CENTRALE
PARIS

ECOLE CENTRALE PARIS

1ère ANNÉE d'ÉTUDES

PROBABILITES ET STATISTIQUE

Thérèse PHAN

2004 - 2005 **Tome 2**

CENTRALE
PARIS

ECOLE CENTRALE PARIS

1ère ANNÉE d'ÉTUDES

STATISTIQUE

Thérèse PHAN

2004 - 2005

Table des matières

Présentation

Ce polycopié consacré à la Statistique comporte deux parties : une première partie traitant de la Statistique exploratoire et une deuxième regroupant toutes les techniques les plus utilisées de la Statistique inférentielle.

La **Statistique exploratoire** regroupe un ensemble de méthodes permettant l'analyse de données de nature aléatoire ou de phénomènes dont le comportement dépend du hasard. Ces phénomènes sont de natures aussi différentes que : réponses à une enquête d'opinion, mesure des taux de pollution de véhicules, relevé journalier des hauteurs d'enneigement dans une station de sports d'hiver…

Ces méthodes ont pour finalité de synthétiser la masse d'informations récoltées et, éventuellement, d'en déduire des conclusions sur la population étudiée, voire de prévoir des règles auxquelles obéiraient ces phénomènes.

Nous débuterons cette première partie du cours de Statistique par un exposé sur la **Statistique descriptive.** Elle a pour but de synthétiser, de résumer, de structurer l'information contenue dans les données sous forme de **tableaux, de graphiques, d'indicateurs numériques**…

L'étude des diverses mesures de liaison entre deux variables fait l'objet du deuxième chapitre de cette partie consacrée à l'analyse exploratoire.

De nombreuses techniques de visualisation des données multidimensionnelles ont enrichi cette branche de la Statistique. C'est **l'analyse des données**. Les méthodes développées sont : soit des méthodes de classification permettant de constituer des groupes homogènes dans la population : **méthodes de partitionnement, classification hiérarchique**…soit des méthodes factorielles réduisant le nombre de variables à l'aide de composantes synthétiques : **Analyse en Composantes Principales, Analyse des Correspondances**…

Certaines de ces méthodes sont étudiées dans des cours d'approfondissement.

La **Statistique inférentielle** a pour objectifs de généraliser à une population toute entière des propriétés constatées sur un échantillon ou de valider ou non des hypothèses émises suite à une première étude.

Le premier objectif est exposé dans les chapitres consacrés à l'échantillonnage ainsi qu'à l'estimation ; le calcul des probabilités y joue un rôle essentiel car les résultats obtenus dans cette partie en découlent.

Le deuxième objectif fait l'objet des chapitres consacrés aux différents tests : tests sur une hypothèse émise ou que l'on conteste, tests d'ajustement d'une distribution théorique sur une distribution empirique, tests de comparaisons de populations, tests d'indépendance de deux caractères… Cette dernière partie se termine par l'étude de la régression.

Première partie

Statistique exploratoire

Ch. I Statistique descriptive

I- 1) DESCRIPTION UNIDIMENSIONNELLE DE DONNEES

Usuellement les données analysées se présentent sous la forme d'un tableau croisé de p variables étudiées sur n individus. Pour étudier ces données, on peut commencer par analyser séparément chaque variable. C'est la description unidimensionnelle, phase préliminaire de toute étude statistique. Cette description utilise des tableaux, des représentations graphiques et des caractéristiques numériques.

I-1-1 Tableaux statistiques

Ces tableaux se présentent de manière différente suivant la nature des variables, en particulier selon si la variable est discrète ou continue.

I-1-1-1. Variable discrète

Un tableau statistique décrivant une variable discrète présente usuellement, pour chaque valeur de la variable, **la fréquence relative ou absolue** (voire les deux) de cette valeur.

Exemple I-1

On a relevé la contamination en fer de l'huile moteur de 60 véhicules en brûlant, pour chaque véhicule, un échantillon et en observant la lumière émise. Les valeurs trouvées sont données, en parts par million, dans la première ligne du tableau I-1. Dans la deuxième ligne, on a indiqué le nombre de véhicules correspondant à chacune des mesures :

Tableau I-1

95	96	97	98	99	100	101	102	103	104	105	106	107	108	109	110
1	3	4	4	7	14	2	8	5	3	1	2	2	2	1	1

I-1-1-2. Variable continue

Les données statistiques d'un tableau décrivant une variable continue sont regroupées en **classes**. Dans un tableau de ce type, figurent les extrémités e_i des classes ainsi que les effectifs de ces classes c'est à dire le nombre d'individus appartenant à chaque classe $[e_{i-1}, e_i[$. Par convention, l'extrémité droite de chaque classe est exclue de cette classe.

L'**amplitude** d'une classe est la longueur de l'intervalle correspondant. On peut aussi indiquer les **fréquences relatives** de chaque classe ainsi que les **fréquences cumulées** représentant les effectifs cumulés des classes d'extrémités inférieures.

Exemple I-2

Le tableau I-2 présente la répartition, en pourcentage P, des 1 800 ingénieurs d'un groupe industriel en fonction du nombre N d'années d'ancienneté dans le groupe :

Tableau I-2

N	[0, 1[	[1, 2[	[2, 3[	[3, 4[	[4, 6[	[6, 8[	[8, 10[	[10,14[	[14,18[	[18,22[	≥ 22
P en %	1,0	2,0	5,2	7,3	8,1	12,3	14,7	16,3	13,6	11,2	8,3

On peut remarquer que les classes ne sont pas toutes de même amplitude. La dernière classe est dite **ouverte** car n'y figure pas de borne supérieure.

Il est fréquent, en statistique, d'avoir à « classer » une série de données. Il faut alors déterminer le nombre « idéal » de classes :

- ✓ trop de classes n'apportent pas de simplification notoire et les effectifs de chaque classe ne sont pas suffisamment représentatifs.
- ✓ trop peu de classes fait perdre des informations.

Usuellement et suivant le nombre de données, on recommande de prendre entre 5 et 20 classes.

La formule de Sturges donne une valeur approchée du nombre k de classes souhaitable en fonction du nombre n d'observations :

$$k \cong 1 + 3{,}222 Log_{10} n$$

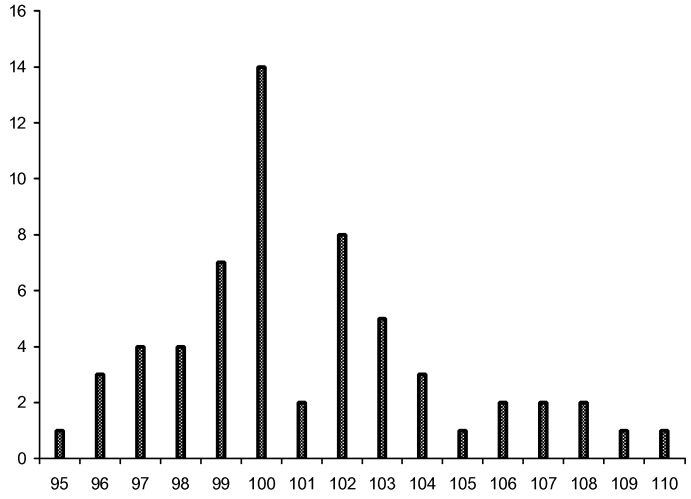
Ainsi, pour n = 100, on obtient k ≅ 7 classes et pour n = 10 000, on obtient k ≅ 14 classes.

I-1-2 Représentations graphiques

I-1-2-1. Diagramme en bâtons

Un diagramme en bâtons est obtenu en portant en ordonnée, pour chaque valeur de la variable mise en abscisse, la fréquence relative correspondante.

Exemple I-1 : diagramme des mesures de contamination en fer de l'huile moteur



I-1-2-2. Diagramme « stem and leaf »

Ce diagramme, appelé aussi diagramme « tige-feuille », dû à Tukey, réalise une sorte d'histogramme couché à partir des valeurs numériques constituant la série étudiée. Chaque donnée est décomposée en deux parties :

- ➢ **La tige** comprenant les chiffres principaux de chaque valeur numérique (usuellement le chiffre des centaines et celui des dizaines)
- ➢ **La feuille** comprenant les autres chiffres (le chiffre des unités).

Exemple II-2

Le tableau II-2 présente le relevé des poids, en grammes, de 24 éprouvettes.

Tableau II-2

263	285	256	258	274	261	250	265	276	271	272	290
260	276	270	279	288	284	253	286	287	281	290	273

Le diagramme « stem and leaf » est composé de la partie tige comprenant les deux premiers chiffres des mesures et de la partie feuille indiquant le chiffre des unités des mesures :

25 | 0 3 6 8
26 | 0 1 3 5
27 | 0 1 2 3 4 6 6 9
28 | 1 4 5 6 7 8
29 | 0 0

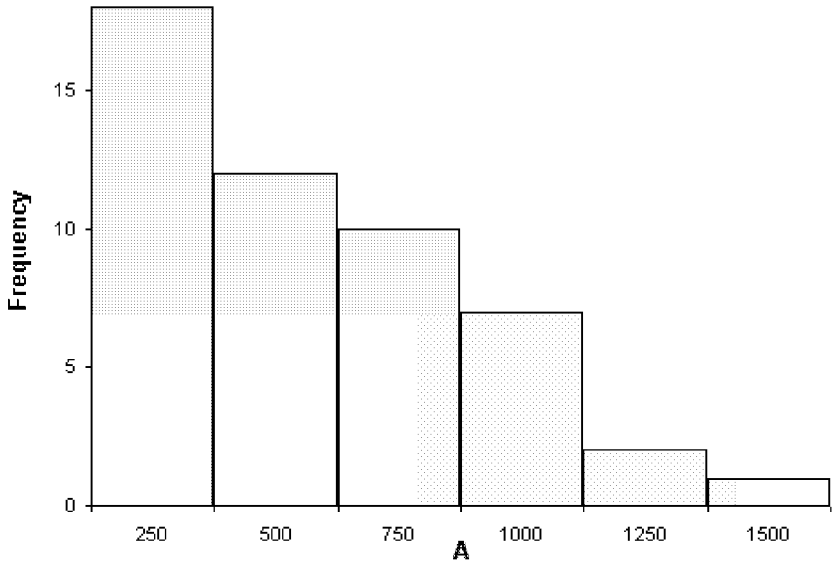
Une lecture rapide du diagramme nous permet de dire que les poids sont tous entre 250 et 290 grammes et que la moyenne se situe entre 270 et 280 grammes.

I-1-2-3. Histogramme

Un histogramme est composé de rectangles dont la largeur de chacun en abscisse est la largeur de la classe correspondante et dont la longueur en ordonnée est telle que la surface du rectangle est proportionnelle à l'effectif de la classe.

Dans le cas de classes de même amplitude, on reporte en ordonnée la fréquence de chaque classe.

Exemple II-3 : histogramme de la durée de vie de 50 lampes



I-1-2-4. Polygone des fréquences

Ce polygone est obtenu en joignant par des segments de droite les milieux des cotés supérieurs des rectangles de l'histogramme correspondant. La courbe obtenue est fermée en créant deux classes fictives d'effectif nul à chaque extrémité.

Il permet de représenter la distribution des fréquences.

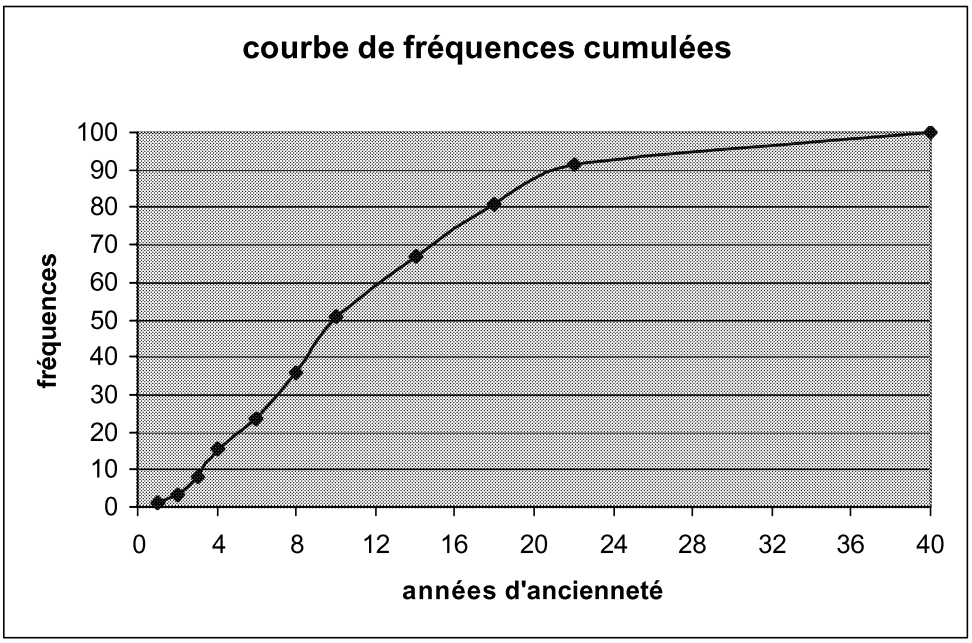
I-1-2-5. Courbes de fréquences cumulées

On considère les deux courbes de fréquences cumulées (croissante ou décroissante) construites à partir des fréquences relatives ou absolues.

La courbe cumulative croissante est construite en joignant les points d'abscisses les limites supérieures des classes et d'ordonnées les fréquences cumulées croissantes.

La courbe cumulative décroissante joint, elle, les points ayant pour abscisses les limites inférieures des classes et pour ordonnées les fréquences cumulées décroissantes.

Exemple I-2 : courbe de fréquences cumulées d'ingénieurs en fonction de l'ancienneté.



I-1-2-6. Courbe de concentration

Elle est très utilisée en statistique économique pour l'étude d'une variable positive cumulative telle que le revenu ou le chiffre d'affaire ou la consommation…

Considérons une variable X correspondant, par exemple, à une distribution de revenus, de fonction de répartition F et de masse totale M. La courbe de concentration est définie par l'ensemble des points tels que, pour chaque valeur x de la variable, l'abscisse est F(x), proportion des individus gagnant moins que x, et l'ordonnée G(x) définie par :

$$G(x) = \frac{Masse\ des\ revenus < x}{Masse\ totale}$$

Remarque :

Pour une distribution quelconque : $F(x) > G(x)$.

La courbe de concentration est donc en dessous de la première bissectrice. Son premier point est l'origine des axes, le dernier le point de coordonnées (1, 1). Elle est toujours située dans le carré de longueur 1 dont deux cotés sont les axes.

L'indice de concentration, ou indice de Gini, est le double de l'aire définie entre la courbe de concentration et la première bissectrice.

Cet indice est compris entre 0 et 1. Plus il est petit, plus la courbe est proche de la bissectrice et donc les valeurs de F et G peu éloignées

I- 2) CARACTERISTIQUES NUMERIQUES

I-2-1 Indicateurs de valeur centrale

I-2-1-1. Médiane

La médiane partage l'ensemble des valeurs observées (classées par valeurs croissantes ou décroissantes) en deux sous-ensembles d'effectifs égaux.

Pour une variable discrète dont les valeurs ont été classées par ordre croissant, et dont la série des valeurs comporte un nombre impair de données égal à 2n +1, la médiane est la $n^{ième}$ valeur. Si la série comporte un nombre pair de données égal à 2n, la médiane est, par convention, la moyenne entre les deux valeurs de rang n et (n+1).

Dans le cas d'une variable continue pour laquelle on a une répartition en classes, on cherche la **classe médiane** $[e_{i-1}, e_i[$ telle que :

$$F(e_{i-1}) < 0,5 \quad \text{et} \quad F(e_i) > 0,5$$

puis on détermine la médiane M par interpolation linéaire à l'intérieur de la classe.

C'est un indicateur de position insensible aux variations des valeurs extrêmes.

I-2-1-2. Moyenne arithmétique

- ✓ **La moyenne arithmétique $\bar{x}$ d'une variable discrète dont les valeurs sont x_1, x_2, …, x_n est définie par:**

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- ✓ **La moyenne arithmétique $\bar{x}$ d'une variable continue présentée en k classes telles que f_i est la fréquence et c_i le centre de la i ème classe, est définie par :**

$$\bar{x} = \sum_{i=1}^{k} f_i c_i$$

La moyenne est très utilisée car elle représente la série par un seul nombre mais elle est peu **robuste** car sensible aux valeurs extrêmes.

I-2-1-3. Médiale

La médiale est la valeur partageant la masse de la variable en deux parties égales.

Dans l'exemple I-1, la médiale partage la masse des ingénieurs du groupe industriel en deux sous-ensembles tels que le nombre total d'années d'ancienneté du premier sous-ensemble soit égal au nombre total d'années d'ancienneté du second.

I-2-1-4. Mode ou classe modale

Le mode est la valeur la plus fréquente d'une variable discrète.

Il n'existe pas toujours, il peut y en avoir plusieurs.

Pour une variable continue, **la classe modale** est la classe correspondant à la longueur la plus grande des rectangles de l'histogramme.

I-2-1-5. Exemples

- Exemple d'une variable discrète

Reprenons l'exemple I-1 du relevé des mesures de contamination en fer de l'huile moteur.

La médiane est égale à 100 et le mode est lui aussi égal à 100. Quant à la moyenne, elle est obtenue par :

$$\bar{x} = \frac{1}{60}(95\times 1 + 96\times 3 + \ldots + 110\times 1) = 80{,}1$$

- Exemple d'une variable continue

L'exemple I-2 présente la répartition des 1 800 ingénieurs en fonction de leurs années d'ancienneté dans le groupe industriel dont ils font partie.

Le calcul de la moyenne se fait ainsi (on fixe à 24 le centre de la dernière classe) :

$$\bar{x} = 0.5\times 0.01 + 1.5\times 0.02 + 2.5\times 0.052 + 3.5\times 0.073 + \ldots + 24\times 0.083$$

Cette moyenne est égale à 11, 37 années ;

La médiane appartient à l'intervalle [8, 10[; en effet, il y a 35, 9% des ingénieurs ayant moins de 8 ans d'ancienneté et 50, 6% ayant moins de 10 ans d'ancienneté. Le calcul de la médiane, par interpolation linéaire, donne le résultat suivant :

$$M = 8 + \frac{10-8}{50{,}6-35{,}9}(50-35{,}9) = 9{,}92$$

La classe modale est la classe [10, 14[

I-2-2 Indicateurs de dispersion

I-2-2-1. Etendue ou range

L'étendue d'une distribution est définie par : $w = \left|x_{\max} - x_{\min}\right|$

L'étendue est un indicateur instable car, par définition, il dépend des valeurs extrêmes.

I-2-2-2. Quartiles

Les trois quartiles Q_1, Q_2 et Q_3 sont les valeurs partageant la série des observations en quatre parties égales.

On appelle distance interquartile la quantité : $|Q_3 - Q_1|$.

Ainsi 25% des valeurs de la série sont inférieures à Q_1, 50% sont inférieures à Q_2, enfin 75% sont inférieures à Q_3. Les quartiles sont des **indicateurs de position**. On remarque que le deuxième quartile est la médiane.

Les quartiles d'une distribution en classes se font là encore, par interpolation linéaire.

On utilise aussi quelquefois les **déciles** qui partagent la série des observations en dix parties égales ou les **centiles** qui, eux, partagent la série en cent parties égales.

I-2-2-3. Diagramme en boîte ou BOX-PLOT

Le diagramme en boîte ou **boîte à moustaches**, due à Tukey, représente schématiquement les principales caractéristiques d'une variable en utilisant les quartiles.

La partie centrale de la distribution est représentée par une boîte de largeur arbitraire et de longueur la distance interquartile. La médiane est tracée à l'intérieur. La boîte est complétée par des « moustaches » correspondant aux valeurs suivantes :

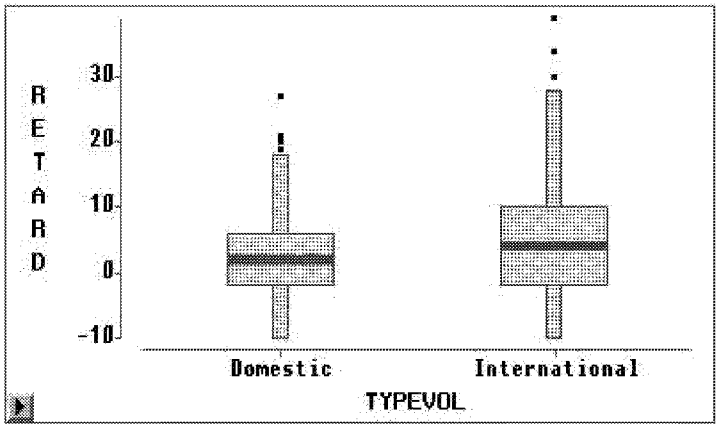
- ✓ **Valeur adjacente supérieure : plus grande valeur inférieure à : $Q_3 + 1{,}5(Q_3 - Q_1)$**
- ✓ **Valeur adjacente inférieure : plus petite valeur supérieure à : $Q_1 - 1{,}5(Q_3 - Q_1)$**

Les valeurs extérieures aux moustaches appelées valeurs aberrantes, sont représentées, en général, par des étoiles ou par des points.

La « box plot » est souvent utilisée pour comparer deux séries de mesure de la même variable.

Exemple II-4

Les retards à l'arrivée des vols d'une compagnie internationale ont été relevés sur une période donnée et on a distingué les vols nationaux (« domestic ») et les vols internationaux ; les boîtes à moustaches spécifiques à chaque catégorie ont été représentées par le logiciel SAS. Une copie d'écran est reproduite ci-dessous :



Le trait noir horizontal à l'intérieur de chaque boîte représente la médiane correspondante, les points isolés : les valeurs aberrantes…On peut remarquer, par exemple, que les médianes des deux catégories sont à peu près égales et que la distribution des retards à l'international est plus étendue.

I-2-2-4. Variance et écart-type

Ce sont les deux mesures de dispersion les plus fréquemment utilisées.

Définitions :

On appelle variance de la série des observations, le nombre positif défini par :

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

On appelle écart-type s le nombre réel positif égal à la racine carré de la variance.

Formule simplifiée :

Si on dispose d'une série de petite taille et que l'on est amené à calculer une variance à la main, on utilise la formule simplifiée qui dit que la variance est la différence entre la moyenne des carrés et le carré de la moyenne :

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}x_i^2 - \bar{x}^2$$

On obtient cette égalité en développant le carré du terme général de la somme dans la définition de la variance et en simplifiant.

I-2-2-5. Coefficient de variation

Le coefficient de variation est défini par :

$$cv = \frac{s}{\bar{x}} \times 100$$

Il ne dépend pas des unités choisies et permet d'apprécier :

- ✓ la représentativité de $\bar{x}$ par rapport à l'ensemble des données
- ✓ l'homogénéité de la distribution : un coefficient de variation d'une valeur inférieure à 15% permet de conclure à une bonne homogénéité de la distribution.

I-2-3 Indicateurs de forme

I-2-3-1. Comparaison entre mode, moyenne et médiane

Une distribution parfaitement symétrique est telle que ces trois valeurs sont égales. Il est donc intéressant de les comparer quand on souhaite ajuster une distribution symétrique telle que celle de la loi de Gauss.

I-2-3-2. Coefficient d'asymétrie ou skewness

Il est défini par :

$$\gamma_1 = \frac{m_3}{s^3} \qquad \text{où } m_3 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3$$

Il est égal à 0 pour une distribution gaussienne et à 2 pour une distribution exponentielle.

I-2-3-3. Coefficient d'aplatissement ou Kurtosis

Par définition, ce coefficient est égal à :

$$\gamma_2 = \frac{m_4}{s^4} \qquad \text{où } m_4 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4$$

Une distribution de Laplace-Gauss a un coefficient d'aplatissement égal à 3 et ce coefficient est égal à 9 pour une distribution exponentielle

Ces deux coefficients sont souvent calculés par les logiciels statistiques. Ils fournissent un premier résultat concernant la comparaison entre la distribution de l'échantillon étudié et une distribution théorique (une distribution normale, par exemple).

Ch. II Mesures de liaison entre variables

II- 1) LIAISON ENTRE DEUX VARIABLES QUANTITATIVES

II-1-1 Etude graphique de la liaison

Supposons que l'on ait observé deux variables X et Y sur un ensemble de n individus. On a obtenu n couples $(x_i, y_i)_i$ soit encore deux vecteurs X et Y de R^n :

$$X = (x_i)_i \qquad \text{et} \qquad Y = (y_i)_i$$

On peut représenter l'ensemble des points de coordonnées (x_i, y_i) dans un repère du plan. Cette représentation fournit des indications sur d'éventuelles liaisons entre les deux variables.

II-1-2 Définition de la corrélation entre deux variables

Deux variables X et Y sont corrélées, ou dépendantes en moyenne, si l'espérance conditionnelle de Y à X = x fixé, $E(Y / X = x)$, est fonction de x.

Si l'espérance conditionnelle $E(Y / X = x)$ est une fonction linéaire de x, on dit que la corrélation est linéaire.

II-1-3 Coefficient de corrélation linéaire

Ce coefficient mesure le caractère linéaire d'une liaison éventuelle entre deux variables.

II-1-3-1. Rappel

La covariance des deux variables aléatoires X et Y est égale à :

$$Cov(X,Y) = \frac{1}{n}\sum_i x_i y_i - \bar{x}\,\bar{y}$$

II-1-3-2. Définition

Le coefficient de corrélation linéaire des variables X et Y est défini par :

$$r = \frac{\mathrm{cov}(X,Y)}{s_X s_Y}$$

II-1-3-3. Propriétés

- $|r| \le +1$, r représente le cosinus de l'angle formé par les vecteurs $X - \bar{X}$ et $Y - \bar{Y}$.
- Si $|r| = 1$, il y a relation linéaire entre les deux variables et réciproquement.
- Le coefficient de corrélation r n'est pas robuste car très sensible aux valeurs extrêmes.
- Si les variables sont indépendantes, le coefficient de corrélation r est nul.
- Si le coefficient r est nul, on ne peut conclure qu'à l'indépendance **linéaire** entre les variables et non à leur indépendance.

II- 2) LIAISON ENTRE DEUX VARIABLES QUALITATIVES

II-2-1 Tableaux croisés

Les tableaux croisés ou tableaux de contingence sont les tableaux statistiques de données pour deux variables qualitatives.

Désignons par X et Y les deux variables étudiées, par p le nombre de modalités de X et q le nombre de celles de Y. Au croisement de la ligne x_i et de la colonne y_j du tableau, figure le nombre n_{ij} de données telles que $X = x_i$ et $Y = y_j$.

Y / X	y_1 …	y_j …	
x_1 …		…	
x_i …	…	n_{ij} …	$n_{i.}$
		$n_{.j}$	

II-2-2 Définitions des fréquences

On appelle fréquences marginales les expressions :

$$n_{i.} = \sum_{j=1}^{q} n_{ij} \qquad\qquad n_{.j} = \sum_{i=1}^{p} n_{ij}$$

La distribution marginale pour la variable X (respectivement pour la variable Y) est décrite dans la dernière colonne (respectivement dans la dernière ligne) du tableau de contingence.

On appelle fréquences conditionnelles les expressions :

$$n_{i/j} = \frac{n_{ij}}{n_{i.}} \qquad\qquad n_{j/i} = \frac{n_{ij}}{n_{.j}}$$

On peut aussi définir de la même façon :

- ✓ la fréquence relative de la modalité (i, j) :

$$f_{ij} = \frac{n_{ij}}{n}$$

- ✓ ainsi que les fréquences relatives marginales $f_{i.}$ et $f_{.j}$ et les fréquences relatives conditionnelles.

On appelle tableau des profils-lignes le tableau des fréquences conditionnelles $n_{i/j}$ et tableau des profils-colonnes les tableau des fréquences conditionnelles $n_{j/i}$.

II-2-3 Indépendance

L'indépendance entre les variables se traduit de manière simple par l'égalité concernant les fréquences relatives :

$$f_{ij} = f_{i.} \times f_{.j}$$

ce qui se traduit en termes de fréquences par :

$$n_{ij} = \frac{n_{i.} \times n_{.j}}{n}$$

II-2-4 Mesures de liaison

II-2-4-1. Définition de la mesure d^2

On appelle mesure de liaison d^2 entre deux variables qualitatives l'expression suivante :

$$d^2 = \sum_i \sum_j \frac{(n_{ij} - \frac{n_{i.} n_{.j}}{n})^2}{\frac{n_{i.} n_{.j}}{n}}$$

On appelle **contribution au d^2**, pour chaque case, le terme : $\frac{1}{d^2} \times \frac{(n_{ij} - \frac{n_{i.} n_{.j}}{n})^2}{\frac{n_{i.} n_{.j}}{n}}$

II-2-4-2. Propriétés

- d^2 est nul quand les deux variables sont indépendantes.
- Les contributions au d^2 mettent en évidence des relations plus ou moins importantes entre modalités de la première variable et modalités de la deuxième variable.
- $\frac{d^2}{n} \leq \inf(p-1, q-1)$

$$d^2 = \sum_i \sum_j \frac{(n_{ij} - \frac{n_{i.} n_{.j}}{n})^2}{\frac{n_{i.} n_{.j}}{n}} = n \left(\sum_i \sum_j \frac{n_{ij}^2}{n_{i.} n_{.j}} - 1 \right)$$

$$\sum_i \sum_j \frac{n_{ij}^2}{n_{i.} n_{.j}} \leq \sum_i \sum_j \frac{n_{ij}}{n_{.j}} \leq \sum_j \frac{\sum_i n_{ij}}{n_{.j}} = q$$

On en déduit que $d^2 \leq n(q-1)$. On montre de même que $d^2 \leq n(p-1)$. D'où le résultat.

II-2-4-3. Autres coefficients

D'autres coefficients ont été définis à partir du d^2. Ce sont :

✓ le coefficient de contingence de Pearson :

$$P = \left(\frac{d^2}{d^2 + n}\right)^{1/2}$$

✓ le coefficient de Cramer :

$$C = \left(\frac{d^2}{n \inf(p-1, q-1)}\right)^{1/2}$$

Leurs mesures ont pour intérêt d'être comprises entre 0 et 1, l'indépendance est représentée par la valeur 0 ; la valeur 1 correspond à une liaison fonctionnelle entre les deux variables.

II-2-4-4. Cas particulier de deux variables ayant chacune deux modalités

Soit X et Y deux variables de modalités respectives : x_1, x_2 et y_1, y_2 et de fréquences :

Y / X	y_1	y_2
x_1	a	b
x_2	c	d

On calcule alors le coefficient d^2 par l'expression :

$$d^2 = \frac{n(ad - bc)^2}{(a+b)(c+d)(a+c)(b+d)}$$

II-2-4-5. Mesure du caractère significatif de la dépendance

Quand on a calculé un des coefficients présentés au paragraphe précédent, il est essentiel de se poser la question du caractère significatif de cette mesure. La démarche repose sur la conduite d'un test. La théorie des tests en Statistique est exposée au chapitre VIII de ce polycopié. Toutefois nous allons exposer ici la marche à suivre pour l'étude de ce coefficient d^2.

On montre, en utilisant les propriétés de la loi multinomiale et de la loi du Chi-deux, que d^2 est une réalisation d'une variable D^2 suivant la loi du Chi-deux à (p-1)(q-1) degrés de liberté.

En se fixant un risque d'erreur α (généralement 5% ou encore 1%), on détermine une certaine valeur d_0, appelée seuil critique, qui a la probabilité α d'être dépassée par une variable suivant la loi du Chi -deux à (p-1)(q-1) degrés de liberté.

On rejette l'hypothèse d'indépendance si d^2 est supérieur à cette valeur critique d_0.

Les logiciels statistiques actuels tels que SAS ne donnent pas ce seuil critique d_0. Ils fournissent la probabilité P qu'a une variable du Chi –deux, à (p-1)(q-1) degrés de liberté, de dépasser le d^2 empirique, calculé à l'aide des données. Si cette probabilité est inférieure à α, alors l'hypothèse d'indépendance est rejetée.

II- 3) VARIABLE QUALITATIVE, VARIABLE QUANTITATIVE

On recherche une liaison éventuelle entre une variable quantitative (par exemple : le salaire des ingénieurs de 30 ans en France) et une variable qualitative (par exemple : l'Ecole d'ingénieurs d'origine). On a alors besoin du rapport de corrélation.

II-3-1 Rappel du coefficient de corrélation théorique

Le rapport de corrélation de la variable Y en la variable X est la mesure de liaison, non symétrique, définie par :

$$\eta^2_{Y/X} = \frac{V[E(Y/X)]}{V(Y)}$$

Rappelons simplement ici que le rapport de corrélation est maximal si T est lié fonctionnellement à X.

II-3-2 Définition du rapport de corrélation empirique

Supposons que la variable X présente k modalités d'effectifs n_1, n_2, …, n_k. Notons $\overline{y_i}$ la moyenne de Y pour la catégorie k et $\bar{y}$ la moyenne totale de Y.

On définit le rapport de corrélation empirique entre la variable qualitative X et la variable quantitative Y par :

$$e^2 = \frac{\frac{1}{n}\sum_{i=1}^{k} n_i(\overline{y_i} - \bar{y})^2}{s_y^2}$$

II-3-3 Propriétés du rapport de corrélation

- $0 \le e^2 \le 1$

Par définition : $s_y^2 = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_i^j - \bar{y})^2$

Par développement et simplification,

$$s_y^2 = \frac{1}{n}\sum_{i=1}^{k} n_i(\overline{y_i} - \bar{y})^2 + \frac{1}{n}\sum_{i=1}^{k} n_i s_i^2$$

- Si $e^2 = 0$, alors il n'y a pas de dépendance en moyenne.

en effet : $e^2 = 0$ implique, $\overline{y_i} = \bar{y}$ pour toutes les valeurs de i.

- Si $e^2 = 1$, pour une modalité i de X, tous les individus ont la même valeur et ceci pour toutes les valeurs de l'indice.

$$s_y^2 = \frac{1}{n}\sum_{i=1}^{k} n_i(\overline{y_i} - \bar{y})^2 \qquad \frac{1}{n}\sum_{i=1}^{k} n_i s_i^2 = 0$$

II-3-4 Caractère significatif du rapport de corrélation

On souhaite connaître le seuil à partir duquel on pourra conclure que le rapport de corrélation est significatif.

$$s_y^2 = \frac{1}{n}\sum_{i=1}^{k} n_i(\overline{y_i} - \overline{y})^2 + \frac{1}{n}\sum_{i=1}^{k} n_i s_i^2$$

$V_A = \frac{1}{n}\sum_{i=1}^{k} n_i(\overline{y_i} - \overline{y})^2$ s'appelle variance intercatégories ou variance expliquée par le facteur contrôlé.

$V_R = \frac{1}{n}\sum_{i=1}^{k} n_i s_i^2$ s'appelle variance intracatégories ou variance résiduelle.

Supposons que les distributions conditionnelles de Y pour chaque valeur de X sont gaussiennes. V_A est alors la réalisation d'une variable Chi-deux à (k-1) degrés de liberté et V_R celle d'une variable du Chi-deux à (n-k) degrés de liberté.

Le rapport de ces deux variables rapportées à leurs degrés de liberté respectifs suit donc une loi de Fisher-Snedecor à (k-1, n-k) degrés de liberté.

Or ce rapport s'écrit en fonction de e^2 :

$$\frac{\frac{V_A}{k-1}}{\frac{V_R}{n-k}} = \frac{\frac{e^2}{k-1}}{\frac{1-e^2}{n-k}}$$

Le test de signification de e^2 consiste alors à comparer la valeur de la variable $\frac{e^2}{k-1} \times \frac{n-k}{1-e^2}$ à la valeur de la variable $F(k-1, n-k)$ ayant la probabilité α d'être dépassée.

II- 4) LIAISON ENTRE DEUX VARIABLES ORDINALES

Il est assez fréquent de disposer de deux classements d'un même ensemble d'objets. L'exemple le plus usuel est celui du classement par deux critiques différents de n films. On dispose alors de ces deux classements :

objets	1	2	…	p	…	n
1er classement	r_1	r_2	…	r_p	…	r_n
2ième classement	s_1	s_2		s_p	…	s_n

Chaque classement est une permutation des n premiers entiers.

Etudier la liaison entre les deux variables revient à comparer les classements générés par ces deux variables.

II-4-1 Coefficient de corrélation des rangs de Spearman

II-4-1-1. Définition

Au début du siècle, le psychologue Spearman a défini le coefficient de corrélation des rangs :

$$r_s = \frac{\mathbf{cov}(r,s)}{s_r s_s}$$

Les rangs étant des permutations des n premiers entiers, on utilise alors les résultats sur la distribution discrète de la loi uniforme sur [1, n] :

$$\bar{r} = \bar{s} = \frac{n+1}{2} \qquad \text{et} \qquad s_r^2 = s_s^2 = \frac{n^2-1}{12}$$

Le coefficient s'écrit alors : $r_s = \dfrac{\frac{1}{n}\sum_{i=1}^{n} r_i s_i - (\frac{n+1}{2})^2}{\frac{n^2-1}{12}}$

Posons $d_i = r_i - s_i$ $\quad \sum_{i=1}^{n} r_i s_i = \frac{1}{2}\sum_{i=1}^{n} u_i^2 + \frac{1}{2}\sum_{i=1}^{n} v_i^2 - \frac{1}{2}\sum_{i=1}^{n} d_i^2$

Or les u_i et les v_i sont des entiers variant entre 1 et n : $\sum_{i=1}^{n} u_i^2 = \sum_{i=1}^{n} v_i^2 = \frac{n(n+1)(2n+1)}{6}$

On en déduit : $\frac{1}{n}\sum_{i=1}^{n} r_i s_i - (\frac{n+1}{2})^2 = -\frac{1}{2n}\sum_{i=1}^{n} d_i^2 + \frac{(n+1)(2n+1)}{6} - (\frac{n+1}{2})^2$

Le calcul du deuxième terme est simple : $\frac{(n+1)(2n+1)}{6} - (\frac{n+1}{2})^2 = \frac{n^2-1}{12}$

D'où l'expression du coefficient de Spearman :

$$r_s = 1 - \frac{6}{n(n^2-1)}\sum_{i=1}^{n} d_i^2$$

II-4-1-2. Propriétés

- $r_s = 1$

  les classements sont identiques, $d_i = 0$ pour tout i.

- $r_s = -1$

  les classements sont inverses l'un de l'autre.

- $r_s = 0$

  les classements sont indépendants.

II-4-1-3. Table

Sous l'hypothèse de concordance des classements, la distribution de r_s a été obtenue en considérant les n ! permutations équiprobables des rangs. Elle a été tabulée pour les petites valeurs de n.

Pour $\alpha = 0{,}05$ et en fonction des diverses valeurs de n, la table ci-dessous nous donne les valeurs de r telles que $P(|r_S| > r) = 0{,}05$:

n	10	20	30	40	50	60	70	80	90	100
r	0,648	0,447	0,362	0,313	0,279	0,255	0,235	0,220	0,207	0,197

Pour les grandes valeurs, ($n \geq 100$) la distribution de r_s est considérée comme une distribution normale de paramètres 0 et $\frac{1}{\sqrt{n-1}}$.

II-4-2 Coefficient de corrélation des rangs de Kendall

II-4-2-1. Définition

Soit r_i, r_j *et* s_i, s_j les rangs, dans les deux classements, d'un couple de deux objets (i, j).

Au couple (i, j) on attribue :

+1 si les deux objets sont dans le même ordre : $r_i < r_j$ *et* $s_i < s_j$

–1 si les deux classements sont discordants : $r_i < r_j$ *et* $s_i > s_j$

On somme les valeurs obtenues pour les $\frac{n(n-1)}{2}$ couples (i, j) distincts. Soit S le résultat.

Le coefficient de Kendall est défini par :

$$\boldsymbol{\tau = \frac{2S}{n(n-1)}}$$

II-4-2-2. Propriétés

- $-1 \leq \tau \leq +1$

en effet d'après la définition de S : $-\frac{n(n-1)}{2} \leq S \leq \frac{n(n-1)}{2}$

- Si $\tau = +1$

  les classements sont identiques.

- Si $\tau = -1$

  les classements sont inversés.

II-4-2-3. Caractère significatif du coefficient

Sous l'hypothèse d'indépendance des deux classements, la loi du coefficient τ est approximativement une loi normale :

$$LG\left[0,\ \sqrt{\frac{2(2n+5)}{9n(n-1)}}\right]$$

On considère que l'approximation est valable pour n ≥ 8.

II-4-2-4. Calcul rapide du coefficient

Une méthode rapide de calcul du coefficient est la suivante :

On ordonne la première série de classement de 1 à n. Pour chaque r_i, on compte le nombre de s_j tels que $s_j > s_i$ parmi ceux pour lesquels j > i. On somme ces nombres, soit R cette somme. Alors :

$$S = 2R - \frac{n(n-1)}{2}$$

$$\tau = \frac{4R}{n(n-1)} - 1$$

II-4-2-5. Conclusion

Ces deux coefficients sont très utilisés pour tester l'indépendance de deux variables. Le coefficient de Kendall peut aussi se définir pour p classements (cf ouvrages de référence).

Ch. III Analyse en composantes principales

L'analyse en composantes principales, due aux statisticiens Pearson et Hotelling, est une méthode puissante d'étude d'un tableau de données. **Cette méthode est utilisée quand on observe un nombre élevé de variables sur un grand nombre d'individus.** L'étude individuelle de chacune de ces variables reste bien sûr très importante et doit être faite mais les liaisons éventuelles existant entre les variables peuvent être essentielles à observer. On est donc amené à étudier ces données dans leur caractère multidimensionnel.

La méthode d'Analyse en Composantes Principales (dite ACP) repose sur une représentation des données dans un espace de dimension faible (deux, trois voire quatre) tout en conservant le maximum d'information. Cette méthode purement descriptive, permet d'étudier les relations entre les variables et les individus mais aussi les relations existant entre les variables ou celles existant entre les individus.

III- 1) REPRESENTATION DES OBSERVATIONS

Les observations de p variables sur n individus sont présentées sous forme d'un tableau rectangulaire X de n lignes et p colonnes où i indice les lignes et j les colonnes.

$$\begin{pmatrix} x_1^1 & \dots & x_1^p \\ \dots & x_i^j & \dots \\ x_n^1 & \dots & x_n^p \end{pmatrix}$$

x_i^j est la valeur de la j^{ème} variable sur le i^{ème} individu. La j^{ème} variable est donc représentée par la colonne d'indice supérieur j :

$$X^j = \begin{pmatrix} x_1^j \\ \dots \\ x_n^j \end{pmatrix}$$

et l'individu E_i par la ligne d'indice inférieur i :

$$E_i = \begin{pmatrix} x_i^1 & \dots & x_i^p \end{pmatrix}$$

III-1-1 Centre de gravité

Chaque individu est un point défini par p coordonnées ; il peut être considéré comme un élément d'un espace vectoriel E appelé espace des individus. L'ensemble de ces n individus est alors un nuage de points dans cet espace E.

Le vecteur g des moyennes arithmétiques de chaque variable définit le point moyen ou centre de gravité du nuage :

$$g' = \begin{pmatrix} \bar{x}^1, \bar{x}^2, \dots, \bar{x}^p \end{pmatrix}$$

III-1-2 Matrice des poids

Quand les individus ont tous même importance (tirage aléatoire équiprobable par exemple), ils interviennent en rapport 1/n dans le calcul des caractéristiques de l'échantillon. Dans certaines applications, il est nécessaire de travailler avec des poids différents suivant les individus (données regroupées…). Ces poids sont des nombre positifs de somme égale à 1 et sont présentés dans une matrice carrée diagonale D :

$$D = \begin{pmatrix} p_1 & & & 0 \\ & p_2 & & \\ & & \ldots & \\ 0 & & & p_n \end{pmatrix}$$

Dans le cas le plus fréquent de poids tous égaux , on a bien sûr :

$$D = \frac{1}{n} I$$

III-1-3 Matrice de variance-covariance, matrice de corrélation

III-1-3-1. La matrice V de variance-covariance

Elle a pour éléments les covariances des variables :

$$v_{ij} = Cov(X^i, X^j)$$

Ses éléments diagonaux sont donc les variances des variables :

$$v_{ii} = S_i^2$$

III-1-3-2. La matrice R des corrélations

Ses éléments sont les coefficients de corrélation linéaire entre les p variables prises deux à deux :

$$r_{ij} = \frac{Cov(X^i, X^j)}{S_i S_j}$$

Ses éléments diagonaux sont égaux à 1.

III-1-3-3. Rappel sdes propriétés du coefficient de corrélation

- ✓ Si les variables X^i et X^j sont indépendantes, alors leur coefficient de corrélation r_{ij} est nul.
- ✓ Si les variables X^i et X^j sont gaussiennes et si leur coefficient de corrélation r_{ij} est nul alors les deux variables sont indépendantes.
- ✓ Si le coefficient de corrélation r_{ij} est voisin de ±1, on peut supposer qu'il existe une relation linéaire entre X^i et X^j.

III- 2) ESPACE DES INDIVIDUS, ESPACE DES VARIABLES

III-2-1 Espace des individus

III-2-1-1. Choix d'une métrique

Pour munir E espace des individus d'une structure euclidienne, il est nécessaire de définir une distance sur cet espace. Le choix de cette distance est essentielle pour l'étude statistique qui est en cours. Les caractères apparaissant en Statistique sont très divers et mesurer une distance entre individus décrits par des critères aussi différents que : âge, profession, catégorie socio-professionnelle, salaire, composition de la famille…est un problème ardu.

La distance entre deux individus E_i et E_j est définie usuellement par la forme quadratique :

$$d^2(E_i, E_j) = (E_i - E_j)' M (E_i - E_j)$$

où M est une matrice carrée symétrique définie positive de taille p.

Le produit scalaire associé est :

$$(E_i, E_j) = E_i' M E_j$$

Le choix de la matrice M définit la métrique associée.

En ACP, on utilise généralement :

- ✓ Soit M = I, on travaille alors avec la métrique usuelle
- ✓ Soit M est égale à la matrice diagonale des inverses des variances. Cette métrique, la plus usitée et celle, par défaut, de nombreux logiciels, revient à diviser chaque observation par son écart-type.

$$M = D_{1/s^2} = \begin{pmatrix} 1/s_1^2 & & & 0 \\ & 1/s_2^2 & & \\ & & \ldots & \\ 0 & & & 1/s_p^2 \end{pmatrix}$$

La distance entre deux individus ne dépend plus des unités de mesure ce qui est essentiel quand les variables n'ont pas les mêmes unités. De plus, chaque variable a la même importance quelle que soit sa dispersion.

On a vu qu'utiliser la métrique diagonale des inverses des variances revient à multiplier les caractères par l'inverse de leur écart-type ; Tout se passe ensuite comme si on utilisait la métrique I sur le tableau des données transformées.

III-2-1-2. Inertie du nuage

On appelle inertie totale du nuage de points la moyenne pondérée des carrés des distances des points au centre de gravité :

$$I_g = \sum_i p_i \|E_i - g\|^2$$

On peut alors observer que l'inertie vérifie :

$$I_g = Trace\ (MV) = Trace\ (VM)$$

Suivant le choix de M, on obtient pour I les valeurs suivantes :

- ✓ Si M est la matrice identité, alors l'inertie est égale à la somme des variances des p variables
- ✓ Si M est la matrice D_{1/s^2}, l'inertie est égale au nombre p des variables. Elle ne dépend donc pas de leurs valeurs.

III-2-2 Espace des variables

Les variables sont représentées par les colonnes du tableau X. De la même façon que pour l'espace des individus, il est nécessaire de munir l'espace des variables d'une métrique pour étudier leurs éventuelles proximités. Toutefois, dans ce cas, il n'y a pas d'hésitation sur la matrice caractérisant la métrique : la matrice D des poids est le meilleur choix ; il en découle les propriétés suivantes :

- ✓ Pour deux variables centrées, le produit scalaire de ces variables est la covariance :

$$\left\langle X^j, X^k \right\rangle = (X^j)' D X^k = \sum_{i=1}^{n} p_i x_i^j x_i^k = Cov(X^j, X^k)$$

- ✓ La norme (ou longueur) d'une variable est égale à son écart-type :

$$\left\| X^j \right\| = S_j$$

- ✓ Le cosinus de l'angle formé par deux variables centrées est égal à leur coefficient de corrélation linéaire :

$$Cos(X^j, X^k) = \frac{\left\langle X^j, X^k \right\rangle}{\left\| X^j \right\| \times \left\| X^k \right\|} = \frac{Cov(X^j, X^k)}{S_j S_k} = r_{jk}$$

Dans l'espace des individus on s'intéressera aux distances entre points tandis que dans l'espace des variables, on privilégiera l'étude des angles.

III- 3) METHODE DE L'ANALYSE

III-3-1 Présentation

Cette méthode repose sur la représentation du nuage de points formé par les n individus dans un espace de dimension plus faible que la dimension de l'espace des variables, en effectuant une projection sur cet espace.

L'espace de projection noté F_k est choisi selon plusieurs critères :

- ✓ le premier critère est pratique : l'espace F_k il doit être de dimension k faible (en pratique deux ou trois) si on veut pouvoir en tirer des résultats de façon simple
- ✓ un deuxième critère consiste à imposer une projection déformant le moins possible les distances entre les individus ; le sous espace F_k recherché est donc tel que la moyenne des carrés des distances entre projections soit la plus grande possible
- ✓ enfin, l'espace F_k est engendré par de nouvelles variables indépendantes.

III-3-2 Inertie du nuage projeté

On souhaite diminuer au minimum la distance entre projections. Il est donc nécessaire que l'inertie du nuage projeté sur F_k soit maximale. Soit P l'opérateur de projection sur F_k.

L'inertie du nuage projeté est alors égal à la trace de la matrice PVP'M.

On peut remarquer :

$$\begin{aligned} \text{Trace}(PVP'M) &= \text{Trace}(PVMP) && \text{car } P'M=MP \\ &= \text{Trace}(VMP^2) && \text{car Trace } (XY)=\text{Trace } (YX) \\ &= \text{Trace}(VMP) && \text{car } P^2=P \end{aligned}$$

Le choix du sous-espace de projection est donc ramené au choix d'un projecteur P orthogonal de rang k maximisant la trace de VMP.

En rappelant que le projecteur associé à la somme directe de deux sous-espaces orthogonaux est la somme des projecteurs associés à chacun de ces sous-espaces, nous obtenons un procédé simple pour obtenir P et donc F_k :

On procédera de proche en proche en cherchant d'abord le sous-espace de dimension 1 d'inertie maximale puis le sous-espace de dimension 1 orthogonal au précédent et d'inertie maximale etc…

Idéalement, le sous-espace F_k sera de dimension 2 ou 3 et le cumul d'inertie dépassera 65 ou 70%.

III-3-3 Axes et Facteurs principaux, composantes principales

III-3-3-1. Axes principaux

On a vu qu'il fallait d'abord chercher l'axe de l'espace R^p passant par le centre de gravité et maximisant l'inertie du nuage projeté sur cet axe puis procéder de proche en proche pour déterminer le sous-espace F_k. On démontre le résultat suivant :

Le sous-espace F_k de dimension k est engendré par les k vecteurs propres de VM associés aux k plus grandes valeurs propres.

On définit alors les axes principaux de la façon suivante :

On appelle axes principaux d'inertie les vecteurs propres de la matrice VM normés. Ils sont au nombre de p.

Soit λ_i une valeur propre de la matrice VM et a_i son vecteur propre associé :

$$VMa_i = \lambda_i a_i$$

III-3-3-2. Facteurs principaux

Les facteurs principaux sont définis à l'aide des axes principaux :

On appelle facteur principal u associé à l'axe principal a le vecteur u = Ma.

D'après la relation précédente, en composant par la matrice M à gauche :

$$MVMa_i = \lambda_i Ma_i$$

$$u_i = Ma_i$$

$$MVu_i = \lambda_i u_i$$

Les facteurs principaux sont les vecteurs propres normés de la matrice MV. Chaque facteur principal est associé à la même valeur propre que l'axe principal qui le définit.

En pratique, on s'intéresse aux facteurs principaux qu'on obtient par diagonalisation de la matrice MV.

III-3-4 Composantes principales

III-3-4-1. Définition

Chaque composante principale c_i est définie comme étant la variable (élément de R^n) composée des coordonnées des projections des individus sur l'axe principal a_i.

Les composantes principales c_i sont liées aux facteurs principaux par la relation :

$$\mathbf{c_i = Xu_i}$$

La variance d'une composante principale c_i est égale à la valeur propre λ_i associée au facteur principal.

Après avoir déterminé les facteurs principaux, on obtient les composantes principales par la relation précédente.

III-3-4-2. Reconstitution

On montre, par un calcul de matrices, qu'on peut reconstituer la matrice des données X à l'aide des composantes principales et des facteurs principaux :

$$X = \sum_{j=1}^{p} c_j u'_j M^{-1}$$

III-3-5 Conséquences du choix de la métrique

On a vu précédemment que deux métriques étaient les plus utilisées ; on va déterminer les facteurs principaux et composantes principales pour ces deux métriques.

III-3-5-1. Métrique identité

$$M = I$$

Les axes principaux et facteur principaux sont confondus. Ce sont les vecteurs propres de la matrice de variance-covariance V.

L'inconvénient majeur de cette métrique est que les résultats obtenus ne sont pas invariants par changement d'unités de mesure des variables.

III-3-5-2. Métrique diagonale des inverses des variances

$$M = D_{1/s^2}$$

On a vu que l'usage de cette métrique est équivalent à la réduction des variables.

En pratique, on associera donc au tableau X le tableau Z des variables centrées-réduites et on utilisera la métrique identité.

Dans ce cas la matrice de variance-covariance des données centrées réduites est la matrice de corrélation.

Les facteurs principaux sont les vecteurs propres de la matrice de corrélation R. Ils sont rangés par valeurs décroissantes des valeurs propres.

$$Ru = \lambda u \qquad \text{et } \|u\| = 1$$

Ayant les facteurs principaux, on obtient les composantes principales :

$$c = Zu$$

On démontre la propriété suivante :

La première composante principale c est la variable la plus liée aux variables X^j d'origine au sens de la somme des carrés des corrélations :

$$\sum_{j=1}^{p} r^2(c, X^j) \text{ maximal.}$$

III- 4) INTERPRETATION D'UNE ACP

On a vu que cette méthode d'analyse consistait en la construction de nouvelles variables. On suppose dans ce paragraphe que nous avons choisi la métrique de la diagonale des inverses des variances.

III-4-1 Corrélation entre composantes et variables initiales

L'ACP remplace les variables d'origine X^1, X^2, …, X^p, corrélées entre elles, par de nouvelles variables c^1, c^2, …, c^p, appelées composantes principales, combinaisons linéaires des variables d'origine ; Les composantes principales sont non corrélées, de variance maximale et liées aux variables d'origine au sens de la somme des carrés des corrélations.

L'ACP est une méthode factorielle linéaire.

III-4-1-1. Relation entre les variables d'origine et les composantes principales

On calcule les coefficients de corrélation linéaire entre les composantes principales et les variables initiales ; ce calcul donne un résultat particulièrement simple avec la métrique choisie :

$$r(c, X^j) = \sqrt{\lambda} \times u_j$$

Lorsque, dans l'ACP, il est décidé de projeter l'espace des individus sur un plan, on met en évidence les corrélations entre variables et composantes principales à l'aide du cercle des corrélations.

III-4-1-2. Cercle des corrélations

Chaque variable initiale X^j est représentée, dans le plan (c_1, c_2), par un point ayant pour coordonnées les deux coefficients de corrélation :

$$r(c_1, X^j) \text{ et } r(c_2, X^j).$$

Les points représentant les variables sont tous situés (par définition du coefficient de corrélation) dans un cercle de rayon 1 appelé cercle des corrélations.

Cette représentation permet de visualiser les corrélations entre variables et composantes principales en se gardant toutefois d'interpréter des proximités de points loin de la circonférence du cercle. Il est alors intéressant de visualiser les proximités ou oppositions de certaines variables sur les axes des composantes principales.

III-4-2 Projection des individus dans l'espace des nouvelles variables

L'analyse des résultats des projections des individus dans le nouvel espace se fait à différents niveaux :

- ✓ La corrélation forte d'une variable X^j avec la première composante principale indique que les individus ayant une coordonnée positive de valeur importante sur le premier axe ont une valeur de cette variable nettement supérieure à la moyenne puisque l'origine des axes principaux est le centre de gravité du nuage.
- ✓ On recherche également à mettre en évidence l'opposition de deux individus sur un axe.

✓ La contribution d'un individu e_i à l'axe k est un renseignement supplémentaire pour l'interprétation des axes : celle-ci est égale à :

$$\frac{p_i c_{ki}^2}{\lambda_k}$$

où c_{ki} est la valeur de la k^{ème} composante pour l'individu e_i

Il faut toutefois faire attention aux individus ayant une trop forte contribution sur les premiers axes. Si cela se produit (contribution supérieure à 0,25), il est conseillé de vérifier cette donnée et si celle-ci est confirmée, il est intéressant de l'enlever de l'analyse et de la reporter en valeur supplémentaire.

III-4-3 Intérêt et limites d'une ACP

L'Analyse en Composantes Principales met en évidence des résultats contenus dans le tableau de données. Les corrélations fortes entre les premières composantes principales et certaines variables n'ont pour sens que celui qui provient de la définition même de ces composantes principales.

En revanche, il sera très intéressant de montrer qu'une variable n'ayant pas servi à l'analyse, est très corrélée avec une composante principale. Cette remarque est quelquefois utilisée pour modifier le processus d'une ACP : on cherche les axes principaux en utilisant certaines variables du tableau de données et on recherche ensuite les liens entre les composantes principales et les variables restantes. Dans ce cas, si ces variables sont numériques, on projette leurs coefficients de corrélation avec les composantes principales sur le cercle de corrélation. Si ces variables restantes sont qualitatives, on les classes en catégories et on représente les centres de gravité de ces catégories dans les plans principaux.

On peut faire de même une partition des individus entre ceux qui participent à l'ACP et ceux qu'on représente sur les axes principaux après.

Deuxième partie

Statistique inférentielle

Ch. IV Echantillonnage

IV- 1) INTRODUCTION

Le débit d'une rivière est un phénomène réel. L'étude de ce débit, pendant 10 ans, relève de la statistique descriptive. Prévoir la crue maximale en prévision de la construction d'un barrage relève de la statistique inférentielle.

L'étude d'une caractéristique d'une pièce fabriquée en grand nombre (telle que la luminosité d'une ampoule, sa durée de vie ou encore le diamètre d'une pièce mécanique) relève, elle aussi, de la statistique descriptive. Il n'est toutefois pas possible de mesurer cette caractéristique sur toutes les pièces produites. Il est alors nécessaire de se limiter à l'étude des éléments d'un échantillon. Cet échantillon devra répondre à des critères particuliers pour pouvoir représenter la population toute entière dans l'étude statistique.

La démarche statistique présente plusieurs étapes :

- **Prélèvement d'un échantillon représentatif de la population ou échantillon aléatoire** par des techniques appropriées. Cela relève de la théorie de l'échantillonnage.
- **Etude des caractéristiques de cet échantillon**, issu d'une population dont on connaît la loi de probabilité. On s'intéresse principalement à ceux issus d'une population gaussienne.

IV-1-1 Définition d'un échantillon aléatoire

Soit une population de taille N dont on étudie une caractéristique mesurable X. La composition de la population vis à vis de ce caractère X est entièrement définie par la connaissance des quantités :

$$F(x) = P(X < x)$$

Si on choisit au hasard un individu i de la population, on peut lui associer une variable aléatoire X_i dont on observe une seule réalisation x_i. On répète n fois cette expérience dans des conditions indépendantes. On associe à ces n expériences, n variables aléatoires X_1, X_2,… X_n. Elles ont toutes même distribution que X et les valeurs x_1, x_2,… x_n sont les réalisations de ces variables. Ces n valeurs (x_1, x_2, …, x_n) représentent à la fois n réalisations de la variable X et une réalisation du n-uplet de variables (X_1, X_2, …, X_n).

Définition :

Un échantillon aléatoire est un n-uplet (X_1, X_2,…, X_n) de n variables aléatoires indépendantes suivant la même loi qu'une variable X appelée variable aléatoire parente. Une réalisation de l'échantillon sera notée (x_1, x_2, …, x_n).

IV-1-2 Définition d'une statistique

Soit X une variable aléatoire. Considérons un n-échantillon (X_1, X_2, …, X_n) de X.

Une statistique T est une variable aléatoire fonction mesurable de (X_1, X_2, …, X_n).

$$\mathbf{T(X) = T(X_1, X_2, \ldots, X_n).}$$

A un échantillon, on peut associer plusieurs statistiques.

IV- 2) DISTRIBUTION D'ECHANTILLONNAGE DE MOYENNE

Dans tout ce paragraphe, on considère un échantillon aléatoire (X_1, X_2,…, X_n) suivant la même loi que la variable parente X.

IV-2-1 Définition de la statistique $\overline{X}$

La statistique $\overline{X}$ ou moyenne empirique de l'échantillon est définie par :

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

IV-2-2 Espérance et variance de $\overline{X}$

Soit m l'espérance et σ^2 la variance de la variable parente X ; l'espérance et la variance de la statistique $\overline{X}$ sont :

$$E(\overline{X}) = m$$

$$V(\overline{X}) = \frac{\sigma^2}{n}$$

$$E(\overline{X}) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n} nm = m$$

$$V(\overline{X}) = \frac{1}{n^2}\sum_{i=1}^{n} V(X_i) = \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n}$$ (les X_i étant supposés indépendants)

IV-2-3 Loi faible des grands nombres

IV-2-3-1. Rappel

Soit X_1, X_2, …, X_n n variables aléatoires indépendantes d'espérances m_1, m_2, …, m_n et de variances σ_1^2, σ_2^2, …, σ_n^2.

Si $\frac{1}{n}\sum_{i=1}^{n} m_i \to m$ et si $\frac{1}{n^2}\sum_{i=1}^{n} \sigma_i^2 \to 0$, quand n tend vers +∞, alors :

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \to m$$ au sens de la convergence en probabilités.

Si les variables X_i suivent toutes une même loi, d'espérance m et de variance σ^2, les hypothèses ci-dessus sont vérifiées et on en déduit la convergence suivante :

IV-2-3-2. Convergence en probabilité

La statistique $\overline{X}$ converge en probabilité vers m quand n tend vers l'infini.

IV-2-4 Loi forte des grands nombres

IV-2-4-1. Rappel

Soit X_1, X_2, …, X_n n variables aléatoires indépendantes d'espérances m_1, m_2, …, m_n et de variances σ_1^2, σ_2^2,…, σ_n^2 telles que :

$$\frac{1}{n}\sum_{i=1}^{n} m_i \to m \quad \text{et} \quad \sum_{i=1}^{\infty} \frac{\sigma_i^2}{i^2} < +\infty \quad \text{quand n} \to +\infty$$

$$\text{alors} \quad \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \to m \quad \text{presque sûrement.}$$

IV-2-4-2. Convergence presque sûre

Dans le cas de notre échantillon, toutes les variables X_i ont leurs espérances égales à l'espérance de la variable parente X et leurs variances égales à la variance σ^2 de X donc :

$$\sum_{i=1}^{n} \frac{\sigma_i^2}{i^2} = \sigma^2 \sum_{i=1}^{n} \frac{1}{i^2}$$

La série du second membre est évidemment convergente. On en déduit :

La statistique $\overline{X}$ converge presque sûrement vers m.

IV-2-5 Théorème de la limite centrale

IV-2-5-1. Rappel

Si X_1, X_2, …, X_n sont n variables aléatoires indépendantes de même loi quelconque d'espérance m et d'écart-type σ, la variable :

$$\frac{X_1 + X_2 + ... + X_n - nm}{\sigma\sqrt{n}}$$

converge en loi vers une variable normale centrée réduite quand n tend vers l'infini.

IV-2-5-2. Convergence vers une variable gaussienne

En posant $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ dans l'expression du quotient et en divisant par n, on obtient :

La variable $\dfrac{\overline{X} - m}{\sigma/\sqrt{n}}$ converge en loi vers une variable normale centrée réduite quand n tend vers l'infini.

IV-2-5-3. Application 1

Pour une taille d'échantillon n suffisamment grande, on peut considérer que $\overline{X}$ a pour loi :

$$L(\overline{X}) \cong LG(m, \frac{\sigma}{\sqrt{n}})$$

IV-2-5-4. Application 2 : loi d'un pourcentage

Soit X la variable aléatoire représentant le nombre de succès au cours d'une suite de n répétitions indépendantes d'une même épreuve dont la probabilité de succès est p.

La loi de X est la loi binomiale de paramètres n et p notée B (n, p). X est la somme de n variables indépendantes de Bernoulli de paramètre p. Notons F la fréquence empirique du nombre de succès parmi les n épreuves :

$$F = \frac{X}{n}$$

F a pour espérance et pour variance :

$$E(F) = p \qquad V(F) = \frac{p(1-p)}{n}$$

En appliquant le théorème de la limite centrale à X somme des variables de Bernoulli :

Pour n suffisamment grand, on peut être considérer que F suit la loi normale :

$$L(F) \cong LG(p, \sqrt{\frac{p(1-p)}{n}})$$

Ce résultat est une autre formulation du théorème de « De Moivre–Laplace ».

IV- 3) ETUDE DE LA STATISTIQUE S^2

IV-3-1 Définition

La statistique S^2 ou variance empirique d'échantillon est définie par :

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

IV-3-2 Propriétés

✓ $S^2 = \frac{1}{n}(\sum_{i=1}^{n} X_i^2) - (\overline{X})^2$

✓ $S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - m)^2 - (\overline{X} - m)^2$

Pour démontrer cette dernière égalité, on élève au carré l'expression $X_i - m = (X_i - \overline{X}) + (\overline{X} - m)$ puis on somme pour toutes les valeurs de i variant de 1 à n.

$$\sum_{i=1}^{n}(X_i - m)^2 = \sum_{i=1}^{n}(X_i - \overline{X})^2 + \sum_{i=1}^{n}(\overline{X} - m)^2 + 2(\overline{X} - m)\sum_{i=1}^{n}(X_i - \overline{X})$$

La dernière somme est clairement nulle ; on divise par n et on obtient le résultat.

IV-3-3 Convergence presque sûre

S^2 converge presque sûrement vers σ^2.

La démonstration repose sur la décomposition : $S^2 = \frac{1}{n}(\sum_{i=1}^{n} X_i^2) - (\overline{X})^2$ et sur les convergences presque sûres des deux termes respectivement vers $E(X^2)$ et $[E(X)]^2$.

IV-3-4 Moments de S^2

IV-3-4-1. Espérance de S^2

$$\boldsymbol{E(S^2) = \frac{n-1}{n}\sigma^2}$$

$$E(S^2) = \frac{1}{n}\sum_{i=1}^{n} E(X_i - m)^2 - E(\overline{X} - m)^2 = \frac{1}{n}\sum_{i=1}^{n} V(X_i) - V(\overline{X})$$

$$E(S^2) = \frac{1}{n}\sum_{i=1}^{n}\sigma^2 - \frac{\sigma^2}{n} = \sigma^2 - \frac{\sigma^2}{n} = \sigma^2 \times \frac{n-1}{n}$$

On peut remarquer que si on pose : $S^{*2} = \frac{n}{n-1} S^2$ alors $E(S^{*2}) = \sigma^2$

IV-3-4-2. Variance de S^2

Une démonstration un peu longue en calcul, mais ne présentant pas de difficulté, nous permet d'écrire la variance de S^2 en fonction de n, σ et de μ_4, moment centré d'ordre 4 de X :

$$\boldsymbol{V(S^2) = \frac{n-1}{n^3}[(n-1)\mu_4 - (n-3)\sigma^4]}$$

IV-3-4-3. Convergence vers une variable gaussienne

La variable $\dfrac{S^2 - \frac{n-1}{n}\sigma^2}{\sqrt{V(S^2)}}$ converge vers une variable normale centrée réduite.

Cette convergence est aussi une conséquence du théorème de la limite centrale.

En prenant les limites de l'espérance et de la variance, pour n suffisamment grand, on peut écrire :

La variable $\dfrac{S^2 - \sigma^2}{\sqrt{\frac{\mu^4 - \sigma^4}{n}}}$ converge vers une variable normale centrée réduite.

IV- 4) ECHANTILLONS GAUSSIENS

Dans ce paragraphe, on considère que la variable X suit une loi normale de moyenne m et d'écart-type σ.

$$L(X) = LG(m,\sigma)$$

Soit $(X_1, X_2,\ldots, X_n)$ un échantillon aléatoire de X.

IV-4-1-1. Etude de la moyenne de l'échantillon

La statistique $\overline{X}$ est une combinaison linéaire de n variables aléatoires gaussiennes indépendantes. C'est donc une variable gaussienne :

$$L(\overline{X}) = LG(m,\frac{\sigma}{\sqrt{n}})$$

Il s'agit ici pour $\overline{X}$ d'une loi exacte et non plus d'une loi limite.

IV-4-1-2. Etude de la statistique S^2

En utilisant la décomposition de S^2 et en divisant par σ^2 on obtient :

$$\sum_{i=1}^{n}(\frac{X_i-m}{\sigma})^2 = \frac{nS^2}{\sigma^2} + (\frac{\overline{X}-m}{\sigma/\sqrt{n}})^2$$

- ✓ Le premier membre est une somme de carrés de variables normales centrées réduites indépendantes donc une variable du Chi-deux $\chi^2(n)$,
- ✓ Le second membre est la somme de deux formes quadratiques sur ces variables, l'une de rang n-1, l'autre de rang 1. Le deuxième terme est le carré d'une variable normale centrée réduite donc une variable du Chi-deux à un degré de liberté.

On en déduit les deux théorèmes suivants (par application du théorème de Cochran) :

Théorème 1

La loi de la variable $\frac{nS^2}{\sigma^2}$ est une loi du Chi-deux à n-1 degrés de liberté:

$$L(n\frac{S^2}{\sigma^2}) = \chi^2(n-1)$$

Théorème 2

Les statistiques $\overline{X}$ et S^2 sont indépendantes.

A ces deux théorèmes, on va en ajouter un troisième ; en effet on peut remarquer que la variable $\frac{\overline{X}-m}{\sigma/\sqrt{n}}$ suit une loi LG(0, 1) et que la variable $\frac{nS^2}{\sigma^2}$ suit une loi χ^2 (n-1).

Or la variable $T = \frac{\overline{X}-m}{S}\sqrt{n-1}$ est le quotient de la variable gaussienne par la racine carrée de la variable du Chi-deux rapportée à son degré de liberté (n-1) :

$$T = \frac{\overline{X} - m}{\sigma/\sqrt{n}} \sqrt{\frac{n-1}{nS^2/\sigma^2}} = \frac{\overline{X} - m}{S}\sqrt{n-1}$$

T est donc une variable de Student à n – 1 degrés de liberté. On en déduit donc le théorème suivant :

Théorème 3

La variable $\frac{\overline{X} - m}{S}\sqrt{n-1}$ suit une loi de Student à n-1 degrés de liberté.

IV- 5) ECHANTILLONS ARTIFICIELS

Fréquemment, on a besoin de recourir à la simulation consistant à fabriquer, à l'aide d'un programme de calcul, une suite de nombres x_1, x_2, …, x_n, chacun suivant la loi voulue indépendamment les uns des autres.

Pour créer de tels échantillons, il est nécessaire de disposer au départ d'une table de nombres aléatoires (suite de tirages de chiffres de 0 à 9 équiprobables avec remise) ou d'un générateur de nombres aléatoires.

IV-5-1 Définition

Un générateur de nombres aléatoires est un algorithme fournissant une suite de nombres compris entre 0 et 1 pseudo-aléatoires (car ils sont nullement aléatoires) mais ayant en apparence toutes les propriétés d'un véritable échantillon aléatoire d'une loi uniforme sur [0, 1].

IV-5-2 Tables de nombres au hasard

Il existe de nombreuses tables de nombres au hasard obtenues par des procédés très variés comme :

- ✓ Les tables de Kendall obtenues par éclairage intermittent d'un disque tournant divisé en secteurs multiples de 10.
- ✓ Les tables de la Rand Corporation obtenues par écrêtage d'un bruit de fond…

Ces tables ne sont pas très faciles d'utilisation donc on a mis au point des procédés d'obtention de nombres au hasard à l'aide des ordinateurs.

IV-5-3 Méthodes d'obtention de nombres pseudo-aléatoires

Les méthodes les plus employées sont basées sur des suites récurrentes (donc hélas aussi périodiques…).

La méthode la plus usitée est la **méthode de Lehmer** :

$$r_{i+1} = ar_i \text{ modulo } m$$

c'est à dire que r_{i+1} est le reste de la division de ar_i par m.

En pratique, on choisit m le plus grand possible afin d'avoir la période la plus grande possible. Celle-ci est de m/4 quand a est de la forme 8t+3 ou 8t-3 et si r_0 est un entier quelconque positif impair.

Les nombres r_i / m-1 compris entre 0 et 1 sont alors considérés comme pseudo-aléatoires. Ils constituent un échantillon de la loi uniforme sur [0, 1].

Sur machine, on prend en général $m = 2^{31}$ et $a = 5^{13}$.

IV-5-4 Echantillon artificiel de n valeurs d'une variable continue

IV-5-4-1. méthode de l'anamorphose

On suppose que la variable aléatoire X dont on cherche un échantillon suit une loi dont la fonction de répartition F est continue, strictement croissante. Soit f sa fonction de densité.

Posons Y = F(X). La densité de Y est alors :

$$g(y) = \frac{f[F(y)]}{F'[F^{-1}(y)]} = 1$$

On en déduit que Y est uniformément distribuée sur [0, 1].

En tirant n nombres au hasard uniformément répartis dans [0, 1], en leur appliquant F^{-1}, on obtient un échantillon artificiel de n valeurs de la variable X.

Cette méthode est particulièrement simple à utiliser pour une loi dont la fonction de répartition est très facile à inverser ; c'est le cas, par exemple, de la loi exponentielle.

Par ailleurs, les logiciels classiques tels que Excel permettent de construire un échantillon de taille choisie pour toutes les lois classiques de Statistique.

IV-5-4-2. Cas particulier d'une loi normale

Si X suit une loi d'espérance m et d'écart-type σ, on a vu, par application du théorème de la limite centrale, que :

$\dfrac{\overline{X} - m}{\sigma/\sqrt{n}}$ converge en loi vers une variable LG(0, 1)

Ce résultat est valable en particulier pour des variables uniformes : la somme de n variables uniformes suit donc à la limite une loi normale d'espérance n/2 et de variance n/12 puisque la loi uniforme sur [0, 1] a pour espérance ½ et pour variance 1/12. En pratique, on peut l'appliquer dès que n = 12.

Pratique de construction de l'échantillon :

On tire 12 nombres au hasard entre 0 et 1, on en fait la somme s. C'est une réalisation d'une Laplace-Gauss de moyenne 6 et d'écart-type 1.

En posant : $x = m + \sigma\,(s - 6)$, on obtient une réalisation de la variable de Laplace-Gauss LG(m, σ). En répétant cette opération n fois, on obtient un n-échantillon de la loi normale.

IV- 6) METHODE DE MONTE-CARLO

La méthode de Monte-Carlo est une méthode très connue de calcul d'intégrales. Nous rappelons ici son principe.

Remarques préliminaires

- Toute intégrale peut se ramener par changement de variables à une intégrale entre 0 et 1.

➢ $I = \int_0^1 f(t)dt$ est l'espérance de $f(U)$ où U est une variable uniforme sur [0, 1].

La méthode consiste alors à estimer l'intégrale I par : $\hat{I} = \frac{1}{n}\sum_{i=1}^{n} f(u_i)$ moyenne de n valeurs de la variable $f(U)$.

On observe que : $E(\hat{I}) = I$.

Ch. V Théorie de l'estimation

Il est fréquent en Statistique de chercher une estimation d'un paramètre. Considérons un phénomène physique pour lequel une des caractéristiques suit une loi, par exemple la loi normale, dont un des paramètres est inconnu. On souhaite donner une estimation de ce paramètre à l'aide d'un échantillon extrait de la population. Cela peut être l'estimation d'une moyenne, d'une proportion, d'une variance…

La qualité de cette estimation est plus ou moins bonne, il s'agit si possible de connaître et de minimiser l'erreur faite. Etudions l'exemple suivant.

Exemple V-1

On veut obtenir une estimation de la teneur (en parts par millions) en manganèse d'un sol pauvre en matière organique (2% et moins). Vingt quatre prélèvements de ce type de sol ont été effectués dans la région concernée. On a obtenu les résultats suivants :

92	77	85	89	93	94	89	85	84	80	86	94
92	80	80	88	87	77	89	84	88	85	94	83

Il est très simple de donner la moyenne des relevés qui paraît une bonne estimation de la teneur recherchée ; on peut aussi rechercher une estimation de cette teneur sous forme d'un intervalle ayant une probabilité égale à 0, 95 de contenir la teneur recherchée.

V- 1) STATISTIQUE EXHAUSTIVE

Dans la recherche d'un estimateur d'un paramètre θ inconnu, un échantillon nous apporte une certaine information sur ce paramètre. Lorsque l'on résume cet échantillon par une statistique (moyenne ou proportion ou variance…), il s'agit de ne pas perdre cette information. Une statistique qui conserve l'information sera dite **exhaustive**.

V-1-1 Définition d'une statistique exhaustive

Soit $X_1, X_2,\ldots, X_n$ un échantillon de la variable aléatoire X dont la loi de probabilité dépend d'un paramètre θ et soit $(x_1, x_2, \ldots, x_n)$ une réalisation de cet échantillon.

V-1-1-1. Vraisemblance d'un échantillon

On appelle vraisemblance de l'échantillon $L(\underline{x},\theta)$ la fonction suivante :

- ✓ $L(\underline{x},\theta) = \prod_{i=1}^{n} f(x_i,\theta)$ **si X est continue**
- ✓ $L(\underline{x},\theta) = \prod_{i=1}^{n} P_\theta(X_i = x_i)$ **si X est discrète**

V-1-1-2. Exhaustivité d'un échantillon

Soit T une statistique fonction du n-uplet $(X_1, \ldots, X_n)$ et soit $g(t, \theta)$ la densité de T. (On suppose la variable T continue ; dans le cas discret, on remplace $g(t, \theta)$ par $P(T = t)$).

La statistique T est dite exhaustive si la vraisemblance de l'échantillon peut se factoriser ainsi :

$$L(\underline{x},\theta) = g(t,\theta)h(\underline{x})$$

T est donc exhaustive si la densité conditionnelle de l'échantillon est indépendante du paramètre. C'est le principe de factorisation.

Remarque :

Il est intéressant de disposer d'une statistique exhaustive indépendante de la taille de l'échantillon.

V-1-1-3. Propriété

Soit T une statistique exhaustive pour le paramètre θ et soit φ une fonction injective de T. Alors $S = \varphi(T)$ est aussi une statistique exhaustive de θ.

V-1-2 Exemples de statistiques exhaustives

V-1-2-1. Loi de Poisson de paramètre λ inconnu.

Soit X une variable de Poisson de paramètre λ inconnu ; la vraisemblance s'écrit :

$$L(\underline{x},\lambda) = \prod_{i=1}^{n} \exp(-\lambda)\frac{\lambda^{x_i}}{x_i!} = \exp(-n\lambda)\prod_{i=1}^{n}\frac{\lambda^{x_i}}{x_i!}$$

Posons $S = X_1 + \ldots + X_n$. S suit une loi de Poisson de paramètre $n\lambda$ dont la densité s'écrit :

$$g(s,\lambda) = \exp(-n\lambda)\frac{(n\lambda)^s}{s!}$$

Le rapport $\frac{L}{g} = \frac{s!}{n^s \prod x_i!}$ est indépendant de λ.

$S = \sum_{i=1}^{n} X_i$ est une statistique exhaustive pour le paramètre λ.

V-1-2-2. Loi normale, m connue, σ inconnu.

Soit X une variable aléatoire suivant une loi normale de moyenne m connue et d'écart-type σ inconnu ; la vraisemblance s'écrit :

$$L(\underline{x},\sigma) = \frac{1}{\sigma^n(\sqrt{2\pi})^n}\exp\left(-\frac{1}{2}\sum_{i=1}^{n}\left(\frac{x_i - m}{\sigma}\right)^2\right)$$

En posant : $T = \sum_{i=1}^{n}(X_i - m)^2$, on sait que $\frac{T}{\sigma^2}$ suit une loi du Chi-deux à n degrés de liberté.

La densité de T est donc :

$$g(t,\sigma) = \frac{1}{2^{n/2}\Gamma(\frac{n}{2})}(\frac{t}{\sigma^2})^{\frac{n}{2}-1}\exp(-\frac{t}{2}\sigma^2)\frac{1}{\sigma^2}$$

Le rapport L/g est indépendant du paramètre σ^2. On peut en déduire :

La statistique $T = \sum_{i=1}^{n}(X_i - m)^2$ est exhaustive pour le paramètre σ^2.

V-1-3 Théorème de Darmois (admis)

Soit X une variable aléatoire dont le domaine de définition ne dépend pas de θ. S'il existe une valeur n telle que l'échantillon (X_1,…, X_n) admette une statistique exhaustive, alors la densité est de la forme suivante :

$f(x,\theta) = \exp[a(x)\alpha(\theta) + b(x) + \beta(\theta)]$ (La variable est de la famille exponentielle).

Si la densité est de cette forme et si de plus l'application : $x_i \rightarrow \sum_{i=1}^{n} a(x_i)$ est bijective et continûment différentiable pour tout i, alors $T = \sum_{i=1}^{n} a(X_i)$ est une statistique exhaustive particulière.

V-1-4 Statistiques exhaustives usuelles

Nous donnons ci-dessous les statistiques exhaustives les plus usuelles :

Loi	Paramètre inconnu	Statistique
Bernoulli	p	$T = \sum_{i=1}^{n} X_i$
Gamma	r	$T = \sum_{i=1}^{n} LnX_i$
Exponentielle	θ	$T = \sum_{i=1}^{n} X_i$
LG(m, σ)	m (σ connu)	$T = \sum_{i=1}^{n} X_i$
LG(m, σ)	σ^2 (m connu)	$T = \sum_{i=1}^{n}(X_i - m)^2$
LG(m, σ)	m, σ^2	$\sum_{i=1}^{n} X_i$, $\sum_{i=1}^{n}(X_i - \overline{X})^2$

V- 2) INFORMATION DE FISHER

V-2-1 Définition

On appelle quantité d'information de Fisher $I_n(\theta)$ apportée par un n -échantillon sur le paramètre θ la quantité suivante positive ou nulle :

$$I_n(\theta) = E[(\frac{\delta \ln L}{\delta \theta})^2]$$

où L est la vraisemblance de l'échantillon.

Remarque :

$$I_n(\theta) = E[(\frac{\delta \ln L}{\delta \theta})^2] = \int_E \left[\frac{\partial LnL(\underline{x},\theta)}{\partial \theta}\right]^2 L(\underline{x},\theta) d\underline{x}$$

V-2-2 Théorème

Si le domaine de définition de X ne dépend pas de θ et si la vraisemblance est deux fois dérivable, alors :

$$I_n(\theta) = -E(\frac{\delta^2 \ln L}{\delta \theta^2})$$

L est une densité de probabilité donc $\int_E L(\underline{x},\theta) d\underline{x} = 1$

On dérive une fois par rapport à θ :

$\int_E \frac{\partial L(\underline{x},\theta)}{\partial \theta} d\underline{x} = 0$ mais le premier terme est aussi égal à : $\int_E \frac{\partial LnL(\underline{x},\theta)}{\partial \theta} L(\underline{x},\theta) d\underline{x} = 0$

On en déduit que :

$$E(\frac{\partial LnL(\underline{x},\theta)}{\partial \theta}) = 0$$, la variable est centrée ;

En dérivant une deuxième fois :

$$\int_E \frac{\partial^2 LnL(\underline{x},\theta)}{\partial \theta^2} L(\underline{x},\theta) d\underline{x} + \int_E \left[\frac{\partial LnL(\underline{x},\theta)}{\partial \theta}\right]^2 L(\underline{x},\theta) d\underline{x} = 0$$

$$E\left[\frac{\partial^2 LnL(\underline{x},\theta)}{\partial \theta^2}\right] + E\left(\left[\frac{\partial LnL(\underline{x},\theta)}{\partial \theta}\right]^2\right) = 0$$

On en déduit alors le résultat annoncé.

V-2-3 Propriétés

Si l'ensemble de définition ne dépend pas du paramètre θ, on a les propriétés suivantes :

V-2-3-1. Additivité

$$I_n(\theta) = n\, I_1(\theta)$$

Cela signifie que chaque observation a la même importance.

Cette propriété est une conséquence du théorème précédent et du fait que la vraisemblance s'écrit, par définition :

$$Ln[L(\underline{X},\theta)] = nLn[f(X,\theta)]$$

V-2-3-2. Dégradation de l'information

Si T est une statistique que l'on substitue à l'échantillon, l'information apportée par la statistique est inférieure ou égale à celle apportée par l'échantillon :

$$I_T(\theta) \leq I_n(\theta)$$

Si T est exhaustive il y a égalité.

Dans le cas d'un domaine indépendant de θ, l'égalité implique l'exhaustivité.

V-2-4 Exemple : information de Fisher pour LG(m,σ) (m inconnu)

Soit X une variable aléatoire normale de moyenne m inconnue et d'écart-type σ connu. On sait que la vraisemblance s'écrit :

$$L(\underline{X},m) = \frac{1}{\sigma^n(\sqrt{2\pi})^n}\exp(-\frac{1}{2}\sum_{i=1}^{n}(\frac{x_i-m}{\sigma})^2)$$

En passant au log : $Ln[L(\underline{X},m)] = -nLn(\sigma\sqrt{2\pi}) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-m)^2$

$$\frac{\partial Ln[L(\underline{X},m)]}{\partial m} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-m)$$

$$\frac{\partial^2 Ln[L(\underline{X},m)]}{\partial m^2} = -\frac{n}{\sigma^2} \qquad I_n(\theta) = -E[-\frac{n}{\sigma^2}] = \frac{n}{\sigma^2}$$

On en déduit que l'information est d'autant plus grande que l'écart-type est plus petit. Cette remarque justifie le mot de « précision ».

Ch. VI Estimation ponctuelle

VI- 1) GENERALITES ET EXEMPLES

L'estimation ponctuelle consiste à chercher des « estimateurs » pour les paramètres d'une population (m, σ, …) à l'aide d'un échantillon de n observations issues de cette population.

Par exemple, les lois des grands nombres permettent de prendre $\overline{X}$ et S^2 pour estimer m et σ^2. De même la fréquence empirique f d'un événement est une estimation de la probabilité p. Les variables $\overline{X}$, S^2 et f sont appelées estimateurs de m, σ^2 et p.

Un même paramètre peut être estimé par plusieurs valeurs plus ou moins satisfaisantes ; afin de déterminer le meilleur des estimateurs, on est donc amené à définir les qualités d'un estimateur.

VI- 2) QUALITES D'UN ESTIMATEUR

On appellera θ le paramètre à estimer et T un estimateur de θ. La première des qualités d'un estimateur est d'être convergent.

VI-2-1 Estimateur convergent

VI-2-1-1. Définition

Un estimateur T d'un paramètre θ est dit convergent si T converge vers θ quand n tend vers l'infini.

VI-2-1-2. Exemple

Les estimateurs précédents : $\overline{X}$, S^2 et f sont des estimateurs convergents.

On peut appliquer, par exemple, l'inégalité de Bienaymé-Tchebychev pour connaître la convergence en probabilité de $\overline{X}$ vers m :

On sait que : $E(\overline{X}) = m$ et $V(\overline{X}) = \dfrac{\sigma^2}{n}$ donc : $P(\left|\overline{X} - m\right| > k) \leq \dfrac{\sigma^2}{nk^2}$

VI-2-1-3. Vitesse de convergence

La convergence de T vers θ peut être une convergence presque sûre ou une convergence en probabilité ou encore une convergence en moyenne quadratique.

La vitesse de convergence peut varier d'un estimateur à l'autre. Elle est liée, pour une taille donnée n d'échantillon, à la précision de l'estimateur.

Dans l'exemple précédent, la vitesse de convergence est au moins en 1/n.

VI-2-2 Estimateur sans biais

Un estimateur T est une variable aléatoire dont on suppose connue la loi de probabilité pour une valeur de θ fixée.

L'erreur d'estimation est la variable aléatoire $T-\theta$ que l'on peut décomposer en :

$$T-\theta = T-E(T)+E(T)-\theta$$

où $E(T)$ est l'espérance de T.

Le premier terme $T-E(T)$ représente les fluctuations aléatoires de T par rapport à sa valeur moyenne, tandis que le deuxième terme $E(T)-\theta$ représente une erreur systématique due au fait que T varie autour de sa valeur moyenne E (T) et non autour de θ, sauf si $E(T)=\theta$.

VI-2-2-1. Définition

Soit T un estimateur du paramètre θ. On appelle biais de l'estimateur T la quantité:

$$E(T)-\theta$$

Si $E(T)=\theta$, l'estimateur T est dit sans biais.

Si $E(T)\neq\theta$, T est dit biaisé.

Si $E(T)\to\theta$, quand n $\to +\infty$, l'estimateur T est dit asymptotiquement sans biais.

Il est évidemment souhaitable de travailler avec des estimateurs sans biais.

VI-2-2-2. Exemples

- On a déjà vu que $\overline{X}$ est un estimateur sans biais de m puisque $E(\overline{X})=m$.
- S^2 est un estimateur biaisé de σ^2 car $E(S^2)=\frac{n-1}{n}\sigma^2$.

Le biais est : $E(S^2)-\sigma^2=-\frac{1}{n}\sigma^2$. S^2 est donc asymptotiquement sans biais.

En revanche, $S^{*2}=\frac{n}{n-1}S^2$ est un estimateur sans biais de σ^2 car $E(S^{*2})=\sigma^2$

Attention :

S^{*2} est un estimateur sans biais de σ^2, mais S^* n'est pas un estimateur sans biais de σ !

VI-2-3 Précision d'un estimateur

VI-2-3-1. Définition

On mesure usuellement la précision d'un estimateur T du paramètre θ par l'erreur quadratique moyenne définie par:

$$E[(T-\theta)^2]$$

VI-2-3-2. Erreur quadratique moyenne et variance

Décomposons l'erreur quadratique moyenne :

$E[(T-\theta)^2] = E[(T - E(T) + E(T) - \theta)^2]$

$E[(T-\theta)^2] = E[(T - E(T))^2 + 2E[(T - E(T))(E(T) - \theta)] + E[(E(T) - \theta)^2]$

$E[(E(T)-\theta)^2] = (E(T)-\theta)^2$ car $E(T)-\theta$ est une constante.

de plus $E[T - E(T)] = 0$ donc le deuxième terme est nul ; Finalement :

$$\boldsymbol{E[(T-\theta)^2] = V(T) + (E(T)-\theta)^2}$$

L'expression ainsi trouvée nous permet d'en déduire que :

Entre deux estimateurs sans biais, le plus précis au sens de l'erreur quadratique moyenne est celui de variance minimale.

En effet, pour un estimateur sans biais $E(T) = \theta$, et l'erreur quadratique moyenne se trouve être égale à la variance de l'estimateur.

Il sera donc utile de chercher des estimateurs sans biais de variance minimale.

VI- 3) ESTIMATION SANS BIAIS DE VARIANCE MINIMALE

VI-3-1 Théorème 1

S'il existe un estimateur de θ sans biais, de variance minimale, il est unique presque sûrement.

Supposons qu'il existe deux estimateurs sans biais T_1et T_2 de θ de variance minimale V. Soit T_3 la demi-somme des deux.

$E(T_3) = \dfrac{E(T_1) + E(T_2)}{2} = \theta$ donc T_3 est sans biais.

Soit ρ le coefficient de corrélation linéaire de T_1 et de T_2 :

$V(T_3) = \dfrac{1}{4}[V(T_1) + V(T_2) + 2\rho\sigma_1\sigma_2]$.

Or $V(T_1) = V(T_2) = \sigma_1 \times \sigma_2 = V$

$$V(T_3) = \frac{V}{2}(1+\rho)$$

Si $\rho < 1$, $V(T_3) < V$ ce qui est contredit l'hypothèse donc :

ρ = 1 c'est à dire $T_1 - E(T_1) = a[T_2 - E(T_2)]$

Or $V(T_1) = V(T_2)$ donc a = 1.

Les espérances étant égales, on a : $T_1 = T_2$ presque sûrement.

VI-3-2 Théorème 2 (Rao-Blackwell)

Soit T un estimateur quelconque sans biais de θ et U une statistique exhaustive pour θ. Alors $T^* = E(T/U)$ est un estimateur sans biais de θ au moins aussi bon que T.

Puisque U est exhaustive, la densité conditionnelle de l'échantillon sachant U ne dépend pas de θ et E (T /U) ne dépend pas de θ. Appliquons les théorèmes de l'espérance totale et de la variance totale :

$E(T^*) = E[E(T/U)] = E(T) = \theta$. On en déduit que T* est sans biais.

$V(T) = V[E(T/U)] + E[V(T/U)] = V(T^*) + E[V(T/U)]$ $\boldsymbol{E(T)=\theta}$ $\boldsymbol{E(T)=\theta}$

$$V(T^*) \leq V(T)$$

de plus si $E[V(T/U)] = 0$ alors $T = f(U)$ presque sûrement.

VI-3-3 Théorème 3

S'il existe une statistique exhaustive U, alors l'estimateur sans biais T de θ de variance minimale ne dépend que de U.

Ce théorème est un corollaire du théorème 2 :

Cet estimateur T, s'il existe, est unique presque sûrement (Théorème 1) ; on ne peut pas l'améliorer (théorème 2) ; de plus si sa variance est égale à celle de T* alors $T = f(U)$ presque sûrement.

Comme il peut exister plusieurs estimateurs sans biais de θ, fonction de U, on n'est pas sûr, par cette méthode, d'obtenir le meilleur.

VI-3-4 Définition d'une statistique complète

On dit qu'une statistique U est complète pour une famille de lois de probabilités dépendant d'un paramètre θ, de densités $f(x,\theta)$, si :

$$\boldsymbol{E[h(U)] = 0 \quad \forall\theta \quad \Rightarrow \quad h = 0} \textbf{ presque sûrement.}$$

On admettra, dans ce cours, le résultat très utile suivant :

Les statistiques exhaustives des familles exponentielles sont complètes.

VI-3-5 Théorème 4 (Lehmann-Scheffé)

Si T^* est un estimateur sans biais de θ, dépendant d'une statistique exhaustive complète U, alors T^* est l'unique estimateur sans biais de variance minimale de θ.

En particulier, si on dispose déjà de T estimateur sans biais de θ, alors :
$$\boldsymbol{T^* = E(T/U)}\textbf{.}$$

T* est un estimateur sans biais de variance minimale donc il est unique presque sûrement.

Il dépend de U donc il est de la forme $T = f(U)$ presque sûrement.

Enfin s'il existait deux estimateurs T_1 et T_2 sans biais fonction de U : $T_1 = f_1(U)$ et $T_2 = f_2(U)$

alors $E[f_1(U)] = E[f_2(U)] = \theta$ puisque les estimateurs sont sans biais.

U étant complète, on en déduit que $f_1 = f_2$ presque sûrement.

VI-3-6 Conclusion

Si on dispose d'un estimateur sans biais fonction d'une statistique exhaustive complète, c'est le meilleur estimateur possible.

Ce résultat est la conclusion des théorèmes précédents.

VI-3-7 Inégalité de Frechet-Darmois-Cramer-Rao

Soit X une variable aléatoire dont la densité dépend d'un paramètre θ et soit $X_1, X_2, \ldots, X_n$ un échantillon aléatoire.

Soit $I_n(\theta)$ la quantité d'information de Fisher.

Si T est un estimateur sans biais de θ et si le domaine de définition de X ne dépend pas de θ alors :

$$V(T) \geq \frac{1}{I_n(\theta)}$$

et si T est un estimateur sans biais de h(θ) :

$$V(T) \geq \frac{[h'(\theta)]^2}{I_n(\theta)}$$

Calculons la covariance de T et de $\frac{\partial LnL}{\partial \theta}$.

On a vu que $\frac{\partial LnL}{\partial \theta}$ est une variable centrée. La covariance recherchée est donc égale à l'espérance du produit :

$$\text{cov}(T, \frac{\partial LnL}{\partial \theta}) = \int_E t \frac{\partial LnL}{\partial \theta} L dx = \int_E t \frac{\partial L}{\partial \theta} dx = \frac{d}{d\theta} \int_E tL dx = \frac{d}{d\theta} E(T) = h'(\theta)$$

L'inégalité de Cauchy-Schwarz nous permet d'écrire :

$$\left[\text{cov}(T, \frac{\partial LnL}{\partial \theta})\right]^2 \leq V(T) V(\frac{\partial LnL}{\partial \theta})$$

Ce qui peut s'écrire aussi :

$$[h'(\theta)]^2 \leq V(T)\, I_n(\theta)$$

VI-3-8 Efficacité

VI-3-8-1. définition

Un estimateur est dit efficace lorsque sa variance vérifie l'égalité :

$$V(T) = \frac{1}{I_n(\theta)}$$

VI-3-8-2. Théorème de l'efficacité (admis)

La borne de FDCR ne peut être atteinte que si la loi de X est de la forme exponentielle :

$$Ln\, f(x,\theta) = a(x)\alpha(\theta) + b(x) + \beta(\theta)$$

car T est nécessairement exhaustif pour θ.

Si la loi est de la forme précédente, il n'existe qu'une seule fonction h (θ) du paramètre qui puisse être estimée efficacement, c'est :

$$h(\theta) = -\frac{\beta'(\theta)}{\alpha'(\theta)}$$

L'estimateur est alors : $T = \frac{1}{n}\sum_{i=1}^{n} a(X_i)$ **de variance minimale :** $V(T) = \frac{h'(\theta)}{n\alpha'(\theta)}$.

VI-3-8-3. Exemple

✓ Dans une loi de Laplace-Gauss LG (m, σ), si m est connue :

σ^2 est le seul paramètre que l'on peut estimer efficacement, par : $T = \frac{1}{n}\sum_{i=1}^{n}(X_i - m)^2$.

L'estimateur : $U = \sqrt{\frac{n}{2}}\frac{\Gamma(\frac{n}{2})}{\Gamma(\frac{n+1}{2})}\sqrt{T}$ est sans biais pour σ, de variance minimale car T est exhaustive, mais n'est pas efficace au sens de la borne de FDCR.

✓ Si m est inconnue, et si S^2 est l'estimateur usuel de σ^2, $V = \sqrt{\frac{n}{2}}\frac{\Gamma(\frac{n-1}{2})}{\Gamma(\frac{n}{2})}S$ est un estimateur sans biais, de variance minimale pour σ.

En pratique on utilise $S^* = \sqrt{\frac{n}{n-1}}S$ qui est très légèrement biaisé.

Si X ne suit pas une loi normale, on ne peut pas donner d'expression universelle d'un estimateur sans biais de σ.

VI- 4) METHODE DU MAXIMUM DE VRAISEMBLANCE

Etant donné un échantillon de valeurs (x_1, x_2, …, x_n), cette méthode consiste à prendre comme estimateur de θ la valeur de ce paramètre qui rend maximale la vraisemblance :

$$L(x_1, \ldots, x_n ; \theta)$$

VI-4-1 Définition

L'estimateur du maximum de vraisemblance est la solution de l'équation :

$$\frac{\delta}{\delta\theta} \ln L(\underline{X};\theta) = 0$$

VI-4-2 Propriété 1

Si T est une statistique exhaustive, l'estimateur $\hat{\theta}$ du maximum de vraisemblance en dépend.

Si T est une statistique exhaustive, $L(\underline{X},\theta) = g(t,\theta)h(x)$ et l'équation de la vraisemblance s'écrit alors :

$$\frac{\delta \ln g}{\delta\theta} = 0 \qquad \text{c'est à dire } \hat{\theta} = f(t).$$

Si de plus l'estimateur ainsi trouvé est sans biais, ce sera la meilleure estimation de θ si les conditions des théorèmes précédents sont réalisées.

VI-4-3 Propriété 2

Si $\hat{\theta}$ est l'estimateur du Maximum de Vraisemblance de θ, alors $f(\hat{\theta})$ est l'estimateur du Maximum de Vraisemblance de $f(\theta)$.

Exemple

Soit X une variable aléatoire suivant une loi normale LG(m, σ) où m est inconnu et σ connu. La vraisemblance s'écrit :

$$L(\underline{x},\sigma) = \frac{1}{\sigma^n(\sqrt{2\pi})^n} \exp(-\frac{1}{2}\sum_{i=1}^{n}(\frac{x_i - m}{\sigma})^2)$$

On a vu que : $\frac{\partial}{\partial m} LnL = \frac{n}{\sigma^2}(\bar{x} - m)$ et $\frac{\partial^2 Ln[L(\underline{x},m)]}{\partial m^2} = -\frac{n}{\sigma^2} < 0$

$\bar{x}$ réalise un maximum strict de L. Il est l'estimateur de m obtenu par la méthode du maximum de vraisemblance.

Ch. VII Estimation par intervalles

VII- 1) DEFINITION

On a vu, dans le chapitre précédent, que l'on cherchait souvent à estimer un paramètre θ. Autant il peut être utile (et possible !) d'estimer ponctuellement un paramètre, autant on cherche souvent un intervalle [a, b] qui a une probabilité donnée de contenir θ.

L'estimation par intervalle de confiance d'un paramètre θ consiste à associer à un n-échantillon, un intervalle aléatoire I choisi de telle sorte que la probabilité que cet intervalle contienne θ soit égale à une valeur convenue d'avance, aussi grande que l'on veut, notée $1-\alpha$:

$$P(\theta \in I) = 1-\alpha$$

1-α s'appelle le seuil ou niveau de confiance, α s'appelle le risque.

Le niveau de confiance représente donc la probabilité donnée que l'intervalle contienne θ. Il faut se garder de dire que 1-α représente la probabilité que le paramètre appartienne à l'intervalle construit car ce paramètre a une valeur, inconnue certes, mais fixée.

VII- 2) CONSTRUCTION D'UN INTERVALLE DE CONFIANCE

VII-2-1 Principe de construction

Soit X une variable aléatoire dont la loi de probabilité dépend d'un paramètre θ et soit un échantillon $(X_1, X_2, \ldots, X_n)$ de X.

Supposons que l'on dispose d'un estimateur T de θ, fonction de l'échantillon.

Sa loi de probabilité dépend, bien sûr, de θ. Grâce à cette loi, on peut déterminer, à α fixé, des valeurs t_1 et t_2 telles que :

$$P(t_1 \le \theta - T \le t_2) = 1-\alpha$$

On a alors :

$$P(T + t_1 \le \theta \le T + t_2) = 1-\alpha$$

La probabilité que l'intervalle $I = [T + t_1 , T + t_2]$ contienne le paramètre θ est égale à 1-α.

L'intervalle I s'appelle intervalle de confiance au seuil de probabilité (1-α).

Les bornes de cet intervalle dépendent de la valeur de l'estimateur calculée à partir de l'échantillon, c'est donc un intervalle aléatoire.

VII-2-2 Seuil d'un intervalle de confiance

Le seuil 1-α étant fixé, supposons que l'on ait pu construire, à l'aide d'un échantillon de taille n, un intervalle de confiance pour un paramètre θ : $P(\theta \in I) = 1-\alpha$.

α représente le risque que l'intervalle de confiance ne contienne pas la vraie valeur du paramètre.

1-α représente le niveau de confiance de l'intervalle. Ce niveau est associé à l'intervalle, à l'estimateur choisi et non au paramètre estimé. Il est clair qu'il est souhaitable de choisir le « meilleur estimateur » possible du paramètre θ.

VII-2-3 Propriétés d'un intervalle de confiance

VII-2-3-1. Intervalle de confiance symétrique

Quand on construit un intervalle de confiance pour un paramètre θ donné, les valeurs t_1 et t_2 ne sont pas uniques. Elles doivent vérifier $P(t_1 \le \theta - T \le t_2) = 1-\alpha$ donc :

$$P(\theta - T < t_1) = \alpha_1 \quad \text{et} \quad P(\theta - T > t_2) = \alpha_2 \qquad \text{où } \alpha_1 + \alpha_2 = \alpha$$

Il y a donc une infinité de couples (α_1, α_2) donc une infinité de couples (t_1, t_2) satisfaisant aux relations ci-dessus. Toutefois un cas particulier retient notre attention :

L'intervalle de confiance est dit à risques symétriques si on choisit :

$$\alpha_1 = \alpha_2 = \frac{\alpha}{2}.$$

Fréquemment on construit des intervalles de confiance à risques symétriques :

Le seuil de confiance 1-α étant choisi, on détermine les bornes de l'intervalle $I = [T + t_1, T + t_2]$ de telle sorte que $P(\theta < T + t_1) = P(\theta > T + t_2) = \frac{\alpha}{2}$

VII-2-3-2. Intervalle de confiance unilatéral

A contrario, on définit les intervalles de confiance à droite ou à gauche en choisissant de poser :

$$\alpha_1 = 0 \text{ et } \alpha_2 = \alpha \qquad \text{ou l'inverse}$$

VII-2-3-3. Largeur de l'intervalle

On peut remarquer que :

- La largeur de l'intervalle de confiance augmente si on diminue la valeur du risque α.
- La largeur de l'intervalle de confiance diminue quand n augmente.

Dans la plupart des cas, on fixe la valeur du risque α à 5% (plus rarement 10% ou 1%). Pour diminuer la largeur de l'intervalle de confiance, on augmente, quand c'est possible, la taille n de l'échantillon.

VII-2-4 Exemple

Un laboratoire est chargé de préciser la résistance à l'éclatement, en kg/cm^2, des réservoirs à essence des camions-citernes d'un constructeur donné.

On considère que la résistance à l'éclatement de ces citernes est une variable normale de moyenne inconnue et d'écart-type égal à 13. Des essais sur 16 réservoirs conduisent à une résistance moyenne à l'éclatement de 1215 kg/cm^2.

On va construire un intervalle de confiance à risques symétriques ayant une probabilité égale à 0,998 de contenir la résistance moyenne à l'éclatement des citernes produites par ce constructeur.

Appelons X la variable : « résistance à l'éclatement de ce type de réservoir », on sait que X suit une loi normale LG (m, 13) où m est le paramètre qu'on cherche à encadrer.

De plus, sur l'échantillon de 16 citernes, on a mesuré la moyenne :

$$\bar{x} = 1215\, kg / cm^2$$

On sait que la moyenne $\overline{X}$ suit la loi normale $LG(m, \frac{13}{\sqrt{16}}) = LG(m, \frac{13}{4})$.

La variable $U = \frac{\overline{X} - m}{13/4}$ suit alors la loi normale centrée réduite. On en déduit :

$$P(|U| < 3{,}09) = 0{,}998$$

$$P(1215 - 3{,}09 \times \frac{13}{4} \le m \le 1215 + 3{,}09 \times \frac{13}{4}) = P(1205 \le m \le 1225) = 0{,}998$$

L'intervalle de confiance cherché est donc : [1205,1225].

VII- 3) MOYENNE D'UNE LOI NORMALE

On cherche un intervalle de confiance pour la moyenne m d'une loi normale LG(m, σ). Cet intervalle est construit différemment suivant si l'écart-type est connu ou estimé. Les bornes de l'intervalle recherché sont déterminées à l'aide de la loi suivie par l'estimateur choisi.

On choisira de construire, dans tous les cas étudiés ici, des intervalles à risques symétriques.

VII-3-1 L'écart-type σ est connu

$\overline{X}$ est le meilleur estimateur de m et on sait que $\boldsymbol{L(\overline{X}) = LG(m, \frac{\sigma}{\sqrt{n}})}$

En utilisant les tables de la loi normale, α et n étant fixés, on trouve a tel que :

$$\boldsymbol{P(-a \le \frac{\overline{X} - m}{\sigma/\sqrt{n}} \le a) = 1 - \alpha}$$

Si $\bar{x}$ est la moyenne d'un échantillon de taille n, l'intervalle de confiance pour m s'écrit :

$$\boldsymbol{\bar{x} - a\frac{\sigma}{\sqrt{n}} \le m \le \bar{x} + a\frac{\sigma}{\sqrt{n}}}$$

VII-3-2 L'écart-type σ est inconnu

$\overline{X}$ est toujours l'estimateur de m mais, σ étant inconnu, la variable :

$$\frac{\overline{X}-m}{S/\sqrt{n-1}}$$ suit une loi de Student à (n-1) degrés de liberté.

La loi de Student étant tabulée, on connait, à α et n fixés, la valeur $t_{\alpha/2}$, telle que :

$$P(-t_{\alpha/2} \le \frac{\overline{X}-m}{S/\sqrt{n-1}} \le t_{\alpha/2}) = 1-\alpha$$

A l'aide d'un échantillon de taille n pour lequel $\bar{x}$ et s sont la moyenne et l'écart-type relevés, l'intervalle de confiance de m s'écrit alors :

$$\boxed{\bar{x} - t_{\alpha/2}\frac{s}{\sqrt{n-1}} \le m \le \bar{x} + t_{\alpha/2}\frac{s}{\sqrt{n-1}}}$$

Si l'estimateur de σ est s*, l'intervalle est alors :

$$\boxed{\bar{x} - t_{\alpha/2}\frac{s^*}{\sqrt{n}} \le m \le \bar{x} + t_{\alpha/2}\frac{s^*}{\sqrt{n}}}$$

puisque nous disposons de la relation $\frac{nS^2}{\sigma^2} = \frac{(n-1)S^{*2}}{\sigma^2}$ entre s et s*.

VII- 4) MOYENNE D'UNE LOI QUELCONQUE

Si la taille de l'échantillon est suffisamment grande, on prend pour estimateur de m la statistique $\overline{X}$ et grâce au théorème de la limite centrale on construit les mêmes intervalles de confiance pour m que ceux construits avec la loi normale. On distingue, là encore, le cas où σ est connu du cas où il est inconnu.

Reprenons l'exemple de VI-2-4 sur la résistance à l'éclatement des réservoirs à essence.

On suppose dorénavant que la loi suivie par cette variable est inconnue, de moyenne m à estimer et d'écart-type σ inconnu. On fait 50 essais sur lesquels on mesure une résistance moyenne à l'éclatement de 1215 kg/cm^2 et un écart-type s = 13 kg/cm^2. Construisons un intervalle de confiance pour m au seuil 95%.

N = 50 est suffisamment grand pour appliquer le théorème de la limite centrale et on peut donc écrire que la variable $U=\frac{\overline{X}-m}{13/7}$ suit une loi de Student à 49 degrés de liberté.

L'intervalle de confiance pour m, au risque 5%, est donné par :

$$1215 - 1{,}99 \times \frac{13}{7} < m < 1215 + 1{,}99 \times \frac{13}{7}$$

$$1211{,}3 < m < 1218{,}7$$

VII- 5) VARIANCE σ^2 D'UNE LOI NORMALE

Là encore deux situations différentes se présentent, suivant si la moyenne de la loi est connue ou inconnue.

VII-5-1 La moyenne m est connue

Le meilleur estimateur de la variance σ^2 est la statistique :

$$T = \frac{1}{n}\sum_{i=1}^{n}(X_i - m)^2 .$$

On sait de plus, que la variable $\frac{nT}{\sigma^2}$ suit une loi du Chi-deux à n degrés de liberté.

Après lecture de la table du chi-deux qui nous donne, à α et n fixés, les bornes a et b (non symétriques) de l'intervalle de probabilité :

$$P(a \le \frac{nT}{\sigma^2} \le b) = 1 - \alpha$$

On en déduit l'intervalle pour σ^2 :

$$\boxed{\frac{nT}{b} \le \sigma^2 \le \frac{nT}{a}}$$

VII-5-2 La moyenne m n'est pas connue

La statistique $\frac{nS^2}{\sigma^2}$ suit une loi du Chi-deux à n-1 degrés de liberté.

On détermine comme précédemment, grâce aux tables de la loi du Chi-deux, les valeurs a et b telles que :

$$P(a < \frac{nS^2}{\sigma^2} < b) = 1 - \alpha$$

L'intervalle de confiance est alors :

$$\boxed{\frac{nS^2}{b} \le \sigma^2 \le \frac{nS^2}{a}}$$

On en déduit un intervalle de confiance pour σ avec le même risque :

$$\sqrt{\frac{nS^2}{b}} \le \sigma \le \sqrt{\frac{nS^2}{a}}$$

On peut, là encore, construire ces intervalles avec s* puisque : $\frac{nS^2}{\sigma^2} = \frac{(n-1)S^{*2}}{\sigma^2}$:

$$\sqrt{\frac{(n-1)S^{*2}}{b}} \le \sigma \le \sqrt{\frac{(n-1)S^{*2}}{a}}$$

Attention :

Quelque soit la taille de l'échantillon, l'intervalle de confiance construit pour la variance d'une loi normale ne peut pas être considéré comme un intervalle de confiance pour la variance d'une loi quelconque.

VII- 6) DIFFERENCE DES MOYENNES DE LOIS NORMALES

Soient X et Y deux variables suivant des lois LG (m_1, σ) et LG (m_2, σ), c'est à dire des lois normales de **même écart-type**.

On cherche à estimer $m = m_1 - m_2$

On prend comme estimateur de m : $\overline{X_1} - \overline{X_2} = \overline{D}$

Cette variable $\overline{D}$ suit une loi normale, de moyenne m et d'écart-type : $\sigma' = \sqrt{\frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}}$

La statistique : $\frac{\sum_{k=1}^{n_1}(X_k - \overline{X_1})^2 + \sum_{h=1}^{n_2}(X_h - \overline{X_2})^2}{n_1 + n_2 - 2}$ est un estimateur sans biais de la variance commune des deux distributions.

Finalement : la variable $\frac{(\overline{X_1} - \overline{X_2}) - (m_1 - m_2)}{\sqrt{\sum_i (X_i - \overline{X_1})^2 + \sum_i X_i - \overline{X_2})^2}} \times \frac{\sqrt{n_1 + n_2 - 2}}{\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$

suit une loi de Student à $n_1 + n_2 - 2$ degrés de liberté ; à α fixé, la lecture de la table de Student permet de déterminer une valeur a telle que la variable précédente appartienne à l'intervalle [-a, a] au seuil 1-α.

On en déduit l'intervalle de confiance à risques symétriques pour $m = m_1 - m_2$:

$$\left| m - (\overline{x_1} - \overline{x_2}) \right| < a \times \sqrt{\sum_i (x_i - \overline{x_1})^2 + \sum_i (x_i - \overline{x_2})^2} \times \frac{\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}{\sqrt{n_1 + n_2 - 2}}$$

VII- 7) RAPPORT DES VARIANCES DE LOIS NORMALES

Soient X et Y deux variables suivant les lois respectives : LG (m_1, σ_1) et LG (m_2, σ_2).

Nous savons que $\frac{S_1^{*2}}{\sigma_1^2} \times \frac{\sigma_2^2}{S_2^{*2}}$ suit une loi de Fisher $F(n_1 - 1, n_2 - 1)$.

On peut alors déterminer un intervalle de confiance au seuil 1 - α en lisant sur les tables de la loi de Fisher les valeurs des bornes de l'intervalle :

$$P(a \le \frac{S_1^{*2}}{\sigma_1^2} \times \frac{\sigma_2^2}{S_2^{*2}} \le b) = 1 - \alpha$$

On en déduit l'intervalle de confiance au seuil 1 - α pour le rapport des variances :

$$\boxed{\frac{s_2^{*2}}{s_1^{*2}} a \le \frac{\sigma_2^2}{\sigma_1^2} \le \frac{s_2^{*2}}{s_1^{*2}} b}$$

Remarque :

Soit Z une variable de Fisher de paramètres n et p.

Les tables de la loi de Fisher disponibles permettent de connaître les seuils f_1 et f_2 tels que :

$$P(Z < f_1) = 0{,}95 \text{ ou } P(Z < f_2) = 0{,}99$$

Quand on cherche à encadrer Z, on peut remarquer que, par définition de cette loi, 1/Z suit une loi de Fisher de paramètres p et n.

On en déduit :

$$P(a < Z) = 0{,}95 = P(\frac{1}{Z} < \frac{1}{a})$$

La valeur a est donc obtenue en lisant son inverse dans la table de la loi de Fisher de paramètres p et n.

VII- 8) INTERVALLE DE CONFIANCE POUR UNE PROPORTION

VII-8-1 Construction

Etant donné une population d'effectif très grand où une proportion p d'individus possèdent un certain caractère, il s'agit de trouver un intervalle de confiance pour p à partir de l'estimation de p notée *f* et calculée grâce à un échantillon de taille n.

On sait que n*f* suit une loi binomiale B (n, p) et que, pour n assez grand, on peut utiliser l'approximation de la loi binomiale par la loi normale :

$$L(nf) \cong LG\,[\,np, \sqrt{np(1-p)}\,]$$

On a déjà vu que cette approximation peut s'écrire : $\boldsymbol{L(f) \cong LG(p, \sqrt{\frac{p(1-p)}{n}})}$

La table de la loi normale nous permet d'obtenir la valeur a, à α fixé, telle que :

$$P(\frac{|p-f|}{\sqrt{\frac{p(1-p)}{n}}} \le a) = 1 - \alpha$$

L'intervalle de confiance à risques symétriques au seuil 1-α est donc donné par l'équation :

$$\boxed{(p-f)^2 \le a^2 \frac{p(1-p)}{n}}$$

On peut remarquer que le paramètre p dont on cherche un intervalle pouvant le contenir avec un seuil de probabilité 1-α, figure dans la partie droite de l'inégalité. Trois méthodes se présentent alors pour déterminer l'intervalle de confiance :

- ✓ On remplace p par la valeur 1/2 dans le second membre de l'inéquation, cette valeur maximisant le produit p(1-p).
- ✓ On remplace p dans le second membre par son estimateur *f*.
- ✓ on résout l'inéquation du second degré où p est l'inconnue (méthode de l'ellipse).

Il est clair que la troisième méthode est la plus employée actuellement et la plus mathématique mais les deux premières donnent, en général, des intervalles très proches de celui obtenu par la méthode de l'ellipse.

VII-8-2 Exemple

Dans un échantillon de 100 automobilistes pris au hasard, on constate que 20 d'entre eux ont un véhicule plus polluant que les normes admises. Donner un intervalle de confiance pour la proportion d'automobilistes ayant un véhicule dépassant les normes de pollution.

Une estimation de cette proportion est $f = \frac{20}{100} = 0{,}2$.

Le nombre d'automobilistes est très grand donc l'approximation par la loi normale est pleinement justifiée. On peut donc écrire que p vérifie :

$$\boxed{(p-f)^2 \leq a^2 \frac{p(1-p)}{n}}$$

où a est le seuil donné par la table de la loi normale. Ainsi, si on choisit un risque $\alpha = 0{,}05$ alors $a = 1{,}96$.

On obtient l'inéquation :

$$|p-0{,}2| < 1{,}96 \frac{\sqrt{p(1-p)}}{10}$$

- ✓ on remplace p par f = 0,2 dans la racine et on obtient l'intervalle : 0,122 < p < 0,278
- ✓ on remplace p par 0,5 dans la racine et on obtient : 0,102 < p < 0,298
- ✓ on résout l'équation de second degré : $(p-0{,}2)^2 < 0{,}196^2\, p(1-p)$. L'intervalle obtenu n'est plus symétrique autour de 0,2 et il vaut : 0,1333 < p < 0,2888

On peut remarquer que le dernier intervalle ne nécessite pas d'approximation de p dans l'inégalité et que le premier intervalle obtenu a une étendue pratiquement semblable au troisième.

VII-8-3 Application

Détermination de la taille d'un échantillon en fonction de la précision souhaitée :

Supposons que l'on désire connaître p avec une précision donnée Δp pour un niveau de confiance donné 1 - α à risques symétriques.

D'après les résultats précédents, la précision obtenue est la différence entre f et p en valeur absolue c'est à dire :

$$\Delta p = |p - f| \leq u_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}$$

Le second membre est majoré par l'expression calculée pour la valeur ½ du paramètre p, cette valeur maximisant p (1 – p).

Exemple :

On cherche à estimer p avec une précision de 1/100. On choisit un risque α égal à 0,05, la valeur de la table de la loi normale nous donne u = 1,96 et :

$\sqrt{\frac{1}{4n}} \times 1{,}96 \leq 0{,}01$, on obtient $\sqrt{4n} \geq 196$ soit une valeur de la taille n : $n \geq 98^2$ = 9604.

Ch. VIII Tests d'hypothèses

Nous avons, dans les chapitres précédents, estimé ponctuellement ou par intervalle des paramètres d'une population. Nous allons maintenant aborder une nouvelle partie de la Statistique : celle des tests. Diverses catégories de tests apparaissent :

- ✓ tests sur une hypothèse posée sur la valeur d'un paramètre,
- ✓ tests d'ajustement d'une distribution théorique connue sur une distribution empirique,
- ✓ tests de comparaison de deux populations,
- ✓ tests d'indépendance de deux caractères…

Dans cette première partie, nous allons nous intéresser aux tests d'hypothèses et nous introduisons ce chapitre par un exemple très instructif ; il est emprunté à G. Saporta qui le présentait dans son cours à l'Ecole, il y a une quinzaine d'années.

VIII- 1) EXEMPLE INTRODUCTIF

Des relevés faits durant de nombreuses années en Beauce ont permis de mettre en évidence que le niveau naturel des pluies, en mm par an, suit une loi normale LG (600, 100). Dans les années cinquante, des ***faiseurs de pluie*** prétendirent pouvoir augmenter de 50 mm le niveau moyen de pluie par insémination des nuages au moyen de chlorure d'argent. Durant les années 1951 à 1959, on mit le procédé à l'essai et on releva les hauteurs d'eau suivantes :

Année	1951	1952	1953	1954	1955	1956	1957	1958	1959
Mm	510	614	780	512	501	534	603	788	650

La question était de savoir si l'insémination avait un effet ou non ; donc deux hypothèses s'affrontaient : soit le procédé avait un effet certain sur le niveau moyen de pluie soit il n'en avait pas.

On peut formaliser le problème ainsi :

soit X la variable aléatoire égale au niveau annuel de pluie et soit m son espérance ; les deux hypothèses en présence se résument par :

H_0 : m = 600 mm

H_1 : m = 650 mm

Les agriculteurs trouvaient le procédé onéreux, donc se tenaient à l'hypothèse H_0 (hypothèse conservatoire) et voulaient être convaincus par des faits expérimentaux, avant de changer d'avis et adopter l'hypothèse H_1 (hypothèse alternative).

Ils décidèrent qu'ils adopteraient le procédé si le résultat obtenu par les mesures faisait partie d'une éventualité qui n'avait que 5% de chance de se produire. Ils assumaient alors le risque de 5% de se tromper.

On désire tester la moyenne m donc on s'intéresse à la moyenne $\overline{X}$ des relevés.

Sous l'hypothèse H_0 que les agriculteurs adoptent au départ, la moyenne de l'échantillon suit une loi normale :

$$L(\overline{X})=LG(600,\frac{100}{\sqrt{9}})=LG(600,\frac{100}{3})\,.$$

Le choix des agriculteurs était donc le suivant :

Si $\overline{X}$ est trop grand, c'est-à-dire s'il est supérieur à un seuil qui n'a que 5% de chances d'être dépassé, on adoptera le procédé avec un risque de 5% de chances de se tromper.

Il reste à déterminer ce seuil et la loi normale nous permet de le faire :

$$P(\overline{X}>k)=0,05=P(\frac{\overline{X}-600}{100/3}>\frac{k-600}{100/3}) = \frac{3(k-600)}{100}=1,64$$

Le seuil cherché était donc :

$$k = (600 + 55)\text{ mm} = 655\text{ mm}.$$

On aboutit à la règle de décision suivante :

$\overline{X} > 655$ mm on adopte H_1, donc le procédé

$\overline{X} < 655$ mm on conserve H_0

Le domaine $\{\overline{X} > 655\}$ s'appelle la **région critique** ou région de rejet de H_0

L'ensemble complémentaire $\{\overline{X} < 655\}$ s'appelle **la région d'acceptation** de H_0 ou plus précisément région de non rejet de H_0.

Les données relevées indiquaient $\overline{X}$ = 610,2 mm. La conclusion était donc de conserver H_0 et de rejeter le procédé.

Toutefois rien ne dit que conserver H_0 n'était pas une erreur. Les faiseurs de pluie avaient peut-être raison mais on ne l'a pas vu. S' ils avaient raison, la loi de $\overline{X}$ était :

$$L(\overline{X})=LG(650,\frac{100}{3})$$

On commettait une erreur si $\overline{X}$ était inférieure à 655 mm car, alors on choisissait de conserver H_0 alors que les faiseurs avaient raison.

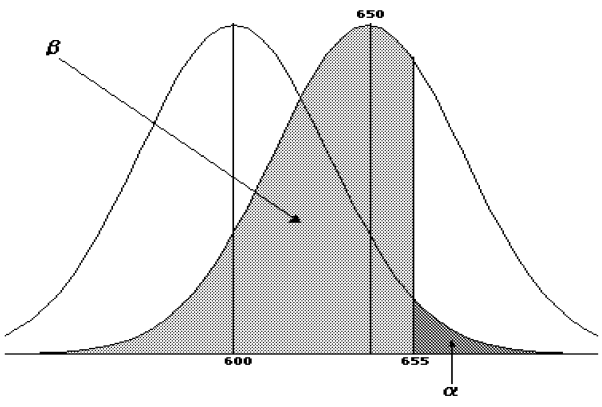
$$\beta=P(\frac{\overline{X}-650}{100/3}<\frac{655-650}{100/3}=0,15)$$

$\beta = 0,56$.

La probabilité de commettre une erreur était grande…

α s'appelle le **risque de première espèce,**

β s'appelle le **risque de deuxième espèce.**



VIII- 2) NOTIONS GENERALES SUR LES TESTS

VIII-2-1 Définition

Un test est un mécanisme qui permet de choisir entre deux hypothèses à l'aide des résultats d'un échantillon.

On est donc en présence de deux hypothèses dont une seule est vraie ; on souhaite pouvoir décider entre les deux en choisissant de prendre un risque α de se tromper ; résumons le problème dans le tableau suivant :

Vérité / Décision	H_0	H_1
H_0	$1-\alpha$	β
H_1	α	$1-\beta$

- ✓ α est la probabilité de choisir H_1 alors que H_0 est vraie ; il est appelé **risque de première espèce.**
- ✓ β est la probabilité de conserver H_0 alors que H_1 est vraie ; il est appelé **risque de deuxième espèce.**

Dans l'exemple précédent, α est le risque d'acheter le procédé alors qu'il ne vaut rien et β est le risque de laisser passer un procédé valable.

VIII-2-2 Choix accompagnant un test d'hypothèses

VIII-2-2-1. Choix du risque α

Au départ de tout test d'hypothèses, le risque α doit être choisi : il représente la probabilité de rejeter H_0 alors que H_0 est vraie donc il est choisi petit : 10%, plus usuellement 5% voire 1% mais le choisir trop petit présente un inconvénient :

En voulant diminuer le risque de première espèce α, on aboutit à une règle de décision beaucoup plus stricte qui aboutit à n'abandonner l'hypothèse H_0 que très rarement.

VIII-2-2-2. Choix de l'hypothèse H_0

Le choix de H_0 relève de diverses raisons :

- ✓ H_0 peut être une hypothèse solidement établie, non contredite jusqu'à présent,
- ✓ H_0 peut être une hypothèse à laquelle on tient particulièrement,
- ✓ Elle est souvent une hypothèse de prudence : quand il est nécessaire de partir d'une hypothèse défavorable à un changement : test d'innocuité d'un nouveau traitement…
- ✓ H_0 est la seule hypothèse facile à formuler…

Le risque de première espèce α étant choisi au départ, H_0 joue un rôle prépondérant.

VIII-2-2-3. Choix de la variable de décision

Quand H_0 et α sont déterminés, on doit choisir une variable de décision permettant la conduite du test. Cette variable est déterminée en fonction du problème et sa loi doit être connue sous H_0 et doit être différente sous H_1.

Dans l'exemple des faiseurs de pluie, on s'intéressait aux relevés annuels des niveaux de pluie; il s'imposait donc comme variable de décision de prendre le relevé moyen mesuré $\overline{X}$. On en connaît la loi et cette loi était différente suivant H_0 ou H_1.

En dehors des variables très usuelles comme $\overline{X}$ ou S^2, d'autres seront choisies et une méthode sera proposée pour déterminer, dans certains cas, la meilleure variable de décision.

VIII-2-3 Conduite du test et conclusion

VIII-2-3-1. Région critique

On appelle région critique l'ensemble des valeurs W de la variable de décision conduisant à écarter H_0 au profit de H_1.

Elle est déterminée par la connaissance de la variable de décision et le choix de la valeur du risque de premier espèce :

$$P(W / H_0) = \alpha$$

La région d'acceptation de H_0 est définie par son complémentaire $\overline{W}$ telle que :

$$P(\overline{W} / H_0) = 1 - \alpha$$

VIII-2-3-2. Calcul de la valeur expérimentale et conclusion

A partir de l'expérimentation accompagnant le test, on détermine la valeur de la variable de décision sur l'échantillon et on peut donc conclure sur l'appartenance ou non de la valeur trouvée à la région critique et donc sur l'hypothèse à retenir.

Dans l'exemple des faiseurs de pluie, la valeur trouvée était de 610,2 mm et les agriculteurs rejetèrent le procédé proposé.

VIII-2-3-3. Puissance du test

La puissance du test est définie par :

$$P(W / H_1) = 1 - \beta$$

Le risque de deuxième espèce β représente le risque de rejeter H_1 alors que H_1 est vraie. Quand on diminue le risque α, β augmente et la puissance du test, qui est la probabilité de choisir H_1 alors que H_1 est vraie, diminue.

Dans l'exemple des faiseurs de pluie, la puissance du test était de 44% ce qui n'est pas suffisant. Une puissance supérieure à 80% est souhaitable pour que le test d'hypothèses soit satisfaisant.

VIII-2-4 Démarche générale d'un test

Les étapes de la démarche d'un test d'hypothèses sont les suivantes :

- **Choix du risque α**
- **Choix des hypothèses H_0 et H_1**
- **Détermination de la variable de décision**
- **Détermination de la région critique W : $P(W / H_0) = \alpha$**
- **Calcul éventuel de la puissance du test : $P(W / H_1) = 1 - \beta$**
- **Calcul, sur l'échantillon, de la valeur expérimentale de la variable de décision**
- **Conclusion du test : rejet ou acceptation de H_0.**

VIII-2-5 Les catégories de tests paramétriques

Un test paramétrique a pour but de tester une hypothèse relative à un ou plusieurs paramètres d'une variable aléatoire de loi connue ou non.

VIII-2-5-1. Tests robustes, tests libres

Très fréquemment on considère que la variable de décision suit une loi normale.

Si les résultats sont encore valables lorsque la variable n'est plus supposée suivre une loi normale, on dit que le test est **robuste.**

Ainsi les tests de moyenne sont robustes.

Si les tests sont valables quelque soit la loi de la variable X, on dit que ce sont des **tests libres**.

VIII-2-5-2. Hypothèses simples

On appelle hypothèse simple une hypothèse du type :

H : $\theta = \theta_0$ où θ_0 est une valeur isolée du paramètre.

Un test paramétrique est dit à hypothèses simples s'il est composé de deux hypothèses H_0 et H_1 simples.

L'exemple vu au début du chapitre est à hypothèses simples.

VIII-2-5-3. Hypothèses composites

On appelle hypothèse composite une hypothèse du type :

H : $\theta \in A$ où A est une partie, non singleton, de R.

La plupart des hypothèses composites se ramènent aux cas :

$\theta > \theta_0$, ou $\theta < \theta_0$, ou $\theta \neq \theta_0$

On construit la région critique en utilisant la valeur θ_0 seule. Dans le cas d'une hypothèse alternative composite, la puissance du test est variable et on parle alors de fonction puissance :

$$P(\theta) = 1 - \beta(\theta).$$

VIII- 3) TEST ENTRE DEUX HYPOTHESES SIMPLES

VIII-3-1 Méthode de Neyman et Pearson

Le problème du choix de la meilleure variable de décision a été résolu par les statisticiens Neyman et Pearson dans les années 1930.

Qui dit meilleure variable de décision dit région critique optimale c'est-à-dire recherche du maximum de la puissance du test pour une valeur choisie du risque de premier espèce.

Cette méthode s'applique pour les tests entre deux hypothèses simples :

H_0 $\quad\quad$ $\theta = \theta_0$

H_1 $\quad\quad$ $\theta = \theta_1$

Soit X une variable aléatoire de densité f (x, θ) où θ est un paramètre réel inconnu et soit un échantillon $(X_1, X_2, \ldots, X_n)$ de cette variable.

Rappelons que si L ($\underline{X}$, θ) désigne la densité de l'échantillon :

$$\boldsymbol{L(\underline{x},\theta) = f(x_1,\ldots,x_n)} \text{ si la variable X est continue}$$

$$\boldsymbol{L(\underline{x},\theta) = P(X_1 = x_1,\ldots,X_n = x_n)} \text{ si X est discrète}$$

α désignant le risque de première espèce, la région critique est définie par :

$$\boldsymbol{P(W / H_0) = \alpha = \int_W L(\underline{x},\theta_0)\, d\underline{x}}$$

VIII-3-1-1. Théorème de Neyman et Pearson

La région critique optimale est définie par l'ensemble des points de R^n tels que :

$$\boldsymbol{\frac{L(\underline{x},\theta_1)}{L(\underline{x},\theta_0)} > k_\alpha}$$

Le risque de première espèce étant fixé, la méthode de Neyman et Pearson consiste à rendre maximum la puissance du test c'est-à-dire :

$$1 - \beta = P(W / H_1) = \int_W L(\underline{x},\theta_1)\, d\underline{x}$$

$$1 - \beta = P(W / H_1) = \int_W \frac{L(\underline{x},\theta_1)}{L(\underline{x},\theta_0)} L(\underline{x},\theta_0)\, d\underline{x}$$

- Montrons d'abord l'existence de cette région c'est-à-dire l'existence de k_α.

Pour toute constante k positive donnée, on définit la région A (k) par :

$$A(k) = \left\{ \underline{x} \in R^n / L(\underline{x},\theta_1 > kL(\underline{x},\theta_0 \right\}$$

$P(A(k) / H_0) = \int_{A(k)} L(\underline{x},\theta_0)\, d\underline{x}$ est une fonction de k, continue, monotone.

Si k = 0, $L(\underline{x},\theta_1)$ étant toujours positif (comme densité), $P(A(k) / H_0) = 1$

Si $k \to +\infty$, $L(\underline{x},\theta_1)$ étant bornée (comme densité), $P(A(k) / H_0) \to 0$.

Il existe donc une valeur intermédiaire k_α telle que $P(A(k_\alpha)/H_0) = \alpha$

- Montrons l'unicité de cette région ou simplement de k_α.

Soit W' une autre région de R^n telle que $P(W'/H_0) = \alpha$. W' diffère alors de W par des points où : $L(\underline{x},\theta_1) \le k_\alpha L(\underline{x},\theta_0)$.

Pour comparer les deux intégrales : $\int_W L(\underline{x},\theta_1)d\underline{x}$ et $\int_{W'} L(\underline{x},\theta_1)d\underline{x}$, il suffit de comparer les deux intégrales :

$$I = \int_A \frac{L(\underline{x},\theta_1)}{L(\underline{x},\theta_0)} L(\underline{x},\theta_0)d\underline{x} \text{ et } I' = \int_{A'} \frac{L(\underline{x},\theta_1)}{L(\underline{x},\theta_0)} L(\underline{x},\theta_0)d\underline{x}$$

où A est l'ensemble des points de W n'appartenant pas à W' et A' est celui des éléments de W' n'appartenant pas à W.

On applique le théorème de la moyenne aux deux intégrales :

$$I = \frac{L(\eta,\theta_1)}{L(\eta,\theta_0)} P(A/H_0) \text{ où } \eta \in A \quad \text{et } I' = \frac{L(\eta',\theta_1)}{L(\eta',\theta_0)} P(A'/H_0) \text{ où } \eta' \in A'$$

Or les deux probabilités $P(A/H_0)$ et $P(A'/H_0)$ sont égales car W et W' ont même mesure sous H_0.

η' appartenant à W' \ W et W' différant de W par des points où $L(\underline{x},\theta_1) \le k_\alpha L(\underline{x},\theta_0)$ alors :

$$\frac{L(\eta',\theta_1)}{L(\eta',\theta_0)} \le k_\alpha < \frac{L(\eta,\theta_1)}{L(\eta,\theta_0)}$$

Il en résulte que la région W réalise bien le maximum de $(1-\beta)$. Le théorème est démontré.

VIII-3-1-2. Etude de la puissance 1-β du test par cette méthode

La puissance 1-β du test vérifie : $1-\beta > \alpha$.

Le test est dit sans biais.

Par définition, $1-\beta = P(W/H_1) = \int_W L(\underline{x},\theta_1)d\underline{x} > k_\alpha \int_W L(\underline{x},\theta_0)d\underline{x} = k_\alpha \alpha$

Si $k_\alpha > 1$ alors le résultat est acquis : $1-\beta > \alpha$

Si $k_\alpha \le 1$, on écrit : $\beta = \int_{\overline{W}} L(\underline{x},\theta_1)d\underline{x} < k_\alpha \int_{\overline{W}} L(\underline{x},\theta_0)d\underline{x}$ car dans $\overline{W}$, on a $\frac{L(\underline{x},\theta_1)}{L(\underline{x},\theta_0)} \le k_\alpha$

Alors $\beta < \int_{\overline{W}} L(\underline{x},\theta_0)d\underline{x} = 1-\alpha$ ce qui achève la démonstration.

On admettra le résultat suivant :

Quand n tend vers l'infini, la puissance du test 1-β tend vers 1.

1-β est une fonction décroissante de k_α, donc :

si le risque de première espèce α décroît alors la puissance du test diminue.

VIII-3-1-3. Exemple

Test sur une moyenne d'une loi de Laplace-Gauss dont l'écart-type σ est supposé connu.

$$H_0 : \quad m = m_0$$

$$H_1 : \quad m = m_1$$

Ecrivons la densité $L(\underline{X},\theta) = \prod_i f(x_i,\theta) = \frac{1}{\sigma^n(\sqrt{2\pi})^n}\prod_i \exp[-\frac{1}{2}(\frac{x_i - m}{\sigma})^2]$

Pour appliquer la méthode de Neyman et Pearson, on étudie le rapport :

$$\frac{L(\underline{X},\theta_1)}{L(\underline{X},\theta_0)} = \exp[-\frac{1}{2\sigma^2}[\sum_i (x_i - m_1)^2 - \sum_i (x_i - m_0)^2]$$

$$\frac{L(\underline{X},\theta_1)}{L(\underline{X},\theta_0)} = \exp[-\frac{1}{2\sigma^2}[\sum_i x_i(m_0 - m_1) + m_0^2 + m_1^2] > k$$

$\frac{L(\underline{X},\theta_1)}{L(\underline{X},\theta_0)} > k$ est équivalent à $\sum_i x_i > K$ dès que $m_1 > m_0$

On en déduit que la variable de décision est $\overline{X}$ et la région critique définie par : $\overline{X} > k_0$

On rejettera H_0 si $\overline{X}$ est trop grand, la valeur k_0, limite de la région critique, est obtenue par :

$$\alpha = P(\overline{X} > k_0 / H_0)$$

Le risque de deuxième espèce est égal à : $\beta = P(\overline{X} < k / H_1)$

VIII-3-2 Cas d'une statistique exhaustive

VIII-3-2-1. Méthode de Neyman et Pearson

S'il existe une statistique exhaustive T pour θ, de densité g(t, θ), alors la densité de l'échantillon s'écrit $L(\underline{x},\theta) = g(t,\theta)h(\underline{x})$ et la région critique est alors définie par :

$$\frac{g(t,\theta_1)}{g(t,\theta_0)} > k_\alpha$$

VIII-3-2-2. Exemple

Reprenons l'exemple précédent du test d'une moyenne de Laplace-Gauss. On sait que $\overline{X}$ est une statistique exhaustive de m et :

$$g(\bar{x}, m) = \frac{1}{\sigma\sqrt{\frac{2\pi}{n}}}\exp(-\frac{1}{2}(\frac{\bar{x} - m}{\sigma/\sqrt{n}})^2)$$

L'inégalité : $\frac{g(\bar{x}, m_1)}{g(\bar{x}, m_0)} > k_\alpha$ revient à écrire : $(\bar{x} - m_0)^2 - (\bar{x} - m_1)^2 > k_\alpha$

En développant on obtient la région critique : $\overline{X} > k_0$ si $m_1 > m_0$ ou $\overline{X} < k_0$ si $m_1 < m_0$

VIII- 4) TESTS ENTRE HYPOTHESES COMPOSITES

VIII-4-1 Test d'une hypothèse simple contre une hypothèse composite

Différents cas sont possibles tels :

$$H_0 : \theta = \theta_0 \qquad H_1 : \theta \neq \theta_0$$

ou

$$H_0 : \theta = \theta_0 \qquad H_1 : \theta > \theta_0$$

L'hypothèse H_1 étant composée de différentes valeurs, la fonction puissance décrit les variations de $1-\beta$ en fonction des valeurs θ possibles de H_1.

Un test est dit uniformément le plus puissant (UPP) si quelque soit la valeur du paramètre θ correspondant à une valeur de H_1, sa puissance est supérieure à la puissance de tout autre test. Il en est ainsi quand la région critique ne dépend pas de la valeur du paramètre θ.

Exemple : Reprenons le test sur une moyenne de loi de Laplace-Gauss avec comme hypothèses :

$$H_0 : m = m_0 \qquad H_1 : m = m_1 > m_0$$

La région critique ne dépend pas explicitement de m_1, elle est donc la même quelque soit m_1. Ce test est donc UPP pour les hypothèses ci-dessus ; en revanche il n'est pas UPP si on prend pour hypothèse $H_1 : m \neq m_0$.

VIII-4-2 Tests entre deux hypothèses composites

L'hypothèse H_0 étant elle-même composite, α dépend de la valeur du paramètre. Il est donc nécessaire d'imposer : $\alpha(\theta) \leq \alpha$ donné.

Il existe un test UPP pour les hypothèses suivantes :

$$\boldsymbol{H_0 : \theta < \theta_0 \qquad H_1 : \theta \geq \theta_0}$$

ou

$$\boldsymbol{H_0 : \theta \leq \theta_1 \qquad H_1 : \theta_1 < \theta \leq \theta_2}$$

Ce théorème dû à Lehman est admis. Il suppose l'existence d'une statistique G telle que le rapport $\dfrac{L(\underline{x}, \theta_1)}{L(\underline{x}, \theta_2)}$ soit une fonction croissante de G si $\theta_1 > \theta_2$. De telles statistiques sont données par les statistiques des lois exponentielles.

VIII-4-3 Test du rapport des vraisemblances maximales

VIII-4-3-1. Test entre deux hypothèses simples

$$\boldsymbol{H_0 : \theta = \theta_0 \qquad H_1 : \theta \neq \theta_0}$$

Posons : $\lambda = \dfrac{L(\underline{x}, \theta_0)}{\sup_\theta L(\underline{x}, \theta)}$; on a : $0 \leq \lambda \leq 1$.

Grâce au principe du maximum de vraisemblance, plus λ est grand, plus H_0 est vraisemblable. Tout se passe comme si on remplaçait, dans H_1, le paramètre θ par son estimation obtenue par la méthode du maximum de vraisemblance.

La région critique du test est donc de la forme $\lambda < k$.

Théorème (admis)

Sous l'hypothèse H_0, la distribution de $-2\ln\lambda$ est, à la limite, celle d'un χ^2_p.

VIII-4-3-2. Test entre deux hypothèse composites

Les résultats ci-dessus sont encore valables si on pose :

$$\lambda = \frac{\sup_{H_0} L(\underline{x}, \theta)}{\sup_{H_1} L(\underline{x}, \theta)}$$

VIII- 5) TESTS SUR UNE MOYENNE

VIII-5-1 Moyenne m d'une loi normale LG(m, σ) où σ est connu

Dans les différents cas étudiés ici, la variable de décision est clairement $\overline{X}$. On sait qu'elle suit la loi $LG(m, \frac{\sigma}{\sqrt{n}})$.

VIII-5-1-1. Test unilatéral

Considérons le test suivant :

$$H_0 : m = m_0 \qquad H_1 : m = m_1 > m_0$$

La région critique est définie par :

$$\overline{X} > k$$

La valeur de k se déduit de la définition de α et de la loi normale centrée réduite :

$$\alpha = P(\overline{X} > k) = P(\frac{\overline{X} - m_0}{\sigma/\sqrt{n}} > \frac{k - m_0}{\sigma/\sqrt{n}})$$

Si la valeur de $\overline{X}$ obtenue sur l'échantillon vérifie $\overline{X} > k$, l'hypothèse H_0 est rejetée et on adopte $m = m_1$.

La valeur de k étant déterminée, on peut alors calculer le risque de deuxième espèce β :

$$\beta = P(\overline{X} < k / H_1) = P(\frac{\overline{X} - m_1}{\sigma/\sqrt{n}} < \frac{k - m_1}{\sigma/\sqrt{n}})$$

Remarque

Si on considère le test : $H_0 : m = m_0 \qquad H_1 : m = m_1 < m_0$, la variable de décision est la même et il suffit d'inverser les inégalités pour déterminer k puis β.

VIII-5-1-2. Test bilatéral

$$H_0 : m = m_0 \qquad H_1 : m = m_1 \neq m_0 .$$

La région critique est alors définie par :

$$P(\left|\overline{X} - m_0\right| > k) = \alpha$$

ce qui permet de calculer k grâce aux tables de la loi normale : $\alpha = P(\sqrt{n}\left|\frac{\overline{X} - m_0}{\sigma}\right| > \sqrt{n}\,\frac{k}{\sigma})$

De la même façon, on obtient le risque de deuxième espèce β.

VIII-5-2 Moyenne m d'une loi normale LG(m, σ) où σ est inconnu

La variable de décision est toujours $\overline{X}$.

La loi suivie par cette variable est la loi de Student : $T_{n-1}(m, \frac{S}{\sqrt{n-1}})$ où S est l'estimateur usuel de σ.

On procède de la même façon que dans le cas précédent, en utilisant cette fois les tables de la loi de Student.

VIII-5-3 Moyenne m d'une loi quelconque

Si la variable parente X suit une loi inconnue, les tests sur la moyenne d'une loi normale, σ connu ou inconnu, s'appliquent encore, dès que n est grand (n > 30 en pratique) par application du théorème de la limite centrale.

VIII- 6) TESTS SUR L'ECART-TYPE σ D'UNE LOI NORMALE

On considère la loi LG(m, σ). On teste la valeur de l'écart-type σ, à m connu ou inconnu.

VIII-6-1-1. Ecart-type d'une loi normale où m est connue

La variable de décision est la statistique $D = \frac{1}{n}\sum_{i=1}^{n}(X_i - m)^2$ estimateur sans biais de la variance. On sait que la variable $\frac{nD}{\sigma^2}$ suit une loi χ^2_n.

Si on considère le test : $H_0 : \sigma = \sigma_0 \qquad H_1 : \sigma = \sigma_1 > \sigma_0$

la région critique est définie par $D > k$ où k est déterminé par :

$$\alpha = P(D > k) = P(\frac{nD}{\sigma_0^2} > \frac{nk}{\sigma_0^2})$$

La valeur de k est obtenue grâce aux tables de la loi du χ^2.

Le test $H_0 : \sigma = \sigma_0 \qquad H_1 : \sigma = \sigma_1 < \sigma_0$ se conduit de la même manière en inversant le signe de l'inégalité pour la région critique.

VIII-6-1-2. Ecart-type d'une loi normale où m est inconnue

La variable de décision est la statistique $S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$ pour laquelle on sait que $\frac{(n-1)S^2}{\sigma^2}$ suit une loi χ^2_{n-1}.

Si on considère le test : $H_0 : \sigma = \sigma_0 \qquad H_1 : \sigma = \sigma_1 > \sigma_0$, la région critique est définie par : $S^2 > k$ où k est déterminé par :

$$\alpha = P(S^2 > k) = P(\frac{nS^2}{\sigma_0^2} > \frac{nk}{\sigma_0^2})$$

Ce cas est le plus fréquent. En effet, il est rare d'avoir à tester la variance d'une loi normale dont on connaît la moyenne, alors que ce test est usuel dans le cas où la moyenne n'est pas connue.

Les deux tests sur la variance d'une loi normale ne sont valables que dans le cas d'une loi normale.

VIII- 7) TEST SUR UNE PROPORTION p

Considérons le test suivant où p est une proportion d'individus possédant une propriété particulière dans une population :

$$H_0 : p = p_0 \qquad H_1 : p \neq p_0$$

La fréquence empirique *F,* proportion d'individus présentant le caractère considéré dans un échantillon de taille n, est un estimateur sans biais de p qui suit approximativement une loi normale pour n suffisamment grand (n > 40) :

$$L(F) = LG(p, \sqrt{\frac{p(1-p)}{n}})$$

la région critique est déterminée par :

$$\alpha = P(|F - p_0| > k) = P\left(\frac{|F - p_0|}{\sqrt{\frac{p(1-p)}{n}}} > \frac{k}{\sqrt{\frac{p(1-p)}{n}}}\right)$$

Le calcul de la fréquence empirique f valeur de F sur un échantillon donné permet de conclure sur l'acceptation de l'hypothèse H_0

Exemple :

Un sondage sur un échantillon de 625 cadres révèle qu'il y a 48% d'entre eux qui utilise Internet. Or une hypothèse a été émise comme quoi la moitié des cadres utilise Internet. Le sondage est-il contradictoire avec cette hypothèse au niveau de confiance 95% ?

Le niveau de confiance étant de 95%, le risque α est égal à 0,05. H_0 est l'hypothèse conservatoire donc les hypothèses sont les suivantes :

$$H_0 : \quad p = p_0 = 0{,}5$$

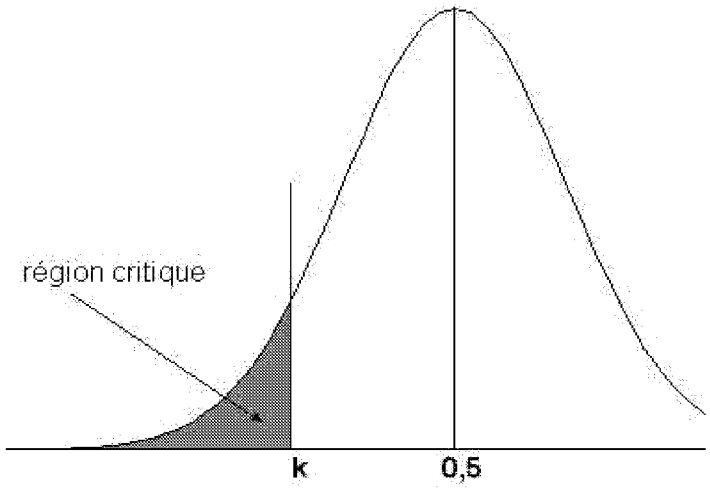
$$H_1 : \quad p < 0{,}5$$

La variable F suit une loi normale $LG(0{,}5\,;\sqrt{\frac{0{,}5\times 0{,}5}{625}}) = LG(0{,}5\,;0{,}02)$.

La région critique est déterminée par : $\alpha = 0{,}05 = P(F < k) = P(\frac{F-0{,}5}{0{,}02} < \frac{k-0{,}5}{0{,}02})$

On en déduit que $\frac{k-0{,}5}{0{,}02} = -1{,}645$ (table de la loi normale) soit $k = 0{,}467$.

Le sondage a donné $f = 0{,}48$ >k donc on garde H_0.



Ch. IX Tests d'ajustement

En Statistique, il est certaines fois nécessaire d'apprécier si un ensemble d'observations, réalisées sur un ou plusieurs échantillons, peut correspondre à une distribution théorique, cet ajustement devant être valable pour la distribution entière.

Aussi des méthodes d'ajustement ont été, depuis longtemps, mises au point : méthodes empiriques, méthodes graphiques et enfin tests statistiques.

Si on appelle F* la fonction de répartition de la variable échantillonnée et F la fonction de répartition de la distribution théorique, on est amené à choisir une des hypothèses suivantes :

$$H_0 \quad F^*(x) = F(x)$$

$$H_1 \quad F^*(x) \neq F(x)$$

On rappellera ici les procédures empiriques usuelles, puis on verra les ajustements graphiques les plus courants avant de développer l'étude des tests statistiques appropriés.

IX- 1) METHODES EMPIRIQUES

IX-1-1 Forme de l'histogramme

Cette étude permet principalement d'éliminer certains modèles, en particulier pour des absences de symétrie. Toutefois, une forme symétrique ne conduit pas systématiquement à une loi normale : d'autres lois ont le même type de courbe de densité (Student, Cauchy…).

Une forme vraiment dissymétrique peut faire penser à une loi log-normale, ou à une loi exponentielle ou encore à une loi de Weibull.

Il est clair que le choix entre deux lois de même forme tiendra compte du phénomène étudié.

IX-1-2 Propriétés mathématiques

Pour un certain nombre de lois, des relations existent entre les paramètres :

- Si X suit une loi de Poisson, on a $E(X) = V(X)$. On s'assure alors que, sur l'échantillon :

$$\overline{X} \approx S^{*2}$$

Il en est de même pour la loi Gamma. Ces deux lois sont, par ailleurs, très différentes puisque l'une est discrète et l'autre est continue.

- Dans le cas d'une loi de Laplace-Gauss :

$$\text{le coefficient d'aplatissement : } \gamma_1 = \frac{E[(X-m)^3]}{\sigma^3} \text{ est égal à 0}$$

$$\text{et le coefficient d'asymétrie } \gamma_2 = \frac{E[(X-m)^4]}{\sigma^4} \text{ est égal à 3.}$$

On vérifie alors sur l'échantillon que les coefficients empiriques correspondent à peu près aux valeurs théoriques. On utilise pour cela des tables donnant les valeurs critiques de ces coefficients en fonction de la taille de l'échantillon.

IX- 2) AJUSTEMENTS GRAPHIQUES

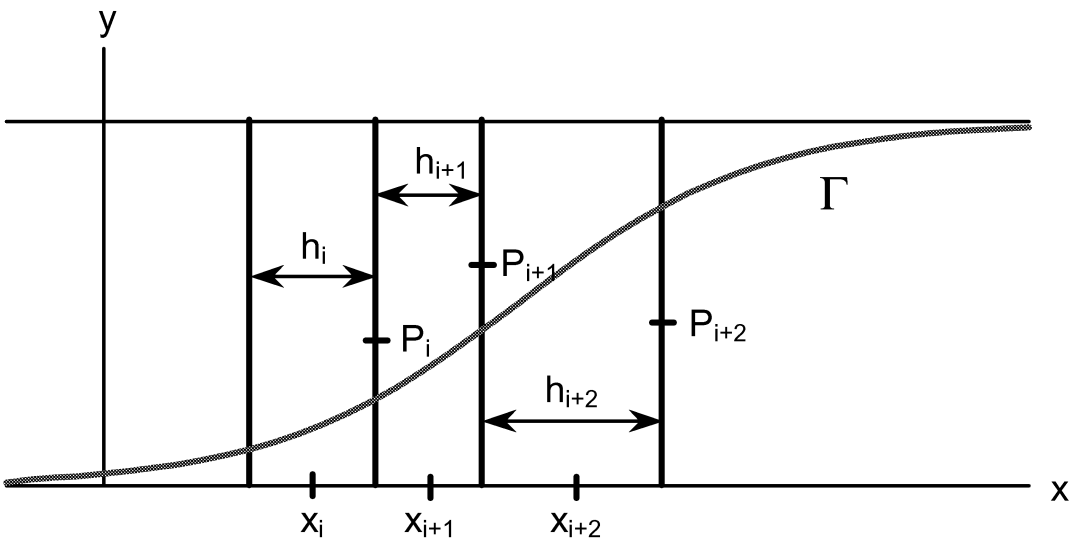
Dès que la taille de l'échantillon sur lequel on cherche à ajuster une distribution connue est grande, la fonction de répartition empirique de l'échantillon diffère peu de la fonction de répartition théorique F.

IX-2-1 Fonctions de répartition empirique et théorique

On considère une série de n observations réparties en k classes. On peut alors construire la fonction de répartition empirique F* de l'échantillon et elle doit être peu différente de la fonction de répartition théorique F.

Soit x_i le centre d'une classe et h_i l'étendue de cette classe;

le point P_i de coordonnées $(x_i + \frac{h_i}{2}, \frac{1}{n}\sum_i n_i)$ est un point de la fonction de répartition empirique.



Si les points P_i ne sont pas trop éloignés de la courbe Γ de la fonction de répartition F on peut admettre que la loi suivie par les observations est voisine de la loi correspondante.

Cette méthode présente un inconvénient majeur : il est, en général, difficile de juger de la distance de points (les points P_i) à une courbe (ici la courbe Γ de la fonction de répartition théorique).

Aussi un procédé mathématique est utilisé. Pour la plupart des lois de probabilité, une fonction simple permet de transformer la courbe de répartition en une droite. Il reste à vérifier l'adéquation des données au modèle en s'assurant de la linéarité des points représentant la fonction de répartition empirique sur un papier à échelle adaptée à la transformation. **Ce papier est dit à échelle fonctionnelle.**

Pour rendre cette estimation plus facile, on cherche donc, pour chaque distribution usuelle, une transformation mathématique simple permettant de représenter la fonction de répartition par une droite. Le cas le plus connu est celui de la droite de Henry pour la loi Laplace-Gauss. Nous allons voir les trois cas les plus usuels.

IX-2-2 Ajustement graphique spécifique à la loi exponentielle

On sait que la fonction de répartition de cette loi exponentielle de paramètre λ est donnée par :

$$F(x) = 1 - e^{-\lambda x}$$

$$\text{soit } Ln[1 - F(x)] = -\lambda x$$

Pour un échantillon de taille n, on reporte pour chaque valeur x de la variable, la valeur de $1 - F^*(x)$ sur du papier logarithmique.

Ainsi supposons que l'on étudie la durée de vie d'un composant électronique. On cherche à montrer que cette durée de vie s'écrit $1 - F(x) = e^{-\lambda x}$ pour une valeur de λ à déterminer.

Pour un échantillon de taille n, on reporte alors, pour chaque valeur x du temps de fonctionnement, le pourcentage de composants encore en activité à cette date sur du papier logarithmique.

On reporte donc les points : $[\, x_i, -\log(1 - \frac{i-1}{n}) \,]$ **pour $1 \leq i \leq n$, où les x_i sont ordonnés.**

Si les points semblent alignés, on peut conclure à l'adéquation du modèle exponentiel.

La pente de la droite tracée fournit une estimation graphique du paramètre λ de la loi.

IX-2-3 Ajustement graphique spécifique à la loi de Weibull

La loi de Weibull a pour fonction de répartition :

$$F(x) = 1 - e^{-\lambda x^{\beta}} .$$

On a donc : $P(X > x) = e^{-\lambda x^{\beta}}$ soit $Log(P(X > x) = -\lambda x^{\beta}$

$$Log\,[-Log(P(X > x)] = Log\lambda + \beta\, Logx$$

En pratique, si on dispose d'un échantillon (x_1, x_2, …, x_n), on reporte sur un papier spécifique, appelé papier d'Alan Plait, les points de coordonnées :

$$[\log x_i, \log(-\log(1 - \frac{i-1}{n}))\,].$$

Là encore, si l'alignement des points est satisfaisant, on peut conclure à l'adéquation du modèle choisi (loi de Weibull).

La pente de la droite obtenue fournit une estimation graphique de β et son ordonnée à l'origine une estimation de $\log \lambda$.

Un exemplaire de feuille de papier d'Alan Plait est situé en fin de chapitre.

IX-2-4 Ajustement graphique spécifique à la loi de Laplace-Gauss

IX-2-4-1. Présentation

Si X est une variable normale de paramètres m et σ, on a vu que la variable centrée réduite associée :

$$U = \frac{X - m}{\sigma}$$

suit une loi normale LG(0, 1).

La transformée de la fonction de répartition théorique de la loi normale dans le plan (U, X) est une droite de pente 1/σ appelée **Droite de Henry** du nom de son inventeur, commandant d'artillerie à Fontainebleau (1894).

Pour vérifier qu'une population suit une loi normale, on utilise alors un papier spécial dit Gauss-linéaire ou Gausso-arithmétique (voir un exemplaire en fin de chapitre) sur lequel l'axe des ordonnées est gradué selon les valeurs de F proportionnellement aux valeurs de U.

IX-2-4-2. Application pratique

Considérons n observations (x_1, x_2, …, x_n) provenant d'une variable normale LG (m, σ) ; les valeurs $u_i = \frac{x_i - m}{\sigma}$ ($1 \le i \le n$) constituent alors un échantillon d'une variable normale centrée réduite.

On ordonne ces observations en classes et on porte sur le papier gausso-arithmétique :

en abscisses, les valeurs des observations, limites supérieures des classes

en ordonnées, les fréquences cumulées correspondantes.

Si les points obtenus sont sensiblement alignés, on peut considérer que le modèle théorique est bien une loi de Laplace-Gauss.

La valeur de la moyenne m est obtenue en cherchant l'abscisse du point d'ordonnée : $F = 0{,}50$.

La valeur de σ est obtenue par l'une des deux valeurs de F suivantes :

✓ **Pour $F = 0{,}8415$ $(U = 1)$, on lit l'abscisse correspondante x_i et $\sigma = x_i - m$**

✓ **Pour $F = 0{,}1585$ $(U = -1)$, x_i étant encore l'abscisse correspondante $\sigma = -x_i + m$**

On peut comparer les valeurs lues sur le graphique pour m et σ avec les estimations de ces paramètres calculées sur l'échantillon.

Remarques

- **Dans chaque classe, on doit avoir au moins 5 objets. Si cette condition n'est pas remplie, on regroupe certaines classes.**
- **En abscisse, on doit porter les limites supérieures des classes.**

IX- 3) TEST DU χ^2

IX-3-1 Présentation

Soit une variable aléatoire X discrète ou discrétisée, divisée en k classes de probabilités respectives p_i, $1 \leq i \leq k$.

Soit un échantillon de cette variable fournissant les effectifs aléatoires N_i, $1 \leq i \leq k$, dans chacune des classes précédentes.

On a bien sûr : $E(N_i) = np_i$.

On considère la statistique :

$$D^2 = \sum_{i=1}^{k} \frac{(N_i - np_i)^2}{np_i}$$

D^2 est une mesure relative de l'écart entre les effectifs réalisés et les effectifs théoriques.

Les termes de la statistique D^2 ne sont pas indépendants puisque :

$$\sum_{i=1}^{k} N_i = n$$

Le nombre de degrés de liberté de D^2est égal à (k-1).

D'après un résultat sur la loi multinomiale, pour n suffisamment grand, D^2 suit une loi du χ^2 à k – 1 degrés de liberté.

On en déduit le test du χ^2.

IX-3-2 Conduite du test

Considérons une variable X pour laquelle on cherche à prouver que sa distribution est une distribution théorique particulière, discrète voire continue, de fonction de répartition F. Les hypothèses du test sont les suivantes :

H_0 : la distribution de X est la distribution théorique proposée

H_1 : la distribution de X n'est pas cette distribution proposée

Soit $(x_1, x_2, \ldots, x_n)$ un échantillon de cette variable réparti en k classes, soit $N_1, \ldots, N_k$ les effectifs empiriques de ces k classes et soit $(p_1, \ldots, p_k)$ les probabilités théoriques :

$$p_i = F(x_{i+1}) - F(x_i)$$

On choisit un risque de première espèce α et on détermine le seuil critique k tel que :

$$P(\chi^2_{k-1} > k) = \alpha$$

On calcule la statistique : $d^2 = \sum_{i=1}^{k} \frac{(N_i - np_i)^2}{np_i}$ **sur l'échantillon et on compare à k.**

Si $d^2 = \sum_{i=1}^{k} \frac{(N_i - np_i)^2}{np_i} > k$ **, on rejette H_0 et dans le cas contraire, on garde H_0.**

Un exemple de test est proposé en fin de chapitre.

IX-3-3 Compléments

IX-3-3-1. Cas de paramètres estimés

Il arrive fréquemment que l'on veuille tester un ajustement avec une loi dont seule la forme de la distribution est spécifiée et dont les paramètres font l'objet d'une estimation (par exemple le paramètre d'une loi de Poisson estimé par la moyenne des observations de l'échantillon…). Ces estimations doivent être faites avec la méthode du maximum de vraisemblance.

Dans ce cas, si r est le nombre de paramètres estimés, le nombre de degrés de liberté de la variable du χ^2 est k – r –1.

IX-3-3-2. Effectifs par classe

La loi de D^2 peut être assimilée à la loi du χ^2 dès que np_i est supérieur à 5 pour chaque classe. S'il y a des classes à effectifs trop faibles, on les regroupe ensemble, voire avec la classe suivante ou la classe précédente.

IX-3-3-3. Lois continues

Si X est une variable continue, on découpera en classes d'égales amplitudes plutôt qu'en classes équiprobables mais le test du χ^2 n'est pas le plus approprié pour tester une distribution continue.

IX- 4) TEST DE KOLMOGOROV, TEST DE CRAMER

IX-4-1 Test de Kolmogorov-Smirnov

Il s'agit d'un test non paramétrique d'ajustement à une distribution continue dont la fonction de répartition est F (x).

Soit F_n^* la fonction de répartition empirique d'un échantillon de taille n de X.

La variable de décision est la variable aléatoire :

$$D_n = Sup_{x \in R} \left| F_n^*(x) - F(x) \right|$$

Les travaux de Kolmogorov et Glivenko ont montré que la fonction de répartition de la variable K_n définie par : $K_n = \sqrt{n} D_n$ converge, quand n $\to +\infty$, vers :

$$K(y) = \sum_{k=-\infty}^{+\infty} (-1)^k \exp(-2k^2 y^2) \qquad \text{pour } y > 0$$

$$K(y) = 0 \qquad \text{pour } y \le 0$$

Cette distribution a été tabulée.

La règle de décision est donc :

On rejette l'hypothèse H_0 dès que la valeur de la statistique D_n est supérieure à une valeur d_n n'ayant qu'une probabilité α d'être dépassée. Cette valeur d_n sera cherchée dans les tables du test de Kolmogorov.

Dans le cas contraire, on garde H_0 et on considère que la distribution proposée est acceptable.

Remarque :

Le test de Kolmogorov est préférable à celui du χ^2 pour des variables continues car la variable de décision utilise l'échantillon tel qu'il est fourni au départ sans regrouper les données en classes.

IX-4-2 Test de Cramer, Von-Mises

Posons :

$$n\omega_n^2 = \int_{-\infty}^{+\infty} [F_n^*(x) - F(x)]^2 \, dF(x)$$

c'est une variable aléatoire dont la loi, indépendante de F, sert à tester :

$$H_0 : F^*(x) = F(x)$$

$$H_1 : F^*(x) \neq F(x)$$

Cette variable aléatoire représente la mesure de l'écart entre une répartition théorique et une répartition empirique.

On démontre que :

$$n\omega_n^2 = \frac{1}{12n} + \sum_{i=1}^{n} [\frac{2i-1}{2n} - F(x_i)]^2$$

où les x_i sont les valeurs ordonnées croissantes de l'échantillon.

Règle de décision :

On rejette l'hypothèse H_0 dès que la valeur de la statistique $n\omega_n^2$ est supérieure à une valeur n'ayant qu'une probabilité α d'être dépassée.

Au seuil $\alpha = 0{,}05$, on rejette H_0 dès que $n\omega_n^2 > 0{,}46136$.

IX- 5) EXEMPLES D'APPLICATION

IX-5-1 Test du caractère exponentiel d'une loi de fiabilité

On dispose de n matériels identiques et on note les durées de vie x_i en heures de chacun.

Si la durée de vie est supposée être une variable exponentielle :

$$F(x) = P(X < x) = 1 - e^{-\lambda x}$$

On estime le paramètre de la loi grâce aux valeurs observées sur l'échantillon puis on applique le test de Kolmogorov en utilisant la variable de décision :

$$D_n = Sup_{x \in R} |F_n^*(x) - F(x)|$$

Exemple :

On a relevé les durées de vie x_i, en heures, de cinq matériels identiques :

133	169	8	122	58

On estime le paramètre de la loi exponentielle :

$$\lambda = \frac{1}{\bar{x}} = \frac{1}{98}$$

On calcule les valeurs de la fonction de répartition théorique :

$$F(x) = P(X < x) = 1 - e^{-\lambda x} = 1 - e^{-\frac{x}{98}}$$

ainsi que les valeurs F_i.

Les valeurs de la fonction de répartition empirique et celles de la fonction de répartition théorique sont présentées dans le tableau ci-dessous :

x_i	8	58	122	133	169
$F(x_i)$	0,079	0,447	0,711	0,743	0,821
F_i	0	0,2	0,4	0,6	0,8

Le calcul de la statistique D_n, $D_n = Sup_{x \in R} |F_n^*(x) - F(x)|$, nous donne : $D_n = 0{,}311$. Elle correspond à la valeur $x_i = 122$.

Si on choisit un risque de $\alpha = 0{,}05$, le seuil critique est de 0,56 ; on garde l'hypothèse H_0 : la distribution exponentielle est acceptée.

IX-5-2 Test du caractère poissonnien d'une file d'attente

On souhaite montrer que la variable : « nombre d'arrivées dans une file d'attente pendant un temps donné »suit une loi de Poisson.

Sur l'échantillon, on évalue le paramètre de la loi par la moyenne puis on applique le test du χ^2 car la distribution est discrète.

Exemple : On relève durant 50 intervalles n_i de deux minutes, le nombre X_i de voitures se présentant à un péage.

X_i	0	1	2	3	≥4
n_i	21	18	7	3	1

On veut montrer que la variable X suit une loi de Poisson. On calcule la moyenne et la variance empiriques égales respectivement à 0,9 et 0,97. On en déduit une estimation du paramètre : $\lambda = 0{,}9$.

On calcule les probabilités p_i de la loi de Poisson :

X_i	0	1	2	3	≥4
n_i	21	18	7	3	1
p_i	0,407	0,366	0,165	0,049	0,013
p_i	0,407	0,366	0,227		

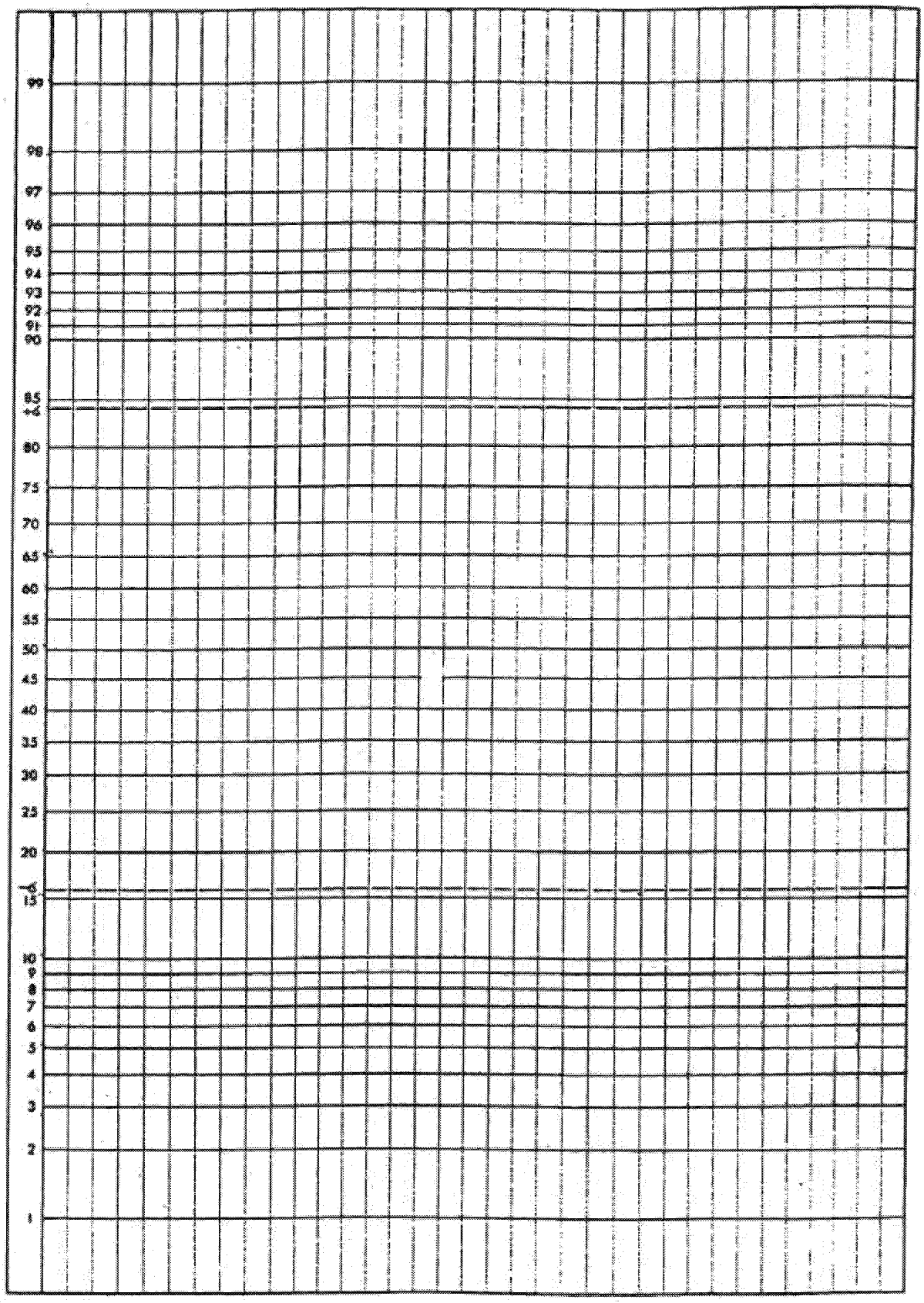
On regroupe les classes à effectifs inférieurs à 5.

$$d^2 = \sum_i \frac{(n_i - np_i)^2}{np_i} = 0{,}033$$

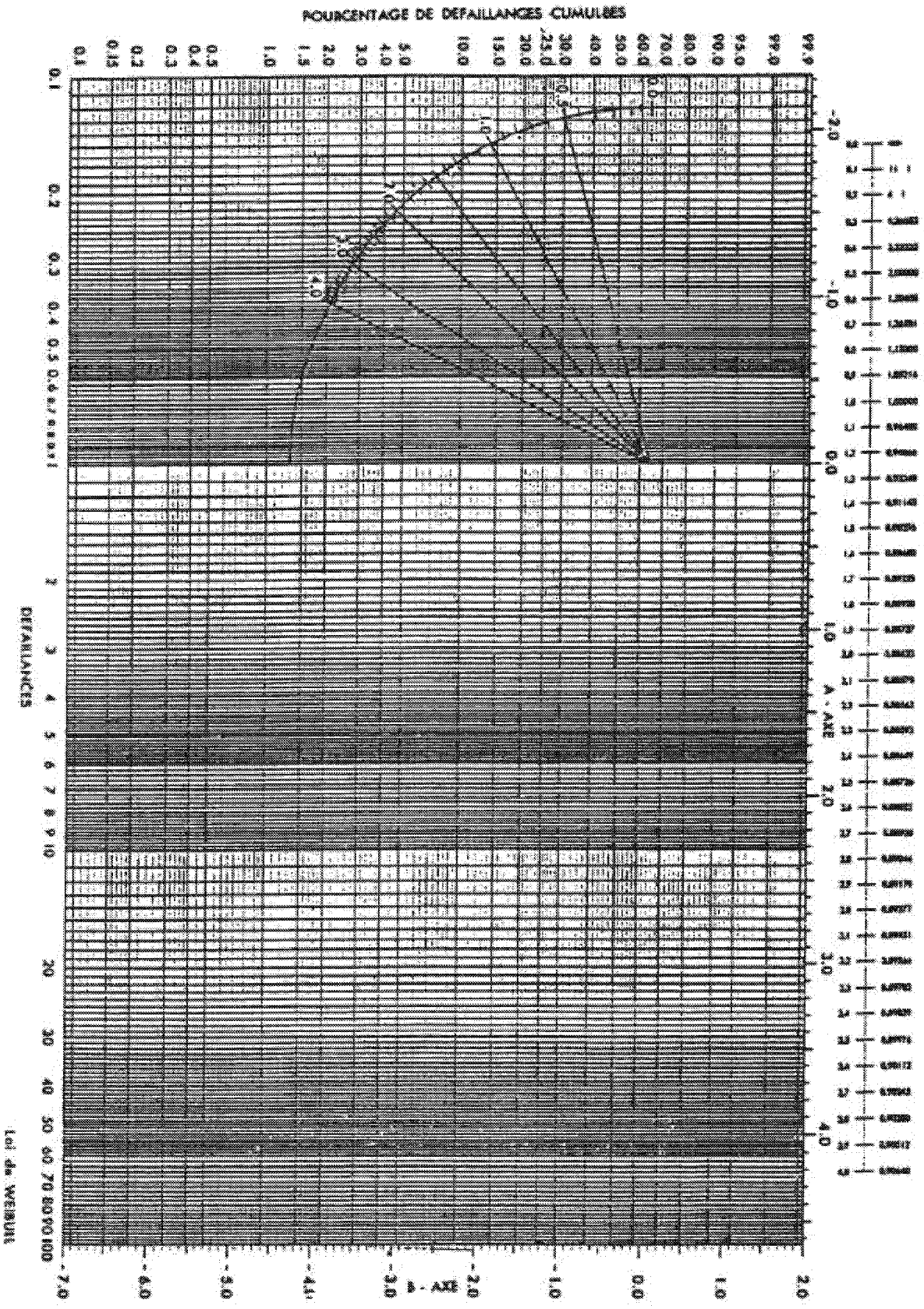
On a 3 classes et un paramètre estimé donc le nombre de degrés de liberté de la loi du Chi-deux est 3-1-1 = 1. On compare donc d^2 avec $\chi^2_{0,95}(1) = 3{,}84$.

On admet que la distribution de X est une distribution poissonnienne.

PAPIER GAUSSO-ARITHMETIQUE



PAPIER D'ALAN PLAIT



Ch. X Tests de comparaison

Soit deux échantillons de tailles n_1 et n_2. On suppose qu'ils ont été prélevés indépendamment l'un de l'autre. La question, qu'il est naturel de se poser alors, est de savoir s'ils sont issus de la même population.

Soit X_1 la variable parente du premier échantillon, de fonction de répartition F_1, et X_2 celle du deuxième échantillon, de fonction de répartition F_2. Le test peut alors se formaliser ainsi :

$$H_0 : \quad F_1(x) = F_2(x)$$

$$H_1 : \quad F_1(x) \neq F_2(x)$$

Dans la pratique, on se contente de vérifier l'égalité des espérances et des variances des deux variables X_1 et X_2 en disposant des moyennes et des variances empiriques des deux échantillons.

X- 1) PARAMETRES D'ECHANTILLONS GAUSSIENS

Il est nécessaire de comparer les variances des deux échantillons avant de pouvoir comparer les moyennes.

X-1-1 Test des variances de Fisher-Snedecor

Le test se présente ainsi :

$$H_0 : \sigma_1 = \sigma_2$$

$$H_1 : \sigma_1 \neq \sigma_2$$

En appliquant les résultats obtenus dans le chapitre de l'échantillonnage, on sait que la variable $\frac{nS^2}{\sigma^2}$ suit une loi du Chi-deux : χ^2_{n-1}. Posons $F = \frac{S_1^2}{S_2^2}$. F est le rapport des deux estimateurs de σ_1^2 et σ_2^2. Si ces variances sont égales, le rapport F doit être peu différent de 1. **F est la variable de décision**.

Sous l'hypothèse H_0, et d'après la définition de la loi de Fisher :

$$L\left(\frac{n_1 S_1^2 \,/\, n_1 - 1}{n_2 S_2^2 \,/\, n_2 - 1}\right) = F(n_1 - 1, n_2 - 1)$$

En pratique, on met toujours au numérateur la plus grande des estimations.

La région critique est alors de la forme : $F > k_0$

Le risque de première espèce α étant choisi, k_0 est déterminé par la valeur correspondante de la variable de Fisher de degrés de liberté $n_1 - 1$ et $n_2 - 1$.

Si la valeur empirique de F calculée sur l'échantillon est inférieure à k_0 alors on peut conclure à l'égalité des variances, sinon on rejette cette égalité.

Si le test de Fisher-Snedecor aboutit à la conservation de H_0, on passe au test des espérances.

X-1-2 Test des espérances de Student

On suppose désormais que les variances sont égales mais que l'on ne connaît pas leur valeur.

On teste les hypothèses :

$$H_0 : m_1 = m_2$$

$$H_1 : m_1 \neq m_2$$

Pour chaque échantillon :

$$L(\frac{nS^2}{\sigma^2}) = \chi^2_{n-1}$$

$$L(\overline{X}) = LG(m, \frac{\sigma}{\sqrt{n}})$$

On en déduit : $$L(\frac{nS_1^2 + nS_2^2}{\sigma^2}) = \chi^2_{n_1+n_2-2}$$

$$L(\overline{X_1} - \overline{X_2}) = LG(m_1 - m_2, \sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}})$$

Considérons la variable : $T = \frac{\overline{X_1} - \overline{X_2} - (m_1 - m_2)}{\sqrt{(n_1 S_1^2 + n_2 S_2^2)(\frac{1}{n_1} + \frac{1}{n_2})}} \sqrt{n_1 + n_2 - 2}$. Elle suit une loi de Student $T_{n_1+n_2-2}$.

La région critique est alors déterminée, le risque α étant choisi, par la lecture dans la table de la loi de Student du seuil k_0 tel que la probabilité $P(|T| > k_0) = \alpha$.

Sous l'hypothèse H_0 : $m_1 = m_2$, on cherche la valeur empirique de T obtenue avec les deux échantillons et on peut conclure sur le rejet ou non de l'hypothèse H_0 donc sur l'égalité ou non des moyennes.

X-1-3 Paramètres d'échantillons non Gaussiens

Le test de variances n'est plus valable.

Toutefois, pour n_1 et n_2 suffisamment grands (supérieurs à 30), que σ_1 soit ou non différent de σ_2, on peut tester les moyennes par le test de Student (le test est dit robuste).

X- 2) TESTS NON PARAMETRIQUES DE COMPARAISON

X-2-1 Test de Smirnov

Soit deux échantillons de tailles n_1 et n_2 de fonctions de répartition empiriques F* et G*.

Le test est le suivant:

$$H_0 : F(x) = G(x)$$

$$H_1 : F(x) \neq G(x)$$

On pose : $H(x) = \sqrt{\frac{n_1 n_2}{n_1 + n_2}} Sup\left| F(x) - G(x) \right|$

Sous l'hypothèse H_0 supposant que les deux échantillons sont issus d'une même variable, alors la variable H suit la loi de la variable K vue au chapitre précédent (test de Kolmogorov).

La variable de décision sera donc : $Sup\left| F^*(x) - G^*(x) \right|$. On rejettera H_0 dès que la valeur calculée de cette variable sur les échantillons sera trop grande.

X-2-2 Test de Wilcoxon

Soit $(x_1, x_2, \ldots, x_n)$ et $(y_1, y_2, \ldots y_m)$ deux échantillons. Le test repose sur le fait qu'en mélangeant les deux séries par valeurs croissantes on doit obtenir un mélange homogène.

Les deux suites étant réordonnées, on compte le nombre U de couples (x_i, y_j) où x_i a un rang plus grand que y_j ou bien tels que $x_i > y_j$ si les variables sont quantitatives. Ce nombre U varie de 0 à nm.

Si les deux distributions sont issues de la même population :

$$E(U) = \frac{nm}{2} \text{ et } V(U) = \frac{nm(n+m+1)}{12}$$

et la distribution de U est asymptotiquement gaussienne. L'approximation est tout à fait acceptable pour n et m supérieurs ou égaux à 8.

La variable de décision est donc $U - \frac{nm}{2}$, variable normale centrée.

La région critique est définie par : $\left| U - \frac{nm}{2} \right| > k_0$ où k_0 est déterminé, le risque α étant choisi, par la lecture de la table de la loi de Laplace-Gauss de la variable centrée réduite associée. L'hypothèse H_0 sera alors rejetée.

Méthode pratique :

On calcule la somme des rangs des individus d'un des deux groupes (le premier par exemple) : Σ_X. Il est simple de montrer que :

$$\Sigma_X = U + \frac{n(n+1)}{2}$$

Sous l'hypothèse nulle, $E(\Sigma_X) = \frac{n(n+m+1)}{2}$ $\qquad V(\Sigma_X) = \frac{nm(n+m+1)}{12}$.

Pour des effectifs n et m supérieurs à 8, la loi de la variable centrée réduite associée tend vers une loi normale LG(0, 1). Cette variable Σ_X (ou la variable centrée réduite associée) est donc la variable de décision.

La région critique est définie par $\Sigma_X > k_0$, k_0 étant déterminé, le risque α étant choisi, à l'aide de la table de la loi de Laplace-Gauss. L'hypothèse H_0 est rejetée dès que Σ_X est supérieure au seuil critique.

X- 3) COMPARAISON DE PLUSIEURS ECHANTILLONS

On dispose de k échantillons décrits par une **variable qualitative** prenant r modalités.

Les données sont présentées dans un tableau :

	Modalité 1	Modalité 2	…	Modalité r	Total
Echantillon 1	n_{11}	n_{12}		n_{1r}	$n_{1.}$
Echantillon 2	n_{21}	n_{22}		n_{2r}	$n_{2.}$
…					
Echantillon k	n_{k1}	n_{k2}		n_{kr}	$n_{k.}$
Total	$n_{.1}$	$n_{.2}$		$n_{.r}$	N

n_{ij} est le nombre d'individus de l'échantillon i possédant la modalité j de la variable, $n_{.j}$ est le nombre total d'individus possédant la modalité j et $n_{i.}$ est l'effectif de l'échantillon i.

On cherche à déterminer si les échantillons proviennent ou non de la même population.

Sous l'hypothèse H_0, on a les probabilités $p_1, p_2, ..., p_r$ de posséder les modalités 1, 2,..., r. On est donc amené à comparer les effectifs constatés n_{ij} aux effectifs espérés $n_{i.}p_j$; pour cela on calcule :

$$d_0^2 = \sum_{i=1}^{k} \sum_{j=1}^{r} \frac{(n_{ij} - n_{i.}\, p_j)^2}{n_{i.}\, p_j}$$

d_0^2 est la réalisation d'une variable suivant un loi du Chi-deux à $kr - k$ degrés de liberté car on a kr termes liés par k relations.

En général, les probabilités p_j ne sont pas connues, elles sont estimées par les rapports $\frac{n_{.j}}{N}$.

Comme elles sont liées par leur somme égale à 1, on estime donc $r - 1$ paramètres.

Le calcul de la statistique d_0^2 donne alors : $d_0^2 = \sum_{i=1}^{k} \sum_{j=1}^{r} \frac{(n_{ij} - \frac{n_{i.}n_{.j}}{N})^2}{\frac{n_{i.}n_{.j}}{N}} = N(\sum_{i=1}^{k} \sum_{j=1}^{r} \frac{(n_{ij})^2}{n_{i.}n_{.j}} - 1)$

et elle suit une loi du Chi-deux à $kr - k - (r-1) = (k-1)(r-1)$ degrés de liberté.

On rejettera l'hypothèse H_0 dès que d^2 est supérieur au seuil critique déterminé, pour un risque α donné, par la valeur de la variable du Chi-deux.

X- 4) TEST DE COMPARAISON DE DEUX POURCENTAGES

On peut utiliser dans la plupart des cas le test du Chi-deux de comparaisons d'échantillons. Toutefois, on peut aussi, dans le cas de grands échantillons, appliquer le test suivant appelé **test de différence de proportions**.

Dans deux échantillons de grandes tailles n_1 et n_2, on relève les pourcentages f_1 et f_2 d'individus présentant un certain caractère. Soit p_1 et p_2 les probabilités correspondantes ; on est donc amené à tester :

$$H_0 : \quad p_1 = p_2 = p$$

$$H_1 : \quad p_1 \neq p_2$$

Sous H_0, on a déjà vu que f_1 et f_2 sont des réalisations des variables F_1 et F_2 dont les lois sont :

$$L(F_1) = LG(p, \sqrt{\frac{p(1-p)}{n_1}}) \text{ et } L(F_2) = LG(p, \sqrt{\frac{p(1-p)}{n_2}})$$

$$L(F_1 - F_2) = LG[0, \sqrt{p(1-p)(\frac{1}{n_1} + \frac{1}{n_2})}]$$

On rejettera H_0 si pour un risque α donné, la variable centrée réduite associée dépasse en valeur absolue le seuil critique associé, fourni par la table de la loi normale.

Fréquemment p est inconnu ; on le remplace alors par son estimation : $\hat{p} = \dfrac{n_1 f_1 + n_2 f_2}{n_1 + n_2}$.

Exemple :

2500 chefs d'entreprise sont interrogés à six mois d'écart sur les perspectives d'exportation pour leurs entreprises. Les résultats sont donnés dans le tableau ci-dessous. Peut-on conclure que les perspectives d'exportation ont changé entre les deux sondages ?

	Sondage 1	Sondage 2
Augmentation	60%	58%
Diminution	40%	42%

On va faire un test de différence de proportions.

$f_1 = 0{,}6$, $f_2 = 0{,}58$, n = 2500 donc $\hat{p} = 0{,}59$

$$\frac{f_1 - f_2}{\sqrt{0{,}59 \times 0{,}41 \times \frac{2}{2500}}} = \frac{0{,}2 \times 50}{\sqrt{0{,}59 \times 0{,}41 \times 2}} \cong \frac{1}{0{,}7} \cong 1{,}44$$

Cette valeur doit être comparée à la valeur 1,64 correspondant à $\alpha = 0{,}05$ dans la table de la loi normale. On peut considérer qu'il n'y a pas eu de changement entre les deux sondages.

Appliquons le test du Chi-deux :

Pour cela reprenons le tableau des données en faisant apparaître les effectifs observés, les effectifs théoriques étant égaux à la demi-somme pour chaque ligne

	S_1	S_2	Somme
Augmentation	1500	1450	2950
Diminution	1000	1050	2050
Somme :	2500	2500	5000

Calculons la valeur de la variable de décision du test du Chi-deux :

$$d^2 = \frac{(1500-1475)^2}{1475} + \frac{(1450-1475)^2}{1475} + \frac{(1000-1025)^2}{1025} + \frac{(1050-1025)^2}{1025} \cong 2{,}07$$

Or la valeur du seuil critique donné par la loi du Chi-deux à un degré de liberté est : 3,84.

On garde H_0. Les perspectives ne sont pas différentes.

X- 5) TESTS DE MOYENNES D'ECHANTILLONS APPARIES

Supposons qu'un même échantillon d'individus soit soumis à deux mesures successives d'une même variable.

Exemple : double correction d'un paquet de copies, passage d'un même test d'aptitude à deux endroits différents...

On veut tester l'hypothèse que les deux séries de valeurs sont semblables. On appelle X_1 et X_2 les deux variables correspondant aux deux séries. En fait on se contente de tester l'hypothèse :

$$E(X_1) = E(X_2)$$

On suppose que X_1 et X_2 sont gaussiennes et que $L(X_1 - X_2) = LG(m_1 - m_2, \sigma)$.

Le test est le suivant :

$$H_0 : \quad m_1 = m_2$$

$$H_1 : \quad m_1 \neq m_2$$

On forme les différences : $d_i = x_{i1} - x_{i2}$ et on fait un test de Student sur la moyenne des d_i (σ inconnu) car :

$$L\left(\frac{\overline{d}}{s_d}\sqrt{n-1}\right) = T(n-1)$$

On rejette l'hypothèse H_0 si la variable vérifie $\left|\frac{\overline{d}}{s_d}\sqrt{n-1}\right| > k_0$.où k_0 a été lu sur la table de la variable de Student à n-1 degrés de liberté, α ayant été fixé.

X- 6) ANALYSE DE VARIANCE

X-6-1 Généralités

Le résultat d'une mesure est, en général, influencé par un certain nombre de facteurs. Ceux-ci peuvent être en grand nombre et d'ordre tout à fait différent.

Prenons l'exemple de la mesure de la limite de l'élasticité d'un acier AFNOR XC 18S. Les propriétés de l'acier résultent de toutes les opérations d'élaboration, de traitements thermiques, de mises en forme qu'il a subies. Elles vont donc dépendre : de la présence de métaux « calmant » l'acier, de traces de gaz occlus tel que l'azote, d'alignements de sulfures, de phosphures (fragilisants), de la structure, de la morphologie dues aux traitements thermomécaniques, de l'état d'écrouissage...

Les mesures seront influencées par : la forme de l'éprouvette, son usinage, son alignement sur la machine de traction, le protocole d'essai...

Il est difficile de les analyser un par un. On choisit un nombre réduit de facteurs, appelés **facteurs contrôlés**, semblant justifier une part importante de la dispersion des mesures. A un facteur sous contrôle, on associe plusieurs **modalités** ; dans le cas de l'éprouvette d'acier, ce sont, par exemple, les diverses teneurs en un composant. Différentes questions se posent :

- ✓ Le phénomène étudié est-il ou non influencé par le facteur contrôlé?
- ✓ Si le facteur est influent, quelle est alors la modalité la plus intéressante?

Les méthodes d'**Analyse de Variance** sont développées dans le but de répondre à ces questions. Cette analyse prend toute son importance dans l'étude de l'influence simultanée de plusieurs facteurs de variabilité, elle met aussi en évidence des interactions entre facteurs. Elle s'applique à tous les domaines où l'on fait des observations quantifiables.

Cette technique suppose que les résultats des mesures sont distribués selon une même loi de Laplace-Gauss quelle que soit la valeur du facteur étudié susceptible d'avoir une influence sur ces résultats. Cette condition mathématique, difficile à vérifier sur quelques nombres, peut généralement être admise sur la base de considérations physiques. L'analyse de la variance revient dans le cas simple à comparer plusieurs **moyennes d'échantillons gaussiens**. On se contentera, dans ce cours, d'exposer la méthode d'analyse de la variance à simple entrée.

X-6-2 Présentation

On isole un facteur A, celui vis à vis duquel on va chercher à prendre une décision,- on l'appellera **facteur contrôlé**-, et on rejette dans le paquet des facteurs non contrôlés tous les autres facteurs. On rangera parmi ceux-ci aussi bien ceux qu'on ne connaît pas que ceux qu'on connaît mais sur lesquels on ne se propose pas, à ce stade, de prendre une décision.

On note A_i le niveau i du facteur contrôlé. Soit k le nombre de niveaux du facteur contrôlé A. Pour le niveau i, il y a n_i résultats de mesure. Chacun de ces résultats de mesure est noté x_{ij}. Le premier indice i est donc celui qui identifie le niveau du facteur contrôlé, le deuxième indice j est le numéro du résultat de la $j^{ième}$ mesure pour ce niveau.

On remarquera qu'il n'est pas nécessaire que le nombre des résultats de mesure soit identique pour chaque niveau du facteur contrôlé, ce qui facilite l'utilisation de cette analyse. On suppose que le facteur A influe uniquement sur les moyennes des distributions et non sur leurs variances. Ces variances seront donc supposées égales. Il s'agit alors d'un test de confusion des k moyennes $x_{1.}, x_{2.}, ..., x_{n.}$.

X-6-3 Tableau du relevé des mesures

Niveaux du facteur contrôlé	Résultats des mesures	Moyennes par niveau	Nombres de mesures
A_1	$x_{11}\ x_{12}\ \dots\ x_{1n_1}$	$x_{1.}$	n_1
A_2	$x_{21}\ x_{22}\ \dots\ x_{2n_2}$	$x_{2.}$	n_2
…			
A_i	x_{ij}	$x_{i.}$	n_i
A_k	$x_{k1}\ x_{k2}\ \dots\ x_{kn_k}$	$x_{k.}$	n_k

En considérant chaque échantillon comme issu d'une variable aléatoire X_i suivant une loi de Laplace - Gauss LG (m_i, σ), le problème est de tester :

$$H_0 : \quad m_1 = m_2 = \dots = m_k$$

$$H_1 : \quad \exists (i, j) \ / \ m_i \neq m_j$$

X-6-4 Définition des variations

Chaque mesure x_{ij} présente un écart $x_{ij} - \bar{x}$ par rapport à la moyenne générale $\bar{x}$:

$$x_{ij} - \bar{x} = (x_{ij} - x_{i.}) + (x_{i.} - \bar{x})$$

La variation totale T **est la somme des carrés des écarts de toutes les mesures par rapport à la moyenne générale** $\bar{x}$ **:**

$$T = \sum_i \sum_j (x_{ij} - \bar{x})^2$$

On peut remarquer que $T = N S^2$ où $N = \sum_i n_i$

Il est alors naturel de distinguer, dans cette variation totale, la contribution du facteur contrôlé et celle des facteurs non contrôlés. Ces derniers contribuent, au sein de chaque niveau A_i du facteur contrôlé A, à une variation : $\sum_j (x_{ij} - x_{i.})^2$

Etendue à tous les niveaux du facteur contrôlé la variation résiduelle ou variation intraclasse vaut :

$$R = \sum_i \sum_j (x_{ij} - x_{i.})^2$$

Elle est constatée indépendamment de tout effet du facteur contrôlé A.

La **variance résiduelle, notée S_R^2, est le quotient** $\frac{R}{N}$.

La différence T-R rend compte de la contribution éventuelle du facteur contrôlé A puisqu'elle rend compte d'une variation qui se manifeste en plus de celle des facteurs non contrôlés.

La variation interclasse, somme des carrés des écarts des différentes moyennes par rapport à la moyenne générale :

$$A = T - R = \sum_i \sum_j (x_{i.} - \bar{x})^2 = \sum_i n_i (x_{i.} - \bar{x})^2$$

La variance due au facteur contrôlé est le quotient $S_A^2 = \frac{A}{N}$.

X-6-5 Interprétation

On fait les hypothèses suivantes :

- ✓ Le facteur contrôlé agit sur les moyennes et non sur les variances, ce qui doit être vérifié.
- ✓ Chaque variable X_i suit une loi normale LG (m_i, σ).

Posons $n_i S_i^2 = \sum_j (x_{ij} - x_{i.})^2$, la statistique $\frac{n_i S_i^2}{\sigma^2}$ suit une loi $\chi^2_{n_i - 1}$. La variance résiduelle S_R^2 vérifie :

$$L\left(\frac{NS_R^2}{\sigma^2} = \sum_i \frac{n_i S_i^2}{\sigma^2}\right) = \chi^2_{N-k}$$

On en déduit que $\frac{NS_R^2}{N-k}$ est une estimation de la variance σ^2 à (N - k) degrés de liberté.

Sous l'hypothèse $\mathbf{H_0}$, on en déduit :

$$\frac{NS^2}{\sigma^2} \text{ suit la loi } \chi^2_{N-1}$$

$$\frac{NS_A^2}{\sigma^2} \text{ suit la loi } \chi^2_{k-1}$$

les statistiques S_A^2 et S_R^2 sont indépendantes.

On en déduit alors que : $$L\left(\frac{S_A^2}{k-1} \times \frac{N-k}{S_R^2}\right) = F(k-1, N-k)$$

Au seuil α choisi, la valeur critique k_α est donnée par les tables de Fisher.

Le facteur contrôlé a une influence significative, c'est à dire que la population n'est pas homogène, dès que :

$$\frac{S_A^2}{k-1} \times \frac{N-k}{S_R^2} > k_\alpha$$

X-6-6 Pratique de l'analyse

X-6-6-1. Test d'égalité des variances

L'analyse de la variance nécessite que les variances des différents niveaux du facteur contrôlé soient égales. Pour le vérifier, nous utiliserons le test de **Fisher-Snedecor**.

Pour chaque niveau i du facteur contrôlé nous avons calculé la variance S_i^2. Appelons S_g^2 la variance la plus grande et S_p^2 la plus petite. Les populations étant gaussiennes, on sait que :

$$L\left(\frac{n_g S_g^2}{n_g - 1} \times \frac{\sigma_p^2}{\sigma_g^2} \times \frac{n_p - 1}{n_p S_p^2}\right) = F(n_g - 1, n_p - 1)$$

Sous l'hypothèse H_0 $\sigma_g = \sigma_p$, on est amené à comparer :

$$\frac{n_g S_g^2}{n_g - 1} \times \frac{n_p - 1}{n_p S_p^2} \qquad \text{à} \qquad F_\alpha(n_g - 1, n_p - 1)$$

Si les variances sont déclarées égales, nous pouvons faire l' analyse de variance.

X-6-6-2. Tableau d'analyse

On calcule rapidement les différentes statistiques par les formules simplifiées suivantes :

$$\boxed{\Delta = \frac{1}{N}\left(\sum_{i=1}^{k}\sum_{j=1}^{n_i} x_{ij}\right)^2} \qquad \boxed{T = NS^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i} x_{ij}^2 - \Delta}$$

$$\boxed{A = NS_A^2 = \sum_{i=1}^{k}\frac{1}{n_i}\left(\sum_{j=1}^{n_i} x_{ij}\right)^2 - \Delta}$$

$$\boxed{R = NS_R^2 = NS^2 - NS_A^2 = T - A}$$

Ces résultats, les degrés de liberté et les quotients correspondants sont résumés dans le tableau d'analyse de la variance.

X-6-6-3. Tableau d'analyse de la variance

Variations :	Sommes	Degrés de liberté	Quotients
due au facteur	$A = NS_A^2$	k-1	$v_A = \frac{NS_A^2}{k-1}$
résiduelle	$R = NS_R^2$	N-k	$v_R = \frac{NS_R^2}{N-k}$
totale	$T = NS^2$	N-1	

X-6-6-4. Conclusion

Le facteur contrôlé n'a pas d'influence c'est à dire qu'on garde l'hypothèse H_0 « la population est homogène » si :

$$\frac{\nu_A}{\nu_R} \leq F_\alpha(k-1, N-k)$$

On obtient alors :

- ✓ une **estimation de la moyenne m** à l'aide de la moyenne de toutes les observations,
- ✓ une **estimation de la variance** grâce au quotient ν_R.

X-6-6-5. Intervalle de confiance pour la moyenne

Si l'hypothèse H_0 est gardée, on construit un **intervalle de confiance pour m** en rappelant que la variable :

$$\frac{m-\bar{x}}{\sqrt{NS_R^2}}\sqrt{N-k}$$

suit une loi de **Student** à **(**N – k) degrés de liberté.

Ch. XI Régression

XI- 1) INTRODUCTION

Dans de multiples domaines de Sciences Appliquées, on observe l'existence d'une liaison entre deux ou plusieurs variables (dépenses et revenus d'une famille, surface, lieu et prix d'un appartement…). Nous sommes donc amenés, en Statistique, à rechercher sur des échantillons où plusieurs variables sont mesurées, une relation éventuelle entre certaines de ces variables.

Considérons un couple de variables aléatoires numériques (X, Y). Si X et Y ne sont pas indépendantes, la connaissance de la valeur prise par X change notre incertitude concernant la réalisation de Y.

Le théorème de la variance totale :

$$V(Y) = E[V(Y/X)] + V[E(Y/X)]$$

nous permet d'écrire que :

$$V(Y) \geq E[V(Y/X)]$$

c'est-à-dire que la variance de la distribution conditionnelle de Y sachant X est en moyenne inférieure à la variance de Y. Il es alors intéressant de mesurer l'intensité de la liaison entre les deux variables.

De plus quand on a l'intuition que la connaissance du phénomène de X sert à prédire celui représenté par Y, on cherche une relation de prévision de Y par X de la forme :

$$Y^* = f(X)$$

telle que l'erreur de prévision $\varepsilon = Y - Y^*$ soit :

- ✓ sans biais $E(Y - Y^*) = 0$
- ✓ de variance minimale.

Le modèle mathématique décrivant la relation entre X et Y peut être un modèle linéaire ou exponentiel…

X est appelée **variable explicative** et Y est la **variable expliquée**.

XI- 2) MODELE DE LA REGRESSION SIMPLE

XI-2-1 Rappel de l'approximation conditionnelle

On a vu dans le cours de Probabilités (Tome 1, chapitre VII) que, deux variables X et Y étant données, $E(Y/X)$ est la projection orthogonale de Y sur l'espace V_X des variables ne dépendant que de la variable X c'est à dire de la forme $\varphi(X)$.

V_X est un sous-espace de l'espace de Hilbert V des variables aléatoires définies sur un même domaine. V_X contient la droite des variables constantes.

$E(Y/X)$ réalise donc le minimum de $E[(Y - \varphi(X))^2]$ à φ variable.

C'est la meilleure approximation de Y par une fonction de X comme $E(Y)$ est la meilleure approximation de Y par une constante.

La qualité de l'approximation de Y par $E(Y/X)$ est mesurée par le rapport de corrélation :

$$\eta^2_{Y/X} = \frac{V[E(Y/X)]}{V(Y)} = \cos^2\theta$$

où θ est l'angle entre les variables X-E(X) et Y-E(Y).

La fonction qui, à toute valeur x de la variable X, associe $E(Y/X=x)$ s'appelle fonction de régression de Y en X, son graphe est la courbe de régression de Y en X.

Si on pose $Y^* = E(Y/X)$ et $\varepsilon = Y - Y^*$, on peut remarquer que :

- ✓ ε est orthogonal à V_X donc non corrélé à X et à Y* (Y* étant la projection de Y sur V_X)
- ✓ $E(\varepsilon) = E(Y) - E[E(Y/X)] = 0$ (théorème de l'espérance totale)
- ✓ $V(\varepsilon) = (1-\eta^2_{Y/X})V(Y)$ (théorème de la variance totale)

XI-2-2 Modèle de régression linéaire simple

Un cas très fréquent dans la pratique est celui où la relation entre Y et X est une relation linéaire telle que :

$$Y^* = Y - \varepsilon = E(Y/X) = \alpha + \beta X$$

La représentation graphique de Y* en fonction de X est une droite de pente β appelée droite de régression. Nous allons exprimer les paramètres α et β à l'aide de coefficients statistiques.

Prenons l'espérance des deux derniers membres. Par le théorème de l'espérance totale :

$$E[E(Y/X)] = E(Y) = \alpha + \beta\, E(X)$$

ce qui veut dire que :

la droite de régression passe par le point de coordonnées $(E(X), E(Y))$

On peut aussi déduire des deux relations précédentes :

$$Y - E(Y) = \beta(X - E(X)) + \varepsilon\,.$$

On multiplie cette égalité par $X - E(X)$ et on prend l'espérance des deux termes :

$$E[(X-E(X))(Y-E(Y))] = \beta \times E[(X-E(X))^2] + E(\varepsilon(X-E(X))]$$

$Cov(X,Y) = \beta\, V(X) + Cov(\varepsilon, X)$ puisque $E(\varepsilon) = 0$

Or ε est non corrélée avec X donc leur covariance est nulle, on en déduit la valeur de β :

$$\beta = \frac{Cov(X,Y)}{V(X)} = \rho\frac{\sigma_Y}{\sigma_X}$$

Finalement l'équation de la droite de régression est :

$$\boldsymbol{Y^* = E(Y) + \rho\frac{\sigma_Y}{\sigma_X}(X - E(X))}$$

XI- 3) AJUSTEMENT SUR DES DONNEES EXPERIMENTALES

XI-3-1 Présentation

Dans le modèle de régression linéaire simple, on dispose de n couples (x_i, y_i) constituant un n-échantillon d'observations indépendantes du couple de variables aléatoires (X, Y). Après étude du nuage de points, on suppose que la corrélation entre X et Y est significative c'est-à-dire que le modèle $E(Y/X) = \alpha + \beta X$ convient :

$$Y = \alpha + \beta X + \varepsilon$$

On cherche à estimer les paramètres α et β.

La méthode adoptée reste valable si on suppose que X n'est plus aléatoire mais imposée par la nécessité (différentes valeurs d'une intensité de courant, dates des mesures…). Il faut alors supposer que, pour chaque observation, on a $y_i = \alpha + \beta x_i + \varepsilon_i$ où les ε_i sont des réalisations indépendantes d'une variable ε d'espérance nulle et de variance constante. On parle alors de modèle linéaire et non plus de régression linéaire, le terme de corrélation étant réservé au cas où X est aléatoire.

On peut remarquer que d'autres modèles peuvent se ramener à celui-ci par transformations simples :

En économétrie, $Y = \alpha X^{\beta}$ devient linéaire en passant aux logarithmes

Le modèle exponentiel $Y = \alpha e^{\beta X}$ devient linéaire en posant Y'=LogY…

D'autres modèles ne peuvent pas se ramener à un modèle linéaire simple : le relation entre X et Y ne peut être linéarisée ou alors il y a deux variables explicatives…

La méthode couramment adoptée pour obtenir une droite qui s'ajuste le mieux possible sur le nuage de points est la **méthode des moindres carrés**, méthode consistant à rendre minimale la somme des carrés des écarts des mesures aux points correspondants sur la droite.

XI-3-2 Méthode des moindres carrés

XI-3-2-1. Estimation des paramètres

Au nuage de points (x_i, y_i), on cherche à ajuster une droite $y^* = a + bx$ de telle sorte que la somme des carrés des écarts des y_i aux $y_i^* = a + bx_i$ soit minimale.

Posons $e_i = y_i - y_i^* = y_i - (a + bx_i)$.

On choisit donc a et b rendant minimale la quantité :

$$S = \sum_{i=1}^{n} e_i^2 .$$

On annule les dérivées partielles de S par rapport à a et b respectivement :

$$\frac{\partial S}{\partial a} = \frac{\partial}{\partial a}\sum_{i=1}^{n}(y_i - a - bx_i)^2 = 2\sum_{i=1}^{n}(y_i - a - bx_i) = \sum_{i=1}^{n} e_i = 0$$

$$\frac{\partial S}{\partial b} = \frac{\partial}{\partial b}\sum_{i=1}^{n}(y_i - a - bx_i)^2 = 2\sum_{i=1}^{n} x_i(y_i - a - bx_i) = \sum_{i=1}^{n} x_i e_i = 0$$

ce qui donne les relations suivantes :

$$na + b\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$$

$$a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} {x_i}^2 = \sum_{i=1}^{n} x_i y_i$$

Le système à deux inconnues se résout très simplement :

$$b = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} {x_i}^2 - (\sum_{i=1}^{n} x_i)^2} = \frac{Cov(X,Y)}{s_X^2}$$

$$a = \frac{1}{n}(\sum_{i=1}^{n} y_i - b\sum_{i=1}^{n} x_i) = \overline{Y} - b\overline{X}$$

On obtient l'équation :

$$y^* = \overline{y} + \frac{Cov(X,Y)}{s_X^2}(x - \overline{x})$$

Soit, en exprimant la covariance en fonction du coefficient r de corrélation linéaire :

$$\boldsymbol{y^* = \overline{y} + r\frac{s_Y}{s_X}(x - \overline{x})}$$

XI-3-2-2. Décomposition de la variation totale

On peut écrire :

$$y_i - \overline{y} = y_i - y_i^* + y_i^* - \overline{y}$$

En élevant au carré et en sommant :

$$\sum_{i=1}^{n}(y_i - \overline{y})^2 = \sum_{i=1}^{n}(y_i - y_i^*)^2 + \sum_{i=1}^{n}(y_i^* - \overline{y})^2 + 2\sum_{i=1}^{n}(y_i - y_i^*)(y_i^* - \overline{y})$$

Or, $\sum_{i=1}^{n}(y_i - y_i^*)(y_i^* - \overline{y}) = b\sum_{i=1}^{n}(y_i - y_i^*)(x_i - \overline{x})$

$e_i = y_i - y_i^*$ $\qquad\qquad$ $\sum_{i=1}^{n}(y_i - y_i^*)(y_i^* - \overline{y}) = b[\sum_{i=1}^{n} x_i e_i - \overline{x}\sum_{i=1}^{n} e_i]$

Les deux sommes de l'expression précédente sont nulles car ce sont les deux dérivées partielles écrites ci-dessus. Finalement, on a la décomposition suivante :

$$\boldsymbol{\sum_{i=1}^{n}(y_i - \overline{y})^2 = \sum_{i=1}^{n}(y_i^* - \overline{y})^2 + \sum_{i=1}^{n}(y_i - y_i^*)^2}$$

variation totale = variation expliquée + variation résiduelle

XI-3-2-3. Contribution d'une observation à la droite des moindres carrés

On peut écrire la pente de la droite des moindres carrés de la façon suivante :

$$b = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \times \frac{y_i - \bar{y}}{x_i - \bar{x}}$$

Cette pente est la moyenne pondérée des pentes des droites reliant le centre de gravité $(\bar{x}, \bar{y})$ à chaque observation.

La pondération de la i^{ième} observation permet de mesurer la contribution de cette observation dans le calcul de la pente ; elle est égale à :

$$\frac{(x_i - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

On appelle levier de la i^{ième} observation, la quantité :

$$\boldsymbol{h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

La moyenne des leviers est $\bar{h} = \frac{2}{n}$.

XI-3-2-4. Propriétés des estimations

Nous avons les résultats suivants :

a et b sont des estimateurs sans biais de variance minimale de α et β.

y* est un estimateur sans biais de $E(Y / X = x) = \alpha + \beta x$

$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(y_i - y_i^*)^2}{n-2}$ est une estimation sans biais de σ^2.

On démontre que a et b sont des estimateurs sans biais de α et β en calculant les espérances conditionnelles $E(b / X = x_i)$ et $E(a / X = x_i)$ et en appliquant le théorème de l'espérance totale. De même, on calcule les variances conditionnelles et on montre que :

$$\sigma_B^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\sigma_A^2 = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)$$

XI-3-2-5. Ecarts résiduels

On a déjà montré que les écarts résiduels $e_i = y_i - y_i^*$ vérifient :

$$\sum_{i=1}^{n} e_i = 0$$

Les écarts ne sont donc pas des réalisations indépendantes.

La variance des écarts, notée $s^2_{Y/X}$, est égale à :

$$s^2_{Y/X} = \frac{1}{n}\sum_{i=1}^{n} e_i^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - y_i^*)^2$$

Or la formule de décomposition de la variation totale nous permet d'écrire que :

$$\sum_{i=1}^{n}(y_i - y_i^*)^2 = \sum_{i=1}^{n}(y_i - \bar{y})^2 - \sum_{i=1}^{n}(y_i^* - \bar{y})^2$$

$$s^2_{Y/X} = \frac{1}{n}\sum_{i=1}^{n}(y_i - y_i^*)^2 = s_Y^2 - b^2 s_X^2 = s_Y^2 - r^2 s_Y^2 = s_Y^2(1 - r^2)$$

On obtient alors le résultat suivant :

$$\boldsymbol{s^2_{Y/X} = s_Y^2(1 - r^2)}$$

XI-3-2-6. Cas du Résidu suivant une loi normale

Supposons que le résidu ε soit une variable normale de loi LG(0, σ) ; alors :

a, b et $\hat{\sigma}^2$ sont les estimateurs de variance minimale de α, β et σ^2

$\dfrac{(n-2)\hat{\sigma}^2}{\sigma^2} = \dfrac{n s^2_{Y/X}}{\sigma^2}$ est une réalisation d'une variable du Chi-deux à n-2 degrés de liberté.

Les paramètres a et b suivent des lois normales comme combinaisons linéaires de variables normales mais leurs écart-types dépendent de σ.

Le deuxième résultat a alors comme conséquence très importante de donner la possibilité d'écrire des intervalles de confiance pour α et β. On obtient en effet :

$(b-\beta)\dfrac{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}{s_{Y/X}}\sqrt{n-2}$ suit une loi de Student T_{n-2}.

$(a-\alpha)\dfrac{1}{\hat{\sigma}\sqrt{\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)}}\sqrt{n-2}$ suit une loi de Student T_{n-2}.

XI-3-3 Test du modèle linéaire : analyse de variance de la régression

On a déjà vu la décomposition de la variation totale :

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(y_i^* - \bar{y})^2 + \sum_{i=1}^{n}(y_i - y_i^*)^2$$

On suppose que ε suit une loi LG(0, σ), alors, d'après le résultat du paragraphe précédent :

$$L\left(\frac{\sum_{i=1}^{n}(y_i - y_i^*)^2}{\sigma^2}\right) = \chi^2_{n-2}$$

Posons les hypothèses du test d'analyse de variance :

H_0 : $\beta = 0$ (non régression)

H_1 : $\beta \neq 0$ (régression linéaire)

Sous l'hypothèse H_0 :

$$L\left(\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{\sigma^2}\right) = \chi^2_{n-1}$$

$$L\left(\frac{\sum_{i=1}^{n}(y_i^* - \bar{y})^2}{\sigma^2}\right) = \chi^2_1 .$$

Le théorème de Cochran s'applique et les deux variables :

$$\sum_{i=1}^{n}(y_i - y_i^*)^2 \text{ et } \sum_{i=1}^{n}(y_i^* - \bar{y})^2 \text{ sont indépendantes.}$$

On en déduit que **sous H_0** :

$$\boldsymbol{L\left(\frac{\sum_{i=1}^{n}(y_i^* - \bar{y})^2}{\sum_{i=1}^{n}(y_i - y_i^*)^2}(n-2)\right) = F(1, n-2)}.$$

Le test significatif de la régression découle de ce résultat :

On calculera, sur les données, la valeur de la variable définie dans l'égalité précédente et on comparera à la valeur de la loi de Fisher fournie par les tables, au risque a fixé.

XI-3-4 Test de linéarité

Ce test consiste à valider l'hypothèse d'un modèle linéaire.

Il paraît donc primordial et devrait être effectué avant toute autre démarche. Or il nécessite plusieurs observations de Y pour chaque valeur de X ce qui n'est pas souvent réalisé.

En effet tester l'hypothèse $E(Y/X) = \alpha + \beta X$ revient à montrer que la courbe des points d'abscisses x_i et d'ordonnées les moyennes conditionnelles $\overline{y_i}$ sachant $X = x_i$ est une droite.

On compare le coefficient de corrélation linéaire empirique r^2 avec le rapport de corrélation empirique e^2 défini par :

$$e^2 = \frac{\frac{1}{n}\sum_{i=1}^{n} n_i (\overline{y_i} - \bar{y})^2}{s_Y^2}$$

Dans ce test, on adopte pour hypothèse H_0, l'hypothèse de linéarité de la régression, c'est à dire $E(Y/X) = \alpha + \beta X$ ou encore $\eta^2_{Y/X} = \rho^2$; alors en notant k le nombre de valeurs distinctes de X :

$$L\left(\frac{e^2 - r^2 / k-1}{1 - e^2 / n-k}\right) = F(k-1, n-k)$$

On sera amené à rejeter l'hypothèse de linéarité pour un rapport trop grand.

XI-3-5 Prévision d'une valeur

A l'aide du modèle de régression linéaire, il est possible de prévoir une valeur y_0^* de la variable Y pour une valeur x_0 de X non observée et c'est une des applications de ce modèle. La valeur de y_0^* est donnée par :

$$y_0^* = \alpha + \beta\, x_0$$

Il est intéressant, en général, d'encadrer cette valeur grâce à un intervalle de prévision.

La loi de Y* est la loi normale :

$$LG\left(\alpha + \beta x, \sigma \sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_i (x_i - \bar{x})^2}}\right)$$

La loi de $Y / X = x_0$ est la loi normale :

$$LG(\alpha + \beta x_0, \sigma)$$

Les variables Y et Y* sont indépendantes, la première ne dépendant que de X et la deuxième dépendant des valeurs observées.

On peut en déduire que la variable Y - Y* suit la loi normale :

$$LG\left(0, \sigma \sqrt{1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_i (x_i - \bar{x})^2}}\right)$$

L'écart-type σ n'étant pas connu mais estimé par $\hat{\sigma}$, il en résulte que :

$$L\left(\frac{y-y^*}{\hat{\sigma}\sqrt{1+\frac{1}{n}+\frac{(x-\bar{x})^2}{\sum_i (x_i-\bar{x})^2}}}\right) = T_{n-2}$$

On en déduit l'intervalle de prévision pour y_0. On remarquera que cet intervalle sera d'autant plus grand que x_0 sera éloigné de $\bar{x}$.

Bibliographie

BAILLARGEON G.
Méthodes Statistiques de l'Ingénieur - Editions S.M.G. 1990

BENZECRI J.C. et coll.
L'Analyse des Données [Tomes 1 et 2] - Dunod 1976

BERGER J.O.
Statistical Decision Theory and Bayesian Analysis - Springer Verlag 1985

BERNARDO J.M. & SMITH A.F.M.
Bayesian Theory - John Wiley & Sons 1994

BUCKLEW J.A.
Large Deviation Techniques in Decision, Simulation, and Estimation - John Wiley & Sons 1990

COMETS F., EL KAROUI N., NEVEU J.
Probabilités - Département de Mathématiques appliquées - Ecole Polytechnique

DOOB J.L.
Stochastic Process – John Wiley & Sons 1990

FISHER FOURGEAUD & FUCHS A.
Statistique – Dunod 1967

GIRSCHIG R.
Analyse Statistique – Ecole Centrale de Paris

GUMBEL E.J.
Statistics of Extremes – Columbia University Press 1958

LEBART L., MORINEAU A. & FENELON J.P.
Traitement des Données Statistiques, méthodes et programmes – Dunod 1979

LEBART L., MORINEAU A. & TABARD N.
Techniques de la Description Statistique des Données – Dunod 1977

LAMOUREUX C.
Cours de Mathématiques – Ecole Centrale de Paris

METIVIER M. & NEVEU J.

Probabilités – Ecole Polytechnique 1979

PELLAUMAIL J.

Probabilités, Statistiques, Files d'Attente – Dunod 1986

PILZ J.

Bayesian Estimation and Experimental Design in Linear Regression Models – John Wiley & Sons 1989

PROCACCIA H. & PIEPSZOWNIK L.

Fiabilité des Equipements et Théorie de la Décision, Statistique Fréquentielle et Bayésienne – Eyrolles 1992

SAPORTA G.

Probabilités - Analyse des données et statistique- Editions Technip 1990

SAPORTA G.

Probabilités & Statistique – Ecole Centrale de Paris 1984

SAVILLE D.J. & WOOD G.R.

Statistical Methods: The Geometric Approach – Springer Verlag 1993

TASSI Ph. & LEGAIT S.

Théorie des Probabilités en vue des Applications Statistiques – Editions Technip 1990

VEYSSEYRE R.

Statistique et probabilités pour l'ingénieur – Dunod - L'Usine Nouvelle 2001