta}_0 &= \bar{Y}-\hat{\beta}_1\bar{X}.\end{aligned}$$

Nous pouvons alors écrire :

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

où $\hat{Y}_i$ est la valeur estimée pour un X_i donné lorsque $\hat{\beta}_0$ et $\hat{\beta}_1$ sont déterminés.

Mesure de la fiabilité de l'estimation de Y

Nous avons calculé une estimation de la valeur de Y, soit $\hat{Y}$, à l'aide de la méthode des moindres carrés, en nous basant sur un modèle linéaire pour traduire la relation qui lie Y à X. Mais jusqu'à quel point pouvons-nous nous fier à ce modèle ?

Pour répondre à cette question, il est utile d'effectuer une **analyse de variance** et de tester les hypothèses sur les paramètres β_0 et β_1 de la droite de régression. Pour effectuer ces tests, nous devons faire les suppositions suivantes :

- Pour chaque valeur de X, Y est une **variable aléatoire** distribuée selon la **loi normale**.
- La **variance** de Y est la même pour tout X ; elle est égale à σ^2 (inconnu).
- Les différentes observations sur Y sont indépendantes les unes des autres mais conditionnées par les valeurs de X.

Analyse de variance

Le tableau d'analyse de variance que nous devons construire est le suivant :

Analyse de variance

Source de variation	Degré de liberté	Somme des carrés	Moyenne des carrés
Régression	1	$\sum_{i=1}^{n}\left(\hat{Y}_i-\bar{Y}\right)^2$	$\sum_{i=1}^{n}\left(\hat{Y}_i-\bar{Y}\right)^2$
Résiduelle	$n-2$	$\sum_{i=1}^{n}\left(\hat{Y}_i-\bar{Y}\right)^2$	$\frac{\sum_{i=1}^{n}\left(\hat{Y}_i-\bar{Y}\right)^2}{n-2}$
Total	$n-1$	$\sum_{i=1}^{n}\left(\hat{Y}_i-\bar{Y}\right)^2$	

Si le modèle est correct,

$$S^2 = \frac{\sum_{i=1}^{n}\left(\hat{Y}_i-\bar{Y}\right)^2}{n-2}$$

est un estimateur sans biais de σ^2.

L'analyse de variance nous permet de tester l'**hypothèse nulle** :

$$H_0 \quad : \quad \beta_1 = 0$$

contre l'**hypothèse alternative** :

$$H_1 \quad : \quad \beta_1 \neq 0$$

en calculant la **statistique** :

$$F = \frac{MC_{reg}}{MC_{res}} = \frac{MC_{reg}}{S^2}.$$

Cette statistique F doit être comparée avec la valeur $F_{\alpha,1,n-2}$ de la **table de Fisher**, où α est le **seuil de signification** du test.

$\Rightarrow$ si $F \leq F_{\alpha,1,n-2}$, nous acceptons H_0,
si $F > F_{\alpha,1,n-2}$, nous rejetons H_0 au profit de H_1.

Le **coefficient de détermination** R^2 se calcule de la manière suivante :

$$R^2 = \frac{SC_{reg}}{SC_{tot}} = \frac{\hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{Y} - n\bar{\mathbf{Y}}^2}{\mathbf{Y}'\mathbf{Y} - n\bar{\mathbf{Y}}^2}$$

où

SC_{reg} est la somme des carrés de la régression et
SC_{tot} est la somme des carrés totale.

Hypothèse sur la pente β_1

Dans le cas d'une régression linéaire simple, la statistique F permet de tester une **hypothèse** sur un seul des paramètres de l'équation de régression, soit β_1, la pente de la droite. Une autre façon d'effectuer le même test est la suivante :

Si H_0 est vraie ($\beta_1 = 0$), la statistique t :

$$t = \frac{\hat{\beta}_1}{S_{\hat{\beta}_1}}$$

suit une **loi de Student** avec $(n-2)$ degrés de liberté.

$S_{\hat{\beta}_1}$ est l'**écart type** de $\hat{\beta}_1$, estimé à partir de l'**échantillon** :

$$S_{\hat{\beta}_1} = \sqrt{\frac{S^2}{\sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2}}.$$

La statistique t doit être comparée avec la valeur $t_{\frac{\alpha}{2},n-2}$ de la **table de Student**, où α est le seuil de signification et $n-2$ le nombre de degrés de liberté. La règle de décision est alors la suivante :

si $\lvert t\rvert \leq t_{\frac{\alpha}{2},n-2}$	$\Rightarrow$ nous acceptons H_0 ($\beta_1 = 0$) ;
si $\lvert t\rvert > t_{\frac{\alpha}{2},n-2}$	$\Rightarrow$ nous rejetons H_0 au profit de H_1 ($\beta_1 \neq 0$).

Il est possible de démontrer que, dans le cas d'une régression linéaire simple, t^2 est égal à F.

La statistique t permet de calculer un **intervalle de confiance** pour β_1.

Hypothèse sur l'ordonnée à l'origine β_0

De façon similaire, nous pouvons construire un intervalle de confiance pour β_0 et tester les hypothèses

$$\begin{array}{lll} H_0 & : & \beta_0 = 0, \\ H_1 & : & \beta_0 \neq 0. \end{array}$$

La statistique t

$$t = \frac{\hat{\beta}_0}{S_{\hat{\beta}_0}}$$

suit également une loi de Student avec $(n-2)$ degrés de liberté.

L'écart type estimé de $\hat{\beta}_0$, noté par $S_{\hat{\beta}_0}$, est défini par :

$$S_{\hat{\beta}_0} = \sqrt{\frac{S^2 \sum_{i=1}^{n} X_i^2}{n \sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2}}.$$

La règle de décision est la suivante :

si $\lvert t\rvert \leq t_{\frac{\alpha}{2},n-2}$	$\Rightarrow$ nous acceptons H_0 ($\beta_0 = 0$) ;
si $\lvert t\rvert > t_{\frac{\alpha}{2},n-2}$	$\Rightarrow$ nous rejetons H_0 au profit de H_1 ($\beta_0 \neq 0$).

$t_{\frac{\alpha}{2},n-2}$ est obtenu à partir de la table de Student pour $(n-2)$ degrés de liberté et un seuil de signification α.

Le **coefficient de détermination** R^2 se calcule de la manière suivante :

$$R^2 = \frac{SC_{reg}}{SC_{tot}} = \frac{\hat{\beta}' X'Y - n\bar{Y}^2}{Y'Y - n\bar{Y}^2}$$

où

SC_{reg}	est la somme des carrés de la régression et
SC_{tot}	est la somme des carrés totale.

DOMAINES ET LIMITATIONS

Le cas de la régression linéaire simple est un cas particulier de la **régression linéaire multiple**. L'approche matricielle exposée dans la régression linéaire multiple (**modèle** comprenant plusieurs variables indépendantes) est donc également valable pour ce cas particulier où nous ne sommes en présence que d'une seule variable indépendante.

Voir **analyse de régression**.

EXEMPLES

Le tableau suivant représente le produit national brut (PNB) et la demande des biens de première nécessité obtenus pour la période 1969-1980 dans un certain pays.

Année	PNB	Demande des biens de première nécessité
	X	Y
1969	50	6
1970	52	8
1971	55	9
1972	59	10
1973	57	8
1974	58	10
1975	62	12
1976	65	9
1977	68	11
1978	69	10
1979	70	11
1980	72	14

Nous cherchons à estimer la demande des biens de première nécessité en fonction du PNB suivant le **modèle** :

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad i = 1, \ldots, n.$$

L'**estimation** des paramètres β_0 et β_1 par la méthode des **moindres carrés** nous donne les estimateurs suivants :

$$\begin{aligned} \hat{\beta}_1 &= \frac{\sum_{i=1}^{12} \left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{\sum_{i=1}^{12} \left(X_i - \bar{X}\right)^2} = 0,226 \\ \hat{\beta}_0 &= \bar{Y} - \hat{\beta}_1 \bar{X} = -4,04 \end{aligned}$$

La droite estimée peut donc s'écrire :

$$\hat{Y} = -4,04 + 0,226X.$$

Analyse de variance

Nous allons calculer les degrés de liberté, ainsi que les sommes des carrés et les moyennes des carrés afin d'établir le tableau d'**analyse de variance** :

$$\begin{aligned} dl_{reg} &= 1 \\ dl_{res} &= n - 2 = 10 \\ dl_{tot} &= n - 1 = 11 \end{aligned}$$

$$\begin{aligned} SC_{reg} &= \sum_{i=1}^{12} \left(\hat{Y}_i - \bar{Y}\right)^2 \\ &= (7,20 - 9,833)^2 + \cdots \\ &\quad + (12,23 - 9,833)^2 \\ &= 30,457 \end{aligned}$$

$$\begin{aligned} SC_{res} &= \sum_{i=1}^{12} \left(Y_i - \hat{Y}_i\right)^2 \\ &= (6 - 7,20)^2 + \cdots + (14 - 12,23)^2 \\ &= 17,21 \end{aligned}$$

$$\begin{aligned} SC_{tot} &= \sum_{i=1}^{12} \left(Y_i - \bar{Y}\right)^2 \\ &= (6 - 9,833)^2 + \cdots + (14 - 9,833)^2 \\ &= 47,667 \end{aligned}$$

$$\begin{aligned} MC_{reg} &= \frac{SC_{reg}}{dl_{reg}} = 30,457 \\ S^2 = MC_{res} &= \frac{SC_{res}}{dl_{res}} \\ &= 1,721. \end{aligned}$$

Nous avons donc le tableau suivant :

Analyse de variance

Source de variation	Degré de liberté	Somme des carrés	Moyenne des carrés	F
Régression	1	30,457	30,457	17,687
Résiduelle	10	17,211	1,721	
Total	11	47,667		

Avec un **seuil de signification** $\alpha = 5\ \%$, nous trouvons dans la **table de Fisher** la valeur

$$F_{\alpha,1,n-2} = F_{0,05,1,10} = 4,96.$$

Par conséquent comme $F > F_{0,05,1,10}$, nous rejetons l'**hypothèse nulle** $H_0 : \beta_1 = 0$, ce qui signifie que β_1 est significativement différent de zéro.

Pour calculer le **coefficient de détermination** de cet exemple, il suffit de se reporter au tableau d'analyse de variance puisqu'il contient tous les éléments requis :

$$\begin{aligned} R^2 &= \frac{SC_{reg}}{SC_{tot}} = \frac{30,457}{47,667} \\ &= 0,6390 = 63,90\,\%. \end{aligned}$$

Nous pouvons conclure que selon le modèle choisi, 63,90 % de la variation de la demande des biens de première nécessité est expliquée par la variation du PNB.

Il est évident que la valeur de R^2 ne peut pas dépasser 100 ; la valeur 63,90 est assez grande mais elle n'est pourtant pas suffisamment proche de 100 pour ne pas tenter d'améliorer le modèle.

Cela peut signifier :

- qu'en plus du PNB, d'autres variables devraient être prises en considération pour une détermination plus fine de la fonction de demande en biens de première nécessité, donc que le PNB n'explique qu'une partie de la variation ;
- qu'il faut tester un autre modèle (modèle sans terme constant, modèle non linéaire, etc.).

Les tests d'hypothèse sur les paramètres vont nous permettre de déterminer si ceux-ci sont significativement différents de zéro ou non.

Hypothèse sur la pente β_1

Dans la table d'analyse de variance, nous avons obtenu :

$$S^2 = MC_{res} = 1,721$$

ce qui nous permet de calculer l'**écart type** $S_{\hat{\beta}_1}$ de $\hat{\beta}_1$:

$$\begin{aligned} S_{\hat{\beta}_1} &= \sqrt{\frac{S^2}{\sum_{i=1}^{12} \left(X_i - \bar{X}\right)^2}} \\ &= \sqrt{\frac{1,721}{(50-61,416)^2 + \cdots + (72-61,416)^2}} \\ &= \sqrt{\frac{1,721}{596,92}} = 0,054. \end{aligned}$$

Nous pouvons calculer la statistique :

$$t = \frac{\hat{\beta}_1}{S_{\hat{\beta}_1}} = \frac{0,226}{0,054} = 4,19.$$

En choisissant un seuil de signification $\alpha = 5$ %, nous obtenons

$$t_{\frac{\alpha}{2},n-2} = t_{0,025,10} = 2,228.$$

En comparant cette **valeur** de t avec la valeur de la table, nous obtenons :

$$|t| > t_{\frac{\alpha}{2},n-2}$$

ce qui indique que l'hypothèse nulle

$$H_0 \quad : \quad \beta_1 = 0$$

doit être rejetée en faveur de l'**hypothèse alternative**

$$H_1 \quad : \quad \beta_1 \neq 0.$$

Nous concluons qu'il existe une relation linéaire entre le PNB et la demande des biens de première nécessité. Si H_0 avait été acceptée, cela aurait signifié qu'il n'y avait pas de relation linéaire entre ces deux variables.

Remarque : nous pouvons constater que

$$\begin{aligned} F &= t^2, \\ 17,696 &= 4,19^2 \end{aligned}$$

(sans tenir compte des erreurs d'arrondis). Par conséquent, nous pouvons utiliser indifféremment le test t ou le test F pour tester l'hypothèse $H_0 : \beta_1 = 0$.

Hypothèse sur l'ordonnée à l'origine β_0

L'**écart type** $S_{\hat{\beta}_0}$ de $\hat{\beta}_0$ est égal à :

$$\begin{aligned} S_{\hat{\beta}_0} &= \sqrt{\frac{S^2 \sum_{i=1}^{12} \bar{X}_i^2}{n \sum_{i=1}^{12} \left(X_i - \bar{X}\right)^2}}. \\ &= \sqrt{\frac{1,721\,(50^2 + \cdots + 72^2)}{(12 \cdot 596,92)}} \\ &= \sqrt{\frac{1,721 \cdot 45\,861}{7\,163,04}} = 3,31. \end{aligned}$$

Nous pouvons calculer la **statistique** :

$$t = \frac{\hat{\beta}_0}{S_{\hat{\beta}_0}} = \frac{-4,04}{3,31} = -1,22.$$

En comparant cette **valeur** de t avec la valeur de la **table de Student** $t_{0,025;10} = 2,228$, nous obtenons :

$$|t| < t_{\frac{\alpha}{2},n-2},$$

ce qui indique que l'hypothèse nulle

$$H_0 \quad : \quad \beta_0 = 0$$

doit être acceptée, au seuil de signification $\alpha = 5$ %. Nous concluons donc que la valeur de β_0 n'est pas significativement différente de zéro, et que la droite passe donc par l'origine.

Nouveau modèle : régression passant par l'origine

Le nouveau modèle est le suivant :

$$Y_i = \beta_1 X_i + \varepsilon_i \quad i = 1, \ldots, 12.$$

En effectuant l'estimation du paramètre β_1 selon la méthode des moindres carrés, nous avons :

$$\sum_{i=1}^{12} \varepsilon_i^2 = \sum_{i=1}^{12} \left(Y_i - \beta_1 X_i\right)^2.$$

En annulant la dérivée par rapport à β_1, nous obtenons :

$$\frac{\partial \sum_{i=1}^{12} \varepsilon_i^2}{\partial \beta_1} = -2 \sum_{i=1}^{12} X_i \left(Y_i - \beta_1 X_i\right) = 0$$

ou

$$\sum_{i=1}^{12} \left(X_i Y_i\right) - \beta_1 \sum_{i=1}^{12} X_i^2 = 0.$$

La valeur de β_1 qui satisfait l'équation est l'estimateur $\hat{\beta}_1$ de β_1 :

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{12} \left(X_i Y_i\right)}{\sum_{i=1}^{12} X_i^2} = \frac{7\,382}{45\,861} = 0,161.$$

Par conséquent, la nouvelle droite de régression est décrite par :

$$\begin{aligned} \hat{Y}_i &= \hat{\beta}_1 X_i \\ \hat{Y}_i &= 0,161 X_i. \end{aligned}$$

Le tableau d'analyse de variance relatif à ce nouveau modèle est le suivant :

Analyse de variance

Source de variation	Degré de liberté	Somme des carrés	Moyenne des carrés	F
Régression	1	1 188,2	1 188,2	660,11
Résiduelle	11	19,8	1,8	
Total	12	1 208,0		

Notons que dans un modèle sans constante, le nombre dedegrés de liberté pour la variation totale est égal au nombre d'observations n. Le nombre de degrés de liberté pour la variation résiduelle est donc égal à $n - 1$.

Nous pouvons alors déterminer R^2 :

$$\begin{aligned} R^2 &= \frac{SC_{reg}}{SC_{tot}} = \frac{1\,188,2}{1\,208,0} = 0,9836 \\ &= 98,36\,\%. \end{aligned}$$

Nous avions trouvé avec le premier modèle la valeur $R^2 = 63,90$ %. Nous pouvons donc constater que ce nouveau modèle sans terme constant est bien meilleur que le premier.

MOTS CLÉS

Analyse de régression *(Regression analysis)*
Analyse des résidus *(Analysis of residuals)*
Coefficient de détermination *(Coefficient of determinatin)*
Corrélation *(Correlation)*
Équations normales *(Normal equations)*
Moindres carrés *(Least squares)*
Régression linéaire multiple *(Multiple linear regression)*
Résidu *(Residual)*

RÉGRESSION LOGISTIQUE

Logistic regression

Lorsque la **variable dépendante** d'un **modèle** de régression est une **variable dichotomique** (ou binaire), la régression logistique donne la possibilité d'ajuster un modèle de régression. La régression logistique permet donc d'étudier la **relation** entre une proportion et un ensemble de variables aléatoires indépendantes quantitatives ou qualitatives.

ASPECTS MATHÉMATIQUES

Le modèle d'une régression logistique avec variable dépendante Y et variables indépendantes X_1, X_2, …, X_p s'exprime par :

$$\begin{aligned}\pi\left((X_{ij})_{i=1,\ldots,p}\right) &= \log\left(\frac{P(Y_j=1)}{P(Y_j=0)}\right)\\ &= \beta_0+\sum_{i=1}^{p}\beta_i X_{ij} \quad j=1,\ldots,n\end{aligned}$$

où les Y_j $(j=1,\ldots,n)$ correspondent aux réalisations (autrement dit observations) de la variable Y et X_{ij} aux observations des variables indépendantes et π est la fonction de régression logistique :

$$\pi\left((X_{ij})_{i=1,\ldots,p}\right)=\frac{\exp\left(\beta_0+\sum_{i=1}^{p}\beta_i X_{ij}\right)}{1+\exp\left(\beta_0+\sum_{i=1}^{p}\beta_i X_{ij}\right)}.$$

Une formulation équivalente est donnée par :

$$P(Y_j=1)=\frac{\exp\left(\beta_0+\sum_{i=1}^{p}\beta_i X_{ij}\right)}{1+\exp\left(\beta_0+\sum_{i=1}^{p}\beta_i X_{ij}\right)},$$

ce qui permet de calculer la probabilité (comprise entre 0 et 1) que la variable Y prenne la valeur 1 en fonction des réalisations des variables indépendantes X_1, X_2, …, X_p.

Une proprieté intéressante est que les paramètres du modèle $(\beta_0,\beta_1,\ldots,\beta_p)$ peuvent prendre n'importe quelle valeur, puisque toute valeur de la transformée $logit\left(\pi(Y)\right)=\log\left(\frac{\pi(Y)}{1-\pi(Y)}\right)$ correspond à un numéro entre 0 et 1 pour $P(Y=1)$.

Les paramètres du modèle peuvent être estimés par la méthode du **maximum de vraisemblance,** en maximisant la fonction suivante :

$$\begin{aligned}L(\beta_0,\beta_1,\ldots,\beta_n) &= \prod_{j=1}^{n}\pi\left((X_{ij})_{i=1,\ldots,p}\right)^{Y_i}\cdot\\ &\quad\cdot\left(1-\pi\left((X_{ij})_{i=1,\ldots,p}\right)\right)^{1-Y_j}\end{aligned}$$

ou, plus facilement, en maximisant la fonction de log-vraisemblance :

$$\begin{aligned}&\log L(\beta_0,\beta_1,\ldots,\beta_n)\\ &= \sum_{j=1}^{n}Y_j\log\pi\left((X_{ij})_{i=1,\ldots,p}\right)\\ &\quad+(1-Y_j)\log\left(1-\pi\left((X_{ij})_{i=1,\ldots,p}\right)\right).\end{aligned}$$

On peut en conséquence tester des hypothèses sur les valeurs des paramètres (par exemple $\beta_0=0$, $\beta_3=\beta_2$, etc.) en effectuant un **test du rapport de vraisemblance**.

L'interprétation des paramètres correspondants aux variables indépendantes dichotomiques est expliquée dans le paragraphe qui suit. Pour les variables indépendantes continues X, les paramètres expriment le risque relatif à l'augmentation de X d'une unité. Dans le cas de variables indépendantes à plus de deux catégories, il faut les transformer (comme décrit dans données binaires) en variables binaires. L'interprétation des coefficients est alors claire.

DOMAINES ET LIMITATIONS

La régression logistique est souvent utilisée en **épidémiologie** à cause de :

- la forme sigmoïde (de l'expression donnée auparavant pour $P(Y_j=1)$) qui correspond à la forme de la relation souvent observée entre une dose X et la fréquence d'une maladie Y ;
- la facilité de son interprétation : la mesure de l'association entre une maladie et un facteur M (qui correspond à une variable binaire X_i) s'exprime par l'**odds ratio**. Ceci est une bonne approximation du **risque relatif** (si les probabilités $P(Y=1\,|X_i=1)$ et $P(Y=1\,|X_i=0)$ sont petites) et se calcule tout simplement comme l'exponentiel du paramètre associé à la variable X_i, donc :

$$odds\ ratio \text{ du facteur } M=\exp(\beta_i).$$

EXEMPLE

On considère ici un exemple d'après Hosmer et Lemeshow. La variable dépendante indique la présence ou l'absence d'une maladie cardiaque pour 100 sujets et la variable indépendante leur âge.

Le modèle obtenu en effectuant une régression logistique à l'aide d'un logiciel est :

$$\pi(X) = -5,3 + 0,11 \cdot X$$

avec écart type pour la constante de $1,1337$ et de $0,241$ pour le coefficient β de le variable X. Les deux paramètres sont donc significatifs.

Le risque d'avoir une maladie cardiaque augmente donc de $\exp(0,11) = 1,12$ si l'âge augmente d'un an.

Il est aussi possible de calculer un **intervalle de confiance** prenant l'exponentiel de l'intervalle pour β :

$$(\exp(1,11 - 1,96 \cdot 0,241), \exp(1,11 + 1,96 \cdot 0,241)) = (1,065, 1,171).$$

Dans le tableau suivant, il y a les valeurs de la log-vraisemblance et des paramètres à estimer obtenus en maximisant la fonction de la log-vraisemblance avec un algorithme de nature itérative :

	log-vraisemblance	constante	β
1	$-54,24$	$-4,15$	$0,872$
2	$-53,68$	$-5,18$	$1,083$
3	$-53,67$	$-5,31$	$1,109$
4	$-53,67$	$-5,31$	$1,109$

MOTS CLÉS

Analyse de régression *(Regression analysis)*
Donnée binaire *(Binary data)*
Maximum de vraisemblance *(Maximum likelihood)*
Modèle *(Model)*
Odds et odds ratio *(Odds and odds ratio)*
Test du rapport de vraisemblance *(Likelihood ratio test)*
Table de contingence *(Contingency table)*

RÉFÉRENCES

Bishop, Y.M.M., Fienberg, S.E., and Holland, P.W. (1975). *Discrete Multivariate Analysis : Theory and Practice.* MIT Press, Cambridge, Mass.

Cox, D.R. (1970). *The analysis of binary data.* Methuen, London.

Hosmer, D.W., Lemeshow S. (1989). *Applied logistic regression.* Wiley Series in Probability and Statistics.

Douglas G. Altman. (1991). *Practical Statistics for Medical Research*, Chapman & Hall, London.

RÉGRESSION NON LINÉAIRE

Non-linear regression

Une **analyse de régression** où la **variable dépendante** Y dépend d'une ou plusieurs variables indépendantes X_1, ..., X_k est appelée régression non linéaire si l'équation de régression $Y = f(X_1, \ldots, X_k ; \beta_0, \ldots, \beta_p)$ n'est pas linéaire en ses paramètres β_0, ..., β_p.

HISTORIQUE

Les origines de la régression non linéaire remontent aux années 1920 et sont dues à **Ronald Aylmer Fisher** et à Mackenzie. Par contre, l'usage et l'investigation plus détaillés de ces modèles a dû attendre les progrès dans le calcul automatique des années 1970.

ASPECTS MATHÉMATIQUES

Soient Y_1, Y_2, ..., Y_r et X_{11}, ..., X_{1r}, X_{21}, ..., X_{2r}, ..., X_{k1}, ..., X_{kr} les r observations (respectivement) de la **variable dépendante** Y et des variables indépendantes X_1, ..., X_k.

Le but est d'estimer les paramètres β_0, ..., β_p du modèle :

$$Y_i = f(X_{1i}, \ldots, X_{ki} ; \beta_0, \ldots, \beta_p) + \varepsilon_i \quad i = 1, \ldots, r.$$

Par la méthode des **moindres carrés** et sous les hypothèses correspondantes concernant les erreurs (voir **analyse des résidus**), nous estimons les paramètres β_0, ..., β_p par les valeurs $\hat{\beta}_0$, ..., $\hat{\beta}_p$ qui minimisent la somme des erreurs au carré, notée $S(\beta_0, \ldots, \beta_p) = \sum_{i=1}^{r} (Y_i - f(X_{1i}, \ldots, X_{ki} ; \beta_0, \ldots, \beta_p))^2$:

$$\min_{\beta_0,\ldots,\beta_p} \sum_{i=1}^{r} \varepsilon_i^2 = \min_{\beta_0,\ldots,\beta_p} S\left(\beta_0,\ldots,\beta_p\right).$$

Puisque f est non linéaire la résolution des **équations normales** (elles aussi non linéaires) :

$$\sum_{i=1}^{r} \left(Y_i - f\left(X_{1i},\ldots,X_{ki}\,;\,\beta_0,\ldots,\beta_p\right)\right) \cdot \left[\frac{\partial f\left(X_{1i},\ldots,X_{ki}\,;\,\beta_0,\ldots,\beta_p\right)}{\partial \beta_j}\right] = 0$$

$(j = 0, \ldots, p)$ peut s'avérer difficile.

Un problème qui peut se poser est lorsque les équations données n'ont pas nécessairement une solution, ou en ont plus d'une.

Dans la suite, nous discutons des approches différentes à la résolution du problème.

1. *Gauss-Newton.*

   Le procédé, de nature itérative, consiste à la linéarisation de f à l'aide de l'expansion de Taylor et dans l'estimation successive des paramètres par la **régression** linéaire.

   Nous allons développer cette méthode plus en détail dans l'**exemple** qui suit.

2. *Steepest descent.*

   L'idée de cette méthode est de fixer, de façon approximative, les paramètres β_j à estimer et ensuite d'y ajouter une approximation de :

   $$-\frac{\partial S\left(\beta_0,\ldots,\beta_p\right)}{\partial \beta_j}$$

   qui correspond au maximum de descente (*steepest descent* en anglais) par rapport à la fonction S.

   Itérativement, nous déterminons les paramètres qui minimisent S.

3. *Marquard-Levenburg.*

   Cette méthode, que nous ne développons pas ici, intègre les avantages des deux procedés mentionnés ci-dessus en forçant et en accélérant la convergence de la suite des approximations des paramètres à estimer.

DOMAINES ET LIMITATIONS

Comme expliqué dans Domaines et limitations du mot **régression linéaire multiple**, il existe des fonctions qui ne sont pas linéaires sur les paramètres mais qui deviennent linéaires après une transformation. Si ce n'est pas le cas, nous les appelons fonctions non linéaires de manière intrinsèque et on les traite avec les méthodes non linéaires.

Un exemple d'une fonction non linéaire de manière intrinsèque est donnée par :

$$Y = \frac{\beta_1}{\beta_1 - \beta_2}\left[\exp\left(-\beta_2 X\right) - \exp\left(-\beta_1 X\right)\right] + \varepsilon.$$

EXEMPLES

Soient $P_1 = (x_1, y_1), P_2 = (x_2, y_2)$, $P_3 = (x_3, y_3)$ trois points dans le plan et d_1, d_2, d_3 les distances approximatives (les « approximations » étant distribuées selon une loi de Gauss de même variance) entre les points donnés et un point $P = (x, y)$ de coordonnées inconnues et recherchées.

Un **modèle** de régression possible pour estimer les coordonnées de P est le suivant :

$$d_i = f\left(x_i, y_i\,;\,x, y\right) + \varepsilon_i, \quad i = 1, 2, 3.$$

où la fonction f exprime la distance entre le point P_i et le point P, c'est-à-dire :

$$f\left(x_i, y_i\,;\,x, y\right) = \sqrt{\left(x - x_i\right)^2 + \left(y - y_i\right)^2}.$$

La fonction distance est clairement non linéaire et ne peut pas être ramenée à une forme linéaire, donc la régression appropriée est non linéaire. Nous appliquons la méthode de *Gauss-Newton.*

En considérant f comme fonction des paramètres à estimer x et y, le développement de Taylor de f jusqu'au terme linéaire en (x_0, y_0) est donné par :

$$\begin{aligned} f(x_i, y_i\,;\, x, y) \;\simeq\; & f(x_i, y_i\,;\, x_0, y_0) \\ & + \left.\frac{\partial f(x_i, y_i\,;\, x, y)}{\partial x}\right|_{\substack{x=x_0\\y=y_0}} \cdot (x - x_0) \\ & + \left.\frac{\partial f(x_i, y_i\,;\, x, y)}{\partial x}\right|_{\substack{x=x_0\\y=y_0}} \cdot (y - y_0) \end{aligned}$$

où

$$\left.\frac{\partial f(x_i, y_i\,;\, x, y)}{\partial x}\right|_{\substack{x=x_0\\y=y_0}} = \frac{x_i - x_0}{f(x_i, y_i\,;\, x_0, y_0)} \quad \text{et}$$

$$\left.\frac{\partial f(x_i, y_i\,;\, x, y)}{\partial x}\right|_{\substack{x=x_0\\y=y_0}} = \frac{y_i - y_0}{f(x_i, y_i; x_0, y_0)}.$$

Soit donc $(x_0\,;\, y_0)$ une première estimation de P, trouvée par exemple géometriquement.

Le modèle de régression linéarisé s'exprime par :

$$\begin{aligned} d_i \;=\; & f(x_i, y_i\,;\, x_0, y_0) \\ & + \frac{x_i - x_0}{f(x_i, y_i\,;\, x_0, y_0)} \cdot (x - x_0) \\ & + \frac{y_i - y_0}{f(x_i, y_i\,;\, x_0, y_0)} \cdot (y - y_0) + \varepsilon_i,\ i = 1, 2, 3. \end{aligned}$$

À l'aide d'une **régression linéaire multiple** (sans constante), nous pouvons estimer les paramètres $(\Delta x\,;\, \Delta y) = (x - x_0\,;\, y - y_0)$ du modèle suivant (les « observations » du modèle sont indiquées entre crochets) :

$$\begin{aligned} [d_i - f(x_i, y_i\,;\, x_0, y_0)] \;=\; & \Delta x \cdot \left[\frac{x_i - x_0}{f(x_i, y_i\,;\, x_0, y_0)}\right] \\ & + \Delta y \cdot \left[\frac{y_i - y_0}{f(x_i, y_i\,;\, x_0, y_0)}\right] \\ & + \varepsilon_i, \quad i = 1, 2, 3. \end{aligned}$$

Soit $\left(\hat{\Delta}x\,;\, \hat{\Delta}y\right)$ l'estimation trouvée de $(\Delta x\,;\, \Delta y)$.

Nous obtenons ainsi une meilleure estimation des coordonnées de P par $\left(x_0 + \hat{\Delta}x\,;\, y_0 + \hat{\Delta}y\right)$.

En utilisant comme point de départ ce nouveau point, nous pouvons à nouveau utiliser la même procédure. La suite des coordonnées trouvées de cette façon converge vers le point désiré, c'est-à-dire vers le point qui approche au mieux (dans le sens des moindres carrés) les distances approximatives par rapport aux points donnés.

Prenons maintenant un exemple concret : $P_1 = (0\,;\, 5)$; $P_2 = (5\,;\, 0)$; $P_3 = (10\,;\, 10)$; $d_1 = 6$; $d_2 = 4$; $d_3 = 6$.

Si nous choisissons comme point de départ du procédé *Gauss-Newton* le point suivant $P = \left(\frac{0+5+10}{3}\,;\, \frac{5+0+10}{3}\right) = (5\,;\, 5)$, nous obtenons la suite de coordonnées suivante :

$$\begin{aligned} & (5\,;\, 5) \\ & (6,2536\,;\, 4,7537) \\ & (6,1704\,;\, 4,7711) \\ & (6,1829\,;\, 4,7621) \\ & (6,1806\,;\, 4,7639) \end{aligned}$$

qui converge vers le point $(6,1806\,;\, 4,7639)$.

Les distances entre le point trouvé et P_1, P_2, P_3 sont respectivement de $6,1856$, $4,9078$ et $6,4811$.

MOTS CLÉS

Analyse de régression *(Regression analysis)*
Équations normales*(Normal equations)*
Modèle *(Model)*
Moindres carrés*(Least squares)*
Régression linéaire multiple *(Multiple linear regression)*
Régression linéaire simple *(Simple linear regression)*

RÉFÉRENCES

Bates, D. M. and Watts D. G. (1988). *Nonlinear regression analysis and its applications*, John Wiley and Sons.

Draper, N.R. and Smith, H. (1998). *Applied Regression Analysis*, 3rd edition. John Wiley & Sons, New York.

Fisher, R.A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London.* Series A, Vol. 222, pp. 309-368.

Fisher, R. A. and Mackenzie, W. A. (1923). The manurial response of different potato varieties. *Journal of Agricultural Science*, 13, pp. 311-320.

Fletcher, R. (1997). *Practical methods of optimization*, John Wiley & Sons, New York.

RÉGRESSION RIDGE

Ridge regression

Le but de la méthode ridge est de minimiser l'erreur quadratique moyenne des estimateurs, c'est-à-dire de choisir un compromis entre le biais et la variance. En considérant le phénomène de colinéarité comme un problème de quasi-singularité de $\mathbf{W'W}$, avec $\mathbf{W}$ matrice des variables explicatives $\mathbf{X}$ standardisées (voir donnée standardisée), on est amené à utiliser la méthode de la régression ridge, qui modifie la matrice $\mathbf{W'W}$ de manière à l'éloigner significativement d'une situation de singularité. Les données restituant une matrice $\mathbf{W'W}$ la moins singulière possible, sont les données où toutes les variables explicatives sont non corrélées entre elles.

HISTORIQUE

La régression ridge fut introduite en 1962 par A.E. Hoerl. Il avait constaté que l'existence de corrélations entre variables explicatives pouvait entraîner des erreurs dans l'estimation des paramètres lors de l'application de la méthode des moindres carrés. Comme alternative à cette méthode traditionnelle, il développa la méthode de la régression ridge. Cette méthode permet le calcul et l'utilisation d'estimateurs biaisés mais de variance plus faible que les estimateurs des moindres carrés. Alors que la méthode des moindres carrés fournit des estimateurs non biaisés, la méthode ridge minimise l'**erreur quadratique moyenne** des estimateurs.

ASPECTS MATHÉMATIQUES

Si on considère le modèle de la **régression linéaire multiple**

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

où

- $\mathbf{Y} = (y_1, \ldots, y_n)'$ est le vecteur $(n \times 1)$ des observations relatives à la **variable dépendante** (n observations),
- $\boldsymbol{\beta} = \left(\beta_0, \ldots, \beta_{p-1}\right)'$ est le vecteur $(p \times 1)$ des paramètres à estimer,
- $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)'$ est le vecteur $(n \times 1)$ des erreurs

et $\mathbf{X} = \begin{pmatrix} 1 & X_{11} & \ldots & X_{1(p-1)} \\ \vdots & \vdots & & \vdots \\ 1 & X_{n1} & \ldots & X_{n(p-1)} \end{pmatrix}$ est la **matrice** $(n \times p)$ ayant trait aux variables indépendantes,

l'**estimation** de $\boldsymbol{\beta}$ par la méthode des **moindres carrés** est donnée par :

$$\widehat{\boldsymbol{\beta}} = \left(\widehat{\beta}_0, \ldots, \widehat{\beta}_{p-1}\right)' = \left(\mathbf{X'X}\right)^{-1}\mathbf{X'Y}.$$

Si l'on considère la **matrice** des données standardisées :

$$\mathbf{W} = \begin{pmatrix} X^s_{11} & X^s_{12} & \cdots & X^s_{1p-1} \\ X^s_{21} & X^s_{22} & \cdots & X^s_{2p-1} \\ \vdots & \vdots & \ddots & \vdots \\ X^s_{n1} & X^s_{n2} & \cdots & X^s_{np-1} \end{pmatrix},$$

alors le modèle devient :

$$\mathbf{Y} = \boldsymbol{\mu} + \mathbf{W}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}$$

où

$$\begin{aligned} \boldsymbol{\mu} &= (\gamma_0, \ldots, \gamma_0)' = (\beta^s_0, \ldots, \beta^s_0)' \\ \boldsymbol{\gamma} &= \left(\gamma_1, \ldots, \gamma_{p-1}\right)' = \left(\beta^s_1, \ldots, \beta^s_{p-1}\right)'. \end{aligned}$$

Les estimateurs des moindres carrées sont donnés par $\widehat{\boldsymbol{\gamma}}_{\mathbf{0}} = \widehat{\boldsymbol{\beta}}^{\boldsymbol{s}}_{\mathbf{0}} = \overline{Y}$ et par :

$$\begin{aligned} \widehat{\boldsymbol{\gamma}} &= \left(\widehat{\gamma}_1, \ldots, \widehat{\gamma}_{p-1}\right)' \\ &= \left(\widehat{\beta}^s_1, \ldots, \widehat{\beta}^s_{p-1}\right)' \\ &= (\mathbf{W'W})^{-1}\mathbf{W'Y}. \end{aligned}$$

Ces estimateurs sont non biaisés, c'est-à-dire que l'on a :

$$\begin{aligned} E\left(\widehat{\beta}_j\right) &= \beta_j, \quad j = 0, \ldots, p-1, \\ E\left(\widehat{\gamma}_j\right) &= \gamma_j, \quad j = 0, \ldots, p-1. \end{aligned}$$

Parmi les estimateurs linéaires non biaisés, il s'agit des estimateurs qui ont la **variance** minimale. On sait que

$$\begin{aligned} Var\left(\widehat{\boldsymbol{\beta}}\right) &= \sigma^2\left(\mathbf{X}'\mathbf{X}\right)^{-1} \\ Var\left(\widehat{\boldsymbol{\gamma}}\right) &= \sigma^2\left(\mathbf{W}'\mathbf{W}\right)^{-1} \end{aligned}$$

où σ^2, est la variance des ε_i. On remarque que $\mathbf{W}'\mathbf{W}$ est égale à $(n-1)$ fois la matrice de corrélation. Par conséquent, le cas le plus favorable — ne présentant aucune colinéarité — survient lorsque $\mathbf{W}'\mathbf{W}$ est un multiple de $\mathbf{I}$. Afin que $\mathbf{W}'\mathbf{W}$ soit plus proche du cas le plus favorable, on remplace $\mathbf{W}'\mathbf{W}$ par $\mathbf{W}'\mathbf{W}+k\mathbf{I}$ où k est un nombre positif.

La régression ridge remplace l'estimateur des moindres carrés

$$\widehat{\boldsymbol{\gamma}} = \left(\mathbf{W}'\mathbf{W}\right)^{-1}\mathbf{W}'\mathbf{Y}$$

par l'estimateur ridge

$$\begin{aligned} \widehat{\boldsymbol{\gamma}}_R &= \left(\mathbf{W}'\mathbf{W}+k\mathbf{I}\right)^{-1}\mathbf{W}'\mathbf{Y} \\ &= \left(\widehat{\gamma}_{R_1},\ldots,\widehat{\gamma}_{R_{(p-1)}}\right). \end{aligned}$$

Hoerl et Kennard ont montré, en 1970, qu'il existe toujours une **valeur** de k pour laquelle l'**erreur quadratique moyenne totale** du vecteur des estimateurs ridge est plus petite que l'erreur quadratique moyenne totale du vecteur des estimateurs des moindres carrés. En 1975, Hoerl, Kennard et Baldwin ont suggéré que $k = \frac{(p-1)\sigma^2}{\boldsymbol{\gamma}'\boldsymbol{\gamma}}$ donne vraisemblablement une erreur quadratique moyenne totale plus petite, où $(p-1)$ est le nombre de variables explicatives. Pour le calcul de k, il faut estimer les valeurs de σ et $\boldsymbol{\gamma}$ par s et $\widehat{\boldsymbol{\gamma}}$. Ensuite $\boldsymbol{\gamma}$ est ré-estimé par $\widehat{\boldsymbol{\gamma}}_R$ en utilisant la régression ridge avec $k = \frac{(p-1)s^2}{\widehat{\boldsymbol{\gamma}}'\widehat{\boldsymbol{\gamma}}}$. On remarque cependant que le k utilisé dans cette procédure est une variable aléatoire, puisqu'il s'agit d'une fonction des données.

DOMAINES ET LIMITATIONS

Propriétés des estimateurs ridge

Les estimateurs ridge sont biaisés car les espérances $E\left(\widehat{\boldsymbol{\gamma}}_{R_j}\right)$ des estimateurs $\widehat{\boldsymbol{\gamma}}_{Rj}$ se trouvent dans le **vecteur**

$$E\left(\widehat{\boldsymbol{\gamma}}_{R_j}\right) = \left(\mathbf{W}'\mathbf{W}+k\mathbf{I}\right)^{-1}\mathbf{W}'\mathbf{W}\boldsymbol{\gamma} \neq \boldsymbol{\gamma}, \forall k > 0.$$

On note que la valeur $k = 0$ correspond aux estimateurs des moindres carrés. En soit, le biais est un inconvénient, mais en acceptant un biais dans les estimateurs de ridge, on est capable d'obtenir des variances plus faibles. On a, en effet :

$$Var\left(\widehat{\boldsymbol{\gamma}}_R\right) = \sigma^2\mathbf{V}_R$$

avec :

$$V_R = \left(\mathbf{W}'\mathbf{W}+k\mathbf{I}\right)^{-1}\mathbf{W}'\mathbf{W}\left(\mathbf{W}'\mathbf{W}+k\mathbf{I}\right)^{-1}.$$

On peut montrer que les éléments diagonaux de V_R sont plus petits que les éléments diagonaux correspondants de $\left(\mathbf{W}'\mathbf{W}\right)^{-1}$, ce qui implique que les estimateurs ridge ont des variances plus faibles que les estimateurs des moindres carrés. On peut le voir facilement dans le cas de la régression simple car on a alors :

$$\mathbf{V}_R = \frac{n-1}{n-1+k^2} \leq \frac{1}{n-1} = \left(\mathbf{W}'\mathbf{W}\right)^{-1}, \forall k > 0.$$

Ainsi, les estimateurs ridge ont des variances plus petites que les estimateurs des moindres carrées. Est-ce que leurs biais sont acceptables ? Pour mesurer un compromis entre le biais et la variance d'un estimateur, on calcule généralement son erreur quadratique moyenne (en anglais *mean squared error*) symbolisé par MSE. L'erreur quadratique moyenne d'un estimateur $\widehat{\theta}$ de θ est définie par :

$$\begin{aligned} MSE\left(\widehat{\theta}\right) &= E\left(\left(\widehat{\theta}-\theta\right)^2\right) \\ &= Var\left(\widehat{\theta}\right) + \left(E\left(\widehat{\theta}\right)-\theta\right)^2. \end{aligned}$$

Par ailleurs, on définit l'**erreur quadratique moyenne totale** (en anglais *total mean squared error*, TMSE) d'un vecteur d'estimateurs par la somme des erreurs quadratiques moyennes de ses composantes. On a ainsi :

$$\begin{aligned} TMSE\left(\widehat{\boldsymbol{\gamma}}_R\right) &= \sum_{j=1}^{p-1} MSE\left(\widehat{\boldsymbol{\gamma}}_{Rj}\right) \\ &= \sum_{j=1}^{p-1}\left(E\left(\widehat{\boldsymbol{\gamma}}_{Rj}-\boldsymbol{\gamma}_j\right)^2\right) \end{aligned}$$

$$
\begin{aligned}
&= \sum_{j=1}^{p-1} \left[Var\left(\widehat{\boldsymbol{\gamma}}_{Rj}\right) + \left(E\left(\widehat{\boldsymbol{\gamma}}_{Rj}\right) - \boldsymbol{\gamma}_j\right)^2 \right] \\
&= (p-1)\,\sigma^2 Trace\left(V_R\right) \\
&\quad + \sum_{j=1}^{p-1} \left(E\left(\widehat{\boldsymbol{\gamma}}_{Rj}\right) - \boldsymbol{\gamma}_j\right)^2 .
\end{aligned}
$$

Cette méthode s'appelle régression ridge adaptative. Les propriétés des estimateurs ridge ne sont vraies que dans le cas où k est une constante et donc ne s'appliquent pas aux estimateurs ridge adaptatifs. Cependant, les propriétés des estimateurs ridge suggèrent que les estimateurs ridge adaptatifs sont de bons estimateurs. En effet, des études statistiques ont trouvé que pour beaucoup d'ensembles de données, une estimation ridge adaptative est meilleure au sens de l'erreur quadratique moyenne qu'une estimation par les moindres carrés.

EXEMPLE

Dans l'exemple suivant, les observations sont relatives à treize fournées de ciment. Chaque fournée est composée de quatre ingrédients cités dans le tableau. Le but de l'expérience est de déterminer comment les quantités x_{i1}, x_{i2}, x_{i3} et x_{i4} de ces quatre ingrédients affectent la quantité y_i de chaleur émise par le durcissement du ciment.

Chaleur émise par le ciment

Fournée	Ingrédient				Chaleur
	1	2	3	4	
i	X_{i1}	X_{i2}	X_{i3}	X_{i4}	Y_i
1	7	26	6	60	78,5
2	1	29	15	52	74,3
3	11	56	8	20	104,3
4	11	31	8	47	87,6
5	7	52	6	33	95,9
6	11	55	9	22	109,2
7	3	71	17	6	102,7
8	1	31	22	44	72,5
9	2	54	18	22	93,1
10	21	47	4	26	115,9
11	1	40	23	34	83,9
12	11	66	9	12	113,3
13	10	68	8	12	109,4

où

- Y_i quantité de chaleur émise par le durcissement de la i-ème fournée (en joules) ;
- X_{i1} quantité de l'ingrédient 1 (aluminate de tricalcium) dans la i-ème fournée ;
- X_{i2} quantité de l'ingrédient 2 (silicate de tricalcium) dans la i-ème fournée ;
- X_{i3} quantité de l'ingrédient 3 (alumino-ferrite de tetracalcium) dans la i-ème fournée ;
- X_{i4} quantité de l'ingrédient 4 (silicate de dicalcium) dans la i-ème fournée.

On effectue tout d'abord une régression linéaire multiple. Le modèle de la régression linéaire simple s'écrit

$$y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + \varepsilon_i.$$

On obtient les résultats suivants :

Variable	Coefficient	Écart type	t_c
Constante	62,4100	70,0700	0,89
X_1	1,5511	0,7448	2,08
X_2	0,5102	0,7238	0,70
X_3	0,1019	0,7547	0,14
X_4	−0,1441	0,7091	−0,20

Le vecteur d'estimateurs obtenu à partir des **données standardisées** est :

$$\widehat{Y} = (9,12\,;\, 7,94\,;\, 0,65\,;\, -2,41)'$$

avec des écarts types valant respectivement 4,381 ; 11,26 ; 4,834 et 11,87. Par ailleurs, l'estimateur s de σ est $s = 2,446$. Si on procède à une analyse de type ridge en utilisant :

$$
\begin{aligned}
k &= \frac{(p-1)\,\sigma^2}{\boldsymbol{\gamma}'\boldsymbol{\gamma}} \\
&= \frac{4 \cdot (2,446)^2}{(9,12)^2 + (7,94)^2 + (0,65)^2 + (2,41)^2} \\
&= 0,157,
\end{aligned}
$$

il s'agit d'une régression ridge adaptative car k dépend des données. On utilise cependant le terme « régression ridge » également pour une régression ridge adaptative. On obtient :

$$\widehat{Y}_R = (7,64\,;\, 4,67\,;\, -0,91\,;\, -5.84)'$$

et les écarts types valant respectivement $1,189$; $1,377$; $1,169$ et $1,391$. On a donc les valeurs estimées :

$$\widehat{Y}_i = 95,4 + 7,64X^s_{i1} + 4,67X^s_{i2} - 0,91X^s_{i3} - 5,84X^s_{i4},$$

l'estimation de la constante ne différant pas de celle obtenue par la méthode des **moindres carrés**.

MOTS CLÉS

Colinéarité *(Collinearity)*
Données standardisées *(Standardised data)*
Erreur quadratique moyenne *(Mean square error)*
Simulation *(Simulation)*

RÉFÉRENCES

Hoerl A.E. (1962). Application of Ridge Analysis to Regression Problems. *Chemical Engineering progress*, vol. 58(3), pp. 54-59.

Hoerl A.E. and Kennard R.W. (1970). Ridge Regression: Biased estimation for nonorthogonal Problems. *Technometrics*, 12, pp. 55-67.

Hoerl A.E., Kennard R.W. and Baldwin K.F. (1975). Ridge Regression: Some Simulations. *Communications in Statistics*, vol. 4, pp. 105-123.

RELATION

Relationship

La notion de relation exprime le lien qui existe entre couples d'éléments. C'est une des notions les plus importantes en **statistique**. Elle apparaît dans des concepts tels que la **classification**, la **corrélation**, la **dépendance**, les **modèles** mathématiques, l'**analyse de régression**, les **séries chronologiques**, etc.

MOTS CLÉS

Analyse de régression *(Regression analysis)*
Classification *(Classification)*
Corrélation *(Correlation)*
Dépendance *(Dependence)*
Modèle *(Model)*
Série chronologique *(Time series)*

REPRÉSENTATION GRAPHIQUE

Graphical representation

La représentation graphique comporte un ensemble complexe de techniques pour élucider, interpréter et analyser des données au moyen de points, de segments de droites, d'aires et d'autres formes géométriques et symboles.

Le graphique sert à fournir une image permettant de visualiser rapidement l'allure générale du phénomène étudié et de mettre en évidence certains faits essentiels. Il sert par exemple à traduire ou à compléter un **tableau de fréquences**.

La représentation graphique est donc une forme de présentation des données.

HISTORIQUE

L'idée de déterminer la position d'un point dans l'espace à l'aide de coordonnées remonte au moins à l'époque de la Grèce antique, mais il faudra attendre celle de René Descartes, pour voir les mathématiciens développer ce concept.

Selon E. Royston (1970), un mathématicien allemand du nom de A.W. Crome fut parmi les premiers à utiliser des représentations graphiques en **statistique** dont il se servit d'abord comme outil pédagogique.

Dans ses ouvrages *Geographisch-statistische Darstellung der Staatskräfte* (1820) et *Ueber die Grösse und Bevölkerung der sämtlichen Europäischen Staaten* (1785), A.W. Crome employa différents systèmes de représentations graphiques, dont le **diagramme circulaire**.

E. Royston (1970) cite également W. Playfair dont l'ouvrage *The Commercial and Political Atlas* (1786) fait aussi appel à des représentations graphiques et notamment au **diagramme en bâtons**. Très intéressé par la balance du commerce interna-

tional, W. Playfair illustra ses études par différents graphiques tels l'**histogramme** et le diagramme circulaire.

Le terme histogramme fut employé pour la première fois par **Karl Pearson**, et Guerry (1833) semble figurer parmi les premiers utilisateurs du diagramme en bâtons ; dans son *Essai sur la statistique morale de la France* publié à Paris en 1833, celui-ci décrivait les fréquences des crimes par certaines caractéristiques telles que l'âge ; cette étude constituerait l'un des premiers usages de distributions de fréquences.

Calvin F. Schmid (1954) présente et décrit les différents types de représentations graphiques dans son *Handbook of Graphic Presentation*.

ASPECTS MATHÉMATIQUES

Selon le type de **données**, la nature de la **variable** étudiée et le but de l'étude, on utilise différents graphiques :

- **graphique quantitatif** permettant notamment de représenter des variables qualitatives catégorielles, comme, par exemple, le **diagramme circulaire**, le **diagramme en barres** ou le **diagramme figuratif** ;

- graphique de fréquences permettant de représenter la **distribution de fréquences** d'une **variable quantitative**, discrète ou continue ; comme, par exemple, l'**histogramme**, le **diagramme stem and leaf** ;

- graphique cartésien, utilisant un système d'axes de coordonnées cartésiennes rectangulaires, c'est-à-dire constitué par deux droites perpendiculaires se coupant à l'origine.

  En fonction de l'échelle utilisée sur les **axes**, on trouve les graphiques suivants :

  - graphique arithmétique ;
  - **graphique semi-logarithmique** ;
  - **graphique logarithmique**.

MOTS CLÉS

Graphique quantitatif *(Quantitative graph)*
Tableau de fréquences *(Frequency table)*

RÉFÉRENCES

Crome, A.F.W. (1785). *Ueber die Grösse und Bevölkerung der sämtlichen Europäischen Staaten.* Leipzig.

Crome, A.F.W. (1820). *Geographisch-statistische Darstellung der Staatskräfte.* Leipzig.

Fienberg, S.E. (1979). Graphical method in statistics. *The American Statistician*, 33, pp. 165-178.

Guerry, A.M. (1833). *Essai sur la statistique morale de la France.* Crochard, Paris.

Playfair, W. (1786). *The Commercial and Political Atlas.* London.

Royston, E. (1970). A note on the history of the graphical presentation of data. *Studies in the History of Statistics and Probability.* Vol. I. Edited by E.S. Pearson and M. Kendall. Charles Griffin & Company Ltd.

Schmid, C.F. (1954). *Handbook of Graphic Presentation*, Ronald Press Co, New York.

RÉSIDU

Residual

Les résidus sont définis comme les différences entre les valeurs observées et les valeurs estimées par un **modèle** de régression.

Ils représentent donc la partie non expliquée par l'équation de régression.

HISTORIQUE

Voir **erreur**.

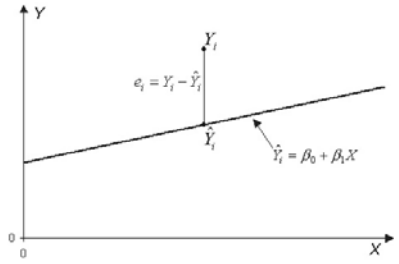
ASPECTS MATHÉMATIQUES

Les résidus e_i sont donnés par :

$$e_i = Y_i - \hat{Y}_i, \quad i = 1, \ldots, n$$

où Y_i est une **observation** et $\hat{Y}_i$ est la **valeur** estimée correspondante, obtenue par l'ajustement d'un **modèle** de régression.

Le graphique suivant donne une représentation des résidus dans le cas d'une **régression linéaire simple** :



EXEMPLES

Voir **analyse des résidus**.

MOTS CLÉS

Analyse de régression *(Regression analysis)*
Analyse des résidus *(Analysis of residuals)*
Erreur *(Error)*
Matrice H *(Hat matrix)*
Point levier *(Leverage point)*
Régression linéaire multiple *(Multiple linear regression)*
Régression linéaire simple *(Simple linear regression)*

RICE STUART ARTHUR

Rice Stuart Arthur

Stuart Arthur Rice (1889-1969) effectua ses études à l'Université de Washington (il y obtint sa licence ès sciences en 1912 et son postgrade en 1915). Son travail académique concerna tout particulièrement le développement d'une méthodologie en science sociale, dans l'étude de la politique.

Au début des années 1930, Rice entra dans l'administration de Franklin Delano Roosevelt. Il devint directeur assistant du Bureau du recensement de 1933 jusque dans les années 1950. Sa contribution majeure en statistique consista en la modernisation de l'utilisation de la statistique dans le gouvernement. Il favorisa le développement du COGSIS (COmmittee on Government Statistics and Information Services) et travailla dans le sens d'amener des techniques modernes d'enquêtes par échantillonnage ainsi que les mathématiques statistiques aux Federal agencies.



En 1955, il créa Stuart A. Rice Associates (plus tard devenu Surveys & Research Corporation), une agence de consultation statistique pour les gouvernements et des agences privées à but lucratif ou non. Il prit sa retraite au milieu des années 1960.

MOTS CLÉS

Recensement *(Census)*
Statistique officielle *(Offical statistics)*

RISQUE

Risk

En épidémiologie, le risque est la probabilité d'apparition de nouveaux cas de maladie dans une population susceptible de développer cette maladie, au cours d'un intervalle de temps donné. Le risque peut être défini pour la population générale ou pour un sous-groupe homogène d'individus du point de vue de l'exposition à une cause supposée de la maladie étudiée. La notion de risque est par conséquent intimement liée à celle d'**incidence** (passer de l'incidence au risque, c'est aller de la fréquence à la probabilité).

HISTORIQUE

Jerome Cornfield, en 1951, mit clairement en évidence et pour la première fois la notion de risque, suite à une étude au cours de laquelle les facteurs de risque d'une population malade ont été comparés à ceux d'une population saine. Depuis lors, cette statistique à été largement utilisée au cours les investigations épidémiologiques, comme une mesure du lien entre une maladie et les facteurs de risque y prédisposant.

ASPECTS MATHÉMATIQUES

Le risque par rapport à un certain facteur est la probabilité d'occurrence de cas incidents au cours d'une période de temps fixée, calculée parmi l'ensemble des individus exposés au facteur étudié. Pratiquement, on estime ce risque par la proportion :

$$risque = \frac{\text{cas incidents (parmi la population exposée)}}{\text{population à risque}}$$

au cours de la période donnée. Autrement dit, le numérateur est le nombre de nouveaux cas (ou cas incidents) durant la période fixée dans la population étudiée et le dénominateur est le nombre de personnes à risque au début de la période de suivi.

DOMAINES ET LIMITATIONS

Intuitivement, la notion de risque est de nature probabiliste, quand bien même elle fait référence à une population et une période bien déterminée. Les compagnies d'assurance y recourent pour prédire la fréquence de décès ou de maladie dans une population définie, par exemple selon le sexe, l'âge, la taille, le poids... Le risque peut aussi être interprété comme la probabilité pour un individu donné de contracter une affection, sachant qu'il est ou non exposé à un facteur. On applique à un individu le risque moyen de son groupe. Les compagnies d'assurance s'en servent pour déterminer les cotisations de chaque assuré, le médecin pour estimer le risque (ou probabilité *a priori*) qu'a son patient de développer une certaine maladie, de faire une complication ou de guérir.

Il est important de noter que la notion de risque est toujours définie par rapport à une durée : le risque sur deux, trois, dix ans.

Le risque minimal est de 0 %, le risque maximal de 100.

EXEMPLES

Exemple 1 : on suppose que le 1er janvier 1992 un groupe de 1 000 hommes âgés de 60 ans soit sélectionné et que l'on observe 10 cas d'infarctus au cours de l'année qui suit. Le risque d'infarctus dans cette population d'hommes de 60 ans est donc de $10/1\,000 = 1$ % sur un an.

Exemple 2 : on prend le cas du risque de cancer du sein sur deux ans dans une population de 100 000 femmes, réparties entre 3 groupes de niveaux d'exposition au facteur « tabagisme ». La table de contingence suivante illustre cet exemple :

Groupe	Nombre de cas incidents	Effectifs du groupe
Non exposées	40	35 100
Fumeuses passives	140	55 500
Fumeuses actives	164	59 400
Total	344	150 000

Le risque pour chaque groupe est calculé en faisant le rapport entre le nombre des cas incidents et les effectifs. Ainsi, le risque des fumeuses passives est de $140/55\,500 = 0,002523 = 252,3/100\,000$ sur deux ans. On a le tableau suivant :

Groupe	Nombre de cas incidents	Effectifs du groupe	Risque sur deux ans (/100 000)
Non exposées	40	35 100	114, 0
Fumeuses passives	140	55 500	252, 3
Fumeuses actives	164	59 400	276, 1
Total	344	150 000	229, 3

La notion de risque implique celle de prédiction : parmi les femmes d'une population donnée, quelle proportion va développer un cancer du sein au cours des deux ans à venir ? On infère du tableau précédent que $229,3/100\,000$ peuvent développer cette maladie dans les deux ans à venir.

MOTS CLÉS

Cause et effet en épidémiologie *(Cause and effect in epidemiology)*
Odds et odds ratio *(Odds and odds ratio)*
Risque attribuable *(Attribuable risk)*
Risque évitable *(Avoidable risk)*
Risque relatif *(Relative risk)*
Taux d'incidence *(Incidence rate)*
Taux de prévalence *(Prevalence rate)*

RÉFÉRENCES

Cornfield, J. (1951). A method of estimating comparative rates from clinical data. Applications to cancer of the lung, breast, and cervix. *J Natl Cancer Inst*, vol. 11, pp. 1269-75.

Lilienfeld A.M. and D.E. Lilienfeld. (1980). *Foundations of epidemiology.* 2nd ed New York: Clarendon Press, Oxford.

MacMahon, B. and Pugh, T.F. (1970). *Epidemiology: Principles and Methods.* Boston: Little Brown.

Morabia A. (1996). *Épidémiologie causale.* Éditions Médecine et Hygiène, Genève.

Morabia A. (1996). *L'Épidémiologie clinique.* Presses universitaires de France, Que sais-je ? Paris.

RISQUE ATTRIBUABLE

Attributable Risk

Le **risque attribuable** à un facteur d'exposition est la différence par soustraction entre le **risque** chez les individus exposés à ce facteur et le risque chez les individus qui n'y sont pas exposés. C'est l'opposé du **risque évitable**. Il mesure l'effet absolu d'une cause (c'est-à-dire l'excès de **risque** ou de cas de maladie).

HISTORIQUE

Voir sous **risque**.

ASPECTS MATHÉMATIQUES

On a par définition :

$$\text{risque attribuable} = \text{risque des exposés} - \text{risque des non exposés.}$$

DOMAINES ET LIMITATIONS

On peut, pour un risque attribuable, calculer un **intervalle de confiance** à $c\,\% = (1-\alpha)$.

L'intervalle de confiance à $c\,\% = (1-\alpha)$ d'un risque attribuable est équivalent à l'intervalle de confiance de la différence entre les proportions p_E et p_{NE} où p_E et p_{NE} représentent respectivement le risque des individus exposés et des individus non exposés au facteur étudié. Si n_E et n_{NE} sont respectivement les effectifs des populations exposée et non exposée, alors l'intervalle de confiance admet pour bornes, pour un degré de confiance de $c\,\% = (1-\alpha)$, avec z_α la **valeur** obtenue par la **table normale** (par exemple, pour un intervalle de confiance à 95 %, alors $\alpha = 0,05$ et $z_\alpha = 1,96$), les valeurs :

$$(p_E - p_{NE}) \pm z_\alpha \sqrt{\frac{p_E \cdot (1-p_E)}{n_E} \cdot \frac{p_{NE} \cdot (1-p_{NE})}{n_{NE}}}$$

et pour l'intervalle de confiance à $c\,\% = (1-\alpha)$ d'un risque évitable admet pour bornes :

$$(p_{NE} - p_E) \mp z_\alpha \sqrt{\frac{p_E \cdot (1-p_E)}{n_E} \cdot \frac{p_{NE} \cdot (1-p_{NE})}{n_{NE}}}.$$

Il est requis que n_E et n_{NE} soient grands. Si l'intervalle de confiance inclut la valeur 0, il n'est pas permis de rejeter la nullité du risque attribuable au facteur d'exposition.

EXEMPLES

On prend le cas de l'étude du risque de cancer du sein attribuable au tabagisme chez la femme :

Groupe	Taux d'incidence (/100 000 /an)	Risque attribuable au tabagisme (A) (/100 000 /an)
Non-exposées	57,0	$57,0 - 57,0 = 0$
Fumeuses passives	126,2	$126,2 - 57,0 = 69,2$
Fumeuses actives	138,1	$138,1 - 57,0 = 81,1$
Total	114,7	$114,7 - 57,0 = 57,7$

Les risques attribuables au tabagisme passif et au tabagisme actif sont respectivement de 69 et de 81 (/100 000 /an). En d'autres termes, si l'exposition au tabac était supprimée, le taux d'incidence des fumeuses actives (138/100 000 par an) pourrait être réduit de 81/100 000 par an et celui des fumeuses passives (126/100 000 par an) de 69/100 000 par an. Les taux d'incidence dans ces deux catégories de fumeuses deviendraient égaux à celui des femmes non exposées (57/100 000 par an). On remarque alors que le taux d'incidence du groupe des femmes non exposées n'étant pas nul, cet indice représente l'influence d'autres facteurs que le tabagisme.

Groupe	Nb d'ind. suivis sur 2 ans	Cas att. au tabagisme (par 2 ans)	Cas att. au tabagisme (par an)
Non-exposées	70 160	0,0	0,0
Fum. passives	110 860	76,7	38,4
Fum. actives	118 636	96,2	48,1
Total	299 656	172,9	86,5

On peut directement calculer le nombre de cas de cancer du sein attribuables au tabac en multipliant le nombre d'individus suivis par année par le risque attribuable. Une fois obtenu le nombre de cas incidents attribuables au tabagisme sur deux ans, on calcule, en le divisant par deux, le nombre de cas attribuables au tabagisme sur un an et on peut déterminer le risque attribuable au tabagisme dans la population, noté RAP, comme montré dans l'exemple suivant. Le tableau ci-dessus montre les détails du calcul.

On précise le calcul pour les fumeuses passives : dans l'étude qui s'est déroulée sur deux ans, on dénombrait 110 860 fumeuses passives. Le risque attribuable au tabagisme passif était de 69,2/100 000 par an. Il résulte que le nombre des cas attribuables au tabagisme sur deux ans est de $(110\,860 \cdot 69,2)\,/\,100\,000 = 76,7$. Si on veut calculer le nombre de cas attribuables au tabagisme passif sur un an, il faut alors diviser cette dernière valeur par 2 et l'on obtient finalement 38,4. En outre, on peut calculer le risque attribuable au tabagisme sur un an en divisant simplement le nombre de cas attribuables au tabagisme sur deux ans (172,9) par le nombre des individus suivis au cours de l'étude durant deux ans (299 656 personnes). On obtient alors une mesure du risque attribuable au tabagisme de 57,7/100 000 par an. On note que l'on aurait obtenu le même résultat en prenant la différence entre le taux d'incidence total (114,7/100 000 par an, voir l'exemple traité sous les mots taux d'incidence, taux de prévalence) et le taux d'incidence du groupe des non exposés (57,0/100 000 par an).

Le risque du cancer du sein attribuable au tabagisme dans la population (RAP) est le rapport entre le

nombre de cas de cancer du sein attribuables à l'exposition au tabac et le nombre de cas de cancer du sein diagnostiqués dans la population (voir tableau ci-dessous). Le risque attribuable, dans la population, est de 22.3 % (38, 4/172) pour le tabagisme passif et de 28 % (48, 1/172) pour le tabagisme actif. Pour les deux formes d'exposition, il est de 50, 3 % (22, 3 % + 28 %). Ainsi, la moitié des cas de cancer du sein diagnostiqués chaque année dans cette population serait attribuable au tabagisme (actif ou passif).

Groupe	Cas att. au tabagisme (par 2 ans)	Cas att. au tabagisme (par an)	RAP
Non-exposées	0, 0	20	0, 0
Fumeuses passives	38, 4	70	22, 3
Fumeuses actives	48, 1	82	28, 0
Total	86, 5	172	50, 3

MOTS CLÉS

Cause et effet en épidémiologie *(Cause and effect in epidemiology)*
Odds et odds ratio *(Odds and odds ratio)*
Risque *(Risk)*
Risque évitable *(Avoidable risk)*
Risque relatif *(Relative risk)*
Taux d'incidence *(Incidence rate)*
Taux de prévalence *(Prevalence rate)*

RÉFÉRENCES

Voir **risque**.

RISQUE ÉVITABLE

Avoidable Risk

Le risque évitable (sous-entendu, évitable si l'on neutralisait l'effet du facteur d'exposition sous-jacent) est l'opposé du **risque attribuable**, c'est-à-dire la différence entre le **risque** chez les individus non exposés au facteur et le risque chez les individus exposés à ce facteur.

HISTORIQUE

Voir **risque**.

ASPECTS MATHÉMATIQUES

On a par définition :

$$\begin{aligned}\text{risque évitable} &= \text{risque des non exposés} \\ &\quad -\text{risque des exposés.}\end{aligned}$$

DOMAINES ET LIMITATIONS

Le risque évitable a été introduit pour parler de **risque attribuable** négatif, ce qui est relativement inconfortable. Il permet de calculer le nombre de patients à traiter, car :

$$\text{nombre de patients à traiter} = \frac{1}{\text{risque évitable}}.$$

Voir aussi risque attribuable.

EXEMPLES

On prend comme exemple une étude dont l'objectif est de déterminer l'efficacité d'un médicament pour le traitement d'une maladie.

Les 223 patients inclus dans l'étude sont tous des patients à risque, mais il n'ont encore jamais contracté la maladie en question. On les sépare en deux groupes, le premier, formé de 114 personnes, recevant le médicament et le second, composé de 109 patients, un placebo. La période de suivi est de deux ans. Au total, 11 cas sont diagnostiqués dans le premier groupe et 27 dans le groupe placebo.

Groupe	Cas incidents (A)	Effectif du groupe (B)	Risque sur deux ans ($\frac{A}{B}$ en %)
1er groupe	11	114	9, 6 %
2e groupe	27	109	24, 8 %

Ainsi, le risque évitable grâce à ce médicament est de $24, 8 - 9, 6 = 15, 2$ % pour deux ans.

MOTS CLÉS

Cause et effet en épidémiologie *(Cause and effect in epidemiology)*
Odds et odds ratio *(Odds and odds ratio)*
Risque *(Risk)*

Risque attribuable *(Attribuable risk)*
Risque relatif *(Relative risk)*
Taux d'incidence *(Incidence rate)*
Taux de prévalence *(Prevalence rate)*

RÉFÉRENCES

Voir **risque**.

RISQUE RELATIF

Relative risk

Le **risque relatif** est défini comme étant le rapport entre le **risque** concernant des individus exposés à un facteur et le risque afférent aux non exposés à ce facteur. Il mesure l'effet relatif d'une cause (c'est-à-dire la force d'association entre la maladie et la cause).

HISORIQUE

Voir **risque**.

ASPECTS MATHÉMATIQUES

Le risque relatif est, formellement, un rapport de probabilité, que l'on estime par la quantité :

$$\begin{aligned}\text{risque relatif} &= \frac{\text{risque des exposés}}{\text{risque des non exposés}} \\ &= \frac{\text{taux d'incidence des exposés}}{\text{taux d'incidence des non exposés}}.\end{aligned}$$

On notera que le risque relatif est une grandeur sans unité, puisqu'étant définie comme le rapport de deux quantités exprimées dans la même unité.

DOMAINES ET LIMITATIONS

Le risque relatif exprime le risque des exposés comme un multiple du risque des non exposés :

$$\begin{aligned}\text{risque des exposés} &= \text{risque des non exposés} \\ &\quad \cdot \text{ risque relatif.}\end{aligned}$$

On désire le étudier le lien qui unit un facteur d'exposition à une maladie. En l'absence d'une telle association, le risque relatif est théoriquement égal à 1. Si le facteur d'exposition accroît la probabilité d'occurrence de la maladie, alors le risque relatif devient supérieur à 1. Si, en revanche, le facteur d'exposition minore la probabilité d'occurrence de la maladie, alors le risque relatif est compris entre 0 et 1 (strictement).

Il est tout à fait possible de calculer un **intervalle de confiance** à $P = (1 - \alpha)$ % pour un risque relatif. On considère à cet effet le cas général suivant :

L'intervalle de confiance est calculé à l'échelle logarithmique puis transformé à l'échelle arithmétique. Les lettres minuscules se rapportent au tableau suivant :

	Malades	Non malades	Total
Exposés	a	$n - a$	n
Non exposés	c	$m - c$	m

L'intervalle de confiance à $P = (1 - \alpha)$ % du risque relatif, RR, est obtenu ainsi :

$$\exp\left\{\ln(RR) \pm z_\alpha \sqrt{\frac{1}{a} - \frac{1}{n} + \frac{1}{c} - \frac{1}{m}}\right\}$$

où « exp » représente l'exponentielle de l'expression entre accolades et z_α la **valeur** de la **table normale**. Il est requis que a, n, c et m soient grands. Si l'intervalle de confiance du risque relatif inclut la valeur $1,0$, alors on conclut que le facteur d'exposition considéré ne joue pas de rôle statistiquement significatif quant à la probabilité d'occurrence de la maladie étudiée.

EXEMPLE

On reprend l'exemple précédent vu sous le **risque attribuable**. Le risque relatif de cancer du sein associé à l'exposition au tabagisme correspond au **risque** (ou le **taux d'incidence**) pour chaque niveau d'exposition, divisé par le risque (ou le taux d'incidence) chez les non exposés :

Groupe	Taux d'incidence (/100 000 /an) (A)	Risque relatif ($\frac{A}{57,0}$)
Non-exposées	$57,0$	$1,0$
Fumeuses passives	$126,2$	$2,2$
Fumeuses actives	$138,1$	$2,4$

Le risque relatif est de $2,2$ pour le tabagisme passif et de $2,4$ pour le tabagisme actif. Comparé au risque des non exposées, le risque des fumeuses actives est ainsi $2,4$ fois plus élevé et celui des fumeuses passives est $2,2$ fois plus élévé. Il s'agit d'une association de force relativement modeste, reflétant le fait que le cancer du sein dépend de facteurs d'exposition étrangers au tabagisme. Le risque relatif ne donne donc qu'une vision partielle de la relation entre le tabagisme et le cancer su sein. À lui seul, il ne permet pas de réaliser que le tabagisme est probablement responsable de $50\,\%$ des cas de cancer du sein dans la population.

MOTS CLÉS

Cause et effet en épidémiologie *(Cause and effect in epidemiology)*
Odds et odds ratio *(Odds and odds ratio)*
Risque *(Risk)*
Risque attribuable *(Attribuable risk)*
Risque évitable *(Avoidable risk)*
Taux d'incidence *(Incidence rate)*
Taux de prévalence *(Prevalence rate)*

RÉFÉRENCES

Voir **risque**.



S

SÉRIE CHRONOLOGIQUE

Time series

Une série chronologique est constituée par une succession d'observations, sur un même sujet ou sur un même phénomène, régulièrement espacées dans le temps. Les séries chronologiques sont mensuelles, trimestrielles ou annuelles, parfois hebdomadaires, journalières, voire horaires (étude de trafic routier, trafic téléphonique) ou au contraire biennales, décennales (**recensement** de la **population**).

L'analyse des séries chronologiques est un outil statistique de **prévision** parmi ceux dont dispose le conjoncturiste pour planifier et faire face au changement.

Les séries chronologiques peuvent être décomposées en quatre composantes, chacune exprimant un aspect particulier du mouvement des valeurs de la série chronologique.

Ces quatres composantes sont :

- la **tendance séculaire**, qui traduit le mouvement à long terme ;
- les variations saisonnières, qui représentent des changements saisonniers ;
- les fluctuations cycliques, qui correspondent à des variations périodiques mais non saisonnières ;
- les variations irrégulières, qui sont les autres sources non aléatoires de variations de la série.

L'analyse d'une série chronologique consiste à faire une description mathématique des éléments qui la composent, c'est-à-dire à estimer séparément les quatre composantes.

HISTORIQUE

Dans un article consacré à l'image au Moyen-Âge, H.G. Funkhauser (1936) reproduisait un diagramme présentant l'inclinaison des planètes en fonction du temps. Pour Maurice George Kendall (1973), il s'agit là du plus ancien diagramme temporel connu dans le monde occidental. C'est d'ailleurs en astronomie qu'apparaissent les premières séries chronologiques constituées volontairement à des fins d'analyse.

Il est assez normal de trouver dans l'astronomie la source de l'histoire des séries chronologiques. L'**observation** des astres était fort répandue dans l'Antiquité. La Grèce notamment, a légué en ce domaine de précieux témoignages, même si le fatal incendie de la bibliothèque d'Alexandrie, en 640, détruisit de nombreux documents essentiels.

À la naissance du christianisme, le développement des sciences en Occident connut une pause qui dura près de quinze siècles. Au XVI^{e} siècle, l'astronomie fut à nouveau la source de découvertes importantes. Rappelons à ce propos le travail effectué par Tycho Brahe (1546-1601) qui constitua un ensemble de données sur le mouvement des planètes tel qu'il permit à Johannes Kepler (1571-1630) de formuler ses lois sur le sujet. Une analyse statistique des données, toute empirique qu'elle ait pu être à l'époque, est à la base des résultats de ce grand savant.

Le développement des connaissances mathématiques aux XVIII^{e} et XIX^{e} siècles a permis de dépasser la visualisation graphique pour en arriver aux premières techniques d'analyse de séries temporelles. C'est par l'approche fréquentielle que se développèrent les premiers travaux dans ce domaine.

Les trente dernières années ont été caractérisées par une diversité accrue des objectifs et de la finalité des développements de l'analyse des séries chronologiques.

L'analyse fréquentielle (encore appelée analyse harmonique) avait à l'origine pour but de mettre en évidence une ou plusieurs « composantes » cycliques d'une série chronologique, malgré l'existence d'autres éléments intervenant dans l'évolution d'un phénomène. C'est ainsi que W.M. Persons a proposé, en 1919, une décomposition d'une série chronologique en termes de composantes tendancielle (**tendance**

séculaire), cyclique **fluctuation cyclique**), saisonnière (**variation saisonnière**) et accidentelle (**variation irrégulière**). De nombreux ouvrages ont été consacrés à la détermination et à l'élimination de l'une ou l'autre d'entre elles.

La détermination de la tendance a souvent été rattachée aux problèmes de l'analyse de régression ; on développa également les recherches relatives à des courbes de tendance non linéaire, comme par exemple la courbe de Gompertz (établie en 1825 par Benjamin Gompertz) ou la courbe logistique (introduite en 1844 par Pierre François Verhulst).

Une autre approche de la tendance concerne l'utilisation des moyennes mobiles, car elles permettent d'aborder le problème de la tendance dans la mesure où elles servent à éliminer des variations irrégulières ou périodiques.

L'usage des moyennes mobiles a été développé dans deux directions : pour les problèmes de la graduation en actuariat et pour l'étude des séries chronologiques en économie. L'utilisation des moyennes mobiles s'est cependant particulièrement développée dans le cadre des méthodes de désaisonnalisation. Plus précisément la séparation des variations saisonnières des autres composantes de la série chronologique a été étudiée à l'origine par M.T. Copeland en 1915 et W.M. Persons en 1919 et reprise par F.R. Macauley en 1930.

Ce dernier introduit la méthode des rapports à la moyenne mobile encore couramment utilisée de nos jours (voir variation saisonnière).

ASPECTS MATHÉMATIQUES

Soit Y_t une série chronologique, elle peut être décomposée à l'aide de ses quatre composantes, à savoir :

- la **tendance séculaire** T_t ;
- les variations saisonnières S_t ;
- les fluctuations cycliques C_t ;
- les variations irrégulières I_t.

On suppose que les valeurs prises par la **variable aléatoire** Y_t sont déterminées par une **relation** entre les quatre composantes précédentes.

On distingue trois modèles de composition :

1. le modèle additif :

$$Y_t = T_t + C_t + S_t + I_t.$$

   Les quatre composantes sont supposées indépendantes les unes des autres ;

2. le modèle multiplicatif :

$$Y_t = T_t \cdot C_t \cdot S_t \cdot I_t.$$

   À l'aide du logarithme on peut passer du modèle multiplicatif au modèle additif ;

3. le modèle mixte :

$$Y_t = S_t + (T_t \cdot C_t \cdot I_t)$$

   ou

$$Y_t = C_t + (T_t \cdot S_t \cdot I_t)\,.$$

   Ce genre de modèle est très peu utilisé.

On choisit le modèle additif lorsque les variations saisonnières sont à peu près constantes et affectent donc la tendance indépendamment de son niveau.

Dans le cas où les variations saisonnières sont d'amplitudes à peu près proportionnelles à la tendance séculaire, on choisit le modèle multiplicatif ; c'est généralement le cas.

L'analyse d'une série chronologique consiste à déterminer les valeurs prises par chacune des quatre composantes. On commence toujours par évaluer la tendance séculaire, puis on estime les variations saisonnières et finalement les fluctuations cycliques. Toutes les fluctuations qui ne peuvent être attribuées à l'une des trois composantes précédentes sont allouées aux variations irrégulières.

À partir du modèle utilisé, l'analyste peut, après avoir déterminé une composante, ajuster les données suivant celle-ci par soustraction ou par division. Ainsi, lorsqu'il a estimé la tendance séculaire T_t, en

chaque temps t, il ajuste la série chronologique de la façon suivante :

— s'il utilise le **modèle** additif $Y_t = T_t + C_t + S_t + I_t$:

$$Y_t - T_t = C_t + S_t + I_t \ ;$$

— pour le modèle multiplicatif $Y_t = T_t \cdot C_t \cdot S_t \cdot I_t$:

$$\frac{Y_t}{T_t} = C_t \cdot S_t \cdot I_t.$$

En évaluant successivement chacune des trois composantes et en ajustant la série chronologique suivant chacune d'elles, il obtiendra les valeurs attribuées aux variations irrégulières.

DOMAINES ET LIMITATIONS

Lorsque la série chronologique ne porte pas sur une trop longue période, on regroupe la **tendance séculaire** et les fluctuations cycliques en une seule composante appelée **mouvement extra-saisonnier**.

Représentations graphiques

La **représentation graphique** d'une série chronologique contribue à analyser les valeurs observées en fonction du temps.

Pour ce faire, un certain nombre de précautions doivent être prises :

1. Le point représentatif doit être reporté exactement à la place convenable. Si la valeur représente un stock, à une date déterminée, le point se trouve à l'aplomb de la date de référence. S'il s'agit d'un flux, par exemple mensuel, le point est fixé à la verticale du milieu du mois.

2. Il arrive qu'on doive reporter sur un même graphique des données annuelles et des données mensuelles. Les points représentatifs des résultats mensuels devraient être douze fois plus serrés que ceux relatifs aux résultats annuels. Il faut donc prendre garde à reporter les points représentatifs à la place convenable.

3. Si plusieurs séries sont représentées sur un même graphique, elles devront être de nature identique. Sinon, il faut utiliser des échelles différentes ou disposer les graphiques l'un en-dessous de l'autre.

Les principaux objectifs de l'étude des séries chronologiques sont :

- la description : détermination et élimination des diverses composantes ;
- la modélisation stochastique ;
- la **prévision** ;
- le filtrage : élimination ou conservation de certaines caractéristiques ;
- le contrôle : **observation** du passé d'un processus pour agir sur son évolution future.

Dans un processus de prévision, l'analyse des séries chronologiques s'appuie sur les hypothèses suivantes :

- les grandes tendances observées historiquement se maintiennent dans le futur, plus ou moins proche ;
- les fluctuations mesurables d'une **variable** se reproduisent à intervalles réguliers.

L'analyse de séries chronologiques sera justifiée tant et aussi longtemps qu'elle permettra de réduire le niveau d'incertitude dans l'élaboration des prévisions.

On distingue :

1. *La prévision à long terme :*

   elle est basée sur la tendance séculaire et ne prend généralement pas en considération les fluctuations cycliques.

2. *La prévision à moyen terme :*

   pour tenir compte de l'effet probable de la composante cyclique, il faut multiplier la valeur prévue de la tendance séculaire par une **estimation**

de la variation relative escomptée attribuable à des fluctuations cycliques.

3. *La prévision à court terme :*

   en général, on n'essaie pas de prévoir les variations irrégulières, ni dans les prévisions à court terme, ni dans les prévisions portant sur des périodes plus longues. On multiplie alors la valeur prévue de la tendance séculaire par une estimation des fluctuations cycliques et par la variation saisonnière de la date correspondante.

Lors d'une analyse de séries chronologiques, on est confronté aux problèmes suivants :

1. *Le modèle classique semble-t-il approprié ?*

   Il faut s'attendre à ce que la qualité des résultats de l'analyse soit proportionnelle à la précision et à la justesse du **modèle** lui-même. Le modèle multiplicatif $Y_t = T_t \cdot S_t \cdot C_t \cdot I_t$ suppose les quatre composantes indépendantes, ce qui est rarement conforme à la réalité.

2. *Nos hypothèses de constance et de régularité sont-elles valables ?*

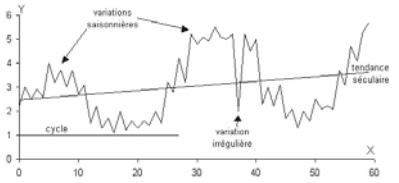
   L'analyste doit rectifier ses calculs à la lumière de facteurs subjectifs et qualitatifs qui se trouvent dans presque toutes les situations.

3. *Peut-on se fier aux données disponibles ?*

   Les problèmes de rupture de séries provenant d'insuffisance de données ou de changement de qualité des variables peuvent survenir.

EXEMPLE

Une série chronologique se présente d'une façon générale comme suit :



On y distingue les quatre composantes, à savoir :

- la **tendance séculaire**, légèrement croissante dans le cas présent ;
- les variations saisonnières assez prononcées ;
- les fluctuations cycliques, apparaissant sous forme de cycles d'amplitude approximative de 27 unités de temps ;
- les variations irrégulières, généralement assez faibles, excepté pour $t = 38$ qui présente une chute très brusque qui ne saurait être rationnelle.

MOTS CLÉS

Fluctuation cyclique *(Cyclical fluctuation)*
Graphique logarithmique *(Logarithmic plot)*
Mouvement extra-saisonnier *(Extra-seasonal movement)*
Prévision *(Forecasting)*
Représentation graphique *(Graphical representation)*
Tendance séculaire *(Secular trend)*
Variation irrégulière *(Irregular variation)*
Variation saisonnière *(Seasonal variation)*

RÉFÉRENCES

Copeland, M.T. (1915). Statistical indices of business conditions. *The Quarterly Journal of Economics*, vol. 29, pp. 522-562.

Funkhauser, H.G. (1936). A Note on a 10th Century Graph. *Osiris*, vol. 1, pp. 260-262. Bruges.

Kendall, M.G. (1973). *Time Series.* Charles Griffin & Company Ltd.

Macauley, F.R. (1930). The smoothing of time series. *National Bureau of Economic Research*, pp. 121-136.

Persons, W.M. (1919). An index of general business conditions. *Review of Economics and Statistics*, vol. 1, pp. 111-205.

SEUIL DE SIGNIFICATION

Signifiance level

Le seuil de signification est un **paramètre** du **test d'hypothèse** et sa **valeur** est fixée par l'utilisateur avant la réalisation du test.

Le seuil de signification d'un test d'hypothèse, dénoté par α, est la **probabilité** de rejeter l'**hypothèse nulle** H_0 alors que celle-ci est vraie :

$$\alpha = P\{\text{rejeter } H_0 | H_0 \text{ est vraie}\}.$$

Le seuil de signification est aussi appelé risque d'**erreur** d'un test, ou plus particulièrement **erreur de première espèce**.

HISTORIQUE

Lorsqu'il développa le **test du chi-carré**, **Karl Pearson** (1900) utilisait le seuil de signification comme une mesure d'adéquation entre les données et les hypothèses. Ainsi, des résultats non significatifs indiquaient une bonne convenance de l'**hypothèse nulle**, mais sans la notion de vérité.

Voir **test d'hypothèse**.

DOMAINES ET LIMITATIONS

Soucieux de la possibilité d'**erreur** en rejetant l'**hypothèse nulle** alors qu'elle est vraie, le statisticien construit le test d'hypothèse tel que la **probabilité** d'erreur de ce type ne dépasse pas une valeur fixe, donnée d'avance. Cette **valeur** correspond au seuil de signification du test.

Il faut noter que le statisticien faisant un **test d'hypothèse** doit faire face à un deuxième type d'**erreur**, appelé **erreur de seconde espèce**. Celui-ci se produit quand l'hypothèse nulle n'est pas rejetée alors qu'en réalité elle est fausse.

Il ne suffit donc pas de diminuer le seuil de signification pour diminuer le risque global d'erreur dans la prise de décision, mais il faut trouver un compromis entre le seuil de signification et l'erreur de seconde espèce. Une des méthodes consiste à se baser sur l'étude de la puissance du test.

EXEMPLE

Dans l'exemple suivant, nous illustrons l'influence du choix du seuil de signification dans la réalisation d'un **test d'hypothèse**.

Considérons un **test unilatéral** à droite sur la **moyenne** μ d'une **population** dont la **variance** σ^2 est égale à 36.

Posons les hypothèses suivantes :

hypothèse nulle	$H_0 : \mu = 30,$
hypothèse alternative	$H_1 : \mu > 30.$

Supposons que la moyenne $\bar{x}$ d'un **échantillon** de taille $n = 100$ issu de cette population est égale à 31.

Cas 1

Choisissons un seuil de signification $\alpha = 5$ %. Nous trouvons la **valeur critique** de $z_\alpha = 1,65$ dans la table normale. La borne supérieure de la **région d'acceptation** est déterminée par :

$$\mu_{\bar{x}} + z_\alpha \cdot \sigma_{\bar{x}}$$

où

- $\mu_{\bar{x}}$ est la moyenne de la **distribution d'échantillonnage** des moyennes et
- $\sigma_{\bar{x}}$ est l'**erreur type** (ou l'**écart type** de la distribution d'échantillonnage des moyennes),

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

Cette borne supérieure est donc égale à :

$$30 + 1,65 \cdot \frac{6}{\sqrt{100}} = 30,99.$$

La **région de rejet** de l'hypothèse nulle correspond donc à l'**intervalle** $[30,99\,;\,\infty[$ et la région d'acceptation à l'intervalle $]-\infty\,;\,30,99[$.

Cas 2

Si nous fixons un seuil de signification plus petit, par exemple $\alpha = 1\ \%$, la valeur de z_α de la **table normale** est égale à $2,33$ et la borne supérieure de la région d'acceptation devient égale à :

$$\begin{aligned} \mu_{\bar{x}} + z_\alpha \cdot \sigma_{\bar{x}} &= 30 + 2,33 \cdot \frac{6}{\sqrt{100}} \\ &= 31,4. \end{aligned}$$

La région de rejet de l'hypothèse nulle correspond donc à l'intervalle $[31,4\,;\,\infty[$ et la région d'acceptation à l'intervalle $]-\infty\,;\,31,4[$.

À partir de la moyenne observée $\bar{x} = 31$, nous constatons que dans le premier cas ($\alpha = 5\ \%$), nous sommes conduits à rejeter l'hypothèse nulle puisque la moyenne échantillonnale (31) est située dans la région de rejet. Alors que dans le second cas ($\alpha = 1\ \%$), nous sommes amenés à prendre la décision inverse : l'hypothèse nulle ne peut pas être rejetée puisque la moyenne échantillonnale est située dans la région d'acceptation.

Nous constatons donc que le choix du seuil de signification a une influence directe sur la détermination de la région d'acceptation de l'hypothèse nulle et en conséquence sur la prise de décision. Un seuil de signification plus petit (donc une **probabilité** d'**erreur de première espèce** plus petite) restreint la région de rejet et augmente la région d'acceptation de l'hypothèse nulle.

À un risque d'**erreur** plus faible est donc associée une précision moins grande.

MOTS CLÉS

Erreur de première espèce *(Type I error)*
Test d'hypothèse *(Hypothesis testing)*

RÉFÉRENCE

Pearson, K. (1900). 1948. On the Criterion, That a Given System of Deviations from the Probable in the Case of a Correlated System of Variables is such that it can be reasonably supposed to have arisen from random Sampling. pp. 339-357 in Karl Pearson, *Karl Pearson's Early Statistical Papers.* Cambridge Univ. Press. First published in the *Philosophical Magazine*, 5th Series, 50, pp. 157-175.

SIGNE, TEST DU

Sign test

Voir **test du signe**.

SIMULATION

Simulation

La simulation est la méthode statistique de recherche opérationnelle permettant la reconstitution fictive de l'évolution d'un phénomène. C'est une **expérience** qui suppose la construction d'un **modèle** théorique présentant une similitude de propriétés ou de relations avec le phénomène faisant l'objet de l'étude.

Le problème fondamental de la simulation réside dans la construction des échantillons artificiels relatifs à des variables aléatoires de loi connue statistiquement. La connaissance de ces lois est en général empirique : elle résulte d'une étude statistique à partir de laquelle on peut déterminer les lois de probabilité des variables aléatoires caractérisant le phénomène.

Une fois ces lois connues, la construction d'échantillons se fait en engendrant des nombres aléatoires.

HISTORIQUE

La méthode de simulation a une longue histoire. Ses débuts remontent au moins aux travaux de Student (1908) qui découvrit les distributions d'échantillonnage de la **statistique** t et le **coefficient de corrélation**.

Plus récemment, grâce aux ordinateurs, les méthodes de simulation et aussi les méthodes de Monte-Carlo ont pu se développer rapidement.

EXEMPLE

Considérons le phénomène de file d'attente aux caisses d'un grand magasin. Supposons que l'on veuille connaître le meilleur « système de caisse » c'est-à-dire celui pour lequel la somme des coûts d'inactivité des caisses et des coûts d'attente des clients soit la plus faible.

L'évolution du phénomène dépend essentiellement de la loi d'arrivée des clients aux caisses et de la loi des temps de service. La simulation consiste donc à construire un **échantillon** d'arrivée des clients et un échantillon des temps de service. On peut alors calculer le coût d'un système. Il suffit de répéter l'**expérience** pour différents systèmes et de choisir le meilleur.

MOTS CLÉS

Bootstrap *(Bootstrap)*
Génération de nombres aléatoires *(Generation of random numbers)*
Méthode de Monte-Carlo *(Monte Carlo method)*
Méthode du jackknife *(Jackknife method)*
Nombre aléatoire *(Random number)*

RÉFÉRENCES

Gosset, S.W. « **Student** » (1908). The Probable Error of a Mean. *Biometrika*, 6, pp. 1-25.

Kleijnen, J.P.C. (1974). Statistical Techniques in Simulation (in two parts). Vol. 9 of *Statistics: Textbooks and Monographs.* Marcel Dekker, New York.

Kleijnen, J.P.C. (1987). Statistical Tools for Simulation Practitioners. Vol. 76 of *Statistics: Textbooks and Monographs.* Marcel Dekker, New York.

Naylor, T.H., Balintfy, J.L., Burdick, D.S. and Chu, K. (1967). *Computer Simulation Techniques.* John Wiley & Sons, New York.

SNEDECOR GEORGE WADDEL

Snedecor George Waddel



George Waddel Snedecor naquit en 1881 à Memphis (Tennesse) et mourut en 1974 à Amherst (Massachusetts). En 1899, il entra au Alabama Polytechnic Institute à Auburn, y resta durant deux ans, puis effectua deux années de préparation à une école d'enseignement. Quand sa famille déménagea à Tuscaloosa en 1903, Snedecor fut transféré à l'Université d'Alabama, où il reçut, en 1905, sa licence ès sciences en mathématiques et physique. Il accepta alors son premier poste académique à la Selma Military Academy où il enseigna de 1905 à 1907. Dès 1907 et jusqu'en 1910, il enseigna les mathématiques et le grec au Austin College à Sherman.

En 1910, il se déplaça à Ann Arbor, Michigan, où Snedecor parfit son éducation à l'Université du Michigan et, en 1913, il obtint un postgrade. Cette même année, il s'installa à Ames dans l'Iowa, où il resta jusqu'en 1958.

En 1927, le Mathematics Statistical Service fut

inauguré à Ames avec Snedecor et A. E. Brandt à sa tête. Ce service fut le précurseur du Statistical Laboratory inauguré en 1933 dont Snedecor fut le directeur et dans lequel travailla **Gertrude Cox**.

En 1931, Snedecor fut nommé professeur au département de mathématique de l'Iowa State College. Durant cette période, Snedecor se lia d'amitié avec **Ronald Aylmer Fisher**.

En 1947, Snedecor quitta le Statistical Laboratory et devint professeur de statistiques au Iowa State College et resta à ce poste jusqu'à sa retraite en 1958.

Ouvrage principal de George Waddel Snedecor :

1937 *Statistical Methods Applied to Experiments in Agriculture and Biology.* Collegiate Press, Inc. Ames, Iowa.

SONDAGE

Survey

Une enquête par sondage, ou plus simplement un sondage, est une enquête effectuée sur une partie restreinte de la **population**. Cette fraction de la population constitue l'**échantillon** et les méthodes qui permettent de construire cet échantillon s'appellent méthodes d'**échantillonnage**.

En procédant par sondage, on n'obtient des informations que sur les unités formant l'échantillon. Ensuite, à l'aide de la **statistique inférentielle**, on généralisera ces informations à toute la population. Procéder à cette généralisation signifie introduire une certaine **erreur**. L'importance de cette erreur dépendra de la **taille** de l'échantillon, de la façon dont il a été choisi et de l'**estimateur** utilisé. Plus l'échantillon sera représentatif de la population, plus l'erreur sera petite. Les méthodes d'échantillonnage servent à former des échantillons permettant d'obtenir une erreur minimale.

Une autre technique d'enquête consiste à observer tous les individus d'une population. On parle alors de **recensement**. Par rapport au recensement, le sondage présente des avantages de coût et de rapidité qui compensent largement l'erreur due à l'échantillonnage. De plus, il est parfois impossible d'avoir recours au recensement. C'est le cas lorsque le fait d'observer une unité de la population entraîne sa destruction, ou lorsque la population est infinie.

MOTS CLÉS

Collecte des données *(Data collection)*
Échantillon *(Sample)*
Échantillonnage *(Sampling)*
Panel *(Panel)*
Population *(Population)*
Recensement *(Census)*
Statistique inférentielle *(Inferential statistics)*

RÉFÉRENCES

Gourieroux, C. (1981). *Théorie des sondages.* Economica, Paris.

Grosbras, J.M. (1987). *Méthodes statistiques des sondages.* Economica, Paris.

SPEARMAN, TABLE DE

Spearman table

Voir **table de Spearman**.

STATISTIQUE

Statistics

Dans une acception précise et scientifique, on appelle statistique une mesure calculée à partir de données provenant d'un **échantillon**.

La **moyenne** échantillonnale, dénotée par $\bar{x}$, l'**écart type** d'un échantillon, représenté par le symbole S, et le **pourcentage** p calculé sur un échantillon, sont chacun une statistique.

Dans une acception plus globale, la notion de statistique représente l'ensemble de la théorie

statistique qui traite des propriétés des populations dans leurs ensembles.

Si nous considérons une population d'individus, nous ne serons pas intéressés, en théorie statistique, de savoir si tel ou tel individu possède une voiture ou s'il est fumeur. Par contre, nous serons intéressés de savoir combien d'individus possèdent une voiture ou sont fumeurs, et s'il existe une **relation** entre la possession d'une voiture et la propension à fumer dans la population étudiée.

Nous nous intéressons donc aux propriétés de la population dans sa globalité, sans nous préoccuper des propriétés individuelles de chaque personne ou de chaque objet faisant partie de cette population.

Le mot statistique, dérivé du latin, se réfère à la notion d'État (*status*) : « qui est relatif à l'État ». Il est certain que les gouvernements avaient et ont grand besoin de compter et de mesurer de nombreuses choses et activités, telles que l'évolution de la population, des naissances, de l'immigration ou de l'émigration, l'évolution des emplois, des entreprises, etc.

Dans cette perspective, le terme « statistique », ou « statistiques », est utilisé pour indiquer un ensemble de données disponibles à propos d'un phénomène déterminé (par exemple : statistiques des chômeurs).

Dans une acception plus moderne et plus complète du mot, la « statistique » est considérée comme une discipline ayant trait à des données numériques. Elle comprend un ensemble de techniques devant conduire à l'acquisition de connaissances générales à partir de données incomplètes, à partir d'un système scientifique rigoureux guidant le recueil de données, leur organisation, leur analyse et leur interprétation, pour autant qu'on puisse leur donner une forme numérique.

On distingue deux sous-ensembles de techniques :

- celles relatives à la **statistique descriptive** ;
- celles relatives à la **statistique inférentielle**.

Le but essentiel des statistiques descriptives est de représenter l'information d'une façon compréhensible et utilisable. La statistique inférentielle, de son côté, a pour fonction d'aider à la généralisation de cette information ou, plus spécifiquement, de faire des inférences (à propos de populations) basées sur des échantillons de ces populations.

HISTORIQUE

Le terme « statistique », issu du vocable latin *status* (état) a été utilisé pour la première fois, selon Kendall (1960), par l'historien italien Girolamo Ghilini en 1589. La statistique à cette époque, et bien avant déjà, était liée à l'idée de dénombrement, d'inventaire et même de **recensement** (voir HISTORIQUE de la notice Recensement).

L'arithmétique politique, conçue dans un contexte démographique et économique et utilisant les informations recueillies à des fins plus élaborées, constitue, selon Kendall (1960) et Stigler (1986), la véritable origine de la statistique moderne ; cette dernière étant fixée à 1660.

Les méthodes statistiques ainsi que les lois statistiques connurent leur développement principalement au XIX^{e} siècle grâce au rôle important de la statistique dans les domaines des sciences expérimentales et humaines.

Au XX^{e} siècle, la statistique devenait une discipline à elle toute seule, à cause de la richesse et de la diversité des méthodes qu'elle renferme. Ces raisons expliquent également la naissance d'encyclopédies consacrées à la statistique comme l'*International Encyclopedia of Statistics* (1978) publiée en deux volumes et l'*Encyclopedia of Statistical Sciences* (1982-1988) en neuf volumes.

DOMAINES ET LIMITATIONS

La statistique intervient dans de nombreux domaines, fort différents les uns des autres. Ainsi trouve-t-on sa place autant dans la production industrielle que dans la recherche océanographique, dans la conception de systèmes de contrôle du trafic aérien, que dans les services de l'État afin d'estimer le coût de la vie.

Passons en revue quelques-uns des domaines susceptibles d'intéresser un statisticien :

- *Recherche fondamentale en statistique :* recherche en probabilités, méthodes statistiques et théorie.
- *Biologie :* recherche fondamentale et expérimentations relatives aux principaux phénomènes présentés par les organismes vivants, biométrie.
- *Commerce :* gestion des données, volume de ventes, management, procédures d'inventaires, implantation industrielle, communications et contrôle théorique, procédures de vérification comptable.
- **Démographie** : étude de la croissance des populations humaines (taux de natalité, taux de mortalité, mouvement migratoire), étude de la structure de la population (caractéristiques personnelles, sociales et économiques).
- *Économie :* mesure du volume de la production, du commerce, des ressources, de l'emploi et du niveau de vie ; analyse du comportement des consommateurs et des producteurs, des réponses du marché aux changements de prix, de l'impact de la publicité et des politiques gouvernementales.
- *Éducation :* mesures et tests liés au processus d'apprentissage, et études auprès d'organismes à vocation éducative.
- *Ingénierie :* recherche et expérimentation incluant le design et les tests d'efficacité, l'amélioration des méthodes de tests, les questions relatives à l'**inférence**, aux tests de précision ; amélioration des méthodes de contrôle de qualité.
- *Santé :* coûts des accidents, des soins médicaux, de l'hospitalisation.
- *Assurances :* détermination de la mortalité, des taux d'accidents concernant les assurés et la population en général ; détermination des taux de primes pour chaque **catégorie** de risque ; application des techniques statistiques à la gestion des programmes d'assurance.
- *Marketing et recherche sur la consommation :* problèmes liés à l'importance des marchés, au circuit de distribution, à la détection de nouveaux débouchés ; étude des préférences des consommateurs et de leurs comportements.
- *Médecine :* épidémiologie, recherche fondamentale et expériences sur les causes, le diagnostic, le traitement et la prévention des maladies.
- *Gestion et administration :* problèmes liés à la gestion du personnel, à celle de l'équipement, aux méthodes de production et aux conditions de travail.
- *Psychologie et psychométrie :* problèmes relatifs à l'évaluation de la capacité d'apprentissage, de l'intelligence, des caractéristiques personnelles et au comportement normal ou anormal de l'individu ainsi qu'à l'établissement d'échelles et d'instruments de mesure.
- *Sciences sociales :* établissement de techniques d'**échantillonnage** et de tests théoriques concernant les systèmes sociaux et le bien-être social. Analyse du coût des assurances sociales ; analyse des différences culturelles sur le plan des valeurs, des comportements sociaux.
- *Recherche spatiale :* interprétation d'expériences et de données recueillies lors de missions spatiales.
- *Sciences (en général) :* recherche fondamentale dans les sciences naturelles, en physique et en sciences sociales.

MOTS CLÉS

Statistique bayesienne *(Bayesian statistics)*
Statistique descriptive *(Descriptive statistics)*
Statistique génétique *(Genetical statistics)*
Statistique inférentielle *(Inferential statistics)*
Statistique officielle *(Official statistics)*
Statistique spatiale *(Spatial statistics)*

RÉFÉRENCES

Kendall, M.G. (1960). Where shall the history of Statistics begin ?, *Biometrika*, 47, pp. 447-449.

Kotz, S. and Johnson, N.L. (1982-1988). *Encyclopedia of Statistical Sciences.* Vol. 1-9. John Wiley & Sons, New York.

Kruskal, W.H. and Tanur, J.M. (1978). *International Encyclopedia of Statistics.* Vol. 1-2. The Free Press, New York.

Stigler, S. (1986). *The History of Statistics, the Measurement of Uncertainty before 1900.* The Belknap Press of Harvard University Press. London, England.

STATISTIQUE, TABLE

Statistical table

Voir **table statistique**.

STATISTIQUE BAYESIENNE

Bayesian statistics

La statistique bayesienne est un large domaine de la statistique qui se distingue par une axiomatisation des statistiques qui lui confère une certaine cohérence interne.

L'idée de base est d'interpréter la **probabilité** d'un **événement** dans le sens commun, c'est-à-dire comme l'incertitude qui lui est liée. L'approche classique considère par contre la probabilité d'un événement comme la limite de sa fréquence relative (voir probabilité pour une approche plus formelle).

L'aspect plus connu de l'**inférence** bayesienne est la possibilité de calculer la **loi de probabilité** (ou **fonction de densité**) **conjointe** $f(\theta\,|X = x_1, \ldots, X = x_n)$ d'un ou de plusieurs paramètres θ (soit un paramètre, soit un vecteur de paramètres) ayant observé des données x_1, …, x_n tirées de manière indépendante d'une variable aléatoire X, de laquelle dépend θ. (À remarquer que cela permet aussi de calculer la loi de probabilité d'une nouvelle observation x_{n+1}.)

La statistique bayesienne traite les paramètres inconnus comme variables aléatoires non pas à cause d'une possible variabilité (en effet les paramètres inconnus sont considérés fixes), mais à cause de notre ignorance ou incertitude en ce qui les concerne. À noter que ce n'est pas possible dans la statistique classique à cause de l'interprétation de la probabilité d'un événement.

Le problème qui se pose lors du calcul de la loi $f(\theta\,|X = x_1, \ldots, X = x_n)$ (en anglais *posterior distribution*) est le passage imminent par $f(X = x_1, \ldots, X = x_n\,|\theta)$ et $f(\theta)$.

La première expression ne cause pas de problème, car c'est la fonction qu'on utilise d'habitude dans la statistique classique et connue sous le nom de fonction de vraisemblance (voir **maximum de vraisemblance**).

La deuxième partie, par contre, suppose une distribution *a priori* (dite en anglais *prior distribution*) de θ. On utilise souvent la distribution initiale de θ pour incorporer des éventuelles informations supplémentaires sur le paramètre d'intérêt. En absence de ces informations, on utilise une fonction de référence, qui maximise l'information manquante (et qui est donc la plus « objective » ou « non informative », selon un usage commun mais impropre).

Une fois qu'on a calculé la distribution $f(\theta\,|x_1, \ldots, x_n)$, toute information sur le(s) paramètre(s) d'intérêt est disponible. Ainsi, on peut calculer des valeurs plausibles pour le paramètre inconnu (soit la **moyenne**, la **médiane** ou autre **mesure de tendance centrale**), son **écart type**, des intervalles de confiance ou bien faire des tests d'hypothèse sur sa valeur.

HISTORIQUE

Voir **Bayes Thomas** et **théorème de Bayes**.

ASPECTS MATHÉMATIQUES

Soit D l'ensemble des données $X = x_1$, …, $X = x_n$ issues indépendamment d'une variable aléatoire X de loi inconnue. On va considérer le cas simple où il n'y a qu'un paramètre θ d'intérêt qui dépend de X.

Une procédure bayesienne standard peut donc s'exprimer par :

1. Identifier les quantités connues $x_1, \ldots, x_n$.
2. Spécifier un modèle pour les données, c'est-à-dire une famille paramétrique $f(x\,|\theta)$ de distributions qui décrit la génération des données.
3. Spécifier l'incertitude concernant θ par une fonction de répartition initiale $f(\theta)$.
4. On peut ensuite calculer la distribution $f(\theta\,|D)$ (dite *distribution finale*) par le **théorème de Bayes**.

Les premiers deux points sont communs à toute **inférence** statistique.

Le troisième point est plus problématique. En l'absence d'informations supplémentaires sur θ, l'idée est de calculer une *distribution de réference* $f(\theta)$ en maximisant une fonction qui spécifie l'information manquante concernant le paramètre θ. Une fois ce problème résolu, le quatrième point découle aisément à l'aide du théorème de Bayes.

Le théorème de Bayes peut s'exprimer, dans sa forme continue, par :

$$f(\theta\,|D) = \frac{f(D\,|\theta)\cdot f(\theta)}{f(D)} = \frac{f(D\,|\theta)\cdot f(\theta)}{\int f(D\,|\theta)\, f(\theta)\, d\theta}.$$

Comme les x_i sont indépendants on peut écrire :

$$f(\theta\,|D) = \frac{\prod_{i=1}^{n} f(x_i\,|\theta)\cdot f(\theta)}{\int \prod_{i=1}^{n} f(x_i\,|\theta) f(\theta)\, d\theta}.$$

Maintenant, on a les moyens pour calculer la **fonction de densité** d'une nouvelle observation (indépendante) x_{n+1}, étant donné $x_1, \ldots, x_n$:

$$\begin{aligned} f(X = x_{n+1}\,|D) &= \frac{f(X = x_1, \ldots, X = x_{n+1})}{f(X = x_1, \ldots, X = x_n)} \\ &= \frac{\int f(X = x_1, \ldots, X = x_{n+1}\,|\theta)\cdot f(\theta)\, d\theta}{\int f(D\,|\theta)\cdot f(\theta)\, d\theta} \\ &= \frac{\int \prod_{i=1}^{n+1} f(X = x_i\,|\theta)\cdot f(\theta)\, d\theta}{\int \prod_{i=1}^{n} f(X = x_i\,|\theta)\cdot f(\theta)\, d\theta} \\ &= \int f(X = x_{n+1}\,|\theta)\cdot f(\theta\,|D)\, d\theta. \end{aligned}$$

On va maintenent expliquer rapidement les méthodes qui permettent de :

- donner *une* valeur, plus plausible que d'autres, du paramètre à estimer ;
- donner des intervalles de confiance pour θ et
- faire des tests d'hypothèse ;

et qui sont strictement liées à la théorie de la décision, qui joue en effet un rôle considérable dans la statistique bayesienne.

Estimation ponctuelle du paramètre

Pour l'**estimation ponctuelle** du paramètre θ, on spécifie une fonction de perte d'information $l\left(\hat{\theta}, \theta\right)$ qui dérive de l'utilisation de $\hat{\theta}$ (la valeur estimée) au lieu de la vraie valeur (inconnue) θ. Ensuite, on minimise cette fonction par rapport à $\hat{\theta}$. Par exemple si θ est réel et « la perte est quadratique » (c'est-à-dire $l\left(\hat{\theta}, \theta\right) = \left(\hat{\theta} - \theta\right)^2$), alors l'estimation ponctuelle de θ sera la moyenne de la distribution calculée.

Intervalles de confiance

Le concept d'**intervalle de confiance** est remplacé dans la statistique bayesienne par le concept de région α-crédible (α étant le « seuil de confiance ») qui est simplement définie comme un intervalle I tel que :

$$\int_I f(\theta\,|D)\, d\theta = \alpha.$$

Souvent, on demande aussi que l'intervalle soit de longueur minimale.

Tests d'hypothèse

Un **test d'hypothèse** général :

$$\begin{aligned} H_0 &: \theta \in I \\ \text{contre } H_1 &: \theta \notin I \end{aligned}$$

revient à nouveau à un problème de théorie de la décision. On définit ainsi une fonction $l(a_0, \theta)$ de perte d'information en acceptant H_0 (si la vraie valeur du paramètre est θ) et une fonction $l(a_1, \theta)$ de perte d'information en acceptant H_1.

Si la valeur esperée de perte d'information obtenue en acceptant H_0, c'est-à-dire :

$$\int l(a_0, \theta)\, d\theta,$$

est inférieure à celle obtenue en rejetant H_0, alors il convient d'accepter H_0.

Dans le cas contraire, on rejette H_0 et donc aussi la restriction du modèle aux valeurs de θ qui appartiennent à I.

EXEMPLE

L'exemple suivant concerne l'estimation du paramètre θ d'une **loi de Bernoulli** X à l'aide de n-observations indépendantes x_1, ..., x_n prenant la valeur 1 dans le cas d'un succès et 0 dans le cas d'un échec. Soit r le nombre de succès et $n - r$ le nombre d'échecs parmi les observations.

On a donc :

$$\begin{aligned} L(\theta) &= P(X = x_1, \ldots, X = x_n \,|\theta) \\ &= \prod_{i=1}^{n} P(X = x_i \,|\theta) \\ &= \theta^r \cdot (1-\theta)^{n-r}. \end{aligned}$$

L'estimateur du **maximum de vraisemblance** de θ, noté $\hat{\theta}_{mle}$, maximise cette fonction. Pour ce faire on considère le logarithme de cette fonction, c'est-à-dire le log-likelihood :

$$\log L(\theta) = r \log(\theta) + (n-r) \log(1-\theta)$$

et on le maximise en posant sa dérivée par rapport à θ égale à zéro :

$$\frac{\partial (\log L)}{\partial \theta} = \frac{r}{\theta} - \frac{n-r}{1-\theta} = 0.$$

Cela est équivalent à $r(1-\theta) = (n-r)\theta$ et se simplifie en $\hat{\theta}_{mle} = \frac{r}{n}$. L'estimateur du maximum de vraisemblance, qui est typiquement classique, de θ est donc tout simplement la proportion de succès observés.

Maintenant, on revient à la méthode bayesienne.

Dans le cas d'une loi de Bernoulli, la distribution de référence *a priori* du paramètre θ s'exprime par :

$$f(\theta) = c \cdot (h(\theta))^{\frac{1}{2}}$$

où c est une constante adéquate (telle que $\int f(\theta)\, d\theta = 1$) et où $h(\theta)$ est dite information de Fisher de X :

$$\begin{aligned} h(\theta) &= -E\left[\frac{\partial^2 L}{\partial \theta^2}\right] = -E\left[-\frac{X}{\theta^2} - \frac{1-X}{(1-\theta)^2}\right] \\ &= \frac{E[X]}{\theta^2} + \frac{1-E[X]}{(1-\theta)^2} = \frac{\theta}{\theta^2} + \frac{1-\theta}{(1-\theta)^2} \\ &= \frac{1}{\theta} + \frac{1}{1-\theta} = \frac{1}{\theta(1-\theta)}. \end{aligned}$$

La distribution *a priori* sera donc :

$$f(\theta) = c \cdot \theta^{-\frac{1}{2}} \cdot (1-\theta)^{-\frac{1}{2}}.$$

La fonction de répartition de θ, les observations $x_1, \ldots, x_n$ étant données, peut s'exprimer par :

$$\begin{aligned} & f(\theta \,|X = x_1, \ldots, X = x_n) \\ &= \tfrac{1}{d} \prod_{i=1}^{n} P(X = x_i \,|\theta) \cdot f(\theta) \\ &= \tfrac{1}{d} \cdot \theta^r \cdot (1-\theta)^{n-r} \cdot c \cdot \theta^{-\frac{1}{2}} \cdot (1-\theta)^{-\frac{1}{2}} \\ &= \tfrac{c}{d} \cdot \theta^{r-\frac{1}{2}} \cdot (1-\theta)^{n-k-\frac{1}{2}} \\ &= \tfrac{c}{d} \cdot \theta^{(r+\frac{1}{2})-1} \cdot (1-\theta)^{(n-r+\frac{1}{2})-1}. \end{aligned}$$

Il s'agit d'une **loi bêta** de paramètres

$$\alpha = r + \frac{1}{2} \quad \text{et} \quad \beta = n - r + \frac{1}{2}$$

et avec constante

$$\frac{c}{d} = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\,\Gamma(\beta)}$$

où Γ est la fonction gamma (voir **loi gamma**).

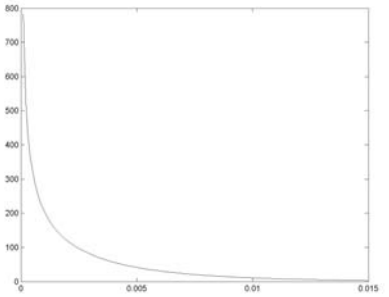
On considère le cas concret dans lequel on désire estimer la proportion d'étudiants séropositifs. On effectue un test sur 200 étudiants et personne ne se révèle séropositif. La proportion d'étudiants séropositifs est estimée égale à 0 par le maximum de vraisemblance. Les intervalles de confiance ne sont pas très utiles dans ce cas, car, suivant la formule usuelle, ils se calculent par :

$$0 \pm 1,96 \cdot \sqrt{p\,(1-p)}.$$

Mais comme p est la proportion de succès observée, l'intervalle de confiance se réduit à 0.

Suivant une méthodologie bayesienne, on obtient par contre une distribution, calculée uniquement à l'aide des données, qui décrit l'incertitude concernant le paramètre à estimer. Plus n est grand, plus il a de sécurité sur θ ; la distribution finale de référence concernant θ est donc plus concentrée sur la vraie valeur de θ.

Dans ce cas, on trouve comme distribution finale de référence une loi bêta de paramètres $\alpha = 0,5$ et $\beta = 200,5$, qui résume l'information sur θ (les valeurs qui correspondent aux pics de la distribution étant les plus probables). Une représentation graphique est souvent utile pour interpréter un tel résultat :



On arrive à lire par exemple :

- que la probabilité que la proportion d'intérêt soit plus petite que $0,015$ est pratiquement 1 ;
- que la probabilité que la proportion d'intérêt soit plus petite que $0,005$ est approximativement $0,84$ et
- que la médiane, qui peut être choisie comme mesure de tendance centrale adéquate à cause de la forte asymétrie de la distribution, est $0,011$.

On remarque qu'il y a une différence qualitative entre le résultat classique (la proportition peut être estimée zéro) et la solution bayesienne au problème, qui permet de calculer la loi de probabilité du paramètre, traduisant mathématiquement l'incertitude à son égard. Cette méthode nous dit en particulier que, disposant de l'information donnée, l'estimation correcte de la proportion d'intérêt est $0,011$. À noter que l'estimation bayesienne dépend (comme l'incertitude) du nombre de cas observés, tandis que classiquement, il est équivalent d'observer 0 cas de sida sur 2, sur 200 ou sur 2 000 étudiants.

MOTS CLÉS

Bayes Thomas*(Bayes Thomas)*
Fonction de densité conjointe *(Joint density function)*
Inférence *(Inference)*
Maximum de vraisemblance *(Maximum likelihood)*
Probabilité *(Probability)*
Probabilité conditionelle *(Conditional probability)*
Théorème de Bayes *(Bayes' theorem)*

RÉFÉRENCES

Bernardo, J.M. and Smith, A.F.M. (1994). *Bayesian theory.* John Wiley & Sons, Chichester.

Berry, D.A. (1996). *Statistics, a Bayesian perspective.* Belmont, California : Wadsworth Publishing Company.

Box, G.E.P. and Tiao, G.P. (1973). *Bayesian inference in statistical analysis.* Addison-Wesley, Reading, Mass.

Carlin, B., Louis T.A. (2000). *Bayes & Empirical Bayes Methods*, Chapman & Hall/CRC, London.

de Finetti, B. (1975). *Theory of probability ; a critical introductory treatment.* Translated by Antonio Mach and Adrian Smith. John Wiley & Sons, New York.

Gelman A., Carlin, J.B., Stern H.S., Rubin D.B. (2003). *Bayesian Data Analysis*, Chapman & Hall/ CRC, London.

Lindley, D.V. (2000). The philosophy of statistics. *The Statistician*, 49, pp. 293-337.

Robert, C.P. (1992). *L'Analyse statistique bayesienne.* Economica, Paris.

STATISTIQUE DESCRIPTIVE

Descriptive statistics

Les informations recueillies dans la plupart des études et des recherches que l'on entreprend peuvent prendre des formes différentes : données de fréquences (nombre de voix aux élections), données d'échelle, etc.

Ces données sont, à la base, disparates et par conséquent peu lisibles et difficiles à interpréter.

La **statistique** descriptive offre un certain nombre de procédures permettant de représenter les données traitées sous une forme utilisable et significative.

Ces procédures permettent, d'une part, une **représentation graphique** des données :

— **histogramme**,

— **diagramme en bâtons**,

— **diagramme circulaire**,

— **diagramme stem and leaf**,

— **box plot**, etc.

et d'autre part, d'obtenir un ensemble d'indicateurs qui, d'une certaine façon, résument les données de base :

— **moyenne**,

— **écart type**,

— **coefficient de corrélation**,

— **indice**, etc.

La statistique descriptive comprend également les méthodes d'**analyse de données**, telles que l'**analyse factorielle des correspondances**, qui consiste à représenter graphiquement les associations entre lignes et colonnes d'une **table de contingence**.

HISTORIQUE

La première forme de **statistique** descriptive a été le **recensement** à l'époque des grandes civilisations de l'Antiquité, où l'on s'intéressait à la liste des hommes, des professions et des biens.

Voir **statistique** et **représentation graphique**.

MOTS CLÉS

Analyse de données *(Data analysis)*
Analyse factorielle des correspondances *(Correspondence analysis)*
Box plot *(Box plot)*
Dendrogramme *(Dendrogram)*
Diagramme circulaire *(Pie chart)*
Diagramme en bâtons *(Line chart)*
Diagramme stem and leaf *(Stem and leaf diagram)*
Graphique quantitatif *(Quantitative plot)*
Histogramme *(Histogram)*
Indice *(Index number)*
Recensement *(Census)*
Représentation graphique *(Graphical representation)*

STATISTIQUE GÉNÉTIQUE

Genetical statistics

La statistique génétique s'intéresse à l'analyse des données génétiques en utilisant les méthodes et les concepts de la **statistique** classique ainsi que ceux provenant de la théorie des processus stochastiques. La génétique étudie les caractères héréditaires et les variations accidentelles ; elle contribue à l'explication

du transformisme, dans le domaine pratique, à l'amélioration des espèces.

HISTORIQUE

Depuis plus de 100 ans, la génétique et la statistique se sont apparentées. Sir **Francis Galton** était très intéressé par l'étude de la génétique humaine plus particulièrement de l'eugénisme, c'est-à-dire, la science qui met en œuvre les méthodes susceptibles d'améliorer les caractères propres des populations humaines, essentiellement fondée sur les connaissances acquises en hérédité. Il est l'inventeur des coefficients de régression et de corrélation comme outils statistiques d'investigation dans la génétique. Ces méthodes ont été ensuite developpées par **Karl Pearson**. En 1900, le travail de Mendel a été redécouvert et les outils mathématiques et statistiques d'investigation classiques de la génétique de Mendel ont été perfectionnés grâce aux contributions de Sir **Ronald Aylmer Fisher**, Sewall Wright et J.B.S. Haldane dans la période 1920-1950.

DOMAINES ET LIMITATIONS

La génétique statistique est utilisée dans des domaines comme la bio-mathématique, la bio-informatique, la biologie, l'**épidémiologie**, la génétique, etc. Les méthodes standards d'**estimation** et des tests d'hypothèses sont utilisés dans le but d'estimer des paramètres génétiques et de tester ces paramètres. La théorie des processus stochastiques est utilisée afin d'étudier les propriétés de l'évolution d'un sujet d'une population en tenant compte du changement aléatoire des fréquences des gènes.

EXEMPLE

Loi de Hardy-Weinberg

La loi de Hardy-Weinberg est le modèle théorique central de la génétique des populations. Formulée indépendamment par le mathématicien anglais Godfrey H. Hardy et par le médecin allemand W. Weinberg (1908), les équations de cette loi montrent qu'une variante génétique apparue d'abord chez quelques individus d'une population ne disparaîtra pas sous l'effet de sa transmission héréditaire mais au contraire se maintiendra au fil des générations, exactement dans cette même proportion, conformément aux lois de Mendel.

La loi de Hardy-Weinberg permet, sous certaines conditions, le calcul des fréquences génotypiques à partir des fréquences allèliques. Par génotype, on entend le patrimoine génétique d'un individu dépendant des gènes hérités de ses parents ; par exemple, les vrais jumeaux ont le même génotype. Un allèle est une version possible d'un gène, constitué d'un enchainement de nucléotides. Donc à un gène peut correspondre, dans une **population** donnée, plusieurs allèles. Chez un individu, chaque gène est représenté par deux allèles qui sont soit identiques dans leur composition nucléotidique — l'individu est alors homozygote pour ce gène — soit différent dans leur composition — l'individu est alors hétérozygote pour ce gène.

Dans sa version d'origine, la loi stipule que si un gène est contrôlé par deux allèles A et B, dont la fréquence d'apparition dans la population à la génération t est de p et respectivement de $q = 1 - p$ alors la fréquence des génotypes à la génération $t + 1$ est donnée par la relation suivante :

$$p^2AA + 2pqAB + q^2BB.$$

L'hypothèse de base est que la taille de la population N est grande pour minimiser les variations d'échantillonnage ; il ne doit y avoir ni s élection, ni mutation, ni migration (pas de perte/gain d'allèle) et les génétations successives doivent être discrètes (pas de croisement entre générations différentes).

Pour qu'un individu soit AA, il faut qu'il ait reçu un allèle de type A de ses deux parents. Si le processus se fait au hasard (hypothèse d'**indépendance**), cet événément se réalise avec la probabilité :

$$P(AA) = pp = p^2.$$

Le raisonnement est identique pour le génotype BB et

$$P(BB) = qq = q^2.$$

Enfin pour le génotype AB, deux cas sont possibles : l'individu a reçu A de son père et B de sa mère ou l'inverse, d'où :

$$P(AB) = pq + qp = 2pq.$$

Ainsi dans une population idéale, les proportions de Hardy-Weinberg sont données par :

AA	AB	BB
p^2	$2pq$	q^2

Cette situation se généralise à un gène avec plusieurs allèles A_1, A_2, ..., A_k. Les fréquences des homozygotes A_iA_i sont égales à :

$$f(A_iA_i) = p_i^2, i = 1, \ldots, k$$

et celles des hétérozygotes A_iA_j sont égales à :

$$f(A_iA_j) = 2p_ip_j, i \neq j, i = 1, \ldots, k,\ j = 1, \ldots, k.$$

Ce qui nous intéresse est de savoir quelle est la fréquence p' de l'allèle A à la génération $t+1$.

Par simple comptage on a :

$$\begin{aligned} p' &= \frac{2p^2N + 2pqN}{2N} = p^2 + pq \\ &= p \cdot (p+q) = p \cdot (1-q+q) = p. \end{aligned}$$

La fréquence de l'allèle A à la génération $t+1$ est donc identique à celle de la génération précédente et donc aussi à celle de la génération initiale.

MOTS CLÉS

Donnée *(Data)*
Épidémiologie *(Epidemiology)*
Estimation *(Estimation)*
Indépendance *(Independence)*
Paramètre *(Parameter)*
Population *(Population)*
Processus stochastique *(Stochatic process)*
Statistique *(Statistics)*
Test d'hypothèse *(Hypothesis testing)*

RÉFÉRENCES

Galton, F. (1877). Typical laws of heredity. *Proceedings of the Royal Institution of Great Britain* 8 (1877), pp. 282-301, réédité dans S.M. Stigler, *The History of Statistics : The Measurement of Uncertainty Before 1900*, Cambridge, MA : Belknap Press of Harvard University Press 1986, CFig. 8.6, p. 280.

Galton, F. (1888). *Co-relations and their measurement, chiefly from anthropological data. Proceedings of the Royal Society of London.* 45, pp. 135-145.

Hardy, G.H. (1908). Mendelian Proportions in a Mixed Population. *Science*, 28, pp. 49-50.

Smith, C.A.B. (1983). Statistics in Human Genetics. In Kotz S. and Johnson N.L. (editors). *Encyclopedia of Statistical Sciences.* Vol 3. John Wiley & Sons, New York.

STATISTIQUE INFÉRENTIELLE

Inferential statistics

La statistique inférentielle ou déductive est complémentaire à la **statistique descriptive**, car le but de la plupart des recherches n'est pas d'établir un certain nombre d'indicateurs sur un échantillon donné, mais plutôt d'estimer un certain nombre de paramètres caractérisant la population associée à l'échantillon traité.

En **statistique** inférentielle, des hypothèses sont formulées sur les paramètres d'une **population**. Ces hypothèses sont ensuite testées sur la base d'observations faites sur un **échantillon** représentatif de cette population.

HISTORIQUE

L'origine de la **statistique** inférentielle coïncide avec celle de la théorie des probabilités et correspond notamment aux travaux de **Thomas Bayes** (1763), d'**Abraham de Moivre** (1718), de **Carl Friedrich Gauss** (1809) et de **Pierre Simon de Laplace** (1812).

Par la suite, les recherches dans le domaine de la statistique inférentielle ont été nombreuses. Citons par exemple les travaux de **Francis Galton** (1889) relatifs à la **corrélation**, ainsi que le développement des tests d'hypothèses dû principalement à **Karl Pearson** (1900) et à **William Sealy Gosset** dit « Student » (1908). **Jerzy Neyman** et **Egon Sharpe Pearson** (1928), ainsi que **Ronald Aylmer Fisher**

(1956) ont également contribué de façon essentielle au développement de la statistique inférentielle.

MOTS CLÉS

Échantillonnage *(Sampling)*
Estimation*(Estimation)*
Maximum de vraisemblance *(Maximum likelihood)*
Optimisation *(Optimization)*
Probabilité *(Probability)*
Sondage*(Survey)*
Statistique*(Statistics)*
Statistique descriptive *(Descriptive statistics)*
Test d'hypothèse *(Hypothesis testing)*
Théorème de Bayes *(Bayes' theorem)*

RÉFÉRENCES

Bayes, T. (1763). An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London*, 53, pp. 370-418. Publié deux ans après la mort de son auteur à l'instigation de R. Price. Réédité avec une biographie de **George A. Barnard** en 1958 et dans **Egon Sharpe Pearson** et Maurice George Kendall (eds) 1970, *Studies in the History of Statistics and Probability*, pp. 131-153.

Fisher, R.A. (1956). *Statistical Methods and Scientific Inference.* Oliver & Boyd, Edinburgh.

Galton, F. (1889). *Natural Inheritance.* The Macmillan Company, London.

Gauss, C.F. (1809). *Theoria motus corporum cœlestium.* Werke 7.

Gosset, S.W. « Student » (1908). The Probable Error of a Mean. *Biometrika*, 6, pp. 1-25.

Laplace, P.S. de (1812). *Théorie analytique des probabilités.* Courcier, Paris.

Moivre, A. de (1718). *The Doctrine of Chances: or, A Method of Calculating the Probability of Events in Play.* W. Pearson, London.

Neyman, J. and Pearson, E.S. (1928). On the use and interpretation of certain test criteria for purposes of statistical inference. I, II. *Biometrika*, 20A, pp. 175-240 and 263-294.

Pearson, K. (1900). 1948. On the Criterion, That a Given System of Deviations from the Probable in the Case of a Correlated System of Variables is such that it can be reasonably supposed to have arisen from random Sampling. pp. 339-357 in Karl Pearson, *Karl Pearson's Early Statistical Papers.* Cambridge Univ. Press. First published in the *Philosophical Magazine*, 5th Series, 50, pp. 157-175.

STATISTIQUE OFFICIELLE

Official statistics

L'adjectif *officiel* signifie : « qui émane du gouvernement, de l'administration, qui a un caractère légal ». Ainsi le mot « officiel », ajouté à la suite du mot « statistique » indique un lien avec l'État. Par ailleurs, l'étymologie de « statistique », du latin *status* (état), rappelle encore ce lien.

Le terme *statistique officielle* fait référence à un concept très général qui peut être abordé de différentes manières. En effet, on entend par statistique officielle à la fois :

— les informations chiffrées produites par un gouvernement et nécessaires à la gestion de l'État ;

— l'organisation chargée de produire ces résultats ;

— l'ensemble des méthodes statistiques utilisées pour produire ces informations.

Les statistiques offcielles, en tant qu'informations chiffrées, sont le résultat d'un long travail effectué généralement dans les instituts nationaux de statistique.

Il faut tout d'abord fixer les objectifs, le type d'enquête et son envergure. Ensuite, il s'agit de collecter les données à l'aide de l'une ou l'autre des méthodes suivantes : l'enquête, le relevé, l'exploitation de sources administratives, les statistiques de synthèses (les comptes nationaux), l'enquête par sondage.

Une fois en possession des données, on passe à la saisie des données et à leur dépouillement. Puis les données sont analysées, interprétées et commentées. Vient encore la diffusion de l'information statistique. Et, finalement, il faut prévoir le stockage des données.

HISTORIQUE

Le **recensement**, la statistique officielle la plus connue, est une pratique presque aussi vieille que l'organisation sociale des peuples. Depuis l'Antiquité, les rois de Babylone et d'Égypte, les chefs chinois, grecs, hébreux, persans et romains cherchent à connaître le nombre de soldats sous leurs ordres et les biens dont ils disposent. Les premiers essais statistiques datent donc de plus de 6 000 ans, mais ceux-ci n'étaient que des tentatives sporadiques de comptage des peuples et des biens. Les statistiques, au sens moderne du terme, n'apparaissent qu'à la fin du XVII^{e} siècle avec l'affirmation de l'État moderne et les besoins de l'administration coloniale en France et en Angleterre. Au XVIII^{e} siècle, dans l'*Encyclopédie* de Diderot et de d'Alembert, il est question de « politique arithmétique » qui est définie comme « l'application des calculs arithmétiques aux sujets et aux usages de la politique ». La Révolution française, puis les débuts de la démocratie aux États-Unis justifient de nouveaux développements des statistiques officielles. En effet, les nouveaux États, pour connaître le poids politique de leurs différentes entités, ont besoin de connaître leur population. C'est pourquoi le premier des recensements d'une série régulière de recensements décennaux s'est fait aux États-Unis en 1790.

Depuis la Deuxième Guerre mondiale, les empires, les colonies et les société tribales sont remplacés peu à peu par des États nationaux. Afin d'organiser élections et assemblées, et favoriser le développement d'entreprises modernes, ces nouveaux pays doivent mettre en place des systèmes statistiques qui sont les conditions essentielles de leur développement.

Les Nations unies, et en particulier sa Commission statistique, ont beaucoup fait pour la promotion et le développement des statistiques officielles dans les différents pays.

Sous leur auspice, le premier recensement mondial de la population a été organisé en 1950. Un autre a suivi en 1960.

Actuellement les statistiques officielles se développent un peu partout. L'Organisation pour la coopération et le développement économique (OCDE) veut produire des statistiques susceptibles d'aider à la formulation d'un jugement concis, exhaustif et équilibré sur l'état des principaux secteurs de la société. Ces statistiques, appelées indicateurs, font l'objet, depuis les années 60, de nombreuses recherches en économie, dans le domaine de la santé et celui de l'enseignement. Le but étant d'installer, pour les politiciens, un système d'informations statistiques permettant de contrôler les effets de leur politique économique, sociale et éducative.

EXEMPLES

La statistique officielle traite de très nombreux domaines dont voici les plus importants :

- *Population* (structure de la population, composition des familles et des ménages, naissances, adoptions, reconnaissances, décès, mariages, divorces, migrations...)
- *Espace, paysage et environnement* (sol, climat, paysage, eau, bruit, déchets, faune et flore...)
- *Emploi et vie active* (entreprises et établissements, emploi, chômeurs, salaires, conditions de travail...)
- *Comptes nationaux* (comptabilité nationale, balance des paiements...)
- *Prix* (prix et *indice* des prix des biens et des services...)
- *Production, commerce et consommation* (résultats comptables des entreprises, production, chiffre d'affaires, commerce extérieur...)
- *Agriculture et sylviculture* (exploitations, main-d'œuvre, machines, cheptel, forêts, pêche...)
- *Énergie* (bilan énergétique, production et consommation d'énergie...)

- *Construction et logement* (structure de parcs d'immeubles et de logements, conditions d'habitation, logements vacants...)
- *Tourisme* (offre et demande dans l'hôtellerie et la para-hôtellerie, voyages...)
- *Transports et communications* (véhicules, installations, trafic, accidents de la circulation routière, comptes ferroviaires...)
- *Masse monétaire, marché financier et banques* (système banquaire, masse monétaire, comptabilité financière...)
- *Assurances* (assurance vieillesse, assurance maladie, assurances privées...)
- *Santé* (état de santé, cause de maladies et de décès, personnel, équipement, coûts de la santé...)
- *Éducation et science* (école, formation professionnnelle et formation complémentaire, recherche et développement...)
- *Sport, culture et conditions de vie* (langue, confession, activités culturelles et sport...)
- *Politique* (élections, participation à la vie politique...)
- *Finances publiques* (recettes et dépenses de l'État et des régions, charge fiscale...)
- *Droit et justice* (enquêtes et instructions pénales, jugements, exécution des peines, récidives...).

DOMAINES ET LIMITATIONS

L'histoire montre l'influence mutuelle des statistiques sur l'évolution de la nature de l'État et du pouvoir sur le développement des statistiques. Cette interdépendance explique, en partie du moins, la difficulté qu'éprouvent les statisticiens à fixer des limites claires à cette discipline. En effet, tout au long de l'histoire et aujourd'hui encore, le pouvoir délimite le champ d'étude des statistiques officielles, élargissant ou restreignant leur contenu en fonction de ses propres besoins.

Choix politique

Le problème du choix politique représente l'une des particularités essentielles des statistiques officielles par rapport aux statistiques théoriques. En effet, le choix politique intervient dans de nombreux domaines des statistiques officielles : lors de la définition des concepts comme : chômage, pauvreté, bien-être ; lors du codage de certaines variables comme : catégories professionnelles, niveau des salaires ; lors de l'énumération des différents articles entrant dans la définition du panier du consommateur pour le calcul de l'indice des prix... Ce problème du choix conduit à deux autres problèmes :

- *La comparabilité des statistiques*
  Les besoins et les objectifs de l'État dictent les choix dans ce domaine. Les besoins n'étant pas toujours les mêmes d'un État à l'autre, les choix seront parfois différents et les statistiques internationales ne seront pas toujours comparables. La différence de moyens des États pour la récolte et l'exploitation des données explique également les disparités des résultats entre les pays.
- *L'objectivité*
  La statistique officielle revendique cependant l'objectivité. S'il est vrai que la part de volonté politique ne peut être complètement éliminée, l'évolution des théories statistiques et l'amélioration des outils permettent des récoltes de données beaucoup plus importantes et plus fiables. La qualité et la quantité des données traitées ont beaucoup augmenté. Les informations y ont gagné en objectivité.

MOTS CLÉS

Donnée *(Data)*
Indice *(Index)*
Indicateur *(Indicator)*
Population *(Population)*
Recensement *(Census)*

RÉFÉRENCES

Duncan, J.W. and Gross, A.C. (1995). *Statistics for the 21st Century, Proposals for Improving Statistics for Better Decision Making.* Irwin Professional Publishing.

Dupâquier, J. et Dupâquier, M. (1985). *Histoire de la démographie.* Académie Perrin, Paris.

INSEE (1977). *Pour une histoire de la statistique. Tome 1: Contributions.* INSEE, Economica, Paris.

INSEE (1987). *Pour une histoire de la statistique. Tome 2: Matériaux.* (J. Affichard éd.) INSEE, Economica, Paris.

ISI (1994). *The Future of Statistics: An International Perspective.* Voorburg.

Stigler, S. (1986). *The History of Statistics, the Measurement of Uncertainty before 1900.* The Belknap Press of Harvard University Press. London, England.

United Nations Statistical Commission, *Fundamental Principles of Official Statistics.*

STATISTIQUE SPATIALE

Spatial statistics

La statistique spatiale confère un modèle aux données à référence spatiale (**données spatiales**) variant dans le temps. La statistique spatiale couvre l'ensemble des techniques utilisées pour explorer et démontrer la présence de la dépendance spatiale entre les observations reparties dans l'espace. Les données spatiales peuvent provenir de la géologie, des sciences de la terre, du traitement des images, de l'**épidémiologie**, l'agriculture, l'écologie, l'astronomie ou des sciences forestières. Les données sont supposées aléatoires et parfois leur emplacement est aussi supposé aléatoire.

HISTORIQUE

La première manifestation de la statistique traitant des données spatiales appartient à Edmond Halley (1686). Il superposa sur une carte des reliefs les directions des vents et moussons entre et autour des tropiques, et essaya d'expliquer leur causes. Les modèles de la statistique spatiale sont apparus bien plus tard. **Ronald Aylmer Fisher** (1920-1930), durant les recherches menées à la station expérimentale de Rothamsted en Angleterre, formulera les bases des principes du choix aléatoire, de l'analyse par blocs ainsi que de la replication.

ASPECTS MATHÉMATIQUES

L'analyse statistique des données spatiales forme le sujet principal de la statistique spatiale. Les principaux outils de l'analyse statistique sont : l'histogramme, la courbe du cumul des fréquences, le **Q-Q plot**, la **médiane**, la **moyenne** de la distribution, les coefficients de variation et la **mesure d'asymétrie**. Dans le cas de plusieurs variables, la représentation sous forme de scatterplot des données ainsi que l'étude de la corrélation sont les outils statistiques les plus utilisés. La particularité que présente l'analyse spatiale est la prise en compte de l'information spatiale et de la relation qui existe entre les données spatiales. Deux données proches l'une de l'autre auront une probabilité de ressemblance plus grande que si elles étaient plus éloignées. Ainsi, lorsqu'on analyse une carte de concentrations de polluants dans un sol, on s'attendra à ce que les valeurs apparaissent non pas d'une manière aléatoire mais, au contraire, que les faibles valeurs du polluant soient groupées de même que les grandes valeurs et que la transition entre ces classes de valeurs se fasse de manière continue.

DOMAINES ET LIMITATIONS

La représentation spatiale sous forme des cartes et plans et la statistique spatiale peuvent être réunies pour offrir une compréhension visuelle de l'influence qu'ont les valeurs voisines les unes sur les autres. Le fait de superposer l'information statistique sur un support cartographié permet une analyse plus aisée de l'information et de tirer les conclusions appropiées.

Les domaines d'application de la statistique spatiale sont nombreux. On peut traiter des données relatives à la pollution terrestre, du traitement de l'image satellitaire, des données concernant la migration des maladies. Des nombreux exemples d'applications se trouvent dans la littérature : modélisation du taux de fertilité d'une région, études de la migration des peuplades, etc.

EXEMPLES

Voir **données spatiales**.

MOTS CLÉS

Autocorrélation et autocorrélation partielle *(Autocorrelation and partial autocorrelation)*
Données spatiales *(Spatial data)*
Géostatistique *(Geostatistics)*

RÉFÉRENCES

Cressie, Noel A.C. (1991). *Statistics for Spatial Data.* John Wiley & Sons, New York.

Goovaerts, P. (1997). *Geostatistics for Natural Resources Evaluation.* Oxford : Oxford University Press.

Isaaks, E.H. and Srivastava, R.M. (1989). *An Introduction to Applied Geostatistics*, Oxford : Oxford University Press.

Journel, A. and Huijbregts, C.J. (1981). *Mining Geostatistics*, Academic Press, New York.

Matheron, G. (1965). *Les variables régionalisées et leur estimation.* Masson. Paris.

Oliver, M.A. and Webster, R. (1990). Kriging : a method of interpolation for geographical information system, *Int. J. Geographical Information Systems*, Vol. 4, No. 3, pp. 313-332.

STEM AND LEAF

Stem and leaf

Voir **diagramme stem and leaf**.

STUDENT

Student

Voir **Gosset William Sealy**.

STUDENT, LOI DE

Student distribution

Voir **loi de Student**.

STUDENT, TABLE DE

Student table

Voir **table de Student** et annexe E.

STUDENT, TEST DE

Student test

Voir **test de Student**.

STUDENT POUR OBSERVATIONS PAIRÉES, TEST DE

Matched pairs Student test

Voir **test de Student pour observations pairées**.



T

TABLE BINOMIALE

Binomial table

La table binomiale donne les valeurs de la **fonction de répartition** d'une **variable aléatoire** suivant une **loi binomiale**.

HISTORIQUE

Voir **loi binomiale**.

ASPECTS MATHÉMATIQUES

Soit la **variable aléatoire** X suivant une **loi binomiale** de paramètres n et p. Sa **fonction de probabilité** est donnée par :

$$P\left(X=x\right)=C_n^x\cdot p^x\cdot q^{n-x},\ x=1,2,\ldots,n$$

où C_n^x est le coefficient binomial égal à $\frac{n!}{x!(n-x)!}$, le paramètre p est la **probabilité** de « succès » et $q=1-p$ est la probabilité complémentaire qui correspond à la probabilité d'« échec » (voir **loi normale**).

La **fonction de répartition** de la variable aléatoire X est définie par :

$$P(X\leq x)=\sum_{i=0}^{x}C_n^i\cdot p^i\cdot q^{n-i},\ 0\leq x\leq n.$$

La table binomiale fournit les valeurs de $P(X\leq x)$ en fonction de x et pour différentes valeurs de n et p.

Pour de grandes valeurs de n, le calcul devient fastidieux. Il existe alors de très bonnes approximations. Si $\min(np,n(1-p))>10$, on peut faire une approximation par la loi normale :

$$P(X\leq x)=\phi\left(\frac{x+\frac{1}{2}-np}{\sqrt{npq}}\right)$$

où ϕ est la fonction de répartition de la loi normale standard, incluant la correction de continuité $\frac{1}{2}$.

DOMAINES ET LIMITATIONS

La table binomiale est utilisée pour effectuer des tests non paramétriques basés sur une **statistique** distribuée selon une **loi binomiale**, notamment le **test du signe** et le **test binomial**.

Le National Bureau of Standards (1950) a publié les probabilités individuelles et cumulées de la loi binomiale pour $n\leq 49$. Les probabilités cumulées de la loi binomiale pour $n\leq 1\,000$ se trouvent dans les *Annals of the Computation Laboratory* (1955).

EXEMPLES

Voir annexe B.

On peut vérifier que pour $n=2$ et $p=0,5$:

$$\begin{aligned}P\left(X\leq 1\right) &= \sum_{i=0}^{1}C_2^i\left(0,5\right)^i\left(0,5\right)^{2-i}\\ &= 0,75\end{aligned}$$

ou pour $n=5$ et $p=0,05$:

$$\begin{aligned}P\left(X\leq 3\right) &= \sum_{i=0}^{3}C_5^i\left(0,05\right)^i\left(0,95\right)^{5-i}\\ &= 1,0000.\end{aligned}$$

Pour un exemple de l'utilisation de la table binomiale, voir **test binomial**.

MOTS CLÉS

Loi binomiale *(Binomial distribution)*
Table statistique *(Statistical table)*
Test binomial *(Binomial test)*
Test du signe *(Sign test)*

RÉFÉRENCES

Annals of the Computation Laboratory. (1955). *Tables of the Cumulative Binomial Probability Distribution.* Vol. 35, Harvard University.

National Bureau of Standards. (1950). Tables of the Binomial Probability Distribution. *Applied Mathematics Series* 6.

TABLE DE CONTINGENCE

Contigency table

Une table de contingence, appelée aussi tableau de contingence, est un tableau croisé dans lequel sont répertoriés les différents attributs de la population ou de l'échantillon observé. L'analyse des tables de contingence consiste à découvrir et à étudier les relations (si elles existent) entres les attributs.

Une table de contingence peut être un tableau à double entrée, comprenant r lignes et c colonnes, relatif à deux variables qualitatives catégorielles possédant respectivement r et c catégories ou aussi, quand le nombre de variables qualitatives catégorielles est supérieur à deux, de forme multi-dimensionnelle : si, par exemple, les éléments d'une **population** ou d'un **échantillon** sont caractérisés par trois attributs, la table de contingence associée est de dimension $I \times J \times K$, où I représente le nombre de catégories définissant le premier attribut, J le nombre de catégories du deuxième attribut et K le nombre de catégories du troisième attribut.

HISTORIQUE

Le terme de contingence, utilisé en rapport avec des tableaux croisés de données catégorielles, semble avoir son origine avec **Karl Pearson** (1904), pour lequel la contingence est une mesure de la **déviation** totale par rapport à l'**indépendance**.

Voir aussi **test d'indépendance du chi-carré**.

ASPECTS MATHÉMATIQUES

Si on considère un tableau à double entrée, alors pour deux variables qualitatives catégorielles X et Y possédant respectivement r et c catégories la table de contingence est :

		Catégories de la variable Y			
		Y_1	...	Y_c	Total
Catégories de la variable X	X_1	n_{11}	...	n_{1c}	$n_{1.}$
	...	...	...	...	...
	X_r	n_{r1}	...	n_{rc}	$n_{r.}$
	Total	$n_{.1}$	...	$n_{.c}$	$n_{..}$

où

- n_{ij} représente la **fréquence** observée pour la catégorie i de la **variable** X et la catégorie j de la variable Y,
- $n_{i.}$ représente la somme des fréquences observées pour la catégorie i de la variable X,
- $n_{.j}$ représente la somme des fréquences observées pour la catégorie j de la variable Y et
- $n_{..}$ indique le nombre total d'observations.

Dans le cas d'un tableau multidimensionnel, les éléments de la table sont notés par n_{ijk} et représentent la fréquence observée pour la catégorie i de la variable X, la catégorie j de la variable Y et la catégorie k de la variable Z.

DOMAINES ET LIMITATIONS

L'**indépendance** des deux variables qualitatives catégorielles représentées dans une table de contingence peut être testée par un **test d'indépendance du chi-carré**.

MOTS CLÉS

Distribution de fréquences *(Frequency distribution)*
Test d'indépendance du chi-carré *(Chi-square test of independence)*

RÉFÉRENCES

Fienberg, S.E. (1980). *The Analysis of Cross-Classified Categorical Data*, 2nd ed. MIT Press, Cambridge, Mass.

Pearson, K. (1904). On the theory of contingency and its relation to association and normal correlation. *Drapers' Company Research Memoirs, Biometric Ser.* I. pp. 1-35.

TABLE DE FISHER

Fisher table

La table de Fisher donne les valeurs de la **fonction de répartition** d'une **variable aléatoire** suivant une **loi de Fisher**.

HISTORIQUE

Voir **loi de Fisher**.

ASPECTS MATHÉMATIQUES

Soit la **variable aléatoire** X suivant une **loi de Fisher** avec m et n degrés de liberté. La **fonction de densité** de la variable aléatoire X est donnée par :

$$f(t) = \frac{\Gamma\left(\frac{m+n}{2}\right)}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)}\left(\frac{m}{n}\right)^{\frac{m}{2}} \cdot \frac{t^{\frac{m-2}{2}}}{\left(1+\frac{m}{n}t\right)^{\frac{m+n}{2}}}, \quad t \geq 0$$

où Γ représente la fonction gamma (voir **loi gamma**).

Pour les nombres de degrés de liberté m et n, la table de Fisher permet de déterminer la **valeur** x pour laquelle la **probabilité** $P(X \leq x)$ est égale à une valeur prédéterminée.

Les tables les plus courantes donnent les valeurs de x pour

$$P(X \leq x) = 95\,\% \text{ et } (X \leq x) = 97,5\,\%.$$

On désigne généralement par $x = F_{m,n,\alpha}$, la valeur de la variable aléatoire X pour laquelle

$$P(X \leq F_{m,n,\alpha}) = 1 - \alpha.$$

DOMAINES ET LIMITATIONS

La table de Fisher est utilisée dans les tests d'hypothèse faisant intervenir une **statistique** distribuée selon une **loi de Fisher** et notamment dans l'**analyse de variance** et dans l'**analyse de régression linéaire simple** et **multiple**.

Merrington et Thompson (1943) donnent des tables allant jusqu'à 5 décimales et Fisher et Yates (1953) des tables à 2 décimales.

EXEMPLES

Voir annexe C.

Ainsi, si X suit une **loi de Fisher** avec $m = 10$ et $n = 15$ degrés de liberté, la **valeur** x correspondant à une **probabilité** de $0,95$ est égale à $2,54$:

$$P(X \leq 2,54) = 0,95.$$

De même, la valeur correspondant à une probabilité de $0,975$ est égale à $3,06$:

$$P(X \leq 3,06) = 0,975.$$

Pour un exemple de l'utilisation de la table de Fisher, voir **test de Fisher**.

MOTS CLÉS

Analyse de variance *(Analysis of variance)*
Loi de Fisher *(Fisher distribution)*
Régression linéaire multiple *(Multiple linear regression)*
Régression linéaire simple *(Simple linear regression)*
Table statistique *(Statistical table)*

RÉFÉRENCES

Fisher, R.A. and Yates, F. (1953). *Statistical Tables for Biological, Agricultural and Medical Research*, (4^{th} ed.). Oliver & Boyd, London.

Merrington, M. and Thompson, C.M. (1944). Tables of percentage points of the inverted beta (F) distribution. *Biometrika*, 33, pp. 73-88.

TABLE DE KRUSKAL-WALLIS

Kruskal-Wallis table

La table de Kruskal-Wallis donne les valeurs théoriques de la **statistique** H du **test de Kruskal-Wallis** sous l'**hypothèse** qu'il n'y a pas de différence entre les k ($k \geq 2$) populations que l'on désire comparer.

HISTORIQUE

Voir **test de Kruskal-Wallis**.

ASPECTS MATHÉMATIQUES

Soient k échantillons de tailles éventuellement différentes $n_1, n_2, \ldots, n_k$. On désigne par N le nombre total d'observations :

$$N = \sum_{i=1}^{k} n_i.$$

On classe ces N observations par ordre croissant sans tenir compte de l'appartenance aux échantillons. On attribue ensuite le rang 1 à la plus petite **valeur**, le rang 2 à la valeur juste supérieure et ainsi de suite jusqu'au rang N attribué à la plus grande valeur.

On note R_i la somme des rangs attribués aux observations de l'échantillon i :

$$R_i = \sum_{j=1}^{n_i} R\left(X_{ij}\right),\ i = 1, 2, \ldots, k$$

où X_{ij} représente l'observation j de l'échantillon i. Lorsque plusieurs observations sont identiques et ont donc le même rang, on leur attribue un rang moyen (voir **test de Kruskal-Wallis)**. S'il n'y a pas de rangs moyens, le test statistique est défini de la façon suivante :

$$H = \left(\frac{12}{N(N+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i}\right) - 3\,(N+1)\,.$$

Pour savoir que faire s'il y a des rangs moyens, voir **test de Kruskal-Wallis**.

La table de Kruskal-Wallis donne les valeurs de la **statistique** H du test de Kruskal-Wallis dans le cas de trois échantillons, pour différentes valeurs de n_1, n_2 et n_3 (avec $n_1, n_2, n_3 \leq 5$).

DOMAINES ET LIMITATIONS

La table de Kruskal-Wallis est utilisée pour les tests non paramétriques qui font intervenir le rang et plus particulièrement pour le test du même nom.

Lorsque le nombre d'échantillons est supérieur à 3, on peut faire une approximation de la **valeur** de la table de Kruskal-Wallis par la **table du chi-carré** avec $k-1$ degrés de liberté.

EXEMPLES

Voir annexe D.

Pour un exemple de l'utilisation de la table de Kruskal-Wallis, voir **test de Kruskal-Wallis**.

MOTS CLÉS

Table du chi-carré *(Chi-square table)*
Table statistique *(Statistical table)*
Test de Kruskal-Wallis *(Kruskal-Wallis test)*

RÉFÉRENCE

Kruskal, W.H. and Wallis, W.A. (1952). Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association*, 47, pp. 583-621 and errata, *ibid.*, 48, pp. 907-911.

TABLE DE SPEARMAN

Spearman table

La table de Spearman donne les valeurs théoriques du coefficient de corrélation des rangs de Spearman sous l'**hypothèse** d'**indépendance** de deux variables aléatoires.

HISTORIQUE

Voir **coefficient de corrélation des rangs de Spearman**.

MOT CLÉ

Coefficient de corrélation des rangs de Spearman *(Spearman rank correlation coefficient)*

TABLE DE STUDENT

Student table

La table de Student donne les valeurs de la **fonction de répartition** d'une **variable aléatoire** suivant une **loi de Student**.

HISTORIQUE

William Sealy Gosset, dit « Student », développa, en 1908, le **test de Student** et la table portant son nom.

ASPECTS MATHÉMATIQUES

Soit la **variable aléatoire** T suivant une **loi de Student** avec v degrés de liberté. La **fonction de densité** de la

variable aléatoire T est donnée par :

$$f(s) = \frac{\Gamma\left(\frac{v+1}{2}\right)}{\sqrt{v\pi} \cdot \Gamma\left(\frac{v}{2}\right)} \left(1 + \frac{s^2}{v}\right)^{-(v+1)/2}$$

où Γ représente la fonction gamma (voir **loi gamma**). La **fonction de répartition** de la variable aléatoire T est définie par :

$$F(t) = P(T \leq t) = \int_{-\infty}^{t} f(s)\, ds.$$

Pour chaque **valeur** de v, la table de Student fournit la valeur de la fonction de répartition $F(t)$.

La loi de Student étant symétrique par rapport à l'origine, on a :

$$F(t) = 1 - F(-t),$$

ce qui permet de déterminer $F(t)$ pour une valeur négative de t.

La table de Student est souvent utilisée en sens inverse : pour trouver la valeur de t correspondant à une **probabilité** donnée. Dans ce cas, la table de Student donne la **valeur critique** pour le nombre de **degrés de liberté** v et le **seuil de signification** α. On note généralement $t_{v,\alpha}$ la valeur de la variable aléatoire T pour laquelle :

$$P(T \leq t_{v,\alpha}) = 1 - \alpha.$$

DOMAINES ET LIMITATIONS

Voir **test de Student**.

EXEMPLES

Voir annexe E.

On considère un **test de Student** unilatéral sur un **échantillon** de taille n. Pour le nombre de degrés de liberté $v = n - 1$ donné, la table de Student permet de déterminer la **valeur critique** $t_{v,\alpha}$ du test pour le **seuil de signification** α donné.

Pour un exemple d'utilisation de la table de Student, voir **test de Student**.

MOTS CLÉS

Loi de Student *(Student distribution)*
Table statistique *(Statistical table)*
Test de Student *(Student test)*

RÉFÉRENCES

Fisher, R.A. (1925). Applications of "Student's" Distribution. *Metron*, 5, pp. 90-104.

Gosset, S.W. « Student » (1908). The Probable Error of a Mean. *Biometrika*, 6, pp. 1-25.

Gosset, S.W. « Student » (1925). New tables for testing the significance of observations. *Metron*, 5, pp. 105-120.

TABLE DE WILCOXON

Wilcoxon table

La table de Wilcoxon donne les valeurs théoriques de la **statistique** T du **test de Wilcoxon** sous l'**hypothèse** qu'il n'y a pas de différence entre les distributions des deux populations comparées.

HISTORIQUE

Les valeurs critiques pour $N \leq 20$ (N est le nombre total d'observations) ont été calculées par Frank Wilcoxon (1947).

ASPECTS MATHÉMATIQUES

Soient $(X_1, X_2, \ldots, X_n)$ un **échantillon** de taille n provenant d'une **population** 1 et $(Y_1, Y_2, \ldots, Y_m)$ un échantillon de taille m provenant d'une population 2.

On obtient ainsi $N = n + m$ observations que l'on va classer dans un ordre croissant sans tenir compte de l'appartenance aux échantillons. On attribue un rang à chaque valeur : le rang 1 à la plus petit valeur, le rang 2 à la valeur suivante et ainsi de suite jusqu'au rang N à la plus grande valeur.

On définit le test statistique T de la façon suivante :

$$T = \sum_{i=1}^{n} R(X_i)$$

où $R(X_i)$ est le rang attribué à l'observation X_i, $i = 1, 2, \ldots, n$ par rapport à l'ensemble des deux échantillons $(X_1, \ldots, X_n)$ et $(Y_1, \ldots, Y_m)$.

Pour différentes valeurs de n, de m et de α (le **seuil de signification**), la table de Wilcoxon donne les valeurs théoriques de la **statistique** T du **test de Wilcoxon**, sous l'**hypothèse** que les deux populations sous-jacentes sont distribuées identiquement.

DOMAINES ET LIMITATIONS

La table de Wilcoxon est utilisée dans les tests non paramétriques qui font intervenir le rang et notamment dans le **test de Wilcoxon**.

EXEMPLE

Voici un extrait de la table de Wilcoxon pour $n = 14$, $m = 11$ et $\alpha = 0,05$ et $0,95$:

n	m	α	valeur critique
14	11	0,05	152
14	11	0,95	212

Pour un exemple de l'utilisation de la table de Wilcoxon, voir **test de Wilcoxon**.

MOTS CLÉS

Table statistique *(Statistical table)*
Test de Wilcoxon *(Wilcoxon test)*

RÉFÉRENCE

Wilcoxon, F. (1947). Probability Tables for Individual Comparisons by Ranking Methods. *Biometrics*, 3, pp. 119-122.

TABLE DE WILCOXON SIGNÉ

Wilcoxon signed table

La table de Wilcoxon signé donne les valeurs théoriques de la **statistique** T du **test de Wilcoxon signé** appliqué à deux échantillons appariés, sous l'**hypothèse** que les deux populations suivent la même loi.

HISTORIQUE

Owen (1962) a publié une table de Wilcoxon signé pour $n \leq 20$ (n est les nombre de paires d'observations) pour le **test de Wilcoxon signé** développé par Frank Wilcoxon (1945).

La table généralement utilisée est celle de Harter et Owen (1970) établie pour $n \leq 50$.

ASPECTS MATHÉMATIQUES

Soit un ensemble de n paires d'observations $((X_1, Y_1), \ldots, (X_n, Y_n))$. On calcule les différences absolues $|X_i - Y_i|$ auxquelles on attribue un rang (c'est-à-dire que l'on associe le rang 1 à la plus petite valeur, le rang 2 à celle qui suit et ainsi de suite, jusqu'au rang n pour la plus grande valeur). Puis on donne à ce rang le signe correspondant à la différence $X_i - Y_i$. On calcule alors la somme des rangs positifs :

$$T = \sum_{i=1}^{k} R_i$$

où k est le nombre de paires présentant une différence positive $(X_i - Y_i > 0)$.

Pour différentes valeurs de m et de α (le **seuil de signification**), la table de Wilcoxon signé donne les valeurs théoriques de la **statistique** T du **test de Wilcoxon signé**, sous l'**hypothèse** que les deux populations sous-jacentes sont distribuées identiquement.

DOMAINES ET LIMITATIONS

La table de Wilcoxon signé est utilisée dans les tests non paramétriques qui font intervenir le rang et notamment dans le **test de Wilcoxon signée**.

EXEMPLES

Voici un extrait de la table de Wilcoxon signé pour m allant de 10 à 15 et $\alpha = 0,05$ d'une part, et $\alpha = 0,975$ d'autre part :

m	$\alpha = 0,05$	$\alpha = 0,975$
10	11	46
11	14	55
12	18	64
3	22	73
14	26	83
15	31	94

Pour un exemple de l'utilisation de la table de Wilcoxon signé, voir **test de Wilcoxon signé**.

MOTS CLÉS

Table statistique *(Statistical table)*
Test de Wilcoxon signé *(Wilcoxon signed test)*

RÉFÉRENCES

Harter, H.L. and Owen, D.B. (1970). *Selected Tables in Mathematical Statistics*, Vol. 1. Markham, Chicago. (5.7 Appendix).

Owen, D.B. (1962). *Handbook of Statistical Tables.* Addison-Wesley, Reading, Mass. (3.3, 5.7, 5.8, 5.10).

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics*, 1, pp. 80-83. (5.1, 5.7).

TABLE DU CHI-CARRÉ

Chi-square table

La table du chi-carré donne les valeurs de la **fonction de répartition** d'une **variable aléatoire** suivant une **loi du chi-carré**.

HISTORIQUE

Une des premières tables du chi-carré a été publiée en 1902, par Elderton. Elle contient les valeurs de la **fonction de répartition** à six décimales.

En 1922, **Karl Pearson** édita la table de la fonction gamma incomplète, à sept décimales.

ASPECTS MATHÉMATIQUES

Soit la **variable aléatoire** X suivant une **loi du chi-carré** avec v degrés de liberté. La **fonction de densité** de la variable aléatoire X est donnée par :

$$f(t) = \frac{t^{\frac{v}{2}-1} \exp\left(-\frac{t}{2}\right)}{2^{\frac{v}{2}} \Gamma\left(\frac{v}{2}\right)}, \quad t \geq 0$$

où Γ représente la fonction gamma (voir **loi gamma**).

La **fonction de répartition** de la variable aléatoire X est définie par :

$$F(x) = P(X \leq x) = \int_0^x f(t)\, dt.$$

La table du chi-carré fournit les valeurs de la fonction de répartition $F(x)$ pour différentes valeurs de v.

On utilise souvent la table du chi-carré, en sens inverse, pour trouver la valeur de x correspondant à une **probabilité** donnée.

On note généralement $\chi^2_{v,\alpha}$ la valeur de la variable aléatoire X pour laquelle

$$P\left(X \leq \chi^2_{v,\alpha}\right) = 1 - \alpha.$$

Remarque : la notation χ^2 se lit « chi-carré ».

EXEMPLES

Voir annexe F.

La table du chi-carré permet, pour un nombre de degrés de liberté v donné, de déterminer :

1. la **valeur** de la **fonction de répartition** $F(x)$, x étant donné ;
2. la valeur de $\chi^2_{v,\alpha}$, la **probabilité** $P\left(X \leq \chi^2_{v,\alpha}\right)$ étant donnée.

MOTS CLÉS

Loi du chi-carré *(Chi-square distribution)*
Table statistique *(Statistical table)*
Test d'adéquation du chi-carré *(Chi-square goodness of fit test)*
Test d'indépendance du chi-carré *(Chi-square test of independence)*
Test du chi-carré *(Chi-square test)*

RÉFÉRENCES

Elderton, W.P. (1902). Tables for testing the goodness of fit of theory to observation. *Biometrika*, 1, pp. 155-163.

Pearson, K. (1922). *Tables of the Incomplete* Γ*-function*. H.M. Stationery Office (depuis 1934 : Cambridge University Press, Cambridge), London.

TABLE NORMALE

Normal table

La table normale, aussi appelée table de Gauss ou table de la **loi normale**, donne les valeurs de la **fonction de répartition** d'une **variable aléatoire** suivant une loi normale standard, c'est-à-dire de **moyenne** égale à 0 et de **variance** égale à 1.

HISTORIQUE

Pierre Simon de Laplace (1774) obtint la **loi normale** à partir de la **loi hypergéométrique** et, dans un second ouvrage (daté de 1778 mais publié en 1781), il donna une table normale.

Egon Sharpe Pearson et Hartley (1948) éditèrent une table normale basée sur les valeurs calculées par Sheppard (1903 et 1907).

Des listes exhaustives de tables normales sur le marché sont données par le National Bureau of Standards (jusqu'en 1952) et par Greenwood et Hartley (jusqu'en 1958).

DOMAINES ET LIMITATIONS

Voir **théorème central limite**.

ASPECTS MATHÉMATIQUES

Soit la **variable aléatoire** Z suivant une **loi normale** de **moyenne** 0 et de **variance** 1.

La **fonction de densité** $f(t)$ est donc égale à :

$$f(t) = \frac{1}{\sqrt{2\pi}} \cdot \exp\left(-\frac{t^2}{2}\right).$$

La **fonction de répartition** de la variable aléatoire Z est définie par :

$$F(z) = P(Z \leq z) = \int_{-\infty}^{z} f(t)\, dt.$$

La table normale présentée ici fournit les valeurs de $F(z)$ pour les différentes valeurs de z comprises entre 0 et $3,5$.

La **loi normale** étant symétrique par rapport à la moyenne, nous avons :

$$F(z) = 1 - F(-z),$$

ce qui nous permet de déterminer $F(z)$ pour une valeur négative de z.

Nous sommes parfois amenés à utiliser la table normale, en sens inverse, pour trouver la valeur de z correspondant à une probabilité donnée.

Nous notons généralement $z = z_\alpha$ la valeur de la variable aléatoire Z pour laquelle

$$P(Z \leq z_\alpha) = 1 - \alpha.$$

EXEMPLES

Voir annexe G.

Exemples d'utilisation de la table normale :

$$P(Z \leq 2,5) = 0,9938$$

$$\begin{aligned} P(1 \leq Z \leq 2) &= P(Z \leq 2) - P(Z \leq 1) \\ &= 0,9772 - 0,8413 \\ &= 0,1359 \end{aligned}$$

$$\begin{aligned} P(Z \leq -1) &= 1 - P(Z \leq 1) \\ &= 1 - 0,8413 \\ &= 0,1587 \end{aligned}$$

$$\begin{aligned} P(-0.5 \leq Z \leq 1,5) &= P(Z \leq 1,5) - P(Z \leq -0,5) \\ &= P(Z \leq 1,5) \\ &\quad -[1 - P(Z \leq 0,5)] \\ &= 0,9332 - [1 - 0,6915] \\ &= 0,9332 - 0,3085 \\ &= 0,6247. \end{aligned}$$

Inversément, nous pouvons utiliser la table normale dans le but de déterminer la borne z_α pour laquelle la **probabilité** que la **variable aléatoire** Z prenne une valeur inférieure à cette borne soit égale à une certaine **valeur** donnée.

$$\begin{aligned} P(Z \leq z_\alpha) = 0,95 \quad &\Rightarrow \quad z_\alpha = 1,645 \\ P(Z \leq z_\alpha) = 0,975 \quad &\Rightarrow \quad z_\alpha = 1,960. \end{aligned}$$

Ceci nous permet, dans un **test d'hypothèse**, de déterminer la **valeur critique** z_α relative à un **seuil de signification** α donné.

Exemple d'application

Dans le cas d'un **test unilatéral**, si le seuil de signification α est égal à 5 %, nous pouvons déterminer la valeur critique $z_{0,05}$ de la manière suivante :

$$\begin{aligned} P(Z \leq z_{0,05}) &= 1 - \alpha \\ &= 1 - 0,05 \\ &= 0,95 \\ \Rightarrow z_{0,05} &= 1,645. \end{aligned}$$

Dans le cas d'un **test bilatéral**, il faut rechercher la valeur $z_{\alpha/2}$. Avec un seuil de signification de 5 %, la valeur critique $z_{0,025}$ est donc obtenue de la manière suivante :

$$\begin{aligned} P(Z \leq z_{0,025}) &= 1 - \frac{\alpha}{2} \\ &= 1 - 0,025 \\ &= 0,975 \\ \Rightarrow z_{0,025} &= 1,960. \end{aligned}$$

MOTS CLÉS

Loi normale *(Normal distribution)*
Table statistique *(Statistical table)*
Test bilatéral *(Two-tail test)*
Test d'hypothèse *(Hypothesis testing)*
Test unilatéral *(One-tail test)*

RÉFÉRENCES

Greenwood, J.A. and Hartley, H.O. (1962). *Guide to Tables in Mathematical Statistics.* Princeton University Press, Princeton NJ.

Laplace, P.S. de (1774). *Mémoire sur la probabilité des causes par les événements.* Mémoire de l'Académie royale des sciences présenté par divers savants, 6, pp. 621-656 ou Vol. 8, pp. 27-65. In P.S. de Laplace, *Œuvres complètes* (1891). Gauthier-Villars, Paris.

Laplace, P.S. de (1781). Mémoire sur les probabilités. *Mémoires de l'Académie royale des sciences de Paris*, 1778, pp. 227-332 ou Vol. 9 pp. 383-485 dans P.S. de Laplace, *Œuvres complètes* (1891). Gauthier-Villars, Paris.

National Bureau of Standards. (1952). *A Guide to Tables of the Normal Probability Integral.* U.S. Department of Commerce. Applied Mathematics Series, 21.

Pearson, E.S. and Hartley, H.O. (1948). *Biometrika Tables for Statisticians*, 1 (2nd ed.). Cambridge University Press, Cambridge.

Sheppard, W.F. (1903). New tables of the probability integral. *Biometrika*, 2, pp. 174-190.

Sheppard, W.F. (1907). Table of deviates of the normal curve. *Biometrika*, 5, pp. 404-406.

TABLE STATISTIQUE

Statistical table

Une table statistique donne les valeurs de la **fonction de répartition** ou de la **probabilité** individuelle d'une **variable aléatoire** suivant une **loi de probabilité** déterminée.

DOMAINES ET LIMITAIONS

La nécessité d'une table statistique découle du fait que certaines fonctions de répartition ne sont pas exprimables sous une forme mathématique simple à utiliser. Dans d'autres cas, il est impossible (ou trop complexe) de trouver une primitive de la **fonction de densité** pour le calcul direct de la fonction de répartition correspondante.
Les valeurs de la fonction de répartition sont donc calculées préalablement et présentées dans une table statistique afin d'en faciliter l'utilisation.

EXEMPLES

La **table normale**, la **table de Student**, celle de Fisher ou encore celle du chi-carré sont des tables statistiques.

MOTS CLÉS

Fonction de répartition *(Distribution function)*
Table binomiale *(Binomial table)*
Table de Fisher *(Fisher table)*
Table de Kruskal-Wallis *(Kruskal-Wallis table)*
Table de Student *(Student table)*
Table de Wilcoxon *(Wilcoxon table)*
Table de Wilcoxon signé *(Wilcoxon signed table)*
Table du chi-carré *(Chi-square table)*
Table normale *(Normal table)*

TABLEAU DE DISTANCES

Distance table

Pour un ensemble de r individus, le tableau de distances est une **matrice** carrée d'ordre r de terme général (i, j) égal à la **distance** entre le i-ème et le j-ème individu. Grâce aux propriétés de la distance, le tableau de distances est symétrique et sa diagonale est nulle.

ASPECTS MATHÉMATIQUES

Soient r individus notés $X_1, X_2, \ldots, X_r$; le tableau de distances construit à partir de ces r individus est la **matrice D** carrée d'ordre r de terme général

$$d_{ij} = d(X_i, X_j)$$

où $d(X_i, X_j)$ est la **distance** entre le i-ème et le j-ème individu.

Les propriétés de la distance nous permettent d'affirmer qu'en particulier :

- les éléments diagonaux d_{ii} de **D** sont nuls car $d(X_i, X_i) = 0$ (pour $i = 1, 2, \ldots, r$) ;
- **D** est symétrique, c'est-à-dire $d_{ij} = d_{ji}$ pour tout i et j allant de 1 à r.

DOMAINES ET LIMITATIONS

Les tableaux de distances sont à la base des méthodes de **classification automatique**, lesquelles ont pour objet d'identifier les individus les plus proches ou les plus semblables afin de les regrouper.

Les tableaux de distances se construisent généralement à l'aide de la **distance** euclidienne si nous ne précisons pas une distance spécifique.

EXEMPLE

Prenons l'exemple des notes obtenues par 5 étudiants à une session d'examens portant sur 4 cours, à savoir l'anglais, le français, les mathématiques et la physique.

Ces notes, de 1 à 6, indiquées au demi-point sont résumées dans le tableau suivant :

	Anglais	Français	Maths	Physique
Alain	5,0	3,5	4,0	4,5
Jean	3,5	4,0	5,0	5,5
Marc	4,5	4,5	4,0	3,5
Paul	6,0	5,5	5,5	5,0
Pierre	4,0	4,5	2,5	3,5

Nous allons chercher à mesurer la **distance** qui sépare les étudiants, pris deux à deux. Pour ce faire, nous comparerons à chaque fois deux lignes du tableau (relatives à deux étudiants) : plus précisément, chaque note de la première ligne considérée avec la note correspondante sur la deuxième ligne.

Par exemple, si nous calculons la distance d'Alain à Jean en utilisant la distance euclidienne, nous

obtenons :

$$\begin{aligned} d(Alain, Jean) &= ((5,0-3,5)^2 + (3,5-4,0)^2 \\ &\quad + (4,0-5,0)^2 + (4,5-5,5)^2)^{\frac{1}{2}} \\ &= \sqrt{2,25+0,25+1+1} \\ &= \sqrt{4,5} = 2,12. \end{aligned}$$

De même, la distance entre Alain et Marc est égale à :

$$\begin{aligned} d(Alain, Marc) &= ((5,0-4,5)^2 + (3,5-4,5)^2 \\ &\quad + (4,0-4,0)^2 + (4,5-3,5)^2)^{\frac{1}{2}} \\ &= \sqrt{0,25+1+0+1} \\ &= \sqrt{2,25} = 1,5. \end{aligned}$$

Et ainsi de suite,

$$\begin{aligned} &d(Alain, Paul) = \sqrt{7,5} = 2,74 \\ &d(Alain, Pierre) = \sqrt{5,25} = 2,29 \\ &d(Jean, Marc) = \sqrt{6,25} = 2,5 \\ &d(Jean, Paul) = \sqrt{9} = 3 \\ &d(Jean, Pierre) = \sqrt{10,75} = 3,28 \\ &d(Marc, Paul) = \sqrt{7,75} = 2,78 \\ &d(Marc, Pierre) = \sqrt{2,5} = 1,58 \\ &d(Paul, Pierre) = \sqrt{16,25} = 4,03. \end{aligned}$$

En rangeant ces résultats dans un tableau, nous obtenons le tableau de distances suivant :

	Alain	Jean	Marc	Paul	Pierre
Alain	0	2,12	1,5	2,74	2,29
Jean	2,12	0	2,5	3	3,28
Marc	1,5	2,5	0	2,78	1,58
Paul	2,74	3	2,78	0	4,03
Pierre	2,29	3,28	1,58	4,03	0

Nous pouvons vérifier que les valeurs des cases diagonales sont égales à zéro et que la **matrice** est symétrique. Cela signifie d'une part que la distance, de chaque étudiant avec lui-même est nulle et, d'autre part, que la distance qui sépare un étudiant d'un autre est identique à celle qui sépare le second du premier. Ainsi, $d(X_i, X_i) = 0$ et $d(X_i, X_j) = d(X_j, X_i)$.

MOTS CLÉS

Classification automatique *(Cluster analysis)*
Dendrogramme *(Dendrogram)*
Distance *(Distance)*
Matrice *(Matrix)*
Vecteur *(Vector)*

TABLEAU DE FRÉQUENCES

Frequency table

Le tableau de fréquences est un outil de représentation d'une **distribution de fréquences**. Il offre une possibilité de présentation d'observations statistiques.

La **fréquence** d'une **valeur** est le nombre de fois que cette valeur apparaît. Une **distribution de fréquences** peut donc être définie comme une liste des fréquences des différentes valeurs prises par une **variable**.

ASPECTS MATHÉMATIQUES

Considérons un ensemble de n unités décrit par une **variable** X prenant les k valeurs $x_1, x_2, \ldots, x_k$.

Désignons par n_i le nombre d'unités ayant la valeur x_i ; n_i est donc la **fréquence** de la valeur x_i.

La fréquence relative de x_i est égale à $f_i = \frac{n_i}{n}$.

Comme les valeurs sont distinctes et exhaustives, la somme des fréquences n_i est égale à l'effectif total n de l'ensemble, ou encore la somme des fréquences relatives f_i est égale à l'unité, c'est-à-dire :

$$\sum_{i=1}^{k} n_i = n \text{ et } \sum_{i=1}^{k} f_i = 1.$$

Le tableau de fréquence est le tableau suivant :

Valeurs de la variable X	Fréquences	Fréquences relatives
x_1	n_1	f_1
x_2	n_2	f_2
$\ldots$	$\ldots$	$\ldots$
x_k	n_k	f_k
Total	n	1

Parfois la notion de fréquence cumulée est importante à indiquer, il s'agit de la somme de la fréquence f_i et

des fréquences précédentes : $\sum_{j=1}^{i} f_j$.

DOMAINES ET LIMITATIONS

Construire un tableau de fréquences est la méthode la plus simple et la plus utilisée pour présenter des données. Certaines règles doivent cependant être respectées :

- un tableau doit avoir un titre clair et concis incluant l'indication des unités employées ;
- les en-têtes de lignes et de colonnes doivent être précis et brefs ;
- les totaux des colonnes doivent être mentionnés ;
- la source de référence du tableau doit être précisée.

En règle générale, le tableau doit être compréhensible par lui-même, sans avoir recours au texte qui l'accompagne.

Pour représenter des données groupées en classes, le choix du nombre de classes dépend de l'effectif total de l'ensemble étudié. Un nombre élevé de classes fera apparaître des irrégularités provenant du faible nombre d'unités par classe, alors qu'un nombre trop restreint de classes conduit à une perte d'information.

Il n'y a pas de règle générale pour déterminer le nombre de classes dans le but de construire un tableau de fréquences. Pour des ensembles de données relativement restreints ($n \leq 200$), il est recommandé de choisir entre 7 et 15 classes ; toutefois, cette règle n'a rien d'absolu.

Par souci de simplicité, il est fréquent de recourir à des classes d'égale amplitude, les intervalles de classes doivent être mutuellement exclusifs. Dans certains cas, par exemple pour une variable dont l'étendue des valeurs est assez grande, il peut être utile que la dernière classe soit ouverte, c'est-à-dire du type 1 000 et plus.

EXEMPLES

Le tableau suivant présente la situation matrimoniale des personnes résidant en Australie, au 30 juin 1981.

Distribution des résidents en Australie (en milliers) selon leur situation matrimoniale, au 30 juin 1981

Situation matrimoniale	Fréquence (en milliers)	Fréquence relative
Célibataire	6 587	0,452
Marié	6 837	0,469
Divorcé	403	0,028
Veuf	749	0,051
Total	14 576	1,000

Source : Bureau de statistique australien, Australian Pocket Year Book 1984, p.11.

Les données sont présentées sous la forme d'un tableau de fréquences. La variable « situation matrimoniale » peut prendre quatre valeurs différentes : célibataire, marié, divorcé ou veuf. Pour chacune d'entre elles, la **fréquence**, exprimée en milliers, ainsi que la fréquence relative sont données.

Dans l'exemple suivant, nous traitons le cas d'observations groupées en classes. Ce tableau présente la répartition, en 1955, de 3 014 contribuables du point de vue de leur revenu imposable défini par tranches.

Revenu imposable de 3014 *contribuables en* 1955 *(en dollars)*

Revenu imposable	n_i	Fréquence relative	Fréquence relative **cumulée**
moins de 1 000	261	0,0866	0,0866
1 000 – 1 999	331	0,1098	0,1964
2 000 – 2 999	359	0,1191	0,3155
3 000 – 3 999	384	0,1274	0,4429
4 000 – 4 999	407	0,1350	0,5780
5 000 – 7 499	703	0,2332	0,8112
7 500 – 9 999	277	0,0919	0,9031
10 000 et plus	292	0,0969	1,0000
Total	3 014	1,0000	

Dans ce tableau, nous relevons par exemple que 57,8 % des 3 014 contribuables ont un revenu imposable inférieur à 5 000 dollars.

MOTS CLÉS

Distribution de fréquences *(Frequency distribution)*
Fréquence *(Frequency)*

TAILLE D'ÉCHANTILLON

Sample size

La taille d'échantillon est le nombre d'individus ou d'éléments appartenant à un **échantillon**.

HISTORIQUE

J.A. Nordin (1944) traite le problème de l'**estimation** de la taille d'échantillon en théorie des décisions à travers un exemple relatif aux ventes potentielles d'un marché auquel un artisan souhaite prendre part.

J. Cornfield (1951) illustre de façon précise l'estimation de la taille d'échantillon pour former des classes connaissant les proportions.

J. Sittig (1951) discute le choix de la taille d'échantillon pour un problème typiquement économique, tenant compte du coût de l'inspection et du coût contracté par des pièces défectueuses dans un lot accepté et de pièces correctes figurant dans un lot rejeté.

De nombreuses recherches ont également été effectuées par D.R. Cox (1952), qui s'est inspiré des travaux de C. Stein (1945), pour déterminer la taille d'échantillon de façon idéale.

ASPECTS MATHÉMATIQUES

Soit $X_1, \ldots, X_n$, un **échantillon** aléatoire tiré sans remise d'une **population** de taille N, ayant une **moyenne** μ inconnue. Nous supposons que la distribution d'échantillonnage de la statistique suit une loi normale.

L'**inégalité de Tchebychev**, exprimée en fonction du nombre n d'observations, permet d'estimer la moyenne μ d'après la moyenne empirique des valeurs obtenues $\overline{X_n}$.

Pour autant que N soit grand, la taille n de l'échantillon peut être déterminée à l'aide de l'inégalité de Tchebychev, de sorte que la probabilité que $\overline{X_n}$ soit incluse dans l'intervalle $[\mu - \varepsilon, \mu + \varepsilon]$ ($\varepsilon > 0$ fixé est la précision exigée) soit aussi grande que nons le désirons. Si nous appelons α le **seuil de signification**, nous obtenons :

$$P\left(|\bar{x} - \mu| \geq \varepsilon\right) = \alpha \text{ ou } P\left(|\bar{x} - \mu| < \varepsilon\right) = 1 - \alpha,$$

que nous pouvons aussi écrire

$$P\left(-z_{\alpha/2} \leq \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} \leq z_{\alpha/2}\right) = 1 - \alpha$$

où $\sigma_{\bar{x}}$ l'**erreur type** de la moyenne est estimé par $\frac{\sigma}{\sqrt{n}}$ où σ est l'**écart type** de la population et $z_{\alpha/2}$ la valeur critique centrée réduite associée au **niveau de confiance** $1 - \alpha$.

Nous construisons alors l'**intervalle de confiance** à l'aide de la **table normale** pour le niveau de confiance $1 - \alpha$:

$$-z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \leq \bar{x} - \mu \leq z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}.$$

Comme la **loi normale** est symétrique, nous nous intéressons uniquement à l'inégalité de droite. Cette dernière indique que $z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$ est la plus grande valeur que peut prendre $\bar{x} - \mu$. Or cette limite est également donnée par la précision exigée. D'où :

$$\varepsilon = z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

c'est-à-dire :

$$n = \left(z_{\alpha/2} \cdot \frac{\sigma}{\delta}\right)^2.$$

Si la précision exigée peut être exprimée en **pourcentage** par rapport à la moyenne μ, soit $k\,\%$, nous avons donc :

$$\begin{aligned} \delta &= k \cdot \mu \\ \text{et } n &= \left(\frac{z_\alpha}{k} \cdot \frac{\sigma}{\mu}\right)^2 \end{aligned}$$

où $\frac{\sigma}{\mu}$ est appelé **coefficient de variation**. Il mesure la dispersion relative de la **variable** étudiée.

Dans le cas où la **distribution d'échantillonnage**

ne suit pas une loi normale, ces formules sont quand même approximativement valables pour autant que la taille de l'échantillon, n, soit suffisamment grande ($n \geq 30$).

DOMAINES ET LIMITATIONS

La détermination de la taille de l'**échantillon** est une étape importante lors de l'organisation d'une enquête statistique. Celle-ci implique non seulement des calculs complexes mais aussi la prise en compte des différentes exigences propres à l'enquête.

Ainsi, la précison souhaitée des résultats de l'enquête est à considérer. En régle générale, plus la précision est élevée, plus la taille de l'échantillon doit être grande. Le coût de l'enquête constitue aussi une autre contrainte importante. En effet, plus le budget est faible, plus la taille de l'échantillon doit être restreinte. De plus, la disponibilité d'autres ressources telles que l'existence d'une base de **sondage** ou d'un réseau d'enquêteurs ou encore la durée de l'enquête, ainsi que le territoire à couvrir, affectent sensiblement la structure de l'échantillon.

La théorie des sondages accorde une large place à la détermination de la taille « optimale » de l'échantillon. Il s'agit essentiellement d'optimiser (maximiser ou minimiser) une fonction objectif dans laquelle les principales exigences sont formulées d'une manière mathématique.

Par ailleurs, il est important de souligner qu'en général la taille optimale de l'échantillon ne dépend pas (ou très peu) de la taille de la **population** d'origine.

Dans le domaine des plans d'expérience, le problème de la taille optimale d'un échantillon se pose également. Il s'agit dans ce cas de trouver le nombre minimal d'observations à prendre en compte afin de pouvoir détecter une différence significative (s'il y a lieu) entre des traitements différents tels que plusieurs médicaments ou procédés de fabrication.

EXEMPLE

Une usine emploie plusieurs milliers d'ouvriers. D'après une expérience antérieure, un chercheur sait que les salaires hebdomadaires sont distribués selon une **loi normale** avec un **écart type** de 40 euros. Ce dernier veut estimer le salaire hebdomadaire moyen à 10 euros près, avec un **niveau de confiance** de 99 %. Quelle taille minimale doit avoir l'**échantillon** ?

Pour un niveau de confiance de 99 % ($\alpha = 0,01 = 1 - 0,99$), la **valeur** de $z_{0,01}$ que nous trouvons dans la **table normale** est égale à $2,575$.

L'application directe de la formule nous donne :

$$\begin{aligned} n &\geq \left[z_\alpha \cdot \frac{\sigma}{\delta}\right]^2 \\ n &\geq \left[2,575 \cdot \frac{40}{10}\right]^2 \\ n &\geq 106,09. \end{aligned}$$

Nous aurons donc la précision désirée si la taille de l'échantillon est supérieure ou égale à 107 ouvriers.

MOTS CLÉS

Échantillon *(Sample)*
Population *(Population)*

RÉFÉRENCES

Cochran, W.G. (1963). *Sampling Techniques*, 2nd ed. John Wiley & Sons, New York.

Cornfield, J. (1951). Methods in the Sampling of Human Populations. The determination of sample size. *Amer. Jour. Pub. Health*, 41, pp. 654-661.

Cox, D.R. (1952). Estimation by double sampling. *Biometrika*, 39, pp. 217-227.

Harnett, D.L. and Murphy, J.L. (1975). *Introductory Statistical Analysis.* Addison-Wesley, Reading, Mass.

Nordin, J.A. (1944). Determining sample size. *Journal of the American Statistical Association*, 39, pp. 497-506.

Sittig, J. (1952). The economic choice of sampling system in acceptance sampling. *Bull. Int. Stat. Inst.*, 33, V, pp. 51-84.

Stein, C. (1945,). A two-sample test for a linear hypothesis whose power is independent of the variance. *Annals of Mathematical Statistics*, 16, pp. 243-258.

TAUX D'INCIDENCE

Incidence page

L'**incidence** d'une maladie est définie comme étant le nombre de nouveaux cas de maladie survenant au cours d'une période déterminée parmi les individus d'une **population**. Cette notion est donc à rapprocher de la notion de « flux ».

Le taux d'incidence est défini comme étant l'incidence rapportée à la taille de la population et au temps ; cette quantité s'exprime donc par relation à un nombre d'individus et à une durée (par exemple, nous parlons d'un taux d'incidence de 400 cas pour 100 000 femmes et par an).

Étant donné que le **risque** est l'incidence rapportée au temps, le taux d'incidence est la mesure du risque par unité de temps.

HISTORIQUE

Voir **taux de prévalence**.

ASPECTS MATHÉMATIQUES

Nous avons, formellement :

$$\text{taux d'incidence} = \frac{\text{risque}}{\text{durée d'observation}}.$$

DOMAINES ET LIMITATIONS

Un taux d'**incidence** élevé calculé au sein d'une **population** et au cours d'une période donnée pourra ainsi être interprété comme l'indice de l'existence d'une épidemie parmi cette population.

EXEMPLE

Prenons le cas du calcul du taux d'**incidence** du cancer du sein dans une population de 150 000 femmes observées sur deux ans, conformément au tableau suivant :

Groupe	Nombre de cas incidents $[A]$	Nombre de sujets $[B]$	Taux d'incidence /100 000/an $[\frac{A}{2B}]$
Non-exposées	40	35 100	57, 0
Fumeuses passives	140	55 500	126, 1
Fumeuses actives	164	59 400	138, 0
Total	344	150 000	114, 7

Le numérateur du taux d'incidence est le même que celui du **risque** : le nombre de nouveaux cas (ou cas incidents) de cancer du sein diagnostiqués au cours des deux ans d'observation parmi le groupe de référence. Le dénominateur du taux d'incidence est obtenu en multipliant l'effectif du groupe de référence par la durée d'observation (ici, deux ans). Il est convenu d'exprimer le taux par année.

MOTS CLÉS

Cause et effet en épidémiologie *(Cause and effect in epidemiology)*
Odds et odds ratio *(Odds and odds ratio)*
Risque *(Risk)*
Risque attribuable *(Attribuable risk)*
Risque évitable *(Avoidable risk)*
Risque relatif *(Relative risk)*
Taux de prévalence *(Prevalence rate)*

RÉFÉRENCES

Lilienfeld A.M. and D.E. Lilienfeld. (1980). *Foundations of epidemiology.* 2nd ed New York: Clarendon Press, Oxford.

MacMahon, B. and Pugh, T.F. (1970). *Epidemiology: Principles and Methods.* Boston: Little Brown.

Morabia A. (1996). *Épidémiologie causale.* Éditions Médecine et Hygiène, Genève.

Morabia A. (1996). *L'Épidémiologie clinique.* Presses universitaires de France, Que sais-je ? Paris.

TAUX DE PRÉVALENCE

Prevalence rate

La **prévalence** d'une maladie est le nombre d'individus atteints par cette affection dans la population globale à un moment donné. Cette notion est donc à rapprocher de la notion de « stock ».

Le taux de prévalence de la maladie est la proportion d'individus malades dans la population à un moment donné (par exemple, nous parlons d'un taux de prévalence d'une maladie donnée dans la population totale de 8 pour 10 000 habitants en l'année 2001).

HISTORIQUE

William Farr, pionnier dans l'utilisation de la statistique en épidémiologie et introducteur des concepts de taux de mortalité et de dose-réponse, montra que la **prévalence** d'une maladie est équivalente au produit de l'**incidence** par la durée de cette maladie.

MacMahon et Pugh (1970) ont illustré la relation des taux d'incidence et de prévalence en se servant de données récoltées entre 1948 et 1952 sur la leucémie aiguë au sein de la population blanche de Brooklin. Le taux d'incidence était de $32,5$ cas par million d'habitants et par an, le taux de prévalence de $6,7$ cas par million d'habitants. La durée de la leucémie aiguë devait être donc être de $0,21$ an ou 2 mois et demi :

$$\text{durée de la léucémie aiguë} = \frac{\text{taux de prévalence}}{\text{taux d'incidence}} = \frac{6.7}{32.5} = 0.21 \text{ an}$$

ASPECTS MATHÉMATIQUES

Le taux de prévalence est défini de la manière suivante :

$$\text{taux de prévalence} = \frac{\text{nombre de malades}}{\text{taille de la population}}.$$

Le taux de prévalence et le **taux d'incidence** d'une maladie sont approximativement associés l'un à l'autre par la durée moyenne de cette maladie (c'est-à-dire la durée moyenne de survie à la maladie ou bien le délai moyen de guérison). Cette relation, valable moyennant les hypothèses de stabilité de l'**incidence** et de la durée moyenne de la maladie dans le temps, s'écrit :

$$\text{taux de prévalence} \simeq (\text{taux d'incidence}) \cdot (\text{durée moyenne de la maladie}).$$

DOMAINES ET LIMITATIONS

La **prévalence** est une notion très différente de celle de **risque**. Un risque est une quantité qui a une valeur prédictive, étant donné qu'elle se rapporte à une période d'observation ponctuelle et contient de ce fait une information homogène par rapport au temps. *A contrario*, la prévalence est tributaire de l'histoire de la maladie au sein d'une population : elle ne peut fournir d'indication concernant le risque dans la mesure où elle agrège, à un moment donné, des cas diagnostiqués récemment ou fort loin dans le passé.

Remarquons que la prévalence d'une maladie est d'autant plus grande qu'est longue la survie moyenne à cette maladie si elle est incurable ou qu'est long le délai de guérison si la maladie est curable.

Notons enfin que le taux de prévalence fait souvent référence non pas à une maladie bien déterminée, mais à un intervalle de valeurs d'un paramètre biologique. Par exemple, le taux de prévalence de l'hypercholestérolémie peut être défini comme la proportion d'individus ayant une cholestérolémie supérieure à $6,5$ mmol/l dans la population à un moment donné. Ainsi, les percentiles d'un **paramètre** biologique, considérés comme mesurant la proportion des individus d'une population pour lesquels le paramètre admet une valeur comprise dans un certain intervalle, peuvent être interprétés comme des taux de prévalence. Par exemple, le taux de prévalence relatif aux valeurs inférieures ou égales à la **médiane** (ou percentile 50) d'un paramètre biologique est de 50 %.

MOTS CLÉS

Cause et effet en épidémiologie *(Cause and effect in epidemiology)*

Odds et odds ratio *(Odds and odds ratio)*
Risque *(Risk)*
Risque attribuable *(Attribuable risk)*
Risque évitable *(Avoidable risk)*
Risque relatif *(Relative risk)*
Taux d'incidence *(Incidence rate)*

RÉFÉRENCES

Cornfield, J. (1951). A method of estimating comparative rates from clinical data. Applications to cancer of the lung, breast, and cervix. *J Natl Cancer Inst*, vol. 11, pp. 1269-75.

Lilienfeld A.M. and D.E. Lilienfeld. (1980). *Foundations of epidemiology.* 2nd ed New York: Clarendon Press, Oxford.

MacMahon, B. and Pugh, T.F. (1970). *Epidemiology: Principles and Methods.* Boston: Little Brown.

Morabia A. (1996). *Épidémiologie causale.* Éditions Médecine et Hygiène, Genève.

Morabia A. (1996). *L'Épidémiologie clinique.* Presses universitaires de France, Que sais-je ? Paris.

TCHEBYCHEV PAFNOUTI LVOVITCH

Chebyshev Pafnouty Lvovitch

Pafnouti Lvovitch Tchebychev (1821-1894) entra à l'Université de Moscou en 1837 où il fut sous l'influence de Nikolaï Efimovich Zernov (qui fut le premier Russe à obtenir un doctorat ès sciences mathématiques) et Nikolaï Dmitrievich Brashman. Après son mémoire, ne trouvant pas d'emploi dans l'enseignement à Moscou, il alla à Saint-Pétersbourg où il fit des conférences sur l'algèbre et la théorie des nombres. En 1859, il reprit le cours de théorie des probabilités de Viktor Yakovlevich Buniakovsky à l'Université de Saint-Pétersbourg.

On lui doit l'inégalité connue en général sous le nom d'**inégalité de Tchebychev** (connue aussi sous le nom d'inégalité de Bienaymé-Tchebychev) dont la preuve fut publiée en version française juste après l'article d'**Irénée-Jules Bienaymé**, dans le *Journal de mathématiques pures et appliquées* (également appelé *Journal de Liouville*).



Il fut l'initiateur d'un travail rigoureux sur une version générale du **théorème central limite** et est considéré comme le fondateur de l'école de mathématique de Saint-Pétersbourg.

Quelques ouvrages et articles principaux de Pafnouti Lvovitch Tchebychev :

1845 *An Essay on Elementary Analysis of the Theory of Probabilities* (thèse).

1867 Preuve de l'inégalité de Tchebychev publiée dans le *Journal de mathématiques pures et appliquées,* 12, pp. 177-184.

MOT CLÉ

Théorème central limite *(Central limit theorem)*

RÉFÉRENCE

Heyde, C.E. and Seneta, E. (1977). *I.J. Bienaymé. Statistical theory anticipated.* Springer-Verlag, New York.

TENDANCE SÉCULAIRE

Secular trend

La tendance séculaire forme l'une des quatre composantes de base de la **série chronologique**.

Elle décrit le mouvement à long terme d'une série chronologique qui globalement peut être croissant, décroissant ou stable.

La tendance séculaire peut être linéaire ou non. Dans ce dernier cas, il faut choisir le type de fonction (exponentielle, quadratique, logistique, etc.) qui se prête le mieux à la répartition observée sur la **représentation graphique** de la série chronologique, avant de pouvoir estimer la tendance séculaire.

HISTORIQUE

Voir **série chronologique**.

ASPECTS MATHÉMATIQUES

Soit Y_t une **série chronologique**. Elle peut être décomposée à l'aide de ses quatre composantes, à savoir :

- la tendance séculaire T_t ;
- les **fluctuations cycliques** C_t ;
- les **variations saisonnières** S_t ;
- les **variations irrégulières** I_t.

Lors de l'étude d'une série chronologique, nous commençons par analyser la tendance séculaire en décidant d'abord si cette tendance peut être supposée linéaire ou non.

Si nous pouvons la supposer linéaire, l'évaluation de la tendance séculaire peut s'effectuer de plusieurs façons :

1. *Méthode graphique ou à main levée*

   Après avoir effectué une **représentation graphique** de la série chronologique, elle consiste à ajuster une droite ou une courbe de tendance à main levée sur le graphe. Il s'agit de l'approximation de la tendance séculaire.

2. *Méthode des moindres carrés*

   Elle permet d'identifier mathématiquement la droite qui représente le mieux l'ensemble des données, à savoir :

   $$Y_t = a + b \cdot t$$

   où

Y_t	est la **valeur** de la série chronologique en un temps donné,
a	est la valeur de la droite lorsque t est à l'origine,
b	est la pente de la droite, et
t	correspond à toute période choisie.

   La droite des **moindres carrés** $\hat{Y}_t$ correspond à l'estimation de la tendance séculaire T_t.

   Calcul de la tendance séculaire

   En supposant que l'unité de temps de la série chronologique est l'année, nous partons de l'idée de déplacer l'origine du temps à l'année correspondant au milieu des années sur lesquelles ont été effectuées les mesures.

   (a) *nombre impair d'années* n :

   soit $j = \dfrac{n-1}{2}$, posons $x_i = t_i - t_j$ pour $i = 1, 2, \ldots, n$; le milieu de l'année t_j est la nouvelle origine et l'unité de la nouvelle **variable** x est l'année.

   Alors $\hat{a} = \dfrac{\sum_{i=1}^{n} Y_i}{n}$ et $\hat{b} = \dfrac{\sum_{i=1}^{n} x_i Y_i}{\sum_{i=1}^{n} x_i^2}$

   nous donne $\hat{Y}_t = \hat{a} + \hat{b}x$

   ou bien $\hat{Y}_t = (\hat{a} - \hat{b}t_j) + \hat{b}t$

   où t est l'année sans changement d'origine.

(b) *nombre pair d'années* n :
soit $j = \frac{n}{2} + 1$; t_j est la nouvelle origine, posons $x_i = 2 \cdot (t_i - t_j) + 1$ pour $i = 1, 2, \ldots, n$

$$\text{alors } \hat{a} = \frac{\sum_{i=1}^{n} Y_i}{n} \text{ et } \hat{b} = \frac{\sum_{i=1}^{n} x_i Y_i}{\sum x_i^2}$$

nous donne $\hat{Y}_t = \hat{a} + \hat{b}x$. L'unité de x est la demi-année comptée à partir du 1er janvier de l'année t_j.

3. *Méthode des moyennes mobiles*
Soit $Y_1, Y_2, Y_3, \ldots$ une série chronologique donnée.
Construisons une nouvelle série :

$$\frac{Y_1 + \ldots + Y_n}{n} ; \frac{Y_2 + \ldots + Y_{n+1}}{n} ; \ldots ;$$
$$\frac{Y_j + \ldots + Y_{n+j-1}}{n} ; \ldots$$

dont l'estimation de la tendance séculaire se fait par l'une des deux méthodes précédentes.

4. *Méthode des semi-moyennes*
Soit $Y_1, Y_2, \ldots, Y_n$ une série chronologique donnée. Notons k la partie entière de $\frac{n}{2}$.

Nous calculons :

$$y_1 = \frac{Y_1 + \ldots + Y_k}{k} \text{ et } y_2 = \frac{Y_{k+1} + \ldots + Y_n}{n-k}$$
$$x_1 = \frac{k+1}{2} \text{ et } x_2 = \frac{n+k+1}{2}$$

et déterminons la droite de tendance séculaire passant par les deux points :

$$(x_1 ; y_1) \text{ et } (x_2 ; y_2).$$

Lorsque la tendance séculaire ne peut pas être supposée linéaire, nous raffinons la méthode des moindres carrés. Il s'agit de l'appliquer en remplaçant $Y_t = a + bt$ par une autre courbe de tendance comme par exemple :

- une exponentielle : $Y_t = a \cdot b^t$;
- une exponentielle modifiée : $Y_t = c + a \cdot b^t$;
- une courbe logistique : $Y_t = (c + a \cdot b^t)^{-1}$;
- une courbe de Gompertz : $Y_t = c \cdot a^{b^x}$.

DOMAINES ET LIMITATIONS

L'analyse de la tendance séculaire poursuit les buts suivants :

- élaborer des modèles descriptifs d'une situation passée ;
- faire des projections de structures constantes ;
- éliminer la tendance séculaire pour étudier les autres composantes de la **série chronologique**.

Les quatre méthodes d'**estimation** de la tendance séculaire présentent certains inconvénients, citons les principaux pour chacune d'entre elles :

1. *Méthode graphique*
Elle est trop subjective.

2. *Méthode des moindres carrés*
Elle ne peut être utilisée que pour une période où le mouvement a lieu entièrement dans le même sens. Dans le cas où une première partie des données correspond à un mouvement ascendant, puis une autre partie à un mouvement descendant, deux droites de longue durée devront être ajustées à ces données, chacune d'elles se rapportant à une période à sens unique.
Elle présente toutefois certaines propriétés intéressantes :

   - Nous retrouvons sur le même graphique l'ensemble des valeurs Y_i de la série chronologique et la droite de tendance. La somme des écarts entre les valeurs observées Y_i et les valeurs estimées de la tendance séculaire $\hat{Y}_i$ correspondantes sera toujours égale à zéro :

$$\sum_{i=1}^{n} \left(Y_i - \hat{Y}_i \right) = 0.$$

- La droite de tendance minimise la somme des carrés des écarts :

$$\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2 = \text{valeur minimale.}$$

3. *Méthode des moyennes mobiles*
   Les données du début et de la fin de la série chronologique initiale sont « perdues ». Les moyennes mobiles peuvent engendrer des cycles et d'autres mouvements qui n'étaient pas présents dans les données d'origine ; de plus, elles sont fortement affectées par les observations aberrantes accidentelles.

   La méthode des moyennes mobiles permet de déceler les changements de direction.

4. *Méthode des semi-moyennes*
   Même si cette méthode est simple à appliquer, elle peut conduire à des résultats sans valeur.

   Bien que valable dans le cas où les données peuvent être classées en deux groupes dans lesquels la tendance est linéaire, la méthode n'est applicable que là où la tendance est globalement linéaire ou approximativement linéaire.

   Si le procédé est d'une application facile, il manque de rigueur.

EXEMPLE

Proposons-nous d'estimer la tendance séculaire des ventes annuelles de la biscuiterie Surfin, à l'aide des données suivantes :

Ventes annuelles de la biscuiterie Surfin (en millions de $)

Année	Vente
1975	7,6
1976	6,8
1977	8,4
1978	9,3
1979	12,1
1980	11,9
1981	12,3

Utilisons la méthode des **moindres carrés**.

Déplaçons l'origine au milieu des sept années, elle se situera donc en 1978. Complétons le tableau suivant :

Année	Vente	Code années		
X_t	Y_t	t	tY_t	t^2
1975	7,6	−3	−22,2	9
1976	6,8	−2	−13,6	4
1977	8,4	−1	−8,4	1
1978	9,3	0	0,0	0
1979	12,1	1	12,1	1
1980	11,9	2	23,8	4
1981	12,3	3	36,9	9
Total	68,4	0	28,0	28

Posons $t_i = X_i - 1978$ pour $i = 1, 2, \ldots, 7$ le milieu de l'année 1978 est la nouvelle origine et l'unité de t est l'année.

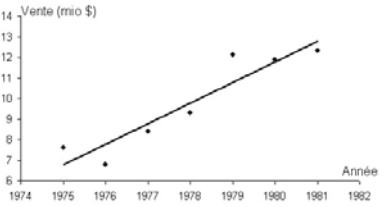
Alors

$$\hat{a} = \frac{\sum_{t=-3}^{3} Y_t}{7} = \frac{68,4}{7} = 9,7714 \text{ (millions de \$)}$$

$$\text{et } \hat{b} = \frac{\sum_{t=-3}^{3} tY_t}{\sum_{t=-3}^{3} t^2} = \frac{28}{28} = 1,0 \text{ (million de \$)}$$

ce qui nous donne $T_t = \hat{Y}_t = 9,7714 + t$ pour l'équation de la droite de tendance séculaire.

Ce qui donne graphiquement :



MOTS CLÉS

Fluctuation cyclique *(Cyclical fluctuation)*
Moindres carrés *(Least squares)*
Moyenne mobile *(Moving average)*
Représentation graphique *(Graphical representation)*
Série chronologique *(Time series)*

TEST BILATÉRAL

Two-tail test

Un test bilatéral concernant une **population** est un **test d'hypothèse** qui s'applique lorsque nous cherchons à comparer une estimation et une valeur donnée ou deux valeurs données.

ASPECTS MATHÉMATIQUES

C'est un **test d'hypothèse** concernant une **population** du type :

hypothèse nulle	H_0	:	$\theta = \theta_0$
hypothèse alternative	H_1	:	$\theta \neq \theta_0$

où θ est un **paramètre** de la population dont la **valeur** est inconnue, et θ_0 est la valeur présumée de ce paramètre.

Pour un test concernant la comparaison de deux populations, les hypothèses sont :

hypothèse nulle	H_0	:	$\theta_1 = \theta_2$
hypothèse alternative	H_1	:	$\theta_1 \neq \theta_2$

où θ_1 et θ_2 sont les paramètres inconnus des deux populations sous-jacentes.

DOMAINES ET LIMITATIONS

Dans un test bilatéral, la **région de rejet** est divisée en deux parties, à droite et à gauche du **paramètre** considéré.

EXEMPLES

- Une entreprise fabrique des câbles d'acier. Elle veut vérifier si le diamètre μ des câbles d'un lot de production est conforme aux normes qui imposent un diamètre de $0,9$ cm.

  Elle prélève un **échantillon** de 100 câbles sur lequel le diamètre moyen mesuré est égal à $0,93$ cm avec un **écart type** $S = 0,12$.

  Les hypothèses sont les suivantes :

hypothèse nulle	H_0	:	$\mu = 0,9$,
hypothèse alternative	H_1	:	$\mu \neq 0,9$.

  Comme la taille de l'échantillon est suffisamment grande, la **distribution d'échantillonnage** de la **moyenne** peut être approximée par une **loi normale**.

  Pour un **seuil de signification** $\alpha = 5\,\%$, nous obtenons la **valeur critique** $z_{\alpha/2}$ dans la **table normale**: $z_{\alpha/2} = 1,96$.

  Nous trouvons la **région d'acceptation** de l'hypothèse nulle (voir **région de rejet**) :

  $$\mu \pm z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

  où σ est l'écart type de la population. L'écart type σ étant inconnu, nous l'estimons par l'écart type S de l'échantillon ($S = 0,12$), ce qui donne :

  $$0,9 \pm 1,96 \cdot \frac{0,12}{\sqrt{100}} = \begin{cases} 0,8765 \\ 0,9235. \end{cases}$$

  La région d'acceptation de l'hypothèse nulle est donc égale à l'**intervalle** $[0,8765\,;\,0,9235]$. La moyenne de l'échantillon ($0,93$ cm) étant située à l'extérieur de cet intervalle, on est conduit à rejeter l'hypothèse nulle au profit de l'hypothèse alternative. Nous concluons donc qu'à un seuil de signification de $5\,\%$ et en fonction de l'échantillon étudié, le diamètre des câbles n'est pas conforme aux normes.

- Une étude est réalisée en vue de comparer la durée de vie moyenne de deux marques de pneus. Un échantillon de 50 pneus de la marque 1 donne une durée de vie moyenne de 72 000 km, alors qu'un échantillon de 40 pneus de la marque 2 donne une durée de vie moyenne de 74 400 km. En supposant que l'écart type pour la marque 1 est $\sigma_1 = 3\,200$ km et celui pour

la marque 2 est $\sigma_2 = 2\,400$ km, nous désirons savoir s'il existe une différence significative de durée de vie entre les deux marques, à un seuil de signification α de 1 %.

Les hypothèses sont les suivantes :

$$\begin{array}{lccl} \text{hypothèse nulle} & H_0 & : & \mu_1 = \mu_2, \\ \text{hypothèse alternative} & H_1 & : & \mu_1 \neq \mu_2. \end{array}$$

La taille des échantillons étant suffisamment grande, et les écarts types des populations étant connus, nos pouvons approcher la distribution d'échantillonnage par une loi normale.

La région d'acceptation de l'hypothèse nulle est donnée par (voir région de rejet) :

$$\mu_1 - \mu_2 \pm z_{\alpha/2} \cdot \sigma_{\bar{x}_1 - \bar{x}_2}$$

où $\sigma_{\bar{x}_1 - \bar{x}_2}$ est l'écart type de la distribution d'échantillonnage de la différence des deux moyennes, ou **erreur type** des deux moyennes.

La valeur $z_{\alpha/2}$ se trouve dans la **table normale**. Pour un test bilatéral avec $\alpha = 1\,\%$, nous obtenons $z_{\alpha/2} = 2.575$, ce qui donne :

$$\begin{aligned} & 0 \pm 2,575 \cdot \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \\ = \;& 0 \pm 2,575 \cdot \sqrt{\frac{3\,200^2}{50} + \frac{2\,400^2}{40}} \\ = \;& 0 \pm 2,575 \cdot 590,59 \\ = \;& \pm 1\,520,77. \end{aligned}$$

La région d'acceptation de l'hypothèse nulle est donc donnée par l'**intervalle** $[-1\,520.77\,;\, 1520,77]$. La différence des moyennes des échantillons ($72\,000 - 74\,400 = -2\,400$) étant extérieure à la région d'acceptation, nous sommes conduits à rejeter l'hypothèse nulle au profit de l'hypothèse alternative. Nous pouvons donc conclure qu'il y a une différence significative de durée de vie entre les pneus des deux marques testées.

MOTS CLÉS

Distribution d'échantillonnage *(Sampling distribution)*
Hypothèse alternative *(Alternative hypothesis)*
Hypothèse nulle *(Null hypothesis)*
Région d'acceptation *(Acceptance region)*
Région de rejet *(Rejection region)*
Seuil de signification *(Significance level)*
Test d'hypothèse *(Hypothesis testing)*
Test unilatéral *(One-tail test)*

TEST BINOMIAL

Binomial test

Le test binomial est un **test d'hypothèse paramétrique** qui s'applique quand la **population** peut être scindée en deux classes, chaque **observation** de cette population appartenant à l'une ou à l'autre de ces deux catégories.

ASPECTS MATHÉMATIQUES

On considère un **échantillon** de n tirages indépendants. Chaque tirage appartient soit à la classe C_1, soit à la classe C_2. On note n_1 le nombre d'observations dans la classe C_1 et n_2 le nombre d'observations dans la classe C_2.

Chaque tirage a la **probabilité** p d'appartenir à la classe C_1, où p est identique pour les n tirages, et la probabilité $q = 1 - p$ d'appartenir à la classe C_2.

Hypothèses

Le test binomial peut être un **test bilatéral** ou un **test unilatéral**. Si p_0 est la valeur présumée de p, $(0 \leq p_0 \leq 1)$, les hypothèses sont exprimées ainsi suivant le cas :

A : *Cas bilatéral*

$$\begin{array}{lcl} H_0 & : & p = p_0 \\ H_1 & : & p \neq p_0 \end{array}$$

B : *Cas unilatéral*

$$\begin{array}{lcl} H_0 & : & p \leq p_0 \\ H_1 & : & p > p_0 \end{array}$$

C : *Cas unilatéral*

$$\begin{array}{lcl} H_0 & : & p \geq p_0 \\ H_1 & : & p < p_0 \end{array}$$

Règles de décision

Cas A

Pour $n < 25$, on utilise la **table binomiale** avec n et p comme paramètres.

Il s'agit tout d'abord de trouver dans cette table la **valeur** la plus proche de $\frac{\alpha}{2}$ (que l'on note α_1) où α est le **seuil de signification**. On note t_1, la valeur correspondant à α_1.

Ensuite, il faut trouver, toujours dans la table, la valeur de $1 - \alpha_1 = \alpha_2$. De même, on note t_2 la valeur correspondant à α_2.

On rejette H_0 au seuil α si

$$n_1 \leq t_1 \text{ ou } n_1 > t_2$$

avec n_1 le nombre d'observations appartenant à la classe C_1.

Quand n est supérieur à 25, on peut utiliser la **loi normale** comme approximation de la **loi binomiale**.

Les paramètres de la loi binomiale étant :

$$\begin{aligned} \mu &= n \cdot p_0 \\ \sigma &= \sqrt{n \cdot p_0 \cdot q_0}, \end{aligned}$$

la **variable aléatoire** Z qui suit une loi normale centrée réduite est donc égale à :

$$Z = \frac{X - \mu}{\sigma} = \frac{X - n \cdot p_0}{\sqrt{n \cdot p_0 \cdot q_0}}$$

où X est une variable aléatoire binomiale de paramètres n et p_0.

L'approximation de la valeur de t_1 et t_2 est alors :

$$\begin{aligned} t_1 &= n \cdot p_0 + z_{\alpha_1} \cdot \sqrt{n \cdot p_0 \cdot q_0} \\ t_2 &= n \cdot p_0 + z_{\alpha_2} \cdot \sqrt{n \cdot p_0 \cdot q_0} \end{aligned}$$

où z_{α_1} et z_{α_2} sont les valeurs à chercher dans la **table normale** correspondant aux seuils α_1 et α_2.

Cas B

Pour $n < 25$, t est la valeur de la table binomiale correspondant à $1 - \alpha$ où α est le **seuil de signification** (ou la valeur la plus proche) avec n et p_0 comme paramètres. On rejette H_0 au seuil α si

$$n_1 > t$$

où n_1 est le nombre d'observations de la classe C_1.

Pour $n \geq 25$, on fait l'approximation :

$$t = n \cdot p_0 + z_\alpha \sqrt{n \cdot p_0 \cdot q_0}$$

où z_α est à chercher dans la table normale au seuil $1 - \alpha$.

Ensuite la règle de décision est la même que pour $n < 25$.

Cas C

Pour $n < 25$, t est la valeur de la table binomiale correspondant à α où α est le seuil de signification (ou la valeur la plus proche) avec n et p_0 comme paramètres.

On rejette H_0 au seuil α si

$$n_1 \leq t$$

où n_1 est le nombre d'observations de la classe C_1.

Pour $n \geq 25$, on fait l'approximation :

$$t = n \cdot p_0 + z_\alpha \sqrt{n \cdot p_0 \cdot q_0}$$

où z_α est à chercher dans la table normale au seuil α. Ensuite la règle de décision est la même.

DOMAINES ET LIMITATIONS

Les deux conditions de base qui doivent être respectées avant d'effectuer un test binomial sont :

1. les n observations sont mutuellement indépendantes,

2. chaque observation a la **probabilité** p d'être dans la première classe. Cette probabilité est la même pour toutes les observations.

EXEMPLES

Exemple 1 : une machine est considérée comme opérationnelle si au plus 5 % des pièces qu'elle produit sont défectueuses. L'**hypothèse nulle** notée H_0 exprime cette situation tandis que l'**hypothèse alternative** notée H_1 signifie que la machine est défaillante :

$$\begin{array}{rcl} H_0 & : & p \leq 0,05, \\ H_1 & : & p > 0,05. \end{array}$$

Pour effectuer le test, on tire un **échantillon** de 10 pièces et on constate qu'il y a trois pièces défectueuses ($n_1 = 3$). Comme les hypothèses correspondent au **test unilatéral** (*cas B*), la règle de décision B va être utilisée. Si l'on choisit un **seuil de signification** $\alpha = 0,05$, la **valeur** de t dans la **table binomiale** est égale à 1 (pour $n = 10$ et $p_0 = 0,05$).

On rejette H_0 car $n_1 = 3 > t = 1$ et l'on conclut que la machine n'est pas opérationnelle.

Exemple 2 : on reprend le même exemple mais cette fois en testant 100 pièces. On constate qu'il y a 12 pièces défectueuses. Dans ce cas, la valeur de t peut être approchée par :

$$t = n \cdot p_0 + z_\alpha \sqrt{n \cdot p_0 \cdot q_0}$$

où z_α est à chercher dans la **table normale** au seuil $1 - \alpha = 0,95$. On a alors :

$$\begin{array}{rcl} t & = & 100 \cdot 0,05 + 1,64 \sqrt{100 \cdot 0,05 \cdot 0,95} \\ & = & 8,57. \end{array}$$

On rejette encore H_0 car $n_1 = 12 > t = 8,57$.

MOTS CLÉS

Table binomiale *(Binomial table)*
Test bilatéral *(Two-tail test)*
Test d'adéquation *(Goodness of fit test)*
Test d'hypothèse *(Hypothesis testing)*
Test paramétrique *(parametric test)*
Test unilatéral *(One-tail test)*

TEST D'ADÉQUATION

Goodness of fit test

Un test d'adéquation effectué sur la base d'un **échantillon** permet de déterminer s'il est correct d'approcher la distribution observée, par une **loi de probabilité** donnée (**loi normale**, **loi de Poisson**, etc.). On veut donc savoir si l'échantillon observé peut être issu d'une **population** qui suit cette loi de probabilité.

HISTORIQUE

Voir **test d'adéquation du chi-carré** et **test de Kolmogorov-Smirnov**.

ASPECTS MATHÉMATIQUES

L'objectif d'un test d'adéquation est de déterminer si un **échantillon** observé peut être considéré comme issu d'une **population** qui suit une **loi de probabilité** particulière.

Le problème peut être posé sous la forme d'un **test d'hypothèse**, de la manière suivante :

$$\begin{array}{lcl} \textbf{hypothèse nulle} & H_0 & : \quad F = F_0, \\ \textbf{hypothèse alternative} & H_1 & : \quad F \neq F_0, \end{array}$$

où F est la **fonction de répartition** inconnue de la population sous-jacente et F_0, la fonction de répartition présumée.

Le test d'adéquation consiste à comparer la distribution empirique avec la distribution présumée. On rejette l'hypothèse nulle, si la distribution empirique n'est pas suffisamment proche de la distribution présumée. Les règles précises de rejet ou d'acceptation de l'hypothèse nulle dépendent du type de test utilisé.

EXEMPLES

Il existe plusieurs tests d'adéquation dont le **test d'adéquation du chi-carré** et le **test de Kolmogorov-Smirnov**.

MOTS CLÉS

Test d'adéquation du chi-carré *(Chi-square goodness of fit test)*

Test d'hypothèse *(Hypothesis testing)*
Test de Kolmogorov-Smirnov *(Kolmogorov-Smirnov test)*

TEST D'ADÉQUATION DU CHI-CARRÉ

Chi-square goodness of fit test

Le test d'adéquation du chi-carré est, avec le **test de Kolmogorov-Smirnov**, l'un des tests d'adéquation les plus utilisés.

Ce test vise à déterminer s'il est correct d'approcher une distribution observée par une **loi de probabilité** particulière (**loi normale**, **loi de Poisson**, etc.).

HISTORIQUE

Le test d'adéquation du chi-carré est le plus ancien et le plus connu des tests d'adéquation, il a été présenté pour la première fois en 1900 par **Karl Pearson**.

ASPECTS MATHÉMATIQUES

Soient $X_1, \ldots, X_n$, un **échantillon** de n observations.

Les étapes d'un test d'adéquation du chi-carré sont les suivantes :

1. Poser les hypothèses. L'**hypothèse nulle** sera de la forme

   $$H_0 \quad : \quad F = F_0$$

   où F_0 est la **fonction de répartition** présumée de la distribution.

2. Répartir les observations en k classes disjointes

   $$[a_{i-1}, a_i].$$

   On note n_i le nombre d'observations contenues dans la i-ème classe, $i = 1, \ldots, k$.

3. Calculer les probabilités théoriques pour chaque classe sur la base de la fonction de répartition présumée F_0 :

   $$p_i = F_0(a_i) - F_0(a_{i-1}), \quad i = 1, \ldots, k.$$

4. En déduire les fréquences estimées pour chaque classe

   $$e_i = n \cdot p_i, \quad i = 1, \ldots, k$$

   où n est la taille de l'**échantillon**.

5. Calculer la **statistique** χ^2 (chi-carré)

   $$\chi^2 = \sum_{i=1}^{k} \frac{(n_i - e_i)^2}{e_i}.$$

   Si H_0 est vraie, la statistique χ^2 suit une **loi du chi-carré** avec v degrés de liberté où

   $$v = (k - 1 - \textit{ nombre de paramètres estimés}).$$

   Par exemple, dans un test d'adéquation à une **loi normale**, le nombre de degrés de liberté sera égal à :

   - $k - 1$ si la **moyenne** μ et l'**écart type** σ de la **population** sont connus ;
   - $k - 2$ si un des deux paramètres μ ou σ est inconnu et sera estimé en vue d'effectuer le test ;
   - $k - 3$ si les deux paramètres μ et σ sont inconnus et tous deux estimés par les valeurs correspondantes de l'échantillon.

6. Rejeter H_0 si l'écart entre les fréquences observées et les fréquences estimées est grand, c'est-à-dire :

   $$\text{si } \chi^2 > \chi^2_{v,\alpha}.$$

   où $\chi^2_{v,\alpha}$ est la **valeur** donnée dans la **table du chi-carré** pour un **seuil de signification** α particulier.

DOMAINES ET LIMITATIONS

En vue d'appliquer le test d'adéquation du chi-carré, il faut que n soit assez grand et que les fréquences estimées, e_i, ne soient pas trop petites. On admet habituellement que les fréquences estimées doivent être supérieures à 5, sauf dans les classes extrêmes où elles peuvent être inférieures à 5, mais supérieures à 1. Si cette contrainte n'est pas satisfaite, il faut regrouper les classes pour satisfaire cette règle.

EXEMPLES

Adéquation à la loi binomiale

On effectue quatre lancers d'une pièce de monnaie et l'on compte le nombre de côtés « face » apparus.

Cette **expérience** est répétée 160 fois. Les fréquences observées sont les suivantes :

Nombre de « face » x_i	Nombre de lancers observés (n_i)
0	17
1	52
2	54
3	31
4	6
Total	160

1. Si l'**expérience** s'est déroulée correctement et que la pièce de monnaie n'est pas truquée, la distribution du nombre de « face » devrait suivre une **loi binomiale**.

   On pose donc comme **hypothèse nulle** que la distribution observée peut être approchée par la loi binomiale, et on va effectuer un **test d'adéquation** en vue de déterminer si cette **hypothèse** doit être acceptée ou non.

2. Dans cet exemple, les différentes valeurs 0, 1, 2, 3 et 4 sont chacune considérées comme une classe.

3. La **variable aléatoire** X (nombre de « face » obtenu par quatre lancers d'une pièce de monnaie) suit une loi binomiale si

   $$P(X = x) = C_n^x \cdot p^x \cdot q^{n-x}$$

   avec

n	nombre d'épreuves indépendantes $= 4$
p	**probabilité** d'un succès (« face ») $= 0,5$
q	probabilité d'un échec (« pile ») $= 0,5$
C_n^x	nombre de combinaisons de x objets parmi n.

   On a donc pour quatre lancers les probabilités théoriques suivantes :

   $$P(X = 0) = \frac{1}{16}$$
   $$P(X = 1) = \frac{4}{16}$$
   $$P(X = 2) = \frac{6}{16}$$
   $$P(X = 3) = \frac{4}{16}$$
   $$P(X = 4) = \frac{1}{16}$$

4. Comme on répète l'expérience 160 fois, le nombre de lancers estimé pour chaque **valeur** possible de X est égal à :

   $$e_i = 160 \cdot P(X = x_i)$$

   On obtient donc le tableau suivant :

Nombre de « face » x_i	Nombre de lancers observé (n_i)	Nombre de lancers estimé (e_i)
0	17	10
1	52	40
2	54	60
3	31	40
4	6	10
Total	160	160

5. La **statistique** χ^2 (chi-carré) est égale à :

   $$\chi^2 = \sum_{i=1}^{k} \frac{(n_i - e_i)^2}{e_i}$$

   où k est le nombre de valeurs possibles de X.

   $$\chi^2 = \frac{(17-10)^2}{10} + \ldots + \frac{(6-10)^2}{10} = 12,725.$$

6. En choisissant un **seuil de signification** α de 5 %, on trouve la valeur $\chi^2_{v,\alpha}$ dans la **table du chi-carré** avec $k - 1 = 4$ degrés de liberté :

   $$\chi^2_{4,0,05} = 9,488.$$

   Comme la valeur calculée du χ^2 est supérieure à la valeur de la table, on doit rejeter l'**hypothèse nulle** et conclure que la loi binomiale ne fournit pas une bonne approximation de notre distribution observée. On peut donc conclure que les pièces de monnaie devaient être truquées, ou qu'elles n'étaient pas correctement lancées.

Adéquation à la loi normale

Une étude a été effectuée en vue de déterminer la distribution des diamètres de câbles fabriqués dans une usine.

Le **tableau de fréquences** de la distribution observée est le suivant :

Diamètre des câbles (en mm)	Fréquences observées n_i
$19,70 - 19,80$	5
$19,80 - 19,90$	12
$19,90 - 20,00$	35
$20,00 - 20,10$	42
$20,10 - 20,20$	28
$20,20 - 20,30$	14
$20,30 - 20,40$	4
Total	140

1. On effectue un **test d'adéquation** à la **loi normale**. On pose donc comme hypothèse nulle que la distribution observée peut être approchée par une loi normale.

2. La répartition des observations en classes est déjà donnée dans le tableau précédent.

3. Si la variable aléatoire X (diamètre des câbles) suit une loi normale, la variable aléatoire

$$Z = \frac{X - \mu}{\sigma}$$

suit une loi normale standard. La **moyenne** μ et l'**écart type** σ de la **population** étant inconnus, on les estime par la moyenne, $\bar{x}$ et l'écart type S de l'**échantillon** :

$$\bar{x} = \frac{\sum_{i=1}^{7} \delta_i \cdot n_i}{n} = \frac{2806,40}{140} = 20,05$$

$$S = \sqrt{\frac{\sum_{i=1}^{7} n_i \cdot (\delta_i - \bar{x})^2}{n-1}} = \sqrt{\frac{2,477}{139}} = 0,134$$

où les δ_i sont les centres de classes (dans cet exemple, le diamètre moyen d'une classe, ce qui donne pour $i = 1$: $\delta_1 = 19,75$) et n est le nombre total d'observations.

On peut alors calculer les probabilités théoriques associées à chaque classe. Le calcul détaillé pour les deux premières classes donne :

$$\begin{aligned} p_1 &= P(X \leq 19,8) \\ &= P\left(Z \leq \frac{19,8 - \bar{x}}{S}\right) \\ &= P(Z \leq -1,835) \\ &= 1 - P(Z \leq 1,835) \end{aligned}$$

$$\begin{aligned} p_2 &= P(19,8 \leq X \leq 19,9) \\ &= P\left(\frac{19,8 - \bar{x}}{S} \leq Z \leq \frac{19,9 - \bar{x}}{S}\right) \\ &= P(-1,835 \leq Z \leq -1,089) \\ &= P(Z \leq 1,835) - P(Z \leq 1,089) \end{aligned}$$

Ces probabilités peuvent être trouvées en consultant la **table normale**. On obtient :

$$\begin{aligned} p_1 &= P(X \leq 19,8) = 0,03325 \\ p_2 &= P(19,8 \leq X \leq 19,9) = 0,10476 \\ p_3 &= P(19,9 \leq X \leq 20,0) = 0,22767 \\ p_4 &= P(20,0 \leq X \leq 20,1) = 0,29083 \\ p_5 &= P(20,1 \leq X \leq 20,2) = 0,21825 \\ p_6 &= P(20,2 \leq X \leq 20,3) = 0,09622 \\ p_7 &= P(X > 20,3) = 0,02902 \end{aligned}$$

4. Les fréquences estimées pour chaque classe sont donc égales à :

$$e_i = n \cdot p_i,$$

ce qui donne le tableau suivant :

Diamètre des câbles (en mm)	Fréquences observées n_i	Fréquences estimées e_i
$19,70 - 19,80$	5	$4,655$
$19,80 - 19,90$	12	$14,666$
$19,90 - 20,00$	35	$31,874$
$20,00 - 20,10$	42	$40,716$
$20,10 - 20,20$	28	$30,555$
$20,20 - 20,30$	14	$13,471$
$20,30 - 20,40$	4	$4,063$
Total	140	140

5. La statistique χ^2 (chi-carré) est égale à :

$$\chi^2 = \sum_{i=1}^{k} \frac{(n_i - e_i)^2}{e_i}$$

où $k = 7$ est le nombre de classes.

$$\begin{aligned} \chi^2 &= \frac{(5-4,655)^2}{4,655} + \frac{(12-14,666)^2}{14,666} \\ &\quad + \ldots + \frac{(4-4,063)^2}{4,063} \\ &= 1,0927. \end{aligned}$$

6. En choisissant un seuil de signification α de 5 %, on trouve la valeur de $\chi^2_{v,\alpha}$ dans la table du chi-carré avec $k - 3 = 7 - 3 = 4$ degrés de liberté:

$$\chi^2_{4,0,05} = 9,49.$$

Comme la valeur calculée de χ^2 est inférieure à la valeur de la table, on ne rejette pas l'hypothèse nulle et on conclut qu'au **seuil de signification** de 5 %, la différence entre la distribution observée et la loi normale n'est pas significative.

MOTS CLÉS

Loi du chi-carré *(Chi-square distribution)*
Table du chi-carré *(Chi-square table)*
Test d'adéquation *(Goodness of fit test)*
Test d'hypothèse *(Hypothesis testing)*
Test de Kolmogorov-Smirnov *(Kolmogorov-Smirnov test)*

RÉFÉRENCE

Pearson, K. (1900). 1948. On the Criterion, That a Given System of Deviations from the Probable in the Case of a Correlated System of Variables is such that it can be reasonably supposed to have arisen from random Sampling. pp. 339-357 in Karl Pearson, *Karl Pearson's Early Statistical Papers.* Cambridge Univ. Press. First published in the *Philosophical Magazine*, 5th Series, 50, pp. 157-175.

TEST D'ANDERSON-DARLING

Anderson-Darling test

Le test d'Anderson-Darling est un test d'adéquation entre la fonction de distribution théorique d'une **variable aléatoire** (absolument) continue et la fonction de distribution empirique observée sur un échantillon issu de cette variable aléatoire. En particulier, il permet de tester si la distribution empirique obtenue est conforme ou non à une **loi normale**.

HISTORIQUE

Theodore W. Anderson et Darling ont utilisé la statistique d'Anderson-Darling, notée A^2, tout d'abord pour le test de conformité à une loi de distribution dont les paramètres sont parfaitement spécifiés (1952 et 1954). Ultérieurement, au cours des années 1960 et surtout 1970, Stephens (1974) essentiellement et quelques autres auteurs ont adapté le test à une gamme plus large de lois de distributions dont les paramètres sont ou ne sont pas connus.

ASPECTS MATHÉMATIQUES

Considérons la **variable aléatoire** X suivant une loi normale d'espérance μ et de variance σ^2, de fonction de répartition $F_X(x\,;\theta)$ où θ est un paramètre (ou un ensemble de paramètres) déterminant F_X et supposé connu.

Une observation d'un échantillon de taille n issu de la variable X fournit une fonction de répartition $F_n(x)$. La statistique d'Anderson-Darling, notée A^2 est alors construite comme la somme pondérée des carrés des écarts $F_X(x\,;\theta) - F_n(x)$:

$$A^2 = \frac{1}{n}\left(\sum_{i=1}^{n} \left(F_X(x\,;\theta) - F_n(x)\right)\right).$$

Partant du fait que A^2 est une variable aléatoire positive qui suit une certaine loi de distribution sur l'intervalle $[0\,;+\infty[$, il est possible de tester, pour un seuil de signification fixé *a priori*, si $F_n(x)$ est ou non la réalisation de la variable aléatoire $F_X(X\,;\theta)$,

c'est-à-dire si X est bien distribuée selon une loi de probabilité de fonction de répartition $F_X(x\,;\theta)$.

Procédure de calcul de la statistique A^2

Nous organisons les observations $x_1,\ x_2,\ \ldots,\ x_n$ de l'échantillon issu de X de manière croissante, de sorte que $x_1 < x_2 < \ldots < x_n$. Notons alors $z_i = F_X(x_i\,;\theta)$, $(i = 1, 2, \ldots, n)$. Compte tenu des définitions précédentes, A^2 est définie formellement par :

$$A^2 = -\frac{1}{n}\left(\sum_{i=1}^{n}(2i-1)\left(\ln(z_i)+\ln\left(1-z_{n+1-i}\right)\right)\right) - n.$$

Dans la situation que nous privilégeons ici (X suit une loi normale d'espérance μ et de variance σ^2), nous pouvons dénombrer quatre cas de figure, relatifs à la connaissance des paramètres μ et σ^2 (F est la fonction de répartition de la loi normale centrée-réduite) :

1. μ et σ^2 sont connues et $F_X\left(x\,;\left(\mu,\sigma^2\right)\right)$ est alors parfaitement spécifiée. Nous avons alors naturellement $z_i = F(w_i)$ où $w_i = \frac{x_i-\mu}{\sigma}$;
2. σ^2 est connue et μ est inconnue et estimée par $\overline{x} = \frac{1}{n}\left(\sum_i x_i\right)$, moyenne de l'échantillon. Posons alors $z_i = F(w_i)$, où $w_i = \frac{x_i-\overline{x}}{\sigma}$;
3. μ est connue mais σ^2 est inconnue et estimée par $s'^2 = \frac{1}{n}\left(\sum_i (x_i-\overline{x})^2\right)$. Dans ce cas, posons $z_i = F(w_i)$, où $w_i = \frac{x_{(i)}-\mu}{s'}$;
4. μ et σ^2 sont inconnues et estimées respectivement par $\overline{x}$ et $s^2 = \frac{1}{n-1}\left(\sum_i (x_i-\overline{x})^2\right)$. Posons alors $z_i = F(w_i)$, où $w_i = \frac{x_{(i)}-\overline{x}}{s}$.

Les lois des distributions asymptotiques de la statistique A^2 ont été explicitées par Anderson et Darling pour le cas 1 et Stephens pour les cas 2 et 3. Pour le cas 4, Stephens a déterminé la distribution asymptotique de la transformée $A^* = A^2(1,0 + \frac{0{,}75}{n} + \frac{2{,}25}{n^2})$ de A^2.

Ainsi, nous disposons d'une table donnant, selon les cas 1 à 4, pour des seuils de signification de 10 %, 5 %, 2,5 %, et 1 %, les valeurs-limites de A^2 (et A^* pour le cas 4) au-delà desquelles l'hypothèse de normalité est rejetée :

	Seuil de signification			
Cas	0,1	0,050	0,025	0,01
1 : $A^2 =$	1,933	2,492	3,070	3,857
2 : $A^2 =$	0,894	1,087	1,285	1,551
3 : $A^2 =$	1,743	2,308	2,898	3,702
4 : $A^* =$	0,631	0,752	0,873	1,035

DOMAINES ET LIMITATIONS

La loi de distribution de A^2 n'étant explicitée qu'asymptotiquement, le test requiert que la taille n de l'échantillon soit relativement importante. En toute rigueur, lorsque nous sommes confrontés aux cas 1 et 2, la distribution de A^2 n'est pas connue et il faudrait procéder à une transformation du type de celle de $A^2 \longmapsto A^*$, où A^* serait susceptible d'être déterminée. Toutefois, lorsque $n > 20$, nous pouvons éviter une telle transformation et alors les données du tableau ci-dessus sont valides.

Nous avons insisté sur l'application de la statistique d'Anderson-Darling à un test de conformité à la loi normale. Cependant, le test d'Anderson-Darling présente l'avantage de s'appliquer avec une notable puissance sur une grande variété de lois de probabilité (non seulement **loi normale** mais aussi lois exponentielle, logistique, gamma, etc.). Ceci permet, lorsque, dans un premier temps, le test a rejeté l'hypothèse selon laquelle la variable aléatoire étudiée suit une loi de distribution particulière, de le réitérer par rapport à une large gamme de distributions alternatives.

EXEMPLE

Les données suivantes permettent d'illustrer le test d'Anderson-Darling pour les hypothèses de normalité.

Soit un échantillon relatif à la taille (en cm) de 25 étudiants de sexe masculin : le tableau suivant récapitule les observations ordonnées ainsi que les w_i et z_i.

Par ailleurs, calculons $\overline{x}$ et à s partir de ces données : $\overline{x} = 177,36$ et $s = 4,98$.

Notons F la fonction de répartition normale

centrée réduite, nous avons :

Obs.	x_i	$w_i = \frac{x_i - \overline{x}}{s}$	$z_i = F(w_i)$
1	169	$-1,678$	$0,047$
2	169	$-1,678$	$0,047$
3	170	$-1,477$	$0,070$
4	171	$-1,277$	$0,100$
5	173	$-0,875$	$0,191$
6	173	$-0,875$	$0,191$
7	174	$-0,674$	$0,250$
8	175	$-0,474$	$0,318$
9	175	$-0,474$	$0,318$
10	175	$-0,474$	$0,318$
11	176	$-0,273$	$0,392$
12	176	$-0,273$	$0,392$
13	176	$-0,273$	$0,392$
14	179	$0,329$	$0,629$
15	180	$0,530$	$0,702$
16	180	$0,530$	$0,702$
17	180	$0,530$	$0,702$
18	181	$0,731$	$0,767$
19	181	$0,731$	$0,767$
20	182	$0,931$	$0,824$
21	182	$0,931$	$0,824$
22	182	$0,931$	$0,824$
23	185	$1,533$	$0,937$
24	185	$1,533$	$0,937$
25	185	$1,533$	$0,937$

Nous obtenons alors $A^2 \cong 0,436$, d'où $A^* = A^2 \cdot (1,0 + \frac{0,75}{25} + \frac{0,25}{625}) = A^2 \cdot (1,0336) \cong 0,451$.

Comme nous nous trouvons dans le cas 4, pour un seuil de signification fixé *a priori* à 1 %, la valeur calculée de A^* est très inférieure à la valeur tabulée $(1,035)$. Au seuil de 1 %, l'hypothèse de normalité ne peut donc pas être rejetée.

MOTS CLÉS

Histogramme *(Histogram)*
Loi normale *(Normal distribution)*
Statistique *(Statistics)*
Test d'adéquation *(Goodness of fit test)*

RÉFÉRENCES

Anderson, T.W. and Darling, D.A. (1952). Asymptotic Theory of Certain Goodness of Fit Criteria Based on Stochastic Processes. *Annals of Mathematical Statistics*, 23, pp. 193-212.

Anderson, T.W. and Darling, D.A. (1954). A test of goodness of fit. *Journal of the American Statistical Association*, 49, pp. 765-769.

Durbin, J., Knott, M. and Taylor, C.C. (1975). Components of Cramer-Von Mises statistics,II. *Journal of the Royal Statistical Society.* Series B, 37, pp. 216-237.

Stephens, M.A. (1974). EDF Statistics for Goodness of Fit and Some Comparisons. *Journal of the American Statistical Association*, 69, pp. 730-737.

TEST D'HOMOGÉNÉITÉ

Homogeneity test

Dans le cadre d'un ensemble de **données catégoriques** pour plusieurs groupes, un problème qui se pose fréquemment est de vouloir tester l'homogénéité des groupes, c'est-à-dire d'identifier s'il y a des différences significatives entre les groupes considérés par rapport à une ou plusieurs variables qualitatives catégorielles. Le test d'homogénéité indique si les différences sont significatives ou pas.

ASPECTS MATHÉMATIQUES

On considère ici le test d'homogénéité du chi-carré qui est un cas particulier du **test du chi-carré**.

Soient I le nombre de groupes consid'érés et J le nombre de catégories considérées, ainsi que :

- $n_{i.} = \sum_{j=1}^{J} n_{ij}$ correspond à la taille du groupe i ;
- $n_{.j} = \sum_{i=1}^{I} n_{ij}$;
- $n = \sum_{i=1}^{I} \sum_{j=1}^{J} n_{ij}$ est le nombre total d'observations ;
- n_{ij} la **fréquence** empirique (c'est-à-dire le nombre d'occurences observées) correspondant au groupe i et à la catégorie j ;
- m_{ij} la fréquence théorique correspondant au

groupe i et à la catégorie j, qui, sous l'hypothèse d'homogénéité parmi les groupes, est égale à :

$$m_{ij} = \frac{n_{i.} \cdot n_{.j}}{n}.$$

Si on représente les données sous forme d'une **table de contingence** avec I lignes et J colonnes, on peut calculer les $n_{i.}$ en sommant tous les éléments de la ligne i et les $n_{.j}$ en sommant tous les éléments de la colonne j.

On calcule

$$\chi_c^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(n_{ij} - m_{ij})^2}{m_{ij}}$$

et le test d'homogénéité du chi-carré s'exprime alors par : on rejette l'hypothèse d'homogénéité (au seuil de signification 5 %) si la valeur χ_c^2 est supérieure à la valeur de la distribution du χ^2 (chi-carré) avec $(J-1) \cdot (I-1)$ degrés de liberté.

Remarque : si on veut tester les proportions p_{ij} (pour le groupe i et la catégorie j) au lieu des proportions égales il suffit de remplacer m_{ij} par $n_{i.} \cdot p_{ij}$.

EXEMPLES

On considère une étude dans le domaine pharmaceutique qui porte sur 100 personnes souffrant d'une maladie particulière. Afin d'examiner l'effet d'un traitement médical, 100 personnes ont été choisies aléatoirement. La moitié d'entre elles constitue le groupe contrôle et un placebo leur a été administré. Les autres patients ont reçu le traitement médical. On a compté ensuite, pour chaque groupe, le nombre de personnes rétablies dans les 24 heures qui ont suivi l'administration de ce traitement. Les résultats sont présentés dans le tableau suivant.

Tableau 2×2

Fréquences observées	Rétabli dans les 24 heures	Non rétabli	Total
Placebo	2	48	50
Traitement	9	41	50

Les fréquences théoriques sont obtenues en supposant que le taux de rétablissement est le même dans le groupe « placebo » et le groupe « traitement ». On obtient $m_{11} = m_{21} = (50 \cdot 11)/100 = 5,5$ et $m_{12} = m_{22} = (50 \cdot 89)/100 = 44,5$.

En comparant les fréquences théoriques avec les fréquences observées, on obtient comme valeur calculée du χ^2 :

$$\begin{aligned} \chi_c^2 &= \frac{(2-5,5)^2}{5,5} + \frac{(48-44,5)^2}{44,5} \\ &\quad + \frac{(9-5,5)^2}{5,5} + \frac{(41-44,5)^2}{44,5} \\ &= 5,005. \end{aligned}$$

Si l'on se réfère à la table de la distribution du χ^2 avec 1 degré de liberté, on obtient pour un seuil de signification de 5 % la valeur $\chi^2_{0,05} = 3,84$, qui est inférieure à la valeur obtenue $\chi_c^2 = 5,005$. On en déduit que le traitement a été efficace.

MOTS CLÉS

Analyse de données catégoriques *(Analysis of categorical data)*
Analyse de variance *(Analysis of variance)*
Données catégoriques *(Categorical data)*
Fréquence *(Frequency)*
Loi du chi-carré *(Chi-square distribution)*
Test du chi-carré *(Chi-square test)*

RÉFÉRENCES

Voir **analyse de données catégoriques**.

TEST D'HYPOTHÈSE

Hypothesis testing

Un test d'hypothèse est une procédure permettant d'aboutir, en fonction de certaines règles de décision, au non-rejet d'une **hypothèse** de départ, appelée **hypothèse nulle** ou au rejet de l'hypothèse nulle en faveur de l'**hypothèse alternative**.

HISTORIQUE

Le développement des tests d'hypothèse s'est ef-

fectué graduellement. Ils consistaient initialement en de vagues exposés concernant la significativité ou non d'un ensemble d'observations. Des exemples isolés datent du XVIII^{e} siècle avec J. Arbuthnott (1710), Daniel Bernoulli (1734) et **Pierre Simon de Laplace** (1773). Ils devinrent plus fréquents au XIX^{e} siècle avec, par exemple, Gavarett (1840) et **Francis Y. Edgeworth** (1885).

Il faut mettre le développement des tests d'hypothèse en parallèle avec celui de la théorie de l'**estimation**. L'élaboration des tests d'hypothèse semble avoir été réalisée en premier lieu au niveau des sciences expérimentales et dans le domaine de la gestion. C'est ainsi que, par exemple, le **test de Student** a été développé par **William Sealy Gosset** dans le cadre de son activité professionnelle aux brasseries Guinness.

C'est à **Jerzy Neyman** et à **Egon Sharpe Pearson** que l'on doit le développement de la théorie mathématique des tests d'hypothèse, dont la présentation se trouve dans leur article publié en 1928 dans la revue *Biometrika*. Ils furent les premiers à reconnaître que le choix rationnel d'un test d'hypothèse devait prendre en considération non seulement l'**hypothèse nulle** que l'on souhaite tester, mais aussi l'**hypothèse alternative**. Un second article fondamental sur la théorie des tests d'hypothèse a été publié en 1933 par les deux mêmes mathématiciens ; ils y introduisent également la distinction entre l'**erreur de première espèce** et l'**erreur de seconde espèce**.

Les travaux résultant de la collaboration de J. Neyman et E.S. Pearson sont décrits dans le recueil de l'œuvre de Pearson (1966) et dans la biographie de Neyman publiée par Reid (1982).

ASPECTS MATHÉMATIQUES

La démarche à suivre pour effectuer un test d'hypothèse, à partir d'un **échantillon**, comprend généralement les étapes suivantes :

1. Formuler les hypothèses :
   - l'**hypothèse nulle** H_0 ;
   - l'**hypothèse alternative** H_1.
2. Déterminer le **seuil de signification** α du test.
3. Déterminer la **loi de probabilité** correspondant à la **distribution d'échantillonnage**.
4. Calculer la **valeur critique** de l'hypothèse nulle et en déduire la **région de rejet** ou la **région d'acceptation**.
5. Établir les règles de décision :
   - si la **statistique** observée sur l'échantillon appartient à la région d'acceptation, on ne rejette pas l'hypothèse nulle H_0 ;
   - si la statistique observée sur l'échantillon appartient à la région de rejet, on rejette l'hypothèse nulle H_0 au profit de l'hypothèse alternative H_1.
6. Prendre la décision d'accepter ou de rejeter l'hypothèse nulle sur la base de l'échantillon observé.

DOMAINES ET LIMITATIONS

Les tests d'hypothèse les plus fréquents sont les suivants :

1. *Test d'hypothèse sur un échantillon*

   On veut tester si la **valeur** d'un **paramètre** θ de la **population** est identique à une valeur présumée.

   Les hypothèses seront du type :

   $$\begin{array}{lcl} H_0 & : & \theta = \theta_0 \\ H_1 & : & \theta \neq \theta_0, \end{array}$$

   θ_0 étant la valeur présumée du paramètre inconnu θ.

2. *Test d'hypothèse sur deux échantillons*

   Le but est de savoir si deux populations décrites par un paramètre particulier sont différentes.

   Soient θ_1 et θ_2, les paramètres décrivant les populations 1 et 2 respectivement, on formule les hypothèses suivantes :

   $$\begin{array}{lcl} H_0 & : & \theta_1 = \theta_2 \\ H_1 & : & \theta_1 \neq \theta_2, \end{array} \quad \text{ou} \quad \begin{array}{lcl} H_0 & : & \theta_1 - \theta_2 = 0 \\ H_1 & : & \theta_1 - \theta_2 \neq 0. \end{array}$$

3. *Test d'hypothèse sur plus de deux échantillons*

   À l'instar d'un test sur deux échantillons, un test d'hypothèse sur plus de deux échantillons a pour but de déterminer si plusieurs populations sont différentes sur la base de la comparaison d'un paramètre des populations testées.

   Dans ce cas, on teste les hypothèses suivantes :

   $$\begin{array}{lcl} H_0 & : & \theta_1 = \theta_2 = \ldots = \theta_k \\ H_1 & : & \text{les valeurs de } \theta_i \quad (i = 1, 2, \ldots, k) \\ & & \text{ne sont pas toutes identiques,} \end{array}$$

   $\theta_1, \ldots, \theta_k$ étant les paramètres inconnus des populations et k le nombre de populations à comparer.

Dans la théorie des tests d'hypothèse et en pratique, on distingue deux types de tests : les tests paramétriques et les tests non paramétriques.

Tests paramétriques

Un test paramétrique est un test d'hypothèse qui suppose une forme paramétrique des distributions relatives aux populations sous-jacentes. Tel est le cas lorsque les populations étudiées suivent une **loi normale**.

Le **test de Student** est un exemple de test paramétrique. Celui-ci vise en effet à comparer les moyennes de deux populations qui suivent une distribution normale.

Tests non paramétriques

Un test non paramétrique est un test d'hypothèse pour lequel il n'est pas nécessaire de spécifier la forme paramétrique de la distribution des populations sous-jacentes.

Quelques exemples de ce type de test : le **test du signe**, le **test de Wilcoxon**, le **test de Wilcoxon signé**, le **test de Mann-Whitney**, le **test de Kruskal-Wallis**, le **test de Kolmogorov-Smirnov**.

EXEMPLES

Pour des exemples de test d'hypothèse paramétrique, voir **test binomial**, **test de Fisher** ou **test de Student**. Pour des exemples de test d'hypothèse non paramétrique, voir **test de Kolmogorov-Smirnov**, **test de Kruskal-Wallis**, **test de Mann-Whitney**, **test de Wilcoxon**, **test de Wilcoxon signé** ainsi que **test du signe**.

MOTS CLÉS

Distribution d'échantillonnage *(Sampling distribution)*
Hypothèse alternative *(Alternative hypothesis)*
Hypothèse nulle *(Null hypothesis)*
Région d'acceptation *(Acceptance region)*
Région de rejet *(Rejection region)*
Seuil de signification *(Significance level)*
Test bilatéral *(Two-tail test)*
Test non paramétrique *(Nonparametric test)*
Test paramétrique *(Parametric test)*
Test unilatéral *(One-tail test)*

RÉFÉRENCES

Arbuthnott, J. (1710). An argument for Divine Providence, taken from the constant regularity observed in the births of both sexes. *Philosophical Transactions* 27, pp. 186-190.

Bernoulli, D. (1734). Quelle est la cause physique de l'inclinaison des planètes (...). *Recueil des pièces qui ont remporté le prix de l'Académie royale des sciences* 3, pp. 95-122.

Edgeworth F.Y. (1885). *Methods of Statistics.* Jubilee volume of Jubelee Volume of the Royal Statistical Society, London.

Gavarret, J. (1840). *Principes généraux de statistique médicale.* Beché Jeune & Labé, Paris.

Laplace, P.S. de (1773). Mémoire sur l'inclinaison moyenne des orbites des comètes. *Mémoire de l'Académie royale des sciences*, Paris, 7, pp. 503-524.

Lehmann, E.L. (1986). *Testing statistical hypotheses.* 2nd ed. Wiley, Series in Probability and Mathematical Statistics.

Neyman, J. and Pearson, E.S. (1928). On the use

and interpretation of certain test criteria for purposes of statistical inference. I, II. *Biometrika*, 20A, pp. 175-240 and 263-294.

Neyman, J. and Pearson, E.S. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London.* Series A, Vol. 231, pp. 289-337.

Pearson, E.S. (1966). "The Neyman-Pearson story : 1926-34" in *Research Papers in Statistics : Festschrift for J. Neyman.* Ed. F.N. David. John Wiley & Sons, New York.

Reid, C. (1982). *Neyman - from Life.* Springer-Verlag, New York.

TEST D'INDÉPENDANCE

Test of independence

Un test d'indépendance est un **test d'hypothèse** dont l'objet est de déterminer si deux variables aléatoires sont indépendantes ou ne le sont pas.

HISTORIQUE

Voir **coefficient de corrélation des rangs de Kendall**, **coefficient de corrélation des rangs de Spearman** et **test d'indépendance du chi-carré**.

ASPECTS MATHÉMATIQUES

Soient $(X_1, X_2, \ldots, X_n)$ et $(Y_1, Y_2, \ldots, Y_n)$ deux échantillons de taille n. L'objectif est de tester les deux hypothèses suivantes :

Hypothèse nulle
H_0 : les deux **variables** X et Y sont indépendantes.
Hypothèse alternative
H_1 : les deux variables X et Y ne sont pas indépendantes.

Le test d'indépendance consiste à comparer la distribution empirique avec la distribution théorique par l'intermédiaire du calcul d'un indicateur.

Les tests d'indépendance appliqués à deux variables continues sont fondés sur les rangs des observations. Tel est le cas des tests basés sur le **coefficient de corrélation des rangs de Spearman** ou sur le **coefficient de corrélation des rangs de Kendall**.

Dans le cas de deux variables catégoriques, le test le plus utilisé est le **test d'indépendance du chi-carré**.

DOMAINES ET LIMITATIONS

L'exigence préalable pour pouvoir effectuer un test d'indépendance est que chaque couple (X_i, Y_i), $i = 1, 2, \ldots, n$, provienne de la même **population** bivariée.

EXEMPLES

Voir **test d'indépendance du chi-carré**.

MOTS CLÉS

Coefficient de corrélation des rangs de Kendall *(Kendall rank correlation coefficient)*
Coefficient de corrélation des rangs de Spearman *(Spearman rank correlation coefficient)*
Test d'hypothèse *(Hypothesis testing)*
Test d'indépendance du chi-carré *(Chi-square test of independence)*
Test non paramétrique *(Nonparametric test)*

TEST D'INDÉPENDANCE DU CHI-CARRÉ

Chi-square test of independence

Le test d'indépendance du chi-carré vise à déterminer si deux variables observées sur un **échantillon** sont indépendantes ou non. Les variables étudiées sont des variables qualitatives catégorielles.

Le test d'indépendance du chi-carré s'effectue sur la base d'une **table de contingence**.

HISTORIQUE

Les premières tables de contingence s'utilisaient uniquement dans un but d'énumération ou de comptage. Puis, avec **Adolphe Quetelet** (1849), on commença à s'intéresser aux associations qui pou-

vaient exister entre les variables. Ce furent les travaux de **Karl Pearson** (1900) qui constituèrent les fondements de l'étude des tables de contingence.

George Udny Yule (1900) proposa une autre approche de l'étude des tables de contingence qui fut très controversée par **Karl Pearson**, qui se disputait également avec **Ronald Aylmer Fisher** quant au nombre de degrés de liberté pour le test d'indépendance du chi-carré. Chacun donnait un nombre différent et il fut donné raison à R.A. Fisher (1922).

ASPECTS MATHÉMATIQUES

On considère deux variables qualitatives catégorielles X et Y. On dispose d'un **échantillon** de n observations sur ces variables.

Les observations obtenues peuvent être représentées par une **table de contingence**.

On note n_{ij}, la **fréquence** observée pour la **catégorie** i de la variable X et la catégorie j de la variable Y.

		Catégories de la variable Y			
		Y_1	...	Y_c	Total
Catégories de la variable X	X_1	n_{11}	...	n_{1c}	$n_{1.}$
	...	...	...	...	...
	X_r	n_{r1}	...	n_{rc}	$n_{r.}$
	Total	$n_{.1}$	...	$n_{.c}$	$n_{..}$

Les hypothèses à tester sont :

Hypothèse nulle

H_0 : les deux **variables** X et Y sont indépendantes.

Hypothèse alternative

H_1 : les deux variables X et Y ne sont pas indépendantes.

Étapes du test

1. Calculer les fréquences estimées notées e_{ij} de chaque case de la **table de contingence** sous l'**hypothèse** d'**indépendance** :

$$e_{ij} = \frac{n_{i.} \cdot n_{.j}}{n_{..}}$$

où $n_{i.} = \sum_{k=1}^{c} n_{ik}$ et $n_{.j} = \sum_{k=1}^{r} n_{kj}$

avec c représentant le nombre de colonnes (ou nombre de catégories de la variable X dans la table de contingence) et r, le nombre de lignes (ou nombre de catégories de la variable Y).

2. Calculer la **valeur** de la **statistique** χ^2 (chi-carré) qui est en réalité une mesure de l'écart entre les fréquences observées n_{ij} et les fréquences estimées e_{ij} :

$$\chi^2 = \sum_{i=1}^{c} \sum_{j=1}^{r} \frac{(n_{ij} - e_{ij})^2}{e_{ij}}.$$

3. Choisir le **seuil de signification** α du test et comparer la **valeur** de χ^2 calculée avec la valeur $\chi^2_{v,\alpha}$ qui peut être obtenue dans la **table du chi-carré**. Le nombre de degrés de liberté correspond au nombre de cases du tableau pour lesquelles il est possible de donner des valeurs arbitraires ; les autres valeurs étant imposées par les totaux sur les lignes et sur les colonnes. Ainsi, le nombre de degrés de liberté est égal à :

$$v = (r-1)(c-1).$$

4. Si le χ^2 calculé est inférieur au $\chi^2_{v,\alpha}$ de la table, on ne rejette pas l'hypothèse nulle : on considère que les deux variables sont indépendantes.

   Si, par contre, le χ^2 calculé est supérieur au $\chi^2_{v,\alpha}$ de la table, on rejettera l'hypothèse nulle au profit de l'hypothèse alternative. On concluera alors que les deux variables ne sont pas indépendantes.

DOMAINES ET LIMITATIONS

Certaines conditions doivent être remplies pour pouvoir appliquer un test d'indépendance du chi-carré :

1. l'**échantillon** formé des n observations doit être un échantillon aléatoire ;

2. chaque individu doit être classé de telle sorte qu'il se trouve dans une et une seule **catégorie** de chaque **variable**. Chaque individu ne doit donc apparaître que sur une seule ligne et une seule colonne de la **table de contingence**.

Il est à noter que le test d'indépendance du chi-carré est peu fiable pour des échantillons de petite taille, et notamment lorsque certaines fréquences estimées sont petites (< 5). Pour éviter cette situation, il convient de regrouper plusieurs catégories en une seule, pour autant que ce regroupement ait un sens.

EXEMPLE

On cherche à déterminer si le fait de fumer est indépendant du sexe des individus.

Les deux variables à étudier sont des variables qualitatives catégorielles qui comptent deux catégories chacune :

— pour la variable « sexe » : M ou F ;

— pour la variable « fumer » : « fume » ou « ne fume pas ».

Les hypothèses sont donc :

H_0 : fumer est indépendant du sexe ;
H_1 : fumer n'est pas indépendant du sexe.

La **table de contingence** obtenue à partir d'un **échantillon** de 100 individus ($n = 100$) est la suivante :

		Fumer		
		« fume »	« ne fume pas »	Total
Sexe	M	21	44	65
	F	10	25	35
	Total	31	69	100

On note n_{ij} ($i = 1, 2$, $j = 1, 2$) les fréquences observées.

On estime les fréquences de chaque case du tableau sous l'**hypothèse** de l'**indépendance** entre les deux variables. On note e_{ij}, ces fréquences estimées :

$$e_{ij} = \frac{n_{i.} \cdot n_{.j}}{n_{..}}.$$

On obtient :

$$\begin{aligned} e_{11} &= \frac{65 \cdot 31}{100} = 20,15 \\ e_{12} &= \frac{65 \cdot 69}{100} = 44,85 \\ e_{21} &= \frac{35 \cdot 31}{100} = 10,85 \\ e_{12} &= \frac{35 \cdot 69}{100} = 24,15. \end{aligned}$$

En définitive, le **tableau de fréquences** estimées est le suivant :

		Fumer		
		« fume »	« ne fume pas »	Total
Sexe	M	20,15	44,85	65
	F	10,85	24,15	35
	Total	31	69	100

Si l'**hypothèse nulle** H_0 est vraie, la **statistique**

$$\chi^2 = \sum_{i=1}^{2} \sum_{j=1}^{2} \frac{(n_{ij} - e_{ij})^2}{e_{ij}}$$

suit une **loi du chi-carré** avec $(r-1)(c-1) = (2-1)(2-1) = 1$ degré de liberté et

$$\begin{aligned} \chi^2 &= 0,036 + 0,016 + 0,066 + 0,030 \\ &= 0,148. \end{aligned}$$

En choisissant un **seuil de signification** de 5 %, la **valeur** $\chi^2_{1,0,05}$ de la **table du chi-carré** est égale à $3,84$.

Comme la valeur de χ^2 calculée est nettement inférieure à la valeur trouvée dans la table du chi-carré, on ne rejette pas l'hypothèse nulle et l'on conclut que les deux variables étudiées sont indépendantes.

MOTS CLÉS

Loi du chi-carré *(Chi-square distribution)*
Table de contingence *(Contingency table)*
Table du chi-carré *(Chi-square table)*
Test d'indépendance *(Test of independence)*

RÉFÉRENCES

Fisher, R.A. (1922). On the interpretation of χ^2 from contingency tables, and the calculation of P. *Journal of the Royal Statistical Society.* Series A, 85, pp. 87-94.

Pearson, K. (1900). 1948. On the Criterion, That a Given System of Deviations from the Probable in the Case of a Correlated System of Variables is such that it can be reasonably supposed to have arisen from random Sampling. pp. 339-357 in Karl Pearson, *Karl Pearson's Early Statistical Papers.* Cambridge Univ. Press. First published in the *Philosophical Magazine*, 5th Series, 50, pp. 157-175.

Quetelet, A. (1849). *Letters addressed to H.R.H. the Grand Duke of Saxe Coburg and Gotha, on the Theory of Probabilities as Applied to the Moral and Political Sciences.* (Traduction du français par Olinthus Gregory Downs). Charles & Edwin Layton, London.

Yule, G.U. (1900). On the association of attributes in statistics: with illustration from the material of the childhood society. *Philosophical Transactions of the Royal Society of London.* Series A, Vol. 194, pp. 257-319.

TEST DE DURBIN-WATSON

Durbin-Watson test

Le test de Durbin-Watson introduit une **statistique** d pour tester l'autocorrélation d'ordre 1 des résidus obtenus à partir d'un modèle de régression linéaire. Le problème apparaît souvent lors de l'application d'un modèle linéaire à une **série chronologique** et que l'on désire tester l'indépendance des résidus ainsi obtenus.

HISTORIQUE

L'invention de ce test est due à J. Durbin et G.S. Watson en 1950.

ASPECTS MATHÉMATIQUES

On prend le cas du modèle de la **régression linéaire multiple** comprenant $p-1$ variables indépendantes. Le modèle s'écrit :

$$Y_t = \beta_0 + \sum_{j=1}^{p-1} \beta_j X_{jt} + \varepsilon_t, \quad t = 1, \ldots, T$$

où

- Y_t est la **variable dépendante**,
- X_{jt}, avec $j = 1, \ldots, p-1$, sont les variables indépendantes
- β_j, avec $j = 1, \ldots, p-1$, sont les paramètres à estimer,
- ε_t avec $t = 1, \ldots, T$, est le terme d'erreur aléatoire non observable.

Sous forme matricielle, le modèle s'écrit :

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}.$$

où

- $\mathbf{Y}$ est le vecteur $(n \times 1)$ des observations relatives à la variable dépendante (n observations),
- $\boldsymbol{\beta}$ est le vecteur $(p \times 1)$ des paramètres à estimer,
- $\boldsymbol{\varepsilon}$ est le vecteur $(n \times 1)$ des erreurs.

et $\mathbf{X} = \begin{pmatrix} 1 & X_{11} & \ldots & X_{1(p-1)} \\ \vdots & \vdots & & \vdots \\ 1 & X_{n1} & \ldots & X_{n(p-1)} \end{pmatrix}$ est la **matrice** $(n \times p)$ du plan ayant trait aux variables indépendantes.

Les résidus obtenus par la méthode des **moindres carrés** sont donnés par :

$$\mathbf{e} = \mathbf{Y} - \widehat{\mathbf{Y}} = \left[\mathbf{I} - \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\right]\mathbf{Y}.$$

La **statistique** d est définie par le rapport suivant :

$$d = \frac{\sum_{t=2}^{T}\left(e_t - e_{t-1}\right)^2}{\sum_{t=1}^{T} e_t^2}.$$

La statistique d teste l'hypothèse d'indépendance des erreurs contre l'alternative qu'ils suivent un processus autorégressif d'ordre 1 :

$$\varepsilon_t = \rho\varepsilon_{t-1} + u_t$$

où $|\rho| < 1$ et les u_t sont indépendants et suivent une **loi normale** de **moyenne** nulle et **variance** σ^2. On appelle le terme ρ **autocorrélation**. En termes de ρ, le **test d'hypothèse** s'écrit comme suit :

$$\begin{aligned} H_0 &: \rho = 0, \\ H_1 &: \rho \neq 0. \end{aligned}$$

DOMAINES ET LIMITATIONS

De part sa construction, cette statistique varie entre 0 et 4. On a $d = 2$ lorsque $\widehat{\rho} = 0$. On note par $\widehat{\rho}$ l'observation de ρ. Afin de tester l'hypothèse H_0, Durbin et Watson ont tabulé les valeurs critiques de d au seuil de 5 % en fonction du nombre d'observations

T et du nombre de variables explicatives $(p-1)$. La lecture de la table permet de déterminer deux valeurs d_L et d_U, comprises entre 0 et 2 qui délimitent l'espace entre 0 et 4 selon le schéma suivant :

$$\begin{array}{rcl} d_U < d < 4 - d_U & : & \text{on accepte } H_0 \Rightarrow \rho = 0 \\ 0 < d < d_L & : & \text{on rejette } H_0 \Rightarrow \rho > 0 \\ 4 - d_L < d < 4 & : & \text{on rejette } H_0 \Rightarrow \rho < 0 \\ d_L < d < d_U & : & \text{zone d'indétermination} \\ 4 - d_U < d < 4 - d_L & : & \text{zone d'indétermination.} \end{array}$$

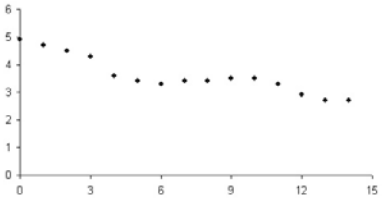
Le modèle doit comporter impérativement un terme constant, car généralement les tables de Durbin-Watson sont établies pour des modèles ayant un terme constant. La variable à expliquer ne doit pas figurer parmi les variables explicatives (en tant que variable retardée).

EXEMPLES

On considère l'exemple du nombre d'accidents sur l'autoroute aux États-Unis par million de miles parcourus par voiture entre 1970 et 1984. Pour chaque année, les données sont :

$$\begin{array}{llllll} 4,9\,; & 4,7\,; & 4,5\,; & 4,3\,; & 3,6\,; & 3,4\,; \\ 3,3\,; & 3,4\,; & 3,4\,; & 3,5\,; & 3,5\,; & 3,3\,; \\ 2,9\,; & 2,7\,; & 2,7. & & & \end{array}$$

La figure suivante illustre la répartition des données en fonction du temps. Le chiffre 0 correspond à l'année 1970, et le 15 à l'année 1984.



On a :

$$\text{nombre d'accidents} = a + b \cdot t,$$

avec $t = 0, 1, \ldots, 14$.

La droite de la régression linéaire simple obtenue est la suivante :

$$\text{nombre d'accidents} = 4,59 - 0,141 \cdot t.$$

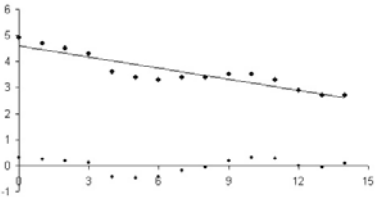
On obtient les résidus suivants :

$$\begin{array}{llllll} 0,31\,; & 0,25\,; & 0,19\,; & 0,13\,; & -0,43\,; & -0,49\,; \\ -0,45\,; & -0,21\,; & -0,07\,; & 0,17\,; & 0,32\,; & 0,26\,; \\ 0,00\,; & -0,06\,; & 0,08. & & & \end{array}$$

La valeur de la statique d est de $0,53$. Dans la table de Durbin-Watson pour une régression à une variable on trouve les valeurs :

$$\begin{array}{rcl} d_{L,0,05} & = & 1,08 \\ d_{U,0,05} & = & 1,36. \end{array}$$

On est dans le cas où $0 < d < d_L$ et donc on rejette l'hypothèse H_0 en faveur d'une autocorrélation positive entre les résidus du modèle. La figure suivante illustre la corrélation positive entre les résidus en fonction du temps :



MOTS CLÉS

Autocorrélation et autocorrélation partielle *(Autocorrelation and partial autocorrelation)*
Régression linéaire multiple *(Multiple linear regression)*
Régression linéaire simple *(Simple linear regression)*
Résidu *(Residual)*
Série chronologique *(Time series)*
Statistique *(Statistics)*

RÉFÉRENCES

Bourbonnais, R. (1998). *Économétrie, manuel et exercices corrigés.* 2e édition. Dunod, Paris.

Durbin, J. (1969). Alternative method to d-test. *Biometrika*, 56, pp. 1-15.

Durbin, J. and Watson, G.S. (1950). Testing for Serial Correlation in Least Squares Regression I, *Biometrika*, 37, pp. 409-428.

Durbin, J. and Watson, G.S. (1951). Testing for Serial Correlation in Least Squares Regression. II. *Biometrika*, 38, pp. 159-177.

Harvey, A.C. (1981). *The Econometric Analysis of Time Series.* Philip Allan, Oxford ; John Wiley & Sons, New York.

TEST DE FISHER

Fisher test

On appelle test de Fisher un **test d'hypothèse** basé sur une **valeur** observée d'une **statistique** de la forme :

$$F = \frac{U}{m} \cdot \frac{n}{V}$$

où U et V sont des variables aléatoires indépendantes suivant chacune une **loi du chi-carré** avec respectivement m et n degrés de liberté.

HISTORIQUE

Voir **loi de Fisher**.

ASPECTS MATHÉMATIQUES

On considère le **modèle** linéaire pour un **facteur** à t niveaux :

$$Y_{ij} = \mu + \tau_i + \varepsilon_{ij},$$

$$i = 1, 2, \ldots, t \ ; \ j = 1, 2, \ldots, n_i$$

où

- Y_{ij} représente la j-ème **observation** recevant le **traitement** i,
- μ la **moyenne** générale commune à tous les traitements,
- τ_i l'effet réel sur l'observation du traitement i et
- ε_{ij} l'**erreur** expérimentale de l'observation Y_{ij}.

Les variables aléatoires ε_{ij} sont indépendantes et distribuées selon une **loi normale** de moyenne 0 et de **variance** σ^2 : $\mathcal{N}(0, \sigma^2)$.

Dans ce cas, l'**hypothèse nulle** pour déterminer s'il existe une différence significative entre les t traitements, est la suivante :

$$H_0 \quad : \quad \tau_1 = \tau_2 = \ldots = \tau_t.$$

L'**hypothèse alternative** est formulée comme suit :

$$H_1 \quad : \quad \text{les valeurs de } \tau_i \ (i = 1, 2, \ldots, t) \text{ ne sont pas toutes identiques.}$$

Pour comparer la variabilité à l'intérieur de chaque **échantillon** avec la variabilité entre les échantillons, on forme le ratio dont le numérateur est une **estimation** de la **variance** entre les échantillons et le dénominateur une estimation de la variance à l'intérieur des échantillons. La **statistique** du test de Fisher, appelée aussi ratio F, est alors donnée par :

$$F = \frac{\displaystyle\sum_{i=1}^{t} \frac{n_i \cdot \left(\bar{Y}_{i.} - \bar{Y}_{..}\right)^2}{t-1}}{\displaystyle\sum_{i=1}^{t} \sum_{j=1}^{n_i} \frac{\left(Y_{ij} - \bar{Y}_{i.}\right)^2}{N-t}}$$

où

- n_i est le nombre d'observations de l'échantillon i $(i = 1, \ldots, t)$,
- $N = \displaystyle\sum_{i=1}^{t} n_i$ est le nombre total d'observations,
- $\bar{Y}_{i.} = \displaystyle\sum_{j=1}^{n_i} \frac{Y_{ij}}{n_i}$ est la moyenne à l'intérieur du groupe i, pour $i = 1, \ldots, t$ et
- $\bar{Y}_{..} = \frac{1}{N} \displaystyle\sum_{i=1}^{t} \sum_{j=1}^{n_i} Y_{ij}$ est la moyenne globale.

Après avoir choisi un **seuil de signification** α pour le test, on compare la valeur calculée F à la **valeur critique** appropriée de la **table de Fisher** pour $t-1$ et $N-t$ degrés de liberté, notée $F_{t-1,N-t,\alpha}$, pour un **test unilatéral**.

Si $F \geq F_{t-1,N-t,\alpha}$, on rejette l'hypothèse nulle, ce qui signifie qu'au moins un des traitements diffère des autres.

Si $F < F_{t-1,N-t,\alpha}$, on ne rejette pas l'hypothèse nulle et les traitements ne présentent pas de différence significative, ce qui revient également à dire que les t échantillons proviennent de la même **population**.

DOMAINES ET LIMITATIONS

Le test de Fisher est particulièrement utilisé en **analyse de variance**, en **analyse de covariance** et en **analyse de régression**.

EXEMPLES

Voir **analyse de variance pour un facteur** et **analyse de variance pour deux facteurs**.

MOTS CLÉS

Analyse de régression *(Regression analysis)*
Analyse de variance *(Analysis of variance)*
Loi de Fisher *(Fisher distribution)*
Table de Fisher *(Fisher table)*
Test d'hypothèse *(Hypothesis testing)*

TEST DE KOLMOGOROV-SMIRNOV

Kolmogorov-Smirnov test

Le test de Kolmogorov-Smirnov est un **test d'adéquation** non paramétrique qui vise à déterminer si les fonctions de répartition de deux populations sont identiques. Il est utilisé lorsque l'on est en présence de deux échantillons provenant de deux populations pouvant être différentes. Contrairement au **test de Mann-Whitney** ou au **test de Wilcoxon** dont l'objet est de détecter des différences entre deux moyennes ou médianes, le test de Kolmogorov-Smirnov a l'avantage de prendre en considération les fonctions de répartition dans leur ensemble. Le test de Kolmogorov-Smirnov peut aussi être utilisé comme **test d'adéquation**. Dans ce cas, on est en présence d'un seul échantillon aléatoire tiré d'une **population** dont la fonction de répartition est en fait une fonction de répartition spécifique et connue.

HISTORIQUE

Le **test d'adéquation** pour un **échantillon** est dû à Andreï Nicolaïévitch Kolmogorov (1933).

Le test de Kolmogorov-Smirnov pour deux échantillons est dû à Vladimir Ivanovitch Smirnov (1939).

Chez Massey (1952), on trouve une table de Smirnov pour le test de Kolmogorov-Smirnov pour deux échantillons et, chez Miller (1956), on trouve une table de Kolmogorov pour le test d'adéquation.

ASPECTS MATHÉMATIQUES

Soient deux échantillons aléatoires indépendants : $(X_1, X_2, \ldots, X_n)$, un échantillon de taille n provenant d'une **population** 1 et $(Y_1, Y_2, \ldots, Y_m)$, un échantillon de taille m provenant d'une population 2. On note respectivement $F(x)$ et $G(x)$ leur **fonction de répartition** inconnue.

Hypothèses

Les hypothèses à tester sont les suivantes :
A. Cas bilatéral :

H_0 : $F(x) = G(x)$ pour tout x,
H_1 : $F(x) \neq G(x)$ pour au moins une **valeur** de x.

B. Cas unilatéral :

H_0 : $F(x) \leq G(x)$ pour tout x,
H_1 : $F(x) > G(x)$ pour au moins une valeur de x.

C. Cas unilatéral :

H_0 : $F(x) \geq G(x)$ pour tout x,
H_1 : $F(x) < G(x)$ pour au moins une valeur de x.

Dans le cas A, on fait l'hypothèse qu'il n'y a pas de différence entre les fonctions de répartition des deux populations. Les deux populations peuvent donc être vues comme une seule population.

Dans le cas B, on fait l'hypothèse que la fonction de répartition de la population 1 est inférieure à celle de la population 2. On dit parfois d'une manière générale que les X tendent à être plus petits que les Y.

Dans le cas C, on fait l'hypothèse que les X sont plus grands que les Y.

On note $H_1(x)$ la fonction de répartition empirique de l'**échantillon** $(X_1, X_2, \ldots, X_n)$ et $H_2(x)$ la fonction de répartition empirique de l'échantillon $(Y_1, Y_2, \ldots, Y_m)$. Le test statistique est défini différemment selon les hypothèses posées.

A. Cas bilatéral
Le test statistique T_1 est défini comme la plus grande **distance** verticale entre les deux fonctions de répartition empiriques :

$$T_1 = \sup_x |H_1(x) - H_2(x)|.$$

B. Cas unilatéral
Le test statistique T_2 est défini comme la plus grande distance verticale quand $H_1(x)$ est supérieur à $H_2(x)$:

$$T_2 = \sup_x [H_1(x) - H_2(x)].$$

C. Cas unilatéral
Le test statistique T_3 est défini comme la plus grande distance verticale quand $H_2(x)$ est supérieur à $H_1(x)$:

$$T_3 = \sup_x [H_2(x) - H_1(x)].$$

Règle de décision

On rejette H_0 au **seuil de signification** α si le test statistique approprié (T_1, T_2 ou T_3) est supérieur à la **valeur** de la table de Smirnov ayant pour paramètres n, m et $1-\alpha$, que l'on note $t_{n,m,1-\alpha}$, c'est-à-dire si

$$T_1 \text{ (ou } T_2 \text{ ou } T_3) > t_{n,m,1-\alpha}.$$

Si on veut tester l'adéquation de la fonction de répartition inconnue $F(x)$ d'un seul échantillon aléatoire tiré d'une population à une fonction de répartition spécifique et connue $F_o(x)$, alors les hypothèses sont les mêmes que pour le test à deux échantillons sauf que $F(x)$ et $G(x)$ sont remplacées par $F(x)$ et $F_o(x)$. Si $H(x)$ est la fonction de répartition empirique de l'échantillon aléatoire, les tests statistiques T_1, T_2 et T_3 sont définis comme suit :

$$\begin{aligned} T_1 &= \sup_x |F_o(x) - H(x)|, \\ T_2 &= \sup_x [F_o(x) - H(x)], \\ T_3 &= \sup_x [H(x) - F_o(x)]. \end{aligned}$$

La règle de décision est la suivante :

rejeter H_0 au seuil de signification α si T_1 (ou T_2 ou T_3) est supérieur à la valeur de la table de Kolmogorov ayant pour paramètres n et $1-\alpha$, que l'on note $t_{n,1-\alpha}$, c'est-à-dire si

$$T_1 \text{ (ou } T_2 \text{ ou } T_3) > t_{n,1-\alpha}.$$

DOMAINES ET LIMITATIONS

Les exigences préalables à respecter pour pouvoir effectuer le test de Kolmogorov-Smirnov sont les suivantes :

1. les deux échantillons sont des échantillons aléatoires tirés de leur **population** respective ;
2. il y a **indépendance** mutuelle entre les deux échantillons ;
3. l'échelle de mesure est au moins ordinale ;
4. pour que ce test soit exact, les variables aléatoires doivent être continues, sinon le test est moins précis.

EXEMPLES

Le premier exemple traite du test de Kolmogorov-Smirnov pour deux échantillons et le deuxième du test de Kolmogorov-Smirnov en tant que **test d'adéquation**.

Exemple 1 : dans une classe, on compte 25 élèves : 15 garçons et 10 filles. On fait passer à cette classe un test de calcul mental pour voir si les garçons ont tendance à être meilleurs que les filles dans cette discipline.

Les données se trouvent dans le tableau ci-dessous, les scores plus élevés correspondant à de meilleurs résultats au test.

Garçons (X_i)		Filles (Y_i)	
19,8	17,5	17,7	14,1
12,3	17,9	7,1	23,6
10,6	21,1	21,0	11,1
11,3	16,4	10,7	20,3
13,3	7,7	8,6	15,7
14,0	15,2		
9,2	16,0		
15,6			

On teste l'**hypothèse** selon laquelle les distributions des résultats des filles et celles des garçons sont identiques. Cela signifie que la **population** d'où est tiré l'échantillon des X a la même **fonction de répartition** que la population d'où est tiré l'échantillon des Y. D'où l'**hypothèse nulle** :

$$H_0 \quad : \quad F(x) = G(x) \text{ pour tout } x.$$

Puisque le cas bilatéral s'applique ici, on calcule :

$$T_1 = \sup_x |H_1(x) - H_2(x)|$$

où $H_1(x)$ et $H_2(x)$ sont les fonctions de répartition empiriques des échantillons $(X_1, X_2, \ldots, X_{15})$ et $(Y_1, Y_2, \ldots, Y_{10})$ respectivement.
Dans le tableau ci-dessous, on a classé les observations des deux échantillons par ordre croissant pour simplifier le calcul de $H_1(x) - H_2(x)$.

X_i	Y_i	$H_1(x) - H_2(x)$
	7.1	$0 - 1/10 = -0,1$
7,7		$1/15 - 1/10 = -0,0333$
	8,6	$1/15 - 2/10 = -0,1333$
9,2		$2/15 - 2/10 = -0,0667$
10,6		$3/15 - 2/10 = 0$
	10,7	$3/15 - 3/10 = -0,1$
	11,1	$3/15 - 4/10 = -0,2$
11,3		$4/15 - 4/10 = -0,1333$
12,3		$5/15 - 4/10 = -0,0667$
13,3		$6/15 - 4/10 = 0$
14,0		$7/15 - 4/10 = 0,0667$
	14,1	$7/15 - 5/10 = -0,0333$
15,2		$8/15 - 5/10 = 0,0333$
15,6		$9/15 - 5/10 = 0,1$
	15,7	$9/15 - 6/10 = 0$
16,0		$10/15 - 6/10 = 0,0667$
16,4		$11/15 - 6/10 = 0,1333$
17,5		$12/15 - 6/10 = 0,2$
	17,7	$12/15 - 7/10 = 0,1$
17,9		$13/15 - 7/10 = 0,1667$
19,8		$14/15 - 7/10 = 0,2333$
	20,3	$14/15 - 8/10 = 0,1333$
	21,0	$14/15 - 9/10 = 0,0333$
21,1		$1 - 9/10 = 0,1$
	23,6	$1 - 1 = 0$

On a donc :

$$T_1 = \sup_x |H_1(x) - H_2(x)| = 0,2333.$$

La **valeur** de la table de Smirnov pour $n = 15$, $m = 10$ et $1 - \alpha = 0,95$ est égale à $t_{15,10,0,95} = 0,5$.

Par conséquent, $T_1 = 0,2333 < t_{15,10,0,95} = 0,5$ et H_0 ne peut être rejetée. Cela signifie qu'il n'y a pas de différence significative entre le niveau de calcul mental des filles et des garçons.

Exemple 2 : soit l'échantillon aléatoire de taille 10 suivant :
$X_1 = 0,695$, $X_2 = 0,937$, $X_3 = 0,134$, $X_4 = 0,222$, $X_5 = 0,239$, $X_6 = 0,763$, $X_7 = 0,980$, $X_8 = 0,322$, $X_9 = 0,523$, $X_{10} = 0,578$

On désire vérifier par le test de Kolmogorov-Smirnov si cet échantillon provient d'une **loi uniforme**. La fonction de répartition de la loi uniforme est donnée par :

$$F_o(x) = \begin{cases} 0 & \text{si } x < 0 \\ x & \text{si } 0 \leq x < 1 \\ 1 & \text{sinon} \end{cases}$$

L'**hypothèse nulle** H_0 est donc la suivante, si $F(x)$ est la fonction de répartition inconnue de la population associée à l'échantillon :

$$H_0 \quad : \quad F(x) = F_o(x) \text{ pour tout } x.$$

Puisque le cas bilatéral s'applique, on calcule :

$$T_1 = \sup_x |F_o(x) - H(x)|.$$

où $H(x)$ est la fonction de répartition empirique de l'échantillon $(X_1, X_2, \ldots, X_{10})$.

Dans le tableau ci-dessous, on a classé les 10 observations par ordre croissant pour simplifier le calcul de $F_0(x) - H(x)$.

X_i	$F_o(x)$	$H(x)$	$F_o(x) - H(x)$
0,134	0,134	0,1	$0,134 - 0,1 = 0,034$
0,222	0,222	0,2	$0,222 - 0,2 = 0,022$
0,239	0,239	0,3	$0,239 - 0,3 = -0,061$
0,322	0,322	0,4	$0,322 - 0,4 = -0,078$
0,523	0,523	0,5	$0,523 - 0,5 = 0,023$
0,578	0,578	0,6	$0,578 - 0,6 = -0,022$
0,695	0,695	0,7	$0,695 - 0,7 = -0,005$
0,763	0,763	0,8	$0,763 - 0,8 = -0,037$
0,937	0,937	0,9	$0,937 - 0,9 = 0,037$
0,980	0,980	1,0	$0,980 - 1,0 = -0,020$

On obtient donc :

$$T_1 = \sup_x |F_o(x) - H(x)| = 0,078,$$

La valeur de la table de Kolmogorov pour $n = 10$ et $1 - \alpha = 0,95$ est $t_{10,0,95} = 0,409$.

Puisque T_1 est inférieur à $t_{10,0,95}$ $(0,078 < 0,409)$, H_0 ne peut être rejetée. Cela signifie que l'échantillon aléatoire pourrait bien provenir d'une population distribuée selon une loi uniforme.

MOTS CLÉS

Test d'adéquation *(Goodness of fit test)*
Test d'hypothèse *(Hypothesis testing)*
Test non paramétrique *(Nonparametric test)*

RÉFÉRENCES

Kolmogorov, A.N. (1933). Sulla determinazione empirica di una legge di distribuzione. *Giornale dell' Instituto Italiano degli Attuari*, 4, pp. 83-91 (6.1).

Massey, F.J. (1952). Distribution table for the deviation between two sample cumulatives. *Annals of Mathematical Statistics*, 23, pp. 435-441.

Miller, L.H. (1956). Table of percentage points of Kolmogorov statistics. *Journal of the American Statistical Association*, 31, pp. 111-121.

Smirnov, N.V. (1939). Estimate of deviation between empirical distribution functions in two independent samples. (Russian). *Bulletin Moscow Univ.* 2: 2, pp. 3-16 (6.1, 6.2).

Smirnov, N.V. (1948). Table for Estimating the Goodness of Fit of Empirical Distributions. *Annals of Mathematical Statistics*, 19, pp. 279-281 (6.1).

TEST DE KRUSKAL-WALLIS

Kruskal-Wallis test

Le test de Kruskal-Wallis est un **test non paramétrique** visant à déterminer si k populations sont toutes identiques ou si au moins une des populations tend à fournir des observations différentes des autres populations.

Le test est utilisé lorsque l'on est en présence de k échantillons, avec $k \geq 2$, provenant de k populations pouvant être différentes.

HISTORIQUE

Le test de Kruskal-Wallis a été développé en 1952 par W.H. Kruskal et W.A. Wallis.

ASPECTS MATHÉMATIQUES

Les données se présentent en k échantillons. On désigne par n_i la taille de l'**échantillon** i, pour $i = 1, \ldots, k$ et par N, le nombre total d'observations :

$$N = \sum_{i=1}^{k} n_i.$$

On classe les N observations par ordre croissant sans tenir compte de l'appartenance aux échantillons. On attribue ensuite le rang 1 à la plus petite **valeur**, le rang 2 à la valeur juste supérieure et ainsi de suite jusqu'au rang N attribué à la plus grande valeur.

Soit X_{ij}, la j-ème observation de l'échantillon i, $i = 1, \ldots, k$ et $j = 1, \ldots, n_i$, on désigne le rang attribué à X_{ij} par $R(X_{ij})$.

Si plusieurs observations ont la même valeur, on leur attribuera un rang moyen. La somme des rangs attribués aux observations de l'échantillon i étant désignée par R_i, on a :

$$R_i = \sum_{j=1}^{n_i} R(X_{ij}), \quad i = 1, \ldots, k.$$

S'il n'y a pas de rangs moyens (ou bien s'ils sont en nombre limité), le test statistique est défini de la façon suivante :

$$H = \left(\frac{12}{N(N+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} \right) - 3(N+1).$$

Si, au contraire, il y a beaucoup de rangs moyens, il est nécessaire d'effectuer une correction et de

calculer :

$$\widetilde{H} = \frac{H}{1 - \frac{\sum_{i=1}^{g}(t_i^3 - t_i)}{N^3 - N}}$$

où g est le nombre de groupes de rangs moyens et t_i la taille du groupe i.

Hypothèses

Le but du test de Kruskal-Wallis est de déterminer si toutes les populations sont identiques ou si au moins une des populations tend à fournir des observations différentes des autres populations. Les hypothèses sont donc les suivantes :

H_0 : il n'y a pas de différence entre les k populations,
H_1 : au moins une des populations est différente des autres populations.

Règle de décision

S'il y a 3 échantillons, chacun ayant une taille inférieure ou égale à 5 et s'il n'y a pas de rangs moyens (c'est-à-dire si c'est H qui est calculé), on utilise la **table de Kruskal-Wallis** pour tester H_0.

La règle de décision est la suivante : on rejette l'**hypothèse nulle** H_0 au **seuil de signification** α si T est supérieur à la valeur de la table avec les paramètres n_i, $k-1$ et $1-\alpha$ notée $h_{n_1,n_2,n_3,1-\alpha}$ et s'il n'y a pas de table exacte disponible ou s'il y a des rangs moyens, on peut faire une approximation de la valeur de la table de Kruskal-Wallis par la distribution du chi-carré avec $k-1$ degrés de liberté (**loi du chi-carré**), c'est-à-dire si :

$H > h_{n_1,n_2,n_3,1-\alpha}$ (**table de Kruskal-Wallis**)
ou $H > \chi^2_{k-1,1-\alpha}$ (**table du chi-carré**).

La règle de décision correspondante est basée sur : $\widetilde{H} > \chi^2_{k-1,1-\alpha}$ (table du chi-carré).

DOMAINES ET LIMITATIONS

Les exigences préalables qu'il faut respecter pour pouvoir effectuer le test de Kruskal-Wallis sont les suivantes :

1. tous les échantillons sont des échantillons aléatoires tirés de leur **population** respective ;
2. en plus de l'**indépendance** à l'intérieur de chaque échantillon, il y a indépendance mutuelle entre les différents échantillons ;
3. l'échelle de mesure est au moins ordinale.

Si le test de Kruskal-Wallis conduit à rejeter l'**hypothèse nulle** H_0, on peut utiliser le **test de Wilcoxon** pour tous les échantillons pris deux à deux, afin de déterminer quelles paires de populations tendent à être différentes.

EXEMPLE

On fait frire des pommes de terre dans 4 huiles différentes. On veut vérifier si la quantité de graisse absorbée par les pommes de terre dépend du type d'huile utilisé. On réalise cinq expériences différentes avec l'huile 1, six avec l'huile 2, quatre avec l'huile 3 et cinq avec l'huile 4, et on obtient les données suivantes :

Type d'huile			
1	2	3	4
64	78	75	55
72	91	93	66
68	97	78	49
77	82	71	64
56	85		70
	77		

Dans cet exemple, le nombre d'échantillons est égal à 4 ($k = 4$) avec les tailles respectives suivantes :

$$\begin{aligned} n_1 &= 5 \\ n_2 &= 6 \\ n_3 &= 4 \\ n_4 &= 5. \end{aligned}$$

Le nombre total d'observations est donc égal à :

$$N = 5 + 6 + 4 + 5 = 20.$$

On classe les observations par ordre croissant et on leur attribue un rang de 1 à 20 en tenant compte des rangs moyens. On obtient donc le tableau suivant avec entre parenthèses le rang des observations et en bas de chaque échantillon la somme des rangs attribués à l'échantillon correspondant :

Type d'huile			
1	2	3	4
64 (4,5)	78 (14,5)	75 (11)	55 (2)
72 (10)	91 (18)	93 (19)	66 (6)
68 (7)	97 (20)	78 (14,5)	49 (1)
77 (12,5)	82 (16)	71 (9)	64 (4,5)
56 (3)	85 (17)		70 (8)
	77 (12,5)		
$R_1 = 37$	$R_2 = 98$	$R_3 = 53,5$	$R_4 = 21,5$

On calcule :

$$\begin{aligned} H &= \left(\frac{12}{N(N+1)}\sum_{i=1}^{k}\frac{R_i^2}{n_i}\right) - 3(N+1) \\ &= \frac{12}{20(20+1)}\left(\frac{37^2}{5}+\frac{98^2}{6}+\frac{53,5^2}{4}+\frac{21,5^2}{5}\right) \\ &\quad -3(20+1) \\ &= 13,64. \end{aligned}$$

Si l'on fait l'ajustement pour prendre en compte les rangs moyens, on obtient :

$$\begin{aligned} \widetilde{H} &= \frac{H}{1-\frac{\sum_{i=1}^{g}(t_i^3-t_i)}{N^3-N}} \\ &= \frac{13,64}{1-\frac{8-2+8-2+8-2}{20^3-20}} \\ &= 13,67. \end{aligned}$$

On constate que la différence entre H et $\widetilde{H}$ est minime.

Les hypothèses sont les suivantes :

H_0 : il n'y a pas de différence entre les 4 huiles,
H_1 : au moins une des huiles est différente des autres.

La règle de décision est la suivante :
rejeter l'**hypothèse nulle** H_0 au **seuil de signification** α si

$$H > \chi^2_{k-1,1-\alpha},$$

$\chi^2_{k-1,1-\alpha}$ étant la **valeur** de la **table du chi-carré** au niveau $1-\alpha$ et avec $k-1=3$ degrés de liberté.

Si l'on choisit $\alpha = 0,05$, la valeur de $\chi^2_{3,0,95}$ est 7,81 et puisque H est supérieur à $\chi^2_{3,0,95}$ $(13,64 > 7,81)$, on rejette l'hypothèse H_0.

Puisque H_0 est rejetée, on peut utiliser la procédure de comparaisons multiples pour voir quelles paires d'huiles sont différentes.

MOTS CLÉS

Test d'hypothèse *(Hypothesis testing)*
Test de Wilcoxon *(Wilcoxon test)*
Test non paramétrique *(Nonparametric test)*

RÉFÉRENCE

Kruskal, W.H. and Wallis, W.A. (1952). Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association*, 47, pp. 583-621 and errata, *ibid.*, 48, pp. 907-911.

TEST DE LA DIFFÉRENCE SIGNIFICATIVE MINIMALE

Least significant difference test

Le test de la différence significative minimale (en anglais *least significant difference test* et abrégé LSD) s'emploie dans le contexte d'une **analyse de variance**, lorsque le ratio F conduit à rejeter l'**hypothèse nulle** H_0 : c'est-à-dire quand la différence entre les moyennes des populations est significative.

Le test sert à identifier parmi les différentes populations, celles dont les moyennes sont significativement différentes. Le principe est d'effectuer des comparaisons de moyennes prises deux à deux. Ce test peut être utilisé à la suite d'une **analyse de variance pour un facteur** ou deux facteurs, pour autant que les hypothèses nulles aient été rejetées.

HISTORIQUE

Le test de la différence significative minimale a été développé par **Ronald Aylmer Fisher** (1935), pour repérer les traitements dont l'effet est significatif dans une **analyse de variance**.

ASPECTS MATHÉMATIQUES

Si dans une **analyse de variance** le ratio F conduit à rejeter l' **hypothèse nulle** H_0, on peut faire le test de la différence significative minimale afin de détecter quelles moyennes ont entraîné le rejet de l'hypothèse nulle.

Le test consiste à établir des comparaisons de moyennes prises deux à deux. En terme général, l'**écart type** de la différence entre la moyenne i et la moyennes j est égal à :

$$\sqrt{s_I^2\left(\frac{1}{n_i}+\frac{1}{n_j}\right)}$$

où s_I^2 est l'**estimation** de la **variance** à l'intérieur des groupes :

$$s_I^2 = \frac{SCI}{N-k} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(Y_{ij}-\bar{Y}_{i.}\right)^2}{N-k}$$

où

SCI	=	la somme des carrés à l'intérieur des groupes,
N	=	le nombre total d'observations,
k	=	le nombre de groupes (ou d'échantillons),
Y_{ij}	=	la j-ème observation du groupe i,
$\bar{Y}_{i.}$	=	la moyenne du groupe i,
n_i	=	représente le nombre d'observations du groupe i et
n_j	=	celui du groupe j.

Le ratio

$$\frac{\bar{Y}_{i.}-\bar{Y}_{j.}}{\sqrt{s_I^2\left(\frac{1}{n_i}+\frac{1}{n_j}\right)}}$$

suit une **loi de Student** avec $N-k$ degrés de liberté. La différence entre une paire spécifique de moyennes est significative quand

$$\left|\bar{Y}_{i.}-\bar{Y}_{j.}\right| \geq \sqrt{s_I^2\left(\frac{1}{n_i}+\frac{1}{n_j}\right)}\cdot t_{N-k,1-\frac{\alpha}{2}}$$

où $t_{N-k,1-\frac{\alpha}{2}}$ est la **valeur** de la **table de Student** pour $N-k$ degrés de liberté et au **seuil de signification** α.

La quantité $\sqrt{s_I^2(\frac{1}{n_i}+\frac{1}{n_j})}\cdot t_{N-k,1-\frac{\alpha}{2}}$ est appelée LSD.

DOMAINES ET LIMITATIONS

Comme pour l'**analyse de variance**, le test de la différence significative minimale requiert des exigences préalables notamment la normalité des distributions sous-jacentes, l'homogénéité des variances et l'**indépendance** des variables.

EXEMPLE

Pendant leur cuisson, les croissants absorbent de la graisse en quantité variable. On veut vérifier si la quantité absorbée dépend du type de graisse. On prépare quatre genres de graisse différents et on fait cuire six croissants pour chacun d'entre eux.

Voici les résultats obtenus qui expriment les quantités de graisse absorbées par croissant :

Type de graisse			
1	2	3	4
64	78	75	55
72	91	93	66
68	97	78	49
77	82	71	64
56	85	63	70
95	77	76	68

On effectue l'**analyse de variance pour un facteur** qui permet de tester l'**hypothèse nulle**

$$H_0 \quad : \quad \mu_1 = \mu_2 = \mu_3 = \mu_4$$

et celle-ci indique qu'il y a une différence significative entre les moyennes.

Comme H_0 est rejetée, on effectue le test de la différence significative minimale dans le but d'identifier quelles moyennes provoquent le rejet de l'hypothèse nulle.

Dans cet exemple, le nombre d'observations est identique pour chaque groupe ($n_1 = n_2 = n_3 = n_4 = n = 6$). Par conséquent, $\sqrt{s_I^2\left(\frac{1}{n_i}+\frac{1}{n_j}\right)}$ se simplifie en $\sqrt{\frac{2s_I^2}{n}}$.

La **valeur** de LSD correspondante est égale à :

$$\sqrt{\frac{2s_I^2}{n}} \cdot t_{N-k,1-\frac{\alpha}{2}}.$$

Les résultats de l'**analyse de variance** donnent les éléments de calcul pour trouver la valeur de s_I^2 :

$$\begin{aligned} s_I^2 &= \frac{SCI}{N-k} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(Y_{ij}-\bar{Y}_{i.}\right)^2}{N-k} \\ &= \frac{2018}{20} \\ &= 100,9. \end{aligned}$$

Si on choisit un **seuil de signification** $\alpha = 0,05$, la valeur de la **table de Student** pour $N-k=20$ degrés de liberté est égale à :

$$t_{N-k,1-\frac{\alpha}{2}} = t_{20,0,975} = 2,086.$$

La valeur de LSD est donc :

$$\begin{aligned} LSD &= \sqrt{\frac{2s_I^2}{n}} \cdot t_{20,0,975} \\ &= \sqrt{\frac{2 \cdot 100,9}{6}} \cdot 2,086 \\ &= 12,0976. \end{aligned}$$

La différence entre une paire de moyennes $\bar{Y}_i$ et $\bar{Y}_j$ est significative si

$$\left|\bar{Y}_{i.} - \bar{Y}_{j.}\right| \geq 12,0976.$$

On peut résumer les résultats dans un tableau comprenant toutes les différences de moyennes, la valeur de LSD et la conclusion du test de la différence significative minimale :

Différence	LSD	Significative
$\left\|\bar{Y}_{1.} - \bar{Y}_{2.}\right\| = \|72-85\| = 13$	12,0976	oui
$\left\|\bar{Y}_{1.} - \bar{Y}_{3.}\right\| = \|72-76\| = 4$	12,0976	non
$\left\|\bar{Y}_{1.} - \bar{Y}_{4.}\right\| = \|72-62\| = 10$	12,0976	non
$\left\|\bar{Y}_{2.} - \bar{Y}_{3.}\right\| = \|85-76\| = 9$	12,0976	non
$\left\|\bar{Y}_{2.} - \bar{Y}_{4.}\right\| = \|85-62\| = 23$	12,0976	oui
$\left\|\bar{Y}_{3.} - \bar{Y}_{4.}\right\| = \|76-62\| = 14$	12,0976	oui

On vérifie que seules les différences $\bar{Y}_{1.} - \bar{Y}_{2.}$, $\bar{Y}_{2.} - \bar{Y}_{4.}$ et $\bar{Y}_{3.} - \bar{Y}_{4.}$ sont significatives.

MOTS CLÉS

Analyse de variance *(Analysis of variance)*
Loi de Student *(Student distribution)*
Table de Student *(Student table)*

RÉFÉRENCES

Dodge, Y. and Thomas, D.R. (1980). On the performance of non-parametric and normal theory multiple comparison procedures. *The Indian Journal of Statistics*, Vol. 42, Serie B, pp. 11-27.

Fisher, R.A. (1935). *The Design of Experiments.* Oliver & Boyd, Edinburgh.

Miller, R.G. Jr. (1981). *Simultaneous Statistical Inference.* 2nd ed. Springer-Verlag, New York.

TEST DE MANN-WHITNEY

Mann-Whitney test

Le test de Mann-Whitney est un **test non paramétrique** visant à tester l'égalité de deux populations. Il est utilisé lorsque l'on est en présence de deux échantillons provenant de deux populations.

HISTORIQUE

Le test de Mann-Whitney a tout d'abord été introduit pour le cas $m = n$ par **Frank Wilcoxon** (1945).

Par la suite, le **test de Wilcoxon** a été étendu au cas d'échantillons de tailles différentes par C. White et van der Reyden (1952).

Il est à noter qu'un test équivalent à celui de Wilcoxon a été développé de manière indépendante et introduit par L. Festinger (1946). H.B. Mann et D.R. Whitney (1947) furent les premiers à considérer des échantillons de tailles différentes et à fournir des tables adaptées aux petits échantillons.

ASPECTS MATHÉMATIQUES

Soit $(X_1, X_2, \ldots, X_n)$ un **échantillon** de taille n

provenant d'une **population** 1 et soit $(Y_1, Y_2, \ldots, Y_m)$ un échantillon de taille m provenant d'une population 2.

On obtient $N = n + m$ observations que l'on classe dans un ordre croissant en notant chaque fois à quel échantillon appartient l'observation.

La **statistique** de Mann et Whitney, notée U, est définie comme le nombre total de fois qu'un X_i précède un Y_j dans la **classification** par ordre croissant des N observations.

Quand la taille des échantillons est importante, la détermination de U devient longue et fastidieuse mais on peut utiliser la relation suivante qui donne des résultats identiques :

$$U = mn + \frac{n\,(n+1)}{2} - T$$

où T est la somme des rangs attribués aux X, provenant du **test de Wilcoxon**.

Hypothèses

Les hypothèses correspondant au test de Mann-Whitney peuvent être formulées de la façon suivante, selon qu'il s'agit d'un **test bilatéral** ou d'un **test unilatéral** :

A. Cas bilatéral :

$$\begin{aligned} H_0 &: P\,(X < Y) = \tfrac{1}{2}, \\ H_1 &: P\,(X < Y) \neq \tfrac{1}{2}. \end{aligned}$$

B. Cas unilatéral :

$$\begin{aligned} H_0 &: P\,(X < Y) \leq \tfrac{1}{2}, \\ H_1 &: P\,(X < Y) > \tfrac{1}{2}. \end{aligned}$$

C. Cas unilatéral :

$$\begin{aligned} H_0 &: P\,(X < Y) \geq \tfrac{1}{2}, \\ H_1 &: P\,(X < Y) < \tfrac{1}{2}. \end{aligned}$$

Dans le cas A, l'**hypothèse nulle** correspond à la situation où il n'y a pas de différence entre les deux populations. Dans le cas B, l'hypothèse nulle signifie que la population 1 est plus grande que la population 2. Dans le cas C, l'hypothèse nulle indique que la population 1 est plus petite que la population 2.

Règles de décision

Si $m, n < 12$, on compare la **statistique** U avec la **valeur** trouvée dans la table de Mann-Whitney pour tester H_0.

Par contre, si $m, n \geq 12$, la **distribution d'échantillonnage** de U approche très rapidement la distribution normale avec

la **moyenne** $\mu = \dfrac{m \cdot n}{2}$

et l'**écart type** $\sigma = \sqrt{\dfrac{mn\,(m+n+1)}{12}}$.

Et par conséquent, on peut déterminer la signification d'une valeur observée de U par :

$$Z = \frac{U - \mu}{\sigma}$$

où Z est une **variable aléatoire** centrée réduite.

En d'autres termes, on peut comparer la valeur de Z ainsi obtenue avec celle de la **table normale**.

Par ailleurs, il faut savoir que les règles de décision sont différentes selon les hypothèses posées. C'est ainsi que l'on a les règles A, B et C relatives aux cas précédents A, B et C.

Règle de décision A

On rejette l'hypothèse nulle H_0 au **seuil de signification** α si U est inférieur à la valeur de la table de Mann-Whitney avec les paramètres n, m et $\frac{\alpha}{2}$ notée $t_{n,m,1-\frac{\alpha}{2}}$ ou si U est plus grand que la valeur de la table de Mann-Whitney pour n, m et $1 - \frac{\alpha}{2}$, notée $t_{n,m,\frac{\alpha}{2}}$, c'est-à-dire si

$$t_{n,m,1-\frac{\alpha}{2}} < U < t_{n,m,\frac{\alpha}{2}}.$$

Règle de décision B

On rejette l'hypothèse nulle H_0 au seuil de signification α si U est inférieur à la valeur de la table de Mann-Whitney avec les paramètres n, m et α, notée $t_{n,m,\alpha}$, c'est-à-dire si

$$U < t_{n,m,\alpha}.$$

Règle de décision C

On rejette l'hypothèse nulle H_0 au seuil de signification α si U est supérieur à la valeur de la table de Mann-Whitney avec les paramètres n, m et $1-\alpha$, notée $t_{n,m,1-\alpha}$, c'est-à-dire si

$$U > t_{n,m,1-\alpha}.$$

DOMAINES ET LIMITATIONS

Étant donné qu'il existe une relation linéaire entre le **test de Wilcoxon** et le test de Mann-Whitney $\left(U = mn + \dfrac{n(n+1)}{2} - T\right)$, ces deux tests sont équivalents.

Par conséquent, ils ont des propriétés communes ; en particulier, les exigences préalables qu'il faut respecter sont identiques, notamment :

1. les deux échantillons sont des échantillons aléatoires tirés de leur **population** respective.
2. en plus de l'**indépendance** à l'intérieur de chaque échantillon, il y a indépendance mutuelle entre les deux échantillons.
3. l'échelle de mesure est au moins ordinale.

EXEMPLE

Dans une classe, on compte 9 élèves : 5 garçons et 4 filles. On fait passer à cette classe un test de calcul mental pour voir si les gar çons ont tendance à être meilleurs que les filles dans cette discipline.

Les données se trouvent dans le tableau ci-dessous, les scores plus élevés correspondant à de meilleurs résultats au test :

Garçons (X_i)	Filles (Y_i)
$9,2$	$11,1$
$11,3$	$15,6$
$19,8$	$20,3$
$21,1$	$23,6$
$24,3$	

On classe ces scores dans un ordre croissant en notant chaque fois à quel **échantillon** appartient l'**observation**.

Scores	Échantillons	Rang	Nombre de fois que X précède Y
$9,2$	X	1	4
$11,1$	Y	2	-
$11,3$	X	3	3
$15,6$	Y	4	-
$19,8$	X	5	2
$20,3$	Y	6	-
$21,1$	X	7	1
$23,6$	Y	8	-
$24,3$	X	9	0

On trouve $U = 4+3+2+1 = 10$. De même, si on utilise la relation $U = mn + \frac{n(n+1)}{2} - T$, on a :

$$T = \sum_{i=1}^{n} R(X_i) = 1+3+5+7+9 = 25$$

et $U = 5 \cdot 4 + \frac{5 \cdot 6}{2} - 25 = 10$.

Les hypothèses que l'on pose sont les suivantes :

H_0 : les garçons n'ont pas tendance à être meilleurs que les filles en calcul mental.
H_1 : les garçons ont tendance à être meilleurs que les filles en calcul mental.

Cela se traduit par H_0 : $P(X < Y) \geq 1/2$ et correspond au cas C. La règle de décision est donc la suivante :

rejeter H_0 au **seuil de signification** α si

$$U > t_{n,m,1-\alpha},$$

où $t_{n,m,1-\alpha}$ est la **valeur** de la table de Mann-Whitney avec les paramètres n, m et $1-\alpha$. On rappelle que, dans notre cas, $m = 4$ et $n = 5$. Étant donné qu'il existe une **relation** entre le **test de Wilcoxon** et le test de Mann-Whitney, au lieu d'utiliser la table de U on utilise la **table de Wilcoxon** pour T. Si on choisit $\alpha = 0,05$, la **valeur** de la table de Wilcoxon vaut 32. Comme $T = 25 < 32$, on ne rejette pas H_0 et on tire la conclusion que les garçons n'ont pas tendance à être meilleurs que les filles en calcul mental.

MOTS CLÉS

Test d'hypothèse *(Hypothesis testing)*
Test de Wilcoxon *(Wilcoxon test)*
Test non paramétrique *(Nonparametric test)*

RÉFÉRENCES

Festinger, L. (1946). The significance of difference between means without reference to the frequency distribution function. *Psychometrika*, 11, pp. 97-105.

Mann, H.B. and Whitney, D.R. (1947). On a test whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics*, 18, pp. 50-60.

White, C. (1952). The Use of Ranks in a Test of Significance for Comparing Two Treatments. *Biometrics*, 8, pp. 33-41.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics*, 1, pp. 80-83.

TEST DE STUDENT

Student test

Le test de Student est un **test d'hypothèse** paramétrique sur la **moyenne** d'un **échantillon**, ou sur la comparaison des moyennes de deux échantillons.

HISTORIQUE

William Sealy Gosset, dit « Student », développa en 1908 le **test de Student** et la table portant son nom.

ASPECTS MATHÉMATIQUES

Soit X une **variable aléatoire** distribuée selon une **loi normale**. La variable aléatoire T définie ci-dessous suit une **loi de Student** avec $n-1$ degrés de liberté :

$$T = \frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}}$$

où

- $\bar{X}$ est la **moyenne arithmétique** de l'**échantillon**,
- μ est la moyenne de la **population**,
- S est l'**écart type** de l'échantillon et
- n est la **taille d'échantillon**.

Le test de Student consiste à calculer ce ratio pour les données de l'échantillon et à le comparer à la **valeur** théorique de la **table de Student**.

A. Test de Student sur un échantillon

1. Les hypothèses que nous souhaitons tester sont :

   - **hypothèse nulle** :

   $$H_0 \quad : \quad \mu = \mu_0$$

   Il s'agit de tester si l'échantillon provient bien d'une population avec la **moyenne** spécifiée, μ_0, ou s'il y a une différence significative entre la moyenne de l'échantillon et la moyenne présumée de la population,

   - **hypothèse alternative** :

   L'hypothèse alternative correspondant à l'hypothèse nulle H_0 peut prendre trois formes :

   $$\begin{array}{lcl} H_1 & : & \mu > \mu_0 \text{ (\textbf{test unilatéral} à droite)}, \\ H_1 & : & \mu < \mu_0 \text{ (test unilatéral à gauche)}, \\ H_1 & : & \mu \neq \mu_0 \text{ (\textbf{test bilatéral})}. \end{array}$$

2. Calcul du ratio pour les données de l'échantillon :

   $$t = \frac{\bar{x} - \mu_0}{\frac{S}{\sqrt{n}}}$$

   où

   - μ_0 est la moyenne de la population spécifiée par H_0,
   - $\bar{x}$ est la moyenne de l'échantillon,
   - S est l'écart type de l'échantillon et
   - n la taille d'échantillon.

3. Choix du **seuil de signification** α du test.

4. Comparaison de la valeur calculée de t avec la **valeur critique** appropriée de t_{n-1} la valeur de la table de Student avec $n-1$ degrés de liberté. Nous rejetons H_0 si la valeur absolue de t est supérieure à cette valeur critique.

   Les valeurs critiques pour différents degrés de

liberté et différents seuils de signification sont données par la table de Student.

Pour un test unilatéral, nous prenons la valeur $t_{n-1,1-\alpha}$ de la table et pour un test bilatéral, nous prenons la valeur $t_{n-1,1-\frac{\alpha}{2}}$.

B. Test de Student sur deux échantillons

1. Les hypothèses que nous souhaitons tester sont :

   - hypothèse nulle :

   $$H_0 \quad : \quad \mu_1 = \mu_2.$$

   Dans un test d'hypothèse ayant pour but la comparaison de deux échantillons, nous désirons savoir s'il existe une différence significative entre la moyenne des deux populations dont sont tirés les échantillons.

   - hypothèse alternative :

   Elle peut prendre les trois formes suivantes :

   $$\begin{array}{lcl} H_1 & : & \mu_1 > \mu_2 \text{ (test unilatéral à droite)}, \\ H_1 & : & \mu_1 < \mu_2 \text{ (test unilatéral à gauche)}, \\ H_1 & : & \mu_1 \neq \mu_2 \text{ (test bilatéral)}. \end{array}$$

2. Calcul du ratio de l'échantillon :

   $$t = \frac{\bar{x}_1 - \bar{x}_2}{S_p \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

   où n_1 et n_2 sont les tailles respectives des deux échantillons et S_p est l'écart type pondéré des deux échantillons :

   $$\begin{aligned} S_p &= \sqrt{\frac{(n_1 - 1)\, S_1 + (n_2 - 1)\, S_2}{n_1 + n_2 - 2}} \\ &= \sqrt{\frac{\sum_{i=1}^{2} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^{\,2}}{n_1 + n_2 - 2}} \end{aligned}$$

   où S_1 représente l'écart type de l'échantillon 1, S_2 celui de l'échantillon 2 et x_{ij} l'**observation** j de l'échantillon i.

3. Choix du seuil de signification α du test.

4. Comparaison de la valeur calculée de t avec la valeur critique appropriée de $t_{n_1+n_2-2,1-\frac{\alpha}{2}}$ la valeur de la table de Student avec $(n_1 + n_2 - 2)$ degrés de liberté. Nous rejetons H_0 si la valeur absolue de t est supérieure à cette valeur critique.

   Si le test est unilatéral, nous prenons la valeur $t_{n_1+n_2-2,1-\alpha}$ de la table de Student et s'il est bilatéral, la valeur $t_{n_1+n_2-2,1-\frac{\alpha}{2}}$.

Test de la signification d'un coefficient de régression

Le test de Student peut aussi être appliqué dans des problèmes de régression. Considérons un **modèle** linéaire de régression simple pour lequel les données consistent en n paires d'observations (x_i, y_i), $i = 1, \ldots, n$, modélisées par :

$$y_i = \beta_0 + \beta_1 \cdot x_i + \varepsilon_i$$

où les ε_i sont des erreurs aléatoires indépendantes et distribuées selon une loi normale de moyenne 0 et de **variance** σ^2.

Les estimateurs de β_0 et β_1 sont donnés par :

$$\begin{aligned} \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \cdot \bar{x} \\ \hat{\beta}_1 &= \frac{\sum_{i=1}^{n} (x_i - \bar{x})\,(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^{\,2}}. \end{aligned}$$

L'estimateur $\hat{\beta}_1$ est distribué selon une loi normale de moyenne β_1 et de variance

$$Var(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^{n} (x_i - \bar{x})^{\,2}}.$$

Si σ^2 est estimé par la moyenne des carrés des résidus S^2 et est remplacé dans l'expression de la variance de $\hat{\beta}_1$, alors le ratio

$$T = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\widehat{Var}\left(\hat{\beta}_1\right)}}$$

est distribué selon une loi de Student avec $n-2$ degrés de liberté.

À un certain seuil de signification α, nous pouvons donc effectuer un test de Student du coefficient β_1. Les hypothèses sont les suivantes :

hypothèse nulle	H_0	:	$\beta_1 = 0$,
hypothèse alternative	H_1	:	$\beta_1 \neq 0$.

Si la valeur absolue du ratio T calculé est supérieure à la valeur de t dans la table de Student $t_{n-2,1-\frac{\alpha}{2}}$, nous pourrons conclure que le coefficient β_1 est significativement différent de zéro. Dans le cas contraire, nous ne pouvons pas rejeter l'hypothèse nulle ($\beta_1 = 0$). Nous concluons dans ce cas que la pente de la droite de régression n'est pas significativement différente de zéro.

Un test similaire peut être effectué pour la constante β_0.

L'estimateur $\hat{\beta}_0$ est distribué selon une loi normale de moyenne β_0 et de variance

$$Var(\hat{\beta}_0) = \sigma^2 \cdot \left[\frac{\sum_{i=1}^{n} x_i^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right].$$

Si σ^2 est estimé par la moyenne des carrés des résidus S^2 et est remplacé dans l'expression de $Var\left(\hat{\beta}_0\right)$, alors le ratio

$$T = \frac{\hat{\beta}_0 - \beta_0}{\sqrt{\widehat{Var}\left(\hat{\beta}_0\right)}}$$

est distribué selon une loi de Student avec $n-2$ degrés de liberté.

Le test de Student concerne les hypothèses suivantes :

hypothèse nulle	H_0	:	$\beta_0 = 0$,
hypothèse alternative	H_1	:	$\beta_0 \neq 0$.

Si la valeur absolue du ratio T calculé est supérieure à la valeur de t dans la table de Student $t_{n-2,1-\frac{\alpha}{2}}$, nous pourrons conclure que le coefficient β_0 est significativement différent de zéro. Dans le cas contraire, nous ne pouvons pas rejeter l'hypothèse nulle ($\beta_0 = 0$). Nous concluons donc que la constante de la régression n'est pas significativement différente de zéro.

DOMAINES ET LIMITATIONS

En règle générale, il faut être très prudent dans le choix du test à effectuer. Le test de Student est souvent appliqué à tort. En effet, il n'a de validité que si les observations proviennent d'une **population** distribuée selon une **loi normale** ou selon une distribution suffisamment proche de la loi normale. Si cette exigence n'est pas satisfaite, la **valeur critique** du test de Student n'offre pas de garantie absolue.

Si la loi normale est applicable, le test de Student sur un **échantillon** est utilisé lorsque l'**écart type** de la population est inconnu ou lorsque la taille de l'échantillon n'est pas grande ($n < 30$).

De même, quand il s'agit de deux échantillons, le test de Student est utilisé lorsque les écarts types des deux populations sont inconnus et supposés égaux, ou lorsque la somme des tailles des échantillons est relativement restreinte ($n_1 + n_2 < 30$).

EXEMPLES

Test de Student sur un échantillon

Considérons un **échantillon** de taille $n = 10$ dont les observations provenant d'une distribution normale sont les suivantes :

47 51 48 49 48 52 47 49 46 47

Nous voulons tester les hypothèses :

hypothèse nulle	H_0 :	$\mu = 50$,
hypothèse alternative	H_1 :	$\mu \neq 50$.

La **moyenne arithmétique** des observations est égale à :

$$\bar{x} = \frac{\sum_{i=1}^{10} x_i}{n} = \frac{484}{10} = 48,4.$$

Nous résumons les calculs dans le tableau suivant :

x_i	$(x_i - \bar{x})^2$
47	1,96
51	6,76
48	0,16
49	0,36
48	0,16
52	12,96
47	1,96
49	0,36
46	5,76
47	1,96
Total	32,4

L'**écart type** est égal à :

$$S = \sqrt{\frac{\sum_{i=1}^{10} (x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{32,4}{9}} = 1,90.$$

Nous calculons le ratio :

$$t = \frac{\bar{x} - \mu}{\frac{S}{\sqrt{n}}} = \frac{48,4 - 50}{\frac{1,90}{\sqrt{10}}} = -2,67.$$

Le nombre de degrés de liberté associé au test est égal à $n - 1 = 9$.

Si nous choisissons un **seuil de signification** α de 5 la **valeur** $t_{n-1,1-\frac{\alpha}{2}}$ de la **table de Student** est égale à :

$$t_{9,0,975} = 2,26.$$

Comme $|t| = 2,67 > 2,26$, nous rejetons l'hypothèse nulle et concluons qu'à un seuil de signification de 5 %, l'échantillon ne provient pas d'une population de moyenne $\mu = 50$.

Test de Student sur deux échantillons

Une étude est réalisée en vue de comparer la durée de vie moyenne de deux marques de pneus. Un échantillon a été prélevé pour chacune des marques. Les résultats obtenus sont :

- pour la marque 1 :
  durée de vie moyenne: $\bar{x}_1 = 72\,000$ km
  écart type : $S_1 = 3\,200$ km
  taille de l'échantillon : $n_1 = 50$
- pour la marque 2 :
  durée de vie moyenne : $\bar{x}_2 = 74\,400$ km
  écart type : $S_2 = 2\,400$ km
  taille de l'échantillon : $n_2 = 40$

Nous désirons savoir s'il existe une différence significative de durée de vie entre les deux marques de pneus, à un seuil de signification α de 1 %.

Ceci implique les hypothèses suivantes :

— hypothèse nulle

$$H_0 \quad : \quad \mu_1 = \mu_2, \text{ ou } \mu_1 - \mu_2 = 0$$

— hypothèse alternative

$$H_1 \quad : \quad \mu_1 \neq \mu_2, \text{ ou } \mu_1 - \mu_1 \neq 0.$$

L'écart type pondéré S_p est égal à :

$$\begin{aligned} S_p &= \sqrt{\frac{(n_1 - 1) \cdot S_1^2 + (n_2 - 1) \cdot S_2^2}{n_1 + n_2 - 2}} \\ &= \sqrt{\frac{49 \cdot 3\,200^2 + 39 \cdot 2\,400^2}{50 + 40 - 2}} \\ &= 2\,873,07. \end{aligned}$$

Nous pouvons donc calculer le ratio t :

$$\begin{aligned} t &= \frac{\bar{x}_1 - \bar{x}_2}{S_p \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \\ &= \frac{72\,000 - 74\,400}{2\,873,07 \cdot \sqrt{\frac{1}{50} + \frac{1}{40}}} \\ &= -3,938. \end{aligned}$$

Le nombre de degrés de liberté associé au test est égal à :

$$v = n_1 + n_2 - 2 = 88.$$

La valeur de $t_{v,1-\frac{\alpha}{2}}$ trouvée dans la table de Student, pour $\alpha = 1\%$ est :

$$t_{88,0,995} = 2,634.$$

Comme la valeur absolue du ratio calculé est supérieure à cette valeur, $|t| = |-3,938| > 2,634$, nous rejetons l'hypothèse nulle au profit de

l'hypothèse alternative, et concluons qu'à un seuil de signification de 1 %, les deux marques de pneus n'ont pas la même durée de vie moyenne.

MOTS CLÉS

Loi de Student *(Student distribution)*
Table de Student *(Student table)*
Test d'hypothèse *(Hypothesis testing)*
Test paramétrique *(Parametric test)*

RÉFÉRENCES

Fisher, R.A. (1925). Applications of "Student's" Distribution. *Metron*, 5, pp. 90-104.

Gosset, S.W. « **Student** » (1908). The Probable Error of a Mean. *Biometrika*, 6, pp. 1-25.

Gosset, S.W. « **Student** » (1925). New tables for testing the significance of observations. *Metron*, 5, pp. 105-120.

TEST DE STUDENT POUR OBSERVATIONS PAIRÉES

Matched pairs Student test

Le test de Student pour observations pairées est un test d'hypothèse qui sert à comparer les moyennes de deux populations, dont chaque élément de l'une des populations est mis en relation avec un élément de l'autre.

Par exemple, il peut s'agir de comparer deux traitements, les données étant considérées comme des paires d'observations (première observation de la paire recevant le traitement 1 et deuxième observation recevant le traitement 2).

ASPECTS MATHÉMATIQUES

Soit x_{ij} l'**observation** j pour la paire i ($j = 1, 2$ et $i = 1, 2, \ldots, n$). Pour chaque paire d'observations nous calculons la différence

$$d_i = x_{i2} - x_{i1}.$$

Puis nous cherchons l'**erreur type** estimée de la **moyenne** des d_i, notée $\overline{d}$:

$$s_{\bar{d}} = \frac{1}{\sqrt{n}} \cdot S_d$$

où S_d est l'**écart type** des d_i :

$$S_d = \sqrt{\frac{\sum_{i=1}^{n} \left(d_i - \bar{d}\right)^2}{n-1}}.$$

Le test statistique résultant est défini par :

$$T = \frac{\bar{d}}{s_{\bar{d}}}.$$

Hypothèses

Le test de Student pour observations pairées est un **test bilatéral**. Les hypothèses sont :

$$\begin{array}{lll} H_0 & : & \delta = 0 \text{ (il n'y a pas de différence entre les traitements),} \\ H_1 & : & \delta \neq 0 \text{ (il y a une différence entre les traitements),} \end{array}$$

où δ représente la différence entre les moyennes des deux populations ($\delta = \mu_1 - \mu_2$).

Règles de décision

Nous rejetons l'**hypothèse nulle** au **seuil de signification** α si

$$|T| > t_{n-1, \frac{\alpha}{2}}$$

où $t_{n-1, \frac{\alpha}{2}}$ est la **valeur** de la **table de Student** avec $n-1$ degrés de liberté.

EXEMPLE

Supposons que deux traitements soient appliqués sur 10 paires d'observations. Les données obtenues et les différences correspondantes notées d_i se trouvent dans le tableau suivant :

Paire	Traitement		
i	1	2	$d_i = x_{i2} - x_{i1}$
1	110	118	8
2	99	104	5
3	91	85	−6
4	107	108	1
5	82	81	−1
6	96	93	−3
7	100	102	2
8	87	101	14
9	75	84	9
10	108	111	3

La **moyenne** est égale à :

$$\begin{aligned} \bar{d} &= \frac{1}{10}\sum_{i=1}^{10} d_i = \frac{32}{10} \\ &= 3,2. \end{aligned}$$

L'**écart type** se calcule ainsi :

$$\begin{aligned} S_d &= \sqrt{\frac{\sum_{i=1}^{10}\left(d_i - \bar{d}\right)^2}{10-1}} = \sqrt{\frac{323,6}{9}} \\ &= 6,00 \end{aligned}$$

et l'**erreur type** :

$$\begin{aligned} s_{\bar{d}} &= \frac{1}{\sqrt{n}} \cdot S_d = \frac{1}{\sqrt{10}} \cdot 6,00 \\ &= 1,90. \end{aligned}$$

Le test statistique est donc égal à :

$$T = \frac{\bar{d}}{s_{\bar{d}}} = \frac{3,2}{1,90} = 1,69.$$

Si nous choisissons un **seuil de signification** $\alpha = 0,05$, la **valeur** de $t_{9,0,025}$ est $2,26$. Par conséquent, l'**hypothèse nulle** $H_0 : \delta = 0$ ne peut être rejetée puisque $|T| < t_{9,0,025}$.

MOTS CLÉS

Table de Student *(Student table)*
Test d'hypothèse *(Hypothesis testing)*
Test de Student *(Student test)*

TEST DE WILCOXON

Wilcoxon test

Le test de Wilcoxon est un **test non paramétrique**. Il est utilisé lorsque l'on est en présence de deux échantillons provenant de deux populations. Son but est de vérifier s'il existe des différences entre les deux populations sur la base d'échantillons aléatoires tirés de ces populations.

HISTORIQUE

Le test de Wilcoxon porte le nom de son auteur et a été introduit en 1945.

Voir aussi **test de Mann-Withney**.

ASPECTS MATHÉMATIQUES

Soit $(X_1, X_2, \ldots, X_n)$ un **échantillon** de taille n provenant d'une **population** 1 et soit $(Y_1, Y_2, \ldots, Y_m)$ un échantillon de taille m provenant d'une population 2.

On obtient ainsi $N = n + m$ observations que l'on va classer dans un ordre croissant sans tenir compte de l'appartenance aux échantillons. On attribue ensuite un rang de 1 à la plus petite **valeur**, un rang de 2 à la valeur juste supérieure et ainsi de suite jusqu'au rang N attribué à la plus grande valeur. Si plusieurs observations ont exactement la même valeur, on leur attribuera un rang moyen. On note $R(X_i)$ le rang attribué à X_i, $i = 1, \ldots, n$.

La **statistique** W du test est définie par :

$$W = \sum_{i=1}^{n} R(X_i).$$

Si la taille des échantillons est grande ($N \geq 12$), on utilise, selon Gibbons (1971), une approximation en utilisant la statistique W_1 supposée suivre une **loi normale** standard $\mathcal{N}(0,1)$:

$$W_1 = \frac{W - \mu}{\sigma},$$

où μ et σ sont respectivement la **moyenne** et l'**écart type** de la **variable aléatoire** W :

$$\mu = \frac{n\,(N+1)}{2}$$

et

$$\sigma = \sqrt{\frac{mn\,(N+1)}{12}}.$$

Si l'on utilise l'approximation pour les grands échantillons W_1 et qu'il y a des rangs *ex æquo* parmi les N observations, on remplace l'écart type par :

$$\sigma = \sqrt{\frac{mn}{12}\left(N+1-\frac{\sum_{j=1}^{g} t_j\left(t_j^2-1\right)}{N\,(N-1)}\right)}$$

où g est le nombre de groupes de rangs *ex æquo* et t_j la taille du groupe j. (S'il n'y a pas de rangs *ex æquo*, les observations sont vues comme autant de groupes de taille 1. Par conséquent, $g = N$ et $t_j = 1$ pour $j = 1, \ldots, N$ et σ se réduit à $\sqrt{\frac{mn\,(N+1)}{12}}$.)

Hypothèses

Le test de Wilcoxon peut être effectué sur la base d'un **test bilatéral** ou **unilatéral**, selon le type d'**hypothèse**:

A. Cas bilatéral :

$$\begin{array}{lcl} H_0 & : & P\,(X<Y) = \frac{1}{2}, \\ H_1 & : & P\,(X<Y) \neq \frac{1}{2}. \end{array}$$

B. Cas unilatéral :

$$\begin{array}{lcl} H_0 & : & P\,(X<Y) \leq \frac{1}{2}, \\ H_1 & : & P\,(X<Y) > \frac{1}{2}. \end{array}$$

C. Cas unilatéral :

$$\begin{array}{lcl} H_0 & : & P\,(X<Y) \geq \frac{1}{2}, \\ H_1 & : & P\,(X<Y) < \frac{1}{2}. \end{array}$$

Le cas A exprime l'hypothèse qu'il n'y a pas de différence entre les deux populations.

Le cas B représente l'hypothèse que la population 1 (d'où est tiré l'échantillon des X) admet en général des valeurs plus grandes que celles de la population 2 (d'où est tiré l'échantillon des Y).

Le cas C exprime l'hypothèse contraire, que les valeurs de la population 1 ont tendance à être plus petites que celles de la population 2.

Règles de décision

Cas A

On rejette l'**hypothèse nulle** H_0 au **seuil de signification** α si W est inférieur à la valeur de la **table de Wilcoxon** avec les paramètres n, m et $\frac{\alpha}{2}$ notée $t_{n,m,\alpha/2}$ ou si W est plus grand que la valeur de la table pour n, m et $1-\frac{\alpha}{2}$, notée $t_{n,m,1-\alpha/2}$, c'est-à-dire si

$$W < t_{n,m,\alpha/2} \text{ ou } W > t_{n,m,1-\alpha/2}.$$

Si le test utilise la statistique W_1, la comparaison se fait avec la **table normale** :

$$W_1 < z_{\alpha/2} \text{ ou } W_1 > z_{1-\alpha/2}$$

avec $z_{\alpha/2}$ et $z_{1-\alpha/2}$ les valeurs de la table normale avec respectivement les paramètres $\frac{\alpha}{2}$ et $1-\frac{\alpha}{2}$.

Cas B

On rejette l'hypothèse nulle H_0 au seuil de signification α si W est inférieur à la valeur de la table de Wilcoxon avec les paramètres n, m et α, notée $t_{n,m,\alpha}$, c'est-à-dire si

$$W < t_{n,m,\alpha}$$

et dans le cas où la statistique W_1 est utilisée :

$$W_1 < z_\alpha$$

avec z_α la valeur de la table normale avec le paramètre α.

Cas C

On rejette l'hypothèse nulle H_0 au seuil de signification α si W est supérieur à la valeur de la table de

Wilcoxon avec les paramètres n, m et $1-\alpha$, notée $t_{n,m,1-\alpha}$, c'est-à-dire si

$$W > t_{n,m,1-\alpha}$$

et avec la statistique W_1 si

$$W_1 > z_{1-\alpha}$$

avec $z_{1-\alpha}$ la valeur de la table normale avec le paramètre α.

DOMAINES ET LIMITATIONS

Les exigences préalables à l'utilisation du test de Wilcoxon sont les suivantes :

1. les deux échantillons sont des échantillons aléatoires tirés de leur **population** respective ;
2. en plus de l'**indépendance** à l'intérieur de chaque échantillon, il y a indépendance mutuelle entre les deux échantillons ;
3. l'échelle de mesure est au moins ordinale.

EXEMPLE

Dans une classe, on compte 25 élèves : 14 garçons et 11 filles. On fait passer à cette classe un test de calcul mental pour voir si les garçons ont tendance à être meilleurs que les filles dans cette discipline.

Les données se trouvent dans le tableau ci-dessous, les scores plus élevés correspondant à de meilleurs résultats au test.

Garçons (X_i)		Filles (Y_i)	
19,8	17,5	17,7	23,6
12,3	17,9	7,1	11,1
10,6	21,1	21,0	20,3
11,3	16,4	10,6	15,6
14,0	7,7	13,3	
9,2	15,2	8,6	
15,6	16,0	14,1	

On classe ces scores dans un ordre croissant et on leur attribue un rang :

Scores	Échantillons	Rangs	$R(X_i)$
7,1	Y	1	-
7,7	X	2	2
8,6	Y	3	-
9,2	X	4	4
10,6	Y	5,5	-
10,6	X	5,5	5,5
11,1	Y	7	-
11,3	X	8	8
12,3	X	9	9
13,3	Y	10	-
14,0	X	11	11
14,1	Y	12	-
15,2	X	13	13
15,6	X	14,5	14,5
15,6	Y	14,5	-
16,0	X	16	16
16,4	X	17	17
17,5	X	18	18
17,7	Y	19	-
17,9	X	20	20
19,8	X	21	21
20,3	Y	22	-
21,0	Y	23	-
21,1	X	24	24
23,6	Y	25	-

On calcule la **statistique** W:

$$\begin{aligned} W &= \sum_{i=1}^{14} R(X_i) \\ &= 2+4+5,5+8+9+11+13+14,5 \\ &\quad +16+17+18+20+21+24 \\ &= 183. \end{aligned}$$

Si on choisit l'approximation pour les grands échantillons, en tenant compte des rangs moyens, on effectue l'ajustement :

$$W_1 = \frac{W - \dfrac{n(N+1)}{2}}{\sqrt{\dfrac{mn}{12}\left(m+n+1-\dfrac{\displaystyle\sum_{j=1}^{g} t_j\left(t_j^2-1\right)}{(m+n)(m+n-1)}\right)}}$$

Ici $g = 2$ et $t_j = 2$ pour $j = 1, 2$. On a donc :

$$W_1 = \frac{183 - \frac{14\,(25+1)}{2}}{\sqrt{\frac{11 \cdot 14}{12}\left(11+14+1-\frac{2(2^2-1+2\,(2^2-1)}{(11+14)\,(11+14-1)}\right)}} = 0,0548.$$

Les hypothèses que l'on souhaite tester sont :

H_0 : les garçons n'ont pas tendance à être meilleurs que les filles en calcul mental,
H_1 : les garçons ont tendance à être meilleurs que les filles en calcul mental.

Cela se traduit par H_0 : $P(X < Y) \geq \frac{1}{2}$ et correspond au cas C. La règle de décision est donc la suivante :

rejeter H_0 au **seuil de signification** α si

$$W > t_{n,m,1-\alpha}$$
$$\text{ou } W_1 > z_{1-\alpha}.$$

Si l'on choisit $\alpha = 0,05$, la **valeur** théorique dans la **table de Wilcoxon** est $t_{n,m,1-\alpha} = 212$.

La valeur calculée de W ($W = 183$) n'étant pas supérieure à $t_{n,m,1-\alpha} = 212$ et on ne peut rejeter l'**hypothèse nulle** H_0.

La valeur de la **table normale** $z_{1-\alpha}$ est égale à $1,6449$. Par conséquent, W_1 n'est pas supérieur à $z_{1-\alpha}$ $(0,0548 < 1,6449)$ et on tire la même conclusion sur la base de la statistique W: les garçons n'ont pas tendance à être meilleurs que les filles en calcul mental.

MOTS CLÉS

Table de Wilcoxon *(Wilcoxon table)*
Test d'hypothèse *(Hypothesis testing)*
Test de Mann-Whitney *(Mann-Whitney test)*
Test non paramétrique *(Nonparametric test)*
Test de Wilcoxon signé *(Wilcoxon signed test)*

RÉFÉRENCES

Gibbons, J.D. (1971). *Nonparametric Statistical Inference.* McGraw-Hill, New York.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics*, 1, pp. 80-83.

TEST DE WILCOXON SIGNÉ

Wilcoxon signed test

Le test de Wilcoxon signé est un **test non paramétrique**. Comme le **test du signe**, il sert à tester l'existence d'une différence entre deux populations.

Il permet de savoir si la différence entre deux paires d'**observations** est positive ou négative, tout en tenant compte de l'amplitude de cette différence.

HISTORIQUE

Le test de Wilcoxon signé est dû à F. Wilcoxon en 1945. En 1949, il propose encore une méthode pour traiter les différences nulles pouvant apparaître entre deux observations des échantillons considérés.

ASPECTS MATHÉMATIQUES

Soient $(X_1, X_2, \ldots, X_n)$ et $(Y_1, Y_2, \ldots, Y_n)$ deux échantillons de taille n. On considère les n paires d'observations (x_1, y_1), (x_2, y_2), …, (x_n, y_n). On désigne par $|d_i|$ la différence en valeur absolue entre x_i et y_i :

$$|d_i| = |y_i - x_i| \quad i = 1, 2, \ldots, n.$$

On soustrait de la taille de l'échantillon toutes les paires qui ne présentent pas de différence ($d_i = 0$). On note par m le nombre de paires qui restent. On attribue ensuite aux paires d'observations des rangs allant de 1 à m en fonction de la grandeur des $|d_i|$, c'est-à-dire que l'on donne le rang 1 à la plus petite valeur de $|d_i|$ et le rang m à la plus grande valeur de $|d_i|$. On nomme R_i, $i = 1, \ldots, m$, le rang ainsi défini.

Si plusieurs paires présentent la même différence absolue $|d_i|$, on leur attribue un rang moyen.

S'il y a des rangs moyens, la **statistique** du test

est calculée par l'expression suivante :

$$T_1 = \frac{\sum_{i=1}^{m} R_i}{\sqrt{\sum_{i=1}^{m} R_i^2}}$$

où R_i est le rang de la paire (x_i, y_i) pourvu du signe de la différence d_i.

S'il n'y a pas de rangs moyens, il est plus correct d'utiliser la somme des rangs associés à des différences positives :

$$T = \sum R_i$$
(pour i tel que $d_i > 0$).

Hypothèses

Les hypothèses sous-jacentes au test de Wilcoxon signé sont formulées sur la base de d_k, la **médiane** des d_i. Suivant qu'il s'agit d'un **test bilatéral** ou d'un **test unilatéral**, on a les cas suivants :

A. Cas bilatéral :

$$\begin{array}{lcl} H_0 & : & d_k = 0 \\ H_1 & : & d_k \neq 0 \end{array}$$

B. Cas unilatéral :

$$\begin{array}{lcl} H_0 & : & d_k \leq 0 \\ H_1 & : & d_k > 0 \end{array}$$

C. Cas unilatéral :

$$\begin{array}{lcl} H_0 & : & d_k \geq 0 \\ H_1 & : & d_k < 0. \end{array}$$

Dans le cas A, l'hypothèse représente la situation d'absence de différence entre les deux populations.

Dans le cas B, on fait l'hypothèse que les valeurs de la population des X ont tendance à être plus grandes que celles de la population des Y.

Dans le cas C, on fait l'**hypothèse** contraire, à savoir que les valeurs de la **population** des X ont tendance à être plus petites que celles de la population des Y.

Règles de décision

Si T est utilisé et si $m < 50$, on se base sur la **table de Wilcoxon signé** pour tester l'**hypothèse nulle** H_0.

La **variable aléatoire** T est une combinaison linéaire de m variables aléatoires de Bernoulli indépendantes (mais pas identiquement distribuées). Elle a pour **moyenne** et **écart type**

$$\begin{array}{rcl} \mu & = & \dfrac{m\,(m+1)}{4} \quad \text{et} \\ \sigma & = & \sqrt{\dfrac{m\,(m+1)\,(2m+1)}{24}}. \end{array}$$

On en déduit la **variable aléatoire** Z :

$$Z = \frac{T - \mu}{\sigma}.$$

On note par w_α la valeur de la table de Wilcoxon signé obtenue avec les paramètres m et α (où α est le **seuil de signification**).

Si $m > 20$, w_α peut être approximé à l'aide de la valeur correspondante de la **loi normale** :

$$w_\alpha = \frac{m\,(m+1)}{4} + z_\alpha \sqrt{\frac{m\,(m+1)\,(2m+1)}{24}}$$

où z_α est la valeur de la **table normale** au seuil α.

Si T_1 est utilisé, on se base directement sur la table normale pour tester l'hypothèse nulle H_0. Les règles de décision sont différentes selon les hypothèses posées. C'est ainsi que l'on a les règles *A*, *B* et *C* relatives aux cas précédents *A*, *B* et *C*. Si T_1 est utilisé, il faut remplacer T par T_1 et w_α par z_α dans les règles qui suivent.

Cas A
On rejette H_0 au seuil de signification α si T est supérieur à $w_{1-\alpha/2}$ ou inférieur à $w_{\alpha/2}$, c'est-à-dire si

$$T < w_{\alpha/2} \text{ ou } T > w_{1-\alpha/2}.$$

Cas B
On rejette H_0 au seuil de signification α si T est supérieur à $w_{1-\alpha}$, c'est-à-dire si

$$T > w_{1-\alpha}.$$

Cas C
On rejette H_0 au **seuil de signification** α si T est inférieur à w_α, c'est-à-dire si

$$T < w_\alpha.$$

DOMAINES ET LIMITATIONS

Les exigences préalables d'utilisation du test de Wilcoxon signé sont les suivantes :

1. la distribution des d_i est symétrique ;
2. les d_i sont indépendants ;
3. les d_i se mesurent en valeur réelle.

Le test de Wilcoxon signé peut aussi être utilisé comme test de **médiane**. Les données $X_1, X_2, \ldots, X_n$ consistent alors en un seul **échantillon** de taille n.

Les hypothèses correspondant aux cas précédents *A*, *B* et *C* sont les suivantes :

A. Cas bilatéral :

H_0 : la médiane des X est égale à une **valeur** présumée M_d,
H_1 : la médiane des X n'est pas égale à M_d.

B. Cas unilatéral :

H_0 : la médiane des X est $\geq M_d$,
H_1 : la médiane des X est $< M_d$.

C. Cas unilatéral :

H_0 : la médiane des X est $\leq M_d$,
H_1 : la médiane des X est $> M_d$.

Il faut noter que le mot médiane pourrait être remplacé par le mot **moyenne** puisque l'on a admis l'hypothèse de symétrie de la distribution des X.

En ce qui concerne les calculs, on forme les n paires (X_1, M_d), (X_2, M_d), …, (X_n, M_d) et on les traite de la même façon que précédemment. Le reste du test est identique.

EXEMPLE

On fait passer des tests à 8 couples de vrais jumeaux. Le but est de voir si celui qui est né le premier tend à être plus intelligent que le second.

Les données se trouvent dans le tableau ci-dessous, les scores plus élevés correspondant à de meilleurs résultats aux tests.

Couple de jumeaux	Né le premier X_i	Né en second Y_i	d_i	Rang des $\lvert d_i \rvert$	R_i
1	90	88	−2	2	−2
2	75	79	4	4	4
3	99	98	−1	1	−1
4	60	66	6	5	5
5	72	64	−8	6	−6
6	83	83	0	-	-
7	86	86	0	-	-
8	92	95	3	3	3

Comme les couples 6 et 7 ne présentent pas de différence ($d_6 = d_7 = 0$), le nombre de couples qui restent est égal à $m = 6$. Les d_i étant tous différents, des rangs moyens ne sont pas à calculer et on utilisera la somme des rangs positifs :

$$T = \sum R_i \quad (\text{pour } i \text{ tel que } d_i > 0)$$

où R_i est le rang du couple (x_i, y_i) présentant un d_i positif. On a alors :

$$T = 4 + 5 + 3 = 12.$$

Les hypothèses sont les suivantes :

H_0 : il n'y a pas de différence de niveau d'intelligence entre le premier et le second jumeau ($d_m = 0$),
H_1 : il y a une différence ($d_m \neq 0$).

Ces hypothèses correspondent au cas *A* et la règle de décision est la suivante :
rejeter H_0 au seuil α si $T < w_{\alpha/2}$ ou $T > w_{1-\alpha/2}$.

Si l'on choisit le **seuil de signification** $\alpha = 0,05$, on obtient de la **table de Wilcoxon signé** pour $m = 6$:

$$w_{1-\alpha/2} = 20 \text{ et } w_{\alpha/2} = 1.$$

T étant ni supérieur à 20 ni inférieur à 1, on ne peut pas rejeter l'**hypothèse nulle** H_0.

On suppose maintenant que les résultats pour les couples 7 et 8 soient changés avec le nouveau tableau suivant :

Couple de jumeaux	Né le premier X_i	Né en second Y_i	d_i	Rang des $\lvert d_i\rvert$	R_i
1	90	88	-2	4	-4
2	75	79	4	5	5
3	99	98	-1	2	-2
4	60	66	6	6	6
5	72	64	-8	7	-7
6	83	83	0	-	-
7	86	87	1	2	2
8	92	91	-1	2	-2

Cette fois, seul le couple 6 ne présente pas de différence ($d_6 = 0$) ; m est donc égal à 7.

Les couples 3, 7 et 8 présentent la même différence absolue : $|d_3| = |d_7| = |d_8| = 1$. On attribue donc un rang moyen à ces couples qui est égal au rang des *ex æquo* auquel on ajoute la moitié du nombre d'*ex æquo* diminué de un. Ainsi :

$$\begin{aligned} \text{rang moyen} &= \text{rang de } |d_3| + \frac{1}{2}(3-1) \\ &= 2. \end{aligned}$$

Dans ce cas, on utilise le test statistique T_1 calculé par :

$$\begin{aligned} T_1 &= \frac{\sum_{i=1}^{m} R_i}{\sqrt{\sum_{i=1}^{m} R_i^2}} \\ &= \frac{-2}{11,75} = -0,17. \end{aligned}$$

La règle de décision est la suivante :
rejeter H_0 au seuil α si $T_1 < z_{\alpha/2}$ ou $T_1 > z_{1-\alpha/2}$.

Si l'on choisit le seuil $\alpha = 0,05$, les valeurs de la **table normale** sont :

$$z_{1-\alpha/2} = -1,96 \quad \text{et} \quad z_{\alpha/2} = 1,96.$$

Par conséquent, T_1 n'est ni supérieur à $1,96$ ni inférieur à $-1,96$.

On obtient le même résultat qu'au cas précédent et on ne rejette pas l'hypothèse nulle H_0.

MOTS CLÉS

Table de Wilcoxon signé *(Wilcoxon signed table)*
Test d'hypothèse *(Hypothesis testing)*
Test non paramétrique *(Nonparametric test)*

RÉFÉRENCES

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics*, 1, pp. 80-83. (5.1, 5.7).

Wilcoxon, F. (1957). *Some rapid approximate statistical procedures.* American Cyanamid Co., Stamford Research Laboratories, Stamford, Conn.

TEST DU CHI-CARRÉ

Chi-square test

Les tests du chi-carré sont des tests d'hypothèse faisant appel à la **loi du chi-carré** et utilisés sous formes diverses ; en voici, les plus courantes :

- Le **test d'indépendance du chi-carré** est utilisé en vue de déterminer si deux variables qualitatives catégorielles étudiées sur un **échantillon** sont indépendantes.

- Le **test d'adéquation du chi-carré** est utilisé pour déterminer si la distribution observée sur un échantillon peut être approximée par une distribution théorique. On veut par exemple savoir si la distribution observée sur un échantillon correspond à une certaine **loi de probabilité** (**loi normale**, **loi de Poisson**, etc.).

- Le test du chi-carré pour une **variance** inconnue mais dont on aimerait tester si elle prend une certaine **valeur** constante fixée.

- Le test du chi-carré pour tester l'homogénéité de la variance calculée sur plusieurs échantillons issus d'une **population** distribuée selon une loi normale.

- Le test du chi-carré pour tester si une répartition en classes donnée correspond à un ensemble de proportions connues.

HISTORIQUE

En 1937, M.S. Bartlett propose une méthode pour tester l'homogénéité de la **variance** calculée sur plusieurs échantillons issus d'une **population** distribuée selon une **loi normale**.

Voir aussi **test d'indépendance du chi-carré** et **test d'adéquation du chi-carré**.

ASPECTS MATHÉMATIQUES

Pour l'aspect mathématique du **test d'indépendance du chi-carré** et celui du **test d'adéquation du chi-carré**, voir les mots respectifs.

Le test du chi-carré pour une **variance** inconnue mais dont on aimerait tester si elle prend une certaine **valeur** constante fixée s'énonce de la façon suivante :

Soit $(x_1, \ldots, x_n)$ un **échantillon** aléatoire provenant d'une **population** distribuée selon une **loi normale** de **moyenne** μ inconnue et de **variance** σ^2 inconnue.

On a de bonnes raisons de croire que la variance de la population est égale à une valeur présumée σ_0^2. Les hypothèses sont exprimées ainsi suivant le cas :

A. Cas bilatéral :

$$\begin{array}{rcl} H_0 & : & \sigma^2 = \sigma_0^2, \\ H_1 & : & \sigma^2 \neq \sigma_0^2. \end{array}$$

B. Cas unilatéral :

$$\begin{array}{rcl} H_0 & : & \sigma^2 = \sigma_0^2, \\ H_1 & : & \sigma^2 > \sigma_0^2. \end{array}$$

C. Cas unilatéral :

$$\begin{array}{rcl} H_0 & : & \sigma^2 = \sigma_0^2, \\ H_1 & : & \sigma^2 < \sigma_0^2. \end{array}$$

On détermine alors la **statistique** du test du chi-carré donnée par :

$$\chi^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{\sigma_0^2}.$$

Cette statistique est distribuée selon une **loi du chi-carré** avec $n-1$ degrés de liberté. Autrement dit, on cherche la valeur de $\chi^2_{n-1,\alpha}$ dans la **table du chi-carré** que l'on compare à la valeur calculée χ^2.

Selon les cas, les règles de décisions sont les suivantes :

Cas A :

si $\chi^2 \geq \chi^2_{n-1,\alpha_1}$ ou si $\chi^2 \leq \chi^2_{n-1,1-\alpha_2}$ on rejette l'**hypothèse nulle** H_0 au profit de l'**hypothèse alternative** H_1, ayant décomposé le **seuil de signification** α en α_1 et α_2 tels que $\alpha_1 + \alpha_2 = \alpha$. Sinon on ne rejette pas l'hypothèse nulle H_0.

Cas B :

si $\chi^2 < \chi^2_{n-1,\alpha}$ on ne rejette pas l'hypothèse nulle H_0. Si $\chi^2 \geq \chi^2_{n-1,\alpha}$ on rejette l'hypothèse nulle H_0 au profit de l'hypothèse alternative H_1.

Cas C :

si $\chi^2 > \chi^2_{n-1,1-\alpha}$ on ne rejette pas l'hypothèse nulle H_0. Si $\chi^2 \leq \chi^2_{n-1,1-\alpha}$ on rejette l'hypothèse nulle H_0 au profit de l'hypothèse alternative H_1.

D'autres tests du chi-carré sont proposés dans l'ouvrage de B. Ostle (1963).

DOMAINES ET LIMITATIONS

Le calcul de la **statistique** χ^2 (chi-carré) doit être basé sur les fréquences absolues et non relatives.

Il est à noter que le test du chi-carré peut ne pas être fiable pour des échantillons de petite taille, et notamment lorsque certaines fréquences estimées sont petites (< 5). Un tel cas peut toutefois être résolu en regroupant plusieurs catégories en une seule, pour autant que ce regroupement ait un sens.

EXEMPLE

On considère un ensemble de pièces fabriquées par une certaine machine. Elles sont classées en fonction de leur diamètre (en mm) dans le tableau suivant :

Diamètre (mm) x_i	Nombre n_i
59,5	2
59,6	6
59,7	7
59,8	15
59,9	17
60,0	25
60,1	15
60,2	7
60,3	3
60,4	1
60,5	2
Total :	$N = 100$

Il s'agit d'un **échantillon** aléatoire issu d'une **population** distribuée selon une **loi normale** de **moyenne** et de **variance** inconnues. Cependant, il vaudrait mieux pour le vendeur des pièces que la variance σ^2 soit inférieure ou égale à 0,05. On effectue alors le **test d'hypothèse** suivant :

hypothèse nulle $H_0 : \sigma^2 \leq 0,05,$
hypothèse alternative $H_1 : \sigma^2 > 0,05.$

Il s'agit d'un test d'hypothèse unilatéral qui se traite comme si on avait une égalité comme hypothèse nulle.

On commence par calculer la moyenne sur l'échantillon :

$$\bar{x} = \frac{\sum_{i=1}^{11} n_i \cdot x_i}{N} = \frac{5\,995}{100} = 59,95.$$

On peut alors calculer la **statistique** χ^2 :

$$\begin{aligned} \chi^2 &= \frac{\sum_{i=1}^{11} n_i \cdot (x_i - \bar{x})^2}{\sigma_0^2} \\ &= \frac{2 \cdot (-0,45)^2 + \ldots + 2 \cdot (0,55)^2}{0,05} \\ \chi^2 &= \frac{3,97}{0,05} = 79,4. \end{aligned}$$

Pour un **seuil de signification** de $\alpha = 5$ %, on trouve la **valeur** de $\chi^2_{99,0,05} = 123,2$ dans la **table du chi-carré**. Comme $\chi^2 = 79,4 < \chi^2_{99,0,05}$, on ne rejette pas l'hypothèse nulle, ce qui veut dire que le commerçant peut vendre ses pièces sachant qu'elles ne comportent pas une trop forte différence de diamètre.

MOTS CLÉS

Loi du chi-carré *(Chi-square distribution)*
Table du chi-carré *(Chi-square table)*
Test d'adéquation du chi-carré *(Chi-square goodness of fit test)*
Test d'indépendance du chi-carré *(Chi-square test of independence)*

RÉFÉRENCES

Bartlett, M.S. (1937). Some Examples of Statistical Methods of Research in Agriculture and Applied Biology. *Journal of the Royal Statistical Society*, (Suppl.), 4, pp. 137-183.

Ostle, B. (1954). *Statistics in Research ; Basic concepts and techniques for research workers.* Iowa State College Press, Ames.

TEST DU RAPPORT DE VRAISEMBLANCE

Likelihood ratio test

Le test du rapport de vraisemblance est un **test d'hypothèse**. Il permet de tester des hypothèses générales concernant les paramètres d'intérêt d'une loi (paramétrique) sous-jacente aux données ou bien de tester deux modèles différents concernant un jeu de données. L'idée du test est la suivante : calculer la **probabilité** d'observer les données sous l'hypothèse H_0 à tester et sous l'**hypothèse alternative** en utilisant la fonction de vraisemblance.

HISTORIQUE

C'est à **Jerzy Neyman** et **Egon Sharpe Pearson** (1928) que revient l'idée d'utiliser la **statistique** du rapport des fonctions de vraisemblance pour tester des hypothèses. Abraham Wald (1941) a généralisé le

test du rapport de vraisemblance à des hypothèses composites. Le résultat sur la distribution asymptotique de cette statistique sous certaines conditions de régularité est dû à Samuel Stanley Wilks (1962).

ASPECTS MATHÉMATIQUES

Dans le test du rapport de vraisemblance, l'**hypothèse nulle** est rejetée si la vraisemblance sous l'**hypothèse alternative** est significativement supérieure à la vraisemblance sous l'hypothèse nulle. Le problème n'est donc pas de trouver quel paramètre est le plus adéquat, mais de savoir si l'hypothèse alternative offre une **estimation** significativement meilleure ou pas (il va de soi que les deux hypothèses ne sont pas traitées symétriquement).

Le théorème de Neyman-Pearson stipule que, pour tout **seuil de signification** α, le test du rapport de vraisemblance (pour tester des hypothèses simples) est plus puissant que tout autre test.

Soient $x_1, x_2, \ldots, x_n$ n observations indépendantes ($n > 1$) d'une **variable aléatoire** X de **loi de probabilité** dépendant d'un **paramètre** θ (ou un vecteur de paramètres $\boldsymbol{\theta}$).

Notons $L(\theta\,;\, x_1, x_2, \ldots, x_n)$ la fonction de vraisemblance (voir **maximum de vraisemblance**), qui, pour x_i fixés, dépend seulement de θ, et Θ l'espace des valeurs possibles de θ et Θ_0, Θ_1 des sous-espaces de Θ.

Distinguons plusieurs cas pour les hypothèses :

Hypothèse nulle simple contre une hypothèse alternative simple

$$\begin{array}{rcl} H_0 & : & \theta = \theta_0 \\ \text{contre } H_1 & : & \theta = \theta_1. \end{array}$$

Dans ce cas, le rapport de vraisemblance est :

$$\lambda = \lambda(x_1, x_2, \ldots, x_n) = \frac{L(\theta_1\,;\, x_1, x_2, \ldots, x_n)}{L(\theta_0\,;\, x_1, x_2, \ldots, x_n)}.$$

Le test du rapport de vraisemblance de **seuil de signification** α est défini par la règle de décision :

$$\text{Rejet de } H_0 \Leftrightarrow \quad \lambda > \lambda_0$$

où λ_0 est la valeur critique pour le seuil de signification α de la loi que suit λ :

$$\alpha = P\left(\lambda(x_1, x_2, \ldots, x_n) \leq \lambda_0 | H_0\right).$$

Le problème majeur qui se pose est celui de trouver la **région de rejet** ; à ce propos il faut se référer aux exemples qui suivront, car il n'y a pas de procédure générale applicable. L'idée est de chercher une statistique, de loi connue, à laquelle « réduire » le rapport de vraisemblance.

Hypothèse nulle simple contre une hypothèse alternative générale

$$\begin{array}{rcl} H_0 & : & \theta = \theta_0 \\ \text{contre } H_1 & : & \theta \in \Theta_1 \text{ (par exemple } \theta < \theta_0) \end{array}$$

Nous posons alors :

$$\lambda = \frac{\sup_{\theta \in \Theta_1} L(\theta_1\,;\, x_1, x_2, \ldots, x_n)}{L(\theta_0\,;\, x_1, x_2, \ldots, x_n)}$$

où le $\sup_{\theta \in \Theta_0} L(\theta\,;\, x_1, x_2, \ldots, x_n)$ exprime la valeur de la fonction de vraisemblance pour le paramètre qui maximise la fonction de vraisemblance sur l'espace restreint des paramètres et cherchons à nouveau la distribution de λ sous H_0.

Remarque

Si l'hypothèse alternative H_1 est $\theta \neq \theta_0$, nous pouvons exprimer λ comme suit :

$$\frac{L\left(\hat{\theta}\,;\, x_1, x_2, \ldots, x_n\right)}{L(\theta_0\ ;\ x_1, x_2, \ldots, x_n)}$$

où $\hat{\theta}$ est l'estimateur du maximum de vraisemblance de θ.

Hypothèse generale $H_0 : \theta \in \Theta_0$ contre l'hypothèse alternative $H_1 : \theta \in \Theta \backslash \Theta_0 = \{\theta \in \Theta; \theta \notin \Theta_0\}$

Nous considérons la **statistique** suivante :

$$\lambda = \frac{\sup_{\theta \in \Theta_0} L(\theta\,;\, x_1, x_2, \ldots, x_n)}{\sup_{\theta \in \Theta} L(\theta\,;\, x_1, x_2, \ldots, x_n)}$$

et cherchons la région de rejet analoguement aux autres cas.

Le $\sup_{\theta\in\Theta_0} L(\theta\,;x_1,x_2,\ldots,x_n)$ exprime la valeur de la fonction de vraisemblance pour le paramètre qui maximise la fonction de vraisemblance (donc pour l'**estimateur** du maximum de vraisemblance) sur l'espace restreint des paramètres. Le $\sup_{\theta\in\Theta} L(\theta\,;x_1,x_2,\ldots,x_n)$ exprime la même chose mais pour l'espace total des paramètres. Plus la différence entre les deux valeurs sera grande (et donc aussi la valeur donnée comme statistique), plus le modèle restreint ne sera pas approprié (et H_0 sera donc rejetée).

Sous H_0, la distribution de $-2\log(\lambda)$ suit asymptotiquement une **loi du chi-carré** avec u degrés de liberté, où $u = \dim\Theta - \dim\Theta_0$ (c'est-à-dire le nombre de paramètres spécifiés par H_0). Autrement dit :

$$-2\lim_{n\to\infty}\log\left(\frac{\sup_{\theta\in\Theta} L(\theta\,;x_1,x_2,\ldots,x_n)}{\sup_{\theta\in\Theta_0} L(\theta\,;x_1,x_2,\ldots,x_n)}\right)=\chi^2(u)\,.$$

Dans les cas particuliers que nous avons traités auparavant, ce résultat s'exprime par :

$-2\log\lambda\sim\chi^2$ à un degré de liberté si $n\to\infty$.

DOMAINES ET LIMITATIONS

Si la distribution sous-jacente aux données est normale, le test du rapport de vraisemblance est équivalent au **test de Fisher.**

L'avantage de ce test (quand il peut être construit explicitement) est qu'il garantit la meilleure puissance (d'après le théorème de Neyman-Pearson, voir Hogg and Craig, 1970, pp. 274-280) pour tester des hypothèses simples.

EXEMPLES

1. Soit X une variable aléatoire suivant une loi normale de moyenne μ et de variance σ^2. Nous désirons tester les hypothèses suivantes :

$$\begin{array}{lcl} H_0 & : & \mu=0,\sigma^2>0,\\ H_1 & : & \mu\neq 0,\sigma^2>0.\end{array}$$

Si nous considérons l'espace $\Theta = \left\{(\mu,\sigma^2)\,;\ -\infty<\mu<\infty, 0<\sigma^2<\infty\right\}$ et un sous-espace $\Theta_1 = \left\{(\mu,\sigma^2)\in\Theta\,;\,\mu=0\right\}$, les hypothèses peuvent alors s'écrire :

$$\begin{array}{lcl} H_0 & : & \theta=(\mu,\sigma^2)\in\Theta_1,\\ H_1 & : & \theta=(\mu,\sigma^2)\in\Theta\backslash\Theta_1.\end{array}$$

Considérons ensuite x_1, …, x_n n observations de la variable X ($n>1$). La fonction de vraisemblance est (voir exemple sous **maximum de vraisemblance**) :

$$\begin{aligned} L(\theta\,;x_1,\ldots,x_n) &= L\left(\mu,\sigma^2\,;x_1,\ldots,x_n\right)\\ &= \left(\frac{1}{2\pi\sigma^2}\right)^{n/2}\cdot\exp\left(-\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2\sigma^2}\right).\end{aligned}$$

À ce stade, nous estimons la moyenne et la variance empiriques de l'échantillon par les estimateurs du maximum de vraisemblance (voir exemple sous maximum de vraisemblance) :

$$\begin{aligned}\widehat{\mu} &= \frac{1}{n}\cdot\sum_{i=1}^{n}x_i=\overline{x}\\ \widehat{\sigma}^2 &= \frac{1}{n}\cdot\sum_{i=1}^{n}(x_i-\mu)^2\,.\end{aligned}$$

Ainsi, les fonctions de vraisemblance sont, pour l'hypothèse nulle :

$$\begin{aligned} &L\left(0,\widehat{\sigma}^2\,;x_1,\ldots,x_n\right)\\ &= \left(\frac{n}{2\pi\sum_{i=1}^{n}x_i^2}\right)^{n/2}\cdot\exp\left(-\frac{n\sum_{i=1}^{n}x_i^2}{2\sum_{i=1}^{n}x_i^2}\right)\\ &= \left(\frac{n}{2\pi\sum_{i=1}^{n}x_i^2}\right)^{n/2}\exp\left(-\frac{n}{2}\right)\\ &= \left(\frac{n\exp(-1)}{2\pi\sum_{i=1}^{n}x_i^2}\right)^{n/2}\end{aligned}$$

et pour l'hypothèse alternative :

$L\left(\widehat{\mu},\widehat{\sigma}^2\,;x_1,\ldots,x_n\right)$

$$
\begin{aligned}
&= \left(\frac{n}{2\pi \sum_{i=1}^{n} (x_i - \overline{x})^2} \right)^{n/2} \\
&\quad \cdot \exp\left(- \frac{n \sum_{i=1}^{n} (x_i - \overline{x})^2}{2 \sum_{i=1}^{n} (x_i - \overline{x})^2} \right) \\
&= \left(\frac{1}{2\pi \sum_{i=1}^{n} (x_i - \overline{x})^2} \right)^{n/2} \cdot \exp\left(-\frac{n}{2} \right) \\
&= \left(\frac{n \exp(-1)}{2\pi \sum_{i=1}^{n} (x_i - \overline{x})^2} \right)^{n/2}
\end{aligned}
$$

donc, le rapport de vraisemblance est :

$$
\begin{aligned}
\lambda &= \lambda(x_1, x_2, \ldots, x_n) \\
&= \frac{L\left(0, \widehat{\sigma}^2 ; x_1, x_2, \ldots, x_n\right)}{L\left(\widehat{\mu}, \widehat{\sigma}^2 ; x_1, x_2, \ldots, x_n\right)} \\
&= \left(\frac{n \exp(-1)}{2\pi \sum_{i=1}^{n} x_i^2} \right)^{n/2} \Bigg/ \left(\frac{n \exp(-1)}{2\pi \sum_{i=1}^{n} (x_i - \overline{x})^2} \right)^{n/2} \\
&= \left(\frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{\sum_{i=1}^{n} x_i^2} \right)^{n/2} .
\end{aligned}
$$

Il s'agit de reconnaître maintenant la loi que suit λ. Pour ce faire, nous nous rappelons que $\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} (x_i - \overline{x})^2 + n\overline{x}^2$ (cf. propriétés de la **moyenne arithmétique**) et obtenons :

$$
\begin{aligned}
\lambda &= \left(\frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{\sum_{i=1}^{n} (x_i - \overline{x})^2 + n\overline{x}^2} \right)^{n/2} \\
&= 1 \Bigg/ \left(1 + \frac{n\overline{x}^2}{\sum_{i=1}^{n} (x_i - \overline{x})^2} \right)^{n/2} .
\end{aligned}
$$

Nous cherchons λ_0 tel que l'hypothèse nulle soit rejetée si $\lambda \leq \lambda_0$, ce que nous pouvons écrire :

$$
\begin{aligned}
\lambda &\leq \lambda_0 \\
&\Leftrightarrow 1 \Bigg/ \left(1 + \frac{n\overline{x}^2}{\sum_{i=1}^{n} (x_i - \overline{x})^2} \right)^{n/2} \leq \lambda_0 \\
&\Leftrightarrow \left(1 + \frac{n\overline{x}^2}{\sum_{i=1}^{n} (x_i - \overline{x})^2} \right)^{n/2} \geq \lambda_0^{-1} \\
&\Leftrightarrow \frac{n\overline{x}^2}{\sum_{i=1}^{n} (x_i - \overline{x})^2} \geq \lambda_0^{-2/n} - 1.
\end{aligned}
$$

En multipliant chaque terme de l'inégalité par $(n-1)$, nous obtenons :

$$
\lambda \leq \lambda_0 \Leftrightarrow \frac{\sqrt{n}\,|\overline{x}|}{\sqrt{\sum_{i=1}^{n} (x_i - \overline{x})^2 / (n-1)}} \geq \sqrt{(n-1)\left(\lambda_0^{-2/n} - 1\right)}.
$$

Nous rappelons encore que :

$$
T = \frac{\overline{X} - \mu}{\widehat{\sigma} / (n-1)}
$$

où S^2 est la variance de l'échantillon et suit une t distribution avec $(n-1)$ degré de liberté ; nous pouvons aussi écrire :

$$
T = \frac{\sqrt{n}\left(\overline{X} - \mu\right)}{\sqrt{n\widehat{\sigma}^2 / (n-1)}}
$$

ce qui donne avec une moyenne μ nulle :

$$
\begin{aligned}
T &= \frac{\sqrt{n}\overline{X}}{\sqrt{n\widehat{\sigma}^2 / (n-1)}} \\
&= \frac{\sqrt{n}\overline{X}}{\sqrt{\sum_{i=1}^{n} x_i^2 / (n-1)}}.
\end{aligned}
$$

Il suffit donc de chercher la valeur de λ_0 telle que $c := \sqrt{(n-1)\left(\lambda_0^{-2/n} - 1\right)}$ soit la valeur de la

table de Student pour le **seuil de signification** α voulu :

$$\lambda_0 = \left(\frac{c^2}{n-1}+1\right)^{-n/2}.$$

Par exemple, pour un échantillon de taille $n = 6$, un seuil de signifiation $\alpha = 0,05$, nous trouvons dans la table de Student : $z_{5,0,05/2} = 2,751$, il suffit donc que

$$\lambda \leq \left(\frac{(2,751)^2}{5}+1\right)^{-3} = 0,063$$

pour que l'hypothèse nulle soit rejetée.

2. Soient deux échantillons aléatoires indépendants $X_1, \ldots, X_n$ et $Y_1, \ldots, Y_m$ $(m+n>2)$ suivants une loi normale de paramètres $\left(\mu_1, \sigma^2\right)$ et $\left(\mu_2, \sigma^2\right)$ respectivement. Nous désirons tester les hypothèses suivantes :

$$\begin{array}{rcl} H_0 & : & \mu_1 = \mu_2, \\ H_1 & : & \mu_1 \neq \mu_2. \end{array}$$

Nous définissons les deux ensembles :

$$\begin{array}{rcl} \Theta & = & \{(\mu_1, \mu_2, \sigma^2)\,;\, -\infty < \mu_1, \mu_2 < \infty, \\ & & 0 < \sigma^2 < \infty\}, \\ \Theta_1 & = & \left\{(\mu_1, \mu_2, \sigma^2) \in \Theta\,;\, \mu_1 = \mu_2\right\}. \end{array}$$

Nous pouvons alors écrire les hypothèses de la façon suivante :

$$\begin{array}{rcl} H_0 & : & \left(\mu_1, \mu_2, \sigma^2\right) \in \Theta_1, \\ H_1 & : & \left(\mu_1, \mu_2, \sigma^2\right) \in \Theta \backslash \Theta_1. \end{array}$$

Alors, les fonctions de vraisemblance sont :

$$\begin{array}{l} L\left(\mu_1, \mu_1, \sigma^2\,;\, x_1, \ldots, x_n, y_1, \ldots, y_m\right) \\ \quad = \left(\dfrac{1}{\sqrt{2\pi}\sigma}\right)^{n+m} \\ \quad \cdot \exp\left(-\dfrac{\sum\limits_{i=1}^{n}(x_i-\mu_1)^2 + \sum\limits_{i=1}^{m}(y_i-\mu_1)^2}{2\sigma^2}\right) \end{array}$$

$$\begin{array}{l} L\left(\mu_1, \mu_2, \sigma^2\,;\, x_1, \ldots, x_n, y_1, \ldots, y_m\right) \\ \quad = \left(\dfrac{1}{\sqrt{2\pi\sigma^2}}\right)^{n+m} \\ \quad \cdot \exp\left(-\dfrac{\sum\limits_{i=1}^{n}(x_i-\mu_1)^2 + \sum\limits_{i=1}^{m}(y_i-\mu_2)^2}{2\sigma^2}\right). \end{array}$$

En utilisant les extimateurs du maximum de vraisemblance, qui sont pour l'hypothèse nulle :

$$\begin{array}{rcl} \widehat{\mu}_1 & = & \dfrac{\sum\limits_{i=1}^{n} x_i + \sum\limits_{i=1}^{m} y_i}{n+m} = \dfrac{n\overline{x}+m\overline{y}}{n+m}, \\ \widehat{\sigma}^2 & = & \dfrac{\sum\limits_{i=1}^{n}(x_i-\widehat{\mu}_1)^2 + \sum\limits_{i=1}^{m}(y_i-\widehat{\mu}_1)^2}{n+m} \end{array}$$

et pour l'hypothèse alternative :

$$\begin{array}{rcl} \widehat{\mu}_1' & = & \dfrac{\sum\limits_{i=1}^{n} x_i}{n} = \overline{x}, \\ \widehat{\mu}_2 & = & \dfrac{\sum\limits_{i=1}^{m} y_i}{m} = \overline{y}, \\ \widehat{\sigma'}^2 & = & \dfrac{\sum\limits_{i=1}^{n}(x_i-\overline{x})^2 + \sum\limits_{i=1}^{m}(y_i-\overline{y})^2}{n+m}, \end{array}$$

nous trouvons le rapport de vraisemblance :

$$\begin{array}{rcl} \lambda & = & \dfrac{L\left(\widehat{\mu}_1, \widehat{\mu}_1, \widehat{\sigma}^2\,;\, x_1, \ldots, x_n, y_1, \ldots, y_m\right)}{L\left(\widehat{\mu}_1', \widehat{\mu}_2', \widehat{\sigma'}^2\,;\, x_1, \ldots, x_n, y_1, \ldots, y_m\right)} \\ & = & \left(\dfrac{2\pi}{n+m}\right)^{-(n+m)/2} \exp\left(-\dfrac{n+m}{2}\right) \\ & & \cdot \left[\sum_{i=1}^{n}\left(x_i - \dfrac{n\overline{x}+m\overline{y}}{n+m}\right)^2\right. \\ & & \left. + \sum_{i=1}^{m}\left(y_i - \dfrac{n\overline{x}+m\overline{y}}{n+m}\right)^2\right]^{-(n+m)/2} \\ & & \cdot \left(\dfrac{2\pi}{n+m}\right)^{(n+m)/2} \exp\left(\dfrac{n+m}{2}\right) \\ & & \left[\sum_{i=1}^{n}(x_i-\overline{x})^2 + \sum_{i=1}^{m}(y_i-\overline{y})^2\right]^{(n+m)/2} \end{array}$$

$$
\begin{aligned}
&= \left(\sum_{i=1}^{n}\left(x_i-\frac{n\overline{x}+m\overline{y}}{n+m}\right)^2\right.\\
&\quad\left.+\sum_{i=1}^{m}\left(y_i-\frac{n\overline{x}+m\overline{y}}{n+m}\right)^2\right)^{-(n+m)/2}\\
&\quad\cdot\left(\sum_{i=1}^{n}(x_i-\overline{x})^2+\sum_{i=1}^{m}(y_i-\overline{y})^2\right)^{(n+m)/2}\\
&= \left(\sum_{i=1}^{n}(x_i-\overline{x})^2+\sum_{i=1}^{m}(y_i-\overline{y})^2\right)^{(n+m)/2}\\
&\quad\Bigg/\left(\sum_{i=1}^{n}\left(x_i-\frac{n\overline{x}+m\overline{y}}{n+m}\right)^2\right.\\
&\qquad\left.+\sum_{i=1}^{m}\left(y_i-\frac{n\overline{x}+m\overline{y}}{n+m}\right)^2\right)^{(n+m)/2}.
\end{aligned}
$$

Ainsi, la statistique $\lambda^{2/(n+m)}$ est donnée par :

$$
\frac{\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2+\sum_{i=1}^{m}\left(Y_i-\overline{Y}\right)^2}{\sum_{i=1}^{n}\left(X_i-\frac{n\overline{X}+m\overline{Y}}{n+m}\right)^2+\sum_{i=1}^{m}\left(Y_i-\frac{n\overline{X}+m\overline{Y}}{n+m}\right)^2}.
$$

Pour trouver quelle loi cette statistique suit, nous utilisons l'astuce suivante (nous rappelons que $\sum_{i=1}^{n}\left(X_i-\overline{X}\right)=0$) :

$$
\begin{aligned}
&\sum_{i=1}^{n}\left(X_i-\frac{n\overline{X}+m\overline{Y}}{n+m}\right)^2\\
&= \sum_{i=1}^{n}\left(\left(X_i-\overline{X}\right)+\left(\overline{X}-\frac{n\overline{X}+m\overline{Y}}{n+m}\right)\right)^2\\
&= \sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2+n\left(\overline{X}-\frac{n\overline{X}+m\overline{Y}}{n+m}\right)^2\\
&= \sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2+\frac{nm^2}{(n+m)^2}\left(\overline{X}-\overline{Y}\right)^2.
\end{aligned}
$$

De même pour Y :

$$
\begin{aligned}
&\sum_{i=1}^{m}\left(Y_i-\frac{n\overline{X}+m\overline{Y}}{n+m}\right)^2\\
&= \sum_{i=1}^{m}\left(Y_i-\overline{Y}\right)^2+\frac{mn^2}{(n+m)^2}\left(\overline{X}-\overline{Y}\right)^2.
\end{aligned}
$$

Nous pouvons donc écrire :

$$
\begin{aligned}
\lambda^{2/(n+m)} &= \left(\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2+\sum_{i=1}^{m}\left(Y_i-\overline{Y}\right)^2\right)\\
&\quad\cdot\left(\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2+\sum_{i=1}^{m}\left(Y_i-\overline{Y}\right)^2\right.\\
&\quad\left.+\frac{nm}{n+m}\left(\overline{X}-\overline{Y}\right)^2\right)^{-1}\\
&= \left(1+\frac{\frac{nm}{n+m}\left(\overline{X}-\overline{Y}\right)^2}{\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2+\sum_{i=1}^{m}\left(Y_i-\overline{Y}\right)^2}\right)^{-1}.
\end{aligned}
$$

Si l'hypothèse nulle est vérifiée, la variable aléatoire

$$
T=\frac{\sqrt{\frac{nm}{n+m}}\left(\overline{X}-\overline{Y}\right)}{\sqrt{\frac{\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2+\sum_{i=1}^{m}\left(Y_i-\overline{Y}\right)^2}{n+m-2}}}
$$

suit la loi de Student avec $n+m-2$ degrés de liberté. Ainsi, nous obtenons :

$$
\begin{aligned}
\lambda^{2/(n+m)} &= \frac{1}{1+(n+m-2)\,T^2}\\
&= \frac{n+m-2}{n+m-2+T^2}.
\end{aligned}
$$

Nous recherchons λ_0 tel que l'hypothèse nulle soit rejetée dès que $\lambda\leq\lambda_0\leq 1$. Or

$$
\begin{aligned}
\lambda &= \left(\frac{n+m-2}{n+m-2+T^2}\right)^{2/(n+m)}\leq\lambda_0\\
&\Leftrightarrow\quad |T|\geq\sqrt{\left(\lambda_0^{-\frac{n+m}{2}}-1\right)(n+m-2)}:=c.
\end{aligned}
$$

Il suffit donc de prendre dans la table de Student la valeur de c en fonction de la taille des échantillons (n et m) et du **seuil de signification** α désiré $z_{n+m-2,\frac{\alpha}{2}}$. Par exemple, avec $n=10$, $m=6$, et $\alpha=0,05$, nous trouvons :

$$
\begin{aligned}
c &= z_{14,0,025}=2,145\\
\lambda_0 &= \left(\frac{c^2}{n+m-2}-1\right)^{-2/(n+m)}\\
&= \left(\frac{(2,145)^2}{14}-1\right)^{-1/8}=-1,671.
\end{aligned}
$$

MOTS CLÉS

Hypothèse *(Hypothesis)*
Hypothèse alternative *(Alternative hypothesis)*
Hypothèse nulle *(Null hypothesis)*
Maximum de vraisemblance *(Maximum likelihood)*
Test d'hypothèse *(Hypothesis testing)*
Test de Fisher *(Fisher test)*

RÉFÉRENCES

Cox, D.R. and Hinkley, D.V. (1973). *Theoretical statistics.* Chapman & Hall, London.

Edwards, A.W.F. (1972). *Likelihood. An account of the statistical concept of likelihood and its application to scientific inference.* Cambridge University Press, Cambridge.

Hogg, R.V. and Craig, A.T. (1970). *Introduction to mathematical statistics.* 3th ed. Collier-Macmillan.

Kendall, M.G. and Steward, A. (1967). *The advanced Theory of Statistics.* Vol. 2, Charles Griffin & Company Ltd.

Neyman, J. and Pearson, E.S. (1928). On the use and interpretation of certain test criteria for purposes of statistical inference. I, II. *Biometrika*, 20A, pp. 175-240 and 263-294.

Wald, A. (1941). Asymptotically Most Powerful Tests of Statistical Hypotheses. *Annals of Mathematical Statistics*, 12, pp. 1-19.

Wald, A. (1941). Some Examples of Asymptotically Most Powerful Tests. *Annals of Mathematical Statistics*, 12, pp. 396-408.

Wilks, S.S. (1962). *Mathematical Statistics.* John Wiley & Sons, New York.

TEST DU SIGNE

Sign test

Le test du signe est un **test non paramétrique** pour tester si une **population** tend à avoir des valeurs plus grandes que celles d'une autre population.

Il doit son nom au fait qu'il utilise des signes « + » et « − » au lieu d'utiliser des valeurs quantitatives. Il est donc applicable même quand la série de mesures est ordinale.

HISTORIQUE

Dans ce qui fut certainement la première publication se rapportant aux tests non paramétriques, J. Arbuthnott (1710), étudiait la liste des naissances enregistrées à Londres sur une période de 82 ans. Pour chacune des années, il comparait le nombre de nouveau-nés de chacun des deux sexes. Il dénota par « + » l'**événement** « plus de garçons nés que de filles » et par « − » l'événement opposé. (Il n'y avait pas eu d'égalité).

À sa plus grande stupéfaction, il constata qu'il obtenait 82 fois le signe « + » et jamais le signe « − » ; ce qui l'amena à rejeter l'**hypothèse nulle** d'une équité des naissances relativement au sexe du nouveau-né.

ASPECTS MATHÉMATIQUES

Soient n paires d'observations (x_1, y_1), (x_2, y_2), ..., (x_n, y_n). Pour chaque paire (x_i, y_i), on fait une comparaison suivant la grandeur relative de x_i par rapport à y_i :

$$\begin{aligned} &\text{« + » si } x_i < y_i, \\ &\text{« − » si } x_i > y_i, \\ &\text{« = » si } x_i = y_i. \end{aligned}$$

On soustrait de la taille de l'**échantillon** le nombre de signes « = » qui apparaissent, c'est-à-dire que l'on tient compte uniquement des paires qui présentent une différence (positive ou négative). On note m ce nombre de paires.

On comptabilise ensuite le nombre de « + », que l'on désigne par T.

Hypothèses

Selon le type de test, bilatéral ou unilatéral, les hypothèses nulle et alternative correspondant au test du signe sont :

A. Cas bilatéral :

$$\begin{array}{lcl} H_0 & : & P(X < Y) = P(X > Y), \\ H_1 & : & P(X < Y) \neq P(X > Y). \end{array}$$

B. Cas unilatéral :

$$\begin{array}{lcl} H_0 & : & P(X < Y) \leq P(X > Y), \\ H_1 & : & P(X < Y) > P(X > Y). \end{array}$$

C. Cas unilatéral :

$$\begin{array}{lcl} H_0 & : & P(X < Y) \geq P(X > Y), \\ H_1 & : & P(X < Y) < P(X > Y). \end{array}$$

Dans le cas A, on fait l'hypothèse (H_0) que la **probabilité** que X soit inférieur à Y est la même que celle que X soit supérieur à Y.

Dans le cas B, on suppose *a priori* que la probabilité que X soit inférieur à Y est inférieure ou égale à celle que X soit supérieur à Y.

Dans le cas C enfin, on suppose *a priori* que la probabilité que X soit inférieur à Y est supérieure ou égale à celle que X soit supérieur à Y.

Étant donné que dans le cas bilatéral, la probabilité que X soit inférieur à Y (comme celle que $X > Y$) est égale à $\frac{1}{2}$, le test du signe est un cas particulier du **test binomial** avec $p = \frac{1}{2}$, c'est-à-dire que sous l' **hypothèse nulle** H_0 on s'attend à obtenir autant de « + » que de « − ».

Règles de décision

Cas A
On se base sur la **table binomiale** avec une probabilité p égale à $\frac{1}{2}$.

Il convient de trouver dans la table la **valeur** la plus proche de $\frac{\alpha}{2}$, où α est le **seuil de signification**, et de noter par $t_{\alpha/2}$ la valeur correspondant à $\frac{\alpha}{2}$.

On rejette H_0 au seuil α si

$$T \leq t_{\alpha/2} \text{ ou } T \geq m - t_{\alpha/2}$$

avec $T =$ nombre de « + ».

Quand $m > 20$, on peut utiliser la **loi normale** comme approximation de la **loi binomiale**.

On transforme la **statistique** T en une **variable aléatoire** Z suivant une loi normale centrée réduite :

$$Z = \frac{T - \mu}{\sigma}$$

où

$$\begin{array}{rcl} \mu & = & m \cdot p = \dfrac{1}{2} m \\ \text{et} \quad \sigma & = & \sqrt{m \cdot p \cdot q} = \dfrac{1}{2}\sqrt{m} \end{array}$$

sont la **moyenne** et l'**écart type** de la loi binomiale.

On obtient ainsi :

$$Z = \frac{T - \frac{1}{2}m}{\frac{1}{2}\sqrt{m}}.$$

L'approximation de la valeur de la table binomiale est donc :

$$t_{\alpha/2} = \frac{1}{2}\left(m + z_{\alpha/2}\sqrt{m}\right)$$

où $z_{\alpha/2}$ est à chercher dans la **table normale** au seuil $\frac{\alpha}{2}$. Ensuite, la règle de décision est la même.

Cas B
On rejette H_0 au seuil α si

$$T \geq m - t_\alpha$$

où t_α est la valeur de la table binomiale correspondant au **seuil de signification** α (ou la valeur la plus proche).

Pour $m > 20$, on fait l'approximation

$$t_\alpha = \frac{1}{2}\left(m + z_\alpha\sqrt{m}\right)$$

où z_α est à chercher dans la table normale au seuil α.

Cas C
On rejette H_0 au seuil α si

$$T \leq t_\alpha$$

où la valeur de t_α est la même que pour le cas B.

DOMAINES ET LIMITATIONS

Les exigences préalables de l'utilisation du test du signe sont les suivantes :

1. les couples de variables aléatoires (X_i, Y_i), $i = 1$, 2, …, n sont mutuellement indépendants ;

2. l'échelle de mesure est au moins ordinale, c'est-à-dire qu'à chaque paire peut être attribué un « + », un « − » ou un « = » ;

3. les paires (X_i, Y_i), $i = 1, 2, \ldots, n$ présentent une logique entre elles, c'est-à-dire que si $P(\text{« + »}) > P(\text{« - »})$ pour une paire (X_i, Y_i), alors $P(\text{« + »}) > P(\text{« - »})$ pour toutes les autres paires.
   De même si $P(\text{« + »}) < P(\text{« - »})$ ou $P(\text{« + »}) = P(\text{« - »})$.

EXEMPLES

Exemple 1 : un fabricant de chocolat étudie un nouvel emballage pour un de ses produits. Dans ce but, il propose à 10 consommateurs d'attribuer à l'ancien et au nouvel emballage des notes d'appréciation comprises entre 1 et 6 (1 étant la moins bonne note et 6 la meilleure).

On effectue un **test unilatéral** (correspondant au cas *C*) où l'**hypothèse nulle** que l'on souhaite tester s'écrit :

$$H_0 \quad : \quad P(\text{« + »}) \geq P(\text{« - »})$$

par rapport à l'**hypothèse alternative** :

$$H_1 \quad : \quad P(\text{« + »}) < P(\text{« - »}).$$

Le signe « + » signifie que le nouvel emballage est préféré à l'ancien.

Les résultats obtenus sont :

Consommateur	Note d'appréciation		Signe de la différence
	ancien emballage	nouvel emballage	
1	4	5	+
2	3	5	+
3	1	1	=
4	5	4	−
5	4	4	=
6	2	6	+
7	4	6	+
8	4	5	+
9	2	3	+
10	1	4	+

Le tableau montre qu'il y a 2 consommateurs qui n'ont indiqué aucune différence entre le nouveau et l'ancien emballage. Il y a donc 8 différences ($m = 8$) parmi lesquelles le nouvel emballage a été, dans 7 cas, apprécié davantage que l'ancien. On a donc $T = 7$ (nombre de « + »).

On rejette alors H_0 au seuil α si

$$T \leq t_\alpha$$

où t_α est la **valeur** de la **table binomiale**.

On choisit $\alpha = 0,05$; la valeur de t_α est 1 (pour $m = 8$ et $\alpha = 0,0352$). On a alors $T > t_\alpha$ puisque $T = 7$ et $t_\alpha = 1$. On ne rejette donc pas l'**hypothèse** H_0 : $P(\text{« + »}) \geq P(\text{« - »})$, et le fabricant de chocolat peut conclure que les consommateurs préfèrent en général le nouvel emballage.

Exemple 2 : on reprend le même exemple mais cette fois on considère que l'étude est menée sur 100 consommateurs. On suppose que les résultats sont les suivants :

— 83 personnes préfèrent le nouvel emballage ;

— 7 personnes préfèrent l'ancien emballage ;

— 10 personnes sont indifférentes.

On a donc :

$$\begin{aligned} m &= 100 - 10 = 90 \\ T &= 83. \end{aligned}$$

La valeur de t_α peut être approximée par la **loi normale** :

$$t_\alpha = \frac{1}{2}\left(m + z_\alpha\sqrt{m}\right)$$

où z_α est la valeur de la **table normale** pour $\alpha = 0,05$. Cela donne :

$$\begin{aligned} t_\alpha &= \frac{1}{2}\left(90 - 1,64\sqrt{90}\right) \\ &= 37,22. \end{aligned}$$

T étant supérieur à t_α $(83 > 37,22)$, l'hypothèse nulle H_0 : $P(\text{« + »}) \geq P(\text{« - »})$ n'est pas rejetée et le fabricant de chocolat tire la même conclusion que précédemment.

MOTS CLÉS

Loi normale *(Normal distribution)*
Table binomiale *(Binomial table)*
Test binomial *(Binomial test)*
Test d'hypothèse *(Hypothesis testing)*
Test non paramétrique *(Nonparametric test)*

RÉFÉRENCE

Arbuthnott, J. (1710). An argument for Divine Providence, taken from the constant regularity observed in the births of both sexes. *Philosophical Transactions* 27, pp. 186-190. (3.4).

TEST NON PARAMÉTRIQUE

Nonparametric test

Un test non paramétrique est un **test d'hypothèse** pour lequel il n'est pas nécessaire de spécifier la forme de la distribution de la **population** étudiée. Il faut cependant en général que les observations soient indépendantes, c'est-à-dire que la sélection d'un quelconque individu dans la population en vue de former l'**échantillon** ne doit pas influencer le choix des autres individus.

HISTORIQUE

Le premier test non paramétrique remonte aux travaux de J. Arbuthnott (1710), qui introduisit le **test du signe**.

Cependant, la plupart des tests non paramétriques ont été développés entre 1940 et 1955.

On peut notamment citer les articles d'**Andreï Nikolaïévitch Kolmogorov** (1933), de Vladimir Ivanovitch Smirnov (1939), de **Frank Wilcoxon** (1945 et 1947), de H.B. Mann et D.R. Whitney (1947), d'A.M. Mood (1950), et de W.H. Kruskal et W.A. Wallis (1952).

Par la suite, d'innombrables articles se sont encore ajoutés à cette liste. I.R. Savage (1962) a publié une bibliographie d'environ 3 000 articles, écrits avant 1962, se rapportant aux tests non paramétriques.

DOMAINES ET LIMITATIONS

Le développement rapide des tests non paramétriques peut être expliqué par les quelques points suivants :

- les méthodes non paramétriques requièrent peu d'hypothèses concernant la **population** étudiée. Elles ignorent notamment l'hypothèse classique de la normalité de la population ;
- les méthodes non paramétriques sont souvent plus faciles à comprendre et à appliquer que les tests paramétriques équivalents ;
- les méthodes non paramétriques sont applicables dans des situations où les tests paramétriques ne peuvent pas être utilisés, par exemple lorsque les variables ne sont mesurées que sur des échelles ordinales ;
- bien qu'à première vue les méthodes non paramétriques semblent sacrifier une bonne part de l'information contenue dans les échantillons, des recherches théoriques ont montré que l'efficacité des tests non paramétriques n'est que légèrement inférieure à celle de leurs équivalents paramétriques quand la distribution de la population étudiée est spécifiée, par exemple la **loi normale**. Elle est en revanche supérieure à celle des tests paramétriques quand la distribution de la population dévie sensiblement de la distribution spécifiée (normale).

EXEMPLES

Voir **coefficient de corrélation des rangs de Spearman**, **test de Kolmogorov-Smirnov**, **test de Kruskal-Wallis**, **test de Mann-Withney**, **test de Wilcoxon (signé)**, **test du signe.**

MOTS CLÉS

Coefficient de corrélation des rangs de Spearman *(Spearman's rank correlation coefficient)*
Test d'adéquation *(Goodness of fit test)*
Test d'hypothèse *(Hypothesis testing)*
Test d'indépendance *(Test of independence)*
Test de Kolmogorov-Smirnov *(Kolmogorov-Smirnov test)*
Test de Kruskal-Wallis *(Kruskal-Wallis test)*
Test de Mann-Whitney *(Mann-Whitney test)*
Test de Wilcoxon *(Wilcoxon test)*
Test de Wilcoxon signé *(Wilcoxon signed test)*
Test du signe *(Sign test)*

RÉFÉRENCES

Arbuthnott, J. (1710). An argument for Divine Providence, taken from the constant regularity observed in the births of both sexes. *Philosophical Transactions* 27, pp. 186-190.(3.4).

Kolmogorov, A.N. (1933). Sulla determinazione empirica di una legge di distribuzione. *Giornale dell' Instituto Italiano degli Attuari*, 4, pp. 83-91 (6.1).

Kruskal, W.H. and Wallis, W.A. (1952). Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association*, 47, pp. 583-621 and errata, *ibid.*, 48, pp. 907-911.

Mann, H.B. and Whitney, D.R. (1947). On a test whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics*, 18, pp. 50-60.

Mood, A.M. (1950). *Introduction to the theory of statistics.* McGraw-Hill, New York. (chap. 16).

Savage, I.R. (1962). *Bibliography of Nonparametric Statistics.* Harvard University Press, Cambridge, Mass.

Smirnov, N.V. (1939). Estimate of deviation between empirical distribution functions in two independent samples. (Russian). *Bulletin Moscow Univ.* 2: 2, pp. 3-16 (6.1, 6.2).

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics*, 1, pp. 80-83.

Wilcoxon, F. (1957). *Some rapid approximate statistical procedures.* American Cyanamid Co., Stamford Research Laboratories, Stamford, Conn.

TEST PARAMÉTRIQUE

Parametric test

Un test paramétrique est un **test d'hypothèse** qui suppose une forme paramétrique des distributions relatives aux populations sous-jacentes. Tel est le cas lorsque les populations étudiées suivent une **loi normale**.

HISTORIQUE

L'un des premiers tests paramétriques apparus est celui du chi-carré, introduit par Pearson en 1900.

EXEMPLES

Le **test de Student** est un exemple de test paramétrique. Celui-ci vise en effet à comparer les moyennes de deux populations qui suivent une distribution normale. Dans d'autres cas, il peut s'agir de déterminer l'adéquation entre la distribution des valeurs d'un échantillon à une loi de probabilité présupposée. L'un de ces tests est le **test du chi-carré**.

Parmi les tests paramétriques les plus connus on cite le **test binomial** et le **test de Fisher**.

MOTS CLÉS

Test binomial *(Binomial testing)*
Test d'hypothèse *(Hypothesis testing)*
Test de Fisher *(Fisher testing)*
Test de Student *(Student test)*

TEST UNILATÉRAL

One-tail test

Un test unilatéral concernant une **population** est un **test d'hypothèse** qui s'applique lorsque nous cherchons à comparer une estimation et une valeur donnée ou deux valeurs données.

ASPECTS MATHÉMATIQUES

Un test unilatéral sur un **échantillon** est un **test d'hypothèse** dans lequel les **hypothèses** sont de la forme :

$$(1)\ \begin{array}{lcl} H_0 & : & \theta \geq \theta_0 \\ H_1 & : & \theta < \theta_0 \end{array} \quad \text{ou} \quad (2)\ \begin{array}{lcl} H_0 & : & \theta \leq \theta_0 \\ H_1 & : & \theta > \theta_0, \end{array}$$

H_0 étant l'**hypothèse nulle**, H_1 l'**hypothèse alternative**, θ un **paramètre** inconnu de la **population** et θ_0 la **valeur** présumée de ce paramètre.

Dans le cas d'un test unilatéral sur deux échantillons, les hypothèses sont les suivantes :

$$(1)\ \begin{array}{lcl} H_0 & : & \theta_1 \geq \theta_2 \\ H_1 & : & \theta_1 < \theta_2, \end{array} \quad \text{ou} \quad (2)\ \begin{array}{lcl} H_0 & : & \theta_1 \leq \theta_2 \\ H_1 & : & \theta_1 > \theta_2, \end{array}$$

θ_1 et θ_2 étant les paramètres inconnus des deux populations à comparer.

Dans les cas (1), le test est dit unilatéral à gauche et la **région de rejet** de l'hypothèse nulle se trouve entièrement à gauche de l'**intervalle** de variation de θ.

Dans les cas (2), le test est dit unilatéral à droite et la région de rejet se trouve entièrement à droite de l'intervalle de variation de θ.

EXEMPLES

Test unilatéral sur la moyenne d'une population

Un producteur de parfum désire s'assurer que les flacons contiennent bien un minimum de 40 ml de parfum. Un **échantillon** de 50 flacons donne une **moyenne** échantillonnale de 39 ml, avec un **écart type** de 4 ml.

Les hypothèses sont les suivantes :

$$\begin{array}{lcl} \textbf{hypothèse nulle} & H_0 & : \quad \mu \geq 40, \\ \textbf{hypothèse alternative} & H_1 & : \quad \mu < 40. \end{array}$$

La taille de l'échantillon étant suffisamment grande, la **distribution d'échantillonnage** de la moyenne peut être approximée par une **loi normale**.

Pour un **seuil de signification** $\alpha = 1\ \%$, nous obtenons la **valeur** de z_α dans la **table normale** : $z_\alpha = 2,33$.

La **valeur critique** de l'**hypothèse nulle** est calculé par l'expression :

$$\mu_{\bar{x}} - z_\alpha \cdot \sigma_{\bar{x}}$$

où $\mu_{\bar{x}}$ est la moyenne de la distribution d'échantillonnage des moyennes

$$\mu_{\bar{x}} = \mu$$

et $\sigma_{\bar{x}}$ est l'écart type de la distribution d'échantillonnage des moyennes, ou **erreur type** de la moyenne

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}.$$

Nous obtenons donc la valeur critique suivante :

$$40 - 2,33 \cdot \frac{4}{\sqrt{50}} = 38,68.$$

Comme la moyenne échantillonnale $\bar{x} = 39$ est supérieure à la valeur critique pour le test unilatéral à gauche, le producteur ne peut donc rejeter l'hypothèse nulle et peut donc considérer que ses flacons contiennent au moins 40 ml, le contenu moyen de l'échantillon étant attribué au hasard.

Test unilatéral sur le pourcentage de deux populations

Le rapporteur d'un projet de loi sur le trafic routier pense que son projet sera perçu de manière beaucoup plus favorable par la **population** urbaine que par la population rurale. Une enquête a été réalisée sur deux échantillons de 100 personnes provenant respectivement d'un milieu urbain et d'un milieu rural.

Dans le milieu urbain (population 1), 82 personnes étaient favorables à son projet, alors que dans le milieu rural (population 2), 69 personnes seulement se sont prononcées de manière positive.

En vue de confirmer ou d'infirmer l'intuition du rapporteur, nous effectuons un **test d'hypothèse** en posant les hypothèses suivantes :

hypothèse nulle	H_0	:	$\pi_1 \leq \pi_2$
	ou		$\pi_1 - \pi_2 \leq 0$
hypothèse	H_1	:	$\pi_1 > \pi_2$
alternative	ou		$\pi_1 - \pi_2 > 0$

où π_1 et π_2 représentent la proportion favorable de la population urbaine et rurale, respectivement.

En fonction des pourcentages mesurés sur les deux échantillons ($p_1 = 0,82$ et $p_2 = 0,69$), nous sommes en mesure d'estimer la **valeur** de l'écart type de la distribution d'échantillonnage de la différence des pourcentages, ou erreur type de la différence des pourcentages :

$$\begin{aligned}\hat{\sigma}_{p_1-p_2} &= \sqrt{\frac{p_1 \cdot (1-p_1)}{n_1} + \frac{p_2 \cdot (1-p_2)}{n_2}} \\ &= \sqrt{\frac{0,82 \cdot (1-0,82)}{100} + \frac{0,69 \cdot (1-0,69)}{100}} \\ &= 0,06012.\end{aligned}$$

Si nous effectuons un test avec un **seuil de signification** α de 5 %, la **valeur** de z_α dans la **table normale** est égale à 1,645. La valeur critique de l'hypothèse nulle est calculé par :

$$\begin{aligned}\pi_1 - \pi_2 + z_\alpha \cdot \sigma_{p_1-p_2} &= 0 + 1,645 \cdot 0,06012 \\ &= 0,0989.\end{aligned}$$

Étant donné que la différence des proportions observées $p_1 - p_2 = 0,82 - 0,69 = 0,13$ est supérieure à la valeur critique, nous devons rejeter l'hypothèse nulle au profit de l'hypothèse alternative, confirmant ainsi la supposition du rapporteur.

MOTS CLÉS

Distribution d'échantillonnage *(Sampling distribution)*
Hypothèse alternative *(Alternative hypothesis)*
Hypothèse nulle *(Null hypothesis)*
Région d'acceptation *(Acceptance region)*
Région de rejet *(Rejection region)*
Seuil de signification *(Significance level)*
Test bilatéral *(Two-tail test)*
Test d'hypothèse *(Hypothesis testing)*

THÉORÈME CENTRAL LIMITE

Central limit theorem

Le théorème central limite est un théorème fondamental de la théorie statistique. Dans sa forme la plus simple, il prescrit que la somme d'un nombre suffisamment grand de variables aléatoires indépendantes et distribuées identiquement suit approximativement une **loi normale**.

HISTORIQUE

Le théorème central limite a été établi initialement dans le cadre d'une **loi binomiale** par **Abraham de Moivre** (1733). **Pierre Simon de Laplace** (1810) en établit la preuve.

Siméon Denis Poisson (1824) travailla également sur ce théorème et au milieu du XIX^e^ siècle **Pafnouti Lvovitch Tchebychev** (1890-1891) en donna une démonstration rigoureuse.

Au début du XX^e^ siècle, le mathématicien russe Aleksandr Mikhaïlovitch Liapounov (1901) démontra, sous des conditions plus générales, le théorème central limite en introduisant les fonctions caractéristiques. **Andreï Andreïevitch Markov** (1908) y consacra aussi des travaux et serait le premier à avoir généralisé le théorème aux cas où les variables sont dépendantes.

Selon L. Le Cam (1986), le qualificatif de « central » serait dû à George Polyà (1920), qui l'appela ainsi pour le rôle essentiel de ce théorème dans la théorie des probabilités.

ASPECTS MATHÉMATIQUES

Soit $X_1, X_2, \ldots, X_n$ une suite de n variables aléatoires indépendantes de même loi (quelconque) ayant une **moyenne** μ et une **variance** finie σ^2.

On définit la somme $S_n = X_1 + X_2 + \ldots + X_n$ et on établit le ratio :

$$\frac{S_n - n \cdot \mu}{\sigma \cdot \sqrt{n}}$$

où $n \cdot \mu$ et $\sigma \cdot \sqrt{n}$ représentent la moyenne et l'**écart type** de S_n, respectivement.

Le théorème central limite établit que la distribution de ce ratio tend vers la distribution normale lorsque n tend vers l'infini. Rigoureusement, cela signifie que :

$$P\left(\frac{S_n - n \cdot \mu}{\sigma\sqrt{n}} \leq x\right) \underset{n \to +\infty}{\longrightarrow} \Phi(x)$$

où $\Phi(x)$ est la **fonction de répartition** de la **loi normale** centrée réduite, exprimée par :

$$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) dx, \quad -\infty < x < \infty.$$

DOMAINES ET LIMITATIONS

Le théorème central limite fournit une méthode simple pour le calcul approximatif de probabilités liées à des sommes de variables aléatoires.

Outre l'intérêt qu'il présente pour la théorie de l'**échantillonnage** où les sommes et les moyennes occupent une place importante, le théorème central limite est utilisé pour l'approximation des lois usuelles dérivées de sommes de lois identiques. On peut par exemple, à l'aide du théorème central limite, utiliser la **loi normale** comme approximation de la **loi binomiale**, de la **loi de Poisson**, de la **loi gamma**, de la **loi du chi-carré**, de la **loi de Student**, de la **loi hypergéométrique**, de la **loi de Fisher** et de la **loi lognormale**.

EXEMPLES

Dans un lot de pièces électriques, la **probabilité** de tirer une pièce défectueuse est égale à $p = \frac{1}{8}$.

Quelle est la probabilité d'en trouver plus de 4 défectueuses sur 25 tirages ?

Soit X la **variable dichotomique** désignant le résultat d'un tirage :

$$X = \begin{cases} 1 & \text{si la pièce sélectionnée est défectueuse,} \\ 0 & \text{si elle ne l'est pas.} \end{cases}$$

La **variable aléatoire** X suit une **loi de Bernoulli** de **paramètre** p. Par conséquent, la somme $S_n = X_1 + X_2 + \ldots + X_n$ suit une **loi binomiale** de **moyenne** np et de **variance** $np(1-p)$, qui, suivant le théorème central limite, peut être approximée par la **loi normale** de moyenne $\mu = np$ et de variance $\sigma^2 = np(1-p)$.

On évalue ces valeurs :

$$\begin{aligned} \mu &= n \cdot p = 25 \cdot \frac{1}{8} = 3,125 \\ \sigma^2 &= n \cdot p\,(1-p) = 25 \cdot \frac{1}{8} \cdot \left(1 - \frac{1}{8}\right) = 2,734. \end{aligned}$$

On calcule ensuite $P(S_n > 4)$ de deux manières différentes :

1. avec la loi binomiale :

   on peut trouver dans la **table binomiale** la probabilité $P(S_n \leq 4) = 0,8047$. La probabilité $P(S_n > 4)$ est donc égale à :

   $$P(S_n > 4) = 1 - P(S_n \leq 4) = 0,1953 \text{ ;}$$

2. avec l'approximation normale (obtenue par le théorème central limite) :

   pour tenir compte du caractère discret de la variable aléatoire S_n, on doit effectuer une correction de continuité, c'est-à-dire que l'on calcule la probabilité que S_n soit supérieure à $4 + \frac{1}{2} = 4,5$.

   On pose :

   $$\begin{aligned} z &= \frac{S_n - n \cdot p}{\sqrt{n \cdot p\,(1-p)}} \\ &= \frac{4,5 - 3,125}{1,654} \\ &= 0,832. \end{aligned}$$

En se référant à la **table normale**, on obtient alors la probabilité :

$$\begin{aligned} P(Z > z) &= P(Z > 0,832) \\ &= 1 - P(Z \leq 0,832) \\ &= 1 - 0,7967 = 0,2033. \end{aligned}$$

MOTS CLÉS

Convergence *(Convergence)*
Loi binomiale *(Binomial distribution)*
Loi de Fisher *(Fisher distribution)*
Loi de Poisson *(Poisson distribution)*
Loi de probabilité *(Probability distribution)*
Loi des grands nombres *(Law of large numbers)*
Loi de Student *(Student distribution)*
Loi du chi-carré *(Chi-square distribution)*
Loi gamma *(Gamma distribution)*
Loi hypergéométrique *(Hypergeometric distribution)*
Loi lognormale *(Lognormal distribution)*
Loi normale *(Normal distribution)*
Probabilité *(Probability)*
Théorème de convergence *(Convergence theorem)*

RÉFÉRENCES

Laplace, P.S. de (1810). Mémoire sur les approximations des formules qui sont fonctions de très grands nombres et sur leur application aux probabilités. *Mémoires de l'Académie royale des sciences de Paris*, 10. Reproduit dans *Œuvres de Laplace* 12, pp. 301-347.

Le Cam, L. (1986). The Central Limit Theorem around 1935. *Statistical Science*, 1, pp. 78-96.

Liapounov, A.M. (1900). Sur une proposition de la théorie des probabilités. *Bulletin de l'Académie impériale des sciences de Saint-Pétersbourg* 8, pp. 1-24.

Markov, A.A. (1908). Extension des théorèmes limites du calcul des probabilités aux sommes des quantités liées en chaîne. *Mem. Acad. Sci. St Petersburg* 8, pp. 365-397.

Moivre, A. de (1733). Approximatio ad summam terminorum binomii $(a + b)^n$, in Seriem Expansi, Supplementum II to *Miscellanæ Analytica*, pp. 1-7. Photographically reprinted in *A rare pamphlet of Moivre and some of his discoveries.* Publié par Archibald R.C. (1926). *Isis* 8, pp. 671-683.

Poisson, S.D. (1824). Sur la probabilité des résultats moyens des observations. *Connaissance des temps pour l'an 1827*, pp. 273-302.

Polyà, G. (1920). Ueber den zentralen Grenzwertsatz der Wahrscheinlichkeitsrechnung und das Momentproblem. *Mathematische Zeitschrift* 8, pp. 171-181.

Tchebychev, P.L. (1890-1891). Sur deux théorèmes relatifs aux probabilités. *Acta Mathematica* 14, pp. 305-315.

THÉORÈME DE BAYES

Bayes' theorem

Si on considère un ensemble de « causes » et un **événement** qui s'est réalisé, le théorème de Bayes fournit la formule de la **probabilité** que l'événement soit le résultat immédiat d'une certaine cause.

Ainsi le théorème de Bayes s'interprète comme une formule pour la **probabilité conditionnelle** d'une cause sachant qu'un événement s'est produit.

HISTORIQUE

Le théorème de Bayes portant le nom de son auteur, le révérend **Thomas Bayes,** a été développé au milieu du XVIII^{e} siècle. Comme T. Bayes n'a fait aucune publication au cours de sa vie, son œuvre n'a été publiée qu'en 1763, deux ans après sa mort. R. Price présenta l'article principal le 23 décembre 1763 devant la Royal Society de Londres, dont T. Bayes avait été membre durant les vingt dernières années de sa vie.

ASPECTS MATHÉMATIQUES

Soit $\{A_1, A_2, \ldots, A_k\}$ une partition de l'**ensemble fondamental** Ω.

On suppose que chacun des événements A_1, ..., A_k a une **probabilité** non nulle.

Soit E un événement tel que $P(E) > 0$.

Alors pour tout i, $(1 \leq i \leq k)$ le théorème de Bayes dans le cas discret stipule que :

$$P(A_i|E) = \frac{P(A_i) \cdot P(E|A_i)}{\sum_{j=1}^{k} P(A_j) \cdot P(E|A_j)}.$$

Dans le cas continu, si X est une **variable aléatoire** donnée par sa **fonction de densité** $f(x)$, dite également fonction de densité *a priori*, le théorème de Bayes fournit la fonction de densité *a posteriori* :

$$f(x|E) = \frac{f(x) \cdot P(E|X = x)}{\int_{-\infty}^{\infty} f(x) \cdot P(E|X = x)dx}.$$

DOMAINES ET LIMITATIONS

Le théorème de Bayes a été l'objet de multiples controverses sur les possibilités d'utilisation lorsque les valeurs des probabilités utilisées pour déterminer la **fonction de probabilité** *a posteriori* ne sont généralement pas établies de façon précise.

EXEMPLES

Trois urnes contiennent des boules rouges, blanches et noires :

- l'urne A contient 5 boules rouges, 2 boules blanches et 3 boules noires ;
- l'urne B contient 2 boules rouges, 3 boules blanches et 1 boule noire ;
- l'urne C contient 5 boules rouges, 2 boules blanches et 5 boules noires.

En choisissant une urne au hasard, on tire une boule : elle est blanche. On se propose de déterminer la **probabilité** qu'elle ait été tiré dans l'urne A.

On pose A_1 l'**événement** « choisir l'urne A », A_2 l'événement « choisir l'urne B » et A_3 l'événement « choisir l'urne C ». $\{A_1, A_2, A_3\}$ forme une partition de l'**ensemble fondamental**.

Soit l'événement E « la boule tirée est blanche » qui a une probabilité strictement positive.

On a :

$$P(A_1) = P(A_2) = P(A_3) = \frac{1}{3}, \quad P((E|A_1)) = \frac{2}{10}$$

$$P((E|A_2)) = \frac{3}{6} \quad \text{et} \quad P((E|A_3)) = \frac{2}{12}.$$

La formule de Bayes permet de déterminer la probabilité que la boule blanche tirée provienne de l'urne A :

$$\begin{aligned} P(A_1|E) &= \frac{P(A_1) \cdot P(E|A_1)}{\sum_{i=1}^{3} P(A_i) \cdot P(E|A_i)} \\ &= \frac{\frac{1}{3} \cdot \frac{2}{10}}{\frac{1}{3} \cdot \frac{2}{10} + \frac{1}{3} \cdot \frac{3}{6} + \frac{1}{3} \cdot \frac{2}{12}} \\ &= \frac{3}{13}. \end{aligned}$$

MOTS CLÉS

Probabilité *(Probability)*
Probabilité conditionnelle *(Conditional probability)*

RÉFÉRENCE

Bayes, T. (1763). An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London*, 53, pp. 370-418. Publié deux ans après la mort de son auteur à l'instigation de R. Price. Réédité avec une biographie de **George A. Barnard** en 1958 et dans **Egon Sharpe Pearson** et Maurice George Kendall (eds) 1970, *Studies in the History of Statistics and Probability*, pp. 131-153.

THÉORÈME DE BERNOULLI

Bernoulli's theorem

Le théorème de Bernoulli stipule que la **fréquence**

relative des succès dans une suite d'épreuves de Bernoulli s'approche de la **probabilité** de succès quand le nombre d'essais croît indéfiniment.

Il est une forme simplifiée de la **loi des grands nombres** et dérive de l'inégalité de Tchebychev.

HISTORIQUE

Le théorème de Bernoulli, parfois aussi appelé loi faible des grands nombres, est dû à **Jakob Bernoulli** (1713) dans son ouvrage *Ars Conjectandi*, qui fut publié sept ans après sa mort à l'instigation de son neveu Nicolaus.

ASPECTS MATHÉMATIQUES

Si S représente le nombre de succès obtenus lors de n épreuves de Bernoulli et que p est la **probabilité** de succès, alors on a :

$$\lim_{n\to\infty} P\left(\left|\frac{S}{n} - p\right| \geq \varepsilon\right) = 0$$

ou

$$\lim_{n\to\infty} P\left(\left|\frac{S}{n} - p\right| < \varepsilon\right) = 1$$

avec $\varepsilon > 0$ arbitrairement petit.

D'une manière équivalente, on peut écrire :

$$\frac{S}{n} \underset{n\to\infty}{\longrightarrow} p,$$

ce qui signifie que la **fréquence** relative des succès tend vers la probabilité de succès quand n tend vers l'infini.

MOTS CLÉS

Convergence *(Convergence)*
Épreuve de Bernoulli *(Bernoulli trial)*
Loi de Bernoulli *(Bernoulli distribution)*
Loi des grands nombres *(Law of large numbers)*

RÉFÉRENCES

Bernoulli, J. (1713). *Ars Conjectandi, Opus Posthumum. Accedit Tractatus de Seriebus infinitis, et Epistola Gallice scripta de ludo Pilæ recticularis.* Basileæ, Impensis Thurnisiorum, fratrum.

Rényi, A. (1970). *Probability Theory.* North Holland, Amsterdam.

THÉORÈME DE CONVERGENCE

Convergence theorem

Les théorèmes de convergence constituent les résultats théoriques les plus importants des probabilités. Parmi eux, on trouve la **loi des grands nombres** et le **théorème central limite**.

EXEMPLES

Le **théorème central limite**, la **loi des grands nombres** sont des théorèmes de convergence.

La loi des grands nombres énonce les conditions sous lesquelles la **moyenne** d'une somme de variables aléatoires distribuées identiquement converge vers leur **espérance mathématique** commune.

En revanche, le théorème central limite détermine sous quelles hypothèses la distribution de la somme d'un nombre suffisamment grand de variables aléatoires tend vers la **loi normale**.

MOTS CLÉS

Loi des grands nombres *(Law of large numbers)*
Théorème central limite *(Central limit theorem)*

THÉORÈME DE GAUSS-MARKOV

Gauss-Markov theorem

Le théorème de Gauss-Markov stipule que lorsque la **loi de probabilité** des erreurs est inconnue, l'**estimateur** obtenu par la méthode des **moindres carrés** est, parmi tous les estimateurs linéaires sans **biais** des paramètres d'un **modèle** linéaire, celui qui a la **variance** minimale. L'**espérance mathématique** des erreurs est supposée nulle et leur variance constante et inconnue.

HISTORIQUE

C'est **Carl Friedrich Gauss** qui a donné la preuve de ce théorème dans la première partie de son ouvrage *Theoria Combinationis Observationum Erroribus Minimis Obnoxiæ* (1821). **Andreï Andreïevitch Markov** a redécouvert ce théorème en 1900.

Une version du théorème de Gauss-Markov, utilisant les notations modernes, est énoncée par Graybill (1976).

ASPECTS MATHÉMATIQUES

Soit le **modèle** linéaire :

$$\mathbf{Y} = \mathbf{X} \cdot \boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

où

- $\mathbf{Y}$ est le **vecteur** $n \times 1$ des observations,
- $\mathbf{X}$ la **matrice** $n \times p$ des variables indépendantes considérées comme fixes,
- $\boldsymbol{\beta}$ le vecteur $p \times 1$ des paramètres inconnus et
- $\boldsymbol{\varepsilon}$ le vecteur $n \times 1$ des erreurs aléatoires.

Si la **loi de probabilité** des erreurs est inconnue mais que les conditions suivantes sont remplies :

1. l'**espérance mathématique** $E[\boldsymbol{\varepsilon}] = \mathbf{0}$;
2. la **variance** $Var(\boldsymbol{\varepsilon}) = \sigma^2 \cdot \mathbf{I_n}$, avec $\mathbf{I_n}$ la **matrice identité** ;
3. la matrice $\mathbf{X}$ est de rang plein ;

alors l'**estimateur** de $\boldsymbol{\beta}$:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

provenant de la méthode des **moindres carrés** est l'estimateur linéaire sans **biais** de $\boldsymbol{\beta}$ qui a la variance minimale.

MOTS CLÉS

Estimateur *(Estimator)*
Moindres carrés *(Least squares)*

RÉFÉRENCES

Gauss, C.F. (1821, 1823 et 1826). Theoria Combinationis Observationum Erroribus Minimis Obnoxiæ, partie 1, 2 et suppl. *Werke* 4, pp. 1-108.

Graybill, F.A. (1976). *Theory and Applications of the Linear Model.* Duxbury Press, North Scituate, Massachusetts; Waldsworth and Brooks/Cole, Pacific Grove, CA.

Rao, C.R. (1973). *Linear Statistical Inference and Its Applications.* 2nd ed., John Wiley & Sons, New York.

THÉORÈME DE MOIVRE-LAPLACE

Moivre-Laplace theorem

Ce theorème établit un lien entre la **loi binomiale** avec paramètres n et p et la **loi normale** pour n qui tend vers l'infini. Son utilité consiste surtout à estimer la **fonction de répartition** $F(k)$ de la loi binomiale par la fonction de répartition de la loi normale, qui est plus facilement repérable.

HISTORIQUE

Ce théorème remonte environ à 1870, quand **Pierre Simon de Laplace** étudiait, avec d'autres (parmi lesquels **Abraham de Moivre**) des problèmes liés à l'approximation de la loi binomiale et à la théorie des erreurs.

ASPECTS MATHÉMATIQUES

Soit $X_1, X_2, \ldots, X_n, \ldots$ une séquence de variables aléatoires indépendantes et identiquement distribuées selon une **loi binomiale** avec paramètres n et p, où $0 < p < 1$.
On considère la variable ajustée correspondante Z_n, $n = 1, 2, \ldots$ telle que :

$$Z_n = \frac{X_n - np}{\sqrt{npq}},$$

où np représente la **moyenne** d'une loi binomiale et $\sqrt{npq}$ sa **variance** (on remarque que la variable Z_n correspond donc à la **normalisation** de X_n).

Le théorème de Moivre-Laplace s'exprime de la manière suivante :

$$P(Z_n \leq x) \overset{n \to \infty}{\longrightarrow} \Phi(x)$$

où Φ est la **fonction de répartition** de la **loi normale** centrée et réduite.

L'approximation peut encore être améliorée en utilisant le **facteur** de correction « un demi ». Donc de façon générale, l'approximation :

$$P\left(X_n \geq x\right) \cong 1 - \Phi\left(\frac{x - \frac{1}{2} - np}{\sqrt{npq}}\right)$$

est supérieure à celle obtenue sans tenir compte du facteur correctif $\frac{1}{2}$ qui figure au numérateur.

DOMAINES ET LIMITATIONS

Ce théorème est un cas particulier du **théorème central limite**. En effet, une **loi binomiale** (comme celle définie auparavant) peut être écrite comme une somme de n variables aléatoires distribuées selon une **loi de Bernoulli** de **paramètre** p. La **moyenne** d'une variable aléatoire Bernoulli étant égale à p et sa **variance** à $p\left(1-p\right)$, le théorème de Moivre-Laplace se déduit facilement du théorème central limite.

EXEMPLE

On désire calculer la **probabilité** d'obtenir plus de 27 succès dans une **expérience** binomiale de 100 épreuves, chaque épreuve ayant la probabilité de succès $p = 0,2$. La probabilité binomiale recherchée peut être exprimée par la somme :

$$P\left(X_n > 27\right) = \sum_{k=27}^{100} \left(\begin{array}{c} 100 \\ k \end{array} \right) (0,2)^k (0,8)^{100-k}.$$

Le calcul direct de cette probabilité demande l'évaluation de 74 termes, chacun de la forme

$$\left(\begin{array}{c} 100 \\ k \end{array} \right) (0,2)^k (0,8)^{100-k}.$$

En utilisant le théorème de Moivre-Laplace, cette probabilité peut être évaluée approximativement en tenant compte du lien existant entre les lois binomiale et normale. Donc :

$$\begin{aligned} P\left(X_n > 27\right) &\cong 1 - \Phi\left(\frac{27 - np}{\sqrt{npq}}\right) \\ &= 1 - \Phi\left(\frac{27 - 100 \cdot 0,2}{\sqrt{100 \cdot 0,2 \cdot 0,8}}\right) \\ &= 1 - \Phi\left(1,75\right). \end{aligned}$$

En se référant à la **table normale**, on obtient : $\Phi\left(1,75\right) = 0,9599$ et donc :

$$P\left(X_n > 27\right) \cong 0,0401.$$

La **valeur** exacte à 4 décimales près de la probabilité binomiale est : $P\left(X_n > 27\right) = 0,0558$.

La comparaison des deux valeurs $0,0401$ et $0,0558$ montre que l'approximation par la loi normale donne un résultat très proche de la valeur exacte.

On obtient en utilisant le facteur correctif, l'approximation suivante :

$$P\left(X_n > x\right) \cong 1 - \Phi\left(1,625\right) = 0,0521.$$

Ce résultat est effectivement plus proche de la valeur exacte $0,0558$ que celui obtenu sans l'utilisation du facteur correctif $(0,0401)$.

MOTS CLÉS

Loi binomiale *(Binomial distribution)*
Loi normale *(Normal distribution)*
Normalisation *(Normalization)*
Théorème central limite *(Central limit theorem)*

RÉFÉRENCES

Laplace, P.S. de (1810). Mémoire sur les approximations des formules qui sont fonctions de très grands nombres et sur leur application aux probabilités. *Mémoires de l'Académie royale des sciences de Paris*, 10. Reproduit dans *Œuvres de Laplace* 12, pp. 301-347.

Moivre, A. de (1733). Approximatio ad summam terminorum binomii $(a + b)^n$, in Seriem Expansi, Supplementum II to *Miscellanæ Analytica*, pp. 1-7. Photographically reprinted in *A rare pamphlet of Moivre and some of his discoveries.* Publié par Archibald R.C. (1926). *Isis* 8, pp. 671-683.

Sheynin, O.B. (1976). P.S. Laplace's work on probability, *Archive for History of Exact Sciences*, 16, pp. 137-187.

Sheynin, O.B. (1977). P.S. Laplace's theory of errors, *Archive for History of Exact Sciences*, 17, pp. 1-61.

TRACE

Trace

Pour une **matrice** carrée **A** d'ordre n, on définit la trace de **A** comme la somme des termes situés sur sa diagonale. La trace d'une matrice est donc un scalaire.

ASPECTS MATHÉMATIQUES

Soit **A** est une **matrice** carrée d'ordre n, $\mathbf{A} = (a_{ij})$ où $i, j = 1, 2, \ldots, n$. On définit la trace de **A** par :

$$tr(\mathbf{A}) = \sum_{i=1}^{n} a_{ii}.$$

DOMAINES ET LIMITATIONS

Dans la mesure où on ne s'intéresse qu'aux éléments diagonaux, la trace n'est définie que pour des matrices carrées.

1. La trace de la somme de deux matrices est égale à la somme des traces des matrices :

$$tr(\mathbf{A} + \mathbf{B}) = tr(\mathbf{A}) + tr(\mathbf{B}).$$

2. La trace du produit de deux matrices est inchangée sous commutation des matrices.
En terme mathématique :

$$tr(\mathbf{A} \cdot \mathbf{B}) = tr(\mathbf{B} \cdot \mathbf{A}).$$

EXEMPLE

Soient **A** et **B** deux matrices carrées d'ordre 2 :

$$\mathbf{A} = \begin{bmatrix} 1 & 2 \\ 1 & 3 \end{bmatrix} \text{ et } \mathbf{B} = \begin{bmatrix} 5 & 2 \\ -2 & 3 \end{bmatrix}.$$

Le calcul des traces donne :

$$tr(\mathbf{A}) = tr\begin{bmatrix} 1 & 2 \\ 1 & 3 \end{bmatrix} = 1 + 3 = 4$$

et

$$tr(\mathbf{B}) = tr\begin{bmatrix} 5 & 2 \\ -2 & 3 \end{bmatrix} = 5 + 3 = 8$$

$$tr(\mathbf{A} + \mathbf{B}) = tr\begin{bmatrix} 6 & 4 \\ -1 & 6 \end{bmatrix} = 6 + 6 = 12.$$

On vérifie que $tr(\mathbf{A}) + tr(\mathbf{B}) = 12$.

D'autre part,

$$\mathbf{A} \cdot \mathbf{B} = \begin{bmatrix} 1 & 8 \\ -1 & 11 \end{bmatrix} \text{ et } \mathbf{B} \cdot \mathbf{A} = \begin{bmatrix} 7 & 16 \\ 1 & 5 \end{bmatrix}$$

$$tr(\mathbf{A} \cdot \mathbf{B}) = tr\begin{bmatrix} 1 & 8 \\ -1 & 11 \end{bmatrix} = 1 + 11 = 12$$

$$tr(B \cdot \mathbf{A}) = tr\begin{bmatrix} 7 & 16 \\ 1 & 5 \end{bmatrix} = 7 + 5 = 12.$$

Ainsi $tr(\mathbf{A} \cdot \mathbf{B}) = tr(\mathbf{B} \cdot \mathbf{A}) = 12$.

MOT CLÉ

Matrice *(Matrix)*

TRAITEMENT

Treatment

Dans un **plan d'expérience**, un traitement est un choix particulier de niveaux pour chaque **facteur**.

EXEMPLES

Les expériences sont souvent menées pour comparer deux ou plusieurs traitements, par exemple deux engrais différents sur une espèce de plante particulière ou encore plusieurs sortes de médicaments pour traiter une certaine maladie. Un autre exemple consiste à mesurer les temps de coagulation d'échantillons de sang de 16 animaux ayant subi quatre régimes différents notés A, B, C, D.

Dans ces différents exemples les traitements sont respectivement les engrais, les médicaments et les régimes.

Par ailleurs, dans une expérience pour tester un engrais particulier, par exemple l'azote, sur une récolte de blé, on peut considérer différentes quantités du même engrais comme différents traitements. Ici, on a un seul **facteur** (azote) à différents niveaux (quantités), par exemple 30 kg, 50 kg, 100 kg

et 200 kg. Chacun de ces niveaux correspond à un traitement.

Le but d'une expérience est de déterminer s'il y a une différence réelle (significative) entre ces traitements.

MOTS CLÉS

Expérience *(Experiment)*
Facteur *(Factor)*
Niveau *(Level)*
Plan d'expérience *(Design of experiments)*
Unité expérimentale *(Experimental unit)*

TRANSFORMATION

Transformation

Une transformation est un changement d'unités d'une ou de plusieurs variables d'une étude statistique.

On transforme les variables, par exemple, en les remplaçant par leurs logarithmes (**transformation logarithmique**).

DOMAINES ET LIMITATIONS

On transforme les variables pour plusieurs raisons :

1. Lorsque l'on se trouve en présence d'un ensemble de données relatives à plusieurs variables et que l'on souhaite exprimer l'une d'entre elles, Y (**variable dépendante**), à l'aide des autres (dites variables indépendantes) et on désire linéariser la **relation** entre les variables. Par exemple, lorsque la relation entre les variables indépendantes est multiplicative, une **transformation logarithmique** des données la rend additive et permet l'utilisation de la régression linéaire pour estimer les paramètres du **modèle**.

2. Lorsque l'on se trouve en présence d'un ensemble de données relatives à une variable et que la **variance** de celle-ci n'est pas constante, il est, en général possible de rendre à peu près constante la variance de cette dernière à l'aide d'une transformation de cette variable.

3. Lorsqu'une **variable aléatoire** suit une loi quelconque, une transformation adéquate de celle-ci permet d'obtenir une nouvelle variable aléatoire distribuée approximativement suivant une **loi normale**.

Les avantages d'une transformation de variables se résument donc principalement à :

- simplifier la relation existant avec d'autres variables ;
- rendre la variance constante ;
- obtenir une variable distribuée approximativement suivant une loi normale.

MOTS CLÉS

Analyse de régression *(Regression analysis)*
Loi normale *(Normal distribution)*
Transformation logarithmique *(Logarithmic transformation)*
Variable dépendante *(Dependent variable)*
Variable indépendante *(Independent variable)*
Variance *(Variance)*

TRANSFORMATION LOGARITHMIQUE

Logarithmic transformation

Une **transformation** parmi les plus utilisées est la transformation logarithmique. En effet, avant d'effectuer une régression linéaire, il peut être judicieux de remplacer la **variable dépendante** Y par son logarithme $\log(Y)$ ou par une combinaison linéaire de ce dernier $a + b \cdot \log(Y)$ (a et b étant deux constantes ajustées par les **moindres carrés**), afin de stabiliser la **variance** de Y et de rendre la distribution de la **variable** transformée plus proche de la **loi normale**.

Une transformation logarithmique permet aussi de passer d'un **modèle** multiplicatif à un modèle additif, dont les paramètres sont plus simples à déterminer.

Par ailleurs, la transformation logarithmique d'une variable Y distribuée aléatoirement selon une **loi gamma** fournit une autre **variable aléatoire** plus proche de la **loi normale** que la variable initiale Y, qui s'adaptera donc mieux à une **analyse de régression**.

MOTS CLÉS

Analyse de régression *(Regression analysis)*
Analyse de variance *(Analysis of variance)*
Variable dépendante *(Dependent variable)*

TRANSPOSÉE

Transpose

La (matrice) transposée d'une **matrice A** d'ordre $(m \times n)$ est une matrice $(n \times m)$ notée $\mathbf{A}'$ obtenue en écrivant les lignes de **A** en colonnes de $\mathbf{A}'$ et vice-versa.

On effectue donc simplement un « renversement » du tableau.

ASPECTS MATHÉMATIQUES

Si $\mathbf{A} = (a_{ij})$, $i = 1, 2, \ldots, m$ et $j = 1, 2, \ldots, n$ est une **matrice** d'ordre $(m \times n)$, la transposée de **A** est la matrice $\mathbf{A}'$ d'ordre $(n \times m)$ donnée par :

$$\mathbf{A}' = (a_{ji}) \text{ avec} j = 1, 2, \ldots, n \text{ et } i = 1, 2, \ldots, m.$$

La transposition est une opération réversible et réciproque ; ainsi en prenant la transposée de la transposée d'une matrice, on retrouve la matrice initiale.

DOMAINES ET LIMITATIONS

On utilise la transposée d'une **matrice**, ou plus spécialement celle d'un **vecteur** lors du calcul du **produit scalaire**.

EXEMPLE

Soit **A** la **matrice** d'ordre (3×2) suivante :

$$\mathbf{A} = \begin{bmatrix} 1 & 2 \\ 0 & 3 \\ 2 & 5 \end{bmatrix}$$

la transposée de **A** est la matrice (2×3) :

$$\mathbf{A}' = \begin{bmatrix} 1 & 0 & 2 \\ 2 & 3 & 5 \end{bmatrix}.$$

On peut aussi vérifier que la transposée de $\mathbf{A}'$ est égale à la matrice initiale **A**,

$$(\mathbf{A}')' = \mathbf{A}.$$

MOTS CLÉS

Matrice *(Matrix)*
Produit scalaire *(Scalar product)*
Vecteur *(Vector)*

TRIANGLE ARITHMÉTIQUE

Arithmetic triangle

On appelle triangle arithmétique une méthode itérative pour déterminer les coefficients du **binôme** obtenus en calculant le nombre de combinaisons de k objets parmi n et que l'on note C_n^k.

HISTORIQUE

La notion du nombre de combinaisons de k objets parmi n était une identité connue en Inde dès le IX^{e} siècle. On en retrouve déjà la trace dans le *Meru Prastara* de Pingala (environ 200 av. J.-C.).

Entre le XIV^{e} et le XV^{e} siècle, Kashi, un mathématicien originaire de la ville iranienne de Kachan, écrit la *Clé de l'arithmétique*, un ouvrage dans lequel il appelle les coefficients du **binôme** « éléments des exposants ».

Dans son *Traité du triangle arithmétique* publié en 1665, Blaise Pascal (1654) définit les nombres dans le « triangle arithmétique » qui porte aujourd'hui parfois son nom : triangle de Pascal.

Il faut noter que ce triangle fut rendu populaire par Niccolo Fontana Tartaglia en 1556 et les Italiens se réfèrent au triangle de Tartaglia plutôt qu'au triangle de Pascal, même si N.F. Tartaglia n'a pas fait d'études particulières sur le triangle arithmétique.

ASPECTS MATHÉMATIQUES

Le triangle arithmétique se présente sous la forme suivante :

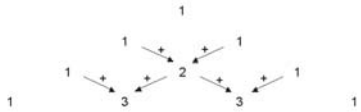
```
                    1
                  1   1
                1   2   1
              1   3   3   1
            1   4   6   4   1
          1   5   10  10  5   1
        1   6   15  20  15  6   1
                ......
```

Chaque élément est le coefficient du **binôme**

$$C_n^k = \frac{n!}{k!\,(n-k)!} = \frac{n\cdot(n-1)\cdot\ldots\cdot(n-k+1)}{1\cdot 2\cdot\ldots\cdot k}$$

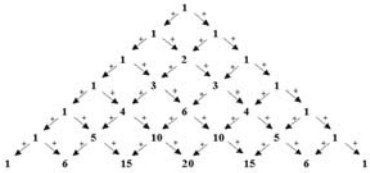
correspondant à l'élément k de la ligne $n+1$, $k = 0, \ldots, n$.

Les nombres d'une ligne s'obtiennent en additionnant les deux nombres adjacents de la ligne précédente.



Par exemple :

$$C_6^4 = C_5^3 + C_5^4 = 10 + 5 = 15$$



Plus généralement, on a la **relation** :

$$C_n^k + C_n^{k+1} = C_{n+1}^{k+1}$$

puisque :

$$\begin{aligned} C_n^k + C_n^{k+1} &= \frac{n!}{(n-k)!\cdot k} + \frac{n!}{(n-k-1)!\cdot(k+1)!} \\ &= \frac{n!\cdot[(k+1)+(n-k)]}{(n-k)!\cdot(k+1)!} \\ &= \frac{(n+1)!}{(n-k)!\cdot(k+1)!} \\ &= C_{n+1}^{k+1}. \end{aligned}$$

MOTS CLÉS

Analyse combinatoire *(Combinatory analysis)*
Binôme *(Binomial)*
Combinaison *(Combination)*
Loi binomiale *(Binomial distribution)*

RÉFÉRENCES

Pascal, B. (1654). *Traité du triangle arithmétique.* (publ. posthume en 1665), Paris.

Pascal, B. (1904-1925). *Œuvres.* Vol. 1-14. L. Brunschvicg, P. Boutroux et F. Gazier (éd.) « Les Grands Écrivains de France », Hachette, Paris.

Pascal, B. (1970). *Œuvres complètes.* Vol. 2. J. Mesnard, éd. Desclée de Brouwer, Paris.

Rashed, R. (1985). La naissance de l'algèbre, in *Le Matin des mathématiciens*, Émile Noël éditeur. Belin-Radio France. (chapitre 12).

Youschkevitch, A.P. (1976). *Les Mathématiques arabes (*VIII^{e}-XV^{e} *siècles).* Traduction partielle par M. Cazenave et K. Jaouiche. Vrin, Paris.

TUKEY JOHN WILDER

Tukey John Wilder

John Wilder Tukey (1915-2000) a étudié la chimie à l'Université de Brown puis obtenu, en 1939, un doctorat en mathématiques à l'Université de Princeton. À l'âge de 35 ans, il devint professeur de mathématiques dans cette même université. Il a dirigé

le groupe de recherches en techniques statistiques de l'Université de Princeton depuis sa formation en 1956. Il fut aussi nommé premier directeur du département de statistique de l'Université de Princeton, en 1965.



John Wilder Tukey a ouvert la voie dans les domaines de « l'analyse exploratoire des données » et des estimations robustes. Ses contributions dans les domaines de l'analyse des séries chronologiques ainsi que dans l'analyse spectrale ont été largement utilisées dans les sciences appliquées. Auteur de *Exploratory Data Analysis* (1977) (actuellement traduit en russe) et d'un recueil d'articles en huit volumes, il est aussi co-auteur de *Statistical Problems of the Kinsey Report on Sexual Behavior in the Human Male*, *Data Analysis and Regression* (aussi traduit en russe), *Index to Statistics and Probability. Citation index ; Permuted Titles ; Locations et Authors* (en quatre volumes) ; *The Measurement of Power Spectra* et *Robust Estimates of Location : Survey and Advances.* Il est co-éditeur des livres suivants : *Understanding Robust and Exploratory Data Analysis*, *Exploring Data Tables*, *Trends Shapes*, *Configural Polysampling*, et *Fundamentals of Exploratory Analysis of Variance.*

Quelques ouvrages et articles principaux de John Wilder Tukey :

1977 *Exploratory Data Analysis.* First Edition. Addison-Wesley, Reading, Mass.

1980 We need both exploratory and confirmatory. *The American Statistician*, 34, pp. 23-25.

1986 Data analysis and behavioral science or learning to bear the quantitative man's burden by shunning badmandments. *The Collected Works of John W. Tukey, Volume III : Philosophy and Principles of Data Analysis, 1949-1964.* (L.V. Jones, ed.) Wadsworth Advanced Books & Software, Monterey, CA.

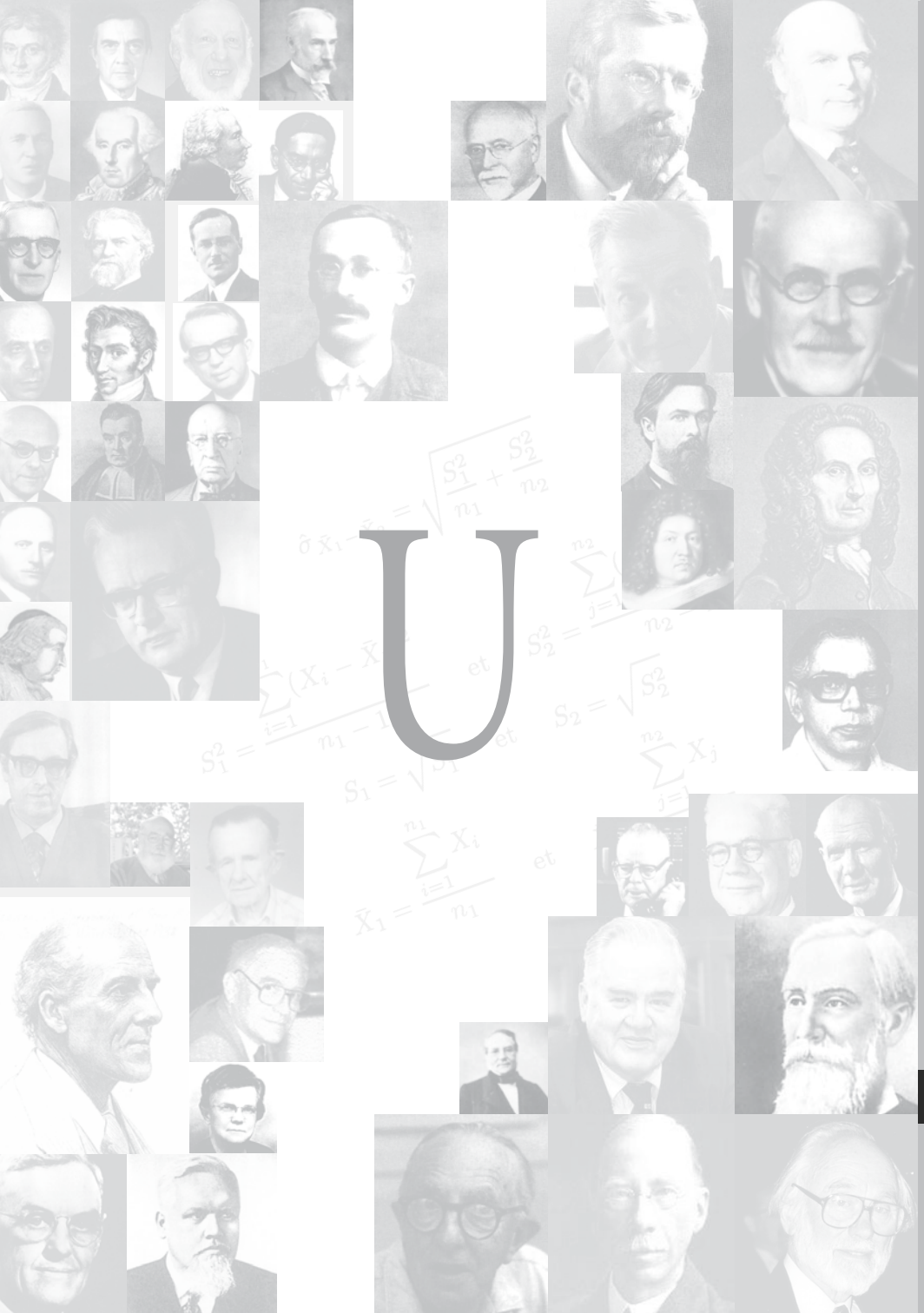
1986 *The Collected Works of John W. Tukey, Volume III : Philosophy and Principles of Data Analysis, 1949-1964.* (L.V. Jones, ed.) Wadsworth Advanced Books & Software, Monterey, CA.

1986 *The Collected Works of John W. Tukey, Volume IV : Philosophy and Principles of Data Analysis, 1965-1986.* (L.V. Jones, ed.) Wadsworth Advanced Books & Software, Monterey, CA.

1988 *The Collected Works of John W. Tukey, Volume V : Graphics, 1965-1985.* (W.S. Cleveland, ed.) Waldsworth and Brooks/Cole, Pacific Grove, CA.

MOTS CLÉS

Analyse exploratoire des données *(Exploratory data analysis)*
Box plot *(Box plot)*
Diagramme stem and leaf *(Stem and leaf diagram)*



U

UNIFORME, LOI

Uniform distribution

Voir **loi uniforme**.

UNIFORME DISCRÈTE, LOI

Discrete uniform distribution

Voir **loi uniforme discrète**.

UNILATÉRAL, TEST

One-tail test

Voir **test unilatéral**.

UNITÉ EXPÉRIMENTALE

Experimental unit

On appelle unité expérimentale la plus petite partie du matériel expérimental sur laquelle est appliqué un **traitement** déterminé.

Un **plan d'expérience** spécifie le nombre d'unités expérimentales soumises aux différents traitements.

MOTS CLÉS

Expérience *(Experiment)*
Facteur *(Factor)*
Niveau *(Level)*
Plan d'expérience *(Design of experiments)*
Traitement *(Treatment)*



V

VALEUR

Value

Une valeur est une mesure quantitative ou qualitative associée à une **variable**.

On parle de l'ensemble des valeurs prises par une variable.

Les valeurs peuvent être des nombres cardinaux si elles sont associées à une **variable quantitative**, ou des nombres nominaux si elles sont associées à une **variable qualitative catégorielle**.

EXEMPLE

Les valeurs prises par une certaine **variable** peuvent être du type suivant :

- soit la **variable quantitative** « nombre d'enfants ». Les valeurs associées à cette variable sont les nombres naturels 0, 1, 2, 3, 4, etc. ;
- soit la **variable qualitative catégorielle** « sexe ». Les valeurs associées sont les catégories « masculin » et « féminin ».

MOTS CLÉS

Variable *(Variable)*
Variable qualitative catégorielle *(Qualitative categorical variable)*
Variable quantitative *(Quantitative variable)*

VALEUR CRITIQUE

Critical value

La valeur critique d'un **test d'hypothèse** est la **valeur** limite à partir de laquelle nous pouvons prendre la décision de rejeter l'**hypothèse nulle** H_0, à un **seuil de signification** donné.

HISTORIQUE

La notion de valeur critique a été introduite par **Jerzy Neyman** et **Egon Sharpe Pearson** en 1928.

ASPECTS MATHÉMATIQUES

La valeur critique dépend du type de test effectué (**test bilatéral** ou **test unilatéral** à droite ou à gauche), de la **loi de probabilité** et du **seuil de signification** α.

DOMAINES ET LIMITATIONS

La valeur critique est déterminée à partir de la **loi de probabilité** de la **statistique** associée au test. Elle se détermine en consultant la **table statistique** correspondant à cette loi de probabilité (**table normale**, **table de Student**, **table de Fisher**, **table du chi-carré**, etc.).

EXEMPLE

Une compagnie fabrique des câbles d'acier. Elle veut vérifier sur un **échantillon** de taille $n = 100$ si le diamètre des câbles est conforme aux normes qui stipulent un diamètre de $0,9$ cm.

L'**écart type** σ de la **population** est connu et vaut $0,05$ cm.

Le **test d'hypothèse** est dans ce cas un **test bilatéral**. Les hypothèses sont les suivantes :

$$\begin{array}{lcl} \textbf{hypothèse nulle} & H_0 & : \quad \mu = 0,9, \\ \textbf{hypothèse alternative} & H_1 & : \quad \mu \neq 0,9. \end{array}$$

À un **seuil de signification** de $\alpha = 5\,\%$, on trouve dans la **table normale** la valeur critique $z_{\frac{\alpha}{2}}$ égale à $1,96$.

MOTS CLÉS

Intervalle de confiance *(Confidence interval)*
Seuil de signification *(Significance level)*
Table statistique *(Statistical table)*
Test d'hypothèse *(Hypothesis testing)*

RÉFÉRENCE

Neyman, J. and Pearson, E.S. (1928). On the use and interpretation of certain test criteria for purposes of statistical inference. I, II. *Biometrika*, 20A, pp. 175-240 and 263-294.

VALEUR p

P-value

La valeur p est définie comme la **probabilité**, calculée sous l'**hypothèse nulle**, d'obtenir un résultat aussi extrême que la **valeur** observée sur l'**échantillon** (dans une direction particulière, c'est-à-dire soit plus petit ou égal, soit plus grand ou égal).

HISTORIQUE

La valeur p a été introduite par Jean D. Gibbons et John W. Pratt en 1975.

ASPECTS MATHÉMATIQUES

Illustrons le cas d'un **test d'hypothèse** effectué sur l'**estimateur** de **moyenne** ; le même principe est applicable pour tout autre estimateur, seules les notations sont différentes.

Supposons que nous souhaitons tester les hypothèses suivantes :

$$\begin{array}{rcl} H_0 & : & \mu = \mu_0, \\ H_1 & : & \mu > \mu_0. \end{array}$$

où μ représente la moyenne d'une **population** distribuée selon une **loi normale** avec un **écart type** σ connu. Un **échantillon** de taille n fournit une moyenne observée $\bar{x}$.

Considérons le cas où $\bar{x}$ dépasse μ_0 et calculons la **probabilité** d'obtenir une **estimation** $\hat{\mu}$ supérieure ou égale à $\bar{x}$ sous l'**hypothèse nulle** $\mu = \mu_0$. La valeur p correspond à cette probabilité. Ainsi :

$$p = P(\hat{\mu} \geq \bar{x} | \mu = \mu_0).$$

La **variable aléatoire** Z centrée réduite donnée par $Z = \dfrac{\hat{\mu} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$ suit une loi normale de moyenne 0 d'**écart type** 1. En introduisant cette **variable** dans l'expression de la valeur p, on trouve :

$$p = P\left(Z \geq \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}}\right)$$

qui sous cette forme peut se lire dans la **table normale**.

Pour un **seuil de signification** α, la comparaison de p à α permet de prendre une décision quant à un éventuel rejet de l'**hypothèse nulle**. En effet, si

- $p \leq \alpha$, nous rejetons l'hypothèse nulle H_0 en faveur de l'**hypothèse alternative** H_1 ;
- $p > \alpha$, nous ne rejetons pas l'hypothèse nulle H_0.

Nous pouvons également calculer la borne supérieure de la **région d'acceptation** de H_0, à savoir la **valeur** C_α telle que

$$P(\hat{\mu} \geq C_\alpha | \mu = \mu_0) = \alpha.$$

En fonction de la variable aléatoire normale centrée réduite, z_α, la valeur C_α est exprimée par :

$$C_\alpha = \mu_0 + z_\alpha \cdot \frac{\sigma}{\sqrt{n}}.$$

DOMAINES ET LIMITATIONS

Nous utilisons fréquemment la valeur p pour rendre compte du résultat d'un **test d'hypothèse**.

Elle s'utilise pour un **test unilatéral**, quand les hypothèses se présentent sous la forme :

$$\begin{array}{rcl} H_0 & : & \mu = \mu_0 \\ H_1 & : & \mu > \mu_0 \end{array} \quad \text{ou} \quad \begin{array}{rcl} H_0 & : & \mu = \mu_0 \\ H_1 & : & \mu < \mu_0 \end{array}$$

ou

$$\begin{array}{rcl} H_0 & : & \mu \leq \mu_0 \\ H_1 & : & \mu > \mu_0 \end{array} \quad \text{ou} \quad \begin{array}{rcl} H_0 & : & \mu \geq \mu_0 \\ H_1 & : & \mu < \mu_0. \end{array}$$

La valeur p peut être interprétée comme le plus bas **seuil de signification** pour lequel l'**hypothèse nulle** ne peut pas être rejetée.

EXEMPLE

Supposons que nous souhaitions effectuer un **test d'hypothèse** unilatéral

$$\begin{array}{rcl} H_0 & : & \mu = 30 \\ \text{contre } H_1 & : & \mu > 30 \end{array}$$

pour une **population** distribuée selon une **loi normale** avec un **écart type** $\sigma = 8$. Un **échantillon** de **taille** $n = 25$ a fourni une **moyenne** observée de 34.

Comme $\bar{x}$ $(= 34)$ dépasse $\mu = 30$ de façon assez nette, l'on serait tenté d'accepter l'**hypothèse alternative**, à savoir $\mu > 30$.

Calculons la **probabilité** d'obtenir une moyenne observée de 34 sous l'**hypothèse nulle** $\mu = 30$. Il s'agit de la valeur p :

$$\begin{aligned} p &= P(\bar{X} \geq 34 | \mu = 30) \\ &= P\left(\frac{\bar{X} - 30}{\frac{8}{5}} \geq \frac{34 - 30}{\frac{8}{5}}\right) \\ &= P(Z \geq 2,5) \end{aligned}$$

où $Z = \dfrac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$ suit une loi normale de moyenne 0 et d'écart type 1 ;

$p = 0,0062$ d'après la **table normale**.

Pour un **seuil de signification** $\alpha = 0,01$, nous rejetons l'hypothèse nulle en faveur de l'hypothèse alternative $\mu > 30$.

Nous pouvons calculer la borne supérieure de la **région d'acceptation** de H_0, à savoir C_α tel que

$$P(\bar{X} \geq C_\alpha | \mu = 30) = 0,01$$

et trouvons $C_\alpha = 30 + 1,64 \cdot \dfrac{8}{5} = 32,624$.

Cela confirme que nous devons en effet rejeter l'hypothèse nulle $\mu = 30$ quand la moyenne observée est égale à 34.

MOTS CLÉS

Région d'acceptation *(Acceptance region)*
Seuil de signification *(Significance level)*
Table normale *(Normal table)*
Test d'hypothèse *(Hypothesis testing)*

RÉFÉRENCE

Gibbons, J.D. and Pratt, J.W. (1975). P-values: Interpretation and Methodology. *The American Statistician*, 29, pp. 20-25.

VALEUR PROPRE

Eigenvalue

Soit $\mathbf{A}$ une **matrice** carrée d'ordre n. S'il existe un **vecteur** $\mathbf{x} \neq \mathbf{0}$ qui en le multipliant par $\mathbf{A}$ donne un multiple $(k \cdot \mathbf{x})$ de lui-même, ce vecteur est appelé **vecteur propre** et le facteur multiplicatif k est appelé valeur propre de la matrice $\mathbf{A}$.

ASPECTS MATHÉMATIQUES

Soit $\mathbf{A}$ une **matrice** carrée d'ordre n et $\mathbf{x}$ un **vecteur propre** non nul, k est valeur propre de $\mathbf{A}$ si

$$\mathbf{A} \cdot \mathbf{x} = k \cdot \mathbf{x}$$

c'est-à-dire : $(\mathbf{A} - k \cdot \mathbf{I}_n) \cdot \mathbf{x} = \mathbf{0}$, où $\mathbf{I}_n$ est la **matrice identité** d'ordre n.

Comme $\mathbf{x}$ est non nul, on détermine les valeurs possibles de k en trouvant les solutions de l'équation $|\mathbf{A} - k \cdot \mathbf{I}_n| = 0$.

DOMAINES ET LIMITATIONS

Pour une **matrice** $\mathbf{A}$ carrée d'ordre n, le **déterminant** $|\mathbf{A} - k \cdot \mathbf{I}_n|$ est un polynôme de degré n en k et possède donc au plus n solutions non nécessairement distinctes. Il y aura alors au maximum n valeurs propres, certaines pouvant être confondues.

EXEMPLE

Considérons la **matrice** carrée d'ordre 3 suivante :

$$\mathbf{A} = \begin{bmatrix} 4 & 3 & 2 \\ 0 & 1 & 0 \\ -2 & 2 & 0 \end{bmatrix}$$

$$\mathbf{A} - k \cdot \mathbf{I}_3 = \begin{bmatrix} 4 & 3 & 2 \\ 0 & 1 & 0 \\ -2 & 2 & 0 \end{bmatrix} - k \cdot \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

$$\mathbf{A} - k \cdot \mathbf{I}_3 = \begin{bmatrix} 4-k & 3 & 2 \\ 0 & 1-k & 0 \\ -2 & 2 & -k \end{bmatrix}.$$

Le **déterminant** de $\mathbf{A} - k \cdot \mathbf{I}_3$ est donné par :

$$|\mathbf{A} - k \cdot \mathbf{I}_3| = (1-k) \cdot [(4-k) \cdot (-k) - 2 \cdot (-2)]$$

$$\begin{aligned} &= (1-k)\cdot(k^2-4\cdot k+4) \\ &= (1-k)\cdot(k-2)^2 \end{aligned}$$

Les valeurs propres sont obtenues en trouvant les solutions de l'équation $(1-k)\cdot(k-2)^2=0$. On trouve

$$k_1 = 1$$

$$\text{et} \quad k_2 = k_3 = 2.$$

Deux des trois valeurs propres sont donc confondues.

MOTS CLÉS

Déterminant *(Determinant)*
Matrice*(Matrix)*
Vecteur *(Vector)*
Vecteur propre *(Eigenvector)*

VARIABLE

Variable

Une variable est une caractéristique mesurable à laquelle on peut attribuer plusieurs valeurs différentes.

On distingue plusieurs types de variables :

- **variable quantitative**, dont on distingue
  - variable discrète, par exemple le résultat à un test,
  - variable continue, par exemple le poids ou le revenu ;
- **variable qualitative catégorielle**, par exemple l'état-civil.

MOTS CLÉS

Donnée *(Data)*
Valeur *(Value)*
Variable aléatoire*(Random variable)*
Variable dichotomique*(Dichotomous variable)*
Variable qualitative catégorielle *(Qualitative categorical variable)*
Variable quantitative *(Quantitative variable)*

VARIABLE ALÉATOIRE

Random variable

Une **variable** dont la **valeur** est déterminée en fonction du résultat d'une **expérience aléatoire** est appelée variable aléatoire.

ASPECTS MATHÉMATIQUES

Une variable aléatoire est généralement désignée par une des dernières lettres (majuscule) de l'alphabet. C'est une fonction à valeurs réelles définie sur l'**ensemble fondamental** Ω. Autrement dit, une variable aléatoire réelle X est une application de Ω dans $\mathbb{R}$:

$$X : \Omega \to \mathbb{R}\,.$$

On parle de :

- variable aléatoire discrète, si l'ensemble des valeurs prises par la variable aléatoire est un ensemble fini ou infini dénombrable ;
- variable aléatoire continue, si l'ensemble des valeurs prises par la variable aléatoire est un ensemble infini non dénombrable.

EXEMPLES

Exemple de variable aléatoire discrète

Si l'on jette deux dés simultanément, les 36 résultats que l'on peut obtenir et qui constituent l'**ensemble fondamental** Ω sont les suivants :

$$\begin{aligned} \Omega = \{&(1,1),(1,2),\ldots,(1,6),(2,1),(2,2),\ldots, \\ &(2,6),\ldots,(6,1),(6,2),\ldots,(6,6)\}. \end{aligned}$$

La variable aléatoire X = « somme des deux dés » est une variable aléatoire discrète. Elle prend ses valeurs dans l'ensemble fini $E = \{2,3,4,\ldots,11,12\}$, c'est-à-dire que la somme des deux dés peut être 2, 3, 4,... ou 12. Dans cet exemple, X est une variable aléatoire discrète entière positive et l'application se fait de Ω dans E.

Il est possible d'attribuer une **probabilité** aux

différentes valeurs que la variable aléatoire X peut prendre.

Ainsi la valeur « somme égale à 2 » n'est obtenue que par le résultat $(1, 1)$. En revanche, la valeur « somme égale à 8 » est obtenue par les résultats $(2, 6)$, $(3, 5)$, $(4, 4)$, $(5, 3)$ et $(6, 2)$.

Comme l'ensemble fondamental Ω contient 36 résultats équiprobables, on peut dire que la « chance » ou probabilité d'obtenir une « somme égale à 2 » est de $\frac{1}{36}$. La probabilité d'obtenir une « somme égale à 8 » est de $\frac{5}{36}$, etc.

Exemples de variable aléatoire continue

De telles variables aléatoires caractérisent généralement une mesure : par exemple la durée d'un appel téléphonique définie sur l'intervalle $]0, \infty[$, la direction du vent définie sur l'intervalle $[0, 360[$, etc.

MOTS CLÉS

Ensemble fondamental *(Sample space)*
Expérience aléatoire *(Random experiment)*
Fonction de densité *(Density function)*
Probabilité *(Probability)*
Valeur *(Value)*

VARIABLE DÉPENDANTE

Dependent variable

La variable dépendante (ou variable réponse) dans un **modèle** d'analyse de régression est la **variable** considérée comme variant en fonction d'autres variables de l'analyse.

MOTS CLÉS

Régression, analyse de *(Regression analysis)*
Régression linéaire multiple *(Multiple linear regression)*
Régression linéaire simple *(Simple linear regression)*
Variable indépendante *(Independent variable)*

VARIABLE DICHOTOMIQUE

Dichotomous variable

Une **variable** est dite dichotomique si elle ne peut prendre que deux valeurs.

L'exemple le plus simple est celui de la **variable qualitative catégorielle** « sexe » qui possède les deux seules valeurs « masculin » et « féminin ».

Notons que les variables quantitatives peuvent toujours être réduites et dichotomisées. La variable « revenu » peut, par exemple, être réduite à deux catégories : « bas revenu » et « haut revenu ».

MOTS CLÉS

Catégorie *(Category)*
Donnée binaire *(Binary data)*
Variable *(Variable)*
Variable qualitative catégorielle *(Qualitative categorical variable)*

VARIABLE INDÉPENDANTE

Independent variable

On appelle variable indépendante dans un **modèle** d'analyse de régression la ou les variables qui sont considérées comme exercant une influence sur la **variable dépendante**, ou qui expliquent les variations de la variable dépendante.

MOTS CLÉS

Régression, analyse de *(Regression analysis)*
Régression linéaire multiple *(Multiple linear regression)*
Régression linéaire simple *(Simple linear regression)*
Variable dépendante *(Dependent variable)*

VARIABLE QUALITATIVE CATÉGORIELLE

Qualitative categorical variable

Une **variable** qualitative catégorielle est une variable dont les modalités sont des catégories, telles que, par exemple, les catégories « homme » et « femme » de la variable « sexe » ; ou les catégories « rouge », « orange », « vert », « bleu », « indigo » et « violet » de la variable « couleur ».

Les modalités d'une variable qualitative catégorielle peuvent être représentées sur une échelle nominale ou sur une échelle ordinale.

Un exemple d'une variable qualitative catégorielle ayant une échelle ordinale est la variable qualificative professionnelle avec catégories « qualifié », « semi-qualifié » et « non qualifié ».

MOTS CLÉS

Catégorie *(Category)*
Variable *(Variable)*

VARIABLE QUANTITATIVE

Quantitative variable

Une variable quantitative est une **variable** dont les modalités ont des valeurs numériques.

Par exemple, le poids, la taille, l'âge, la vitesse ou le temps sont des variables quantitatives.

On distingue les variables discrètes (par exemple le nombre d'enfants par famille) des variables continues (par exemple la longueur d'un saut).

MOT CLÉ

Variable *(Variable)*

VARIANCE

Variance

La variance est une **mesure de dispersion** d'une distribution d'une **variable aléatoire**.

Empiriquement, la variance d'une **variable quantitative** X est définie comme la somme des déviations au carré de chaque **observation** par rapport à la **moyenne arithmétique**, divisée par le nombre d'observations.

HISTORIQUE

L'**analyse de variance** telle que nous l'entendons et pratiquons de nos jours a été développée principalement par **Ronald Aylmer Fisher** (1918, 1925 et 1935). Il introduisit également les termes de variance et d'analyse de variance.

ASPECTS MATHÉMATIQUES

La variance est généralement désignée par S^2 lorsqu'elle est relative à un **échantillon**, et σ^2 lorsqu'elle est relative à une **population**.

Nous notons aussi la variance $Var(X)$ lorsque nous parlons de la **variance d'une variable aléatoire**.

Soit une population de N observations relatives à une **variable quantitative** X. Selon la définition, la variance d'une population se calcule comme suit :

$$\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N},$$

N étant la taille de la population et μ la **moyenne** des observations :

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N}.$$

Lorsque les observations sont ordonnées sous forme d'une **distribution de fréquences**, le calcul de la

variance se fait de la façon suivante :

$$\sigma^2 = \frac{\sum_{i=1}^{k} f_i \cdot (x_i - \mu)^2}{\sum_{i=1}^{k} f_i}$$

où

- x_i sont les différentes valeurs de la **variable**,
- f_i sont les fréquences associées à ces valeurs et
- k est le nombre de valeurs différentes.

Pour calculer la variance de la distribution de fréquences d'une variable quantitative X où les valeurs sont groupées en classes, nous considérons que toutes les observations appartenant à une certaine classe prennent la valeur du centre de la classe. Toutefois, ceci n'est correct que si l'**hypothèse** spécifiant que les observations sont uniformément réparties à l'intérieur de chaque classe est vérifiée. Si cette hypothèse n'est pas vérifiée, la valeur obtenue de la variance ne sera qu'approximative.

Pour des valeurs groupées en classes, nous avons :

$$\sigma^2 = \frac{\sum_{i=1}^{k} f_i \cdot (\delta_i - \mu)^2}{\sum_{i=1}^{k} f_i}$$

où

- δ_i sont les centres des classes,
- f_i sont les fréquences associées à chaque classe et
- k est le nombre de classes.

La formule du calcul de la variance peut être modifiée en vue de diminuer le temps de calcul et d'augmenter la précision. Nous calculons donc de préférence la variance par la formule suivante :

$$\sigma^2 = \frac{N \cdot \sum_{i=1}^{N} x_i^2 - \left(\sum_{i=1}^{N} x_i\right)^2}{N^2}.$$

La variance correspond donc au **moment** centré d'ordre deux.

La variance mesurée sur un échantillon est un **estimateur** de la variance de la population.

Considérons un échantillon de n observations relatives à une variable quantitative X et dénotons par $\bar{x}$ la **moyenne arithmétique** de cet échantillon :

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}.$$

Pour que la variance de l'échantillon soit un estimateur sans **biais** de celle de la population, elle doit se calculer en divisant par $(n-1)$ et non par n :

$$S^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}.$$

EXEMPLE

Cinq étudiants ont passé successivement deux examens auxquels ils ont obtenu les notes suivantes :

Examen 1 :

3,5	4	4,5	3,5	4,5

Examen 2 :

2,5	5,5	3,5	4,5	4

Nous notons que la **moyenne arithmétique** $\bar{x}$ de ces deux ensembles d'observations est identique :

$$\bar{x}_1 = \bar{x}_2 = 4.$$

Toutefois, la dispersion des observations autour de la **moyenne** n'est pas la même.

Pour déterminer la variance, nous calculons tout d'abord les déviations de chaque observation par rapport à la moyenne arithmétique, puis ces déviations sont élevées au carré :

Examen 1

Notes	$(x_i - \bar{x}_1)$	$(x_i - \bar{x}_1)^2$
3,5	−0,5	0,25
4	0,0	0,00
4,5	0,5	0,25
3,5	−0,5	0,25
4,5	0,5	0,25
$\sum_{i=1}^{5} (x_i - \bar{x}_1)^2 =$		1,00

Examen 2

Notes	$(x_i - \bar{x}_2)$	$(x_i - \bar{x}_2)^2$
$2,5$	$-1,5$	$2,25$
$5,5$	$1,5$	$2,25$
$3,5$	$-0,5$	$0,25$
$4,5$	$0,5$	$0,25$
4	$0,0$	$0,00$
$\sum_{i=1}^{5} (x_i - \bar{x}_2)^2 =$		$5,00$

La variance des notes de chaque examen est donc égale à :

Examen 1 :

$$S_1^2 = \frac{\sum_{i=1}^{5} (x_i - \bar{x}_1)^2}{n-1} = \frac{1}{4} = 0,25.$$

Examen 2 :

$$S_2^2 = \frac{\sum_{i=1}^{5} (x_i - \bar{x}_2)^2}{n-1} = \frac{5}{4} = 1,25.$$

La variance des notes du deuxième examen étant plus grande, cela signifie qu'au deuxième examen les notes sont plus dispersées autour de la moyenne arithmétique qu'au premier.

MOTS CLÉS

Analyse de variance *(Analysis of variance)*
Écart type *(Standard deviation)*
Mesure de dispersion *(Measure of dispersion)*
Variance d'une variable aléatoire *(Variance of a random variable)*

RÉFÉRENCES

Fisher, R.A. (1918). The correlation between relatives on the supposition of Mendelian inheritance. *Trans. Roy. Soc.* Edinburgh, 52, pp. 399-433.

Fisher, R.A. (1925). *Statistical Methods for Research Workers.* Oliver & Boyd, Edinburgh.

Fisher, R.A. (1935). *The Design of Experiments.* Oliver & Boyd, Edinburgh.

VARIANCE D'UNE VARIABLE ALÉATOIRE

Variance of a random variable

La **variance** d'une **variable aléatoire** quantitative mesure la grandeur des déviations moyennes des valeurs de cette variable aléatoire par rapport à l'**espérance mathématique** (c'est-à-dire par rapport à sa **moyenne**).

ASPECTS MATHÉMATIQUES

Suivant que la **variable aléatoire** est discrète ou continue, on parle de variance d'une variable aléatoire discrète ou de variance d'une variable aléatoire continue.

Dans le cas d'une variable aléatoire discrète X prenant n valeurs $x_1, x_2, \ldots, x_n$ avec les fréquences relatives $f_1, f_2, \ldots, f_n$, la **variance** est définie par :

$$\begin{aligned} \sigma^2 &= Var(X) = E\left[(X-\mu)^2\right] \\ &= \sum_{i=1}^{n} f_i (x_i - \mu)^2 \end{aligned}$$

où μ représente l'**espérance mathématique** de X.

On peut établir une autre formule pour le calcul de la variance en procédant ainsi :

$$\begin{aligned} Var(X) &= E\left[(X-\mu)^2\right] \\ &= E\left[X^2 - 2\mu X + \mu^2\right] \\ &= E\left[X^2\right] - E\left[2\mu X\right] + E\left[\mu^2\right] \\ &= E\left[X^2\right] - 2\mu E[X] + \mu^2 \\ &= E\left[X^2\right] - 2\mu^2 + \mu^2 \\ &= E\left[X^2\right] - \mu^2 \\ &= E\left[X^2\right] - (E[X])^2 \\ &= \sum_{i=1}^{n} f_i x_i^2 - \left(\sum_{i=1}^{n} f_i x_i\right)^2. \end{aligned}$$

Dans la pratique, cette formule est en général plus commode pour effectuer les calculs.

Si la variable aléatoire X est continue sur l'**intervalle** D, l'expression de la variance devient :

$$\begin{aligned}\sigma^2 &= Var(X) = E\left[(X-\mu)^2\right] \\ &= \int_D (x-\mu)^2 f(x)\,dx\end{aligned}$$

On peut aussi transformer cette formule par les mêmes opérations que dans le cas discret et la variance est alors égale à :

$$Var(X) = \int_D x^2 f(x)\,dx - \left(\int_D x f(x)\,dx\right)^2.$$

Propriétés de la variance

1. Soient a et b deux constantes et X une variable aléatoire :

$$Var(aX+b) = a^2 \cdot Var(X)$$

2. Soient X et Y deux variables aléatoires :

$$\begin{aligned}Var(X+Y) &= Var(X)+Var(Y) \\ &\quad +2Cov(X,Y) \\ Var(X-Y) &= Var(X)+Var(Y) \\ &\quad -2Cov(X,Y),\end{aligned}$$

où $Cov(X,Y)$ est la **covariance** entre X et Y. En particulier, si X et Y sont indépendantes :

$$Var(X+Y) = Var(X)+Var(Y)$$
$$Var(X-Y) = Var(X)+Var(Y).$$

EXEMPLES

On considère deux exemples, l'un concernant une **variable aléatoire** discrète, l'autre une variable aléatoire continue.

- On lance un dé plusieurs fois de suite. On suppose que l'on gagne 1 euro si le résultat obtenu est pair, 2 euros si le résultat est 1 ou 3 et que l'on perde 3 euros si le résultat est 5.

  La variable aléatoire X décrit le nombre d'euros gagnés ou perdus. Le tableau ci-dessous représente les différentes valeurs de X et leur **probabilité** respective :

X	-3	1	2
$P(X)$	$\frac{1}{6}$	$\frac{3}{6}$	$\frac{2}{6}$

  La variance de la variable aléatoire discrète X est égale à :

$$\begin{aligned}Var(X) &= \sum_{i=1}^{3} f_i x_i^2 - \left[\sum_{i=1}^{3} f_i x_i\right]^2 \\ &= (-3)^2\frac{1}{6} + (1)^2\frac{3}{6} + (2)^2\frac{2}{6} \\ &\quad -\left(-3\frac{1}{6} + 1\frac{3}{6} + 2\frac{2}{6}\right)^2 \\ &= \frac{26}{9}.\end{aligned}$$

- Soit une variable aléatoire continue X dont la **fonction de densité** est uniforme :

$$f(x) = \begin{cases} 1 & \text{pour } 0 < x < 1, \\ 0 & \text{sinon.}\end{cases}$$

  L'**espérance mathématique** de cette variable aléatoire est égale à $\frac{1}{2}$.

  Nous pouvons calculer la variance de la variable aléatoire X :

$$\begin{aligned}Var(X) &= \int_0^1 x^2 f(x)\,dx - \left(\int_0^1 x f(x)\,dx\right)^2 \\ &= \int_0^1 x^2 \cdot 1dx - \left(\int_0^1 x \cdot 1dx\right)^2 \\ &= \frac{x^3}{3}\Big|_0^1 - \left(\frac{x^2}{2}\Big|_0^1\right)^2 \\ &= \frac{1}{3} - \left(\frac{1}{2}\right)^2 \\ &= \frac{1}{12}.\end{aligned}$$

MOTS CLÉS

Espérance mathématique *(Expected value)*
Variable aléatoire *(Random variable)*

VARIANCE PONDÉRÉE

Pooled variance

La **variance** pondérée est utilisée pour estimer la **valeur** de la variance de deux ou plusieurs populations lorsque les variances respectives de chaque population sont inconnues mais qu'elles sont considérées comme égales.

ASPECTS MATHÉMATIQUES

Sur la base de k échantillons de tailles n_1, n_2, ..., n_k, la **variance** pondérée S_p^2 est estimée par :

$$S_p^2 = \frac{\sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_{i.})^2}{\left(\sum_{i=1}^{k} n_i\right) - k}$$

avec

x_{ij} j-ème **observation** de l'échantillon i et
$\bar{x}_{i.}$ **moyenne arithmétique** de l'échantillon i.

Connaissant les variances respectives de chaque échantillon, la variance pondérée peut être aussi définie par :

$$S_p^2 = \frac{\sum_{i=1}^{k} (n_i - 1) \cdot S_i^2}{\left(\sum_{i=1}^{k} n_i\right) - k}$$

où S_i^2 est la variance de l'échantillon i :

$$S_i^2 = \frac{\sum_{j=1}^{n_i} (x_{ij} - \bar{x}_{i.})^2}{n_i - 1}.$$

EXEMPLE

Considérons deux échantillons de micro-ordinateurs de marque différente pour lesquels nous avons relevé le temps écoulé (en heures) avant la première panne :

Marque 1	Marque 2
2 800	2 800
2 700	2 600
2 850	2 400
2 650	2 700
2 700	2 600
2 800	2 500
2 900	
3 000	

Le nombre d'échantillons k est égal à 2, et la taille des échantillons est respectivement égale à :

$$n_1 = 8,$$
$$n_2 = 6.$$

Le calcul de la **variance** du premier échantillon nous donne :

$$S_1^2 = \frac{\sum_{j=1}^{n_1} (x_{1j} - \bar{x}_{1.})^2}{n_1 - 1}$$

avec :

$$\bar{x}_{1.} = \frac{\sum_{j=1}^{n_1} x_{1j}}{n_1} = \frac{22\,400}{8} = 2\,800$$

$$S_1^2 = \frac{95\,000}{8 - 1} = 13\,571,428.$$

Pour le deuxième échantillon, nous obtenons :

$$S_2^2 = \frac{\sum_{j=1}^{n_2} (x_{2j} - \bar{x}_{2.})^2}{n_2 - 1}$$

avec :

$$\bar{x}_{2.} = \frac{\sum_{j=1}^{n_2} x_{2j}}{n_2} = \frac{15\,600}{6} = 2\,600$$

$$S_2^2 = \frac{70\,000}{6 - 1} = 14\,000.$$

Si nous considérons que les variances inconnues des deux populations σ_1^2 et σ_2^2 sont identiques, nous pouvons calculer une **estimation** de $\sigma^2 = \sigma_1^2 = \sigma_2^2$ par la variance pondérée :

$$\begin{aligned} S_p^2 &= \frac{\sum_{i=1}^{2} (n_i - 1) \cdot S_i^2}{\sum_{i=1}^{2} n_i - 2} \\ &= \frac{(8 - 1) \cdot 13\,571,428 + (6 - 1) \cdot 14\,000}{8 + 6 - 2} \\ &= 13\,750. \end{aligned}$$

MOT CLÉ

Variance *(Variance)*

VARIATION IRRÉGULIÈRE

Irregular variation

Les variations irrégulières ou variations aléatoires constituent l'une des quatre composantes d'une **série chronologique**. Elles correspondent à des mouvements qui apparaissent irrégulièrement et généralement durant de courtes périodes.

Les variations irrégulières ne suivent pas de modèles particuliers et sont imprévisibles.

En pratique, toutes les composantes d'une série chronologique qui ne peuvent être attribuées à l'influence des fluctuations cycliques, des variations saisonnières ou à celle de la **tendance séculaire** sont classées parmi les variations irrégulières.

HISTORIQUE

Voir **série chronologique**.

ASPECTS MATHÉMATIQUES

Soit Y_t une **série chronologique** ; on peut la décrire à l'aide de ses composantes, à savoir :

- la **tendance séculaire** T_t ;
- les variations saisonnières S_t ;
- les fluctuations cycliques C_t ;
- les variations irrégulières I_t.

On distingue :

- le **modèle** multiplicatif : $Y_t = T_t \cdot S_t \cdot C_t \cdot I_t$;
- le modèle additif : $Y_t = T_t + S_t + C_t + I_t$.

Une fois la tendance séculaire, les variations saisonnières et les fluctuations cycliques déterminées, il est possible d'ajuster les données initiales de la série chronologique suivant ces trois composantes. On obtient ainsi les valeurs des variations irrégulières en chaque temps t, par les expressions suivantes :

$$\frac{Y_t}{S_t \cdot T_t \cdot C_t} = I_t \text{ (modèle multiplicatif),}$$

$$Y_t - S_t - T_t - C_t = I_t \text{ (modèle additif).}$$

DOMAINES ET LIMITATIONS

L'examen des composantes d'une **série chronologique** montre que les variations irrégulières peuvent être classées en deux catégories :

- les plus nombreuses peuvent être attribuées à un grand nombre de petites causes, d'erreurs de mesure notamment qui provoquent des variations de faible amplitude ;
- celles résultant d'événements accidentels isolés de plus grande ampleur tels que grève, décision administrative, krach financier, catastrophe naturelle, etc. Dans ce cas, pour appliquer la méthode des moyennes mobiles, on effectue un traitement préalable des données, pour corriger au mieux les données brutes.

Bien que l'on suppose habituellement que les variations irrégulières ne produisent des variations durables que sur un **intervalle** de temps de courte durée, on peut imaginer qu'elles soient assez puissantes pour prendre l'allure de fluctuations cycliques ou autres.

Pratiquement, on trouve que les variations irrégulières ont tendance à avoir une petite amplitude et à suivre une **loi normale**. Cela signifie qu'il y a fréquemment de petits écarts et rarement de grands écarts.

MOTS CLÉS

Fluctuation cyclique *(Cyclical fluctuation)*
Loi normale *(Normal distribution)*
Série chronologique *(Time series)*
Tendance séculaire *(Secular trend)*
Variation saisonnière *(Seasonal variation)*

RÉFÉRENCES

Bowerman, B.L. and O'Connel, R.T. (1979). *Time Series and Forecasting: An Applied Approach.* Duxbury Press, Wadsworth, Belmont, California.

Harnett, D.L. and Murphy, J.L. (1975). *Introductory Statistical Analysis.* Addison-Wesley, Reading, Mass.

VARIATION SAISONNIÈRE

Seasonal variation

Les variations saisonnières forment l'une des composantes de base de la **série chronologique**.

Il s'agit de variations de nature périodique qui réapparaissent régulièrement. Dans un contexte économique, les variations saisonnières peuvent se produire pour des raisons telles que :

- les conditions climatiques ;
- les coutumes propres à une **population** ;
- les fêtes religieuses.

Ce sont des fluctuations périodiques plus ou moins régulières qui se superposent au **mouvement extra-saisonnier**.

HISTORIQUE

Voir **série chronologique**.

ASPECTS MATHÉMATIQUES

L'évaluation des variations saisonnières s'effectue en déterminant un **indice** de variation saisonnière, appelé **indice saisonnier**.

EXEMPLE

On considère la **série chronologique** d'un phénomène quelconque ; il s'agit de données trimestrielles prises sur une durée de six ans résumées dans le tableau suivant :

Année	1^{er} trim.	2^{e} trim.	3^{e} trim.	4^{e} trim.
1	19, 65	16, 35	21, 30	14, 90
2	28, 15	25, 00	29, 85	23, 40
3	36, 75	33, 60	38, 55	32, 10
4	45, 30	42, 25	47, 00	40, 65
5	54, 15	51, 00	55, 75	49, 50
6	62, 80	59, 55	64, 40	58, 05

On détermine les indices saisonniers à l'aide de la méthode du **pourcentage** à la **moyenne**.

On calcule d'abord les moyennes arithmétiques trimestrielles pour les six années :

Année	Somme	Moyenne
1	72, 2	18, 05
2	106, 4	26, 6
3	141, 0	35, 25
4	175, 2	43, 8
5	210, 4	52, 6
6	244, 8	61, 2

On divise alors les lignes du tableau initial par la moyenne trimestrielle de l'année correspondante et on multiplie par 100.

Ensuite, puisque les résultats montrent d'assez grands écarts pour les valeurs obtenues pour un même trimestre, il est plus judicieux de prendre la **médiane** au lieu de la moyenne arithmétique pour obtenir des chiffres plus significatifs.

Année	1^{er} trim.	2^{e} trim.	3^{e} trim.	4^{e} trim.
1	108, 86	90, 58	118, 01	82, 55
2	105, 83	93, 98	112, 22	87, 97
3	104, 26	95, 32	109, 36	91, 06
4	103, 42	96, 46	107, 31	92, 81
5	102, 95	96, 96	105, 99	94, 11
6	102, 61	97, 30	105, 23	94, 85
Médiane	103, 84	95, 89	108, 33	91, 94

Le pourcentage médian pour chaque trimestre est donné dans la dernière ligne du tableau ci-dessus. La somme de ces pourcentages étant exactement de 400 %, un ajustement n'est pas nécessaire, et les chiffres figurant sur la dernière ligne représentent les indices saisonniers demandés.

On a vu qu'il existe trois autres méthodes pour déterminer les indices saisonniers, ces dernières fournissent les valeurs suivantes pour ce même problème :

Méthode	Trimestre			
	1	2	3	4
Pourcentage à la moyenne	103, 84	95, 89	108, 33	91, 94
Rapport à la tendance	112, 62	98, 08	104, 85	84, 45
Rapport à la moy. mobile	111, 56	98, 87	106, 14	83, 44
Chaînes	112, 49	98, 24	104, 99	84, 28

On constate que les résultats sont en accord les uns avec les autres, malgré la diversité des méthodes utilisées pour les déterminer.

MOTS CLÉS

Fluctuation cyclique *(Cyclical fluctuation)*
Indice *(Index number)*
Indice saisonnier *(Seasonal index)*
Moyenne arithmétique *(Arithmetic mean)*
Moyenne mobile *(Moving average)*
Prévision *(Forecasting)*
Série chronologique *(Time series)*
Tendance séculaire *(Secular trend)*

VECTEUR

Vector

On appelle vecteur de dimension n, une **matrice** $(n \times 1)$, c'est-à-dire un tableau ayant une seule colonne et n lignes. Les éléments figurant dans le tableau sont appelés les coefficients ou composantes du vecteur.

ASPECTS MATHÉMATIQUES

Soient $\mathbf{x} = (x_i)$ et $\mathbf{y} = (y_i)$ pour $i = 1, 2, \ldots, n$ deux vecteurs de taille n.

On définit :

- la somme de deux vecteurs, qui s'effectue composante par composante :

$$\mathbf{x} + \mathbf{y} = (x_i) + (y_i) = (x_i + y_i) \; ;$$

- le produit du vecteur $\mathbf{x}$ par le scalaire k :

$$k \cdot \mathbf{x} = (k \cdot x_i) \, ,$$

chaque composante étant multipliée par k ;

- le **produit scalaire** de $\mathbf{x}$ et $\mathbf{y}$:

$$\mathbf{x}' \cdot \mathbf{y} = (x_i)' \cdot (y_i) = \sum_{i=1}^{n} x_i \cdot y_i \, ;$$

- la **norme** de $\mathbf{x}$:

$$\| \mathbf{x} \| = \sqrt{\mathbf{x}' \cdot \mathbf{x}}$$

$$\| \mathbf{x} \| = \sqrt{\sum_{i=1}^{n} x_i^2}.$$

On appelle vecteur unité un vecteur de norme (ou longueur) 1. On obtient un vecteur unité en divisant un vecteur par sa norme ; ainsi, par exemple :

$$\frac{1}{\| \mathbf{x} \|}(x_i)$$

est un vecteur unité ayant la même direction que $\mathbf{x}$;

- le i-ème vecteur de la base canonique dans un espace à n dimensions est défini par le vecteur de taille n ayant un « 1 » à la ligne i et des « 0 » pour les autres composantes ;

- un vecteur étant une **matrice** de dimension particulière $(n \times 1)$, il est possible de multiplier un vecteur par une matrice $(m \times n)$ selon les règles de multiplication définies pour les matrices.

DOMAINES ET LIMITATIONS

Un vecteur peut être considéré comme une colonne de nombres, d'où le terme parfois utilisé de vecteur-colonne ; de même on utilise le terme de vecteur-ligne pour désigner la **transposée** d'un vecteur.

Ainsi le **produit scalaire** de deux vecteurs correspond au produit d'un vecteur-ligne par un vecteur-colonne.

EXEMPLE

Dans l'espace à trois dimensions, on considère les vecteurs :

$$\mathbf{x} = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} \quad \text{et} \quad \mathbf{y} = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$$

où $\mathbf{y}$ est le deuxième vecteur de la base canonique.

- La **norme** de $\mathbf{x}$ est donnée par :

$$\begin{aligned} \| \mathbf{x} \|^2 &= \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}' \cdot \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} = [1\ 2\ 3] \cdot \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} \\ \| \mathbf{x} \|^2 &= 1^2 + 2^2 + 3^2 = 1 + 4 + 9. \\ \| \mathbf{x} \|^2 &= 14, \end{aligned}$$

  d'où $\| \mathbf{x} \| = \sqrt{14} \approx 3,74$.

- De même, on obtient $\| \mathbf{y} \| = 1$.

- La somme des deux vecteurs $\mathbf{x}$ et $\mathbf{y}$ est égale à :

$$\mathbf{x} + \mathbf{y} = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 1+0 \\ 2+1 \\ 3+0 \end{bmatrix} = \begin{bmatrix} 1 \\ 3 \\ 3 \end{bmatrix}.$$

- La multiplication du vecteur $\mathbf{x}$ par le scalaire 3 donne :

$$3 \cdot \mathbf{x} = 3 \cdot \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} = \begin{bmatrix} 3 \cdot 1 \\ 3 \cdot 2 \\ 3 \cdot 3 \end{bmatrix} = \begin{bmatrix} 3 \\ 6 \\ 9 \end{bmatrix}.$$

- Le **produit scalaire** de $\mathbf{x}$ par $\mathbf{y}$ vaut :

$$\begin{aligned} \mathbf{x}' \cdot \mathbf{y} &= \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}' \cdot \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} = [1\ 2\ 3] \cdot \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} \\ &= 1 \cdot 0 + 2 \cdot 1 + 3 \cdot 0 = 2. \end{aligned}$$

MOTS CLÉS

Matrice *(Matrix)*
Norme *(Norm)*
Produit scalaire *(Scalar product)*
Transposée *(Transpose)*

VECTEUR PROPRE

Eigenvector

Soit $\mathbf{A}$ une **matrice** carrée d'ordre n, on appelle vecteur propre de $\mathbf{A}$, un **vecteur** $\mathbf{x}$ à n composantes (non toutes nulles) tel que le produit matriciel de $\mathbf{A}$ par $\mathbf{x}$ donne un multiple de $\mathbf{x}$, par exemple $k \cdot \mathbf{x}$.

On dit que le vecteur propre $\mathbf{x}$ est associé à la **valeur propre** k.

DOMAINES ET LIMITATIONS

Soit $\mathbf{A}$ une **matrice** carrée d'ordre n, le **vecteur** $\mathbf{x}$ (non nul) est un vecteur propre de $\mathbf{A}$ s'il existe un nombre k tel que

$$\mathbf{A} \cdot \mathbf{x} = k \cdot \mathbf{x}.$$

Comme $\mathbf{x}$ est non nul, on détermine les valeurs possibles de k en trouvant les solutions de l'équation $|\mathbf{A} - k \cdot \mathbf{I_n}| = 0$, où $\mathbf{I_n}$ est la **matrice identité** d'ordre n.

On trouve les vecteurs propres correspondants en résolvant le système d'équations défini par :

$$(\mathbf{A} - k \cdot \mathbf{I_n}) \cdot \mathbf{x} = \mathbf{0}.$$

Tout multiple non nul d'un vecteur propre peut être considéré comme vecteur propre. Prenons c un scalaire non nul

$$\begin{aligned} \mathbf{A} \cdot \mathbf{x} &= k \cdot \mathbf{x} \\ c \cdot (\mathbf{A} \cdot \mathbf{x}) &= c \cdot (k \cdot \mathbf{x}) \\ \mathbf{A} \cdot (c \cdot \mathbf{x}) &= k \cdot (c \cdot \mathbf{x}). \end{aligned}$$

$c \cdot \mathbf{x}$ est donc aussi vecteur propre pour autant que c soit non nul.

On convient d'appeler vecteur propre le vecteur de **norme** 1 ou vecteur unité porté par l'axe des $c \cdot \mathbf{x}$, et l'**axe factoriel** sera l'axe portant ce vecteur propre. Ce vecteur propre est unique (au signe près).

EXEMPLES

On considère la **matrice** carrée d'ordre 3 suivante :

$$\mathbf{A} = \begin{bmatrix} 4 & 3 & 2 \\ 0 & 1 & 0 \\ -2 & 2 & 0 \end{bmatrix}.$$

Les valeurs propres de cette matrice $\mathbf{A}$ s'obtiennent en annulant le **déterminant** de $(\mathbf{A} - k \cdot \mathbf{I_3})$

$$\mathbf{A} - k \cdot \mathbf{I_3} = \begin{bmatrix} 4 & 3 & 2 \\ 0 & 1 & 0 \\ -2 & 2 & 0 \end{bmatrix} - k \cdot \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

$$\mathbf{A} - k \cdot \mathbf{I_3} = \begin{bmatrix} 4-k & 3 & 2 \\ 0 & 1-k & 0 \\ -2 & 2 & -k \end{bmatrix}.$$

Le déterminant de $\mathbf{A} - k \cdot \mathbf{I_3}$ est donné par :

$$\begin{aligned} |\mathbf{A} - k \cdot \mathbf{I_3}| &= (1-k) \cdot [(4-k) \cdot (-k) - 2 \cdot (-2)] \\ &= (1-k) \cdot \left(k^2 - 4 \cdot k + 4\right) \\ &= (1-k) \cdot (k-2)^2. \end{aligned}$$

Les valeurs propres sont donc égales à :

$$k_1 = 1$$

$$\text{et} \quad k_2 = k_3 = 2.$$

On effectue maintenant le calcul des **vecteurs** propres associés à la valeur propre k_1 :

$$k_1 = 1.$$

$$\mathbf{A} - k_1 \cdot \mathbf{I_3} = \begin{bmatrix} 4-1 & 3 & 2 \\ 0 & 1-1 & 0 \\ -2 & 2 & -1 \end{bmatrix} = \begin{bmatrix} 3 & 3 & 2 \\ 0 & 0 & 0 \\ -2 & 2 & -1 \end{bmatrix}.$$

En désignant les composantes du vecteur propre x par

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix},$$

il faut résoudre le système d'équations suivant :

$$\begin{aligned} (\mathbf{A} - k_1 \cdot \mathbf{I_3}) \cdot \mathbf{x} &= 0 \\ \begin{bmatrix} 3 & 3 & 2 \\ 0 & 0 & 0 \\ -2 & 2 & -1 \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} &= \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \end{aligned}$$

c'est-à-dire

$$\begin{aligned} 3x_1 + 3x_2 + 2x_3 &= 0 \\ -2x_1 + 2x_2 - x_3 &= 0, \end{aligned}$$

où la deuxième équation, étant triviale, a été éliminée.

La solution de ce système d'équations en fonction de x_2 est donnée par :

$$\mathbf{x} = \begin{bmatrix} 7 \cdot x_2 \\ x_2 \\ -12 \cdot x_2 \end{bmatrix}.$$

Le vecteur $\mathbf{x}$ est un vecteur propre de la matrice $\mathbf{A}$, pour chaque **valeur** non nulle attribuée à x_2. Pour trouver le vecteur propre unitaire, on pose :

$$\begin{aligned} 1 &= \| \mathbf{x} \|^2 = (7 \cdot x_2)^2 + (x_2)^2 + (-12 \cdot x_2)^2 \\ &= (49 + 1 + 144) \cdot x_2^2 = 194 \cdot x_2^2, \end{aligned}$$

d'où $x_2^2 = \dfrac{1}{194}$

et donc $x_2 = \dfrac{1}{\sqrt{194}} = 0,0718.$

Le vecteur propre unitaire de la matrice $\mathbf{A}$ associé à la valeur propre $k_1 = 1$ est donc égal à :

$$\mathbf{x} = \begin{bmatrix} 0,50 \\ 0,07 \\ -0,86 \end{bmatrix}.$$

MOTS CLÉS

Axe factoriel *(Factorial axis)*
Déterminant *(Determinant)*
Matrice*(Matrix)*
Norme *(Norm)*
Valeur propre *(Eigenvalue)*
Vecteur *(Vector)*

RÉFÉRENCE

Dodge, Y. (2002). *Mathématiques de base pour économistes.* Springer-Verlag France, Paris.

VON MISES RICHARD
von Mises Richard

Richard von Mises naquit en 1883 à Lemberg dans l'Empire austro-hongrois (actuellement L'viv, Ukraine) et mourut en 1953. Jusqu'en 1906, R. von Mises étudia l'ingénierie mécanique à la Technical University de Vienne. Il devint l'assistant du professeur de mécanique Georg Hamel, à Bruenn (maintenant Brno, République tchèque) où il obtint le Venia Legendi en 1908. Il travailla ensuite comme professeur associé de mathématiques appliquées. Après la Première Guerre mondiale, en 1920, il devint directeur de l'Institut de mathématiques appliquées de Berlin.



Richard von Mises est principalement connu pour son travail sur les fondements de la probabilité et les statitisques qui furent réhabilitées dans les années 60. Il a fondé l'École de mathématiques appliquées de Berlin et écrivit son premier ouvrage sur le positivisme philosophique en 1939.

Quelques ouvrages et articles principaux de Richard von Mises :

1928 *Wahrscheinlichkeit, Statistik und Wahrheit.* Springer-Verlag, Wien.

1939 *Probability, Statistics and Truth*, Translated by J. Neyman, D. Scholl and E. Rabinowitsch. Edinburg, London. William Hodge & Co., Glasgow.



W

WILCOXON, TABLE DE

Wilcoxon table

Voir **table de Wilcoxon**.

WILCOXON, TEST DE

Wilcoxon test

Voir **test de Wilcoxon**.

WILCOXON FRANK

Wilcoxon Frank



Frank Wilcoxon (1892-1965) obtint un postgrade en chimie en 1921 à l'Université de Rugers et son doctorat en physique en 1924 à l'Université de Cornell où il fit encore un postdoctorat en 1924-25.

De 1928 à 1929, il fut employé par la Nichols Copper Company à Maspeth dans le Queens et retourna ensuite au Boyce Thompson Institute comme chef de groupe d'une étude concernant les effets des insecticides et fongicides. Il resta à l'Institut jusqu'à la Deuxième Guerre mondiale ; à ce moment et, pour les deux années qui suivirent, il fut chargé de la direction du laboratoire de Ravenna Ordnance Plant qui travaillait en collaboration avec l'Atlas Powder Company en Ohio.

En 1943, Frank Wilcoxon fut engagé comme chef de groupe au laboratoire de l'American Cyanamid Company (à Stamford, Connecticut) qui étudiait les inscticides et les fongicides.

C'est alors qu'il s'intéressa aux statistiques : en 1950, il fut transféré à la Division of Cyanamid aux Laboratoires Lederle où il développa un groupe de consultation statistique. De 1960 jusqu'à sa mort, il enseigna les statistiques appliquées au département de statistiques de l'Université d'État de Floride.

Quelques ouvrages et articles principaux de Frank Wilcoxon :

1945 Individual comparisons by ranking methods. *Biometrics*, 1, pp. 80-83.

1957 *Some rapid approximate statistical procedures.* American Cyanamid Co., Stamford Research Laboratories, Stamford, Conn.

WILCOXON SIGNÉ, TABLE DE

Wilcoxon signed table

Voir **table de Wilcoxon signé**.

WILCOXON SIGNÉ, TEST DE

Wilcoxon signed test

Voir **test de Wilcoxon signé**.

WILLCOX WALTER FRANCIS

Willcox Walter Francis



Walter Francis Willcox (1861-1964) fut diplômé du Amherst College (il obtint sa licence en 1884 et son postgrade en 1888). Il continua ses études à l'Université de Columbia et en Allemagne à la fin des années 1880. Il obtint son doctorat de l'Université de Columbia en 1891, après avoir terminé sa thèse *The Divorce Problem A Study in Statistics*. Dès 1891, Willcox occupa un poste de professeur d'économie et de statistique à Cornell où il resta jusqu'en 1931, date à laquelle il prit sa retraite.

Ses contributions principales en statistique furent dans le domaine de la démographie ainsi que dans celui du développement du système statistique fédéral.

De 1899 à 1902, il dirigea le douzième recensement des Nations unies et présida l'American Statitistical Association en 1912, l'American Economic Association en 1915 ainsi que l'International Statistical Institute en 1947.

Quelques ouvrages et articles principaux de Walther Francis Willcox :

1897 The Divorce Problem: A Study in Statistics. *Studies in History, Economics and Public Law.* Columbia University Press, New York.

1940 *Studies in American Demography*, Cornell University Press, Ithaca, NY.



Y

YOUDEN WILLIAM JOHN

Youden William John

William John Youden (1900-1971) obtint un doctorat en chimie à l'Université de Columbia en 1924.

De 1924 à 1948, il travailla au Boyce Thompson Institute for Plant Research, à Yonkers, New York. De 1948 et jusqu'à sa retraite en 1965, Youden fut un membre de la division de mathématiques appliquées au National Bureau of Standars, à Washington, D.C. C'est durant cette période que Youden développa son « ruggedness » test et écrivit *Statistical Methods for Chemists*.

Ouvrage principal de William John Youden :

1951 *Statistical Methods for Chemists*. John Wiley & Sons, New York.

YULE GEORGE UDNY

Yule George Udny



George Udny Yule naquit en 1871 à Beech Hill près de Haddington (Ecosse) et mourut en 1951 à Cambridge.

Le coefficient d'association de Yule, le paradoxe de Yule (appelé plus tard le paradoxe de Simpson) et le procédé de Yule perpétuent le nom d'un pionnier des statistiques qui a aussi contribué à la théorie de Mendel et à l'analyse des séries chronologiques.

Quelques ouvrages et articles principaux de George Udny Yule :

1897 On the Theory of Correlation. *Journal of the Royal Statistical Society*, 60, pp. 812-854.

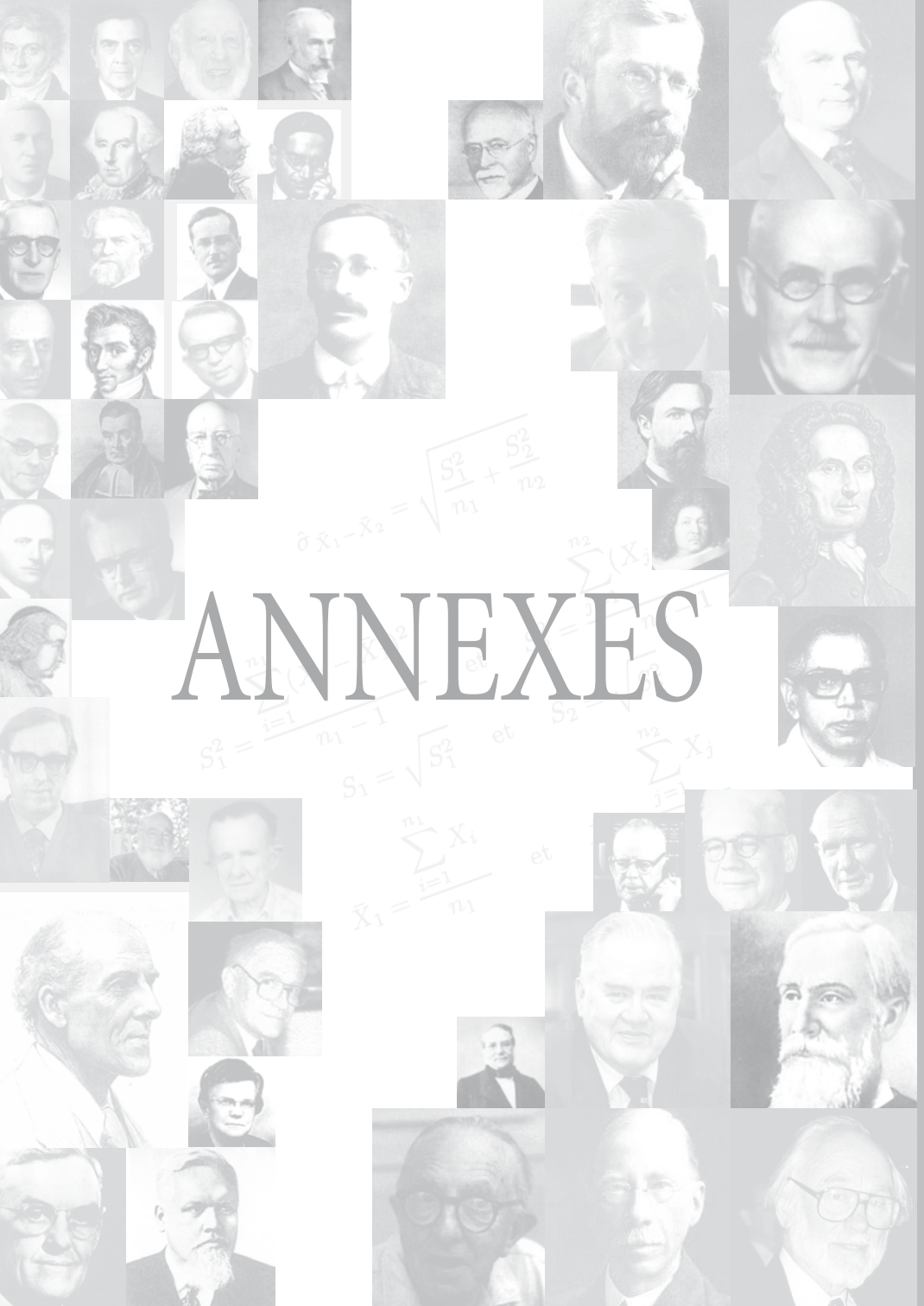
1900 On the association of attributes in statistics: with illustration from the material of the childhood society. *Philosophical Transactions of the Royal Society of London.* Series A, Vol. 194, pp. 257-319.

1926 Why do we sometimes get nonsense-correlations between time-series? A study sampling and the nature of time-series. *Journal of the Royal Statistical Society*, (2) 89, pp. 1-64.

1968 (and Kendall, M.G.) *An Introduction to the Theory of Statistics*. 14th edition. Charles Griffin & Company Ltd.

MOTS-CLÉS

Coefficient de Yule et Kendall *(Yule and Kendall coefficient)*
Mesure d'asymétrie *(Measure of skewness)*



ANNEXES

ANNEXE

TABLES STATISTIQUES

(Statistical tables)

A Table de nombres aléatoires *(Table of random numbers)*

B Table binomiale *(Binomial table)*

C Table de Fisher *(Fisher table)*

D Table de Kruskal-Wallis *(Kruskal-Wallis table)*

E Table de Student *(Student table)*

F Table du chi-carré *(Chi-square table)*

G Table normale *(Normal table)*

Annexe A

Table de nombres aléatoires

1415926535	3305727036	5024459455	8583616035	8164706001	9465764078
8979323846	5759591953	3469083026	6370766010	6145249192	9512694683
2643383279	9218611739	4252230825	4710181942	1732172147	9835259570
5028841971	8193261179	3344685035	9555961989	7235014144	9825822620
6939937510	3105118548	2619311881	4676783744	1973568548	5224894077
5820974944	7446237999	7101000313	9448255379	1613611573	2671947826
5923078164	6274956735	7838752886	7747268471	5255213347	8482601476
6286208995	1885752724	5875332083	4047534640	5741849468	9909026401
8628034825	8912279381	8142061717	6208046684	4385233239	3639443745
3421170679	8301194912	7669147303	2590694912	7394143337	5305068203
8214808651	9833673362	5982534904	3313677028	4547762416	4962524517
3282306647	4406566430	2875546873	8989152104	8625189835	4939965143
9384460952	8602139494	1159562863	7521620569	6948556209	1429809190
5058223172	6395224737	8823537875	6602405803	9219222184	6592509372
5359408128	1907021798	9375195778	8150193511	2725502542	2169646151
4811174502	6094370277	1857780532	2533824300	5688767179	5709858387
8410270193	5392171760	1712268066	3558764024	4946016530	4105978859
8521105559	2931767523	1300192787	7496473263	4668049886	5977297549
6446229489	8467481846	6611195909	9141992726	2723279178	8930161753
5493038196	7669405132	2164201989	4269922792	6085784383	9284681382
4428810975	5681271702	3809525720	6782354781	8279679766	6868386894
6659334461	4526356082	1065485863	6360093417	8145410095	2774155991
2847564823	7785771342	2788659361	2164121992	3883786360	8559252459
3786783165	7577896091	5338182796	4586315030	9506800642	5395943104
2712019091	7363717872	8230301952	2861829745	2512520511	9972524680
4564856692	1468440901	3530185292	5570674983	7392984896	8459872736
3460348610	2249534301	6899577362	8505494588	8412848862	4469584865
4543266482	4654958537	2599413891	5869269956	2694560424	3836736222
1339360726	1050792279	2497217752	9092721079	1965285022	6260991246
2491412733	6892589235	8347913151	7509302955	2106611863	8051243888
7245870066	4201995611	5574857242	3211653449	6744278629	4390451244
6315588170	2129021960	4541506959	8720275596	2039194945	1365497627
4881520920	8640344181	5082953311	2364806657	4712371373	8079771569
9628292540	5981362977	6861727855	4991198818	8696095636	1435997700
9171536436	4771309960	8890750983	3479775356	4371917287	1296160894
7892590360	5187072113	8175463746	6369807426	4677646575	4169486855
1133053052	4999999837	4939319255	5425278625	7396241389	5848406353
4882046652	2978049951	6040092774	5181841757	8658326451	4220722258
1384146951	5973173280	1671139001	4672890977	9958133904	2848864815
9415116094	1609631859	9848824012	7727938000	7802759009	8456028506

Annexe B

Table binomiale

Taille de l'échantillon		Probabilité de réussite p								
		0,01	0,02	0,03	0,04	0,05	0,1	0,2	0,3	0,5
5	0	0,9510	0,9039	0,8587	0,8154	0,7738	0,5905	0,3277	0,1681	0,0313
	1	0,9990	0,9962	0,9915	0,9852	0,9774	0,9185	0,7373	0,5282	0,1875
	2	1.0000	0,9999	0,9997	0,9994	0,9988	0,9914	0,9421	0,8369	0,5000
	3	1.0000	1.0000	1.0000	1,0000	1,0000	0,9995	0,9933	0,9692	0,8125
	4	1	1,0000	1,0000	1,0000	1,0000	1,0000	0,9997	0,9976	0,9688
	5	1	1	1	1	1	1	1	1	1
10	0	0,9044	0,8171	0,7374	0,6648	0,5987	0,3487	0,1074	0,0282	0,0010
	1	0,9957	0,9838	0,9655	0,9418	0,9139	0,7361	0,3758	0,1493	0,0107
	2	0,9999	0,9991	0,9972	0,9938	0,9885	0,9298	0,6778	0,3828	0,0547
	3	1,0000	1,0000	0,9999	0,9996	0,9990	0,9872	0,8791	0,6496	0,1719
	4	1,0000	1,0000	1,0000	1,0000	0,9999	0,9984	0,9672	0,8497	0,3770
	5	1	1,0000	1,0000	1,0000	1,0000	0,9999	0,9936	0,9527	0,6230
	6	1	1	1,0000	1,0000	1,0000	1,0000	0,9991	0,9894	0,8281
	7	1	1	1	1	1,0000	1,0000	0,9999	0,9984	0,9453
	8	1	1	1	1	1	1,0000	1,0000	0,9999	0,9893
	9	1	1	1	1	1	1	1,0000	1,0000	0,9990
	10	1	1	1	1	1	1	1	1	1
15	0	0,8601	0,7386	0,6333	0,5421	0,4633	0,2059	0,0352	0,0047	0,0000
	1	0,9904	0,9647	0,9270	0,8809	0,8290	0,5490	0,1671	0,0353	0,0005
	2	0,9996	0,9970	0,9906	0,9797	0,9638	0,8159	0,3980	0,1268	0,0037
	3	1,0000	0,9998	0,9992	0,9976	0,9945	0,9444	0,6482	0,2969	0,0176
	4	1,0000	1,0000	0,9999	0,9998	0,9994	0,9873	0,8358	0,5155	0,0592
	5	1,0000	1,0000	1,0000	1,0000	0,9999	0,9978	0,9389	0,7216	0,1509
	6	1	1,0000	1,0000	1,0000	1,0000	0,9997	0,9819	0,8689	0,3036
	7	1	1	1,0000	1,0000	1,0000	1,0000	0,9958	0,9500	0,5000
	8	1	1	1	1,0000	1,0000	1,0000	0,9992	0,9848	0,6964
	9	1	1	1	1	1	1,0000	0,9999	0,9963	0,8491
	10	1	1	1	1	1	1,0000	1,0000	0,9993	0,9408
	11	1	1	1	1	1	1	1,0000	0,9999	0,9824
	12	1	1	1	1	1	1	1,0000	1,0000	0,9963
	13	1	1	1	1	1	1	1,0000	1,0000	0,9995
	14	1	1	1	1	1	1	1	1,0000	1,0000
	15	1	1	1	1	1	1	1	1	1

Taille de l'échantillon		Probabilité de réussite p								
		0,01	0,02	0,03	0,04	0,05	0,1	0,2	0,3	0,5
20	0	0,8179	0,6676	0,5438	0,4420	0,3585	0,1216	0,0115	0,0008	0,0000
	1	0,9831	0,9401	0,8802	0,8103	0,7358	0,3917	0,0692	0,0076	0,0000
	2	0,9990	0,9929	0,9790	0,9561	0,9245	0,6769	0,2061	0,0355	0,0002
	3	1,0000	0,9994	0,9973	0,9926	0,9841	0,8670	0,4114	0,1071	0,0013
	4	1,0000	1,0000	0,9997	0,9990	0,9974	0,9568	0,6296	0,2375	0,0059
	5	1,0000	1,0000	1,0000	0,9999	0,9997	0,9887	0,8042	0,4164	0,0207
	6	1,0000	1,0000	1,0000	1,0000	1,0000	0,9976	0,9133	0,6080	0,0577
	7	1	1,0000	1,0000	1,0000	1,0000	0,9996	0,9679	0,7723	0,1316
	8	1	1	1,0000	1,0000	1,0000	0,9999	0,9900	0,8867	0,2517
	9	1	1	1	1,0000	1,0000	1,0000	0,9974	0,9520	0,4119
	10	1	1	1	1	1,0000	1,0000	0,9994	0,9829	0,5881
	11	1	1	1	1	1	1,0000	0,9999	0,9949	0,7483
	12	1	1	1	1	1	1,0000	1,0000	0,9987	0,8684
	13	1	1	1	1	1	1	1,0000	0,9997	0,9423
	14	1	1	1	1	1	1	1,0000	1,0000	0,9793
	15	1	1	1	1	1	1	1,0000	1,0000	0,9941
	16	1	1	1	1	1	1	1,0000	1,0000	0,9987
	17	1	1	1	1	1	1	1	1,0000	0,9998
	18	1	1	1	1	1	1	1	1,0000	1,0000
	19	1	1	1	1	1	1	1	1	1,0000
	20	1	1	1	1	1	1	1	1	1

Taille de l'échantillon		Probabilité de réussite p								
		0,01	0,02	0,03	0,04	0,05	0,1	0,2	0,3	0,5
30	0	0,7397	0,5455	0,4010	0,2939	0,2146	0,0424	0,0012	0,0000	0,0000
	1	0,9639	0,8795	0,7731	0,6612	0,5535	0,1837	0,0105	0,0003	0,0000
	2	0,9967	0,9783	0,9399	0,8831	0,8122	0,4114	0,0442	0,0021	0,0000
	3	0,9998	0,9971	0,9881	0,9694	0,9392	0,6474	0,1227	0,0093	0,0000
	4	1,0000	0,9997	0,9982	0,9937	0,9844	0,8245	0,2552	0,0302	0,0000
	5	1,0000	1,0000	0,9998	0,9989	0,9967	0,9268	0,4275	0,0766	0,0002
	6	1,0000	1,0000	1,0000	0,9999	0,9994	0,9742	0,6070	0,1595	0,0007
	7	1	1,0000	1,0000	1,0000	0,9999	0,9922	0,7608	0,2814	0,0026
	8	1	1,0000	1,0000	1,0000	1,0000	0,9980	0,8713	0,4315	0,0081
	9	1	1	1,0000	1,0000	1,0000	0,9995	0,9389	0,5888	0,0214
	10	1	1	1,0000	1,0000	1,0000	0,9999	0,9744	0,7304	0,0494
	11	1	1	1	1,0000	1,0000	1,0000	0,9905	0,8407	0,1002
	12	1	1	1	1	1,0000	1,0000	0,9969	0,9155	0,1808
	13	1	1	1	1	1	1,0000	0,9991	0,9599	0,2923
	14	1	1	1	1	1	1,0000	0,9998	0,9831	0,4278
	15	1	1	1	1	1	1,0000	0,9999	0,9936	0,5722
	16	1	1	1	1	1	1	1,0000	0,9979	0,7077
	17	1	1	1	1	1	1	1,0000	0,9994	0,8192
	18	1	1	1	1	1	1	1,0000	0,9998	0,8998
	19	1	1	1	1	1	1	1,0000	1,0000	0,9506
	20	1	1	1	1	1	1	1,0000	1,0000	0,9786
	21	1	1	1	1	1	1	1	1,0000	0,9919
	22	1	1	1	1	1	1	1	1,0000	0,9974
	23	1	1	1	1	1	1	1	1,0000	0,9993
	24	1	1	1	1	1	1	1	1,0000	0,9998
	25	1	1	1	1	1	1	1	1	1,0000
	26	1	1	1	1	1	1	1	1	1,0000
	27	1	1	1	1	1	1	1	1	1,0000
	28	1	1	1	1	1	1	1	1	1,0000
	29	1	1	1	1	1	1	1	1	1,0000
	30	1	1	1	1	1	1	1	1	1

Annexe C

Table de Fisher

ν_2	ν_1									
	1	2	3	4	5	6	7	8	9	10
1	161,4	199,5	215,7	224,6	230,2	234,0	236,8	238,9	240,5	241,9
2	18,51	19,00	19,16	19,25	19,30	19,33	19,35	19,37	19,38	19,40
3	10,13	9,55	9,28	9,12	9,01	8,94	8,89	8,85	8,81	8,79
4	7,71	6,94	6,59	6,39	6,26	6,16	6,09	6,04	6,00	5,96
5	6,61	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,77	4,74
6	5,99	5,14	4,76	4,53	4,39	4,28	4,21	4,15	4,10	4,06
7	5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,68	3,64
8	5,32	4,46	4,07	3,84	3,69	3,58	3,50	3,44	3,39	3,35
9	5,12	4,26	3,86	3,63	3,48	3,37	3,29	3,23	3,18	3,14
10	4,96	4,10	3,71	3,48	3,33	3,22	3,14	3,07	3,02	2,98
11	4,84	3,98	3,59	3,36	3,20	3,09	3,01	2,95	2,90	2,85
12	4,75	3,89	3,49	3,26	3,11	3,00	2,91	2,85	2,80	2,75
13	4,67	3,81	3,41	3,18	3,03	2,92	2,83	2,77	2,71	2,67
14	4,60	3,74	3,34	3,11	2,96	2,85	2,76	2,70	2,65	2,60
15	4,54	3,68	3,29	3,06	2,90	2,79	2,71	2,64	2,59	2,54
16	4,49	3,63	3,24	3,01	2,85	2,74	2,66	2,59	2,54	2,49
17	4,45	3,59	3,20	2,96	2,81	2,70	2,61	2,55	2,49	2,45
18	4,41	3,55	3,16	2,93	2,77	2,66	2,58	2,51	2,46	2,41
19	4,38	3,52	3,13	2,90	2,74	2,63	2,54	2,48	2,42	2,38
20	4,35	3,49	3,10	2,87	2,71	2,60	2,51	2,45	2,39	2,35
21	4,32	3,47	3,07	2,84	2,68	2,57	2,49	2,42	2,37	2,32
22	4,30	3,44	3,05	2,82	2,66	2,55	2,46	2,40	2,34	2,30
23	4,28	3,42	3,03	2,80	2,64	2,53	2,44	2,37	2,32	2,27
24	4,26	3,40	3,01	2,78	2,62	2,51	2,42	2,36	2,30	2,25
25	4,24	3,39	2,99	2,76	2,60	2,49	2,40	2,34	2,28	2,24
26	4,23	3,37	2,98	2,74	2,59	2,47	2,39	2,32	2,27	2,22
27	4,21	3,35	2,96	2,73	2,57	2,46	2,37	2,31	2,25	2,20
28	4,20	3,34	2,95	2,71	2,56	2,45	2,36	2,29	2,24	2,19
29	4,18	3,33	2,93	2,70	2,55	2,43	2,35	2,28	2,22	2,18
30	4,17	3,32	2,92	2,69	2,53	2,42	2,33	2,27	2,21	2,16
40	4,08	3,23	2,84	2,61	2,45	2,34	2,25	2,18	2,12	2,08
60	4,00	3,15	2,76	2,53	2,37	2,25	2,17	2,10	2,04	1,99
120	3,92	3,07	2,68	2,45	2,29	2,17	2,09	2,02	1,96	1,91
∞	3,84	3,00	2,60	2,37	2,21	2,10	2,01	1,94	1,88	1,83

ν_2	ν_1								
	12	15	20	24	30	40	60	120	∞
1	243,9	245,9	248,0	249,1	250,1	251,1	252,2	253,3	254,3
2	19,41	19,43	19,45	19,45	19,46	19,47	19,48	19,49	19,50
3	8,74	8,70	8,66	8,64	8,62	8,59	8,57	8,55	8,53
4	5,91	5,86	5,80	5,77	5,75	5,72	5,69	5,66	5,63
5	4,68	4,62	4,56	4,53	4,50	4,46	4,43	4,40	4,36
6	4,00	3,94	3,87	3,84	3,81	3,77	3,74	3,70	3,67
7	3,57	3,51	3,44	3,41	3,38	3,34	3,30	3,27	3,23
8	3,28	3,22	3,15	3,12	3,08	3,04	3,01	2,97	2,93
9	3,07	3,01	2,94	2,90	2,86	2,83	2,79	2,75	2,71
10	2,91	2,85	2,77	2,74	2,70	2,66	2,62	2,58	2,54
11	2,79	2,72	2,65	2,61	2,57	2,53	2,49	2,45	2,40
12	2,69	2,62	2,54	2,51	2,47	2,43	2,38	2,34	2,30
13	2,60	2,53	2,46	2,42	2,38	2,34	2,30	2,25	2,21
14	2,53	2,46	2,39	2,35	2,31	2,27	2,22	2,18	2,13
15	2,48	2,40	2,33	2,29	2,25	2,20	2,16	2,11	2,07
16	2,42	2,35	2,28	2,24	2,19	2,15	2,11	2,06	2,01
17	2,38	2,31	2,23	2,19	2,15	2,10	2,06	2,01	1,96
18	2,34	2,27	2,19	2,15	2,11	2,06	2,02	1,97	1,92
19	2,31	2,23	2,16	2,11	2,07	2,03	1,98	1,93	1,88
20	2,28	2,20	2,12	2,08	2,04	1,99	1,95	1,90	1,84
21	2,25	2,18	2,10	2,05	2,01	1,96	1,92	1,87	1,81
22	2,23	2,15	2,07	2,03	1,98	1,94	1,89	1,84	1,78
23	2,20	2,13	2,05	2,01	1,96	1,91	1,86	1,81	1,76
24	2,18	2,11	2,03	1,98	1,94	1,89	1,84	1,79	1,73
25	2,16	2,09	2,01	1,96	1,92	1,87	1,82	1,77	1,71
26	2,15	2,07	1,99	1,95	1,90	1,85	1,80	1,75	1,69
27	2,13	2,06	1,97	1,93	1,88	1,84	1,79	1,73	1,67
28	2,12	2,04	1,96	1,91	1,87	1,82	1,77	1,71	1,65
29	2,10	2,03	1,94	1,90	1,85	1,81	1,75	1,70	1,64
30	2,09	2,01	1,93	1,89	1,84	1,79	1,74	1,68	1,62
40	2,00	1,92	1,84	1,79	1,74	1,69	1,64	1,58	1,51
60	1,92	1,84	1,75	1,70	1,65	1,59	1,53	1,47	1,39
120	1,83	1,75	1,66	1,61	1,55	1,50	1,43	1,35	1,25
∞	1,75	1,67	1,57	1,52	1,46	1,39	1,32	1,22	1,00

Annexe D

Table de Kruskal-Wallis

n_1	n_2	n_3	**valeur critique**	α
2	1	1	2,7000	0,500
2	2	1	3,6000	0,267
2	2	2	4,5714	0,067
			3,7143	0,200
3	1	1	3,2000	0,300
3	2	1	4,2857	0,100
			3,8571	0,133
3	2	2	5,3572	0,029
			4,7143	0,048
			4,5000	0,067
			4,4643	0,105
3	3	1	5,1429	0,043
			4,5714	0,100
			4,0000	0,129
3	3	2	6,2500	0,011
			5,3611	0,032
			5,1389	0,061
			4,5556	0,100
			4,2500	0,121
3	3	3	7,2000	0,004
			6,4889	0,011
			5,6889	0,029
			5,0667	0,086
			4,6222	0,100
4	1	1	3,5714	0,200
4	2	1	4,8214	0,057
			4,5000	0,076
			4,0179	0,114
4	2	2	6,0000	0,014
			5,3333	0,033
			5,1250	0,052
			4,3750	0,100
			4,1667	0,105
4	3	1	5,8333	0,021
			5,2083	0,050
			5,0000	0,057
			4,0556	0,093
			3,8889	0,129

n_1	n_2	n_3	**valeur critique**	α
4	3	2	6,4444	0,009
			6,4222	0,010
			5,4444	0,047
			5,4000	0,052
			4,5111	0,098
			4,4666	0,101
4	3	3	6,7455	0,010
			6,7091	0,013
			5,7909	0,046
			5,7273	0,050
			4,7091	0,094
			4,7000	0,101
4	4	1	6,6667	0,010
			6,1667	0,013
			4,9667	0,046
			4,8667	0,054
			4,1667	0,082
			4,0667	0,102
4	4	2	7,0364	0,006
			6,8727	0,011
			5,4545	0,046
			5,2364	0,052
			4,5545	0,098
			4,4455	0,103
4	4	3	7,1739	0,010
			7,1364	0,011
			5,5985	0,049
			5,5758	0,051
			4,5455	0,099
			4,4773	0,102
4	4	4	7,6538	0,008
			7,5385	0,011
			5,6923	0,049
			5,6538	0,054
			4,6539	0,097
			4,5001	0,104

n_1	n_2	n_3	**valeur critique**	α
5	1	1	3,8571	0,143
5	2	1	5,2500	0,036
			5,0000	0,048
			4,4500	0,071
			4,2000	0,095
			4,0500	0,119
5	2	2	6,5333	0,008
			6,1333	0,013
			5,1600	0,034
			5,0400	0,056
			4,3733	0,090
			4,2933	0,122
5	3	1	6,4000	0,012
			4,9600	0,048
			4,8711	0,052
			4,0178	0,095
			3,8400	0,123
5	3	2	6,9091	0,009
			6,8606	0,011
			5,4424	0,048
			5,3455	0,050
			4,5333	0,097
			4,4121	0,109
5	3	3	6,9818	0,010
			6,8608	0,011
			5,4424	0,048
			5,3455	0,050
			4,5333	0,097
			4,4121	0,109
5	4	1	6,9545	0,008
			6,8400	0,011
			4,9855	0,044
			4,8600	0,056
			3,9873	0,098
			3,9600	0,102
5	4	2	7,2045	0,009
			7,1182	0,010
			5,2727	0,049
			5,2682	0,050
			4,5409	0,098
			4,5182	0,101

n_1	n_2	n_3	**valeur critique**	α
5	4	3	7,4449	0,010
			7,3949	0,011
			5,6564	0,049
			5,6308	0,050
			4,5487	0,099
			4,5231	0,103
5	4	4	7,7604	0,009
			7,7440	0,011
			5,6571	0,049
			5,6176	0,050
			4,6187	0,100
			4,5527	0,102
5	5	1	7,3091	0,009
			6,8364	0,011
			5,1273	0,046
			4,9091	0,053
			4,1091	0,086
			4,0364	0,105
5	5	2	7,3385	0,010
			7,5429	0,010
			5,7055	0,046
			5,6264	0,051
			4,5451	0,100
			4,5363	0,102
5	5	4	7,8229	0,010
			7,7914	0,010
			5,6657	0,049
			5,6429	0,050
			4,5229	0,099
			4,5200	0,101
5	5	5	8,0000	0,09
			7,9800	0,010
			5,7800	0,049
			5,6600	0,051
			4,5600	0,100
			4,5000	0,102

Annexe E

Table de Student

ν	α					
	0,100	0,050	0,025	0,010	0,005	0,001
1	3,078	6,314	12,706	31,821	63,656	318,289
2	1,886	2,920	4,303	6,965	9,925	22,328
3	1,638	2,353	3,182	4,541	5,841	10,214
4	1,533	2,132	2,776	3,747	4,604	7,173
5	1,476	2,015	2,571	3,365	4,032	5,894
6	1,440	1,943	2,447	3,143	3,707	5,208
7	1,415	1,895	2,365	2,998	3,499	4,785
8	1,397	1,860	2,306	2,896	3,355	4,501
9	1,383	1,833	2,262	2,821	3,250	4,297
10	1,372	1,812	2,228	2,764	3,169	4,144
11	1,363	1,796	2,201	2,718	3,106	4,025
12	1,356	1,782	2,179	2,681	3,055	3,930
13	1,350	1,771	2,160	2,650	3,012	3,852
14	1,345	1,761	2,145	2,624	2,977	3,787
15	1,341	1,753	2,131	2,602	2,947	3,733
16	1,337	1,746	2,120	2,583	2,921	3,686
17	1,333	1,740	2,110	2,567	2,898	3,646
18	1,330	1,734	2,101	2,552	2,878	3,610
19	1,328	1,729	2,093	2,539	2,861	3,579
20	1,325	1,725	2,086	2,528	2,845	3,552
21	1,323	1,721	2,080	2,518	2,831	3,527
22	1,321	1,717	2,074	2,508	2,819	3,505
23	1,319	1,714	2,069	2,500	2,807	3,485
24	1,318	1,711	2,064	2,492	2,797	3,467
25	1,316	1,708	2,060	2,485	2,787	3,450
26	1,315	1,706	2,056	2,479	2,779	3,435
27	1,314	1,703	2,052	2,473	2,771	3,421
28	1,313	1,701	2,048	2,467	2,763	3,408
29	1,311	1,699	2,045	2,462	2,756	3,396
30	1,310	1,697	2,042	2,457	2,750	3,385
40	1,303	1,684	2,021	2,423	2,704	3,307
50	1,299	1,676	2,009	2,403	2,678	3,261
60	1,296	1,671	2,000	2,390	2,660	3,232
70	1,294	1,667	1,994	2,381	2,648	3,211
80	1,292	1,664	1,990	2,374	2,639	3,195
90	1,291	1,662	1,987	2,368	2,632	3,183
100	1,290	1,660	1,984	2,364	2,626	3,174
∞	1,282	1,645	1,960	2,326	2,576	3,090

Annexe F

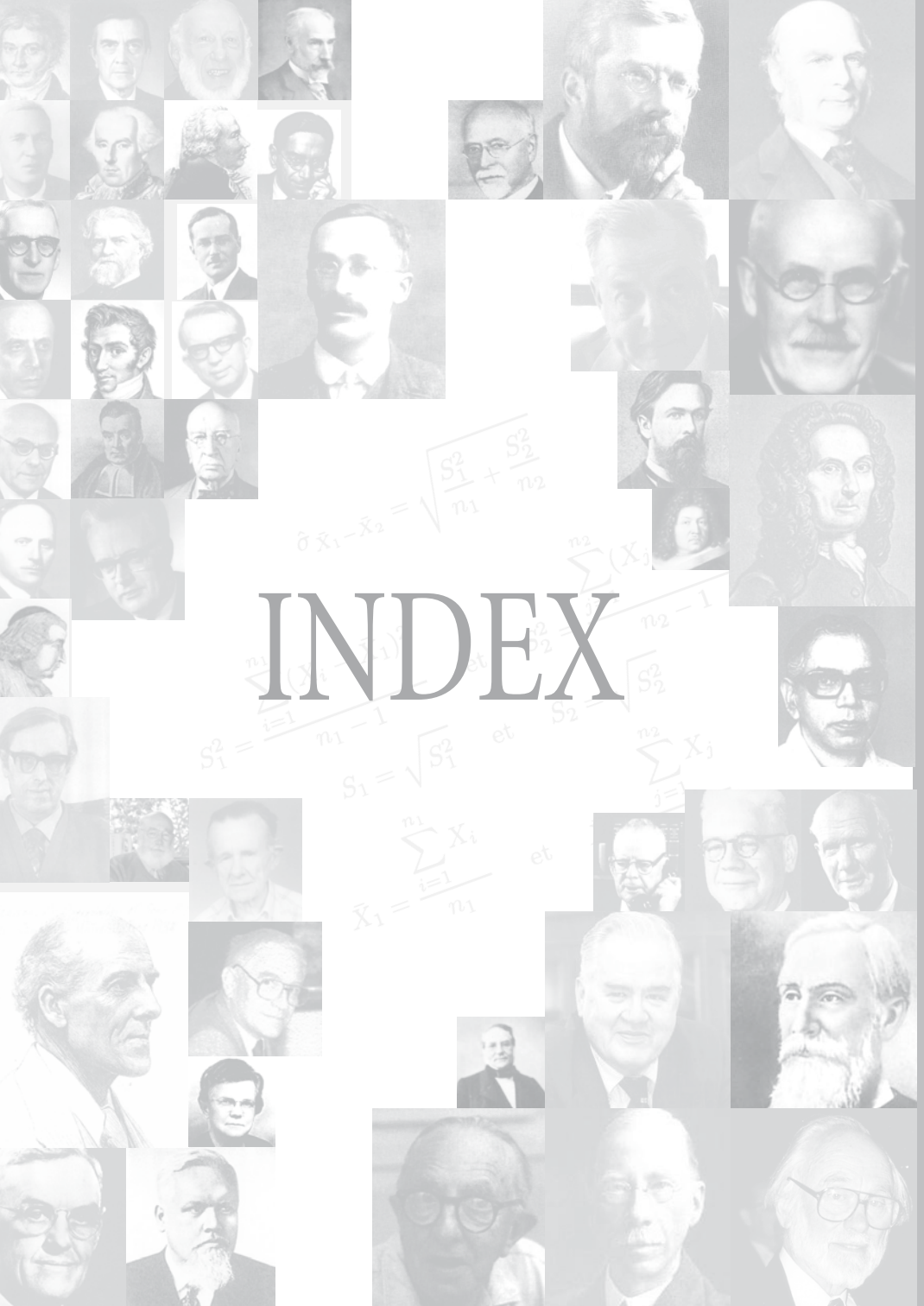
Table du chi-carré

	Seuil de signification					
ν	0,990	0,950	0,10	0,05	0,025	0,01
1	0,000	0,004	2,71	3,84	5,020	6,63
2	0,020	0,103	4,61	5,99	7,380	9,21
3	0,115	0,352	6,25	7,81	9,350	11,34
4	0,297	0,711	7,78	9,49	11,140	13,23
5	0,554	1,145	9,24	11,07	12,830	15,09
6	0,872	1,635	10,64	12,53	14,450	16,81
7	1,239	2,167	12,02	14,07	16,010	18,48
8	1,646	2,733	13,36	15,51	17,530	20,09
9	2,088	3,325	14,68	16,92	19,020	21,67
10	2,558	3,940	15,99	18,31	20,480	23,21
11	3,050	4,580	17,29	19,68	21,920	24,72
12	3,570	5,230	18,55	21,03	23,340	26,22
13	4,110	5,890	19,81	22,36	24,740	27,69
14	4,660	6,570	21,06	23,68	26,120	29,14
15	5,230	7,260	22,31	25,00	27,490	30,58
16	5,810	7,960	23,54	26,30	28,850	32,00
17	6,410	8,670	24,77	27,59	30,190	33,41
18	7,020	9,390	25,99	28,87	31,530	34,81
19	7,630	10,120	27,20	30,14	32,850	36,19
20	8,260	10,850	28,41	31,41	34,170	37,57
21	8,900	11,590	29,62	39,67	35,480	38,93
22	9,540	12,340	30,81	33,92	36,780	40,29
23	10,200	13,090	32,01	35,17	38,080	41,64
24	10,860	13,850	33,20	36,42	39,360	42,98
25	11,520	14,610	34,38	37,65	40,650	44,31
26	12,200	15,380	35,56	38,89	41,920	45,64
27	12,880	16,150	36,74	40,11	43,190	46,96
28	13,570	16,930	37,92	41,34	44,460	48,28
29	14,260	17,710	39,09	42,56	45,720	49,59
30	14,950	18,490	40,26	43,77	46,980	50,89

Annexe G

Table normale

Z	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,5000	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239	0,5279	0,5319	0,5359
0,1	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675	0,5714	0,5753
0,2	0,5793	0,5832	0,5871	0,5910	0,5948	0,5987	0,6026	0,6064	0,6103	0,6141
0,3	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443	0,6480	0,6517
0,4	0,6554	0,6591	0,6628	0,6664	0,6700	0,6736	0,6772	0,6808	0,6844	0,6879
0,5	0,6915	0,6950	0,6985	0,7019	0,7054	0,7088	0,7123	0,7157	0,7190	0,7224
0,6	0,7257	0,7291	0,7324	0,7357	0,7389	0,7422	0,7454	0,7486	0,7517	0,7549
0,7	0,7580	0,7611	0,7642	0,7673	0,7704	0,7734	0,7764	0,7794	0,7823	0,7852
0,8	0,7881	0,7910	0,7939	0,7967	0,7995	0,8023	0,8051	0,8078	0,8106	0,8133
0,9	0,8159	0,8186	0,8212	0,8238	0,8264	0,8289	0,8315	0,8340	0,8365	0,8389
1.0	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531	0,8554	0,8577	0,8599	0,8621
1.1	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749	0,8770	0,8790	0,8810	0,8830
1.2	0,8849	0,8869	0,8888	0,8907	0,8925	0,8944	0,8962	0,8980	0,8997	0,9015
1.3	0,9032	0,9049	0,9066	0,9082	0,9099	0,9115	0,9131	0,9147	0,9162	0,9177
1.4	0,9192	0,9207	0,9222	0,9236	0,9251	0,9265	0,9279	0,9292	0,9306	0,9319
1.5	0,9332	0,9345	0,9357	0,9370	0,9382	0,9394	0,9406	0,9418	0,9429	0,9441
1.6	0,9452	0,9463	0,9474	0,9484	0,9495	0,9505	0,9515	0,9525	0,9535	0,9545
1.7	0,9554	0,9564	0,9573	0,9582	0,9591	0,9599	0,9608	0,9616	0,9625	0,9633
1.8	0,9641	0,9649	0,9656	0,9664	0,9671	0,9678	0,9686	0,9693	0,9699	0,9706
1.9	0,9713	0,9719	0,9726	0,9732	0,9738	0,9744	0,9750	0,9756	0,9761	0,9767
2.0	0,9772	0,9778	0,9783	0,9788	0,9793	0,9798	0,9803	0,9808	0,9812	0,9817
2.1	0,9821	0,9826	0,9830	0,9834	0,9838	0,9842	0,9846	0,9850	0,9854	0,9857
2.2	0,9861	0,9864	0,9868	0,9871	0,9875	0,9878	0,9881	0,9884	0,9887	0,9890
2.3	0,9893	0,9896	0,9898	0,9901	0,9904	0,9906	0,9909	0,9911	0,9913	0,9916
2.4	0,9918	0,9920	0,9922	0,9925	0,9927	0,9929	0,9931	0,9932	0,9934	0,9936
2.5	0,9938	0,9940	0,9941	0,9943	0,9945	0,9946	0,9948	0,9949	0,9951	0,9952
2.6	0,9953	0,9955	0,9956	0,9957	0,9959	0,9960	0,9961	0,9962	0,9963	0,9964
2.7	0,9965	0,9966	0,9967	0,9968	0,9969	0,9970	0,9971	0,9972	0,9973	0,9974
2.8	0,9974	0,9975	0,9976	0,9977	0,9977	0,9978	0,9979	0,9979	0,9980	0,9981
2.9	0,9981	0,9982	0,9982	0,9983	0,9984	0,9984	0,9985	0,9985	0,9986	0,9986
3.0	0,9987	0,9987	0,9987	0,9988	0,9988	0,9989	0,9989	0,9989	0,9990	0,9990



INDEX

INDEX

A

adéquation, test d' 3
  voir test d'adéquation 518
algorithme 3
algorithme de Yates 4
analyse combinatoire 5
analyse de covariance 6
analyse de données 12
analyse de données binaires 14
analyse de données catégoriques 14
analyse de données manquantes 15
  voir observation manquante 384
analyse de régression 15
analyse de variance 17
analyse de variance pour deux facteurs 20
analyse de variance pour un facteur 26
analyse des résidus 29
analyse exploratoire de données 33
analyse factorielle des correspondances 35
analyse multivariée 41
  voir analyse de données 12
  voir analyse factorielle des correspondances 35
  voir classification automatique 73
Anderson-Darling, test d' 41
  voir test d'Anderson-Darling 522
Anderson Oskar 42
Anderson Theodore W. 42
arrangement 43
autocorrélation et autocorrélation partielle 45
axe factoriel 47

B

Barnard George A. 51
Bayes, théorème de 51
  voir théorème de Bayes 571
Bayes Thomas 51
Bernoulli, épreuve de 52
  voir épreuve de Bernoulli 175
Bernoulli, famille 52
Bernoulli, loi de 52
  voir loi de Bernoulli 290
Bernoulli, théorème de 53
  voir théorème de Bernoulli 572
Bernoulli Jakob 53
bêta, loi 54
  voir loi bêta 286
biais 54
Bienaymé Irénée-Jules 55
bilatéral, test 55
  voir test bilatéral 515
binôme 56
binomial, test 56
  voir test binomial 516
binomiale, loi 56
  voir loi binomiale 287
binomiale, table 56
  voir table binomiale 495
  voir annexe B 615
binomiale négative, loi 56
  voir loi binomiale négative 288
biostatistique 57
bloc 58
Bonferroni Carlo E. 58
bootstrap 59
Boscovich Roger J. 62
box plot 62

C

carrés gréco-latins 67
  voir plan carré gréco-latin 397
carrés latins 67
  voir plan carré latin 398
catégorie 67
Cauchy, loi de 67
  voir loi de Cauchy 291
cause et effet en épidémiologie 67
centile 69
central limite, théorème 71
  voir théorème central limite 569
chi-carré, distance du 71
  voir distance du chi-carré 139
chi-carré, loi du 71
  voir loi du chi-carré 300
chi-carré, table du 71
  voir table du chi-carré 501
  voir annexe F 623
chi-carré, test d'adéquation du 71

voir test d'adéquation du chi-carré 519
chi-carré, test d'indépendance 71
voir test d'indépendance du chi-carré ... 528
chi-carré, test du 71
voir test du chi-carré 555
classification 71
classification automatique 73
coefficient de corrélation 76
coefficient de corrélation des rangs de Kendall 80
coefficient de corrélation des rangs de Spearman 84
coefficient de corrélation multiple 87
voir coefficient de corrélation 76
coefficient de détermination 88
coefficient de variation 90
coefficient de Yule et de Kendall 90
coefficient β_1 de Pearson 92
coefficient β_2 de Pearson 94
colinéarité 96
collecte des données 98
combinaison 99
compatibilité 100
complémentaire 101
contingence, table de 102
voir table de contingence 496
contraste 102
convergence 102
convergence, théorème de 104
voir théorème de convergence 573
corrélation partielle 104
corrélation sérielle 105
couple de variables aléatoires 106
courbe de fréquences 107
covariance 108
covariation 109
Cox David R. 112
Cox Gertrude 113
C_p de Mallows 114
Cramér Harald 116

D

Daniels Henry E. 121
décile 121
degré de liberté 123
Deming W. Edwards 124
démographie 124
dendrogramme 125
dépendance 128
déterminant 128
déviation 129
diagramme circulaire 129
diagramme en barres 130
diagramme en bâtons 132
diagramme en points 133
diagramme figuratif 133
diagramme stem and leaf 134
dichotomie 135
différence significative minimale, test de ... 136
voir test de la différence significative minimale 539
dispersion 136
voir mesure de dispersion 332
distance 136
distance de Mahalanobis 138
distance du chi-carré 139
distances, tableau de 140
voir tableau de distances 504
distribution d'échantillonnage 140
distribution de fréquences 144
distribution plurimodale 146
voir distribution de fréquences 144
distribution unimodale 146
voir distribution de fréquences 144
donnée 146
donnée binaire 146
données catégoriques 147
données incomplètes 148
voir observation manquante 384
données spatiales 148
données standardisées 149
Durbin-Watson, test de 151
voir test de Durbin-Watson 531

E

écart géométrique 155
écart médian 155
écart moyen 155
écart type 156
échantillon 159
échantillonnage 159
échantillonnage aléatoire simple 162
échantillonnage par grappes 163
échantillonnage par quotas 164

échantillonnage stratifié 165
échantillonnage systématique 168
économétrie 168
Edgeworth Francis Y. 169
effet principal 170
empan 170
ensemble fondamental 171
épidémiologie 172
épidémiologie causale 173
épidémiologie descriptive 174
épreuve de Bernoulli 175
équations normales 175
erreur 177
erreur de première espèce 179
erreur de seconde espèce 180
erreur quadratique moyenne 180
erreur quadratique moyenne totale 182
erreur type 185
espérance mathématique 188
estimateur 189
estimation 191
estimation L1 193
estimation ponctuelle 194
estimation robuste 195
evénement 196
événement certain 198
voir ensemble fondamental 171
exactitude 198
expérience 198
expérience aléatoire 199
expérience factorielle 200
exponentielle, loi 201
voir loi exponentielle 301
extrapolation 201

F

facteur 205
Fisher, indice de 205
voir indice de Fisher 253
Fisher, loi de 205
voir loi de Fisher 291
voir table de Fisher 496
voir annexe C 618
Fisher, test de 205
voir test de Fisher 533
Fisher Irving 205
Fisher Ronald Aylmer 206
fluctuation cyclique 207
fonction de densité 209
fonction de densité conjointe 210
fonction de densité marginale 211
fonction de probabilité 211
fonction de répartition 212
fonction de répartition conjointe 212
fonction de répartition continue 213
fonction de répartition disctrète 214
fonction de répartition marginale 215
fraction de plan d'expérience factorielle 216
fréquence 216
fréquences, polygone de 217
voir polygone de fréquences 404
fréquences, tableau de 217
voir tableau de fréquences 505

G

Galton Francis 221
gamma, loi 221
voir loi gamma 302
Gauss Carl Friedrich 221
Gauss-Markov, théorème de 222
voir théorème de Gauss-Markov 573
génération de nombres aléatoires 222
voir annexe A 614
géométrique, loi 224
voir loi géométrique 303
géostatistique 224
Gini Corrado 225
Gosset William Sealy 225
grands nombres, loi des 226
voir loi des grands nombres 298
graphique de dispersion 227
graphique gausso-arithmétique 228
graphique logarithmique 230
graphique quantitatif 231
graphique semi-logarithmique 232

H

Hájek Jaroslav 237
histogramme 237
homogénéité, test d' 239
voir test d'homogénéité 524
Hotelling Harold 239
Huber Peter J. 240

hypergéométrique, loi 240
voir loi hypergéométrique 304
hypothèse 240
hypothèse, test d' 241
voir test d'hypothèse 525
hypothèse alternative 241
hypothèse nulle 243

I

incidence 247
voir taux d'incidence 509
indépendance 247
indépendance, test d' 249
voir test d'indépendance 528
indicateur 249
indice 251
indice de Fisher 253
indice de Gini 255
indice de Laspeyres 257
indice de Paasche 258
indice élémentaire 259
indice saisonnier 260
indice synthétique 263
inégalité de Tchebychev 264
voir loi des grands nombres 298
inertie 264
inférence 265
interaction 265
intervalle 266
intervalle de confiance 266
intervalle interquartile 268
inverse généralisé 268
inversion 269

J

jackknife, méthode du 273
voir méthode du jackknife 346
Jeffreys Harold 273

K

Kolmogorov Andreï Nicolaïévitch 277
Kolmogorov-Smirnov, test de 278
voir test de Kolmogorov-Smirnov 534
Kruskal-Wallis, table de 278
voir table de Kruskal-Wallis 497
Kruskal-Wallis, test de 278
voir test de Kruskal-Wallis 537

L

Laplace, loi de 281
voir loi de Laplace 292
Laplace Pierre Simon de 281
Legendre Adrien-Marie 281
logiciel statistique 282
lognormale, loi 286
voir loi lognormale 306
loi bêta 286
loi binomiale 287
loi binomiale négative 288
loi de Bernoulli 290
loi de Cauchy 291
loi de Fisher 291
loi de Laplace 292
loi de Poisson 293
loi de probabilité 294
loi de probabilité continue 295
loi de probabilité discrète 296
loi de Student 296
loi des grands nombres 298
loi du chi-carré 300
loi exponentielle 301
loi gamma 302
loi géométrique 303
loi hypergéométrique 304
loi lognormale 306
loi multinomiale 307
loi normale 309
loi uniforme 310
loi uniforme discrète 312

M

Mahalanobis Prasanta Chandra 317
Mann-Withney, test de 317
voir test de Mann-Whitney 541
Markov Andreï Andreïevitch 317
matrice 318
matrice de variance-covariance 321
matrice d'inertie 323
matrice **H** 324
matrice idempotente 326
voir matrice 318

matrice identité 326
voir matrice 318
matrice inverse 326
maximum de vraisemblance 327
médiane 329
mesure d'aplatissement 331
mesure d'asymétrie 332
mesure de dispersion 332
mesure de dissimilarité 333
mesure de forme 334
mesure de position 335
mesure de similarité 335
mesure de tendance centrale 336
méthode de Monte Carlo 338
méthode des moindres carrés 339
méthode des moindres carrés pondérés 341
méthode des moments 344
méthode du chaînage complet 346
méthode du chaînage simple 346
méthode du jackknife 346
mode 349
modèle 351
modèles *ARMA* 353
moindres carrés 354
voir méthode des moindres carrés 339
moindres carrés pondérés, méthode des 354
voir méthode des moindres carrés pondérés 341
Moivre Abraham de 354
Moivre-Laplace, théorème de 355
voir théorème de Moivre-Laplace 574
moment 355
mouvement extra-saisonnier 357
moyenne 358
moyenne arithmétique 358
moyenne arithméthique pondérée 362
moyenne géométrique 362
moyenne harmonique 364
moyenne mobile 365
multicolinéarité 367
multinomiale, loi 367
voir loi multinomiale 307
multiplicateur de Lagrange 367

N

Nelder John A. 373
Newcomb Simon 373
Neyman Jerzy 374
niveau d'un facteur 375
niveau de confiance 375
nombre aléatoire 377
non paramétrique, test 377
voir test non paramétrique 566
normale, loi 377
voir loi normale 309
normale, table 377
voir table normale 502
voir annexe G 624
normalisation 377
norme d'un vecteur 378

O

observation 383
observation aberrante 383
observation manquante 384
odds et odds ratio 385
ogive 387
Olkin Ingram 387
optimisation 388

P

panel 393
paramètre 394
paramétrique, test 394
voir test paramétrique 567
Pearson Egon Sharpe 394
Pearson Karl 395
permutation 396
Pitman Edwin James George 397
plan carré gréco-latin 397
plan carré latin 398
plan complètement randomisé 399
plan d'expérience 400
plan optimal 400
plan randomisé à blocs 401
point levier 402
Poisson, loi de 404
voir loi de Poisson 293
Poisson Siméon-Denis 404
polygone de fréquences 404
population 406
population-cible 406
pourcentage 406

prévalence 406
voir taux de prévalence 510
prévision 407
probabilité 408
probabilité, loi de 411
voir loi de probabilité 294
probabilité conditionnelle 411
probabilité continue, loi de 412
voir loi de probabilité continue 295
probabilité discrète, loi de 412
voir loi de probabilité discrète 296
processus stochastique 413
produit scalaire 415
programmation linéaire 416

Q

Q-Q plot (Quantile to Quantile Plot) 421
quantile 422
quartile 423
Quetelet Adolphe 425

R

randomisation 429
rang d'une matrice 429
Rao Calyampudi Radhakrishna 430
rapport de vraisemblance, test du 430
voir test du rapport de vraisemblance ... 557
ré-échantillonnage 431
recensement 431
recherche opérationnelle 433
région d'acceptation 434
région de rejet 434
régression, analyse de 436
voir analyse de régression 15
régression LAD 436
régression linéaire généralisée 439
régression linéaire multiple 443
régression linéaire simple 446
régression logistique 451
régression non linéaire 453
régression ridge 456
relation 459
représentation graphique 459
résidu 460
Rice Stuart Arthur 461
risque 462
risque attribuable 463
risque évitable 465
risque relatif 466

S

série chronologique 471
seuil de signification 475
signe, test du 476
voir test du signe 563
simulation 476
Snedecor George Waddel 477
sondage 478
Spearman, table de 478
voir table de Spearman 498
statistique 478
statistique, table 481
voir table statistique 503
statistique bayesienne 481
statistique descriptive 485
statistique génétique 485
statistique inférentielle 487
statistique officielle 488
statistique spatiale 491
steam and leaf 492
voir diagramme stem and leaf 134
Student 492
voir Gosset William Sealy 225
Student, loi de 492
voir loi de Student 296
Student, table de 492
voir table de Student 498
voir annexe E 622
Student, test de 492
voir test de Student 544
Student pour observations pairées, test de . 492
voir test de Student pour observations pairées 548

T

table binomiale 495
voir annexe B 615
table de contingence 496
table de Fisher 496
voir annexe C 618
table de Kruskal-Wallis 497
voir annexe D 620

table de Spearman 498
table de Student 498
voir annexe E 622
table de Wilcoxon 499
table de Wilcoxon signé 500
table du chi-carré 501
voir annexe F 623
table normale 502
voir annexe G 624
table statistique 503
tableau de distances 504
tableau de fréquences 505
taille d'échantillon 507
taux d'incidence 509
taux de prévalence 510
Tchebychev Pafnouti Lvovitch 511
tendance séculaire 512
test bilatéral 515
test binomial 516
test d'adéquation 518
test d'adéquation du chi-carré 519
test d'Anderson-Darling 522
test d'homogénéité 524
test d'hypothèse 525
test d'indépendance 528
test d'indépendance du chi-carré 528
test de Durbin-Watson 531
test de Fisher 533
test de Kolmogorov-Smirnov 534
test de Kruskal-Wallis 537
test de la différence significative minimale . 539
test de Mann-Whitney 541
test de Student 544
test de Student pour observations pairées .. 548
test de Wilcoxon 549
test de Wilcoxon signé 552
test du chi-carré 555
test du rapport de vraisemblance 557
test du signe 563
test non paramétrique 566
test paramétrique 567
test unilatéral 568
théorème central limite 569
théorème de Bayes 571
théorème de Bernoulli 572
théorème de convergence 573
théorème de Gauss-Markov 573
théorème de Moivre-Laplace 574
trace 576
traitement 576
transformation 577
transformation logarithmique 577
transposée 578
triangle arithmétique 578
Tukey John Wilder 579

U

uniforme, loi 583
voir loi uniforme 310
uniforme discrète, loi 583
voir loi uniforme discrète 312
unilatéral, test 583
voir test unilatéral 568
unité expérimentale 583

V

valeur 587
valeur critique 587
valeur p 588
valeur propre 589
variable 590
variable aléatoire 590
variable dépendante 591
variable dichotomique 591
variable indépendante 591
variable qualitative catégorielle 592
variable quantitative 592
variance 592
variance d'une variable aléatoire 594
variance pondérée 596
variation irrégulière 597
variation saisonnière 598
vecteur 599
vecteur propre 600
von Mises Richard 601

W

Wilcoxon, table de 605
voir table de Wilcoxon 499
Wilcoxon, test de 605
voir test de Wilcoxon 549
Wilcoxon Frank 605
Wilcoxon signé, table de 605

voir table de Wilcoxon signé 500
Wilcoxon signé, test de 605
voir test de Wilcoxon signé 552
Willcox Walter Francis 606

Y

Youden William John 609
Yule George Udny 609