

Renée Veyseyre

L'USINE NOUVELLE

Aide-mémoire Statistique et probabilités pour l'ingénieur

2^e édition



DUNOD

Renée Veysseyre

Aide-mémoire

**Statistique
et probabilités
pour l'ingénieur**

2^e édition

L'USINE NOUVELLE

DUNOD

Le pictogramme qui figure ci-contre mérite une explication. Son objet est d'alerter le lecteur sur la menace que représente pour l'avenir de l'écrit, particulièrement dans le domaine de l'édition technique et universitaire, le développement massif du photocopillage.

Le Code de la propriété intellectuelle du 1^{er} juillet 1992 interdit en effet expressément la photocopie à usage collectif sans autorisation des ayants droit. Or, cette pratique s'est généralisée dans les établissements

d'enseignement supérieur, provoquant une baisse brutale des achats de livres et de revues, au point que la possibilité même pour

les auteurs de créer des œuvres nouvelles et de les faire éditer correctement est aujourd'hui menacée. Nous rappelons donc que toute reproduction, partielle ou totale, de la présente publication est interdite sans autorisation de l'auteur, de son éditeur ou du

Centre français d'exploitation du droit de copie (CFC, 20, rue des Grands-Augustins, 75006 Paris).



© Dunod, Paris, 2001, 2006
ISBN 2 10 049994 7

Le Code de la propriété intellectuelle n'autorisant, aux termes de l'article L. 122-5, 2° et 3° a), d'une part, que les « copies ou reproductions strictement réservées à l'usage privé du copiste et non destinées à une utilisation collective » et, d'autre part, que les analyses et les courtes citations dans un but d'exemple et d'illustration, « toute représentation ou reproduction intégrale ou partielle faite sans le consentement de l'auteur ou de ses ayants droit ou ayants cause est illicite » (art. L. 122-4).

Cette représentation ou reproduction, par quelque procédé que ce soit, constituerait donc une contrefaçon sanctionnée par les articles L. 335-2 et suivants du Code de la propriété intellectuelle.

TABLE DES MATIÈRES

Principales notations

XI

A

Statistique descriptive

1 • Représentation graphique et numérique des données	3
1.1 Généralités et principales définitions	3
1.2 Séries numériques à une dimension	7
1.3 Séries numériques à deux dimensions	26

B

Calcul des probabilités

2 • Le modèle probabiliste	33
2.1 Introduction	33
2.2 Les concepts probabilistes	35
2.3 Mesure de probabilité et espace probabilisé	40
2.4 Échantillons et sous-populations	41
3 • Probabilité conditionnelle. Indépendance	42
3.1 Définition	42
3.2 Principe des probabilités composées	44
3.3 Événements indépendants	44

3.4	Indépendance deux à deux et indépendance mutuelle	45
3.5	Théorème de Bayes	46
4	Variables aléatoires réelles	49
4.1	Généralités sur les variables aléatoires	49
4.2	Fonction de répartition	52
4.3	Densité de probabilité	54
4.4	Discontinuités d'une fonction de répartition et lois discrètes	56
4.5	Loi de probabilité d'une variable aléatoire Y fonction d'une variable aléatoire X	57
4.6	Indépendance de deux variables aléatoires	58
4.7	Moments d'une variable aléatoire	59
5	Lois de probabilité discrètes	67
5.1	Définition d'une variable discrète	67
5.2	Loi de Dirac	69
5.3	Loi uniforme	70
5.4	Loi binomiale ou loi des tirages avec remise	71
5.5	Loi multinomiale	77
5.6	Loi hypergéométrique ou loi du tirage exhaustif	80
5.7	Loi de Poisson	83
5.8	Lois limites	84
5.9	Résumé	87
6	Lois de probabilité continues	89
6.1	Généralités	89
6.2	Loi uniforme	90
6.3	Loi exponentielle	92
6.4	Loi gamma	95
6.5	Lois bêta de types I et II	97
6.6	Loi de Laplace-Gauss ou loi normale	100
6.7	Loi log-normale	109
7	Convolution. Fonctions caractéristiques. Convergences stochastiques	112
7.1	Convolution	112

7.2	Fonction caractéristique	116
7.3	Convergence des suites de variables aléatoires	120
7.4	Lois des grands nombres	124
7.5	Théorème central limite	125
8 •	Variables aléatoires simultanées	127
8.1	Étude d'un couple de variables aléatoires discrètes	127
8.2	Étude d'un couple de variables aléatoires continues	132
8.3	Extension à des vecteurs aléatoires	139
8.4	Application : loi normale multidimensionnelle	141
9 •	Processus aléatoires	146
9.1	Définitions	147
9.2	Processus équivalents	148
9.3	Moments	149
9.4	Continuités	149
9.5	Processus stationnaires	150
9.6	Exemples de processus aléatoires	153
9.7	Martingale	154
9.8	Mouvement brownien	156
9.9	Marche au hasard	157
9.10	Processus et chaînes de Markov	158
9.11	Processus ponctuels	166
9.12	Application aux phénomènes d'attente	170

C

Statistique inférentielle

10 •	Caractéristiques d'un échantillon.	
	Application aux échantillons gaussiens	179
10.1	Introduction	179
10.2	Définition d'un échantillon aléatoire	180
10.3	Caractéristiques d'un échantillon aléatoire	181

10.4	Distribution du chi-deux	185
10.5	Distribution de Fisher-Snedecor	188
10.6	Distribution de Student	190
10.7	Cas particulier des échantillons gaussiens	192
11	• Lois des valeurs extrêmes. Échantillons artificiels	195
11.1	Échantillons ordonnés et statistique d'ordre	195
11.2	Loi de la variable $X_{(k)}$, réalisation de rang k	198
11.3	Loi de la variable $X_{(n)}$, plus grande valeur observée	199
11.4	Loi de la variable $X_{(1)}$, plus petite valeur observée	202
11.5	Échantillons artificiels et simulation	203
12	• Théorie de l'estimation	210
12.1	Exposé du problème et exemples	210
12.2	Définition d'une statistique	212
12.3	Statistique exhaustive	213
12.4	Information de Fisher	218
13	• Estimation ponctuelle	220
13.1	Définition d'un estimateur	220
13.2	Principales qualités d'un estimateur	221
13.3	Estimateur sans biais de variance minimale	227
13.4	Précision intrinsèque d'un estimateur et inégalité de Cramer-Rao	228
13.5	Méthode du maximum de vraisemblance (MV)	229
13.6	Extension au cas de plusieurs paramètres	232
14	• Estimation par intervalle de confiance	235
14.1	Définition d'un intervalle de confiance	235
14.2	Exemples d'intervalles de confiance	238
14.3	Estimation et intervalle de confiance dans le cas d'une population d'effectif fini	253
15	• Les tests statistiques	255
15.1	Notions générales sur les tests statistiques	255
15.2	Différentes catégories de tests statistiques	263

15.3	Test entre deux hypothèses simples et méthode de Neyman et Pearson	264
15.4	Tests entre deux hypothèses composites	267
15.5	Principaux tests paramétriques	270
16	• Tests d'ajustement et de comparaison	277
16.1	Tests d'ajustement	277
16.2	Tests de comparaison d'échantillons	289
16.3	Analyse de la variance à simple entrée	299
17	• Tests d'indépendance	306
17.1	Variables quantitatives	306
17.2	Variables ordinales et corrélation des rangs	308
17.3	Concordance de p classements	313
17.4	Liaison entre une variable quantitative et une variable qualitative	314
17.5	Liaison entre deux variables qualitatives	316
18	• Fiabilité	321
18.1	Généralités et principales définitions	321
18.2	Définition mathématique de la fiabilité	322
18.3	Taux de défaillance	324
18.4	Fiabilité d'un matériel usagé	326
18.5	Fiabilité en cas de remplacement préventif	327
18.6	Espérance de vie	328
18.7	Exemples de lois de fiabilité	328
18.8	Fiabilité d'un système en fonction de celle de ses composants	332

D

Analyse des données

19	• Introduction à l'analyse des données	337
19.1	Échantillon d'une variable aléatoire	338
19.2	Échantillon d'un couple de variables aléatoires	343

19.3	Échantillon de p variables aléatoires	345
19.4	Présentation des principales méthodes	348
20	• Régression linéaire simple	352
20.1	Introduction	352
20.2	Mesures de liaison	353
20.3	Choix des variables	354
20.4	Modèle théorique de la régression simple	355
20.5	Ajustement du modèle de régression linéaire sur des données expérimentales	357
20.6	Étude de la régression linéaire (aspects descriptifs)	359
20.7	Étude de la régression linéaire (aspects inférentiels)	363
20.8	Étude d'une valeur prévisionnelle	371
20.9	Conclusions	375
21	• Régression multiple. Modèle linéaire général	376
21.1	Introduction	376
21.2	Régression entre variables aléatoires	377
21.3	Modèle linéaire général	382
21.4	Estimations des paramètres du modèle de régression ($\underline{Y}, \underline{X}\underline{\beta}, \sigma^2, I_n$)	385
21.5	Estimation du paramètre $\underline{\beta}$ du modèle linéaire	387
21.6	Tests dans le modèle linéaire	387
21.7	Intervalle de prévision	390
21.8	Corrélations	390
21.9	Fiabilité de la régression	393
22	• Analyse de la variance	410
22.1	Généralités et but de la théorie	410
22.2	Analyse de la variance à double entrée	411
22.3	Analyse de la variance orthogonale à entrées multiples	419
22.4	Analyse de la variance emboîtée	422
22.5	Carré latin	427

Annexes

Analyse combinatoire	433
Rappels mathématiques	436
Tables statistiques	442
Bibliographie	467
Index	471

PRINCIPALES NOTATIONS

$\mathbb{N}$	Ensemble des entiers positifs ou nuls (on dit aussi les entiers naturels).
$\mathbb{N}^*$	Ensemble des entiers strictement positifs (cet ensemble ne contient pas 0).
$\mathbb{Z}$	Ensemble des entiers de signes quelconques.
$\mathbb{Z}^*$	Ensemble $\mathbb{Z}$ sauf 0.
$\mathbb{R}$	Ensemble des entiers de signes quelconques.
$\mathbb{R}^+$	Ensemble des entiers positifs ou nuls.
$\mathbb{R}^*$	Ensemble des entiers non nuls.

Cardinal d'un ensemble fini (abréviation card) :

L'entier naturel qui indique le nombre de ses éléments.

Cardinal d'un ensemble infini : un nombre appelé *aleph*.

$1_{[a, b]}$ *fonction caractéristique* de l'ensemble $[a, b]$ égale à 1 pour les points de cet ensemble et à 0, sinon.

Notation de la fonction exponentielle :

e^a ou $\exp a$ (la deuxième notation est utilisée pour éviter d'écrire un exposant trop long).

Notation de la fonction logarithme :

ln désigne le logarithme népérien et log le logarithme à base 10 sauf dans le cas de la loi log-normale.

Factorielle $n! = n(n-1)(n-2)\dots 2 \times 1$.

Matrice transposée :

La matrice tA transposée de la matrice A est obtenue en permutant lignes et colonnes.

A

Statistique descriptive

1 • REPRÉSENTATION GRAPHIQUE ET NUMÉRIQUE DES DONNÉES

A

STATISTIQUE DESCRIPTIVE

1.1 Généralités et principales définitions

Ce premier chapitre donne les définitions et les propriétés des principales notions utiles pour comprendre et traiter un problème de statistique.

La statistique descriptive a pour but :

- de dégager les propriétés essentielles que l'on peut déduire d'une accumulation de données ;
- de donner une image concise et simplifiée de la réalité.

Le résultat d'une observation, d'une mesure, n'est pas égale à la valeur théorique calculée ou espérée par l'ingénieur ; la répétition d'une même mesure, réalisée dans des conditions qui semblent identiques, ne conduit pas toujours aux mêmes résultats. Ces fluctuations, dues à des causes nombreuses, connues ou inconnues, contrôlées ou non, créent des difficultés aux ingénieurs et aux scientifiques. Quel résultat doivent-ils prendre ? Quel degré de confiance peuvent-ils accorder à la décision prise ? Les réponses à une enquête varient d'un individu à un autre ; quelles conclusions valables peut-on tirer d'un sondage ? Les méthodes de la statistique descriptive apportent des réponses à ces problèmes.

Pour être soumis à un *traitement statistique*, un tableau de données doit comporter au moins une variable de nature aléatoire. Une définition simple du caractère aléatoire d'une variable est qu'elle peut prendre au hasard des valeurs différentes.

1.1.1 Population et individus

Ensemble statistique ou *population* : réunion des individus sur lesquels on étudie une ou plusieurs propriétés.

Unité statistique : chaque individu.

Une population doit être correctement définie afin que l'appartenance d'un individu à cette population soit reconnue sans ambiguïté.

Exemple 1.1

Une usine fabrique des tiges métalliques utilisées dans l'assemblage de certaines structures. Pour étudier la résistance à la traction de ces tiges, on mesure cette résistance pour un lot de 100 tiges.

Propriété étudiée : la résistance à la traction de tiges métalliques.

Population statistique : l'ensemble des 100 tiges ou des 100 mesures.

Unité statistique : chacune des tiges ou chacune des 100 mesures.

1.1.2 Caractères et variables statistiques

■ Caractères

On s'intéresse à certaines particularités ou *caractères des individus* d'une population statistique :

- un seul caractère étudié, série numérique à une dimension (paragraphe 1.2),
- deux caractères étudiés, série numérique à deux dimensions (paragraphe 1.3),
- plus de deux caractères, on doit utiliser les techniques de l'analyse multidimensionnelle (voir chapitres 19 et suivants).

Les caractères étudiés peuvent être :

- le poids, la taille, le niveau d'études, la catégorie socioprofessionnelle, le lieu d'habitation..., dans le secteur des sciences humaines,
- le poids, la masse, la composition..., dans le secteur des sciences techniques.

■ Modalités

Un caractère peut prendre différentes *modalités*. Ces modalités doivent être incompatibles et exhaustives afin que l'appartenance ou la non-appartenance

d'un individu à une modalité soit définie sans ambiguïté. Un caractère peut être :

- *quantitatif*, les modalités sont mesurables ou repérables,
- *qualitatif*, les modalités ne sont pas mesurables.

■ Variables statistiques ou aléatoires

Une *variable statistique* ou *aléatoire* est un caractère faisant l'objet d'une étude statistique. Elle peut donc être qualitative ou quantitative.

Une variable quantitative est appelée :

- *discrète* si elle prend un nombre fini de valeurs souvent entières,
- *continue* si elle prend toutes les valeurs d'un intervalle fini ou infini.

Remarque

En toute rigueur, une variable statistique ne peut jamais être continue, le degré de précision des mesures ou des appareils entraînant toujours des discontinuités dans les résultats.

Une variable statistique ou aléatoire est notée par une lettre majuscule X , Y , et les valeurs qu'elle prend par des lettres minuscules $x_1, x_2, \dots, y_1, y_2, \dots$

1.1.3 Échantillon

Échantillon : groupe restreint, ou sous-ensemble, issu de la population.

Échantillon aléatoire : les résultats recueillis sur ce sous-ensemble doivent pouvoir être étendus, c'est-à-dire inférés, à la population entière.

Pour définir un tel échantillon, une méthode consiste à prélever, au hasard, un sous-ensemble d'individus, en utilisant, par exemple, des tables de nombres au hasard (chapitre 11, paragraphe 11.5).

1.1.4 Fréquences absolues, relatives, cumulées

Dans le cas des variables discrètes, on appelle :

- *Fréquence absolue* n_i ou *effectif*, associée à une valeur x_i de la variable aléatoire X , le nombre d'apparitions de cette variable dans la population ou dans l'échantillon.

- *Fréquence relative*, associée à la valeur x_i de la variable aléatoire X , le nombre

$$f_i = \frac{n_i}{n}$$

où n_i est la fréquence absolue et n le nombre total de données.

- *Fréquence cumulée absolue*, associée à une valeur x_i de la variable, le nombre d'individus dont la mesure est inférieure ou égale à x_i .

$$N_i = \sum_{k=1}^i n_k$$

On définit la *fréquence cumulée relative* :

$$F_i = \sum_{k=1}^i f_k$$

Exemple 1.2 Défauts relevés sur une pièce de tissu

Un fabricant de tissu essaie une nouvelle machine ; il compte le nombre de défauts sur 75 échantillons de 10 mètres. Il a trouvé les résultats suivants :

Tableau 1.1 – Nombre de défauts sur une pièce de tissus.

Nombre k de défauts	0	1	2	3	4	5
Nombre n_k d'échantillons	38	15	11	6	3	2

Nombre d'individus : les 75 échantillons.

Fréquence absolue associée à la valeur k , le nombre n_k : par exemple, sur les 75 échantillons examinés, 11 présentent $k = 2$ défauts, donc si $k = 2$, $n_k = 11$.

Fréquence relative associée à la valeur k : le quotient n_k/n .

$11/75 = 0,146$ est la fréquence relative associée à la valeur $k = 2$.

Fréquence cumulée absolue associée à la valeur k : le nombre d'échantillons ayant au plus k défauts (k compris).

$38 + 15 + 11 = 64$ est la fréquence cumulée absolue associée à la valeur $k = 2$.

Fréquence cumulée relative associée à la valeur k , le nombre d'échantillons ayant au plus k défauts (k compris) divisé par n .

$64/75 = 0,853$ est la fréquence cumulée relative associée à la valeur $k = 2$.

Les fréquences relatives et les fréquences cumulées relatives peuvent être utilisées pour comparer deux ou plusieurs populations.

Dans le cas d'une distribution continue, les données sont en général regroupées en classes (paragraphe 1.2.1). Les fréquences absolues, relatives et cumulées sont définies par rapport aux classes et non par rapport aux valeurs de la variable.

1.2 Séries numériques à une dimension

1.2.1 Représentation graphique des données

En présence d'un ensemble de données associées à un seul caractère, on doit :

- ranger ces données par valeurs non décroissantes (ou non croissantes) et déterminer les fréquences absolues, relatives et cumulées,
- visualiser ces données à l'aide d'un diagramme en bâtons pour des variables discrètes ou d'un histogramme pour des variables continues.

■ Rangement des données par valeurs non décroissantes

□ Variables discrètes

Tableau 1.2 – Données discrètes.

Valeurs de la variable	Fréquences absolues	Fréquences relatives	Fréquences cumulées absolues	Fréquences cumulées relatives
x_i	n_i	f_i	N_i	$F_i = \sum_{k=1}^i f_k$

Exemple 1.3 Défauts relevés sur une pièce de tissu (suite)

On complète le tableau 1.1 en calculant les fréquences relatives f_i , toutes les fréquences absolues cumulées N_i et les fréquences relatives cumulées F_i .

Tableau 1.3 – Étude statistique du nombre de défauts sur une pièce de tissu.

Nombre de défauts	n_i	f_i	N_i	F_i
0	38	0,506	38	0,506
1	15	0,20	53 = 38 + 15	0,706
2	11	0,146	64 = 53 + 11	0,853
3	6	0,08	70 = 64 + 6	0,933
4	3	0,04	73 = 70 + 3	0,973
5	2	0,026	75 = 73 + 2	1

□ Variables continues

Les données sont regroupées en k classes.

Une classe est définie par ses extrémités e_{i-1} , e_i et son effectif n_i .

□ Effectif d'une classe ou fréquence absolue

Le nombre n_i de valeurs de la variable X telles que : $e_{i-1} \leq X < e_i$.

□ Amplitude d'une classe

La quantité $e_i - e_{i-1}$.

□ Fréquence cumulée relative

$$F_i = \sum_{k=1}^i f_k$$

avec $F_1 = f_1$. Elle donne la proportion des individus tels que $X < e_i$.

Tableau 1.4 – Données continues.

Classes	Effectifs	Fréquences absolues	Fréquences cumulées
$e_{i-1} \leq X < e_i$	n_i	f_i	N_i

Exemple 1.4 Essais de fiabilité de dispositifs électroniques

100 dispositifs identiques ont été soumis à un test de fiabilité ; on a noté la durée de vie, en heures, jusqu'à défaillance (fin de l'aptitude du dispositif à remplir la fonction requise).

Tableau 1.5 – Durée de vie de 100 dispositifs identiques.

Durée de vie (en heures)	Nombre n_i de dispositifs (fréquence absolue)	Fréquence relative f_i	Fréquence cumulée absolue	Fréquence cumulée relative F_i
$0 \leq X < 150$	30	0,30	30	0,30
$150 \leq X < 300$	15	0,15	45	0,45
$300 \leq X < 450$	12	0,12	57	0,57
$450 \leq X < 600$	10	0,10	67	0,67
$600 \leq X < 750$	8	0,08	75	0,75
$750 \leq X < 900$	8	0,08	83	0,83
$900 \leq X < 1\ 050$	8	0,08	91	0,91
$1\ 050 \leq X < 1\ 200$	6	0,06	97	0,97
$1\ 200 \leq X < 1\ 350$	3	0,03	100	1

La variable statistique « durée de vie des dispositifs » est une variable continue.

Les classes peuvent être d'égale amplitude ou non ; on choisit, soit le nombre de classes, soit l'amplitude des classes. Dans l'exemple 1.4, les classes sont d'égale amplitude (150 heures).

Le nombre de classes ne doit pas être trop petit, perte d'informations, ni trop grand, le regroupement en classes est alors inutile et de plus, certaines classes pourraient avoir des effectifs trop faibles.

En général, le nombre de classes est compris entre 5 et 20 ; il dépend du nombre n d'observations et de l'étalement des données. La formule de Sturges donne une valeur approximative du nombre k de classes :

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

d'où le nombre de classes selon les valeurs de n (tableau 1.6).

Tableau 1.6 – Effectif n de l'échantillon et nombre k de classes.

$n \leq 10$	$10 < n \leq 35$	$35 \leq n \leq 70$	$70 \leq n \leq 90$	$90 \leq n \leq 150$	$150 \leq n \leq 300$	$300 \leq n \leq 620$	$620 \leq n \leq 1\,300$
4	5	6	7	8	9	10	11

La première ligne donne l'effectif de l'échantillon étudié et la deuxième ligne, le nombre correspondant k de classes.

□ Amplitude des classes

Elle est égale à E/k où $E = x_{\max} - x_{\min}$ est l'étendue de la série des observations (si les classes sont d'égale amplitude).

Si au contraire, on commence par définir l'amplitude des classes, on ne doit pas choisir cette amplitude trop faible, le nombre de classes est alors trop élevé ni trop grande, le nombre de classes est alors trop petit par rapport à celui que donne la formule de Sturges.

Les valeurs d'une classe sont assimilées à la valeur centrale ou centre de la classe égale à :

$$\frac{e_{i-1} + e_i}{2}$$

Le regroupement en classes fait perdre aux individus leur caractère propre ainsi que les détails fins des distributions.

Exemple 1.5 Essais de fiabilité de dispositifs électroniques (suite)

30 dispositifs ont une durée de vie comprise entre 0 et 150 heures, on admet que ces 30 dispositifs ont tous une durée de vie égale à 75 heures.

De même, 10 dispositifs ont une durée de vie comprise entre 450 et 600 heures que l'on prend égale à 525 heures.

■ Le diagramme en feuilles

On décompose une donnée numérique en deux parties :

- la tige qui comprend le premier ou les deux premiers chiffres,
- la feuille qui comprend les autres chiffres.

On écrit les tiges les unes sous les autres et en regard de chaque tige, les feuilles correspondantes ; tiges et feuilles sont séparées par un trait vertical.

Exemple 1.6 Exemple de diagramme en feuilles

Le tableau 1.7 donne le poids en grammes de 25 éprouvettes.

Tableau 1.7 – Poids de 25 éprouvettes.

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

Comme tige, on choisit les deux premiers chiffres de chaque mesure, c'est-à-dire 25, 26, 27, 28 et 29. Les feuilles sont alors constituées du dernier chiffre de la mesure :

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						

Le diagramme indique que le poids moyen se situe entre 270 et 280 g et qu'il doit être voisin de 270 g.

■ Les différents modes de représentation graphique des données

Les représentations graphiques permettent d'avoir rapidement une vue d'ensemble d'un tableau de données.

□ Variables discrètes : diagramme en bâtons

En abscisses, on porte les différentes valeurs x_i prises par la variable X . Puis, on trace un bâton dont la longueur est proportionnelle à n_i ou à f_i ; dans le deuxième cas, on peut éventuellement comparer deux séries de données.

Exemple 1.7 Classement de 100 familles en fonction du nombre d'enfants

On a relevé le nombre d'enfants de 100 familles choisies au hasard. Le tableau 1.8 donne les principales caractéristiques de cette étude.

Tableau 1.8 – Statistique sur le nombre d'enfants de 100 familles.

x_i	0	1	2	3	4	5	6	7	Total
n_i	20	25	30	10	5	5	3	2	100
f_i	0,20	0,25	0,30	0,10	0,05	0,05	0,03	0,02	1
F_i	0,20	0,45	0,75	0,85	0,90	0,95	0,98	1	

x_i nombre d'enfants compris entre 0 et 7.

n_i nombre de familles ayant x_i enfants.

f_i fréquence relative des familles ayant x_i enfants.

F_i fréquence cumulée des familles ayant au plus x_i enfants.

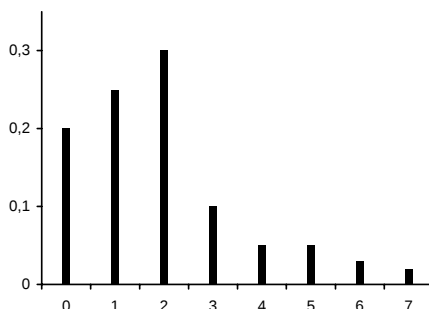


Figure 1.1 – Diagramme en bâtons de la distribution de l'exemple 1.7.

A

STATISTIQUE DESCRIPTIVE

□ Variables continues ou réparties en classes

Histogramme et propriétés

Un histogramme est constitué de rectangles juxtaposés dont la base correspond à l'amplitude de chaque classe et dont la surface est proportionnelle à la fréquence absolue ou relative de cette classe.

L'histogramme est un outil statistique facile à utiliser, donnant rapidement une image du comportement d'un procédé industriel et l'allure globale de la

distribution ; il montre l'étalement des données et apporte ainsi des renseignements sur la dispersion et sur les valeurs extrêmes ; il permet de déceler, éventuellement, des valeurs aberrantes.

Polygone de fréquences

Il permet de représenter sous forme de courbe, la distribution des fréquences absolues ou relatives. Il est obtenu en joignant, par des segments de droite, les milieux des côtés supérieurs de chaque rectangle de l'histogramme. Pour fermer ce polygone, on ajoute à chaque extrémité une classe de fréquence nulle.

Exemple 1.8 Étude de la dispersion d'un lot de 400 résistances

On a contrôlé 400 résistances dont la valeur nominale est égale à 100 k Ω et on a regroupé les résultats en classes d'amplitude 2 k Ω qui représente environ le dixième de la dispersion totale de l'échantillon contrôlé.

Tableau 1.9 – Étude statistique des mesures de la résistance d'un lot de 400 pièces.

Classe	Limites des classes	n_i	N_i	f_i	F_i
I	[92, 94[	10	10	0,025	0,025
II	[94, 96[	15	25	0,0375	0,0625
III	[96, 98[	40	65	0,10	0,1625
IV	[98, 100[	60	125	0,15	0,3125
V	[100, 102[	90	215	0,225	0,5375
VI	[102, 104[	70	285	0,175	0,7125
VII	[104, 106[	50	335	0,125	0,8375
VIII	[106, 108[	35	370	0,0875	0,925
IX	[108, 110[	20	390	0,05	0,975
X	[110, 112[	10	400	0,025	1

Les classes étant toutes de même amplitude, l'histogramme est facile à tracer ; il suffit de construire des rectangles dont l'aire est proportionnelle à la fréquence des résistances de la classe correspondante.

Courbes de fréquences cumulées

Courbe cumulative croissante : on joint les points ayant pour abscisses la limite supérieure des classes et pour ordonnées les fréquences cumulées croissantes

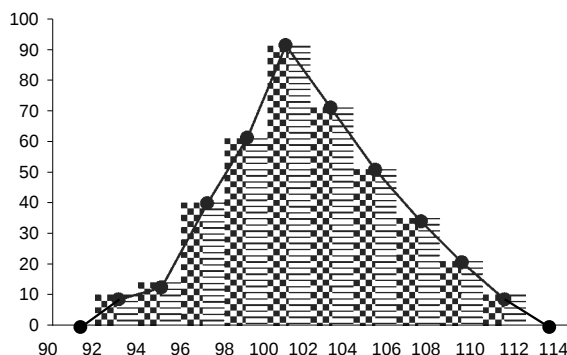


Figure 1.2 – Histogramme de la distribution de l'exemple 1.8 et polygone de fréquence.

correspondant à la classe considérée (pour le premier point, on porte la valeur 0). Elle donne le nombre d'observations inférieures à une valeur quelconque de la série.

Courbe cumulative décroissante : la construction de cette courbe est analogue à la précédente. Les points ont pour abscisses, les limites inférieures des classes et pour ordonnées, les fréquences cumulées décroissantes (pour le dernier point, la valeur est 0). Elle donne le nombre d'observations supérieures à une valeur quelconque de la série.

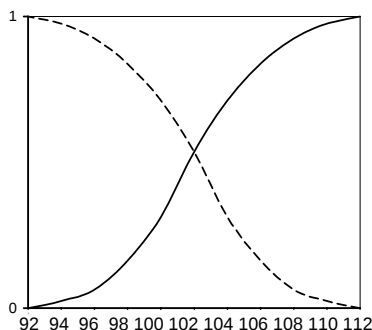


Figure 1.3 – Courbe cumulative croissante (trait plein) et courbe cumulative décroissante (trait pointillé) de la distribution de l'exemple 1.8.

A

STATISTIQUE DESCRIPTIVE

Autres modes de représentations graphiques

On définit des *diagrammes à secteurs circulaires* et des *diagrammes à rectangles horizontaux*.

Le diagramme à secteurs circulaires consiste en un cercle découpé en secteurs circulaires ; l'aire de chaque secteur, représentant la proportion des différentes composantes d'un tout, est proportionnelle aux fréquences, relatives ou absolues.

Le diagramme à rectangles horizontaux est défini de façon analogue.

Un autre mode de représentation est *la boîte à moustaches* ou *box-plot* (voir paragraphe 1.2.2, Quantiles).

1.2.2 Représentation numérique des données

Une série de données peut être résumée par quelques valeurs numériques appelées *caractéristiques des séries statistiques*, classées en quatre grandes catégories :

- les caractéristiques de tendance centrale,
- les caractéristiques de dispersion,
- les caractéristiques de forme,
- les caractéristiques de concentration.

■ Caractéristiques de tendance centrale

Elles donnent une idée de l'ordre de grandeur des valeurs constituant la série ainsi que la position où semblent se concentrer les valeurs de cette série. Les principales caractéristiques de tendance centrale sont la *moyenne arithmétique*, la *médiane*, la *médiale*, le *mode* et les *quantiles*.

□ Moyenne arithmétique

Définition et calcul

Pour calculer la moyenne arithmétique, deux cas sont à distinguer selon la façon dont les données ont été recueillies.

Cas 1 : n données non réparties en classes :

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

Cas 2 : n données réparties en k classes, la classe i étant d'effectif absolu n_i et d'effectif relatif f_i :

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

Changement d'origine et changement d'échelle

On pose pour toutes les données, $y_i = ax_i + b$, a et b étant des constantes ; on obtient :

$$\bar{y} = a\bar{x} + b$$

Propriétés

- La moyenne arithmétique permet de résumer par un seul nombre la série statistique.
- Elle prend en compte toutes les valeurs de la série et elle est facile à calculer.
- Elle est sensible aux valeurs extrêmes, il est parfois nécessaire de supprimer des valeurs extrêmes ou « aberrantes ».

La quantité $e_i = x_i - \bar{x}$ est l'écart de la valeur x_i à la moyenne arithmétique.

La moyenne arithmétique des écarts e_i est nulle.

□ Médiane M_e

Définition et calcul

La médiane est plutôt une moyenne de position.

La médiane est la valeur, observée ou possible, dans la série des données classées par ordre croissant (ou décroissant) qui partage cette série en deux parties comprenant exactement le même nombre de données de part et d'autre de M_e .

Comme pour la moyenne arithmétique, on distingue deux cas.

Cas 1 : n données non réparties en classes :

- pour une série ayant un nombre impair de données, la médiane est une valeur observée de la série ;
- pour une série ayant un nombre pair de données, on peut prendre pour valeur médiane, indifféremment l'une ou l'autre des valeurs centrales ou n'importe quelle valeur intermédiaire entre ces deux valeurs, par exemple, la moyenne arithmétique de ces deux valeurs, mais, dans ces conditions, ce n'est pas une valeur observée.

Cas 2 : n données réparties en k classes. La médiane est obtenue :

- soit par interpolation linéaire à l'intérieur de la classe centrale, si le nombre de classes est impair,
- soit en prenant la moyenne des deux classes « centrales », si le nombre de classes est pair.

Pour faire ce calcul, on suppose implicitement que *la distribution est uniforme* à l'intérieur de chaque classe.

Propriétés

- Le calcul de la médiane est rapide.
- La médiane n'est pas influencée par les valeurs extrêmes ou aberrantes.
- La médiane est influencée par le nombre des données mais non par leurs valeurs, elle ne peut donc pas être utilisée en théorie de l'estimation.
- Si la variable statistique est discrète, la médiane peut ne pas exister ; elle correspond seulement à une valeur possible de cette variable.
- La médiane est le point d'intersection des courbes cumulatives croissante et décroissante.
- La médiane ne se prête pas aux combinaisons algébriques ; la médiane d'une série globale ne peut pas être déduite des médianes des séries composantes.

Exemple 1.9 Dispersion d'un lot de 400 résistances (suite)

Calcul de la moyenne arithmétique :

$$\bar{x} = \frac{1}{400} (93 \times 10 + 95 \times 15 + 97 \times 40 + \dots + 111 \times 10) = 101,90$$

La moyenne arithmétique est égale à 101,90 k Ω .

Médiane : la série des observations comporte un nombre pair de classes. On peut définir une classe médiane comme la moyenne des classes V et VI, c'est-à-dire la classe fictive [101, 103[donc une résistance égale à 102 k Ω .

Un calcul plus précis consiste à chercher la valeur de la résistance de l'individu occupant le rang 200 (ou 200,5 !). Ne connaissant pas la distribution à l'intérieur des classes, on fait une interpolation linéaire. Le tableau de l'exemple 1.8 montre que cet individu appartient à la classe V.

125 résistances ont une valeur nominale inférieure à 100 k Ω et 215 résistances ont une valeur nominale inférieure à 102 k Ω d'où le calcul de la médiane :

$$100 + \frac{2 \times (200 - 125)}{(215 - 125)} = 101,66$$

La médiane est égale à 101,66 k Ω . Donc, 200 résistances ont une valeur nominale inférieure ou égale à 101,66 k Ω et 200 résistances ont une valeur nominale supérieure à 101,66 k Ω .

Le point d'intersection des deux courbes cumulatives a pour abscisse la médiane.

Exemple 1.10 Étude de deux séries d'observations

On considère les séries d'observations suivantes.

Série I : 5 observations classées par ordre croissant, 2, 5, 8, 11, 14

Moyenne arithmétique 8, médiane 8

Série II : 6 observations classées par ordre croissant, 6, 6, 14, 16, 18, 18

Moyenne arithmétique 13, médiane 15

Série III : les deux séries précédentes réunies, 2, 5, 6, 6, 8, 11, 14, 14, 16, 18, 18

Moyenne arithmétique 10,72, médiane 11

□ Mode ou valeur dominante M_0

Le mode est une moyenne de fréquence.

Définition

Le mode est la valeur de la variable statistique la plus fréquente que l'on observe dans une série d'observations.

Si la variable est une variable discrète, le mode s'obtient facilement. Si la variable est une variable continue, on définit une classe modale.

Propriétés

- Le mode n'existe pas toujours et quand il existe, il n'est pas toujours unique.
- Si après regroupement des données en classes, on trouve deux ou plusieurs modes différents, on doit considérer que l'on est en présence de deux ou plusieurs populations distinctes ayant chacune leurs caractéristiques propres ; dans ce cas, la moyenne arithmétique n'est pas une caractéristique de tendance centrale.

Exemple 1.11 Dispersion d'un lot de 400 résistances (suite)

On ne peut pas définir une valeur modale en ne connaissant pas la distribution à l'intérieur de chaque classe.

On définit une classe modale, c'est la classe V.

Exemple 1.12 Suite de l'exemple 1.10

Série I : pas de mode.

Série II : deux modes 6 et 18.

Série III : les deux séries réunies, trois modes 6, 14 et 18.

□ Médiale

La médiale est la valeur centrale qui partage en deux parties égales la masse de la variable.

Par exemple, la médiale partage un ensemble d'employés d'une entreprise en deux groupes tels que la somme totale des salaires perçus par le premier groupe soit égale à la somme totale des salaires perçus par le second groupe.

On vérifie facilement que :

$$\text{médiale} \geq \text{médiane}$$

Remarque

Pour définir n'importe quelle caractéristique (excepté la moyenne arithmétique), il faut que les données soient classées en ordre croissant (ou décroissant). Pour le calcul de la médiane, on peut trouver un résultat différent selon que les données sont classées par ordre croissant ou décroissant.

□ Quantiles

Cette notion est très utilisée dans les sciences humaines.

Définition

Les quantiles sont des caractéristiques de position partageant la série statistique ordonnée en k parties égales.

Pour $k = 4$, les quantiles, appelés *quartiles*, sont trois nombres Q_1 , Q_2 , Q_3 tels que :

- 25 % des valeurs prises par la série sont inférieures à Q_1 ,
- 25 % des valeurs prises par la série sont supérieures à Q_3 ,
- Q_2 est la médiane M_e ,
- $Q_3 - Q_1$ est l'intervalle interquartile, il contient 50 % des valeurs de la série.

Pour $k = 10$, les quantiles sont appelés *déciles*, il y a neuf déciles $D_1, D_2, \dots$ 10 % des valeurs de la série sont inférieures à D_1 ...

Pour $k = 100$, les quantiles sont appelés *centiles*, il y a 99 centiles, chacun correspondant à 1 % de la population.

Application

Le diagramme en *boîte à moustaches* ou *box-plot* (Tukey) permet de représenter schématiquement les principales caractéristiques d'une distribution en utilisant les quantiles.

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 rectangle est complétée par des moustaches correspondant aux valeurs suivantes :

- valeur supérieure : $Q_3 + 1,5(Q_3 - Q_1)$
- valeur inférieure : $Q_1 - 1,5(Q_3 - Q_1)$

Les valeurs extérieures « aux moustaches » sont représentées par des étoiles et peuvent être considérées comme aberrantes.

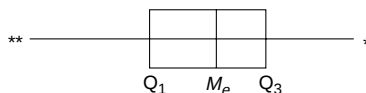


Figure 1.4 – Exemple de boîte à moustaches (les astérisques * représentent les valeurs aberrantes de la distribution).

■ Caractéristiques de dispersion

Ces caractéristiques quantifient les fluctuations des valeurs observées autour de la valeur centrale et permettent d'apprécier l'étalement de la série. Les principales sont : l'*écart-type* ou son carré appelé *variance*, le *coefficient de variation* et l'*étendue*.

□ Variance et écart-type

Définition et calcul

La *variance* d'un échantillon, notée s^2 , est appelée aussi *écart quadratique moyen* ou *variance empirique*. La racine carrée de la variance est appelée *écart-type*.

C'est la moyenne de la somme des carrés des écarts par rapport à la moyenne arithmétique.

La moyenne arithmétique $\bar{x}$ et l'écart-type s s'expriment avec la même unité que les valeurs observées x_i .

Cas 1 : n données non réparties en classes :

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

Formule simplifiée ne faisant apparaître que les données (facile à démontrer) :

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

La variance est donc égale à la moyenne des carrés moins le carré de la moyenne.

Cas 2 : n données réparties en k classes, la classe i étant d'effectif absolu n_i .

Dans ces conditions, on obtient :

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

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

Changement d'origine et d'échelle

On pose, pour toutes les données, $Y = aX + b$, a et b étant des constantes, on obtient :

$$s_{aX+b}^2 = a^2 s_X^2$$

Un changement d'origine n'a donc aucune influence sur le résultat.

Propriétés

- L'écart-type s caractérise la dispersion d'une série de valeurs. Plus s est petit, plus les données sont regroupées autour de la moyenne arithmétique $\bar{x}$ et plus la population est homogène ; cependant avant de conclure, il faut faire attention à l'ordre de grandeur des données.

- L'écart-type permet de trouver le pourcentage de la population appartenant à un intervalle centré sur l'espérance mathématique.
- La variance tient compte de toutes les données, c'est la *meilleure caractéristique de dispersion* (nombreuses applications en statistique).

Exemple 1.13 Séries d'observations de l'exemple 1.10**Série I**

$$\text{Variance : } s^2 = \frac{1}{5} (2^2 + 5^2 + 8^2 + 11^2 + 14^2) - (8)^2 = 18$$

$$\text{Écart-type : } s = 4,24$$

Série II

$$\text{Variance : } s^2 = 26,33$$

$$\text{Écart-type : } s = 5,13$$

Série III (les deux séries réunies)

$$\text{Variance : } s^2 = 28,74$$

$$\text{Écart-type : } s = 5,36$$

□ Coefficient de variation*Définition*

Il s'exprime, sous la forme d'un pourcentage, par l'expression suivante :

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

Propriétés

- Le coefficient de variation ne dépend pas des unités choisies.
- Il permet d'apprécier la représentativité de la moyenne arithmétique $\bar{x}$ par rapport à l'ensemble des données.
- Il permet d'apprécier l'homogénéité de la distribution, une valeur du coefficient de variation inférieure à 15 % traduit une bonne homogénéité de la distribution.
- Il permet de comparer deux distributions, même si les données ne sont pas exprimées avec la même unité ou si les moyennes arithmétiques des deux séries sont très différentes.
- Quelques exemples de coefficient de variation : le coefficient de variation du régime nival est voisin de 0,1 ; celui d'un cours d'eau régulier de 0,3 mais il peut atteindre 0,5 et même 1 pour un cours d'eau irrégulier.

□ Étendue

Définition

L'étendue est la quantité :

$$E = x_{\max} - x_{\min}$$

Propriétés

- L'étendue est facile à calculer.
- Elle ne tient compte que des valeurs extrêmes de la série ; elle ne dépend ni du nombre, ni des valeurs intermédiaires ; elle est très peu utilisée dès que le nombre de données dépasse 10.
- Elle est utilisée en contrôle industriel où le nombre de pièces prélevées dépasse rarement 4 ou 5 ; elle donne une idée appréciable de la dispersion. Cependant, dès que cela est possible, on préfère prélever 15 à 20 unités et utiliser l'écart-type pour apprécier la dispersion.

■ Caractéristiques de forme

□ Distribution symétrique

Une distribution est symétrique si les valeurs de la variable statistique sont également distribuées de part et d'autre d'une valeur centrale. Pour une distribution symétrique :

$$\text{mode} = \text{médiane} = \text{moyenne arithmétique}$$

□ Coefficient d'asymétrie ou de dissymétrie ou *skewness*

$$\gamma_1 = \frac{\mu_3}{s^3} \quad \text{où} \quad \mu_3 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^3$$

□ Coefficient d'aplatissement ou *kurtosis*

$$\gamma_2 = \frac{\mu_4}{s^4} \quad \text{où} \quad \mu_4 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^4$$

Ces deux coefficients sont principalement utilisés pour vérifier qu'une distribution est proche d'une distribution normale (loi de Laplace-Gauss) ; en effet, pour une telle loi, le coefficient d'aplatissement est égal à 3 et le coefficient

d'asymétrie à 0 (chapitre 6, paragraphe 6.6.2). Selon la valeur de ces coefficients, on peut donner quelques caractéristiques sur la forme de la distribution :

- si $\gamma_1 > 0$, la distribution est étalée vers la droite,
- si $\gamma_1 < 0$, la distribution est étalée vers la gauche,
- si $\gamma_1 = 0$, on ne peut pas conclure que la distribution est symétrique mais la réciproque est vraie,
- si $\gamma_2 > 3$, la distribution est moins aplatie qu'une distribution gaussienne,
- si $\gamma_2 < 3$, la distribution est plus aplatie qu'une distribution gaussienne.

■ Caractéristiques de concentration

Ces caractéristiques sont utilisées pour une grandeur positive cumulative telle que le revenu, la consommation...

□ Courbe de concentration

Soit une distribution de consommation X de masse totale M . À chaque valeur x_i de la variable X , on associe le point qui a :

- pour abscisse $F(x_i)$ = Proportion des individus consommant moins de x_i
- pour ordonnée $G(x_i) = \frac{\text{Masse des consommations} < x_i}{\text{Masse totale}}$

Pour une distribution non uniforme, cette courbe est toujours en dessous de la première bissectrice ; en effet, $F(x_i)$ est la proportion des individus consommant moins de x_i ; ils ne peuvent pas globalement consommer autant que les 100 $F(x_i)$ % suivants donc $G(x_i) < F(x_i)$.

La courbe de concentration traduit le pourcentage des individus consommant moins de x_i à la contribution de ces individus à la moyenne $\bar{x}$ de la masse totale.

Indice de concentration ou indice de Gini¹

Plus la distribution de X est inégalement répartie, plus la courbe de concentration s'éloigne de la première bissectrice, la première bissectrice traduisant l'équipartition.

1. Économiste italien né en 1884.

Un indice proposé par Gini est le suivant (figure 1.5) :

$$G = \text{aire ODBC} - \text{aire ODBA}$$

L'indice de Gini est égal au double de l'aire comprise entre la courbe de concentration et la première bissectrice.

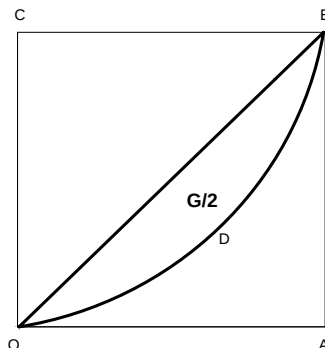


Figure 1.5 – Courbe de concentration et indice de Gini.

Cet indice est donné par l'intégrale double où f est la densité de la loi de la variable X et m son espérance mathématique :

$$G = \frac{1}{2m} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} |x - y| f(x) f(y) dx dy$$

Pour un échantillon de taille n , on obtient :

$$G = \frac{1}{n(n-1)\bar{x}} \sum_{i=1}^n \sum_{j=i+1}^n |x_i - x_j|$$

Exemple 1.14 Étude de la répartition du revenu annuel

On considère la répartition par tranches de revenus annuels des contribuables soumis à l'impôt sur le revenu (rôles émis en 1966 au titre des revenus de 1965).

Le fait que la dernière classe soit ouverte n'entraîne pas d'inconvénients pour les représentations graphiques et les calculs car l'effectif de cette classe représente environ 1 % de l'effectif total (0,009 9).

Tableau 1.10 – Répartition du revenu annuel.

Classes de revenus (en francs)	Nombre de contribuables (en milliers)
$0 \leq R < 50\,000$	549,3
$50\,000 \leq R < 100\,000$	3 087,4
$100\,000 \leq R < 150\,000$	2 229,0
$150\,000 \leq R < 200\,000$	1 056,7
$200\,000 \leq R < 350\,000$	925,0
$350\,000 \leq R < 500\,000$	211,0
$500\,000 \leq R < 700\,000$	90,8
700 000 et plus	81,6
Total	8 230,8

Pour calculer la moyenne arithmétique, on donne une valeur moyenne à cette dernière classe, 775 000 F par exemple.

– La moyenne arithmétique est alors égale à 142 225 F, l'écart-type à 114 640 F. Le coefficient de variation est égal à 0,80.

– La médiane est égale à 110 742 F, elle est représentée par le contribuable qui a pour numéro $n^{\circ} 4\,115,4 \times 1\,000$, le nombre d'observations présentant une valeur inférieure à la médiane est égal au nombre d'observations présentant une valeur supérieure à la médiane.

– Le mode est approximativement égal à 62 500 F.

La distribution est étalée vers la droite :

$$\text{mode} < \text{médiane} < \text{moyenne arithmétique}$$

– Le premier quartile est représenté par le contribuable $n^{\circ} 2057,73 \times 1\,000$ qui a pour revenu 74 433,50 F.

– Le troisième quartile est représenté par le contribuable $n^{\circ} 6173,1 \times 1\,000$ qui a pour revenu 164 536,24 F.

– Pour définir la courbe de concentration, on a divisé, afin de simplifier les calculs, les revenus par 25 000.

Abscisses F_i : fréquences cumulées croissantes

Ordonnées G_i : (masse des revenus des contribuables $\leq x$)/masse totale des revenus)

La masse totale des revenus est égale à 46 824,20 F (ou $46\,824,2 \times 25\,000$).

Tableau 1.11 – Résultats numériques du tableau 1.10.

Classe	Centre	Effectif	Abscisse F_i	Contribution de chaque classe	Ordonnée G_i
[0, 2[	1	549,3	0,0667	549,3	0,00117
[2, 4[	3	3 087,4	0,4418	9 262,2	0,2095
[4, 6[	5	2 229,0	0,7127	11 145	0,4475
[6, 8[	7	1 056,7	0,8410	7 396,9	0,605
[8, 14[	11	925,0	0,9534	10 175	0,823
[14, 20[	17	211,0	0,9791	3 587	0,899
[20, 28[	24	90,8	0,9901	2 179,2	0,946
[28, [	31	81,6	1	2 529,6	1

1.3 Séries numériques à deux dimensions

Soient X et Y les deux caractères étudiés, p le nombre de modalités prises par X , q le nombre de modalités prises par Y et n le nombre total d'observations. On étudie, par exemple, le poids et la taille d'un nombre n d'individus, le temps de travail sans pause et le nombre de pièces assemblées ou le nombre d'accidents survenus pendant cette période.

1.3.1 Représentation graphique des données

■ Tableaux statistiques

On suppose que les deux variables étudiées sont des variables discrètes et que les caractères sont des caractères quantitatifs. Les tableaux statistiques portent le nom de *tableaux croisés* ou *tableaux de contingence*.

Dans chaque case du tableau, on écrit l'effectif n_{ij} de l'échantillon, c'est-à-dire le nombre de données tel que $X = x_i$ et $Y = y_j$.

On définit les fréquences absolues suivantes :

– Les *fréquences marginales* :

$$n_{i.} = \sum_{j=1}^q n_{ij} \quad \text{et} \quad n_{.j} = \sum_{i=1}^p n_{ij}$$

Tableau 1.12 – Tableau de contingence.

XY	x_1		x_i		x_p	Fréquences marginales
y_1	n_{11}		n_{i1}		n_{p1}	$n_{.1}$
y_j	n_{1j}		n_{ij}		n_{pj}	$n_{.j}$
y_q	n_{1q}		n_{iq}			$n_{.q}$
Fréquences marginales	$n_{1.}$		$n_{i.}$		$n_{p.}$	n

- La fréquence marginale $n_{i.}$ est donc le nombre d'individus possédant la modalité i du caractère X quelle que soit la distribution du caractère Y ; par exemple tous les individus ayant le même poids quelle que soit leur taille.
- Les *fréquences conditionnelles* sont définies pour chaque valeur de i et j .
- La *fréquence conditionnelle* $n_{j/i}$ est la distribution de la variable Y quand on a fixé la modalité i pour la variable X ; on s'intéresse, par exemple, à la répartition des tailles des individus ayant tous le même poids. Elle est définie par :

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

- On définit de la même façon la *fréquence conditionnelle* $n_{i/j}$ par :

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

On s'intéresse, par exemple, à la répartition des poids des individus ayant tous la même taille.

- Les *fréquences relatives* f_{ij} , $f_{i.}$ et $f_{.j}$ sont obtenues en divisant les effectifs n_{ij} et les fréquences marginales $n_{i.}$ et $n_{.j}$ par l'effectif total n .
- Les distributions X et Y sont *statistiquement indépendantes* si et seulement si :

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

pour toutes les valeurs des indices i et j .

Différents tests peuvent être mis en œuvre pour vérifier l'indépendance de deux variables statistiques (chapitre 17, tests d'indépendance).

■ Représentations graphiques

- Variables quantitatives : nuage de points dans $\mathbb{R}^2$.
- Variables qualitatives : analyse multidimensionnelle, en particulier théorie de la régression (chapitres 19, 20, 21 et 22).

1.3.2 Mesure de dépendance

L'étude de la distribution simultanée de deux variables a pour but de préciser le type de liaison pouvant exister entre ces deux variables, la nature et l'intensité de cette liaison, à l'aide de différents coefficients.

■ Variables quantitatives

□ Rapport de corrélation linéaire

Soient $\bar{x}$ et $\bar{y}$ les moyennes des valeurs prises par les variables X et Y égales à :

$$\bar{x} = \frac{1}{n} \sum_i n_i \cdot x_i \quad \bar{y} = \frac{1}{n} \sum_j n_j \cdot y_j$$

et s_X et s_Y les écarts-types de ces distributions.

Le rapport de corrélation linéaire est le coefficient symétrique par rapport aux variables X et Y défini par la relation :

$$r = \frac{\frac{1}{n} \sum_i n_{ij} (x_i - \bar{x}) (y_j - \bar{y})}{s_X s_Y}$$

On démontre que $-1 \leq r \leq 1$.

- $r = 0$ non-corrélation linéaire,
- $r = \pm 1$ relation du type $aX + bY + c = 0$ où a , b et c sont des constantes.

□ Rapport de corrélation de Y en X

Le rapport de corrélation de la variable Y par rapport à la variable X est un coefficient non symétrique défini par :

$$e_{Y/X}^2 = \frac{s_{Y/X}^2}{s_Y^2} = \frac{\frac{1}{n} \sum_i n_i \cdot (\bar{y}_i - \bar{y})^2}{\frac{1}{n} \sum_j n_j \cdot (y_j - \bar{y})^2}$$

$\bar{y}_i$ est la moyenne des valeurs prises par la variable Y pour une valeur fixée x_i de la variable X c'est-à-dire :

$$\bar{y}_i = \frac{1}{n_i} \sum_{k=1}^q n_{ik} y_k$$

On démontre que $0 \leq e^2 \leq 1$.

- $e^2 = 0$ non-corrélation,
- $e^2 = 1$ liaison fonctionnelle, à une valeur de la variable X correspond une seule valeur de la variable Y .

Remarques

- Si la variable X est qualitative et la variable Y quantitative, on peut calculer ce rapport de corrélation.
- Si pour toutes les valeurs des indices i et j , l'effectif n_{ij} est égal à 1, alors le rapport de corrélation e^2 est égal à 1 mais, dans ce cas, il n'a aucune signification.

Les propriétés de ces deux coefficients et les tests correspondants sont donnés dans le chapitre 17.

■ Variables qualitatives

Les principaux coefficients sont les suivants :

- Coefficient d^2

$$d^2 = \sum_{i,j} \frac{\left(n_{ij} - \frac{n_i \cdot n_j}{n} \right)^2}{\frac{n_i \cdot n_j}{n}} = n \left(\sum_{i,j} \frac{n_{ij}^2}{n_i \cdot n_j} - 1 \right) = n \left(\sum_{i,j} \frac{f_{ij}^2}{f_i \cdot f_j} - 1 \right)$$

Plus d^2 est petit, plus la liaison entre les variables X et Y est forte. Ses propriétés sont étudiées dans le chapitre 16, paragraphe 16.2.4.

À partir de ce coefficient, on peut en définir d'autres :

- Coefficient de contingence

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

- Coefficient de Pearson

$$\phi^2 = \frac{d^2}{n}$$

- Coefficient de Tschuprow

$$T = \frac{\phi^2}{\sqrt{(p-1)(q-1)}}$$

où p et q désignent le nombre de modalités prises par les variables X et Y respectivement. Ce coefficient est analogue à un coefficient de corrélation linéaire $0 < T < 1$.

Remarque

Dans le cas particulier où $p = q = 2$, le calcul du coefficient d^2 et donc des autres coefficients sont particulièrement simples.

Tableau 1.13 – Calcul du coefficient d^2 ($p = q = 2$).

X Y	x_1	x_2	Fréquences marginales
y_1	n_{11}	n_{21}	$n_{11} + n_{21}$
y_2	n_{12}	n_{22}	$n_{12} + n_{22}$
Fréquences marginales	$n_{11} + n_{12}$	$n_{21} + n_{22}$	n

On obtient pour le coefficient d^2 :

$$d^2 = n \frac{(n_{11} n_{22} - n_{12} n_{21})^2}{(n_{11} + n_{21})(n_{11} + n_{12})(n_{21} + n_{22})(n_{12} + n_{22})}$$

B

Calcul des probabilités

2 • LE MODÈLE PROBABILISTE

B

CALCUL DES PROBABILITÉS

2.1 Introduction

Dans des domaines très différents comme le domaine scientifique, sociologique, médical, les sciences humaines..., on s'intéresse à de nombreux phénomènes dans lesquels apparaît souvent l'effet du hasard. Ces phénomènes sont caractérisés par le fait que les résultats des observations varient d'une expérience à l'autre.

Une expérience est appelée *aléatoire* s'il est impossible de prévoir son résultat et si, répétée dans des conditions identiques, elle peut donner, ou aurait pu donner, si l'expérience est unique, des résultats différents. En général, les résultats obtenus varient dans un certain domaine, certains résultats apparaissant plus fréquemment que d'autres. Ils peuvent être visualisés par des diagrammes, des histogrammes, des courbes cumulatives de fréquences, etc., et être caractérisés par quelques valeurs numériques telles que la moyenne arithmétique, la médiane, le mode, la variance... (voir chapitre 1).

Le mot *probabilité* est passé rapidement dans le langage courant bien que la théorie des probabilités soit une branche relativement récente des théories mathématiques.

Le concept des probabilités semblait être connu des Grecs et des Égyptiens. Cependant, ce n'est que vers le milieu du XVIII^e siècle que l'on peut situer le début de cette théorie. D'abord limitée à l'étude des jeux de hasard (jeux de pile ou face, roulettes, jeux de cartes...), elle s'est rapidement étendue à tous les domaines de la Science, en Physique (théorie du potentiel, physique statistique, physique corpusculaire...), en Informatique, en Économie, en Génétique, en Psychologie... L'influence des jeux de hasard se retrouve encore dans certaines expressions, comme l'espérance mathématique qui était l'espérance du gain,

pouvant être parfois une perte. Le mot *probabilité* ou l'adjectif *probable* est bien souvent synonyme du mot *chance*.

Les premiers résultats mathématiques furent introduits par Pascal et Fermat au milieu du XVII^e siècle. Puis, apparaissent, à la fin du XVII^e siècle, le nom de Huyghens et surtout au XVIII^e siècle, les noms de Bernoulli, De Moivre, Bayes, Laplace, (le tome VII de ses œuvres s'intitule *Calcul des Probabilités*), Gauss et au XX^e siècle, Poincaré, Borel, Fréchet, Lévy, Kolmogorov, Khintchin...

Alors que la théorie du calcul des probabilités s'est développée rapidement au cours du XX^e siècle, le concept de probabilité soulève encore de nombreuses controverses non entièrement résolues. Cependant, on peut distinguer deux Écoles et différents concepts.

2.1.1 L'École objective

La probabilité d'événements répétitifs est définie à partir de la fréquence d'apparitions de ces événements. On distingue différents concepts :

■ L'approche fréquentiste ou fréquentielle

C'est la théorie de Laplace, Von Mises ; elle est fondée sur la notion d'épreuves répétées et indépendantes, la probabilité étant définie comme la limite de la fréquence relative des observations.

Cette fréquence, exprimée comme le rapport $\frac{n_a}{n}$ (n_a étant le nombre d'essais où l'événement A a été réalisé au cours de n essais indépendants, répétés dans des conditions identiques), a des fluctuations autour d'une valeur limite qui est la probabilité de l'événement A (loi des grands nombres). Mais, on suppose implicitement que la fréquence relative tend vers cette limite avec une grande *probabilité* ! C'est-à-dire, que l'on définit la probabilité à partir de la probabilité !

■ La notion de probabilité tirée des jeux de hasard

La probabilité est le quotient du nombre de cas *favorables* par le nombre de cas possibles, mais chaque cas étant supposé également possible, donc *équitable*, on définit encore la probabilité à partir de la probabilité !

■ L'approche axiomatique ou mathématique

Kolmogorov a introduit, au début du XX^e siècle (1933), les concepts probabilistes c'est-à-dire le *modèle probabiliste*. À partir d'axiomes, il a construit une théorie parfaitement logique et cohérente, le mot *hasard* n'intervenant pas. Cette axiomatique repose essentiellement sur des concepts mathématiques généraux, principalement sur la théorie de l'intégration et de la mesure.

Jusqu'à la fin du XIX^e siècle, la seule manière de définir l'intégrale d'une fonction était celle de Riemann avec les sommes de Riemann-Darboux. Grâce au concept de mesure, introduit par Borel (1894, 1897), Lebesgue élabore une théorie plus générale de l'intégration. Puis enfin, grâce à Radon vers 1913, les concepts de mesure et d'intégration, définis sur $\mathbb{R}$ et $\mathbb{R}^n$, vont être étendus à des ensembles plus généraux sur lesquels on a défini une *tribu*. La notion de tribu, les théorèmes de décomposition de Lebesgue-Nikodym et l'existence des densités ont apporté un développement considérable à la théorie des probabilités et lui ont donné sa forme actuelle.

La probabilité étant alors une mesure particulière, tous les résultats de la théorie de la mesure lui sont applicables.

2.1.2 L'École subjective

Elle associe, à la fréquence observée de la réalisation d'un événement, un degré de confiance (ou de croyance) qui permet d'évaluer la probabilité de cet événement. Elle a été développée principalement par Keynes, De Finetti, Savage...

Elle va même jusqu'à nier l'existence de probabilités objectives. Le traité de probabilités de De Finetti commence en effet par *la probabilité n'existe pas*.

Elle prend beaucoup d'importance dans les théories de la décision en associant la probabilité des événements à celle de leurs conséquences. Mais la difficulté est d'évaluer la première probabilité, c'est-à-dire la probabilité *a priori* et l'importance des conséquences dépend des utilisateurs.

2.2 Les concepts probabilistes

À l'origine *probabiliser* consistait à répartir, sur chacun des éléments d'un ensemble, un ensemble de valeurs ou *probabilités* dont la somme était égale à 1.

Si cet ensemble, ou *espace des épreuves*, est de dimension finie, il n'y a pas de difficultés majeures. En revanche, si cet espace a la puissance du continu, le problème d'associer à chacun de ses éléments, une probabilité, est pratiquement sans solution.

Pour formaliser ces notions, trois étapes sont nécessaires :

- définir le cadre dans lequel on observe les manifestations du hasard, c'est-à-dire définir une *expérience aléatoire* et l'*ensemble fondamental* Ω ,
- définir un *événement aléatoire* et la *classe* $\mathcal{C}$ des événements aléatoires,
- définir une *probabilité sur l'espace* $(\Omega, \mathcal{C})$, c'est-à-dire affecter un poids à chaque événement traduisant la chance de réalisation de cet événement.

2.2.1 Expérience aléatoire

Une expérience est dite *aléatoire* s'il est impossible d'en prévoir le résultat, c'est-à-dire, si répétée dans les mêmes conditions, elle peut donner des résultats différents, dans un ensemble d'issues considérées comme possibles :

- succession d'appels à un standard téléphonique non surchargé,
- observation de la durée de vie d'un individu anonyme dans une population humaine,
- observation de la durée de fonctionnement sans panne d'un appareil,
- jeu de pile ou face de durée infinie...

Les résultats d'une expérience aléatoire appartiennent à un *espace fondamental* ou *espace des épreuves* Ω ; un point quelconque ω de Ω est un *résultat élémentaire*.

D'où la définition :

Une expérience aléatoire est un choix au hasard d'un point ω dans un ensemble Ω .

L'ensemble Ω dépend des connaissances que l'on a, *a priori*, sur les résultats possibles de l'expérience aléatoire.

Exemples 2.1

On lance une pièce de monnaie. Pour l'ensemble Ω , on peut choisir :

- soit l'ensemble $\Omega_1 = \{\text{pile, face}\}$,
- soit l'ensemble $\Omega_2 = \{\text{pile, face, tranche}\}$.

On considère la succession des appels à un standard téléphonique non surchargé et on étudie la répartition des instants où le standard reçoit un appel, à partir d'un

instant choisi comme origine (on admet que deux appels ne peuvent se produire rigoureusement au même instant, et que le phénomène n'est pas limité dans le temps).

Une réalisation de cet événement est une suite croissante de nombres réels positifs t_i où t_i désigne l'instant d'enregistrement du $i^{\text{ème}}$ appel :

$$\omega = \{t_1 < t_2 < \dots < t_n < t_{n+1} < \dots\}.$$

Ω est donc une partie de $(\mathbb{R}^+)^{\mathbb{N}}$

On lance deux dés et on s'intéresse à la somme des points apparaissant sur les deux dés. On obtient :

- soit $\Omega_1 = \{2, 3, \dots, 12\}$
- soit $\Omega_2 = \{2, 3, \dots, 12\}^N$

si on recommence N fois la partie.

On lance deux dés et on s'intéresse aux points marqués sur chaque dé :

$$\omega = \{x, y\} \text{ avec } 1 \leq x \leq y \leq 6$$

$$\Omega = \{x, y\}^6 \text{ est une partie de } \mathbb{Z}^2$$

On considère l'expérience aléatoire « durée de vie d'un individu ». L'ensemble Ω est soit l'ensemble $\mathbb{N}$, soit la demi-droite réelle positive $\mathbb{R}$ selon le procédé discontinu ou continu de cette mesure.

Le choix de l'espace Ω peut s'avérer difficile ou même arbitraire. Si on répète l'expérience une infinité de fois, les espaces qui vont intervenir seront $\mathbb{Z}^{\mathbb{N}}$ ou $\mathbb{R}^{\mathbb{N}}$ de dimension infinie. Dans certains cas, il faut même faire intervenir des espaces fonctionnels.

2.2.2 Événement aléatoire

Un événement aléatoire est lié à une expérience aléatoire ; une fois l'expérience réalisée, on peut alors dire si l'événement a été réalisé ou non.

Un événement aléatoire A peut être identifié à la partie de Ω dont les éléments réalisent l'événement A .

Exemple 2.2

On jette deux dés et soit A l'événement :

« le total des points est supérieur ou égal à 11 ».

L'ensemble des résultats possibles est l'ensemble $\Omega = \{(1, 1), (1, 2), \dots, (6, 6)\}$.

Un total supérieur ou égal à 11 est représenté par l'ensemble des trois couples :

$\omega = \{x, y\}$ tels que $x + y \geq 11$, c'est-à-dire les couples $\{5, 6\}$, $\{6, 5\}$, $\{6, 6\}$.

Cet ensemble de trois couples réalise l'événement A.

On pourrait choisir pour l'ensemble des événements, l'ensemble $P(\Omega)$ des parties de Ω , mais comme cet ensemble est en général trop vaste pour être « probabilisé », on se limite à un ensemble strictement contenu dans $P(\Omega)$, vérifiant les propriétés logiques suivantes, qui servent de base axiomatique à la définition mathématique de la notion d'événement aléatoire.

■ Parallélisme entre la terminologie ensembliste et la terminologie probabiliste

- À tout événement A est associé son *contraire*, non A ou $\bar{A}$ ou A^c qui est réalisé si et seulement si A ne l'est pas.
Dans l'espace Ω des événements, A et $\bar{A}$ sont représentés par des ensembles complémentaires au sens ensembliste.
- Pour tout couple d'événements A et B, l'événement « A et B » est réalisé si A et B sont réalisés.
Dans l'espace Ω des événements, l'événement « A et B » est représenté par l'intersection des ensembles réalisant A et B, on le note « A et B » ou « $A \cap B$ ».
- Pour tout couple d'événements A et B, l'événement « A ou B » est réalisé si l'un des deux ou si les deux sont réalisés.
Dans l'espace Ω des événements, il est représenté par la réunion des ensembles réalisant A et B, on le note, *ou n'étant pas exclusif*, « A ou B » ou « $A \cup B$ ».
- Deux événements A et B sont *incompatibles* si la réalisation de l'un implique la non réalisation de l'autre,
Dans l'espace Ω des événements, deux événements incompatibles sont représentés par deux parties disjointes.
- Les événements $A_1, A_2, \dots, A_n$ forment un *système complet* d'événements ou *système exhaustif* si les ensembles qui leur sont associés forment une *partition* de l'espace Ω .

Tableau 2.1 – Terminologies probabiliste et ensembliste.

Terminologie probabiliste	Terminologie ensembliste	Notation
Événement certain	Espace entier	Ω
Événement impossible	Partie vide	$\emptyset$
Événement contraire	Complémentaire	$\bar{A}$ ou A^c
A et B	Intersection	$A \cap B$
A ou B (ou non exclusif)	Réunion	$A \cup B$
Événements incompatibles	Parties disjointes	$A \cap B = \emptyset$
Système complet d'événements	Partition de Ω	$A_i \cap B_j = \emptyset$ $\cup A_i = \Omega$
Implication $A \Rightarrow B$	Inclusion	$A \subset B$

Implication $A \subset B$ ou $A \Rightarrow B$: l'événement A ne peut être réalisé sans que B le soit.

Toutes les opérations précédemment définies s'étendent à plus de deux événements. La classe des événements associés à une expérience aléatoire est donc une *tribu* $\mathcal{C}$ de parties de Ω (tribu ou σ -algèbre). (Voir annexe 2 la définition d'une tribu.)

En résumé :

Un espace probabilisable est un couple $(\Omega, \mathcal{C})$ formé d'un ensemble Ω et d'une tribu $\mathcal{C}$ de parties de Ω (qui sont les événements).

2.2.3 Quantification des résultats

Le résultat d'une expérience aléatoire ne peut pas être prévu avec certitude. La théorie des probabilités doit cependant donner des résultats quantifiés, donc associer à chaque événement un *poids*, c'est-à-dire un nombre qui évalue sa chance de réalisation, ce nombre traduit la *loi* du phénomène étudié.

Historiquement, cette notion s'est dégagée à partir de la notion de fréquence de réalisation d'un événement A lié à une expérience ω , au cours d'une suite de répétitions identiques de ω . Puis l'approche axiomatique, utilisée depuis la fin du siècle dernier, a donné les bases mathématiques à la théorie des probabilités.

2.3 Mesure de probabilité et espace probabilisé

2.3.1 Définition de la probabilité

Intuitivement, si A et B sont deux événements incompatibles, la *chance* de voir se réaliser A ou B doit être égale à la somme des poids traduisant les chances de réalisation de A et B. De même, si (A_n) , n appartenant à $\mathbb{N}$, désigne un ensemble d'événements tel que chacun d'eux est impliqué par le suivant et tel que leur réalisation simultanée est impossible, alors le poids de A_n a une limite nulle quand n tend vers l'infini.

Une *probabilité* $\Pr$ définie sur l'ensemble (Ω, C) , est une application de C dans $[0, 1]$ telle que :

- $\Pr(\Omega) = 1$
- $\Pr(\cup A_i) = \sum_i \Pr(A_i)$ pour toute réunion finie ou dénombrable d'événements incompatibles.

Le triplet $(\Omega, C, \Pr)$ est un *espace probabilisé*, la mesure $\Pr$ ainsi définie est une mesure positive de masse totale égale à 1.

2.3.2 Propriétés élémentaires

Elles se déduisent des axiomes de définition :

- $\Pr(\emptyset) = 0$ mais $\Pr(A) = 0$ n'implique pas $A = \emptyset$
L'événement A tel que $\Pr(A) = 0$ est un événement presque impossible.
- $\Pr(\overline{A}) = 1 - \Pr(A)$
- $\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B)$
- $\Pr(\cup A_i) \leq \sum_i \Pr(A_i)$ (aucune hypothèse particulière sur les événements A_i)
- si la suite des événements A_i tend vers 0 en décroissant, la limite de $\Pr(A_i)$ est nulle.
- si B_i est un système complet d'événements, alors

$$\forall A, \quad \Pr(A) = \sum_i \Pr(A \cap B_i)$$

C'est la première forme du *théorème des probabilités totales*.

Remarque

$\Pr(A) = 1$ n'implique pas $A = \Omega$. L'événement A tel que $\Pr(A) = 1$ est un événement presque certain.

2.4 Échantillons et sous-populations

De nombreux problèmes faisant intervenir le calcul des probabilités se ramènent aux problèmes de tirer des échantillons de taille r dans un ensemble de taille n , appelé population, quelle que soit la nature de ses éléments. Suivant la règle du tirage, cet échantillon est :

- ordonné ou non,
- avec ou sans répétitions (on dit aussi avec ou sans remise).

Deux autres espaces interviennent souvent dans des problèmes élémentaires, l'espace des sous-populations de taille r avec répétitions et l'espace des permutations de n objets.

Remarque

Choisir un élément au hasard, signifie que les divers choix possibles sont équiprobables donc que l'ensemble Ω est muni de la loi de probabilité uniforme. Dans ce cas, tous les calculs sont simples et se ramènent souvent à des calculs d'analyse combinatoire.

Des rappels d'analyse combinatoire sont développés dans l'annexe 1.

3 • PROBABILITÉ CONDITIONNELLE INDÉPENDANCE

3.1 Définition

Soit $(\Omega, \mathcal{C}, \Pr)$ un espace probabilisé. L'intersection de deux événements A et B est l'événement, noté $A \cap B$, réalisé, si et seulement si, les deux événements A et B sont réalisés. Cependant, on peut s'intéresser à la réalisation de l'événement A sachant l'événement B réalisé, si cet événement est de probabilité non nulle, c'est-à-dire on s'intéresse à la probabilité conditionnelle sachant B.

La *probabilité conditionnelle sachant B* est l'application de $\mathcal{C}$ dans $[0, 1]$ définie par :

$$\forall A \in \mathcal{C} \quad \Pr(A/B) = \frac{\Pr(A \cap B)}{\Pr(B)}$$

Cette application définit une probabilité sur le même espace probabilisé $(\Omega, \mathcal{C}, \Pr)$, la probabilité conditionnelle $\Pr(. / B)$ est définie comme la probabilité $\Pr$ sur la tribu $\mathcal{C}$, le terme $\Pr(B)$ est un facteur de normalisation.

Selon les événements A et B, différents cas sont possibles.

3.1.1 Les événements A et B sont incompatibles

L'événement A ne se réalisera pas si l'événement B est réalisé :

$$\Pr(A/B) = 0$$

Exemple 3.1

On lance deux dés et on considère les deux événements :

- A : obtenir un chiffre impair sur les deux dés,
- B : la somme des points obtenus sur les deux dés est un nombre impair.

Ces deux événements sont incompatibles.

3.1.2 Les événements A et B ne sont pas incompatibles

Deux événements peuvent être totalement dépendants ou dépendants.

– *Événements totalement dépendants*

Deux événements A et B sont totalement dépendants si $A \subset B$, ou si l'événement B étant réalisé, la probabilité de réalisation de l'événement A est égale à 1 :

$$\Pr(A/B) = 1$$

On dit que A dépend totalement de B.

Exemple 3.2

Les événements suivants sont totalement dépendants :

– A : le nombre est égal à 4, 6, 8,

– B : le nombre est un nombre pair compris entre 2 et 20.

– *Événements dépendants*

Deux événements A et B sont dépendants si la probabilité de réalisation de l'événement A change selon que B est réalisé ou non.

Exemple 3.3

On lance un dé parfaitement équilibré et on considère les événements suivants :

– A : obtenir la face 6,

– B : obtenir un nombre pair,

– C : obtenir un nombre supérieur ou égal à 3.

$$\Pr(A) = 1/6 \quad \Pr(B) = 1/2 \quad \Pr(C) = 4/6 = 2/3$$

Si l'événement B réalisé, la probabilité de réalisation de A est égale à 1/3.

Si l'événement C réalisé, la probabilité de réalisation de A est égale à 1/4.

Les probabilités conditionnelles de A ne sont donc pas égales à la probabilité de A ni égales entre elles :

$$\Pr(A) = 1/6 \quad \Pr(A/B) = 1/3 \quad \Pr(A/C) = 1/4$$

Les événements A et B d'une part, A et C d'autre part sont dépendants.

3.2 Principe des probabilités composées

Le principe des probabilités composées découle des axiomes et des définitions. Il s'écrit :

$$\Pr(A \cap B) = \Pr(A/B) \Pr(B) = \Pr(B/A) \Pr(A)$$

Cette formule est valable même si les probabilités $\Pr(A)$ et $\Pr(B)$ sont nulles toutes les deux ; mais dans ces conditions, on ne peut pas définir $\Pr(A/B)$ ni $\Pr(B/A)$.

3.3 Événements indépendants

3.3.1 Définition

L'événement A est *indépendant* de l'événement B si la probabilité de réalisation de l'événement A n'est pas modifiée par une information concernant la réalisation de l'événement B , c'est-à-dire si :

$$\Pr(A/B) = \Pr(A)$$

Le principe des probabilités composées entraîne :

$$\Pr(A \cap B) = \Pr(A) \Pr(B) = \Pr(B/A) \Pr(A)$$

$$\Pr(B/A) = \Pr(B)$$

L'événement B est donc également indépendant de l'événement A . Les événements A et B sont indépendants et vérifient la propriété :

$$\Pr(A \cap B) = \Pr(A) \Pr(B)$$

3.3.2 Événements incompatibles et événements indépendants

– La propriété « les événements A et B sont incompatibles » implique :

$$\Pr(A \cup B) = \Pr(A) + \Pr(B)$$

– La propriété « les événements A et B sont indépendants » implique :

$$\Pr(A \cap B) = \Pr(A) \times \Pr(B)$$

Les opérations union $\cup$ et somme semblent jouer le même rôle que les opérations intersection $\cap$ et produit. Cependant, les deux concepts, incompatibles et indépendants, sont totalement différents :

- Le premier « événements incompatibles » est une notion ensembliste.
- Le second « événements indépendants » est une notion probabiliste : deux événements peuvent être indépendants pour une loi de probabilité et non pour une autre loi.

3.4 Indépendance deux à deux et indépendance mutuelle

La notion d'indépendance et le principe des probabilités composées se généralisent à plusieurs événements.

3.4.1 Généralisation du principe des probabilités composées

Ce principe se traduit par la *formule de Poincaré* que l'on démontre par récurrence :

$$\Pr(A_1 \cap A_2 \cap \dots \cap A_n) = \Pr(A_1) \Pr(A_2/A_1) \Pr(A_3/A_1 \cap A_2) \dots \Pr(A_n/A_1 \cap A_2 \cap \dots \cap A_{n-1})$$

3.4.2 Indépendance mutuelle

Les événements A_i , $i \in (1, \dots, n)$, sont *mutuellement indépendants* si, pour toute partie I de l'ensemble des indices, on a :

$$\Pr\left(\bigcap_{i \in I} A_i\right) = \prod_{i \in I} \Pr(A_i)$$

L'indépendance mutuelle implique l'indépendance deux à deux mais c'est une condition plus forte.

Exemple 3.4

On lance deux dés et on considère les événements suivants :

- A : le premier dé donne une face impaire,
- B : le deuxième dé donne une face impaire,

– C : la somme des points apparaissant sur les deux faces est impaire.

Les événements A, B et C sont deux à deux indépendants. En effet :

$$\Pr(A) = 1/2 \quad \Pr(B) = 1/2 \quad \Pr(C) = 1/2$$

$$\Pr(A \cap B) = \Pr(A \cap C) = \Pr(B \cap C) = 1/4$$

Les événements A, B et C ne sont pas indépendants :

$$\Pr(A \cap B \cap C) = 0$$

3.5 Théorème de Bayes

3.5.1 Deuxième forme du théorème des probabilités totales

On considère un événement A de probabilité non nulle et l'ensemble $(C_i)_{i \in \{1, \dots, n\}}$ de toutes les causes possibles de réalisation de cet événement ; cet ensemble forme un ensemble complet d'événements et l'événement A se produit en même temps qu'un et un seul des C_i , c'est-à-dire :

$$A = (A \cap C_1) \cup (A \cap C_2) \cup \dots \cup (A \cap C_n)$$

On en déduit la deuxième forme du théorème des probabilités totales :

$$\Pr(A) = \sum_{i=1}^n \Pr(A \cap C_i) = \sum_{i=1}^n \Pr(A/C_i) \Pr(C_i)$$

3.5.2 Théorème de Bayes

Considérons une des causes susceptibles de réaliser l'événement A, la cause C_k par exemple. Le théorème des probabilités composées donne :

$$\Pr(A \cap C_k) = \Pr(A/C_k) \Pr(C_k) = \Pr(C_k/A) \Pr(A)$$

De la deuxième forme du théorème des probabilités totales, on déduit $\Pr(A)$, puis le *théorème de Bayes* :

$$\Pr(C_k/A) = \frac{\Pr(A/C_k) \Pr(C_k)}{\sum_{i=1}^n \Pr(A/C_i) \Pr(C_i)}$$

Sous cette forme, le théorème de Bayes (publié après sa mort en 1763) apparaît comme une conséquence logique des axiomes et des définitions. Il présente un

grand intérêt, car il permet de modifier notre connaissance des probabilités en fonction d'informations nouvelles, il joue un rôle très important dans la statistique bayésienne.

Exemple 3.5

Trois machines automatiques produisent des pièces de voitures. La machine M_1 produit 40 % du total des pièces, la machine M_2 25 % et la machine M_3 produit 35 %. En moyenne, les pourcentages des pièces non conformes aux critères imposés sont de 10% pour la machine M_1 , de 5 % pour la machine M_2 et de 1 % pour la machine M_3 .

Une pièce est choisie au hasard dans la production totale des trois machines. On constate qu'elle n'est pas conforme aux critères imposés.

Quelle est la probabilité qu'elle ait été produite par la machine M_1 ?

On peut appliquer directement le théorème de Bayes.

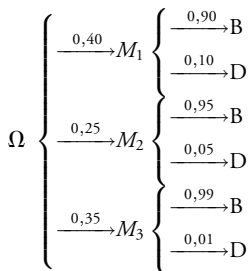
Soit B l'événement « la pièce est bonne » et D l'événement « la pièce est défectueuse ».

Les trois causes possibles de réalisation de l'événement D sont les trois machines.

On connaît les probabilités de ces causes par exemple $\Pr(M_1) = 0,40$ ainsi que les probabilités conditionnelles $\Pr(D/M_1) = 0,10$.

$$\Pr(M_1/D) = \frac{\Pr(D/M_1) \Pr(M_1)}{\sum_{i=1}^3 \Pr(D/M_i) \Pr(M_i)}$$

On peut visualiser ce problème par l'arbre suivant :



$$\Pr(M_1/D) = \frac{\Pr(M_1 \text{ et } D)}{\Pr(D)}$$

$$\Pr(D) = 0,40 \times 0,10 + 0,25 \times 0,05 + 0,35 \times 0,01 = 0,056$$

$$\Pr(M_1/D) = \frac{0,40 \times 0,10}{0,056} = 0,714$$

3.5.3 Signification et rôle de ce théorème

- Les événements C_i constituent l'ensemble de toutes les causes possibles et exclusives de réalisation d'un événement A .
- Les probabilités $\Pr(C_i)$ des événements C_i (pour chaque valeur de l'indice i) sont évaluées compte tenu de notre connaissance relative aux conditions dans lesquelles l'événement A s'est produit ou se produira.
- Les probabilités $\Pr(A/C_i)$ sont les probabilités de réalisation de A dans l'éventualité C_i (pour chaque valeur de l'indice i). L'événement A étant lié aux événements C_i , nos connaissances sur ces liens permettent d'attribuer des valeurs aux probabilités conditionnelles.

L'événement A est réalisé :

- les probabilités $\Pr(A/C_i)$ ne changent pas,
- les probabilités $\Pr(C_i)$ deviennent caduques, on doit les remplacer par les probabilités sachant A réalisé, c'est-à-dire les expressions $\Pr(C_i/A)$,
- on est donc passé des probabilités *a priori* aux probabilités *a posteriori*.

L'expression « *a priori* » ne signifie pas en l'absence de toute information ; les expressions correctes sont probabilités *avant et après information*, car il est impossible de définir des probabilités de réalisation d'événements sur lesquels on n'a aucune information.

3.5.4 Conclusion

La probabilité d'un événement peut être considérée comme une caractéristique de notre information à son sujet que l'on modifie dès que cette information est complétée.

Toute probabilité est donc conditionnelle et dépend de notre connaissance des objets en cause.

Nous devons nous souvenir que la probabilité d'un événement n'est pas une qualité de l'événement lui-même mais un simple mot pour désigner le degré de connaissance que nous, ou quelqu'un d'autre, peut espérer.

J. Stuart Mill (1806-1873)

Cette démarche bayésienne est une des approches possibles de la probabilité ; elle peut servir au diagnostic médical, à la théorie de la décision...

4 • VARIABLES ALÉATOIRES RÉELLES

B

CALCUL DES PROBABILITÉS

4.1 Généralités sur les variables aléatoires

4.1.1 Définition d'une variable aléatoire

Les variables aléatoires constituent un espace fondamental d'éléments aléatoires, un tel élément étant défini par référence à une expérience aléatoire.

Si $(\Omega, \mathcal{C}, \text{Pr})$ désigne un espace probabilisé et $(E, \mathcal{E})$ un espace probabilisable, un *élément aléatoire*, défini sur $(\Omega, \mathcal{C}, \text{Pr})$ et à valeurs dans $(E, \mathcal{E})$, est une application mesurable de $(\Omega, \mathcal{C})$ dans $(E, \mathcal{E})$. Cet élément est appelé :

- variable aléatoire réelle si l'espace $(E, \mathcal{E})$ est l'espace $(\mathbb{R}, \mathcal{B})$, où $\mathcal{B}$ est la tribu de Borel de $\mathbb{R}$,
- variable aléatoire complexe si l'espace $(E, \mathcal{E})$ est l'espace $(\mathbb{C}, \mathcal{C})$,
- variable aléatoire vectorielle ou vecteur aléatoire, de dimension n , si l'espace $(E, \mathcal{E})$ est l'espace $(\mathbb{R}^n, \mathcal{B}^n)$.

Dans ce chapitre, on ne définira que des variables aléatoires réelles. Les propriétés de ces variables sont donc celles des fonctions réelles mesurables.

Exemple 4.1 Variable aléatoire

On jette n fois une pièce de monnaie. L'espace fondamental est $\Omega = (P, F)^n$ où P désigne pile et F face ; la tribu associée est la tribu $\mathcal{P}(\Omega)$ des parties de Ω .

On peut s'intéresser :

- soit aux résultats élémentaires :

$\omega = (\omega_1, \omega_2, \dots, \omega_n)$ où ω_i désigne soit pile, soit face.

On obtient, par exemple, la succession $\omega = (P, F, F, F, P, F)$ pour $n = 6$

– soit au nombre de fois où « pile » est sorti au cours des n jets.

On obtient, par exemple, 2 fois pile quand on a lancé 6 fois la pièce.

On définit une fonction X application de Ω dans $\Omega' = (1, 2, \dots, n)$

où $X(\omega)$ est le nombre de fois où « pile » apparaît dans ω .

Si $\omega = (P, F, F, F, P, F)$, $X(\omega) = 2$. Si la pièce est parfaitement équilibrée, il semble logique de munir $(\Omega, P(\Omega))$ de la loi de *probabilité uniforme* :

$$\Pr(P) = \Pr(F) = 1/2$$

Sur l'espace $(\Omega', P(\Omega'))$, on définit une probabilité $\Pr_X$ ou $\Pr'$, image de $\Pr$ par l'application :

$$\forall A' \in P(\Omega') \quad \Pr'(A') = \Pr(X^{-1}(A'))$$

Cette application X est une variable aléatoire.

4.1.2 Loi de probabilité d'une variable aléatoire réelle X

La loi de probabilité d'une variable aléatoire réelle X est la loi de probabilité $\Pr_X$ définie sur l'espace $(\mathbb{R}, \mathcal{B})$ par :

$$\forall B \in \mathcal{B} \quad \Pr_X(B) = \Pr(\omega/X(\omega) \in B) = \Pr(X^{-1}(B))$$

On montre facilement que $\Pr_X$ est une mesure positive sur $(\mathbb{R}, \mathcal{B})$ et comme :

$$\Pr_X(\mathbb{R}) = \Pr(X^{-1}(\mathbb{R})) = \Pr(\Omega) = 1$$

cette mesure est une probabilité. $\Pr_X$ est la mesure image de $\Pr$ par X .

$(\mathbb{R}, \mathcal{B}, \Pr_X)$ est l'espace de probabilité associé à la variable aléatoire réelle X .

Une variable aléatoire réelle traduit donc l'idée de résultat numérique associé à un phénomène aléatoire.

Exemple 4.2 Loi de probabilité

On jette deux dés équilibrés et on s'intéresse à la somme S des points figurant sur les deux dés. On définit les espaces Ω et Ω' par :

$$- \Omega = (1, 2, \dots, 6)^2$$

$$- \text{et } \Omega' = (2, 3, \dots, 12)$$

Ω est l'espace fondamental ou ensemble des couples $\omega = (n_1, n_2)$, n_1 et n_2 prenant les valeurs entières entre 1 et 6, bornes comprises, et Ω' est l'ensemble des résultats possibles, c'est-à-dire l'ensemble des valeurs que la somme S peut prendre. Soit X l'application de Ω dans Ω' telle que :

$$X(\omega) = (n_1, n_2)$$

$\Pr(\omega) = 1/36$ car tous les éléments de Ω ont la même probabilité de réalisation et le cardinal de Ω est égal à 36.

Par définition, $\Pr'(A') = \Pr(X^{-1}(A'))$

Ainsi, $\Pr'(6) = \Pr\{X^{-1}(6)\} = \Pr\{(1, 5), (2, 4), (3, 3), (4, 2), (5, 1)\} = 5/36$.

La loi $\Pr'$ est constituée de masses ponctuelles, elle peut donc être représentée par un diagramme en bâtons.

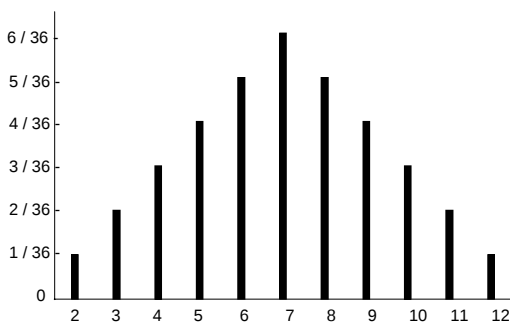


Figure 4.1 – Histogramme de la loi de la variable aléatoire S (somme des points obtenus en lançant deux dés).

B

CALCUL DES PROBABILITÉS

4.1.3 Quelle tribu de $\mathbb{R}$ doit-on choisir ?

Une variable aléatoire réelle est un procédé de mesure d'un phénomène aléatoire. La question essentielle est de connaître la probabilité que X prenne ses valeurs dans un intervalle $[a, b]$ et ceci, quel que soit cet intervalle, car la probabilité que X prenne une valeur donnée, est souvent nulle. $\Pr_X$ permet de donner un sens à cette notion puisque :

$$\Pr_X([a, b]) = \Pr(X \in [a, b])$$

La tribu de Borel $\mathcal{B}$ de $\mathbb{R}$ est la plus petite tribu de $\mathbb{R}$ contenant les intervalles, d'où le choix.

En résumé, les propriétés des variables aléatoires réelles sont donc celles des fonctions mesurables. Il en résulte, en particulier que la composition, la somme, le produit de deux variables aléatoires réelles, la limite d'une suite dénombrable de variables aléatoires réelles est une variable aléatoire réelle.

4.2 Fonction de répartition

Une loi de probabilité est une mesure abstraite sur $\mathbb{R}$, elle est donc en général peu maniable et peu utilisée dans les applications concrètes. Or, la tribu de Borel $\mathcal{B}$ de $\mathbb{R}$ contient les intervalles du type $] -\infty, x[$, on en déduit la notion de fonction de répartition.

4.2.1 Définition

La *fonction de répartition* de la variable aléatoire réelle X est l'application F de $\mathbb{R}$ dans $\mathbb{R}$ définie par :

$$\forall x \in \mathbb{R} \quad F(x) = \Pr_x(-\infty, x) = \Pr \{ \omega / X(\omega) < x \}$$

On écrit plus simplement :

$$\forall x \in \mathbb{R} \quad F(x) = \Pr(X < x)$$

4.2.2 Principales propriétés

Elles se déduisent de la définition et des propriétés d'une probabilité (mesure positive, finie, définie sur $\mathbb{R}$) :

- une fonction de répartition est une fonction F définie sur $\mathbb{R}$ et à valeurs dans $[0, 1]$,
- une fonction de répartition est une fonction croissante au sens large,
- la limite de $F(x)$ quand x tend vers $-\infty$ est égale à 0,
- la limite de $F(x)$ quand x tend vers $+\infty$ est égale à 1,
- une fonction de répartition est continue à gauche, c'est-à-dire $F(x) = F(x^-)$
- si la variable aléatoire réelle est continue, la fonction F est continue à droite et dérivable,

- la fonction de répartition permet de traiter tous les problèmes faisant intervenir une seule variable aléatoire X . La probabilité de tout intervalle de $\mathbb{R}$ est égale à :

$$\Pr(a \leq X < b) = F(b) - F(a)$$

En revanche, si on considère plusieurs variables aléatoires réelles, la seule connaissance des fonctions de répartition de chaque variable est insuffisante pour déterminer la dépendance entre ces variables.

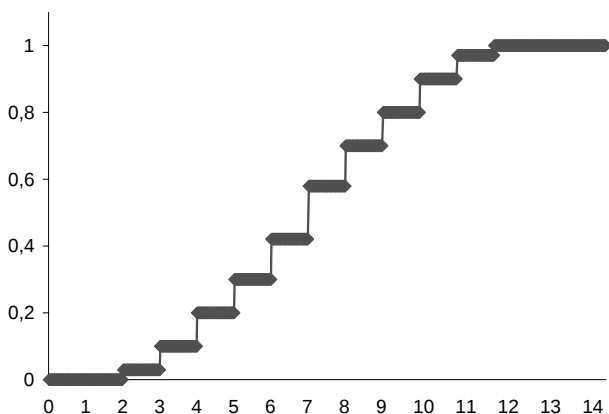


Figure 4.2 – Fonction de répartition de la variable aléatoire S (somme des points obtenus en lançant deux dés).

Remarque

Cette fonction de répartition est aussi appelée *fonction de répartition en non-dépassement* par opposition à une fonction G qui serait définie par :

$$G(x) = \Pr(X \geq x)$$

ou *fonction de répartition en dépassement*. Cette fonction, étant décroissante au sens large, a une dérivée négative qui ne peut pas définir une probabilité. Pour cette raison, c'est la fonction de répartition en non-dépassement qui est utilisée comme fonction de répartition.

4.3 Densité de probabilité

4.3.1 Définition

Si la loi de probabilité $\Pr_x$ d'une variable aléatoire réelle X admet une densité par rapport à la mesure de Lebesgue λ sur $\mathbb{R}$, cette densité est appelée *densité de probabilité de la variable X* . Plus simplement, on peut définir la densité f si elle existe par :

$$f(x) dx = \Pr(x \leq X < x + dx)$$

où dx désigne la mesure de Lebesgue sur $\mathbb{R}$.

4.3.2 Relation entre fonction de répartition et densité

Soit X une variable aléatoire réelle et F sa fonction de répartition. Si la loi de probabilité $\Pr_x$ admet une densité f , on peut écrire :

$$F(x) = \Pr_x]-\infty, x[= \int_{\mathbb{R}} 1_{]-\infty, x[} d\Pr_x = \int_{\mathbb{R}} 1_{]-\infty, x[} f(x) dx$$

$$F(x) = \int_{]-\infty, x[} f(x) dx = \int_{]-\infty, x[} f(x) dx$$

$1_{[a,b]}$ est la fonction caractéristique de l'intervalle $[a, b]$ (voir annexe 2).

Les deux dernières intégrales sont égales car un ensemble réduit à un point est un ensemble de mesure de Lebesgue nulle.

Si de plus f est continue en x , F est dérivable et $F'(x) = f(x)$. On peut alors écrire :

$$\Pr(a \leq X < b) = \int_a^b f(x) dx = F(b) - F(a)$$

Les figures 4.3 (représentation graphique de f) et 4.4 (représentation graphique de F) mettent en évidence la relation existant entre ces deux fonctions.

4.3.3 Caractérisation d'une densité de probabilité

Une application mesurable de $\mathbb{R}$ dans $\mathbb{R}^+$ telle que :

$$\int_{\mathbb{R}} f d\lambda = 1$$

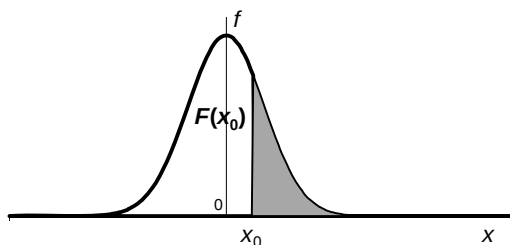


Figure 4.3 – Exemple de densité de probabilité.

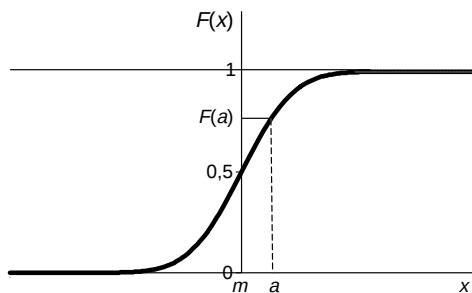


Figure 4.4 – Exemple de fonction de répartition.

peut être considérée comme la densité d'une loi de probabilité. En effet, la fonction F définie par :

$$\forall x \in \mathbb{R} \quad F(x) = \int_{]-\infty, x[} f(x) \, dx$$

est une fonction :

- monotone, croissante,
- continue,
- $F(+\infty) = 1$ et $F(-\infty) = 0$

Cette fonction F a les propriétés d'une fonction de répartition, la fonction f est la densité correspondante. Cependant, une loi de probabilité ne peut admettre de densité que si sa fonction de répartition est continue (condition nécessaire mais non suffisante).

Remarques

– Retrouver f par dérivation de F ou F par intégration de f n'a de sens que pour des variables aléatoires continues.

– La propriété :

$$\int_{\mathbb{R}} f \, d\lambda = 1$$

correspond à la propriété $\Pr(\Omega) = 1$.

– Pour une variable aléatoire réelle discrète, l'intégrale est remplacée par une somme finie ou infinie.

4.4 Discontinuités d'une fonction de répartition et lois discrètes

On suppose que la fonction de répartition F d'une variable aléatoire réelle X admet une discontinuité en x_0 , alors $F(x_0^+) - F(x_0) > 0$ et la loi de probabilité $\Pr_X$ admet en x_0 , une masse ponctuelle égale à $F(x_0^+) - F(x_0)$.

Une *discontinuité* de F en un point entraîne l'existence d'une masse ponctuelle, au point correspondant, pour la distribution.

Ce résultat se généralise au cas où F admet une infinité, au plus dénombrable, de points de discontinuité. Les lois de probabilité constituées d'une somme, au plus dénombrable, de masses ponctuelles sont appelées *lois discrètes*.

Une variable aléatoire discrète est une variable aléatoire dont la loi de probabilité est discrète ; sa fonction de répartition se compose de segments.

Une loi discrète est définie par deux suites numériques (a_n) et (p_n) , $n \in \mathbb{N}$, ayant les propriétés suivantes :

$$\Pr(a_n) = p_n \quad p_n \geq 0 \quad \sum_n p_n = 1$$

Remarque

La variable aléatoire S somme des points marqués sur les deux dés est une variable aléatoire discrète.

4.5 Loi de probabilité d'une variable aléatoire Y fonction d'une variable aléatoire X

X est une variable aléatoire réelle définie sur $(\Omega, \mathcal{C}, \text{Pr})$, admettant F pour fonction de répartition et f pour densité.

φ est une application mesurable de $\mathbb{R}$ dans $\mathbb{R}$, muni de sa tribu de Borel.

L'application composée, $\varphi \circ X$ de $(\Omega, \mathcal{C})$ dans $\mathbb{R}$ est mesurable, elle définit donc une variable aléatoire réelle notée $Y = \varphi(X)$.

Soient Pr_X et Pr_Y les lois de probabilité des variables X et Y respectivement. Pour tout borélien de $\mathbb{R}$, on a :

$$\text{Pr}_Y(B) = \text{Pr}\{Y^{-1}(B)\} = \text{Pr}\{X^{-1}\varphi^{-1}(B)\} = \text{Pr}_X\{\varphi^{-1}(B)\}$$

Pr_Y est par définition la mesure image de Pr_X par l'application φ . Deux cas sont à distinguer selon que l'application est bijective ou non.

- φ est une application *bijective*, ayant une fonction inverse φ^{-1} dérivable. La variable aléatoire Y admet pour fonction de répartition et pour densité les expressions :

$$G(y) = F[\varphi^{-1}(y)] \quad \text{et} \quad g(y) = \frac{f[\varphi^{-1}(y)]}{|\varphi'[\varphi^{-1}(y)]|}$$

- φ est une application *quelconque*, mesurable de $\mathbb{R}$ dans $\mathbb{R}$. La fonction de répartition et la densité de la variable Y sont obtenues cherchant directement l'antécédent ou les antécédents, pour la variable X , de l'événement $Y < y$.

Exemple 4.3

– La variable aléatoire X suit une loi uniforme sur $[-1, 2]$.

Densité de la variable :

- $f(x) \, dx = 1/3$ sur $[-1, 2]$,
- $f(x) = 0$ sinon.

Fonction de répartition :

- $F(x) = 0$ $x \leq -1$,
- $F(x) = (x + 1)/3$ $-1 \leq x \leq 2$,
- $F(x) = 1$ $x \geq 2$.

– On considère la variable $Y = \varphi(X) = X^2$

L'application φ n'est pas bijective sur $[-1, 2]$

- Elle est bijective et croissante sur $]1, 2]$. La formule générale s'applique dans ce cas et donne :

$$x \in]1, 2[\Leftrightarrow y \in]1, 4[\quad g(y) = \frac{1}{6\sqrt{y}} \quad G(y) = \frac{\sqrt{y} + 1}{3}$$

- L'application φ n'est pas bijective sur $[-1, 1]$

$$x \in [-1, 1] \Rightarrow y \in]0, 1]$$

$$G(y) = \Pr(Y < y) = \Pr(-\sqrt{y} < X < \sqrt{y})$$

$$\Pr(X < -\sqrt{y}) = \frac{-\sqrt{y} + 1}{3} \quad \Pr(X < \sqrt{y}) = \frac{\sqrt{y} + 1}{3}$$

D'où :

$$G(y) = \frac{2\sqrt{y}}{3} \quad \text{et} \quad g(y) = \frac{1}{3\sqrt{y}}$$

En résumé :

$$\begin{array}{lll} y \leq 0 & g(y) = 0 & G(y) = 0 \\ y \in]0, 1] & g(y) = \frac{1}{3\sqrt{y}} & G(y) = \frac{2\sqrt{y}}{3} \\ y \in]1, 4] & g(y) = \frac{1}{6\sqrt{y}} & G(y) = \frac{\sqrt{y} + 1}{3} \\ y > 4 & g(y) = 0 & G(y) = 1 \end{array}$$

La fonction de répartition est continue mais la densité est discontinue pour $y = 1$.

4.6 Indépendance de deux variables aléatoires

Soient X et Y deux variables aléatoires réelles définies sur le même espace probabilisé $(\Omega, \mathcal{C}, \Pr)$. Le couple (X, Y) est donc une application mesurable de $(\Omega, \mathcal{C})$ dans $\mathbb{R}^2$, muni de sa tribu de Borel.

X et Y sont deux variables aléatoires *indépendantes* si, pour tout couple de boréliens $\mathcal{B}_i$ et $\mathcal{B}_j$ de $\mathbb{R}$, on a :

$$\Pr\{(X \in \mathcal{B}_i) \cap (Y \in \mathcal{B}_j)\} = \Pr(X \in \mathcal{B}_i) \Pr(Y \in \mathcal{B}_j)$$

La loi de probabilité du couple (X, Y) , ou loi conjointe, c'est-à-dire $\Pr_{xy}$, est égale à la loi produit $\Pr_x \otimes \Pr_y$.

D'où *les propriétés* :

- la fonction de répartition $H(x, y)$ du couple (X, Y) est égale au produit des fonctions de répartition $F(x)$ et $G(y)$ de X et Y , ou *fonctions marginales du couple* :

$$H(x, y) = F(x) G(y)$$

- si les variables X et Y admettent des densités de probabilité f et g respectivement, la *densité du couple* (X, Y) est :

$$h(x, y) = f(x) g(y)$$

Ces deux conditions sont des conditions nécessaires et suffisantes d'indépendance de deux variables aléatoires.

4.7 Moments d'une variable aléatoire

4.7.1 Espérance mathématique

- *Variable aléatoire discrète*

X est une variable aléatoire réelle prenant un ensemble fini ou dénombrable de valeurs sur un espace probabilisé $(\Omega, \mathcal{C}, \Pr)$

$\{\alpha_i\} (i \in I)$ est l'ensemble des valeurs prises par X avec les probabilités

$$p_i = \Pr(X = \alpha_i)$$

L'espérance mathématique de X , notée $E(X)$ est définie par (si la série converge) :

$$E(X) = \sum_{i \in I} p_i \alpha_i$$

– Variable aléatoire continue

L'espérance mathématique de la variable aléatoire X , définie sur l'espace probabilisé $(\Omega, \mathcal{C}, \text{Pr})$, est donnée par l'intégrale, si elle converge :

$$E(X) = \int_{\Omega} X \, d\text{Pr} = \int_{\Omega} x \, \text{Pr}_x \, dx$$

que l'on peut écrire, si f est la densité de probabilité de X :

$$E(X) = \int_{\Omega} x f(x) \, dx$$

■ Propriétés de l'espérance mathématique

- X et Y sont deux variables aléatoires admettant chacune une espérance mathématique, a et b sont deux constantes :

$$E(a) = a$$

$$E(aX + b) = aE(X) + b$$

$$E(X + Y) = E(X) + E(Y)$$

- une variable aléatoire est dite *centrée* si son espérance mathématique est nulle :

$$E(X) = 0$$

Remarques

- L'espérance mathématique peut être considérée comme le centre de gravité de la distribution de masses Pr_x .
 – Il existe des variables aléatoires qui n'ont pas d'espérance mathématique.

Exemple 4.4

La variable aléatoire X suit la loi de Cauchy, de densité de probabilité :

$$f(x) = \frac{1}{\pi (1 + x^2)} \quad \forall x \in \mathbb{R}$$

L'intégrale

$$E(X) = \int_{\mathbb{R}} x \frac{1}{\pi (1 + x^2)} \, dx$$

ne converge pas, une variable de Cauchy n'a pas d'espérance mathématique.

■ Espérance mathématique d'une variable aléatoire Y , fonction d'une variable aléatoire X

Si φ est une fonction réelle (ou complexe) de Lebesgue mesurable, on définit l'espérance mathématique de la fonction $Y = \varphi \circ X$ par l'expression :

$$E(\varphi \circ X) = \int_{\Omega} \varphi(x) f(x) dx$$

■ Espérance mathématique du produit de deux variables aléatoires

Soient X et Y deux variables aléatoires de loi conjointe Pr_{XY} ou $h(x, y)$.

L'espérance mathématique de la variable aléatoire XY , produit des variables aléatoires X et Y , est donnée par l'intégrale, si elle existe :

$$E(XY) = \int_{D_{XY}} xy \, d\text{Pr}_{XY}(xy) = \int_{D_{XY}} xy \, h(x, y) \, dx dy$$

D_{XY} étant le domaine de variation du couple (X, Y) .

□ Cas de deux variables aléatoires indépendantes

Soient X et Y deux variables aléatoires indépendantes de densités respectives f et g . L'espérance mathématique du couple (X, Y) est :

$$\begin{aligned} E(XY) &= \int_{D_{XY}} xy f(x) g(y) \, dx dy \\ &= \int_{D_X} x f(x) \, dx \int_{D_Y} y g(y) \, dy = E(X) E(Y) \end{aligned}$$

Attention, la réciproque est fautive : la propriété $E(XY) = E(X)E(Y)$ n'entraîne pas l'indépendance des variables aléatoires X et Y .

4.7.2 Moments d'ordre supérieur à 1 d'une variable aléatoire

L'espérance mathématique $E(X)$ est le moment d'ordre 1 de la distribution, il apporte peu de renseignements sur cette variable. Les moments d'ordre supérieur à 1 (moments d'inertie) donnent des indications sur l'étalement de la distribution.

■ Définition des moments d'ordre k

Le *moment d'ordre k* de la variable aléatoire X , par rapport au point a , ou *espérance mathématique* de $(X - a)^k$, est donné par l'intégrale, si cette intégrale et la densité f existent :

$$E[(X - a)^k] = \int_{\Omega} (x - a)^k f(x) dx$$

On note, en général, m_k les moments $E(X^k)$

Les moments les plus utilisés sont :

- les moments autour de $a = E(X)$, appelés *moments centrés*, et notés μ_k en général.

Relations entre les moments m_k et les moments centrés μ_k :

$$m_0 = 1 \quad \mu_0 = 1$$

$$m_1 = m = E(X) \quad \mu_1 = 0$$

$$m_2 = \mu_2 + m^2 \quad \mu_2 = m_2 - m^2 = \text{Var}(X)$$

$$m_3 = \mu_3 + 3\mu_2 m + m^3 \quad \mu_3 = m_3 - 3m m_2 + 2m^3$$

$$m_4 = \mu_4 + 4\mu_3 m + 6\mu_2 m^2 + m^4 \quad \mu_4 = m_4 - 4m m_3 + 6m^2 m_2 - 3m^4$$

Ces relations se démontrent facilement.

- le *moment centré d'ordre 2* ou *variance* (ses propriétés sont étudiées dans le paragraphe 4.7.3).

■ Définition des moments absolus d'ordre k

Le moment absolu d'ordre k par rapport à un point a , est égal à, sous réserve de l'existence de l'intégrale :

$$E(|X - a|^k) = \int_{\Omega} |x - a|^k f(x) dx$$

4.7.3 Variance d'une variable aléatoire

■ Définition

La variance (moment centré d'ordre 2), ou carré de l'écart-type σ , est donnée par l'intégrale si cette intégrale et la densité f existent :

$$E[(X - E(X))^2] = \text{Var}(X) = \sigma^2 = \int_{\Omega} [x - E(X)]^2 f(x) dx$$

L'écart-type σ s'exprime avec la même unité que la variable.

La formule de König-Huyghens donne :

$$\begin{aligned} E[(X - a)^2] &= E[(X - E(X) + E(X) - a)^2] \\ &= E[(X - E(X))^2] - 2[E(X) - a][E[X - E(X)]] + [E(X) - a]^2 \end{aligned}$$

ou, comme $E[X - E(X)] = 0$,

$$E[(X - a)^2] = \text{Var}(X) + [E(X) - a]^2$$

La borne inférieure de $E[(X - a)^2]$ est obtenue pour $a = E(X)$, elle est égale à $\text{Var}(X)$.

La variance est le plus petit moment d'ordre deux d'une variable aléatoire.

■ Propriétés de la variance

– Pour $a = 0$, la formule précédente devient :

$$\text{Var}(X) = E(X^2) - [E(X)]^2$$

Cette formule peut être utilisée pour calculer rapidement $\text{Var}(X)$.

– Si a et b sont des constantes :

$$\text{Var}(b) = 0$$

$$\text{Var}(aX + b) = a^2 \text{Var}(X)$$

– $\text{Var}(X) = 0$ implique que X est presque sûrement égale à une constante.

– Si X est une variable aléatoire de carré intégrable, on définit une *variable aléatoire U centrée réduite*, associée à X , par la relation :

$$U = \frac{X - E(X)}{\sigma}$$

$$E(U) = 0 \quad \text{Var}(U) = 1$$

– *Variance d'une somme de variables aléatoires*

$$\begin{aligned} \text{Var}(X + Y) &= E[(X - E(X) + Y - E(Y))^2] \\ &= \text{Var}(X) + \text{Var}(Y) + 2E[(X - E(X))(Y - E(Y))] \\ &= E[(X - E(X)) \cdot (Y - E(Y))] \\ &= E(XY) - E[XE(Y)] - E[YE(X)] + E(X)E(Y) \\ &= E(XY) - E(X)E(Y) \end{aligned}$$

car $E(X)$ et $E(Y)$ sont des constantes.

La quantité $E[\{X - E(X)\} \{Y - E(Y)\}]$ est la *covariance* de X et Y :

$$\text{Cov}(X, Y) = E[\{X - E(X)\} \{Y - E(Y)\}] = E(XY) - E(X) E(Y)$$

La variance d'une somme de deux variables aléatoires est donc égale à :

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) + 2 \text{Cov}(X, Y)$$

Pour un couple de deux variables aléatoires, il existe trois moments différents d'ordre 2 qui sont :

$$\text{Var}(X), \text{Var}(Y) \text{ et } \text{Cov}(X, Y)$$

Cas particulier : la covariance de deux variables aléatoires indépendantes est égale à 0.

En effet : $E(XY) = E(X)E(Y)$.

On en déduit :

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y)$$

La réciproque n'est pas vraie en général, si la covariance de deux variables est nulle, ces deux variables ne sont pas indépendantes sauf si ce sont des variables gaussiennes (chapitre 8, paragraphe 8.4.3).

– *Généralisation : variance d'une somme algébrique de variables aléatoires*

$$\text{Var}\left(\sum_{i=1}^n a_i X_i\right) = \sum_{i=1}^n a_i^2 \text{Var}(X_i) + 2 \sum_{i \neq j, i < j} a_i a_j \text{Cov}(X_i, X_j)$$

les coefficients a_i étant des constantes.

Si les variables aléatoires X_i sont indépendantes, on obtient :

$$\text{Var}\left(\sum_{i=1}^n a_i X_i\right) = \sum_{i=1}^n a_i^2 \text{Var}(X_i)$$

4.7.4 Fractiles

Le *fractile d'ordre* α ($0 < \alpha < 1$) de la loi de X est le nombre x_α tel que :

$$\Pr(X < x_\alpha) = \alpha$$

– Si $\alpha = 1/2$, x_α est la *médiane*. La médiane M_e vérifie les deux inégalités :

$$\Pr(X < M_e) = 0,50 \quad \Pr(X > M_e) = 0,50$$

Il existe toujours au moins une valeur médiane, mais elle peut ne pas être unique, il peut même exister un segment de valeurs médianes.

- Si $\alpha = 0,25$, x_α est le *premier quartile* et pour $\alpha = 0,75$, le troisième quartile.
- Si $\alpha = k/10$ (k entier compris entre 1 et 9), les différentes valeurs de x_α définissent les *déciles*.
- Si $\alpha = k/100$ (k entier compris entre 1 et 99), les différentes valeurs de x_α définissent les *centiles*.

■ Application : inégalité de Bienaymé-Tchebyshev

Pour toute variable aléatoire réelle X de carré intégrable, définie sur un espace de probabilité $(\Omega, \mathcal{C}, \text{Pr})$ et pour tout réel quelconque k strictement positif, on a :

$$\text{Pr}(|X - E(X)| > k \sigma_X) \leq \frac{1}{k^2}$$

Pour démontrer cette inégalité, il suffit de revenir à la définition de la variance de X .

Autres formes de cette inégalité :

$$\text{Pr}(|X - E(X)| > k) \leq \frac{\sigma_X^2}{k^2}$$

$$\text{Pr}(|X - E(X)| < k \sigma_X) \geq 1 - \frac{1}{k^2}$$

Ces différentes formes permettent de comprendre la signification de l'écart-type.

L'écart-type caractérise la dispersion de la distribution autour de l'espérance mathématique.

- Pour $k = 10$, l'événement

$$|X - E(X)| \geq 10 \sigma_X$$

a peu de chances de se réaliser, en effet :

$$\text{Pr}(|X - E(X)| \geq 10 \sigma_X) \leq \frac{1}{100}$$

- Supposons $\sigma = 0$, alors :

$$\text{Pr}(|X - E(X)| > k) = 0$$

X est presque sûrement égale à $E(X)$.

L'inégalité de Bienaymé-Tchebyschev impose seulement l'existence des moments d'ordre 1 et 2. Comme elle ne fait pas intervenir la loi de probabilité suivie par la variable aléatoire considérée, elle peut s'appliquer à de nombreuses lois mais elle donne une majoration de la probabilité beaucoup trop grande.

On peut comparer les majorations qu'elle donne avec celles obtenues en considérant la loi exacte suivie par la variable aléatoire.

Exemple 4.5

Un échantillon d'effectif n doit être extrait d'une distribution de moyenne m et d'écart-type σ . On cherche le plus petit entier n vérifiant la condition :

$$\Pr \left(\left| \bar{X} - m \right| < \frac{\sigma}{4} \right) \geq 0,99$$

– On applique l'inégalité de Bienaymé-Tchebyschev à la variable $\bar{X}$ qui a pour moyenne m et pour variance σ^2/n :

$$\Pr \left(\left| \bar{X} - m \right| \leq \frac{k\sigma}{\sqrt{n}} \right) \geq 1 - \frac{1}{k^2}$$

$$1 - \frac{1}{k^2} = 0,99$$

$$\frac{k}{\sqrt{n}} = \frac{1}{4} \Rightarrow n = 1\,600$$

– En utilisant le théorème central limite (chapitre 6 paragraphe 6.6.6), on détermine une autre valeur pour n , meilleure que la première :

Théorème central limite :

La variable aléatoire $\frac{\bar{X} - m}{\sigma/\sqrt{n}}$ suit une loi normale centrée réduite quand n tend vers l'infini.

D'où :

$$\Pr \left(\left| \frac{\bar{X} - m}{\sigma/\sqrt{n}} \right| < 2,3263 \right) \geq 0,99 \quad \frac{\sigma}{\sqrt{n}} \times 2,3263 = \frac{\sigma}{4} \Rightarrow n \cong 86$$

5 • LOIS DE PROBABILITÉ DISCRÈTES

B

CALCUL DES PROBABILITÉS

Pour trouver un modèle décrivant un ensemble de données, il est nécessaire de connaître parfaitement les lois statistiques les plus utilisées. Le choix d'une loi est lié :

- à la nature du phénomène étudié afin de choisir entre loi discrète et loi continue,
- à la forme de la distribution (histogramme),
- à la connaissance et à l'interprétation des principales caractéristiques de l'ensemble de données : espérance, médiane, variance, écart-type, coefficients d'asymétrie et de dissymétrie, etc.,
- au nombre de paramètres des lois, une loi dépendant de plusieurs paramètres peut s'adapter plus facilement à une distribution.

5.1 Définition d'une variable discrète

Une *variable aléatoire discrète* prend ses valeurs sur un ensemble fini ou dénombrable de points. La loi de probabilité d'une telle variable est appelée *loi discrète*.

Une loi de probabilité discrète est caractérisée par l'énumération des valeurs x_i , appartenant à $\mathbb{R}$ ou à un intervalle de $\mathbb{R}$, prises par la variable aléatoire X et par les probabilités associées, c'est-à-dire les nombres réels positifs p_i tels que :

$$\Pr(X = x_i) = p_i \quad 0 \leq p_i \leq 1 \quad \sum_i p_i = 1$$

La *fonction de répartition* est une fonction en escalier, constante sur tout intervalle $[x_i, x_{i+1}[$, admettant en chaque point x_i un saut égal à $p_{i+1} = \Pr(X = x_{i+1})$.

5.1.1 Moments

$$E(X) = \sum_i x_i p_i \quad \text{Var}(X) = \sum_i x_i^2 p_i - [E(X)]^2$$

5.1.2 Domaine d'utilisation

Les lois discrètes sont utilisées pour modéliser les résultats des jeux de hasard, les sondages d'opinion, les phénomènes biologiques, les processus aléatoires (files d'attente, évolution de l'état de matériels)... Les plus utilisées sont la loi uniforme, la loi binomiale et les lois dérivées, la loi hypergéométrique, la loi de Poisson.

Exemple 5.1

Soit X la variable aléatoire prenant trois valeurs 0, 1, 2 avec les probabilités :

$$\Pr(X = 0) = 1/2 \quad \Pr(X = 1) = 1/3 \quad \Pr(X = 2) = 1/6$$

$$(1/2 + 1/3 + 1/6 = 1)$$

Espérance mathématique :

$$E(X) = 0 \times 1/2 + 1 \times 1/3 + 2 \times 1/6 = 2/3$$

Variance :

$$\text{Var}(X) = (\sigma_x)^2 = E(X^2) - [E(X)]^2 = 1 \times 1/3 + 4 \times 1/6 - (2/3)^2 = 5/9$$

Les figures 5.1 et 5.2 représentent l'histogramme et la fonction de répartition de la loi donnée dans l'exemple 5.1.

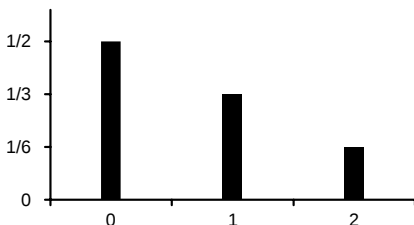


Figure 5.1 – Histogramme de la loi donnée dans l'exemple 5.1.

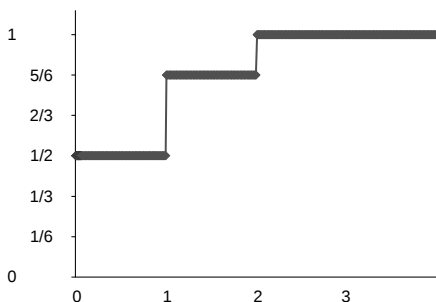


Figure 5.2 – Fonction de répartition de la loi donnée dans l'exemple 5.1.

5.2 Loi de Dirac

5.2.1 Définition

La loi de Dirac au point a de $\mathbb{R}$ est la loi de probabilité δ_a , définie sur $\mathbb{R}$ par :

$$\delta_a(x) = 1 \quad \text{si } x = a \quad \delta_a(x) = 0 \quad \text{si } x \neq a$$

Cette loi est la loi la plus simple, associée à un phénomène déterministe X dont le résultat de toute expérience est égale à a .

5.2.2 Moments

$$E(X) = a \quad \text{Var}(X) = 0$$

5.2.3 Généralisation

Soit A un ensemble de n nombres réels distincts a_i et n nombres réels p_i tels que :

$$0 \leq p_i \leq 1 \quad \sum_{i=1}^n p_i = 1$$

La combinaison linéaire $\sum_{i=1}^n p_i \delta_{a_i}$ définit une loi de probabilité associée à une variable aléatoire discrète X telle que pour tout indice i :

$$\Pr(X = a_i) = p_i$$

Si, pour toutes les valeurs de l'indice i , $p_i = 1/n$, la loi de la variable aléatoire X est la *loi de probabilité uniforme sur A* .

5.3 Loi uniforme

5.3.1 Définition

La *loi uniforme* sur $[1, n]$ est la loi de probabilité d'une variable aléatoire X prenant chaque valeur de l'ensemble $(1, 2 \dots n)$ avec la même probabilité :

$$\Pr(X = k) = 1/n \text{ pour tout entier } k \text{ compris entre } 1 \text{ et } n$$

Plus généralement, soit Ω un ensemble fini de cardinal n . La *loi de probabilité équilibrée* ou *uniforme* sur Ω est la loi définie sur Ω par la probabilité :

$$\Pr(\omega) = 1/n \text{ pour tout élément } \omega \text{ de } \Omega$$

Pour toute partie finie A de Ω , on a :

$$\Pr(A) = \frac{\text{card}(A)}{\text{card}(\Omega)}$$

5.3.2 Moments

$$E(X) = \frac{1}{n} \sum_{k=1}^n k = \frac{n+1}{2}$$

$$\begin{aligned} \text{Var}(X) &= \frac{1}{n} \sum_{k=1}^n k^2 - \left(\frac{n+1}{2} \right)^2 \\ &= \frac{1}{n} \times \frac{n(n+1)(2n+1)}{6} - \frac{(n+1)^2}{4} = \frac{n^2-1}{12} \end{aligned}$$

5.3.3 Domaine d'utilisation

La loi de probabilité uniforme intervient dans de nombreux domaines comme les jeux de pile ou face ou les jeux de dés (avec une pièce ou un dé parfaitement équilibré(e)), les jeux de cartes, les loteries, les sondages...

5.4 Loi binomiale ou loi des tirages avec remise

5.4.1 Définition et propriétés d'une variable de Bernoulli

On considère une expérience dont le résultat ne peut prendre que deux valeurs appelées, par convention, *succès* ou *échec* : un candidat est reçu ou non à un examen, une pièce usinée est bonne ou défectueuse, une porte est ouverte ou fermée...

À une expérience de ce type, est associée une variable aléatoire X prenant la valeur 1 pour le succès et la valeur 0 pour l'échec, avec les probabilités respectives p et $(1 - p) = q$. Cette variable est appelée *variable de Bernoulli*¹.

La *loi de probabilité* d'une variable de Bernoulli est définie par :

$$\Pr(X = 1) = p$$

$$\Pr(X = 0) = 1 - p = q$$

Ses *moments* sont :

$$E(X) = p$$

$$\text{Var}(X) = p(1 - p) = pq$$

Domaine d'utilisation : elle est utilisée pour modéliser des matériels qui seront soit survivants (valeur 1), soit défectueux (valeur 0) à un instant donné.

Elle s'applique aux jeux de hasard de type binaire comme pile ou face...

5.4.2 Définition d'une variable binomiale

On réalise n épreuves indépendantes de la même expérience telles que :

- chaque épreuve ne peut avoir que deux résultats, s'excluant mutuellement, soit le succès, soit l'échec,
- la probabilité p de succès est constante à chaque épreuve, la probabilité d'échec est également constante et égale à $1 - p = q$.

1. Jacques Bernoulli, mathématicien suisse (1654-1705).

■ Probabilité d'obtenir k succès au cours de ces n épreuves

Soit X la variable aléatoire qui « compte » le nombre de succès au cours de n épreuves.

- Si au cours de n épreuves, on obtient k succès, on obtient également $(n - k)$ échecs. La probabilité de réalisation d'un tel événement est égale à $p^k(1 - p)^{n-k}$ (les épreuves sont indépendantes).
- Il y a différentes façons d'obtenir k succès au cours de n épreuves, chaque épreuve pouvant être, indépendamment les unes des autres, un succès ou un échec. Le nombre de réalisations possibles de l'événement « obtenir k succès au cours de n épreuves » est le nombre de combinaisons sans répétitions de n objets pris k à k , soit :

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

D'où :

$$\Pr(X = k) = C_n^k p^k (1 - p)^{n-k} = \frac{n!}{k! (n - k)!} p^k (1 - p)^{n-k}$$

Cette expression étant un terme du développement du binôme $[p + (1 - p)]^n$, la variable X est appelée *variable binomiale*. On vérifie facilement que :

$$\sum_{k=0}^n \Pr(X = k) = 1$$

La loi de la variable X est appelée *loi binomiale de paramètres (n, p)* , notée $B(n; p)$.

Une variable binomiale est égale à la somme de n variables aléatoires indépendantes de Bernoulli, la loi binomiale est donc la *loi des épreuves répétées*. Elle est également la loi *d'un tirage sans remise* dans une urne contenant des boules de deux types différents.

5.4.3 Propriétés de la loi binomiale

■ Histogramme

La loi binomiale étant une loi discrète, son histogramme est en bâtons. La hauteur des bâtons, proportionnelle à la quantité $\Pr(X = k)$, croît de façon

monotone, puis décroît également de façon monotone. En effet :

$$\frac{\Pr(X = k)}{\Pr(X = k-1)} = \frac{C_n^k p^k (1-p)^{n-k}}{C_n^{k-1} p^{k-1} (1-p)^{n-k+1}} = \frac{n-k+1}{k} \times \frac{p}{1-p}$$

$\Pr(X = k) \geq \Pr(X = k-1)$ si $(n-k+1)p \geq k(1-p)$ ou $k \leq p(n+1)$

La probabilité $\Pr(X = k)$ croît si k varie de 0 à une valeur k' égale à la partie entière de $p(n+1)$, puis décroît si k varie de k' à n .

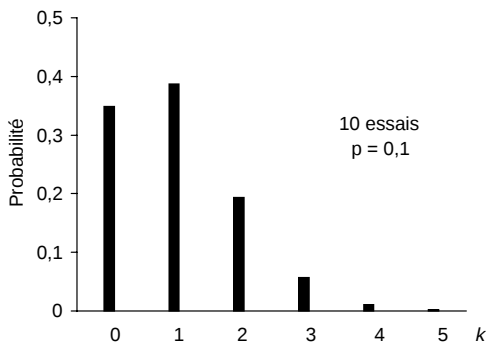


Figure 5.3 – Histogramme de la loi $B(10; 0,10)$.

B

CALCUL DES PROBABILITÉS

■ Tables de la loi binomiale

La table 1 donne les probabilités individuelles, $\Pr(X = k)$, et la table 2, les probabilités cumulées $\sum_{i=0}^k \Pr(X = i)$, pour toutes les valeurs de i et k et pour des valeurs du nombre n d'épreuves et de la probabilité p .

■ Moments

Une variable binomiale, suivant la loi $B(n; p)$, peut être considérée comme la somme de n variables indépendantes de Bernoulli. D'où les résultats :

$$E(X) = np \quad \text{Var}(X) = np(1-p) = npq$$

■ Coefficients d'asymétrie et d'aplatissement

$$\gamma_1 = \frac{1-2p}{\sqrt{npq}} \quad \gamma_2 = 3 + \frac{1-6p}{npq}$$

■ Domaine d'utilisation

- La loi binomiale décrit des phénomènes ne pouvant prendre que deux états s'excluant mutuellement, succès ou échec dans un jeu, bonne pièce ou pièce défectueuse dans une fabrication, lot acceptable ou lot refusé, défaillance ou fonctionnement d'un matériel...
- Elle est utilisée dans le domaine technique pour déterminer la probabilité de défaillance à la sollicitation de matériels, en contrôle qualité, mais elle ne peut s'appliquer rigoureusement que si les expériences sont non exhaustives, c'est la loi du *tirage avec remise*.
- Les événements considérés doivent être indépendants et la probabilité de réalisation d'un événement doit être constante.

■ Somme de variables aléatoires, binomiales, indépendantes et de même paramètre

De la définition résulte la propriété suivante :

La somme de n variables aléatoires binomiales indépendantes et de même paramètre p est une variable aléatoire binomiale.

$$\left. \begin{array}{l} \text{Loi de la variable } X : B(n_1; p) \\ \text{Loi de la variable } Y : B(n_2; p) \\ X \text{ et } Y \text{ indépendantes} \end{array} \right\} \text{loi de } S = X + Y : B(n_1 + n_2; p)$$

Exemple 5.2

On veut réaliser une étude clinique sur des malades se présentant à une consultation hospitalière. Pour cette étude, seuls les malades répondant à un ensemble de critères C sont retenus. Des statistiques antérieures ont montré que 20 % des consultants présentent les critères C .

10 malades viennent consulter le premier jour.

Soit X la variable aléatoire « nombre de malades retenus » c'est-à-dire répondant à l'ensemble des critères C . La variable X suit la loi binomiale $B(10; 0,20)$.

La probabilité qu'aucun malade ne soit recruté ce jour est égale à :

$$\Pr(X = 0) = C_{10}^0 (0,20)^0 (0,80)^{10} = 0,107$$

La probabilité pour qu'il y ait au moins un malade recruté est égale à :

$$\Pr(X \geq 1) = 1 - \Pr(X = 0) = 1 - 0,107 = 0,893$$

5.4.4 Lois voisines de la loi binomiale

■ Loi géométrique

□ Définition

La loi géométrique est la loi de la variable Y « loi du nombre d'essais nécessaires » pour qu'un événement de probabilité p apparaisse pour la première fois, les hypothèses étant les mêmes que pour la loi binomiale, en particulier, la probabilité p reste constante au cours des essais :

$$\Pr(Y = k) = p(1 - p)^{k-1} \quad k \in \mathbb{N}^*$$

(Il y a eu $k - 1$ échecs avant d'obtenir le succès au $k^{\text{ème}}$ essai).

□ Moments

$$E(Y) = \frac{1}{p} \quad \text{Var}(Y) = \frac{1-p}{p^2}$$

□ Coefficients d'asymétrie et d'aplatissement

$$\gamma_1 = \frac{2-p}{\sqrt{1-p}} \quad \gamma_2 = 3 + \frac{(p-3)^2 - 3}{1-p}$$

Exemple 5.3

Un certain matériel a une probabilité $p = 0,02$ constante de défaillance à chaque mise en service. On procède à l'expérience suivante, l'appareil est mis en marche, arrêté, remis en marche, arrêté, jusqu'à ce qu'il tombe en panne. Le nombre d'essais nécessaires pour obtenir la panne est une variable aléatoire suivant la loi géométrique de paramètre p . La probabilité que ce matériel tombe en panne (pour la première fois) au dixième essai est égale à :

$$\Pr(Y = 10) = (0,02)(1 - 0,02)^9 = 0,0167$$

■ Loi de Pascal

□ Définition

La loi de Pascal est la loi de la variable Z « loi du nombre d'essais nécessaires » pour obtenir exactement k fois un événement de probabilité p , les hypothèses étant les mêmes que pour la loi binomiale (la probabilité p est constante au cours des essais).

$$\Pr(Z = z) = p C_{z-1}^{k-1} p^{k-1} (1-p)^{z-k} \quad z \geq k \quad z \in \mathbb{N}^*$$

(On a obtenu un succès à l'essai $n^\circ z$ (de probabilité p) et $k-1$ succès au cours des $z-1$ essais précédents.)

□ Moments

$$E(Z) = \frac{k}{p} \quad \text{Var}(Z) = \frac{k(1-p)}{p^2}$$

□ Coefficients d'asymétrie et d'aplatissement

$$\gamma_1 = \frac{2-p}{\sqrt{k(1-p)}} \quad \gamma_2 = 3 + \frac{p^2 + 6(1-p)}{k(1-p)}$$

■ Loi binomiale négative

□ Définition

À partir de la loi de Pascal, on définit la *loi binomiale négative*, ou loi de la variable aléatoire $T = Z - k$:

$$\Pr(T = t) = C_{k+t-1}^{k-1} p^k (1-p)^t = C_{k+t-1}^t p^k (1-p)^t$$

□ Moments, coefficients d'asymétrie et d'aplatissement

Si on pose $P = (1-p)/p$ et $Q = 1/p$, on obtient :

$$E(T) = kP \quad \text{Var}(T) = kPQ \quad \gamma_1 = \frac{P+Q}{\sqrt{kPQ}} \quad \gamma_2 = 3 + \frac{1+6PQ}{kPQ}$$

valeurs que l'on peut comparer à celles de la loi binomiale $B(k; p)$.

■ Comparaison des lois binomiale et de Pascal

□ Loi binomiale

- Elle compte le nombre de succès au cours de n épreuves.
- Le nombre n d'épreuves est fixé.
- Le nombre de succès est une variable aléatoire pouvant prendre toutes les valeurs entières entre 0 et n .

□ Loi de Pascal

- Elle compte le nombre d'essais nécessaires pour obtenir k succès.
- Le nombre k de succès est fixé.
- Le nombre d'épreuves est une variable aléatoire pouvant prendre toutes les valeurs entières entre k et l'infini.
- Elle compte le nombre d'essais nécessaires pour obtenir k succès.

5.5 Loi multinomiale

5.5.1 Définition

La loi multinomiale est une généralisation de la loi binomiale.

Une population P est composée d'individus appartenant à k types différents, dans des proportions $p_1, p_2, \dots, p_k$ telles que $\sum_{i=1}^k p_i = 1$.

On tire un échantillon de n individus, de façon équiprobable et indépendante et on s'intéresse à la composition de l'échantillon.

Soit X_i la variable aléatoire représentant le nombre d'individus de type i dans l'échantillon. Par définition, le vecteur $\underline{X} = (X_1, \dots, X_k)$ est un vecteur aléatoire suivant une *loi multinomiale de paramètres* $(n; p_1, \dots, p_k)$, notée $M(n; p_1, \dots, p_k)$.

$$\begin{aligned} \Pr[X = (x_1, \dots, x_k)] &= C_n^{x_1} C_{n-x_1}^{x_2} \dots C_{n-(x_1+\dots+x_{k-1})}^{x_k} p_1^{x_1} \dots p_k^{x_k} \\ &= \frac{n!}{x_1! \dots x_k!} p_1^{x_1} \dots p_k^{x_k} \\ \sum_{i=1}^k p_i &= 1 \quad \sum_{i=1}^k x_i = n \end{aligned}$$

Le coefficient $\frac{n!}{x_1! \dots x_k!}$ est le nombre de partitions d'un échantillon de taille n en sous-populations d'effectifs x_i (voir annexe 1).

5.5.2 Propriétés

La loi marginale de X_i est une loi binomiale $B(n; p_i)$. En effet, un élément tiré est :

- soit du type i avec la probabilité p_i ,
- soit de n'importe quel autre type avec la probabilité $(1 - p_i)$.

Le nombre d'individus du type i dans l'échantillon suit donc la loi binomiale $B(n; p_i)$. D'où :

$$E(X_i) = np_i \quad \text{Var}(X_i) = np_i(1 - p_i)$$

Les variables X_i et X_k ne sont pas indépendantes.

Le couple (X_i, X_k) suit une loi multinomiale de dimension 3. En effet, un élément tiré est :

- soit du type i (probabilité p_i),
- soit du type k (probabilité p_k),
- soit de n'importe quel autre type (probabilité $1 - p_i - p_k$).

En partant de ces propriétés, on démontre que :

$$E(X_i X_k) = n(n-1) p_i p_k$$

$$\text{Cov}(X_i, X_k) = E(X_i X_k) - E(X_i) E(X_k) = -n p_i p_k$$

Les variables X_i et X_k ne peuvent donc pas être indépendantes.

5.5.3 Domaine d'utilisation

La loi multinomiale est utilisée en statistique.

Soit X une variable aléatoire continue de densité $f(x)$. On suppose que l'espace D_X des valeurs prises par cette variable est partagé en k classes distinctes C_i d'extrémités e_{i-1} et e_i , par exemple tranches d'âges, de revenus, d'impôts...

On considère un échantillon $(X_1, \dots, X_n)$ de n observations de cette variable et on cherche le nombre de points N_i de l'échantillon dans la classe C_i .

Le vecteur $(N_1, \dots, N_k)$ suit la loi multinomiale de paramètres $(n, p_1, \dots, p_k)$ avec :

$$p_i = \int_{e_i-1}^{e_i} f(x) dx$$

Connaissant p_i pour chaque valeur de i , on en déduit la composition de l'échantillon. C'est la méthode qui peut être utilisée, par exemple, pour construire un histogramme.

Exemple 5.4

Un produit d'éclairage de l'entreprise M peut présenter des défauts regroupés en trois catégories : défaut critique, défaut majeur, défaut mineur.

Un contrôle final est effectué une semaine après la sortie du produit pour vérifier si certaines défauts se seraient développés au cours de cette période. Le résultat du contrôle est le suivant :

- 80 % du produit ne présente aucune défaut (ensemble E_1),
- 10 % du produit présente des défauts mineurs (ensemble E_2),
- 6 % du produit présente des défauts majeurs (ensemble E_3),
- 4 % du produit présente des défauts critiques (ensemble E_4).

Un échantillon de taille $n = 20$ est prélevé au hasard dans un grand lot et vérifié selon les critères précédents.

Soit X_i le nombre d'unités appartenant au sous-ensemble E_i dans l'échantillon contrôlé. L'ensemble (X_1, X_2, X_3, X_4) suit une loi multinomiale qui a pour paramètres les pourcentages donnés par le contrôle. D'où la probabilité :

$$\begin{aligned} \Pr(X_1 = x_1, X_2 = x_2, X_3 = x_3, X_4 = x_4) \\ &= C_{20}^{x_1} C_{20-x_1}^{x_2} C_{20-(x_1+x_2)}^{x_3} C_{20-(x_1+x_2+x_3)}^{x_4} (0,80)^{x_1} (0,10)^{x_2} (0,06)^{x_3} (0,04)^{x_4} \\ &= \frac{20!}{x_1! x_2! x_3! x_4!} (0,80)^{x_1} (0,10)^{x_2} (0,06)^{x_3} (0,04)^{x_4} \end{aligned}$$

Espérance mathématique des variables X_i :

$$\begin{aligned} E(X_1) &= 0,80 \times 20 = 16 & E(X_2) &= 0,10 \times 20 = 2 \\ E(X_3) &= 0,06 \times 20 = 1,2 & E(X_4) &= 0,04 \times 20 = 0,8 \end{aligned}$$

On peut calculer différentes probabilités :

$$\Pr(X_1 = 10, X_2 = 6, X_3 = 3, X_4 = 1) = 0,0001439$$

$$\Pr(X_1 = 20) = 0,0115292$$

$$\Pr(X_1 = 15, X_2 = 5) = 0,0054549$$

5.6 Loi hypergéométrique ou loi du tirage exhaustif

5.6.1 Définition

On considère un tirage *équiprobable sans remise* ou tirage *exhaustif* dans une population d'effectif N , cette population étant composée de deux parties disjointes :

- une partie A à Np éléments (éléments possédant un certain caractère),
- une partie B à $(N - Np)$ éléments (éléments n'ayant pas ce caractère).

C'est le cas, par exemple, d'un lot de N pièces comprenant Np pièces défectueuses et donc $(N - Np)$ pièces fonctionnant bien.

Quelle est la probabilité qu'un sous-ensemble de n éléments contienne x éléments de l'ensemble A ?

Les éléments peuvent être tirés, soit un par un, soit d'un seul coup, mais *sans remise*.

Soit X la variable aléatoire représentant le nombre d'individus ayant le caractère considéré dans l'échantillon. On veut calculer $\Pr(X = x)$.

Le nombre d'échantillons de taille n est égal à C_N^n

Un échantillon de taille n comprend :

- x individus, pris parmi les Np individus ayant le caractère considéré, donc C_{Np}^x choix possibles,
- $(n - x)$ individus pris parmi les $(N - Np)$ individus n'ayant pas ce caractère, donc $C_{N - Np}^{n - x}$ choix possibles.

Le nombre d'échantillons de taille n , comprenant x individus pris parmi les Np individus, est donc égal à $C_{Np}^x C_{N - Np}^{n - x}$

La probabilité cherchée est égale à :

$$\Pr(X = x) = \frac{C_N^x C_{N-N}^{n-x}}{C_N^n}$$

Les valeurs extrêmes de x sont :

$$\min x = \max\{0, n - N(1 - p)\} \quad \max x = \min\{n, Np\}$$

Le quotient n/N est appelé *taux de sondage*.

La variable aléatoire ainsi définie suit *une loi hypergéométrique* $H(N ; n ; p)$.

Cette loi dépend de trois paramètres N , n et p .

5.6.2 Moments

Une variable aléatoire X suivant une loi hypergéométrique $H(N ; n ; p)$ peut être considérée comme une somme de n variables aléatoires de Bernoulli, $X_1 \dots X_n$, non indépendantes, correspondant aux tirages successifs de n individus, tirage sans remise. On en déduit le calcul des moments d'ordre 1 et 2.

□ Espérance mathématique

– La variable aléatoire X_1 correspond au tirage du premier individu :

$$\Pr(X_1 = 1) = p$$

$$\Pr(X_1 = 0) = 1 - p$$

d'où :

$$E(X_1) = p \quad \text{Var}(X_1) = p(1 - p)$$

La variable aléatoire X_2 correspond au tirage du 2^e individu :

$$\Pr(X_2 = 1) = \Pr(X_2 = 1/X_1 = 1) \Pr(X_1 = 1) + \Pr(X_2 = 1/X_1 = 0) \Pr(X_1 = 0)$$

$$\Pr(X_2 = 1) = \frac{Np - 1}{N - 1} p + \frac{Np}{N - 1} (1 - p) = p$$

d'où :

$$\Pr(X_2 = 1) = p$$

$$\Pr(X_2 = 0) = 1 - p$$

On en déduit :

$$\begin{aligned}E(X_2) &= p \\ \text{Var}(X_2) &= p(1-p)\end{aligned}$$

De la même façon, pour toutes les variables aléatoires X_i , on trouve :

$$\begin{aligned}E(X_i) &= p \\ \text{Var}(X_i) &= p(1-p)\end{aligned}$$

On en déduit que :

$$E(X) = \sum_{i=1}^n E(X_i) = np$$

Cette espérance est indépendante de la taille N de la population.

□ Variance

Des calculs analogues conduisent au résultat suivant :

$$\begin{aligned}\text{Var}(X) &= \frac{N-n}{N-1} np(1-p) \\ \frac{N-n}{N-1} &\text{ est le } \textit{facteur d'exhaustivité}.\end{aligned}$$

Remarques

- Les variables aléatoires X_i sont des variables aléatoires de Bernoulli non indépendantes.
- Le facteur d'exhaustivité tend vers 1 si N est grand devant n et la variance de la variable X tend alors vers la variance d'une variable suivant la loi binomiale $B(n; p)$.

5.6.3 Domaine d'utilisation

La loi hypergéométrique est utilisée, en particulier dans les contrôles de qualité où on retire, de la population étudiée, les éléments défectueux.

Exemple 5.5

Dans une assemblée de 30 personnes, il y a 20 hommes et 10 femmes. On tire un échantillon de 15 personnes (tirage sans remise).

Soit X la variable aléatoire « nombre d'hommes » dans cet échantillon.

Valeurs extrêmes de cette variable :

- 5 (la valeur 0 ne peut pas être obtenue car il n'y a pas 15 femmes dans l'assemblée),
- 15 (dans l'échantillon, il n'y aura pas de femmes).

La probabilité d'avoir 10 hommes dans un échantillon de taille 15 est égale à :

$$\Pr(X = 10) = \frac{C_{20}^{10} C_{10}^5}{C_{30}^{15}} = 0,30$$

5.7 Loi de Poisson

5.7.1 Définition

La loi de Poisson de paramètre λ est la loi d'une variable aléatoire discrète réelle X , prenant toutes les valeurs entières non négatives, avec les probabilités :

$$\Pr(X = k) = p_k = \frac{e^{-\lambda} \lambda^k}{k!} \quad k \in [0, +\infty[\quad \sum_{k=0}^{\infty} p_k = 1$$

La loi de Poisson dépend d'un seul paramètre λ . On la note $P(\lambda)$.

La table 3 donne les probabilités individuelles, $\Pr(X = k)$, et la table 4, les probabilités cumulées, $\sum_{i=1}^k \Pr(X = i)$, pour des valeurs du paramètre λ compris entre 0,1 et 18.

5.7.2 Moments

$$E(X) = \text{Var}(X) = \lambda$$

La loi de Poisson est la seule loi discrète possédant la propriété, $E(X) = \text{Var}(X)$

5.7.3 Propriétés et domaine d'utilisation

- Grâce aux propriétés des fonctions caractéristiques (chapitre 7, paragraphe 7.2.1), on démontre que :

La somme de deux variables aléatoires de Poisson, indépendantes, de paramètres λ_1 et λ_2 , est une variable aléatoire de Poisson, de paramètre $\lambda = \lambda_1 + \lambda_2$.

- La loi de Poisson est la loi discrète d'une variable aléatoire représentant un nombre d'événements. Elle est utilisée pour décrire :
 - la réalisation d'événements peu probables, dans une succession d'épreuves très nombreuses, au moins 50,
 - le nombre d'accidents dans un atelier, le nombre de défauts sur un appareil,Elle a des applications dans le domaine des files d'attente (chapitre 9). Elle est la loi limite de la loi binomiale quand n tend vers l'infini et p tend vers zéro, le produit np restant fini (paragraphe 5.8.2).

La loi de Poisson est la *loi des événements rares* ou *loi des petites probabilités*.

Exemple 5.6

Selon les données recueillies depuis plusieurs années, le nombre de pannes hebdomadaires du système informatique d'une entreprise suit une loi de Poisson de paramètre $\lambda = 0,05$.

Soit X la variable aléatoire « nombre de pannes hebdomadaires » :

$$\Pr(X = k) = \frac{e^{-0,05} (0,05)^k}{k!}$$

La probabilité que le système tombe en panne une fois au cours d'une semaine quelconque ($k = 1$) est égale à 0,04756.

La probabilité qu'il fonctionne sans panne ($k = 0$) est égale à 0,95122.

On considère une année (50 semaines) de fonctionnement de ce système. Le nombre de pannes Y obéit à une loi de Poisson de paramètre $\mu = 0,05 \times 50 = 2,5$.

$$\Pr(Y = k) = \frac{e^{-2,5} (2,5)^k}{k!}$$

La probabilité d'observer 2 pannes au cours de l'année ($k = 2$) est égale à 0,2565 et la probabilité d'en observer 4 est égale à 0,1336.

5.8 Lois limites

La loi hypergéométrique dépend de trois paramètres, la loi binomiale de deux et la loi de Poisson d'un seul, les calculs pour la première loi sont donc plus longs que pour la deuxième et que pour la troisième. Il est intéressant de chercher dans quelles conditions, on peut utiliser la loi binomiale comme

approximation de la loi hypergéométrique puis la loi de Poisson comme approximation de la loi binomiale.

Cependant, comme les domaines de définition et d'utilisation de ces trois lois sont différents, il faut définir avec précision dans quelles conditions ces approximations seront valables.

5.8.1 Approximation d'une loi hypergéométrique par une loi binomiale

La variable aléatoire X suit la loi hypergéométrique $H(N; n; p)$. On suppose que l'effectif N de la population devient très grand par rapport à la taille n de l'échantillon prélevé. On cherche, dans ces conditions, la limite de l'expression

$$\Pr(X = x) = \frac{C_{Np}^x C_{N-Np}^{n-x}}{C_N^n}$$

Après avoir développé les différents termes et utilisé la condition n petit devant N , on obtient comme limite :

$$\Pr(X = x) \rightarrow C_n^x p^x (1-p)^{n-x}$$

On reconnaît dans cette limite, l'expression de $\Pr(X = x)$ d'une variable aléatoire X suivant la loi binomiale $B(n; p)$; cette approximation est valable dès que $n < 0,1N$.

Exemple 5.7

Dans une population de 1 000 hommes adultes, 50 ont un poids supérieur à 120 kg. On considère un échantillon de taille 100 et soit X la variable aléatoire « nombre d'individus ayant un poids supérieur à 120 kg » dans cet échantillon.

– La loi suivie par cette variable est la loi hypergéométrique de paramètres $N = 1\,000$, $n = 100$ et $p = 50/1\,000 = 0,05$:

$$\Pr(X = k) = \frac{C_{50}^k C_{950}^{100-k}}{C_{1\,000}^{100}} \quad 0 \leq k \leq 100$$

$$E(X) = 5 \text{ et } \text{Var}(X) = 4,28$$

– La taille de l'échantillon (100) est égale à $1\,000/10$. On peut donc utiliser l'approximation par la loi binomiale de paramètres $n = 100$ et $p = 0,05$:

$$\Pr(X = k) = C_{100}^k (0,05)^k (0,95)^{100-k} \quad 0 \leq k \leq 100$$

$$E(X) = 5 \text{ et } \text{Var}(X) = 4,75$$

5.8.2 Approximation d'une loi binomiale par une loi de Poisson

Soit X une variable aléatoire suivant la loi binomiale $B(n; p)$. On suppose :

- que le nombre n d'essais est grand,
- que la probabilité de succès p est petite,
- que le produit np reste fini.

On pose $np = \lambda$ et on cherche la limite de l'expression :

$$\Pr(X = x) = C_n^x p^x (1 - p)^{n-x}$$

On obtient (en utilisant la même méthode que dans le cas précédent) :

$$\Pr(X = x) \rightarrow \frac{e^{-\lambda} \lambda^x}{x!}$$

On reconnaît l'expression de $\Pr(X = x)$ pour une variable aléatoire X suivant une loi de Poisson $P(\lambda)$, avec $\lambda = np$.

En pratique, cette approximation est valable si $n > 50$ et $p < 0,1$.

Exemple 5.8

Un atelier fabrique un grand nombre d'objets. On admet que la probabilité qu'un objet soit défectueux est égale à 1/100. Combien doit-on contrôler de pièces pour avoir 95 chances sur 100 d'obtenir au moins une pièce défectueuse ?

Soit X la variable aléatoire « nombre de pièces défectueuses » dans un échantillon de taille n . On veut déterminer le nombre n tel que :

$$\Pr(X \geq 1) = 0,95 \quad \text{ou} \quad \Pr(X = 0) = 0,05$$

La loi exacte suivie par la variable X est la loi binomiale $B(n; p = 0,01)$. D'où la condition sur n :

$$\Pr(X = 0) = (1 - 0,01)^n = (0,99)^n = 0,05$$

$n \ln(0,99) = \ln(0,05)$ et $n = 299$ (valeur entière approchée).

Approximation : la taille de l'échantillon est inconnue, la probabilité p est petite. On peut remplacer la loi binomiale par une loi de Poisson de paramètre $\lambda = np$, en justifiant, *a posteriori*, la validité de cette approximation (taille de l'échantillon).

$$\Pr(X = 0) = e^{-\lambda} = 0,05 \quad \lambda = 2,9957$$

n est peu différent de 300. L'approximation est justifiée.

Exemple 5.9 (suite de l'exemple 5.7)

La probabilité p est inférieure à 0,10, la taille de l'échantillon est supérieure à 50, on peut utiliser l'approximation par une loi de Poisson de paramètre $np = 100 \times 0,05 = 5$

$$\Pr(X = k) = \frac{e^{-5} (5)^k}{k!} \quad k \in 0, 1, 2 \dots$$

$$E(X) = \text{Var}(X) = 5$$

Avec cet exemple, on peut comparer pour les trois lois, hypergéométrique, binomiale et de Poisson :

- les espérances mathématiques : elles sont égales ;
- les variances : la variance obtenue par la loi hypergéométrique est plus faible que les deux autres qui sont égales, elle est divisée par le facteur d'exhaustivité qui est inférieure à 1 ;
- les domaines de variation : ils sont différents.

B**CALCUL DES PROBABILITÉS**

5.9 Résumé

5.9.1 Première approximation

$$H(N; n; p) \rightarrow B(n; p) \quad \text{si } n < 0,10 N$$

Domaine de variation de la variable :

- loi hypergéométrique :

$$\min x = \max\{0, n - N(1 - p)\} \quad \text{et} \quad \max x = \min\{n, Np\}$$

- loi binomiale : toutes les valeurs entières entre 0 et n .

5.9.2 Deuxième approximation

$$B(n; p) \rightarrow P(\lambda) \quad \text{si } n \rightarrow \infty \quad p \rightarrow 0 \text{ et } np \rightarrow \lambda$$

Domaine de variation de la variable :

- loi binomiale : toutes les valeurs entières entre 0 et n ,
- loi de Poisson : toutes les valeurs entières positives ou nulle.

Tableau 5.1 – Caractéristiques des principales lois discrètes.

Loi	Formule	Application	Approximation
Hypergéométrique	$\Pr(X = x) = \frac{C_{Np}^x C_{N-Np}^{n-x}}{C_N^n}$	Tirage sans remise	Loi binomiale si $n < 0,1N$
Binomiale	$\Pr(X = x) = C_n^x p^x (1-p)^{n-x}$	Tirage avec remise Probabilité constante	Loi de Poisson si $n > 50$ et $p < 0,10$
Poisson	$\Pr(X = k) = \frac{e^{-\lambda} \lambda^k}{k!}$	Événements instantanés et indépendants	

6 • LOIS DE PROBABILITÉ CONTINUES

B

CALCUL DES PROBABILITÉS

6.1 Généralités

Une variable aléatoire continue prend ses valeurs sur un ensemble infini non dénombrable de points, elle décrit par exemple la durée de vie d'une batterie de voiture, l'heure d'arrivée des voitures à un péage donné d'autoroute...

Il existe une fonction f non négative, définie pour toute valeur x appartenant à $\mathbb{R}$ et vérifiant, pour toute partie A de $\mathbb{R}$, la propriété :

$$\Pr(X \in A) = \int_A f(x) \, dx$$

et telle que :

$$\int_{\mathbb{R}} f(x) \, dx = 1$$

La fonction f est la *densité de probabilité* de la variable aléatoire X .

La *fonction de répartition* de la variable aléatoire X est définie par :

$$F(a) = \Pr(X < a) = \int_{-\infty}^a f(x) \, dx$$

Pour toutes les valeurs a et b appartenant à $\mathbb{R}$, on a donc la relation :

$$\Pr(a \leq X < b) = F(b) - F(a)$$

On en déduit :

$$\Pr(X = x) = 0 \quad \Pr(x \leq X < x + dx) = f(x) \, dx$$

Espérance mathématique :

$$E(X) = \int_{\mathbb{R}} x f(x) \, dx$$

Variance :

$$\text{Var}(X) = E(X^2) - [E(X)]^2 = \int_{\mathbb{R}} x^2 f(x) dx - [E(X)]^2$$

L'espérance et la variance existent si les intégrales sont définies.

La plupart des problèmes rencontrés en statistique peuvent se résoudre à l'aide de quelques lois fondamentales continues, environ une dizaine.

Les principales sont la loi uniforme, la loi exponentielle, les lois gamma, les lois bêta, la loi normale et la loi log-normale auxquelles il faut ajouter les lois du chi-deux, de Fisher-Snedecor et de Student utilisées dans la théorie de l'estimation (lois étudiées, chapitre 10, paragraphes 10.4, 10.5 et 10.6).

6.2 Loi uniforme

6.2.1 Définition

Une variable aléatoire réelle X , suit une *loi uniforme sur l'intervalle* $[a, b]$, si sa loi de probabilité admet *une densité f égale à* :

$$f(x) = \frac{1}{b-a} 1_{[a, b]}$$

$1_{[a, b]}$ est la fonction caractéristique du segment $[a, b]$.

Fonction de répartition :

$$\begin{aligned} F(x) &= 0 & \text{si } x \leq a \\ F(x) &= (x-a)/(b-a) & \text{si } a < x < b \\ F(x) &= 1 & \text{si } x \geq b \end{aligned}$$

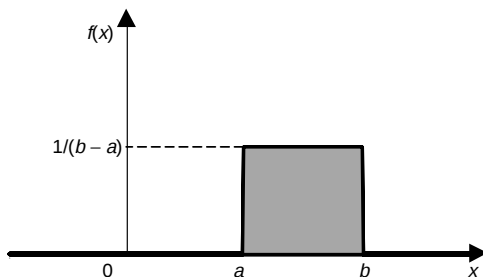
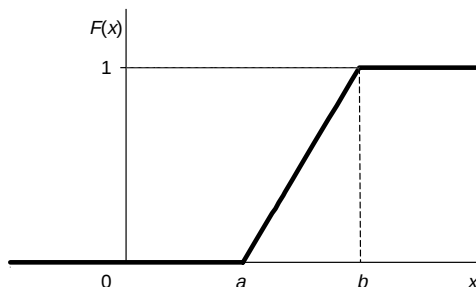


Figure 6.1 – Loi uniforme sur $[a, b]$. Densité.

Figure 6.2 – Loi uniforme sur $[a, b]$. Fonction de répartition.

6.2.2 Moments

$$E(X) = \frac{(b+a)}{2} \quad \text{Var}(X) = \frac{(b-a)^2}{12}$$

6.2.3 Propriétés et domaine d'utilisation

- La somme de deux variables aléatoires, indépendantes ou non, suivant une loi uniforme sur $[a, b]$, ne suit pas une loi uniforme sur $[a, b]$.
- L'image, par sa fonction de répartition, de toute variable aléatoire réelle continue, est une variable aléatoire réelle suivant la loi uniforme sur $[0, 1]$. Cette propriété est utilisée, pour simuler ou engendrer, des échantillons de la loi de la variable aléatoire X (chapitre 11, paragraphe 11.5.4).

Démonstration de cette propriété

Soit X une variable aléatoire dont la fonction de répartition F est continue et strictement croissante.

On considère la variable aléatoire $Y = F(X)$, elle varie de 0 à 1. On désigne par G et g sa fonction de répartition et sa densité :

$$G(y) = \Pr(Y < y) = \Pr(F(X) < y) = \Pr(X < F^{-1}(y)) = F[F^{-1}(y)] = y$$

En résumé :

Si $y \in [-\infty, 0]$

$G(y) = 0$ et $g(y) = 0$

Si $y \in [0, 1]$

$G(y) = y$ et $g(y) = 1$

Si $y \in [1, +\infty]$

$G(y) = 1$ et $g(y) = 0$

On reconnaît la fonction de répartition et la densité d'une variable suivant une loi uniforme sur $[0, 1]$.

- La loi uniforme sur $[a, b]$ traduit l'hypothèse d'équirépartition, ou répartition indifférente, sur $[a, b]$. Les tables, concernant ce type de répartition, sont les tables de nombres au hasard (chapitre 11, paragraphe 11.5.3).
- La loi uniforme est utilisée en statistique bayésienne, pour déterminer les lois de probabilité a priori, dans le cas de l'ignorance totale, dans l'intervalle $[0, 1]$ (elle est dite non informative) ou dans l'intervalle $[a, b]$, en utilisant les résultats de l'expert (elle est dite informative).

6.3 Loi exponentielle

6.3.1 Définition

Une variable aléatoire réelle positive X suit une *loi exponentielle*, de paramètre λ positif, si sa densité de probabilité est donnée par :

$$\begin{aligned} f(x) &= \lambda e^{-\lambda x} & \text{si } x \geq 0 \\ f(x) &= 0 & \text{sinon} \end{aligned}$$

X est appelée *variable exponentielle*.

Fonction de répartition :

$$F(a) = \Pr(X < a) = \int_0^a \lambda e^{-\lambda x} dx = 1 - e^{-\lambda a}$$

6.3.2 Moments

Espérance et variance :

$$E(X) = \frac{1}{\lambda} \quad \text{Var}(X) = \frac{1}{\lambda^2}$$

L'espérance d'une variable exponentielle est égale à son écart-type.

Coefficients d'asymétrie et d'aplatissement :

$$\gamma_1 = 2 \quad \gamma_2 = 9$$

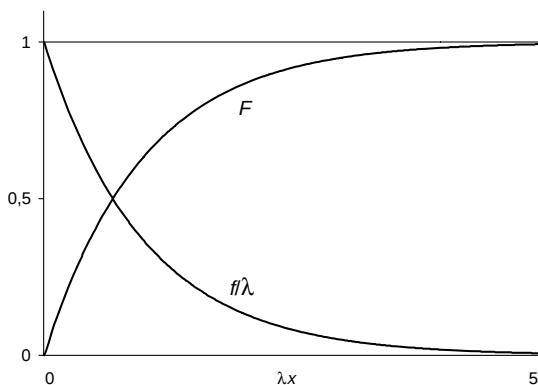


Figure 6.3 – Loi exponentielle. Densité et fonction de répartition de la loi exponentielle (l'axe des abscisses est gradué proportionnellement aux valeurs λx).

B

CALCUL DES PROBABILITÉS

6.3.3 Domaine d'utilisation

- La distribution exponentielle est associée aux processus de Poisson. Un tel processus génère des événements dont les temps d'occurrence sont indépendants et distribués suivant une loi exponentielle (chapitre 9).
- La loi exponentielle est utilisée en *fiabilité* (chapitre 18), le paramètre λ représente le taux de défaillance alors que son inverse $\theta = 1/\lambda$ est le temps moyen de bon fonctionnement MTBF (*Mean Time Between Failure*). Avec le paramètre θ , la densité de probabilité s'écrit :

$$f(x) = \frac{1}{\theta} e^{-\frac{x}{\theta}}$$

et les moments sont égaux à :

$$E(X) = \theta$$

$$\text{Var}(X) = \theta^2$$

- La loi exponentielle s'applique bien aux matériels électroniques, c'est-à-dire aux matériels fonctionnant pratiquement sans usure, aux matériels subissant des défaillances brutales ou à des systèmes complexes dont les

composants ont des lois de fiabilité différentes. Elle permet de décrire la période de fonctionnement durant laquelle le taux de défaillance est constant ou presque constant.

6.3.4 Propriétés

- La somme de deux variables aléatoires indépendantes, suivant des lois exponentielles de paramètres respectifs λ_1 et λ_2 , est une variable aléatoire suivant une loi exponentielle de paramètre $\lambda_1 + \lambda_2$.
- La loi exponentielle est qualifiée de loi « sans mémoire », elle permet la modélisation du comportement des matériels fonctionnant avec un taux de défaillance constant (ou pouvant être considéré comme constant).
- On considère un matériel ayant fonctionné sans défaillance pendant le temps x_1 et on cherche la probabilité qu'il soit encore en état de marche au temps $x + x_1$. La définition de la probabilité conditionnelle donne :

$$\begin{aligned}\Pr(X \geq x + x_1 / X \geq x_1) &= \frac{\Pr(X \geq x + x_1 \text{ et } X \geq x_1)}{\Pr(X \geq x_1)} \\ &= \frac{e^{-\lambda(x+x_1)}}{e^{-\lambda x_1}} = e^{-\lambda x}\end{aligned}$$

Le matériel a « oublié » qu'il avait déjà fonctionné pendant le temps x_1 . Pour ce type de matériel, il est inutile de procéder à un remplacement préventif.

Exemple 6.1

On suppose que le temps, en heures, nécessaire pour réparer une machine est une variable aléatoire suivant une loi exponentielle de paramètre $\lambda = 0,5$.

La densité de probabilité est $f(t) = 0,5 e^{-0,5t}$ et la fonction de répartition $F(t) = 1 - e^{-0,5t}$.

La probabilité pour que le temps de réparation dépasse 2 heures est :

$$\Pr(T > 2) = 1 - \Pr(T < 2) = 1 - F(2) = e^{-1} = 0,368$$

Sachant que la réparation a déjà dépassé 9 heures, quelle est la probabilité qu'elle prenne au moins 10 heures ?

La loi exponentielle étant une loi sans « mémoire », on obtient :

$$\Pr(T > 10 / T > 9) = \Pr(T > 10 - 9 = 1) = e^{-0,5} = 0,606$$

6.4 Loi gamma

6.4.1 Définition

La loi exponentielle est un cas particulier de la famille des lois gamma.

Une variable aléatoire réelle positive X suit une *loi gamma* $\gamma(t; \lambda)$ ou $\Gamma(t; \lambda)$, de paramètres positifs t et λ , si sa densité de probabilité est donnée par :

$$f(x) = \frac{\lambda e^{-\lambda x} (\lambda x)^{t-1}}{\Gamma(t)} \quad \text{si } x \geq 0$$

$$f(x) = 0 \quad \text{sinon}$$

Γ est la *fonction eulérienne* définie par l'intégrale pour $t > 0$ (voir annexe 2) :

$$\Gamma(t) = \int_0^{\infty} e^{-y} y^{t-1} dy$$

Le paramètre t est un paramètre de forme tandis que $1/\lambda$ est un paramètre d'échelle. Pour les représentations graphiques, on peut prendre $\lambda = 1$.

Selon les valeurs du paramètre t , la densité de la loi gamma a différentes formes (figures 6.4 et 6.5).

En particulier, si $t = 1$, on retrouve la loi exponentielle.

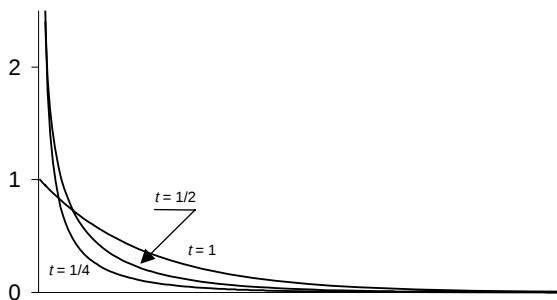
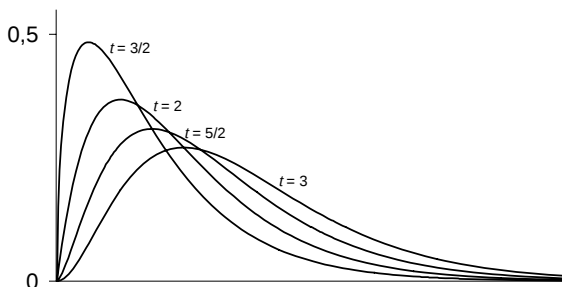


Figure 6.4 – Loi gamma ($t \leq 1$).

Figure 6.5 – Loi gamma ($t > 1$).

Si le paramètre λ est différent de 1, la variable aléatoire $Y = \lambda X$ suit une loi $\gamma(t; 1)$ ou $\gamma(t)$ de densité :

$$f(y) = \frac{e^{-y} y^{t-1}}{\Gamma(t)} \quad \text{si } y \geq 0$$

$$f(y) = 0 \quad \text{sinon}$$

6.4.2 Moments

Par intégrations par parties et en utilisant les propriétés de la fonction Γ , on obtient :

$$E(X) = \frac{t}{\lambda} \quad \text{Var}(X) = \frac{t}{\lambda^2}$$

6.4.3 Propriétés et domaine d'utilisation

La loi gamma dépendant de deux paramètres peut être utilisée pour représenter un grand nombre de distributions. Ainsi :

- dans la théorie des files d'attente, la loi gamma représente la loi de probabilité d'occurrence de t événements (t étant un entier), dans un processus poissonnien. Si le temps T , entre les défaillances successives d'un système, suit une loi exponentielle, le temps cumulé d'apparitions de λ défaillances suit une loi gamma $\gamma(t; \lambda)$,
- en fiabilité, la loi gamma peut être utilisée pour modéliser les temps de défaillance d'un matériel,

– selon les valeurs des paramètres, la loi gamma s'identifie à d'autres lois :

- la loi exponentielle si $t = 1$,
- la loi d'Erlang si t est égal à un entier n supérieur à 1, sa densité est :

$$f(x) = \frac{\lambda e^{-\lambda x} (\lambda x)^{n-1}}{(n-1)!} \quad \text{si } x \geq 0$$

$$f(x) = 0 \quad \text{sinon}$$

En effet, $\Gamma(n) = (n-1)!$ et sa fonction de répartition, qui correspond à l'apparition de n événements en un temps inférieur à x , est donnée par l'expression :

$$F(x) = 1 - e^{-\lambda x} \sum_{i=1}^n \frac{(\lambda x)^{i-1}}{(i-1)!} \quad \text{si } x \geq 0$$

$$F(x) = 0 \quad \text{sinon}$$

- la loi de la variable chi-deux à n degrés de liberté, $\chi^2(n)$, utilisée en statistique, si $\lambda = 1/2$ et $t = n/2$, où n est un entier positif (chapitre 10, paragraphe 10.4),
- la somme de deux variables aléatoires indépendantes, suivant des lois gamma $\gamma(t; \lambda)$ et $\gamma(u; \lambda)$, suit une loi gamma $\gamma(t + u; \lambda)$ (propriété d'additivité des lois gamma).

B

CALCUL DES PROBABILITÉS

6.5 Lois bêta de types I et II

6.5.1 Définitions

■ Loi bêta de type I

Une variable aléatoire réelle X , prenant ses valeurs dans l'intervalle $[0, 1]$, suit une loi bêta de type I, notée $\beta(n; p)$, de paramètres positifs n et p , si sa densité de probabilité est donnée par :

$$f(x) = \frac{1}{B(n; p)} x^{n-1} (1-x)^{p-1} \quad 0 \leq x \leq 1$$

$$f(x) = 0 \quad \text{sinon}$$

où $B(n; p)$ est la fonction eulérienne définie par (voir annexe 2) :

$$B(n; p) = \int_0^1 x^{n-1} (1-x)^{p-1} dx = B(p; n)$$

$$B(n; p) = \frac{\Gamma(n) \Gamma(p)}{\Gamma(n+p)}$$

La forme de la densité de X varie selon la valeur des paramètres n et p .

■ Loi bêta de type II

Soit X une variable aléatoire suivant une loi bêta de type I, $\beta(n; p)$. La variable aléatoire Y , positive ou nulle, définie par $Y = \frac{X}{1-X}$, suit une *loi bêta de type II* dont la densité s'obtient facilement par changement de variables :

$$f(y) = \frac{1}{B(n, p)} \frac{y^{n-1}}{(1+y)^{n+p}} \quad \text{si } y \geq 0$$

$$f(y) = 0 \quad \text{sinon}$$

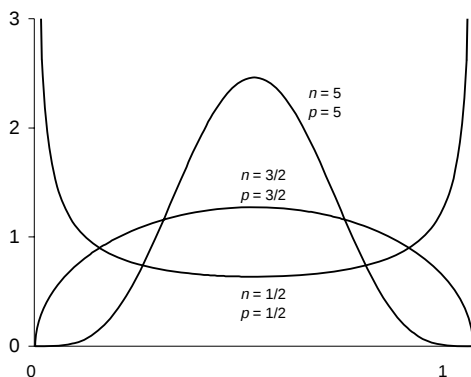


Figure 6.6 – Densité de la loi bêta II, paramètres égaux.

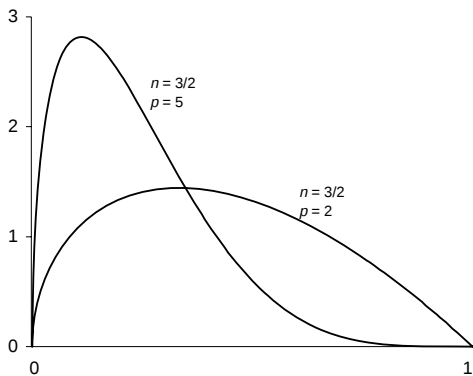


Figure 6.7 – Densité de la loi bêta II, paramètres différents et supérieurs à 1.

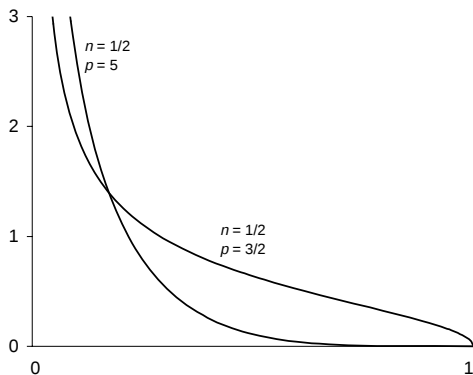


Figure 6.8 – Densité de la loi bêta II, paramètres différents dont un est inférieur à 1.

B

CALCUL DES PROBABILITÉS

6.5.2 Moments

■ Loi bêta de type I

$$E(X) = \frac{n}{n+p} \quad \text{Var}(X) = \frac{np}{(n+p+1)(n+p)^2}$$

■ Loi bêta de type II

$$E(X) = \frac{n}{p-1} \quad \text{Var}(X) = \frac{n(n+p-1)}{(p-1)^2(p-2)^2}$$

6.5.3 Propriétés et domaines d'utilisation

- Le rapport de deux variables aléatoires indépendantes, suivant les lois gamma $\gamma(t; \lambda)$ et $\gamma(u; \lambda)$, suit une loi bêta de type II de paramètres t et u .
- Les lois bêta, dépendant de deux paramètres, s'adaptent bien à la description de nombreux phénomènes aléatoires positifs (temps d'attente, durées de vie...); elles sont liées aux lois de Fisher-Snedecor utilisées en statistique.
- Les lois bêta de type I sont utilisées en fiabilité, en statistique bayésienne pour représenter la distribution *a priori* de la probabilité d'un événement suivant une loi binomiale, la distribution *a posteriori* suit aussi une loi binomiale.

6.6 Loi de Laplace-Gauss ou loi normale

6.6.1 Définition

Une variable aléatoire réelle X , prenant ses valeurs dans $\mathbb{R}$, suit une *loi de Laplace-Gauss ou loi normale*, de paramètres m et σ , si sa densité de probabilité est donnée par :

$$f(x) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(x-m)^2}{2\sigma^2}}$$

La fonction f définit une densité. En effet :

$$\int_{-\infty}^{+\infty} f(x) \, dx = 1$$

Cette loi est notée, en général $N(m; \sigma)$. On dit indifféremment qu'une variable suivant une telle loi est *une variable normale ou gaussienne*.

Fonction de répartition :

$$\Pr(X < a) = \frac{1}{\sigma \sqrt{2\pi}} \int_{-\infty}^a e^{-\frac{(x-m)^2}{2\sigma^2}} \, dx$$

Cette intégrale n'ayant pas d'expression mathématique simple, des tables donnent les valeurs de la fonction de répartition.

Sur la courbe représentant la densité de probabilité d'une variable gaussienne, la valeur de $F(a)$ est représentée par la partie non hachurée. Cette courbe a un axe de symétrie vertical pour $x = m$ et du fait de sa forme, elle est souvent appelée « courbe en cloche ».

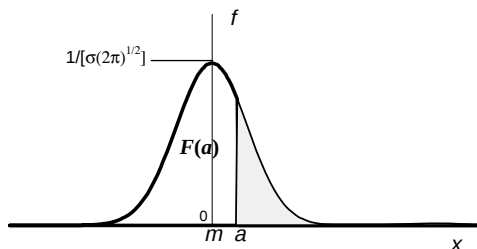


Figure 6.9 – Densité de la loi normale.

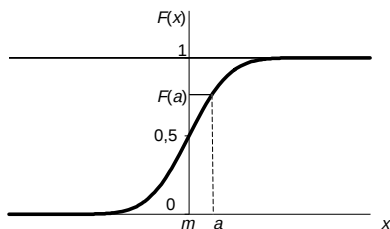


Figure 6.10 – Fonction de répartition de la loi normale.

6.6.2 Moments

Espérance et variance :

$$E(X) = m \quad \text{Var}(X) = \sigma^2$$

Ces résultats justifient le choix des deux paramètres figurant dans l'expression de la densité.

- Les *moments de tous les ordres existent*. En effet, les intégrales :

$$\frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{+\infty} x^k e^{-\frac{(x-m)^2}{2\sigma^2}} dx$$

convergent pour toutes les valeurs de k .

- Les *moments centrés d'ordre impair* sont tous nuls (propriété de symétrie de la densité).
- Les *moments centrés d'ordre pair* ont pour valeurs :

$$\mu_{2k} = 1 \times 3 \times \dots \times (2k-1) \sigma^{2k} = \frac{(2k)!}{2^k k!} \sigma^{2k}$$

- $\mu_3 = 0$, le coefficient d'asymétrie γ_1 est nul,
- $\mu_4 = 3\sigma^4$, le coefficient d'aplatissement γ_2 est égal à 3.

6.6.3 Variable aléatoire centrée réduite

La *variable centrée réduite* associée à la variable aléatoire X est la variable :

$$U = \frac{X - m}{\sigma}$$

Ses moments d'ordre impair sont nuls, en particulier $E(U) = 0$ et les moments d'ordre pair sont égaux à :

$$\mu_{2k} = 1 \times 3 \times \dots \times (2k-1) = \frac{(2k)!}{2^k k!}$$

en particulier $\text{Var}(U) = 1$.

Densité de probabilité de la variable U :

$$f(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}}$$

La variable U suit la loi normale $N(0; 1)$ dont les paramètres sont $m = 0$ et $\sigma = 1$.

Fonction de répartition :

$$\Pr(X < a) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^a e^{-\frac{u^2}{2}} du$$

La table 5.1 donne la fonction de répartition, et la table 5.2 les fractiles de la loi normale réduite car elle ne dépend d'aucun paramètre, les formules de

changement de variables :

$$U = \frac{X - m}{\sigma} \text{ et } X = \sigma U + m$$

permettant de passer d'une variable à l'autre.

Exemple 6.2 Utilisation de la table de la loi normale

Soit X une variable suivant la loi normale $N(3; 2)$, donc de moyenne 3 et d'écart-type 2. On veut calculer les probabilités suivantes : $\Pr(X < 4)$, $\Pr(X < -1)$, $\Pr(X > 1)$ ou les nombres a_i tels que $\Pr(X < a_1) = 0,75$, $\Pr(X > a_2) = 0,85$. On utilise la variable centrée réduite U associée à la variable X :

$$U = \frac{X - 3}{2} \text{ et } X = 2U + 3$$

$$\Pr(X < 4) = \Pr(2U + 3 < 4) = \Pr(U < 0,50) = 0,6915$$

$$\Pr(X < -1) = \Pr(2U + 3 < -1) = \Pr(U < -2) = 0,0228$$

$$\Pr(X > 1) = \Pr(2U + 3 > 1) = \Pr(U > -1) = 0,8413$$

$$\Pr(X < a_1) = \Pr(2U + 3 < a_1) = \Pr\left(U < \frac{a_1 - 3}{2}\right) = 0,75$$

$$\Pr(U < 0,6745) = 0,75 \quad \text{D'où : } a_1 = 4,35$$

$$\Pr(X > a_2) = \Pr(2U + 3 > a_2) = \Pr\left(U > \frac{a_2 - 3}{2}\right) = 0,85$$

$$\Pr(U < -1,0364) = 0,15 \quad \Pr(U > -1,0364) = 0,85$$

$$\text{D'où : } a_2 = -1,0364 \times 2 + 3 = 0,9272$$

Résultats remarquables :

$$\Pr(m - 1,64\sigma < X < m + 1,64\sigma) = 0,90$$

$$\Pr(m - 1,96\sigma < X < m + 1,96\sigma) = 0,95$$

$$\Pr(m - 3,09\sigma < X < m + 3,09\sigma) = 0,998$$

Loi de la variable U^2 : la densité de probabilité g de la variable aléatoire $T = U^2$ est obtenue en utilisant la formule donnant la densité de probabilité

d'une variable aléatoire Y fonction d'une variable X , d'où :

$$g(t) = \frac{1}{\sqrt{2\pi}} e^{-\frac{t}{2}} t^{-\frac{1}{2}} \quad t \geq 0$$

$$g(t) = 0 \quad \text{sinon}$$

La densité de probabilité h de la variable aléatoire $Z = T/2 = U^2/2$ est obtenue de façon analogue :

$$h(z) = \frac{1}{\sqrt{\pi}} e^{-z} z^{-\frac{1}{2}} \quad z \geq 0$$

$$h(z) = 0 \quad \text{sinon}$$

La variable aléatoire $Z = U^2/2$ suit donc une loi $\gamma(1/2 ; 1)$.

6.6.4 Domaine d'utilisation

- La loi normale est une des lois de probabilité la plus utilisée. Elle dépend de deux paramètres, la moyenne m , paramètre de position, et l'écart-type σ , paramètre mesurant la dispersion de la variable aléatoire autour de sa moyenne.
- Elle s'applique à de nombreux phénomènes, en physique, en économie (erreurs de mesure). De plus, elle est la *forme limite de nombreuses distributions discrètes*.
- Elle représente la loi de distribution d'une variable aléatoire X dépendant d'un grand nombre de *facteurs agissant sous forme additive*, chacun ayant une variance faible par rapport à la variance résultante.
- Elle peut représenter la fin de vie des dispositifs subissant un phénomène de vieillissement, usure, corrosion...

Remarque

Les variables utilisées dans le domaine technologique ou économique sont en général positives. La loi normale pourra représenter un tel phénomène si la probabilité d'obtenir des valeurs négatives de la variable est très faible. Il faut, en particulier, éviter de l'utiliser pour les queues des distributions.

6.6.5 Propriété d'additivité

On démontre le résultat suivant grâce aux propriétés des fonctions caractéristiques (chapitre 7, paragraphe 7.2) :

$$\left. \begin{array}{l} \text{Loi de la variable } X : N(m_1 ; \sigma_1) \\ \text{Loi de la variable } Y : N(m_2 ; \sigma_2) \\ X \text{ et } Y \text{ variables indépendantes} \end{array} \right\} \begin{array}{l} \text{loi de la somme } S = X + Y \\ N\left(m_1 + m_2 ; \sqrt{\sigma_1^2 + \sigma_2^2}\right) \end{array}$$

Ce résultat se généralise facilement à la somme de n variables aléatoires gaussiennes, indépendantes.

■ Cas particulier

La somme de deux variables aléatoires gaussiennes, centrées, réduites est une variable gaussienne centrée non réduite :

$$\left. \begin{array}{l} \text{Loi de la variable } X : N(0 ; 1) \\ \text{Loi de la variable } Y : N(0 ; 1) \\ X \text{ et } Y \text{ variables indépendantes} \end{array} \right\} \begin{array}{l} \text{loi de la somme } S = X + Y \\ N(0 ; \sqrt{2}) \end{array}$$

■ Application

Soit $(X_1 \dots X_n)$ un échantillon de n observations indépendantes, issu d'une population suivant la loi $N(m ; \sigma)$. La variable aléatoire $\bar{X}$, moyenne de l'échantillon :

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

est une combinaison linéaire de n variables aléatoires indépendantes suivant la même loi $N(m ; \sigma)$, elle suit donc une loi normale dont les paramètres sont :

$$E(\bar{X}) = m \quad \text{Var}(\bar{X}) = \frac{\sigma^2}{n}$$

La variable aléatoire $\bar{X}$ suit la loi $N(m ; \sigma/\sqrt{n})$.

6.6.6 Théorème central limite

La distribution normale a été introduite par le mathématicien français De Moivre en 1733 ; il l'utilisa comme approximation de la loi binomiale $B(n; p)$ pour n grand.

Ce résultat fut ensuite généralisé par Laplace et par d'autres mathématiciens pour devenir *le théorème central limite* ou *théorème de la limite centrale* qui donne les conditions dans lesquelles une variable aléatoire tend vers une variable normale. La version la plus simple du théorème central limite est la suivante :

Soit (X_n) , $n \geq 1$, une suite de n variables aléatoires indépendantes, de même loi de probabilité, d'espérance mathématique m et de variance σ^2 . On considère la variable aléatoire Y_n définie par :

$$Y_n = \frac{\sum_{i=1}^n X_i - n m}{\sigma \sqrt{n}} = \frac{\frac{1}{n} \sum_{i=1}^n X_i - m}{\sigma / \sqrt{n}}$$

La loi de la variable aléatoire Y_n converge vers la loi $N(0; 1)$ quand n tend vers l'infini.

La démonstration de ce théorème fait appel aux propriétés de la fonction caractéristique d'une distribution (chapitre 7, paragraphe 7.2).

■ Première application : approximation de la loi binomiale

Soit X la variable aléatoire, nombre de succès lors de la réalisation de n épreuves indépendantes, la probabilité de succès pour chaque épreuve étant égale à p . La loi de la variable aléatoire :

$$\frac{X - np}{\sqrt{np(1-p)}}$$

tend vers la loi $N(0; 1)$ quand n tend vers l'infini.

Une forme équivalente de ce résultat est :

La loi de la variable X tend vers la loi normale $N\left(np; \sqrt{np(1-p)}\right)$

En pratique, cette approximation est valable dès que les quantités np et $n(1-p)$ sont supérieures à 5.

Deux remarques importantes doivent être faites :

- une variable aléatoire binomiale est une variable discrète prenant ses valeurs dans l'intervalle $[0, n]$ alors qu'une variable aléatoire gaussienne est une variable continue prenant ses valeurs dans $\mathbb{R}$;
- dans le cas d'une loi binomiale, un point a une mesure ou une probabilité non nulle alors que dans le cas d'une loi normale, un point est un ensemble de mesure nulle.

Pour ces deux raisons, on doit faire une *correction de continuité* quand on utilise l'approximation d'une loi binomiale par une loi normale.

La représentation graphique de la densité de probabilité d'une variable gaussienne est une courbe en « cloche » tandis que celle d'une variable aléatoire binomiale est un diagramme en bâtons.

Une valeur approchée de $h = \Pr(X = k)$ est donnée par l'aire comprise entre la courbe en cloche et les droites d'abscisses $k - 0,5$ et $k + 0,5$; les deux aires curvilignes pouvant être considérées comme égales, cet aire est égale à celle du rectangle de base AB (égale à 1) et de hauteur h égale à $\Pr(X = k)$.

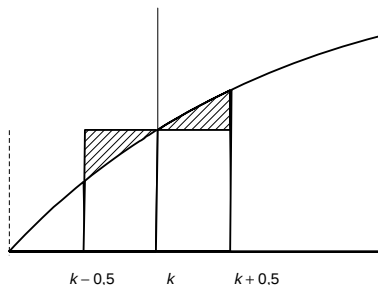


Figure 6.11 – Correction de continuité. Les aires hachurées sont approximativement égales.

Si U est la variable aléatoire normale centrée réduite, on obtient :

$$\Pr(X = k) \cong \Pr\left(\frac{k - 0,5 - np}{\sqrt{np(1-p)}} < U < \frac{k + 0,5 - np}{\sqrt{np(1-p)}}\right)$$

et de la même façon :

$$\Pr(X \leq k) \cong \Pr\left(U < \frac{k + 0,5 - np}{\sqrt{np(1-p)}}\right)$$

Exemple 6.3

Un contrôle au calibre, effectué depuis plusieurs mois sur le diamètre des pièces usinées par une machine outil, indique que le pourcentage de pièces « défectueuses » est égal à 8 %.

Un échantillon de 100 pièces de la production est prélevé et le diamètre de ces pièces est vérifié. Soit X la variable aléatoire « nombre de pièces défectueuses » dans un échantillon de 100 pièces.

La variable X suit la loi binomiale $B(100 ; 0,08)$.

Comme, $np = 100 \times 0,08 = 8 > 5$ et $n(1-p) = 100 \times 0,92 = 92 > 5$, on peut utiliser l'approximation par la loi normale $N(8 ; 2,713)$. Les paramètres sont en effet : la moyenne $m = np = 8$ et la variance est égale à $np(1-p) = 7,36 = (2,713)^2$.

En utilisant cette approximation, on peut calculer, par exemple, la probabilité d'avoir au moins 10 pièces classées défectueuses dans un échantillon de 100 pièces, soit $\Pr(X \geq 10)$ qui devient avec la correction de continuité $\Pr(X > 9,5)$:

$$\Pr(X > 9,5) = \Pr\left(\frac{X - 8}{2,713} > \frac{9,5 - 8}{2,713} = 0,552\right) = 0,2903$$

■ Deuxième application : approximation de la loi de Poisson

Le théorème central limite donne dans ce cas :

Soit X une variable aléatoire suivant la loi de Poisson $P(\lambda)$, la variable aléatoire Y définie par :

$$Y = \frac{X - \lambda}{\sqrt{\lambda}}$$

converge vers la loi $N(0 ; 1)$ quand λ tend vers l'infini.

L'approximation est satisfaisante si le paramètre λ est supérieur à 18. Comme pour la loi binomiale, on doit faire une correction de continuité.

Exemple 6.4

Dans un essai de sérologie préventive d'une maladie, 2 000 enfants ont été partagés en deux groupes de 1 000 ; les enfants d'un groupe recevaient un sérum, les autres ne recevaient rien. On a observé 40 cas de maladies dans le premier groupe et 50 dans le deuxième.

Les lois du nombre de maladies N_1 et N_2 dans chaque groupe sont des lois discrètes, le nombre de maladies est faible 40 et 50 pour un effectif de 1 000 individus. Ce sont donc des événements rares de probabilité 0,04 et 0,05, les variables N_1 et N_2 suivent donc des lois de Poisson $P(40)$ et $P(50)$. Ces lois peuvent être approchées par les lois normales $N(40, \sqrt{40})$ et $N(50, \sqrt{50})$.

■ Résumé**□ Première approximation**

$$B(n; p) \rightarrow N(np; \sqrt{np(1-p)}) \quad \text{si } np > 5 \quad \text{et } n(1-p) > 5$$

Domaine de variation de la variable :

- loi binomiale : toutes les valeurs entières entre 0 et n ,
- loi normale : toutes les valeurs réelles.

□ Deuxième approximation

$$P(\lambda) \rightarrow N(\lambda; \sqrt{\lambda}) \quad \text{si } \lambda > 18$$

Domaine de variation de la variable :

- loi de Poisson : toutes les valeurs entières positives ou nulle,
- loi normale : toutes les valeurs réelles.

B

CALCUL DES PROBABILITÉS

6.7 Loi log-normale

6.7.1 Définition

Soit Y une variable aléatoire suivant la loi $N(m; \sigma)$. La variable aléatoire X définie par $X = e^Y$ suit, par définition, une *loi log-normale*. Cette loi est aussi appelée *loi de Galton* ou *loi de Gibrat*.

La densité de la loi de la variable X se déduit de celle de la variable Y par le changement de variable $x \rightarrow e^y$:

$$f(x) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{(\ln x - m)^2}{2\sigma^2}} \quad x \geq 0$$

$$f(x) = 0 \quad \text{sinon}$$

Le facteur $1/x$ dans l'expression de la densité est un facteur de pondération.

6.7.2 Moments

$$E(X) = \exp\left(m + \frac{\sigma^2}{2}\right)$$

$$\text{Var}(X) = [\exp(\sigma^2) - 1] \exp(2m + \sigma^2)$$

Remarque

Quand l'écart-type σ est petit, la loi de X est proche d'une loi normale.

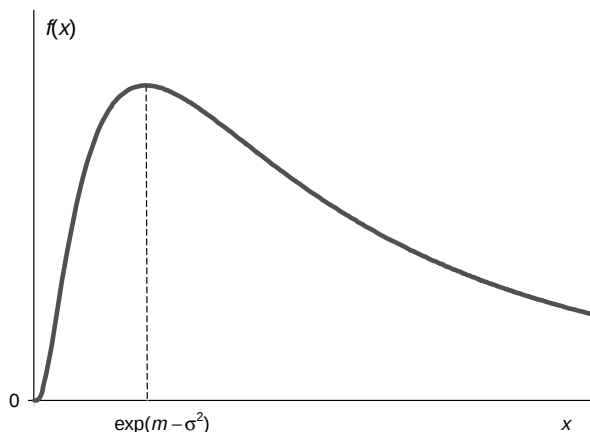


Figure 6.12 – Densité de la loi log-normale.

6.7.3 Domaine d'utilisation

La loi log-normale représente la loi d'une grandeur résultant de l'influence d'un grand nombre de facteurs aléatoires et indépendants agissant sous *forme multiplicative*. Elle est fréquemment utilisée en fiabilité car la variable aléatoire X est positive et grâce au paramètre de forme σ , elle peut avoir des représentations très variées.

Le produit de n variables aléatoires indépendantes suivant une loi log-normale suit une loi log-normale.

7 • CONVOLUTION FONCTIONS CARACTÉRISTIQUES CONVERGENCES STOCHASTIQUES

7.1 Convolution

Problème : connaissant la loi des variables aléatoires X et Y , indépendantes ou non, quelle est la loi de la somme $Z = X + Y$ de ces deux variables ?

La résolution de ce problème dépend de la nature des variables, discrètes ou continues. Il est plus facile à résoudre si les variables sont indépendantes.

7.1.1 Cas de deux variables aléatoires discrètes

Le théorème des probabilités totales donne la solution. En effet :

$$\begin{aligned}\Pr(Z = z) &= \sum_x \Pr[(X = x) \text{ et } (Y = z - x)] \\ &= \sum_y \Pr[(Y = y) \text{ et } (X = z - y)]\end{aligned}$$

- Si les variables aléatoires X et Y sont indépendantes, la formule précédente s'écrit :

$$\begin{aligned}\Pr(Z = z) &= \sum_x \Pr(X = x) \Pr(Y = z - x) \\ &= \sum_y \Pr(Y = y) \Pr(X = z - y)\end{aligned}$$

- Si les variables aléatoires X et Y ne sont pas indépendantes, on peut simplement écrire, en introduisant les probabilités conditionnelles :

$$\begin{aligned}\Pr(Z = z) &= \sum_x \Pr(X = x) \Pr(Y = z - x / X = x) \\ &= \sum_y \Pr(Y = y) \Pr(X = z - y / Y = y)\end{aligned}$$

Remarque

Les limites des variables X et Y doivent être compatibles avec la condition $Z = z$.

Exemple 7.1 Loi de Poisson

X et Y sont deux variables aléatoires indépendantes, suivant chacune une loi de Poisson de paramètres λ et μ respectivement.

Les formules précédentes donnent ($Z = X + Y$) :

$$\begin{aligned}\Pr(Z = z) &= \sum_{x=0}^z \frac{e^{-\lambda} \lambda^x}{x!} \frac{e^{-\mu} \mu^{z-x}}{(z-x)!} = \frac{e^{-(\lambda+\mu)}}{z!} \sum_{x=0}^z \frac{z!}{x! (z-x)!} \lambda^x \mu^{z-x} \\ \Pr(Z = z) &= \frac{e^{-(\lambda+\mu)} (\lambda + \mu)^z}{z!}\end{aligned}$$

On reconnaît l'expression de la densité de probabilité d'une loi de Poisson de paramètre $(\lambda + \mu)$.

En résumé la loi suivie par la somme de deux variables aléatoires indépendantes suivant des lois de Poisson de paramètres λ et μ est une loi de Poisson de paramètre $(\lambda + \mu)$.

La loi de Poisson est stable pour l'addition de variables aléatoires indépendantes.

7.1.2 Cas de deux variables continues

La loi de probabilité de la variable $Z = X + Y$ est la mesure image de $\Pr_{XY}$, loi de probabilité du couple (X, Y) , par l'application de $\mathbb{R}^2$ dans $\mathbb{R}$ définie par $(x, y) \rightarrow x + y$.

Il en résulte que si les *variables aléatoires* X et Y sont *indépendantes*, la loi de probabilité $\Pr_Z$ de la variable aléatoire $Z = X + Y$ est la mesure image de

$\Pr_X \otimes \Pr_Y$ par l'application de $\mathbb{R}^2$ dans $\mathbb{R}$ définie par $(x, y) \rightarrow x + y$. $\Pr_Z$ est le produit de convolution de $\Pr_X$ et $\Pr_Y$.

Pour tout borélien $\mathcal{B}$ de $\mathbb{R}^2$, cette probabilité est définie par :

$$\Pr_Z(\mathcal{B}) = \int_{\mathbb{R}^2} 1_{\mathcal{B}}(x + y) d\Pr_x \otimes d\Pr_y$$

Les variables X et Y jouent des rôles symétriques.

Si les lois de probabilité des variables aléatoires indépendantes X et Y admettent des densités f et g , l'expression précédente s'écrit :

$$\Pr_Z(\mathcal{B}) = \int_{\mathbb{R}^2} 1_{\mathcal{B}}(x + y) f(x) g(y) dx dy$$

Posons $x + y = z$, $x = u$ et appliquons le théorème de Fubini :

$$\begin{aligned} \Pr_Z(\mathcal{B}) &= \int_{\mathbb{R}^2} 1_{\mathcal{B}}(z) f(u) g(z - u) du dz \\ &= \int_{\mathbb{R}} 1_{\mathcal{B}} dz \int_{D_x} f(u) g(z - u) du \end{aligned}$$

D'où la densité de la variable aléatoire Z :

$$k(z) = \int_{D_x} f(x) g(z - x) dx = \int_{D_y} f(z - y) g(y) dy$$

D_x et D_y désignent les domaines de variation des variables aléatoires X et Y , compatibles avec la condition $Z = z$.

Par intégration, on obtient la fonction de répartition de la variable Z :

$$K(z) = \Pr(Z < z) = \int_{D_x} f(x) G(z - x) dx = \int_{D_y} F(z - y) g(y) dy$$

où F et G désignent les fonctions de répartition des variables X et Y .

Exemple 7.2 Loi gamma

X et Y sont deux variables aléatoires indépendantes, suivant des lois gamma γ_r et γ_s . Les densités de probabilité de X et Y ont pour expression :

$$f(x) = \frac{1}{\Gamma(r)} e^{-x} x^{r-1} \quad g(y) = \frac{1}{\Gamma(s)} e^{-y} y^{s-1}$$

La densité $k(z)$ de la variable aléatoire $Z = X + Y$ est donnée par l'intégrale :

$$k(z) = \int_0^z \frac{1}{\Gamma(r)} e^{-x} x^{r-1} \frac{1}{\Gamma(s)} e^{-(z-x)} (z-x)^{s-1} dx$$

$$k(z) = \frac{1}{\Gamma(r+s)} e^{-z} z^{r+s-1}$$

On reconnaît la densité de probabilité d'une loi gamma γ_{r+s} . D'où le résultat :

Si X et Y sont des variables aléatoires indépendantes suivant des lois γ_r et γ_s , la variable aléatoire $Z = X + Y$ suit une loi γ_{r+s} .

La loi gamma est stable pour l'addition des variables aléatoires indépendantes.

Exemple 7.3 Loi uniforme

Soient X et Y deux variables aléatoires indépendantes suivant des lois uniformes sur $(0, 1)$. Leur somme ne suit pas une loi uniforme sur $(0, 2)$.

On peut donner une démonstration géométrique de ce résultat. Le couple (X, Y) est uniformément distribué dans le carré de côté unité ; l'événement $Z < z$ correspond à la zone hachurée. Il suffit de calculer l'aire de ce domaine. On obtient deux expressions différentes pour la densité $k(z)$ et pour la fonction de répartition $K(z)$, selon que z est compris entre 0 et 1 ou entre 1 et 2 (figures 7.1 et 7.2).

Les fonctions $k(z)$ et $K(z)$ sont continues.

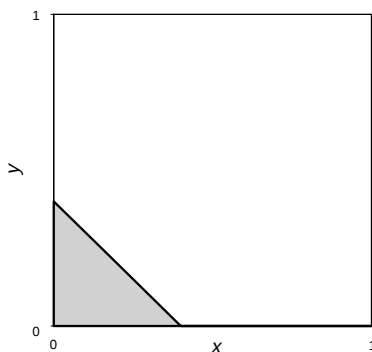


Figure 7.1 – Somme de deux lois uniformes correspondant à $0 \leq z < 1$.

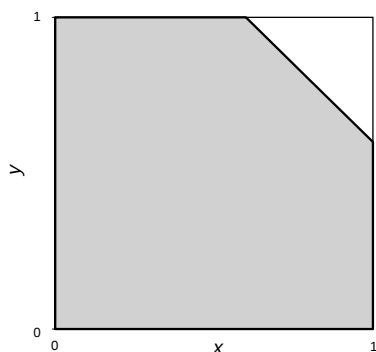


Figure 7.2 – Somme de deux lois uniformes correspondant à $1 \leq z < 2$.

D'où les résultats :

$$z \leq 0 \quad K(Z) = k(z) = 0$$

$$0 \leq z < 1 \quad K(z) = \frac{z^2}{2} \quad k(z) = z$$

$$1 \leq z \leq 2 \quad K(z) = 1 - \frac{(2-z)^2}{2} \quad k(z) = 2 - z$$

$$z \geq 2 \quad K(Z) = 1 \quad k(z) = 0$$

7.2 Fonction caractéristique

La fonction caractéristique (f.c., en abrégé) d'une variable aléatoire est l'espérance mathématique de la variable e^{itx} ; elle est définie par :

$$\varphi_X(t) = E(e^{itX}) = \int_{\mathbb{R}} e^{itx} d\Pr_X \quad \forall t \in \mathbb{R}$$

C'est la transformée de Fourier-Stieltjes de sa mesure de probabilité.

Elle est utilisée, pour étudier le comportement asymptotique de la somme de variables aléatoires indépendantes car des théorèmes d'analyse permettent de remplacer l'étude de la convergence en loi par l'étude de la convergence des

fonctions caractéristiques, elle est également utilisée pour calculer les différents moments d'une distribution...

7.2.1 Principales propriétés

Elles se déduisent des propriétés de la transformée de Fourier d'une fonction intégrable.

- La fonction caractéristique d'une variable aléatoire détermine sans ambiguïté sa loi de probabilité. Deux variables aléatoires ayant la même fonction caractéristique ont la même loi de probabilité. D'où le nom de « caractéristique » donné à la fonction φ .
- Comme Pr_X est une mesure bornée et que $|e^{itx}| = 1$, la fonction $\varphi_X(t)$ existe, est bornée et continue pour toutes les valeurs de t .
- Si la loi de la variable X possède une densité, la fonction caractéristique s'obtient par l'intégrale :

$$\varphi_X(t) = \int_{\mathbb{R}} e^{itx} f(x) dx \quad \forall t \in \mathbb{R}$$

- Propriétés de linéarité :

$$\varphi_{\lambda X}(t) = \varphi_X(\lambda t) \quad \varphi_{X+a}(t) = \varphi_X(t) e^{ita}$$

- En particulier, si U est la variable aléatoire centrée réduite associée à la variable X , $U = \frac{X - m}{\sigma}$, on obtient :

$$\varphi_U(t) = e^{-\frac{itm}{\sigma}} \varphi_X\left(\frac{t}{\sigma}\right) \quad \varphi_X(t) = e^{itm} \varphi_U(\sigma t)$$

- Si φ est une fonction caractéristique, il en est de même de :
 φ^n (n entier) $\overline{\varphi}$ $|\varphi|^2$ partie réelle de φ ...
- Fonction caractéristique de la somme de deux variables aléatoires indépendantes : les variables aléatoires X et Y étant indépendantes, il en est de même des variables aléatoires e^{itX} et e^{itY} D'où le résultat :

$$\varphi_{X+Y}(t) = E[e^{it(X+Y)}] = \varphi_X(t) \varphi_Y(t)$$

La fonction caractéristique de la somme de deux variables aléatoires indépendantes est égale au produit de leurs fonctions caractéristiques.

Exemple 7.4 Loi normale

X et Y sont deux variables aléatoires normales, indépendantes, de paramètres m_1 et σ_1 pour X et m_2 et σ_2 pour Y .

La fonction caractéristique de leur somme $Z = X + Y$ est égale à :

$$\varphi_Z(t) = e^{it(m_1+m_2)} e^{-\frac{t^2(\sigma_1^2+\sigma_2^2)}{2}}$$

On reconnaît la fonction caractéristique d'une variable aléatoire normale qui a pour espérance mathématique $m = m_1 + m_2$ et pour variance $(\sigma)^2 = (\sigma_1)^2 + (\sigma_2)^2$.

La somme de deux variables aléatoires normales indépendantes est une variable aléatoire normale.

7.2.2 Fonction caractéristique et moments

- Si $E(X^k)$ est finie pour un entier $k \geq 1$, alors φ est continûment dérivable jusqu'à l'ordre k inclus et on peut calculer ses dérivées par dérivation sous le signe d'intégration. On obtient :

$$\varphi_X(0) = 1 \quad \varphi_X^k(0) = i^k E(X^k)$$

- Si $\varphi_X(t)$ est de classe C^∞ , la formule de Mac-Laurin donne :

$$\varphi_X(t) = \sum_{k=0}^{\infty} \frac{t^k}{k!} i^k E(X^k)$$

Exemple 7.5 Moments de la loi normale centrée réduite

Le développement de la fonction caractéristique de la loi normale réduite (donnée dans le tableau 7.2) est égal à :

$$\varphi_U(t) = e^{-\frac{t^2}{2}} = 1 - \frac{t^2}{2} + \frac{1}{2!} \left(-\frac{t^2}{2}\right)^2 + \dots + \frac{1}{k!} \left(-\frac{t^2}{2}\right)^k + \dots$$

On identifie à la formule précédente :

$$\varphi_X(t) = \sum_{k=0}^{\infty} \frac{t^k}{k!} i^k E(X^k)$$

On retrouve les résultats donnés dans le chapitre 6, paragraphe 6.6.2 :

- les moments d'ordre impair sont nuls,
- les moments d'ordre pair égaux à $\mu_{2k} = \frac{(2k)!}{2^k k!}$

7.2.3 Inversion de la fonction caractéristique

Les formules d'inversion de la transformée de Fourier permettent d'obtenir la densité de la loi de X connaissant $\varphi_X(t)$.

– X admet une densité $f(x)$, continue, définie par l'intégrale :

$$f(x) = \frac{1}{2\pi} \int_{\mathbb{R}} \varphi_X(t) e^{-itx} dt$$

– Sinon, on a le résultat suivant :

$$f(b) - f(a) = \lim_{T \rightarrow \infty} \frac{1}{2\pi} \int_{-T}^T \varphi_X(t) \frac{e^{itb} - e^{ita}}{it} dt$$

7.2.4 Exemples de fonctions caractéristiques

Tableau 7.1 – Loïs discrètes.

Loi de Bernoulli	$\varphi_X(t) = p e^{it} + q \quad (q = 1 - p)$
Loi binomiale $B(n; p)$	$\varphi_X(t) = (p e^{it} + q)^n$
Loi de Poisson $P(\lambda)$	$\varphi_X(t) = e^{\lambda(t e^{it} - 1)}$

Tableau 7.2 – Loïs continues.

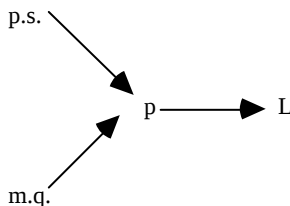
Loi uniforme sur $[a, b]$	$\varphi_X(t) = \frac{\sin at}{at}$
Loi gamma $\Gamma(1)$	$\varphi_{\gamma_1}(t) = \frac{1}{1 - it}$
Loi gamma $\Gamma(r)$	$\varphi_{\gamma_r}(t) = \frac{1}{(1 - it)^r}$
Loi normale réduite $N(0; 1)$	$\varphi_U(t) = e^{-\frac{t^2}{2}}$
Loi normale $N(m, \sigma)$	$\varphi_X(t) = e^{itm} e^{-\frac{t^2 \sigma^2}{2}}$
Loi de Cauchy de densité : $f(x) = \frac{1}{\pi} \frac{1}{1 + x^2}$	$\varphi_X(t) = e^{- t }$

7.3 Convergence des suites de variables aléatoires

Les modes de convergence les plus utilisés en probabilité sont les suivants :

- la convergence en probabilité notée p ,
- la convergence presque sûre notée $p.s.$,
- la convergence en loi notée L ,
- la convergence en moyenne quadratique notée $m.q.$

Ces différents modes de convergence ne sont pas indépendants et satisfont aux implications suivantes :



Grâce à ces théorèmes de convergence, le calcul des probabilités trouve sa justification dans l'étude des phénomènes mettant en jeu des populations ou des observations nombreuses (méthodes de sondage, théorie de l'estimation, des tests...).

7.3.1 Convergence en probabilité

Une suite de n variables aléatoires (X_n) , non nécessairement indépendantes, *converge en probabilité* vers une constante a et on écrit, $X_n \xrightarrow{p} a$, quand n tend vers l'infini si :

$$\forall \varepsilon, \eta > 0 \quad \exists N(\varepsilon, \eta) \text{ tel que } n > N \Rightarrow \Pr(|X_n - a| > \varepsilon) < \eta$$

La convergence en probabilité de la suite (X_n) vers la variable aléatoire X est la convergence de la suite de variables aléatoires $(X_n - X)$ vers 0.

■ Propriétés

- Si $E(X_n) \rightarrow a$, il suffit de montrer que $\text{Var}(X_n) \rightarrow 0$, pour établir la convergence en probabilité de la suite (X_n) vers a (application de l'inégalité de Bienaymé-Tchebycheff, chapitre 4, paragraphe 4.7.4).
- La convergence en probabilité n'entraîne pas celle des moments ; en particulier, elle n'entraîne pas celle des espérances mathématiques.

Exemple 7.6

Considérons la suite de variables aléatoires (X_n) , chaque variable prenant deux valeurs 0 et α_n avec les probabilités suivantes :

$$\Pr(X_n = 0) = 1 - \frac{1}{n} \quad \Pr(X_n = \alpha_n) = \frac{1}{n}$$

La suite (X_n) converge en probabilité vers 0 quand n tend vers l'infini.

$$E(X_n) = \frac{\alpha_n}{n}$$

Quand $n \rightarrow \infty$, selon les valeurs de α_n , on obtient pour $E(X_n)$ différentes limites, finies ou non, ou pas de limite...

$$\alpha_n = \sqrt{n} \quad E(X_n) \rightarrow 0 \quad \alpha_n = n \quad E(X_n) = 1$$

$$\alpha_n = (-1)^n n \quad E(X_n) = (-1)^n \quad \alpha_n = n^r \quad E(X_n) \rightarrow \infty$$

tandis que l'espérance mathématique de la limite de (X_n) est nulle puisque cette limite est égale à 0.

B

CALCUL DES PROBABILITÉS

7.3.2 Convergence presque sûre

- Deux variables aléatoires X et Y sont égales presque sûrement si :

$$\Pr[\omega / X(\omega) \neq Y(\omega)] = 0$$

C'est la définition de l'égalité presque partout des fonctions mesurables.

- La suite de variables aléatoires (X_n) , non nécessairement indépendantes, converge presque sûrement vers X et on écrit $X_n \xrightarrow{\text{p.s.}} X$, quand n tend vers l'infini, si :

$$\Pr[\omega / \lim X_n(\omega) \neq X(\omega)] = 0$$

■ Comparaison de ces deux modes de convergence

- La convergence en probabilité de la suite (X_n) , vers 0 par exemple, implique que, ε et η positifs étant donnés, l'événement $\{|X_n| \leq \varepsilon\}$ est réalisé avec une probabilité supérieure à $(1 - \eta)$ pour tout n fixé, à partir d'un certain rang.
- La convergence presque sûre de la suite (X_n) vers 0, implique qu'à partir d'un certain rang, tous les événements $\{|X_n| \leq \varepsilon\}$ sont réalisés simultanément avec une probabilité supérieure à $(1 - \eta)$.
- La convergence presque sûre est plus stricte que la convergence en probabilité, elle l'entraîne comme on peut facilement le démontrer.
- La convergence en probabilité justifie l'utilisation de la méthode des sondages pour estimer la proportion des individus qui ont le caractère A dans une population donnée.
- La convergence en probabilité implique la loi faible des grands nombres, la convergence presque sûre implique la loi forte des grands nombres (paragraphe 7.4).

7.3.3 Convergence en loi

La suite des variables aléatoires (X_n) , non nécessairement indépendantes, de fonction de répartition F_n , converge en loi vers la variable aléatoire X , de fonction de répartition F , quand n tend vers l'infini, si en tout point de continuité x de F , la limite de la fonction F_n est égale à la fonction F et on écrit :

$$X_n \xrightarrow{L} X$$

La convergence est réalisée aux points de continuité de la fonction F . C'est la convergence ponctuelle de la suite des fonctions de répartition. En un point de discontinuité x_0 de F , différentes situations sont possibles, soit $F_n(x_0)$ n'a pas de limite ou a une limite différente de $F(x_0)$...

Exemple 7.7

Soit X_n la variable aléatoire suivant la loi normale $N\left(0; \frac{1}{n}\right)$.

La suite des variables (X_n) converge en loi vers 0 quand $n \rightarrow \infty$.

Soit F_n la fonction de répartition de X_n , c'est-à-dire la fonction définie par :

$$F_n(x) = \Pr(X_n < x)$$

Quand $n \rightarrow \infty$, $F_n(x)$ a pour limite 0 si $x \leq 0$ et pour limite 1 si $x > 0$.

La suite $F_n(x)$ converge donc, pour toutes les valeurs de x différentes de 0, vers la fonction de répartition $F(x)$, définie par :

$$F(x) = 0 \quad \text{si } x \leq 0 \quad \text{et} \quad F(x) = 1 \quad \text{si } x > 0.$$

Or $\forall n \quad F_n(0) = 0,5$ donc $F(0) = 0 \neq F_n(0)$.

Au point de discontinuité 0, la limite de $F_n(x)$ est différente de $F(0)$.

■ Propriétés

Deux théorèmes donnent les relations entre la convergence en loi d'une suite de variables aléatoires (X_n) et la convergence de la suite (φ_n) des fonctions caractéristiques.

□ Théorème 1

Si la suite de variables aléatoires (X_n) converge en loi vers la variable aléatoire X quand $n \rightarrow \infty$, alors la suite (φ_n) des fonctions caractéristiques converge vers la fonction caractéristique φ de la variable aléatoire X , la convergence étant uniforme dans tout intervalle fini de $\mathbb{R}$.

□ Théorème 2 (Lévy-Cramer-Dugué)

Si la suite (φ_n) de fonctions caractéristiques converge simplement, quand $n \rightarrow \infty$, vers une fonction φ et si la partie réelle de φ est continue à l'origine, alors :

- φ est une fonction caractéristique,
- la suite de fonctions de répartition F_n de X_n converge simplement vers la fonction de répartition F dont φ est la transformée de Fourier-Stieltjes, la convergence a lieu en tous les points de continuité de F .

Si F est continue, la convergence est uniforme.

□ Applications

La convergence en loi est la plus faible mais elle est la plus utilisée car elle permet d'approcher la fonction de répartition de X_n par celle de X . C'est ainsi que l'on justifie la convergence de la loi binomiale vers la loi de Poisson, de la loi binomiale et de la loi de Poisson vers la loi normale (chapitre 6, paragraphe 6.6.6).

7.3.4 Convergence en moyenne quadratique

Une suite de variables aléatoires (X_n) , non nécessairement indépendantes, converge en moyenne quadratique d'ordre q vers X si $E[(X_n - X)^q] \rightarrow 0$ quand $n \rightarrow \infty$. Cette condition implique que le moment d'ordre q existe.

Le cas $q = 2$ est le plus utilisé.

La convergence en moyenne quadratique implique la convergence en probabilité.

7.4 Lois des grands nombres

7.4.1 Loi faible des grands nombres

Soit (X_n) une suite de variables aléatoires, indépendantes, centrées, telles que les variances $(\sigma_i)^2$ existent et vérifient :

$$\frac{1}{n^2} \sum_{i=1}^n \sigma_i^2 \rightarrow 0 \quad \text{quand } n \rightarrow \infty$$

Dans ces conditions, la suite des moyennes :

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

converge en probabilité vers 0 quand n tend vers l'infini.

La suite (X_n) satisfait à la *loi faible des grands nombres*.

Ce résultat se démontre grâce à l'inégalité de Bienaymé-Tchebyshev.

■ Application

Soit (X_n) une suite de variables aléatoires, indépendantes, de même loi, telles que l'espérance et la variance existent. Alors, la suite des moyennes :

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

converge en probabilité vers $E(X)$.

Ainsi, la moyenne d'un échantillon issu d'une population de taille n tend vers la moyenne théorique de la population quand la taille de l'échantillon augmente (application en théorie de l'estimation).

7.4.2 Loi forte des grands nombres

Soit (X_n) une suite de variables aléatoires, indépendantes, centrées, telles que les variances $(\sigma_i)^2$ existent et vérifient :

$$\sum_{k \geq 1} \frac{\sigma_k^2}{k^2} < \infty$$

Dans ces conditions, la suite des moyennes :

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

converge presque sûrement vers 0 quand n tend vers l'infini.

La démonstration de ce théorème nécessite l'utilisation de théorèmes fins d'analyse.

Ces résultats sont utilisés dans la théorie de l'estimation et des tests.

7.5 Théorème central limite

Le théorème central limite ou théorème de la limite centrale établit la convergence en loi de la somme de variables aléatoires indépendantes vers la loi normale, sous des hypothèses faibles.

7.5.1 Théorème central limite (première forme)

Soit (X_n) une suite de variables aléatoires, indépendantes, de même loi, d'espérance mathématique m et d'écart-type σ .

Quand $n \rightarrow \infty$, la loi de la variable aléatoire :

$$\frac{1}{\sqrt{n}} \frac{(X_1 + \dots + X_n) - n m}{\sigma} = \sum_{i=1}^n \frac{X_i - m}{\sigma \sqrt{n}} = \frac{\bar{X} - m}{\sigma / \sqrt{n}}$$

tend vers la loi normale, centrée, réduite $N(0 ; 1)$.

La démonstration de ce théorème est une application des propriétés des fonctions caractéristiques.

Un théorème plus général est dû à Lindeberg.

7.5.2 Théorème central limite ou théorème de Lindeberg

C'est la deuxième forme de ce théorème. (X_i) est une suite de variables aléatoires indépendantes, non nécessairement de même loi, d'espérance mathématique m_i et d'écart-type σ_i . Soit $F_i(x)$ la fonction de répartition de $X_i - m_i$.

Posons :

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

Si la limite de :

$$\frac{1}{S_n^2} \sum_{i=1}^n \int_{|x| > S_n} x^2 dF_i(x)$$

est égale à 0, quand $n \rightarrow \infty$, alors, la variable aléatoire :

$$\frac{1}{S_n} \sum_{i=1}^n (X_i - m_i)$$

converge en loi vers une variable aléatoire U suivant la loi normale $N(0 ; 1)$. Les variables aléatoires qui figurent dans la somme sont *uniformément petites*. Le théorème central limite a des applications nombreuses en théorie de l'estimation.

8 • VARIABLES ALÉATOIRES SIMULTANÉES

B

CALCUL DES PROBABILITÉS

L'étude des variables aléatoires réelles isolées, discrètes ou continues, à une dimension, ne permet pas de résoudre tous les problèmes faisant appel à la théorie statistique. En effet, les événements étudiés peuvent être liés à plusieurs variables aléatoires simultanément, d'où la nécessité de définir des lois de probabilité conjointes, marginales, conditionnelles ainsi que toutes les caractéristiques s'y rapportant. La loi de probabilité d'un couple de variables aléatoires (X, Y) est une application de $(\Omega, \mathcal{C}, \Pr)$ dans $\mathbb{R}^2$, muni de la tribu de Borel.

8.1 Étude d'un couple de variables aléatoires discrètes

X et Y sont deux variables aléatoires prenant soit un nombre fini de valeurs, p pour la variable X et q pour la variable Y , soit un ensemble dénombrable de valeurs, notées x_i et y_j .

8.1.1 Lois associées

- La loi du couple (X, Y) , appelée *loi de probabilité simultanée* ou *loi conjointe*, est définie par l'ensemble des nombres p_{ij} ($0 \leq p_{ij} < 1$) tels que :

$$p_{ij} = \Pr(X = x_i \text{ et } Y = y_j)$$

Ces probabilités p_{ij} vérifient la relation :

$$\sum_{i=1}^p \sum_{j=1}^q p_{ij} = 1$$

- Dans le cas fini, la loi conjointe se met sous la forme d'un *tableau de contingence*.

Les probabilités p_{ij} figurant dans le tableau de contingence définissent la loi du couple et toutes les lois associées.

Tableau 8.1 – Tableau de contingence.

X	Y	y_1		y_j		y_q	Total
x_1		p_{11}		p_{1j}		p_{1q}	$p_{1.}$
x_i		p_{i1}		p_{ij}			$p_{i.}$
x_p		p_{p1}		p_{pj}		p_{pq}	$p_{p.}$
Total		$p_{.1}$		$p_{.j}$		$p_{.q}$	1

- Les *lois marginales* sont les lois de probabilité des variables X et Y prises séparément :

$$\text{Loi marginale de la variable } X \quad \Pr(X = x_i) = \sum_{j=1}^q p_{ij} = p_{i.}$$

$$\text{Loi marginale de la variable } Y \quad \Pr(Y = y_j) = \sum_{i=1}^p p_{ij} = p_{.j}$$

Les quantités $p_{i.}$ et $p_{.j}$ constituent les *marges* du tableau de contingence et vérifient les relations :

$$\sum_{i=1}^p p_{i.} = \sum_{j=1}^q p_{.j} = 1$$

- Les *lois conditionnelles* sont les deux familles de lois suivantes :
 - *loi conditionnelle de X sachant $Y = y_j$* (la valeur de la variable Y est connue) :

$$\Pr(X = x_i / Y = y_j) = \frac{p_{ij}}{p_{.j}} = \frac{\Pr(X = x_i \text{ et } Y = y_j)}{\Pr(Y = y_j)}$$

- *loi conditionnelle de Y sachant X = x_i* (la valeur de la variable X est connue) :

$$\Pr(Y = y_j / X = x_i) = \frac{p_{ij}}{p_{i.}} = \frac{\Pr(X = x_i \text{ et } Y = y_j)}{\Pr(X = x_i)}$$

Remarques

- Ces lois sont parfaitement définies si les quantités $\Pr(Y = y_j)$ ou $\Pr(X = x_i)$ sont différentes de 0.
- Si on connaît les lois conditionnelles, on peut inversement, en déduire la loi du couple.
- Grâce à la formule de Bayes, on peut exprimer une loi conditionnelle en fonction de l'autre. Ainsi par exemple :

$$\Pr(X = x_i / Y = y_j) = \frac{\Pr(Y = y_j / X = x_i) \Pr(X = x_i)}{\sum_{i=1}^p \Pr(Y = y_j / X = x_i) \Pr(X = x_i)}$$

8.1.2 Indépendance

Les variables aléatoires X et Y sont *indépendantes* si et seulement si :

$$\Pr(X = x_i \text{ et } Y = y_j) = \Pr(X = x_i) \Pr(Y = y_j)$$

D'où les relations :

$$\Pr(X = x_i / Y = y_j) = \Pr(X = x_i)$$

$$\Pr(Y = y_j / X = x_i) = \Pr(Y = y_j)$$

La probabilité de réalisation de l'événement $(X = x_i)$ ne dépend pas de la réalisation de l'événement $(Y = y_j)$ et la même propriété est vérifiée pour l'événement $(Y = y_j)$, il ne dépend pas de la réalisation de l'événement $(X = x_i)$.

8.1.3 Moments conditionnels. Théorèmes de l'espérance totale et de la variance totale

■ Espérance conditionnelle de Y sachant X = x, notée E(Y/X = x)

C'est l'espérance de la variable Y par rapport à sa loi conditionnelle :

$$E(Y/X = x) = \sum_y y \Pr(Y = y / X = x)$$

Cette fonction de x , appelée *fonction de régression de Y en X* , est l'ensemble des moyennes conditionnelles de Y sachant X :

$$E(Y/X = x) = \varphi(x)$$

On définit ainsi une variable aléatoire *espérance conditionnelle* qui a des propriétés remarquables.

- Propriétés de linéarité. Si a et b sont des constantes :

$$E[aY_1 + bY_2 / X = x] = aE(Y_1 / X) + bE(Y_2 / X)$$

- Espérance de l'espérance conditionnelle ou théorème de l'espérance totale :

$$E[E(Y/X)] = E(Y)$$

Pour démontrer ce résultat, il suffit d'utiliser la définition et les propriétés de l'espérance.

- Pour $Z = \varphi(X)$, une variable aléatoire fonction de X , on obtient :

$$E[\varphi(X) Y/X] = \varphi(X) E(Y/X)$$

■ Variance conditionnelle de Y sachant $X = x$, notée $\text{Var}(Y/X = x)$

Par définition :

$$\text{Var}(Y/X = x) = E\left[\{Y - E(Y/X = x)\}^2 / X = x\right]$$

La propriété suivante relie l'espérance et la variance conditionnelles.

■ Variance de l'espérance conditionnelle ou théorème de la variance totale

On démontre, soit grâce aux propriétés de l'espérance et de la variance, soit géométriquement, le résultat suivant :

$$\text{Var}(Y) = E[\text{Var}(Y/X)] + \text{Var}[E(Y/X)]$$

On définit de la même façon l'*espérance conditionnelle de X sachant $Y = y$* .

Les théorèmes de l'espérance totale et de la variance totale sont très utiles pour calculer l'espérance et la variance d'une loi compliquée dont les lois conditionnelles sont simples.

Exemple 8.1

On suppose que le nombre d'accidents de la circulation au cours d'un week-end suit une loi de Poisson de paramètre λ et que le nombre de blessés par accident suit également une loi de Poisson de paramètre μ .

Soit X_i la variable aléatoire « nombre de blessés dans l'accident n° i ». La loi de X_i est la loi de Poisson $P(\mu)$.

Soit N la variable aléatoire représentant le nombre d'accidents, la loi de N est la loi de Poisson $P(\lambda)$.

Soit S la variable aléatoire représentant le nombre total de blessés :

$$S = X_1 + X_2 + \dots + X_N$$

S est donc égale à la somme d'un nombre aléatoire de variables aléatoires, indépendantes suivant la même loi de Poisson (on suppose que les accidents se produisent indépendamment les uns des autres, ce qui n'est peut-être pas rigoureusement vrai). La loi de S est donc une loi de Poisson $P(\theta)$ avec $\theta = n\mu$ s'il y a eu n accidents.

La loi de probabilité conditionnelle de $S/N = n$ est la loi de Poisson :

$$\Pr(S = s / N = n) = \frac{e^{-n\mu} (n\mu)^s}{s!}$$

et la loi de S est donc :

$$\Pr(S = s) = \sum_{n=1}^{\infty} \frac{e^{-\lambda} \lambda^n}{n!} \frac{e^{-n\mu} (n\mu)^s}{s!} = \frac{e^{-\lambda} \mu^s}{s!} \sum_{n=1}^{\infty} \frac{\lambda^n n^s e^{-n\mu}}{n!}$$

On ne peut pas donner une expression mathématique simple à ce résultat. Cependant en utilisant les théorèmes de l'espérance et de la variance totales, on calcule l'espérance et la variance de la variable S :

$$- E(S) = E[E(S/N)], \quad E(S/N = n) = n\mu \quad \text{d'où} \quad E(S/N) = N\mu$$

$$E(S) = E(N\mu) = \mu E(N) = \mu\lambda$$

$$- \text{Var}(S) = E[\text{Var}(S/N)] + \text{Var}[E(S/N)]$$

$$E(S/N) = N\mu \quad \text{Var}(S/N) = N\mu \quad E[\text{Var}(S/N)] = E(N\mu) = \mu\lambda$$

$$\text{Var}[E(S/N)] = \text{Var}(N\mu) = \mu^2 \text{Var}(N) = \mu^2 \lambda$$

$$\text{Var}(S) = \mu\lambda + \mu^2 \lambda = \mu\lambda(1 + \mu)$$

Pour tous ces calculs, on a utilisé les propriétés de la loi de Poisson.

8.2 Étude d'un couple de variables aléatoires continues

Le problème est de généraliser les résultats obtenus dans le cas des variables aléatoires discrètes au cas continu, c'est-à-dire lorsque l'événement $X = x$ est de probabilité nulle. En particulier, il faut donner un sens à des expressions telles que $E(Y/X = x)$.

Soit (X, Y) un couple de variables aléatoires réelles continues, définies sur le même espace de probabilité $(\Omega, \mathcal{B}, \Pr)$, c'est-à-dire une application mesurable de $(\Omega, \mathcal{B}, \Pr)$ dans l'espace probabilisable $(\mathbb{R}^2, \mathcal{B}^2)$, muni de la mesure de Lebesgue, notée $dx dy$.

8.2.1 Loïs associées

- La *fonction de répartition conjointe* F du couple (X, Y) est l'application de $\mathbb{R}^2$ dans $[0, 1]$ définie par :

$$F(a, b) = \Pr(X < a \text{ et } Y < b) \quad \forall (a, b) \in \mathbb{R}^2$$

- Le couple (X, Y) est absolument continu, s'il existe une fonction f continue, des deux variables X et Y , appelée *densité de probabilité conjointe du couple* (X, Y) , telle que, pour tout domaine D du plan, on ait :

$$\Pr[(X, Y) \in D] = \int_D f(x, y) \, dx \, dy$$

Si le domaine D est l'ensemble des couples (x, y) tels que $x \leq a$ et $y \leq b$, on obtient :

$$F(a, b) = \int_{-\infty}^b dy \int_{-\infty}^a f(x, y) \, dx$$

Entre la densité f et la fonction de répartition F , il existe la relation suivante :

$$f(x, y) = \frac{d^2 F(x, y)}{dx \, dy}$$

- Les *loïs marginales* sont les loïs des variables X et Y prises séparément. Fonction de répartition marginale de la variable X :

$$\begin{aligned} \Pr(-\infty < X < a \text{ et } -\infty < Y < +\infty) &= H(a) \\ &= \int_{-\infty}^a dx \int_{-\infty}^{+\infty} f(x, y) \, dy \end{aligned}$$

Fonction de répartition marginale de la variable Y :

$$\Pr(-\infty < Y < b \text{ et } -\infty < X < +\infty) = G(b) \\ = \int_{-\infty}^b dy \int_{-\infty}^{+\infty} f(x, y) dx$$

On peut aussi écrire :

$$H(a) = F(a, +\infty) \quad G(b) = F(+\infty, b)$$

– Densités marginales :

$$h(x) = \int_{-\infty}^{+\infty} f(x, y) dy \quad g(y) = \int_{-\infty}^{+\infty} f(x, y) dx$$

– Densités conditionnelles :

$$h(x/y) = \frac{f(x, y)}{g(y)} \quad g(y/x) = \frac{f(x, y)}{h(x)}$$

8.2.2 Variables aléatoires indépendantes

Soient f , h et g les densités de probabilité du couple (X, Y) et des variables X et Y prises séparément.

Les variables aléatoires continues, X et Y , sont *indépendantes* si et seulement si :

$$f(x, y) = h(x) g(y)$$

Remarque sur l'indépendance

En analyse statistique, la question qui se pose souvent est de savoir, au vu d'un tableau répartissant une population de taille n selon des modalités définies par deux variables aléatoires X et Y , si ces variables sont indépendantes ou non. Le test le plus utilisé est le *test du chi-deux* (chapitre 17, paragraphe 17.5) qui consiste à comparer le tableau observé et le tableau théorique que l'on obtiendrait si les variables X et Y étaient indépendantes.

8.2.3 Moments d'une distribution

Le moment d'ordre p par rapport à la variable X et d'ordre q par rapport à la variable Y est l'espérance mathématique m_{pq} de la variable aléatoire $X^p Y^q$.

Il est défini, sous réserve de l'existence de l'intégrale et en désignant par D le domaine de définition des variables aléatoires X et Y , par :

$$m_{pq} = \int_D \int x^p y^q f(x, y) \, dx \, dy$$

■ Cas particuliers

- Les moments des distributions marginales correspondent à :

$$m_{p0} = E(X) \text{ distribution marginale en } X$$

$$m_{0q} = E(Y) \text{ distribution marginale en } Y$$

- Les moments centrés, d'ordre p et q , sont les moments, notés μ_{pq} , des variables aléatoires centrées $[X - E(X)]$ et $[Y - E(Y)]$.
- Les moments centrés d'ordre 2 sont au nombre de trois et représentent la variance de la variable X , celle de la variable Y et la covariance entre ces variables :

$$\mu_{20} = \text{Var}(X) = \sigma_X^2 = \int_{D_X} [X - E(X)]^2 h(x) \, dx$$

$$\mu_{02} = \text{Var}(Y) = \sigma_Y^2 = \int_{D_Y} [Y - E(Y)]^2 g(y) \, dy$$

$$\mu_{11} = \text{Cov}(X, Y) = \int_D [X - E(X)] [Y - E(Y)] f(x, y) \, dx \, dy$$

■ Propriétés des moments centrés d'ordre 2

- La covariance peut être positive, négative ou nulle.
- On montre aisément que :

$$\text{Cov}(X, Y) = E(XY) - E(X) E(Y)$$

et si a et b sont des réels :

$$\text{Cov}(aX, bY) = ab \text{Cov}(X, Y)$$

$$\text{Var}(aX + bY) = a^2 \text{Var}(X) + b^2 \text{Var}(Y) + 2ab \text{Cov}(X, Y)$$

- Si X et Y sont deux variables aléatoires Lebesgue intégrables, on a :

$$[E(XY)]^2 \leq E(X^2)E(Y^2)$$

Cette propriété est la traduction de l'inégalité de Cauchy-Schwarz appliquée aux variables aléatoires $X - E(X)$ et $Y - E(Y)$.

■ Matrice de variances-covariances

On groupe les moments centrés d'une distribution en une matrice V , carrée, symétrique, appelée *matrice de variances-covariances*, qui se présente sous la forme suivante :

$$V = \begin{bmatrix} \mu_{2,0} & \mu_{1,1} \\ \mu_{1,1} & \mu_{0,2} \end{bmatrix} = \begin{bmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{bmatrix}$$

où ρ est le *coefficient de corrélation* défini et étudié dans le paragraphe 8.2.5.

■ Covariance de deux variables aléatoires indépendantes

Si les variables aléatoires X et Y sont indépendantes, leur covariance est nulle. Il en résulte que le coefficient de corrélation est nul aussi.

La réciproque n'est pas vraie en général (exemple 8.2).

Exemple 8.2 Covariance de deux variables aléatoires

Soit la variable aléatoire X prenant les valeurs $-2, -1, 1, 2$ avec les probabilités $1/4$. On considère la variable $Y = X^2$; Y prend donc les valeurs $1, 4$ avec les probabilités $1/2$.

Un calcul rapide donne : $E(X) = 0$; $E(Y) = 5/2$; $E(XY) = 0$

Donc $\text{Cov}(X, Y) = 0$ et cependant les variables X et Y ne sont pas indépendantes.

Si la covariance de deux variables aléatoires est nulle, ces variables ne sont pas nécessairement indépendantes (sauf pour des variables normales, voir paragraphe 8.4.3).

8.2.4 Moments conditionnels d'une distribution

Les résultats qui suivent supposent que l'espérance $E(Y)$ est *finie*. On démontre alors que $E(Y/X)$ existe et est égale à :

$$E(Y/X = x) = \int_{\mathbb{R}} y g(y/x) dy$$

La courbe définie par l'ensemble des couples $\{x, E(Y/X = x)\}$ est la *courbe de régression* de la variable aléatoire Y en la variable aléatoire X .

Les formules de l'espérance totale et de la variance totale démontrées pour les variables aléatoires discrètes s'appliquent aux variables aléatoires continues.

On définit d'une manière analogue l'espérance conditionnelle $E(X/Y)$.

Remarque

Si une des deux variables aléatoires est discrète et si l'autre possède une densité, il suffit de remplacer les intégrales par des sommes finies et les densités par des probabilités ponctuelles dans les expressions relatives à la variable aléatoire discrète.

8.2.5 Corrélation de la variable Y en la variable X

On définit deux mesures de liaison entre deux variables aléatoires.

– Le *coefficient de corrélation linéaire* est une *mesure de dépendance symétrique* :

$$\rho = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}$$

Des inégalités démontrées dans le paragraphe 8.2.3, on déduit que :

$$-1 \leq \rho \leq 1$$

Si $\rho = \pm 1$,

$$|\text{Cov}(X, Y)| = \sigma_X \sigma_Y$$

Il existe alors une *relation linéaire* entre les deux variables aléatoires X et Y :

$$Y - E(Y) = a [X - E(X)]$$

où a est une constante.

La nullité du coefficient de corrélation exclut la relation linéaire mais n'entraîne pas l'indépendance.

– Le *rapport de corrélation* est une *mesure de liaison non symétrique* :

$$\eta_{y/x}^2 = \frac{\text{Var}[E(Y/X)]}{\text{Var}(Y)}$$

Le théorème de la variance totale entraîne que le rapport de corrélation est compris entre 0 et 1.

$$\bullet \eta_{y/x}^2 = 1 \quad \text{si} \quad \text{Var}[E(Y/X)] = \text{Var}(Y)$$

Dans ces conditions (théorème de la variance totale), $E[\text{Var}(Y/X)] = 0$.

Comme $\text{Var}(Y/X) \geq 0$, il en résulte que $\text{Var}(Y/X) = 0$ presque sûrement, ou en d'autres termes, à une valeur de X fixée, $\text{Var}(Y/X) = 0$, Y ne prend qu'une valeur.

Donc, la variable aléatoire Y est *liée fonctionnellement* à la variable aléatoire X :

$$Y = \varphi(X)$$

$$\bullet \eta_{y/x}^2 = 0 \quad \text{si} \quad \text{Var}[E(Y/X)] = 0$$

Il en résulte que $E(Y/X)$ est presque sûrement une constante.

Donc, la variable aléatoire Y est *non corrélée* avec la variable aléatoire X .

C'est le cas, par exemple, si les variables aléatoires X et Y sont indépendantes mais la réciproque n'est pas vraie.

Remarque

Si $\eta_{y/x}^2 = \rho^2$, $E(Y/X) = aX + b$ (voir chapitre 20 sur la régression).

Exemple 8.3

Soit (X, Y) un couple de variables aléatoires continues défini sur le domaine D :

$$0 \leq X \leq 1 \qquad 0 \leq Y \leq 1 \qquad 0 \leq X + Y \leq 1$$

et de densité $f(x, y) = 2$.

– Loi du couple $f(x, y) = 2$ sur le domaine D et $f(x, y) = 0$ sinon,

– Loi marginale et moments de la variable X :

$$\text{Densité : } h(x) = 2 \int_0^{1-x} dy = 2(1-x)$$

$$\text{Fonction de répartition : } H(x) = 2 \int_0^x (1-x) dx = 2x - x^2$$

$$E(X) = 2 \int_0^1 x(1-x) dx = 1/3 \qquad \text{Var}(X) = E(X^2) - E(X)^2$$

$$E(X^2) = 2 \int_0^1 x^2(1-x) dx = 1/6 \qquad \text{Var}(X) = 1/18$$

– Loi marginale et moments de la variable Y , résultats analogues :

$$g(y) = 2(1-y) \quad G(y) = 2y - y^2$$

$$E(Y) = 1/3 \quad \text{Var}(Y) = 1/18$$

– Loi Y/X et moments conditionnels :

$$f(x, y) = g(y/x) b(x) \quad g(y/x) = \frac{1}{1-x}$$

$$E(Y/X) = \int_0^{1-x} \frac{y}{1-x} dy = \frac{1-x}{2}$$

$$\text{Var}(Y/X) = \int_0^{1-x} \frac{y^2}{1-x} dy - \left(\frac{1-x}{2}\right)^2 = \frac{(1-x)^2}{12}$$

– Loi X/Y et moments conditionnels, résultats analogues.

• Théorèmes de l'espérance totale et de la variance totale (vérification) :

$$E[E(Y/X)] = E\left(\frac{1-x}{2}\right) = \frac{1-E(X)}{2} = \frac{1-1/3}{2} = \frac{1}{3}$$

$$\text{Var}(Y) = \text{Var}[E(Y/X)] + E[\text{Var}(Y/X)] = \text{Var}\left(\frac{1-x}{2}\right) + \frac{1}{12}E[(1-x)^2]$$

$$\text{Var}(Y) = \frac{1}{4} \times \frac{1}{18} + \frac{1}{12} \left(1 - \frac{2}{3} + \frac{1}{6}\right) = \frac{1}{18}$$

• Coefficient de corrélation linéaire :

$$\rho = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}$$

$$\text{Cov}(X, Y) = E(XY) - E(X)E(Y) = 2 \int_0^1 x dx \int_0^{1-x} y dy - \frac{1}{9} = -\frac{1}{36}$$

$$\text{D'où } \rho = -\frac{1}{2}$$

• Rapport de corrélation de Y en X :

$$\eta_{Y/X}^2 = \frac{\text{Var}[E(Y/X)]}{\text{Var}(Y)} = \frac{1/72}{1/18} = \frac{1}{4}$$

8.3 Extension à des vecteurs aléatoires

8.3.1 Définition et loi conjointe

Soit $\underline{X} = (X_1, \dots, X_p)$ un vecteur aléatoire, c'est-à-dire une application mesurable de l'espace probabilisé $(\Omega, \mathcal{C}, \Pr)$ dans l'espace probabilisable $(\mathbb{R}^p, \mathcal{B}^p)$ muni de la mesure de Lebesgue λ_p . X_i est la i^{e} coordonnée du vecteur aléatoire $\underline{X}$. La fonction de répartition du vecteur $\underline{X}$ est l'application F de $\mathbb{R}^p$ dans $[0, 1]$ définie par :

$$F(a_1, a_2, \dots, a_p) = \Pr[(X_1 < a_1) \text{ et } (X_2 < a_2) \text{ et } \dots (X_p < a_p)]$$

Si le vecteur $\underline{X} = (X_1, \dots, X_p)$ est continu, il existe une fonction f de p variables, appelée *densité de probabilité* du vecteur $\underline{X}$ telle que pour tout domaine D de $\mathbb{R}^p$, on ait :

$$\Pr[(X_1, X_2, \dots, X_p) \in D] = \int_{(x_i) \in D} f(x_1, x_2, \dots, x_p) dx_1 dx_2 \dots dx_p$$

Entre la densité et la fonction de répartition, il existe la relation suivante :

$$f(x_1, x_2, \dots, x_p) = \frac{d^p F(x_1, x_2, \dots, x_p)}{dx_1, dx_2, \dots, dx_p}$$

que l'on démontre facilement dans le cas $p = 2$.

La distribution multinomiale est un exemple de distribution conjointe.

8.3.2 Moments

■ Espérance mathématique

C'est le vecteur, sous réserve de l'existence des espérances $E(X_i)$:

$$E(\underline{X}) = \begin{bmatrix} E(X_1) \\ \vdots \\ E(X_p) \end{bmatrix}$$

■ Matrice de variance-covariance

La matrice de variance-covariance est une matrice carrée Σ , $p \times p$, symétrique ayant pour terme général :

$$\sigma_{ik} = E\{[X_i - E(X_i)][X_k - E(X_k)]\}$$

La matrice de variance-covariance Σ s'écrit, en notant ${}^t[\underline{X} - E(\underline{X})]$ le vecteur ligne transposé du vecteur colonne $[\underline{X} - E(\underline{X})]$:

$$\Sigma = E \{ {}^t[\underline{X} - E(\underline{X})] [\underline{X} - E(\underline{X})] \}$$

Les éléments de la diagonale sont les variances, $\text{Var}(X_i) = \sigma_{x_i}^2$, des variables aléatoires X_i et les termes non diagonaux les covariances égales à, en introduisant le coefficient de corrélation entre deux variables, $\text{Cov}(X_i, X_k) = \rho_{ik} \sigma_i \sigma_k$.

$$\Sigma = \begin{bmatrix} \sigma_{x_1}^2 & \cdot & \cdot & \cdot & \rho_{1p} \sigma_1 \sigma_p \\ \rho_{21} \sigma_2 \sigma_1 & \sigma_{x_2}^2 & \cdot & \cdot & \rho_{2p} \sigma_2 \sigma_p \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \rho_{p1} \sigma_p \sigma_1 & \cdot & \cdot & \cdot & \sigma_{x_p}^2 \end{bmatrix}$$

8.3.3 Changement de variables

Soit $\underline{X} = (X_1, \dots, X_p)$ un vecteur aléatoire, application mesurable de $(\Omega, \mathcal{C}, \text{Pr})$ dans $(\mathbb{R}^p, \mathcal{B}^p)$ de densité $f \, 1_D$, où 1_D est la fonction caractéristique de l'ouvert D de $\mathbb{R}^p$.

Soit φ une application de l'ouvert D dans l'ouvert D' de $\mathbb{R}^q$. Cette application est supposée inversible, continûment différentiable ainsi que son inverse.

$Y = \varphi \circ X$ est donc un vecteur aléatoire, de dimension q , dont la densité est :

$$h(y) = f \circ \varphi^{-1}(y) \left| J\varphi^{-1} \right| 1_{D'}$$

où $J(\varphi^{-1}) = J^{-1}(\varphi)$ est le jacobien de φ^{-1} (déterminant de la matrice des dérivées premières, voir annexe 2).

Exemple 8.4

Soient X et Y deux variables aléatoires indépendantes suivant la loi normale centrée réduite, $N(0; 1)$. On considère la variable aléatoire $Z = X/Y$ et on veut définir sa densité de probabilité.

X et Y étant deux variables indépendantes, la densité du couple (X, Y) est :

$$f(x, y) = \frac{1}{2\pi} \exp\left(-\frac{x^2 + y^2}{2}\right)$$

On fait le changement de variables $x = x$ et $z = x/y$. Le jacobien de la transformation est :

$$J = \begin{vmatrix} 1 & 1/y \\ 0 & x/y^2 \end{vmatrix} = x/y^2 = \frac{z^2}{x}$$

D'où la densité du couple (X, Z) :

$$g(x, z) = \frac{1}{2\pi} \exp \left[-\frac{1}{2} \left(x^2 + \frac{x^2}{z^2} \right) \right] \frac{|x|}{z^2}$$

La densité de la variable Z s'obtient par intégration :

$$\begin{aligned} g(z) &= -\frac{1}{2\pi} \int_{-\infty}^0 \exp \left[-\frac{1}{2} \left(x^2 + \frac{x^2}{z^2} \right) \right] x/z^2 \, dx \\ &\quad + \frac{1}{2\pi} \int_0^{\infty} \exp \left[-\frac{1}{2} \left(x^2 + \frac{x^2}{z^2} \right) \right] x/z^2 \, dx \\ g(z) &= \frac{1}{\pi} \frac{1}{(1 + z^2)} \end{aligned}$$

On reconnaît la densité d'une variable aléatoire suivant une loi de Cauchy.

B

CALCUL DES PROBABILITÉS

8.4 Application : loi normale multidimensionnelle

8.4.1 Définitions

Soit $\underline{X}$ un vecteur aléatoire de $(\mathbb{R}^p, \mathcal{B}^p)$ et $\underline{U}$ un vecteur quelconque de $\mathbb{R}^p$.

Le vecteur $\underline{X}$ est un *vecteur normal ou gaussien* si ${}^t \underline{U} \underline{X}$ est une *variable aléatoire réelle*, suivant une *loi normale à une dimension*, quel que soit le vecteur $\underline{U}$.

Cette définition est équivalente à la suivante.

Toute *combinaison linéaire des composantes du vecteur $\underline{X}$* est une *variable aléatoire normale*.

On peut aussi définir la *loi normale multidimensionnelle par sa densité*.

Soit $\underline{m}$ un vecteur de $\mathbb{R}^p$ et Σ une matrice, $p \times p$, réelle, symétrique, définie positive. Le vecteur $\underline{X}$ de dimension p est un vecteur normal si sa densité est

donnée par :

$$f(x_1, x_2, \dots, x_p) = \frac{1}{(2\pi)^{p/2} \sqrt{\text{Det } \Sigma}} \exp \left[-\frac{1}{2} {}^t(\underline{X} - \underline{m}) \Sigma^{-1} (\underline{X} - \underline{m}) \right]$$

- $\underline{m}$ est le vecteur espérance mathématique $E(\underline{X})$ du vecteur $\underline{X}$.
- Σ est la matrice de variance-covariance du vecteur $\underline{X}$.

On note la loi du vecteur $\underline{X}$, $N(\underline{m}; \Sigma)$.

Les composantes du vecteur $\underline{X}$ sont indépendantes si et seulement si la matrice Σ est diagonale, c'est-à-dire si les composantes sont non corrélées.

8.4.2 Cas particulier : loi normale à deux dimensions

La matrice Σ de variance-covariance est la matrice :

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \rho \sigma_1 \sigma_2 \\ \rho \sigma_1 \sigma_2 & \sigma_2^2 \end{bmatrix}$$

$$\text{Det } \Sigma = \sigma_1^2 \sigma_2^2 (1 - \rho^2)$$

La densité $f(x_1, x_2)$ du vecteur $\underline{X}$ est donnée par :

$$\frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp \left\{ -\frac{1}{2(1-\rho^2)} \left[\left(\frac{x_1 - m_1}{\sigma_1} \right)^2 - 2\rho \frac{(x_1 - m_1)(x_2 - m_2)}{\sigma_1\sigma_2} + \left(\frac{x_2 - m_2}{\sigma_2} \right)^2 \right] \right\}$$

■ Propriétés

Pour trouver, rapidement et sans calculs, les lois marginales et conditionnelles, il suffit d'écrire cette densité sous l'une des deux formes suivantes, en supposant pour simplifier l'écriture, $m_1 = m_2 = 0$:

$$f(x_1, x_2) = \frac{1}{\sigma_1\sqrt{2\pi}} \exp \left(-\frac{x_1^2}{2\sigma_1^2} \right) \times \frac{1}{\sigma_2\sqrt{1-\rho^2}\sqrt{2\pi}} \exp \left(-\frac{(x_2 - \rho \frac{\sigma_2}{\sigma_1} x_1)^2}{2(1-\rho^2)\sigma_2^2} \right)$$

$$f(x_1, x_2) = \frac{1}{\sigma_2 \sqrt{2\pi}} \exp\left(-\frac{x_2^2}{2\sigma_2^2}\right) \times \frac{1}{\sigma_1 \sqrt{1-\rho^2} \sqrt{2\pi}} \exp\left(-\frac{\left(x_1 - \rho \frac{\sigma_1}{\sigma_2} x_2\right)^2}{2(1-\rho^2)\sigma_1^2}\right)$$

On passe d'une expression à l'autre en permutant les indices 1 et 2.

Sous la première forme, on voit que :

- la *distribution marginale de la variable aléatoire* X_1 est la loi normale de densité :

$$\frac{1}{\sigma_1 \sqrt{2\pi}} \exp\left(-\frac{x_1^2}{2\sigma_1^2}\right)$$

L'espérance mathématique est égale à 0 (ou à m_1) et l'écart-type est égal à σ_1 .

- la *distribution conditionnelle de la variable aléatoire* X_2 liée par X_1 est également une loi normale de densité :

$$\frac{1}{\sigma_2 \sqrt{1-\rho^2} \sqrt{2\pi}} \exp\left(-\frac{\left(x_2 - \rho \frac{\sigma_2}{\sigma_1} x_1\right)^2}{2(1-\rho^2)\sigma_2^2}\right)$$

L'espérance mathématique conditionnelle $E(X_2/X_1)$ a pour valeur $\rho \frac{\sigma_2}{\sigma_1} X_1$ et l'écart-type, $\sigma_2 \sqrt{1-\rho^2}$.

Cet écart-type ne dépend pas des valeurs de la variable X_1 . Les graphes des lois conditionnelles se déduisent donc l'un de l'autre par translation.

La *courbe de régression* est la droite d'équation :

$$E(X_2/X_1) = \rho \frac{\sigma_2}{\sigma_1} X_1$$

Sous la deuxième forme, on montre de même que la loi marginale de X_2 et la loi conditionnelle de la variable aléatoire X_1 liée par X_2 sont des lois normales.

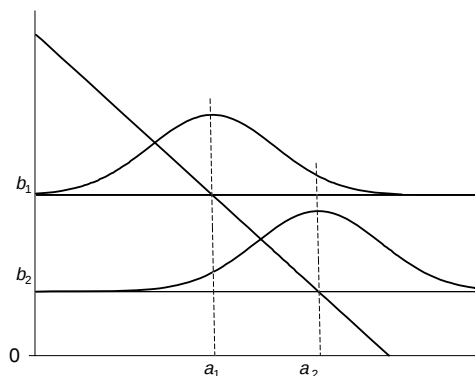


Figure 8.1 – Lois conditionnelles de la loi normale de dimension 2.

8.4.3 Condition nécessaire et suffisante d'indépendance de deux variables normales

Si $X = (X_1, X_2)$ est un *couple gaussien à valeurs dans $\mathbb{R}^2$* , il vérifie la propriété suivante :

$$X_1 \text{ et } X_2 \text{ indépendantes} \Leftrightarrow X_1 \text{ et } X_2 \text{ non corrélées}$$

La première partie du théorème est vraie quelle que soit la loi suivie par les variables aléatoires X_1 et X_2 , l'indépendance implique la non-corrélation ($\rho = 0$).

La réciproque est vraie seulement dans le cas gaussien. En effet, si les variables aléatoires X_1 et X_2 vérifient $\text{Cov}(X_1, X_2) = 0$, le coefficient de corrélation ρ est nul. Alors, la densité du couple (X_1, X_2) s'écrit :

$$f(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2} \exp \left\{ -\frac{1}{2} \left[\left(\frac{x_1 - m_1}{\sigma_1} \right)^2 + \left(\frac{x_2 - m_2}{\sigma_2} \right)^2 \right] \right\}$$

$$f(x_1, x_2) = \frac{1}{\sigma_1\sqrt{2\pi}} \exp \left[-\frac{1}{2} \left(\frac{x_1 - m_1}{\sigma_1} \right)^2 \right]$$

$$\times \frac{1}{\sigma_2\sqrt{2\pi}} \exp \left[-\frac{1}{2} \left(\frac{x_2 - m_2}{\sigma_2} \right)^2 \right]$$

On en conclut que les variables aléatoires X_1 et X_2 sont indépendantes et suivent des lois normales, $N(m_1; \sigma_1)$ et $N(m_2; \sigma_2)$, respectivement.

Attention

Deux variables aléatoires gaussiennes non corrélées ne sont pas nécessairement indépendantes. Pour que cette propriété soit vraie, il faut qu'elles forment un couple gaussien.

Exemple 8.5

X et Y sont deux variables aléatoires réelles centrées réduites, de coefficient de corrélation ρ . Ces variables vérifient :

$$E(X) = E(Y) = 0, \quad \text{Var}(X) = \text{Var}(Y) = 1 \quad \text{d'où} \quad \rho = \text{Cov}(X, Y)$$

Le couple (X, Y) est un couple gaussien. Sa densité est donc égale à :

$$f(x, y) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp \left[-\frac{1}{2(1-\rho^2)} (x^2 - 2\rho xy + y^2) \right]$$

On cherche la densité du couple $[X, (1-\rho^2)^{-1/2}(Y-\rho X)]$:

1) Le vecteur $T = [X, (1-\rho^2)^{-1/2}(Y-\rho X)]$ est un vecteur gaussien. En effet, le vecteur $Z = (X, Y)$ étant un vecteur gaussien, la variable $Y' = (1-\rho^2)^{-1/2}(Y-\rho X)$, qui est une combinaison linéaire des composantes de ce vecteur, est une variable gaussienne. On calcule facilement les moments suivants :

$$E(Y') = 0 \quad \text{Var}(Y') = \frac{1}{1-\rho^2} [1 + \rho^2 - 2\rho \text{Cov}(X, Y)] = 1$$

$$\begin{aligned} \text{Cov}(X, Y') &= E(XY') - E(X)E(Y') = (1-\rho^2)^{-1/2} E(X(Y-\rho X)) \\ &= (1-\rho^2)^{-1/2} [E(XY) - \rho E(X^2)] = 0 \end{aligned}$$

Le coefficient de corrélation entre X et Y' est donc nul.

2) La densité du couple $[X, (1-\rho^2)^{-1/2}(Y-\rho X)]$ est égale à (vecteur gaussien, X et Y' variables centrées réduites et coefficient de corrélation nul) :

$$\begin{aligned} g(x, y') &= \frac{1}{2\pi} \exp \left[-\frac{1}{2} (x^2 + y'^2) \right] \\ &= \frac{1}{\sqrt{2\pi}} \exp \left[-\frac{x^2}{2} \right] \times \frac{1}{\sqrt{2\pi}} \exp \left[-\frac{y'^2}{2} \right] \end{aligned}$$

Les variables X et Y' sont indépendantes.

Les propriétés (couple gaussien et non-corrélation) sont vérifiées.

9 • PROCESSUS ALÉATOIRES

Un système est à *évolution aléatoire* s'il peut prendre au cours du temps une série d'états successifs, sans qu'il soit possible d'en prévoir sa configuration exacte à un instant futur ; son évolution au cours du temps dépend donc du hasard. En d'autres termes, ces situations ne peuvent pas être étudiées en utilisant simplement le calcul des probabilités qui décrit des événements où le résultat possible de chaque épreuve est un nombre.

L'étude de ces systèmes évoluant d'une manière aléatoire avec le temps et présentant parfois un caractère périodique est un vaste sujet. Ces systèmes ont été étudiés par Markov (1906), qui a fait l'hypothèse que le passé et le futur étaient indépendants étant donné le présent (voir *processus et chaînes de Markov*, paragraphe 9.10). Puis les bases théoriques et mathématiques ont été formulées par Paul Lévy (1931), Doob (1933) après que Kolmogorov a élaboré la théorie mathématique du calcul des probabilités, résultant elle-même de la théorie de l'intégration. Ils ont des applications dans de nombreux domaines, en économie, en recherche opérationnelle, et peuvent intervenir dans l'étude de problèmes plus spécifiquement physiques. Donnons quelques exemples :

- l'état de la fortune d'un joueur dans un jeu de hasard,
- le débit journalier d'une rivière,
- en recherche opérationnelle, les problèmes de file d'attente, les arrivées de clients dans un service, le stock de pièces détachées dans un atelier,
- l'évolution démographique d'une population,
- la propagation d'une épidémie, etc.

9.1 Définitions

■ Définition 1

Un *processus stochastique* ou *aléatoire* représente la modélisation d'un phénomène évoluant au cours du temps. C'est une application de l'espace probabilisé $(\Omega, \mathcal{C}, \Pr)$ dans un espace probabilisable de fonctions $(\Omega', \mathcal{C}')$. Un processus associe à tout élément ω de Ω une fonction de la variable $t \in T$ telle que :

$$\omega \rightarrow X_t(\omega)$$

$X_t(\omega)$ étant l'application :

$$t \rightarrow X_t(\omega)(t) = X(\omega, t)$$

Un processus est donc décrit par une suite de variables aléatoires indexées par la variable t , on écrit $(X_t, t \in T)$ ou plus simplement (X_t) .

■ Définition 2

L'espace $(\Omega', \mathcal{C}')$ est appelé *espace des états* du processus.

- Si $(\Omega', \mathcal{C}')$ est l'espace $(\mathbb{R}, \mathfrak{R})$, le processus est *réel*.
- Si $(\Omega', \mathcal{C}')$ est l'espace $(\mathbb{R}^n, \mathfrak{R}^n)$, le processus est *multidimensionnel*.
- Si $\Omega' \subset \mathbb{Z}$, le processus est à *espace d'états discrets*.

■ Définition 3

Pour tout $\omega \in \Omega$ fixé, $X_t(\omega)$ est la *trajectoire* de la variable X pour l'individu ω .

- Si T est l'ensemble $\mathbb{R}$, le processus est *continu*.
- Si T est l'ensemble discret infini $\mathbb{Z}$, le processus est *discret*. $(X_t, t \in T)$ est défini soit par ... X_{m-1}, X_m ... soit par ... X_m, X_{m+1} ... ou encore par ... X_0, X_1 ...
- Si T est un intervalle de $\mathbb{Z}$, le processus $(X_t, t \in T)$ peut être appelé *famille à un paramètre*.
- Si t est fixé, $X_t(\omega)$ définit une variable aléatoire réelle.

Remarques

L'espace T des indices est souvent assimilé au temps, t est alors l'instant d'observation de la variable aléatoire X sur l'individu ω . Le processus est appelé *processus stochastique*.

Si la variable t prend des valeurs discrètes équidistantes, le processus décrit des *séries chronologiques*.

L'espace T ne représente pas toujours le temps, il peut avoir d'autres significations, comme le nombre de kilomètres parcourus, etc.

■ Définition 4

La loi du processus est la loi $\Pr_X$, loi image de $\Pr$ par X . Donc, $\forall t_1, t_2, \dots, t_k$ la loi du vecteur $(X_{t_1}, X_{t_2}, \dots, X_{t_k})$ est la loi marginale correspondante extraite de $\Pr_X$.

■ Définition 5

Soit un processus (X_t) indexé dans un ensemble $T \subset \mathbb{R}$.

La variable aléatoire $X(t_i) - X(t_j)$ où $t_i < t_j$ est l'*accroissement* du processus sur l'intervalle $[t_j, t_i[$.

9.2 Processus équivalents

Soient X et X^* deux processus admettant le même ensemble d'indices, le même espace des états (Ω', C') et définis respectivement sur $(\Omega_1, C_1, \Pr_1)$ et $(\Omega_2, C_2, \Pr_2)$. Ces deux processus sont *équivalents* si la relation :

$$\Pr(X_{t_1} \in B_1, X_{t_2} \in B_2, \dots, X_{t_k} \in B_k) = \Pr(X_{t_1}^* \in B_1, X_{t_2}^* \in B_2, \dots, X_{t_k}^* \in B_k)$$

est vérifiée pour tout système fini d'indices t_i extraits de T et pour tout ensemble fini de parties B_i extraites de C' .

Remarque

De la connaissance de cette relation à différents instants, le statisticien ou le probabiliste peut en déduire la loi d'évolution au cours du temps du processus aléatoire étudié.

9.3 Moments

Si la variable aléatoire (X_t) admet pour toutes les valeurs de t , une espérance mathématique m et une variance σ_t^2 finies, ces moments définissent des fonctions certaines du temps.

Deux processus ayant même moyenne et même variance peuvent avoir des évolutions très différentes au cours du temps. Cette plus ou moins grande régularité au cours du temps, qui provient de la structure temporelle, est résumée par deux caractéristiques, la *fonction de covariance* et la *fonction d'auto-corrélation*.

Ces fonctions sont définies pour deux instants t et s par :

– *Fonction de covariance* :

$$C(t, s) = \text{Cov}(X_t, X_s)$$

– *Fonction d'auto-corrélation* :

$$\rho(t, s) = \frac{\text{Cov}(X_t, X_s)}{\sigma_t \sigma_s}$$

9.4 Continuités

Un processus aléatoire est *continu en probabilité* au point t si :

$$\Pr(|X_{t+h} - X_t| > \varepsilon) \rightarrow 0 \quad \text{si } h \rightarrow 0$$

Cette condition exprime que la probabilité qu'une trajectoire soit discontinue en t est nulle. Cependant, la trajectoire d'un processus continu en probabilité peut être discontinue en tous les points.

Un processus aléatoire est *continu en moyenne quadratique* si :

$$E[(X_{t+h} - X_t)^2] \rightarrow 0 \quad \text{si } h \rightarrow 0$$

La continuité en moyenne quadratique implique la continuité de la moyenne, de la variance et de la fonction de covariance, mais n'implique pas la continuité des trajectoires.

9.5 Processus stationnaires

Les processus stationnaires vérifient des propriétés d'invariance, par translation du temps, très utilisées en économétrie.

9.5.1 Stationnarité stricte

La loi de probabilité d'un processus possédant la propriété de *stationnarité stricte* est invariante par translation sur t ; elle reste la même quand le temps passe. Il en résulte en particulier que X_{t+h} et X_t ont les mêmes caractéristiques quelle que soit la valeur de h à condition que $t + h$ appartient à l'ensemble T . La condition de stationnarité stricte s'écrit :

$$\begin{aligned} \Pr [(X_{t_1} < x_1) \cap \dots \cap (X_{t_n} < x_n)] \\ = \Pr [(X_{t_1+h} < x_1) \cap \dots \cap (X_{t_n+h} < x_n)] \quad \forall (t_i, h) \end{aligned}$$

C'est une condition difficile à réaliser. Elle entraîne les trois propriétés suivantes :

- $\forall t \quad E(X_t) = m$ l'espérance mathématique est constante,
- $\forall t \quad \text{Var}(X_t) = \sigma^2$ la variance est constante,
- $\forall (t, s) \quad C(t, s) = \text{Cov}(X_t, X_s) = \varphi(|t - s|)$.

9.5.2 Stationnarité faible

Un processus est *faiblement stationnaire*, propriété plus faible que la stationnarité stricte, si seulement les trois propriétés précédentes sont vérifiées.

9.5.3 Processus à accroissements stationnaires

Un processus est à *accroissements stationnaires* si la variable aléatoire $(X_{t+h} - X_t)$ est stationnaire pour toutes les valeurs de h .

Remarque

Les conditions de stationnarité, espérance et variance indépendantes de la variable t , sont des conditions difficiles à réaliser en économie. On cherche plutôt à rendre stationnaires les processus étudiés.

Exemple 9.1

Soit le processus aléatoire :

$$X_t = at + b + \varepsilon_t$$

Les variables aléatoires ε_t sont indépendantes et suivent la même loi telle que :

$$\forall t \quad E(\varepsilon_t) = 0 \quad \text{Var}(\varepsilon_t) = \sigma^2$$

On en déduit :

$$E(X_t) = at + b \quad \text{Var}(X_t) = E[(X_t - at - b)^2] = E[(\varepsilon_t)^2] = \text{Var}(\varepsilon_t) = \sigma^2$$

Ce processus *n'est pas stationnaire* ; pour le rendre *stationnaire*, il suffit de considérer le processus défini à partir des différences : $Y_t = X_t - X_{t-1} = a + \varepsilon_t - \varepsilon_{t-1}$. En effet :

$$E(Y_t) = a \quad \text{Var}(Y_t) = 2\sigma^2$$

$\text{Cov}(Y_{t+h}, Y_t) = E[(Y_{t+h})(Y_t)]$ car les variables Y_t sont des variables centrées. D'où les résultats :

$$\begin{aligned} \text{Cov}(Y_{t+h}, Y_t) &= -\sigma^2 & \text{si } h = \pm 1 \\ \text{Cov}(Y_{t+h}, Y_t) &= 2\sigma^2 & \text{si } h = 0 \\ \text{Cov}(Y_{t+h}, Y_t) &= 0 & \text{si } h \in \mathbb{Z} - \{-1, 1, 0\} \end{aligned}$$

Y_t est donc un processus *stationnaire*.

Exemple 9.2

Soit le processus X_t à valeurs dans $(-1, 1)$, tel que, $\forall t$:

$$\Pr(X_t = -1) = \Pr(X_t = 1) = 1/2$$

D'où les résultats :

$$\begin{aligned} E(X_t) &= 0 & \text{Var}(X_t) &= 1 \\ \text{Cov}(X_t, X_{t+h}) &= 1 \times 1/4 + 2(-1) \times 1/4 + 1 \times 1/4 = 0 \end{aligned}$$

Ce processus est donc un processus stationnaire.

Cependant, la meilleure prévision de X_t , c'est-à-dire $E(X_t)$, n'est pas à une valeur prise par X_t . Ce processus est donc *stationnaire mais imprévisible*.

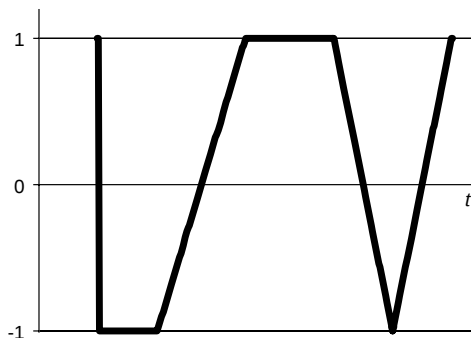


Figure 9.1 – Représentation graphique du processus de l'exemple 9.2.

Exemple 9.3

Soient les processus $X_t = \varepsilon_t Y_t = (-1)^t \varepsilon_t$ où ε_t est un processus stationnaire tel que :

$$E(\varepsilon_t) = 0 \quad \text{Var}(\varepsilon_t) = \sigma^2 \quad E(\varepsilon_t, \varepsilon_{t'}) = 0 \quad \forall t, t'$$

On considère le processus $S_t = X_t + Y_t$

Si t est pair : $S_t = 2\varepsilon_t \quad E(S_t) = 0 \quad \text{Var}(S_t) = 4\sigma^2$

Si t est impair : $S_t = 0 \quad E(S_t) = \text{Var}(S_t) = 0$

La somme de deux processus stationnaires n'est pas nécessairement un processus stationnaire.

Exemple 9.4

Soit le processus X_t défini pour $T \in \mathbb{N} - \{0\}$ et tel que :

$$\Pr(X_t = 0) = 1 - 1/t$$

$$\Pr(X_t = \sqrt{t}) = \Pr(X_t = -\sqrt{t}) = \frac{1}{2t}$$

Caractéristiques de ce processus :

$$E(X_t) = 0 \quad \text{Var}(X_t) = 1 \quad \text{Cov}(X_{t+h}, X_t) = E[(X_{t+h})(X_t)] = 0$$

Le dernier résultat est obtenu en remarquant que la variable X_t est une variable centrée et que les valeurs possibles du produit $X_{t+h}X_t$ sont soit 0, soit $\pm \sqrt{t+h}\sqrt{t}$ avec la même probabilité $1/4t(t+h)$.

Le processus X_t est un processus stationnaire au sens faible. Cependant, quand la valeur de t augmente, ce processus prend soit une valeur nulle avec une probabilité qui tend vers 1, soit les valeurs $\pm\sqrt{t}$ qui tendent vers l'infini mais avec une probabilité de plus en plus faible. En résumé, ce processus est stationnaire mais susceptible d'engendrer des *points aberrants*.

9.6 Exemples de processus aléatoires

De nombreux types de processus peuvent être définis et classés selon leurs propriétés. Parmi les exemples cités ci-dessous, certains seront étudiés dans les paragraphes suivants.

■ Processus à accroissements indépendants du temps

Soit un processus numérique (X_t) indexé dans un ensemble $T \subset \mathbb{R}$.

Un processus est à accroissements indépendants du temps si, quelle que soit la suite croissante d'indices $t_1, t_2, \dots, t_k$, les variables aléatoires $U_i = X(t_i) - X(t_{i-1})$, représentant les accroissements du processus sur les intervalles $[t_{i-1}, t_i]$, sont indépendantes en probabilité.

Si les variables $X_0, X_1 \dots$ définissent un tel processus, les différences $X_1 - X_0, X_2 - X_1 \dots$ sont mutuellement indépendantes, alors $X_n - X_0$ est la $n^{\text{ième}}$ somme partielle de la série $\sum_{m=1}^{\infty} (X_m - X_{m-1})$.

Inversement, si X_0 est une variable aléatoire arbitraire et si $Y_1, Y_2 \dots$ sont des variables aléatoires mutuellement indépendantes alors le processus aléatoire défini par :

$$X_n = X_0 + Y_1 + Y_2 + \dots + Y_n$$

est un processus à accroissements indépendants du temps.

■ Processus de Wiener-Lévy

Soit (X_t) un processus aléatoire défini pour $t \geq 0$, à accroissements indépendants du temps tel que :

- sur tout intervalle $]t', t''[$, l'accroissement de la fonction obéit à une loi normale, centrée, de variance $(t'' - t')$;

- cette fonction est presque sûrement continue mais ne possède pas de graphe car elle est presque sûrement sans dérivée et non bornée sur tout intervalle fini.

Ce schéma a été imaginé par Bachelier (1920) puis étudié par Wiener (1923) et par P. Lévy (1937-1948). Ce processus, très utile en pratique, intervient en particulier dans le calcul différentiel stochastique d'Ito utilisé dans le traitement du signal.

■ Martingale

Les martingales sont étudiées au paragraphe 9.7.

■ Mouvement brownien

Le mouvement brownien est étudié au paragraphe 9.8.

■ Promenades aléatoires ou marches au hasard

Elles sont étudiées au paragraphe 9.9.

■ Processus et chaînes de Markov

Leur étude est faite au paragraphe 9.10.

■ Processus ponctuels

Les processus ponctuels ainsi que leurs applications aux files d'attente sont étudiés au paragraphe 9.11.

9.7 Martingale

Une martingale évoque une stratégie élaborée en vue de gagner aux jeux de hasard. Cette notion a été étudiée la première fois par Abraham de Moivre (*The doctrine of chance*, 1718). Cependant, la martingale la plus classique, dite de d'Alembert, relative à un jeu de pile ou face, ne peut être utilisée pour gagner que dans l'hypothèse où la fortune du joueur est infinie, sinon la probabilité de gagner n'est plus égale à 1. La formulation mathématique est due à P. Lévy (1935-1937) et à J. Ville (1939) car le lien entre la théorie de l'intégration et le calcul des probabilités venait d'être mis en évidence par Kolmogorov. Puis Doob a élaboré la théorie de la convergence des martingales.

9.7.1 Définition mathématique d'une martingale

Une suite (X_n) de variables aléatoires définies sur un espace probabilisé $(\Omega, C, \Pr)$ est une *martingale* si :

- l'espérance $E(X_n|)$ est finie pour toute valeur de t .
- $E(X_{n+1}/X_1 = x_1, \dots, X_n = x_n) = x_n \quad \forall n \geq 1$ p.s. (presque sûrement).

Cette suite est une *sous-martingale* si :

$$E(X_{n+1}/X_1 = x_1, \dots, X_n = x_n) = x_n \quad \forall n \geq 1 \text{ p.s.}$$

9.7.2 Propriétés

Si une suite est une sous-martingale, elle vérifie l'inégalité de martingale :

$$\Pr(\text{Max } X_k > \lambda) \leq 1/\lambda E(X_n^+) \quad 1 \leq k \leq n$$

Si une suite est une martingale et s'il existe un entier $\alpha \geq 1$ tel que :

$$n \geq 1 \quad E(|X|^\alpha) < \infty$$

alors la suite $(|X|^\alpha)$ est une sous-martingale. On en déduit :

$$\Pr(|X_k| > \alpha) \leq 1/\lambda^\alpha E(|X_n|^\alpha) \quad 1 \leq k \leq n$$

Exemple 9.5

Soit une suite de variables aléatoires $X_1, X_2, \dots$. On suppose que ces variables peuvent être écrites sous la forme :

$$X_n = Y_1 + Y_2 + \dots + Y_n \quad n \geq 1$$

Les variables Y_i sont mutuellement indépendantes et telles que $E(|Y_i|) < \infty$ pour $n \geq 1$ et $E(Y_j) = 0$ si $j > 1$. Le processus défini par les variables X_n est une martingale.

9.7.3 Domaines d'application

■ Martingales et processus stochastiques

Les propriétés des martingales sont utilisées dans de nombreux processus intervenant dans l'étude de phénomènes réels comme le mouvement brownien et le processus de Poisson.

■ Martingales en analyse

La théorie des martingales permet d'apporter des solutions dans certains domaines mathématiques comme la recherche de fonctions harmoniques.

■ Problème du filtrage

Pour donner une estimation « raisonnable » d'un signal inconnu évoluant au cours du temps et déformé par des perturbations aléatoires, on peut utiliser la théorie des processus aléatoires. Soit :

- Y_t l'observation au temps t ,
- S_t le signal au temps t en réalité $S_t(t, \omega)$,
- B_t un mouvement brownien, donc une martingale, représentant la perturbation à l'instant t .

Ces fonctions vérifient le système d'équations :

$$\begin{aligned} dS_t(t, \omega) &= A[t, S_t(t, \omega)]dt \\ dY_t &= a(t, S_t, Y_t)dt + b(t)dB_t \end{aligned}$$

Pour ω fixé, la première équation est une équation différentielle ordinaire, il n'en est pas de même pour la deuxième à cause de la présence du terme dB_t . Il faut introduire la théorie des martingales en temps continu.

9.8 Mouvement brownien

9.8.1 Définition

Un mouvement brownien est :

- un processus à accroissements indépendants du temps et stationnaires,
- un processus suivant une loi normale (processus gaussien),
- un processus de Markov,
- une martingale.

9.8.2 Mouvement brownien et fractale

Un tel mouvement est défini de la façon suivante. Une série de direction quelconque et de longueur donnée définit un ensemble de points visités de

dimension fractale égale à R quel que soit l'ensemble de départ si la longueur est aléatoire, de moyenne finie (volume de Raleigh).

On peut généraliser et définir un mouvement brownien fractionnaire : il est représenté par une fonction dont les accroissements ont une distribution normale (gaussienne).

9.9 Marche au hasard

9.9.1 Définition d'une marche au hasard sur un axe

On considère une particule qui se déplace sur un axe, d'une unité à chaque pas, dans un sens ou dans l'autre, avec des probabilités égales, après avoir reçu une légère impulsion :

$$\text{-----} O \text{-----} P_n \text{-----}$$

Au temps $t = 0$, le mobile est en O et au temps $t = n$, il est au point P_n . On pose :

$$\overrightarrow{OP_n} = x_n = U_1 + U_2 + \dots + U_n$$

Les variables aléatoires U_i prennent les valeurs $+1$ ou -1 avec la même probabilité.

On définit ainsi un processus aléatoire à accroissements indépendants du temps car les variables aléatoires $U_i = x_i - x_{i-1}$ sont indépendantes.

Ce processus est un *processus discret*. Pour étudier un *processus continu*, il est nécessaire de faire un passage à la limite. On considère, entre les instants t et $t + 1/n$, un déplacement aléatoire égal soit à $1/\sqrt{n}$ soit à $-1/\sqrt{n}$ avec une probabilité égale à $1/2$ pour chaque déplacement. La fonction caractéristique de ce déplacement est $\cos t/\sqrt{n}$. Les déplacements étant indépendants, la somme de n déplacements pendant la durée h a pour fonction caractéristique $(\cos t/\sqrt{n})^{hn}$ et pour limite :

$$n \rightarrow \infty \quad (\cos t/\sqrt{n})^{hn} \rightarrow \exp(-ht^2/2)$$

On reconnaît la fonction caractéristique d'une loi normale $N(0; \sqrt{h})$. C'est la définition du processus de Wiener-Lévy.

9.9.2 Généralisation

1. On peut généraliser et définir la *marche au hasard dans un plan*, en considérant le déplacement d'un mobile sur un quadrillage du plan. À chaque sommet du quadrillage, le mobile a une probabilité égale à $1/4$ de partir dans une des quatre directions possibles. On définit ainsi un mouvement brownien à deux dimensions. Soit $\Pr(x, y)$ la probabilité pour que le chemin passe par le point de coordonnées (x, y) . Cette probabilité est définie par :

$$\Pr(x, y) = \frac{1}{4} [\Pr(x-1, y) + \Pr(x+1, y) + \Pr(x, y-1) + \Pr(x, y+1)]$$

Si le côté du quadrillage tend vers 0, $\Pr(x, y)$ est solution de l'équation de Laplace :

$$\frac{\partial^2 P}{\partial x^2} + \frac{\partial^2 P}{\partial y^2} = 0$$

Cette solution est une fonction harmonique, solution du problème de Dirichlet dans le plan.

2. Ce problème peut encore se généraliser en considérant un *réseau d'un espace euclidien de dimension n* . En chacun des sommets de ce quadrillage, le mobile a une probabilité égale à $1/2n$ de partir dans une des $2n$ directions possibles. Enfin, une autre généralisation est possible si l'on suppose que les accroissements indépendants, c'est-à-dire les variables U_i , obéissent à une loi quelconque.

Ce processus peut aussi représenter la fortune d'un joueur qui renouvelle des paris successifs ou l'évolution de la situation financière d'une compagnie d'assurances, etc.

9.10 Processus et chaînes de Markov

En calcul des probabilités, on fait très souvent l'hypothèse que toutes les variables aléatoires dont dépend le phénomène étudié sont indépendantes. Or, cette hypothèse très utile conduisant à des résultats corrects, ne peut pas décrire toutes les situations. L'étude des processus aléatoires a montré que l'état d'un système à l'instant t_n dépend en général de son comportement aux instants antérieurs $t_1, \dots, t_n$. Markov a étudié un cas particulier et fait l'hypothèse que l'évolution future du système ne dépend que de l'instant présent.

9.10.1 Définition d'un processus de Markov

Un *processus de Markov* ou *processus markovien* peut se décrire de la façon simple suivante.

L'état d'un système est connu à l'instant t ; on suppose que les informations sur le comportement du système avant l'instant t sont sans influence sur les prédictions relatives à l'évolution de ce système après l'instant t . Pour un « présent » donné, le « passé » et le « futur » sont indépendants. Le temps d'arrêt du système peut lui-même être aléatoire.

En termes mathématiques, cette propriété se traduit de la façon suivante. Soit un système, observé en une suite discrète d'instants $T = (t_1, \dots, t_n)$ et ne pouvant prendre qu'une suite d'états en nombre fini ou dénombrable. On note X_n l'état du système à l'instant t_n .

Un *processus markovien* est tel que :

$$\Pr(X_n = j_n / X_0 = j_0, \dots, X_{n-1} = j_{n-1}) = \Pr(X_n = j_n / X_{n-1} = j_{n-1})$$

L'ensemble T étant une suite discrète, on dit soit *processus* soit *chaîne* de Markov.

C'est une chaîne de Markov *d'ordre 1*. On définit de la même façon une chaîne de Markov *d'ordre 2* par la relation :

$$\begin{aligned} \Pr(X_n = j_n / X_0 = j_0, \dots, X_{n-1} = j_{n-1}) \\ = \Pr(X_n = j_n / X_{n-2} = j_{n-2}, X_{n-1} = j_{n-1}) \end{aligned}$$

et plus généralement une chaîne de Markov *d'ordre r* . Par un changement de variables, il est possible de se ramener à une chaîne de Markov d'ordre 1.

Un processus de Markov est appelé *discret* si la suite T des instants est dénombrable, il est *continu* si la suite des instants est continue.

Relativement à l'ensemble $\mathcal{E}$ des états, on distingue trois cas, selon que l'ensemble $\mathcal{E}$ est fini, c'est-à-dire $\mathcal{E} = (1, 2, \dots, k)$ ou dénombrable $\mathcal{E} = \mathbb{N}$ ou continu $\mathcal{E} = \mathbb{R}$.

9.10.2 Chaîne de Markov homogène

On étudie le cas où l'ensemble $\mathcal{E}$ des états est fini $\mathcal{E} = (1, 2, \dots, k)$ et l'ensemble T discret. Une chaîne de Markov est *homogène* si les lois conditionnelles sont invariantes par translation sur l'échelle des temps, c'est-à-dire si la probabilité conditionnelle :

$$\Pr(X_n = j / X_{n-1} = i)$$

ne dépend que de l'intervalle $t_n - t_{n-1}$ et est indépendante de n . On définit ainsi une probabilité de transition de l'état i à l'état j notée p_{ij} . La *matrice de probabilité de transition* M ou *matrice de passage* ou encore *matrice de Markov* est la matrice carrée, d'ordre k , dont les éléments sont les probabilités (p_{ij}) , i étant l'indice des lignes et j celui des colonnes. Une chaîne de Markov homogène est parfaitement définie si on connaît la loi initiale des états, c'est-à-dire le vecteur colonne P_0 qui a pour composantes les probabilités : $\Pr(X_0 = l)$ pour $l = 1, \dots, k$ ainsi que la matrice M .

■ Propriétés de la matrice de transition

La matrice de transition est une *matrice stochastique* : elle vérifie en effet les conditions suivantes :

$$\forall i, j \quad p_{ij} \geq 0 \quad \sum_{j=1}^k p_{ij} = 1$$

On en déduit les propriétés suivantes.

Les valeurs propres sont égales ou inférieures à 1. Soit V un vecteur dont toutes les composantes sont égales à 1, il possède la propriété suivante : $MV = V$. Donc V est un vecteur propre de la matrice M associé à la valeur propre 1. On démontre que les autres valeurs sont toutes inférieures ou égales à 1.

Produit de deux matrices stochastiques. Le produit de deux matrices stochastiques est une matrice stochastique et plus généralement M^r est une matrice stochastique :

$$MV = V \Rightarrow M^2V = V \Rightarrow \dots \Rightarrow M^rV = V \quad \forall r$$

Matrice de transition en r étapes. La matrice de transition en r étapes est la matrice M^r . En effet :

$$\begin{aligned} p_{ij}^{(2)} &= \Pr(X_n = j / X_{n-2} = i) = \sum_{l=1}^k \Pr(X_{n-2} = j, X_{n-1} = l / X_{n-2} = i) \\ &= \sum_{l=1}^k \Pr(X_n = j / X_{n-2} = i, X_{n-1} = l) \times \Pr(X_{n-1} = l / X_{n-2} = i) \\ &= \sum_{l=1}^k p_{lj} p_{il} \end{aligned}$$

On en déduit, en désignant par $M^{(2)}$, que $M^{(2)} = M^2$ et de même, après r étapes, $M^{(r)} = M^r$.

■ Loi de probabilité de l'état X_n

Supposons connu l'état du système à un instant quelconque, qui peut éventuellement être l'instant initial, et notons $I_{(0)}$ cette distribution de probabilité. Pour la distribution de probabilité $I_{(n)}$ l'état X_n est donné par :

$$\begin{aligned} \Pr(X_n = j) &= \sum_k \Pr(X_0 = k) \times \Pr(X_n = j / X_0 = k) \\ I_{(n)} &= I_{(0)} M^n = I_{(n-1)} M \end{aligned}$$

■ Classifications des états

1. Un état j est *accessible* à partir d'un état i s'il existe un entier $l \geq 0$ tel que $p_{ij}^{(l)} > 0$.
2. Deux états i et j mutuellement accessibles sont appelés *communicants*, on écrit : $i \leftrightarrow j$; il existe donc deux entiers l et r tels que $p_{ij}^{(l)} > 0$ et $p_{ji}^{(r)} > 0$. La propriété *états communicants* définit sur l'ensemble des états de la chaîne une *relation d'équivalence*. En effet, cette relation est :
 - réflexive si on pose : $p_{ij}^{(0)} = \delta_{ij}$ ($\delta_{ij} = 1$ si $i = j$ et $\delta_{ij} = 0$ si $i \neq j$);
 - symétrique, par définition;
 - transitive; pour démontrer cette propriété, il suffit de revenir à la définition.

À partir de cette relation d'équivalence, on définit, sur l'ensemble des états, des *classes d'équivalence* ; il est possible de quitter une classe d'équivalence mais il est impossible d'y retourner.

3. Une chaîne est *irréductible* si elle ne possède qu'une seule classe d'équivalence.

Exemple 9.6

Soit une chaîne de Markov à cinq états notés 1, 2, 3, 4, 5 dont la matrice de probabilité de transition est la suivante :

	1	2	3	4	5
1	0	0	1	0	0
2	0,25	0	0	0,75	0
3	1	0	0	0	0
4	0	0	0	0	0,5
5	0	0	0	0	0,4

C'est une chaîne réductible avec deux classes (1, 2, 3) et (4, 5).

Exemple 9.7

Soit une chaîne de Markov à trois états notés 1, 2, 3 dont la matrice de transition est la suivante :

	1	2	3
1	0,5	0	0,5
2	0,25	0,25	0,5
3	0	0,25	0,75

C'est une chaîne irréductible, elle a une seule classe.

4. Un état i est *périodique* et a pour période $\alpha(i)$ si $\alpha(i)$ est le plus grand commun diviseur de tous les entiers $r \geq 1$ tels que $p_{ii}^{(r)} > 0$.

Deux états communicants ont la même période, donc tous les états d'une même classe ont la même période.

Les états (1, 2, 3) de la chaîne de Markov de l'exemple 9.6 ont pour période 2.

Exemple 9.8

Soit la chaîne de Markov à trois états notés 1, 2, 3 dont la matrice de transition est la suivante :

	1	2	3
1	0	0	1
2	1	0	0
3	0	1	0

Chaque état a pour période 3.

5. Une chaîne est dite *apériodique* si tous les états de la chaîne ont pour période 1.

6. *États récurrents ou persistants ou se renouvelant.* Un état i est *récurrent* si, partant de cet état, on y revient presque sûrement. Un état i est récurrent si et seulement si $\sum_{n=1}^{\infty} p_{ii}^{(n)} = \infty$ (propriété caractéristique).

Partant d'un état récurrent, on ne peut atteindre que des états récurrents. Un état récurrent est atteint presque sûrement une infinité de fois.

Si l'état i est récurrent et si i et j sont deux états communicants, alors j est aussi un état récurrent. Tous les états d'une même classe d'équivalence sont donc tous récurrents ou tous non récurrents.

Un état non récurrent est appelé état *transitoire*.

9.10.3 Théorème limite pour les chaînes apériodiques récurrentes

Une chaîne irréductible, récurrente, apériodique, possède les propriétés suivantes :

$$\lim_{n \rightarrow \infty} p_{ii}^{(n)} = \frac{1}{\mu_i} \quad \text{où} \quad \mu_i = \sum_{n=1}^{\infty} n f_{ii}^{(n)} \quad \text{est le temps moyen de retour en } i$$

$$\lim_{n \rightarrow \infty} p_{ji}^{(n)} = \frac{1}{\mu_i}$$

$f_{ii}^{(n)}$ est la probabilité pour que, partant de l'état i , on y revienne pour la première fois après n transitions.

La matrice de transition d'une telle chaîne tend vers une matrice limite dont toutes les lignes sont égales, cette valeur est l'unique solution du système

$I^* = I^* M$ avec la condition que la somme des composantes de I^* soit égale à l'unité. Ces chaînes tendent donc vers un régime d'équilibre indépendant de l'état initial.

9.10.4 Théorème limite pour les états transitoires

Un état transitoire j vérifie les propriétés suivantes :

- $p_{jj}^{(n)} \rightarrow 0$ car $\sum_{n=1}^{\infty} p_{jj}^{(n)} < \infty$
- Si i est un état récurrent $p_{ij}^{(n)} = 0 \quad \forall n$ donc, d'un état récurrent, on ne peut pas atteindre des états transitoires.
- Si k est un autre état transitoire : $p_{kj}^{(n)} \xrightarrow{n \rightarrow \infty} 0$

9.10.5 Processus de Markov à espace d'états continu

On généralise l'étude des processus de Markov au cas où l'espace des états est continu, le temps T restant discret. On considère des transitions d'un point x à un intervalle A de $\mathbb{R}$ ou en généralisant encore à un domaine D de $\mathbb{R}^n$. On ne définit plus des probabilités de transition mais des densités de transition.

■ Densités de transition

On étudie le cas où l'espace des états est $\mathbb{R}$. Soit X_n (X_n appartient à $\mathbb{R}$), l'état du système à l'instant n . La suite X_n est *une suite de Markov*, d'ordre 1, si :

$$\begin{aligned} \Pr(X_n \in A / X_{n-1} = x_{n-1}, X_{n-2} = x_{n-2}, \dots, X_0 = x_0) \\ = \Pr(X_n \in A / X_{n-1} = x_{n-1}) \end{aligned}$$

quels que soient n, x_i et l'intervalle A de $\mathbb{R}$.

$F(x, A)$ est la probabilité de transition d'un point x à un point appartenant à A , c'est-à-dire :

$$F(x, A) = \Pr(X_n \in A / X_{n-1} = x)$$

Si la fonction $F(x, A)$ a une densité $f(x, y)$, alors $f(x, y)dy$ est la probabilité de transition, en une étape de x à l'intervalle $[y, y + dy]$.

■ Chaînes gaussio-markoviennes

On considère une suite de variables aléatoires $X_i (i = 1, \dots, r)$ vérifiant les conditions suivantes $\forall i, j$:

$$E(X_i) = 0 \quad \text{Var}(X_i) = \sigma_i^2$$

$$\text{Cov}(X_i X_j) = E(X_i X_j) = \rho_{ij} \sigma_i \sigma_j$$

ρ_{ij} étant le coefficient de corrélation linéaire entre les variables X_i et X_j .

On suppose que la suite $X_i (i = 1, \dots, r)$ des variables aléatoires suit une loi normale de dimension r . Cette suite est une *suite markovienne* si pour tout $k \leq r$ la loi conditionnelle de X_k pour $X_1, \dots, X_{k-1}$ fixées est identique à la loi conditionnelle de X_k pour X_{k-1} fixée.

□ Théorème 1

La condition nécessaire et suffisante pour que la suite $X_i (i = 1, \dots, r)$ soit une suite gaussio-markovienne est que pour $k \leq r$:

$$E(X_k / X_1, \dots, X_{k-1}) = E(X_k / X_{k-1})$$

□ Théorème 2

La condition nécessaire et suffisante pour que la suite $X_i (i = 1, \dots, r)$ soit une suite gaussio-markovienne est que pour $j < q < k \leq r$:

$$\rho_{jk} = \rho_{jq} \rho_{qk}$$

9.10.6 Applications

■ Processus gaussiens à accroissements indépendants

Une suite finie ou infinie de variables aléatoires X_k suivant une loi normale centrée (donc $E(X_k) = 0$) est une *chaîne à accroissements indépendants* si pour tout $j < k$, l'accroissement $X_k - X_j$ est indépendant de $X_1, \dots, X_j$. Donc :

$$E[X_j (X_k - X_j)] = E(X_j) E(X_k - X_j) = 0 \quad j < k$$

$$E[X_j (X_k - X_j)] = E(X_j X_k) - E(X_j^2) = 0 \Rightarrow E(X_j X_k) = E(X_j^2) = \sigma_j^2$$

La définition du coefficient de corrélation et les propriétés précédentes entraînent :

$$\rho_{jk} = \frac{\text{Cov}(X_j, X_k)}{\sigma_j \sigma_k} = \frac{E(X_j X_k) - E(X_j) E(X_k)}{\sigma_j \sigma_k} = \frac{\sigma_j^2}{\sigma_j \sigma_k} = \frac{\sigma_j}{\sigma_k}$$

Une telle chaîne vérifie donc le théorème 2, c'est une chaîne markovienne.

■ Modèles auto-régressifs

□ Théorème

La condition nécessaire et suffisante pour que la suite X_i ($i = 1, \dots, r$) soit gaussio-markovienne est qu'il existe une suite de variables aléatoires, mutuellement indépendantes, suivant une loi normale centrée réduite.

Soit une chaîne gaussio-markovienne. Il existe une constante a_k telle que l'accroissement $X_k - a_k X_{k-1}$ soit indépendant de X_{k-1} , donc de $X_1, \dots, X_{k-2}$.

Soit $\lambda_k^2 = \text{Var}(X_k - a_k X_{k-1})$ et soit Z_i la suite des variables aléatoires définies par les relations suivantes :

$$X_1 = \lambda_1 Z_1 \quad \dots \quad X_k = a_k X_{k-1} + \lambda_k Z_k \quad k = 2, 3, \dots$$

$$\text{D'où : } E(Z_i) = 0 \quad \text{Var}(Z_i) = E(Z_i^2) = 1$$

Réciproquement, si on considère une suite de variables aléatoires Z_i indépendantes suivant une loi normale centrée réduite, alors les variables X_i définies par les relations précédentes sont gaussio-markoviennes.

9.11 Processus ponctuels

9.11.1 Définition d'un processus ponctuel

Un processus est *ponctuel* si on peut attribuer une date précise à chaque événement observé et si leur réalisation est supposée instantanée (événement ponctuel).

Exemples

Accidents du travail dans un atelier.

Débuts de pannes d'un ensemble de machines.

Instants de naissance d'une particule ou d'un individu d'une population.

Arrivées des bateaux dans un port, des trains dans une gare, etc.

La réalisation d'un tel phénomène est caractérisée, soit par la donnée d'une suite croissante d'instant t_n , dates de réalisation des événements, soit par la donnée d'une suite de points P_n sur l'échelle des temps avec $\overline{OP} = t_n$ d'où le nom de *processus ponctuel*.

9.11.2 Cas particulier : processus ponctuel de Poisson

■ Définition

Les processus ponctuels de Poisson interviennent dans différents domaines, en particulier dans la théorie des files d'attente.

Une suite d'événements $E_1, E_2, \dots, E_n$ constitue un *processus ponctuel de Poisson* si ces événements satisfont aux trois conditions suivantes :

- Les événements attendus se produisent indépendamment les uns des autres, c'est-à-dire les temps d'attente $(E_1, E_2), \dots, (E_k, E_{k+1})$ sont des variables aléatoires indépendantes. C'est un *processus sans mémoire*.
- La loi du nombre N_t d'événements se produisant dans l'intervalle $(t, t + h)$ ne dépend que de h , c'est-à-dire qu'elle ne dépend pas du nombre de réalisations au cours des intervalles antérieurs. Si $h = 1$, on note c l'*espérance mathématique* ou *cadence* de la loi N_t .
- Deux événements ne peuvent pas se produire simultanément, donc la probabilité que l'événement se réalise plus d'une fois dans un intervalle de temps petit Δt est négligeable.

■ Lois associées à un processus de Poisson

□ Calcul de la probabilité $\text{Pr}_0(h)$

La probabilité $\text{Pr}_0(h)$ qu'il ne se produise aucun événement pendant un intervalle de temps égal à h ne dépend que de h (condition 2). Soient trois instants $t, (t + h), (t + h + k)$. Les conditions 2 et 1 (indépendance des événements) et le théorème des probabilités totales entraînent :

$$\text{Pr}_0(h + k) = \text{Pr}_0(h) \times \text{Pr}_0(k)$$

Cette équation a pour solution :

$$\text{Pr}_0(h) = e^{-ch}$$

□ Étude de la durée T séparant deux événements consécutifs

E_k et E_{k+1}

La probabilité pour qu'aucun événement ne se produise pendant l'intervalle de temps t est égale à :

$$\text{Pr}(T > t) = e^{-ct}$$

D'où la fonction de répartition et la densité de la loi de la variable aléatoire T :

$$F(t) = \Pr(T < t) = 1 - e^{-ct} \quad f(t) = ce^{-ct} \quad \forall t \geq 0$$

La loi de la variable T est donc une loi exponentielle de paramètre c . Les propriétés de cette loi sont données au chapitre 6 (paragraphe 6.3). La variable aléatoire $Z = cT$ a pour densité :

$$g(z) = e^{-z}$$

La loi de la variable Z est donc une *loi gamma* γ_1 (voir chapitre 6, paragraphe 6.4).

□ Étude de la durée Y séparant $(n + 1)$ événements consécutifs

La variable aléatoire Y séparant $(n + 1)$ événements consécutifs est la variable nT . La variable $V = cY = cnT = nZ$ suit donc une *loi gamma* γ_n de densité :

$$g(v) = \frac{1}{\Gamma(n)} \exp(-v) v^{n-1}$$

Il en résulte la densité de la loi Y :

$$f(y) = \frac{c}{\Gamma(n)} \exp(-cy) (cy)^{n-1}$$

La durée Y séparant $(n + 1)$ événements consécutifs suit donc la loi gamma $\gamma(n, c)$ ou *loi d'Erlang* car n est un entier.

□ Étude du nombre N d'événements se produisant pendant un intervalle de temps fixé

La probabilité qu'il se produise n événements pendant une période de temps fixée T est donnée par :

$$\Pr(N = n) = \Pr(N \geq n) - \Pr(N \geq n + 1)$$

La loi de la variable Y est la loi $\gamma(n, c)$ donc :

$$\Pr(Y > y) = \frac{1}{\Gamma(n)} \int_0^y e^{-cy} (cy)^{n-1} c dy$$

D'où :

$$\Pr(N > n + 1/T) = \frac{1}{\Gamma(n)} \int_0^T e^{-cy} (cy)^{n-1} c dy$$

$$\Pr(N = n) = \frac{1}{\Gamma(n)} \int_0^T e^{-cy} (cy)^{n-1} c dy - \frac{1}{\Gamma(n+1)} \int_0^T e^{-cy} (cy)^n c dy$$

Après intégration par parties de la première intégrale, on obtient :

$$\Pr(N = n) = \frac{e^{-cT} (cT)^n}{n!}$$

La loi de N est la *loi de Poisson de paramètre cT* .

Il en résulte que $E(N) = cT$ et, si $T = 1$, $E(N) = c$. On retrouve la propriété du paramètre d'une loi de Poisson, c'est la cadence définie au début du paragraphe, espérance mathématique du nombre d'événements aléatoires observés par unité de temps.

Remarques

Dans un processus de Poisson, la loi de Poisson correspond de manière rigoureuse à la distribution du nombre d'événements pendant un temps donné alors que, dans l'étude des épreuves répétées, la loi de Poisson est une approximation de la loi binomiale.

Si un événement se réalise suivant un processus de Poisson, c'est-à-dire si le nombre d'événements survenant pendant un temps T fixé suit une loi de Poisson de paramètre cT , le temps séparant deux événements consécutifs suit une loi exponentielle de paramètre c .

Réciproquement, si le délai s'écoulant entre deux événements est distribué suivant une loi exponentielle, le processus ponctuel est régi par une loi de Poisson.

□ Étude de la répartition des dates $E_1, E_2, \dots$ des événements dans un intervalle de temps donné T

On calcule facilement les probabilités suivantes en revenant aux définitions :

$$\Pr(t_1 < E_1 < t_1 + dt_1) = c e^{-c t_1} dt_1$$

$$\Pr(t_2 < E_2 < t_2 + dt_2 / E_1) = c e^{-c(t_2 - t_1)} dt_2$$

...

$$\Pr(\text{aucun événement n'arrive après } t_n / E_n) = c e^{-c(T - t_n)} dt_n$$

D'où : $f(t_1, \dots, t_n, n) = c^n e^{-cT}$.

On en déduit la densité de la loi de probabilité conditionnelle :

$$f(t_1, \dots, t_n / N = n) = c^n e^{-cT} \times \frac{n!}{e^{-cT} (cT)^n} = \frac{n!}{T^n}$$

Les temps $t_1, \dots, t_n$ constituent un échantillon ordonné de la loi uniforme sur $[0, T]$. Si on s'intéresse aux dates et non à l'ordre, on doit diviser la densité par $n!$.

□ Étude du processus N_t

Le processus N_t , nombre d'événements se produisant pendant un temps t , suit une loi de Poisson de paramètre ct . On en déduit : $E(N_t) = \text{Var}(N_t) = ct$.

C'est donc un processus non stationnaire, mais à accroissements stationnaires et indépendants car la variable aléatoire $(N_{t+h} - N_t)$ suit une loi de Poisson de paramètre h , quelle que soit la valeur de h .

La fonction de covariance de ce processus est :

Si $s > t$,

$$\begin{aligned} C(t, s) &= \text{Cov}(N_t, N_s) = \text{Cov}(N_t, N_t + X) \\ &= \text{Var}(N_t) + \text{Cov}(N_t, X) = \text{Var}(N_t) = ct \end{aligned}$$

X est une variable indépendante de t car le processus N_t est à accroissements stationnaires.

Si $s < t$, un calcul analogue conduit à $C(t, s) = cs$.

D'où : $C(t, s) = c \inf(t, s)$.

Cette fonction étant continue en $t = s$, le processus est continu en moyenne quadratique mais aucune trajectoire n'est continue puisque N_t est une fonction en escalier.

9.12 Application aux phénomènes d'attente

9.12.1 Définition et généralités

Les phénomènes d'attente ont une grande importance dans l'étude du fonctionnement de centres de services à postes de services ou guichets multiples, dès que l'arrivée des clients ou usagers et les temps de service sont aléatoires. On peut donner quelques exemples :

- les arrivées des clients dans un bureau de poste à plusieurs guichets,
- les arrivées des clients dans une grande surface à plusieurs caisses,

- les arrivées des avions dans un aéroport à plusieurs pistes,
- les arrivées des malades dans un service d'urgence d'un hôpital, etc.

Quand le nombre de guichets est insuffisant pour le service à assurer, il se produit rapidement une file d'attente préjudiciable aux clients.

En revanche, si le nombre de guichets est trop élevé pour le service à assurer, il ne se produit jamais de file d'attente mais le coefficient moyen d'utilisation des guichets est faible, ce qui peut entraîner des frais d'exploitation prohibitifs.

Par conséquent, il doit y avoir un optimum économique entre ces deux intérêts opposés. L'étude des files d'attente a pour but de fournir aux responsables, les éléments d'un choix motivé.

La théorie des files d'attente a fait l'objet de nombreux travaux, prenant en compte :

- les lois de distribution des arrivées et des temps de service,
- éventuellement des règles de priorité,
- la structure des systèmes, certains systèmes pouvant comporter plusieurs étages en cascade, etc.

9.12.2 Schéma étudié

On étudie le cas classique d'un centre possédant k stations (centre à stations multiples). Les clients sont parfaitement disciplinés et respectent l'ordre d'arrivée, c'est-à-dire que, quand il n'y a pas d'urgence, le premier arrivé est le premier servi. On suppose que :

- les arrivées sont distribuées suivant un processus de Poisson, de cadence λ ; ce nombre λ caractérise le taux moyen des arrivées, c'est-à-dire le nombre moyen d'utilisateurs se présentant au centre par unité de temps ;
- les durées de service sont distribuées suivant une loi exponentielle de paramètre μ ; ce paramètre μ représente le nombre moyen d'utilisateurs servis par unité de temps par un même guichet. La durée moyenne de service est donc égale à $1/\mu$. Si T représente le temps de service, on obtient :

$$\Pr(T > t) = e^{-\mu t}$$

Soit n le nombre d'utilisateurs dans le centre, soit en attente, soit entrain de se faire servir. Ce nombre est tel que :

- si $n < k$, il n'y a pas de file d'attente,
- si $n > k$, il se forme une file d'attente de longueur $(n - k)$.

Il est évident que, si l'on pose $\psi = \lambda/\mu$, on doit avoir $\psi < k$ pour éviter le risque de voir se former une file d'attente de longueur infinie, ψ est le *facteur de charge du centre*.

Il est impossible de prévoir l'évolution rigoureuse d'un système, même si la configuration initiale est connue, car les arrivées ainsi que les départs se font d'une manière aléatoire. Cependant, on peut calculer les probabilités $\text{Pr}_n(t)$ qu'il y ait n usagers dans le centre au temps t en fonction de n et de t . On en déduit ensuite les configurations les plus vraisemblables du système au cours du temps : ce sont les *équations d'état du système*.

■ Équations d'état du système

Les probabilités $\text{Pr}_n(t)$ qu'il y ait n usagers dans le système à l'instant t ne sont pas indépendantes, mais liées par des relations ou *équations d'état*.

□ Première équation d'état

Pour calculer la probabilité $\text{Pr}_0(t + dt)$, il n'y a que deux éventualités exclusives à examiner :

1. Il n'y a pas d'utilisateur dans le système à l'instant t et il n'y a pas d'arrivée pendant l'intervalle de temps $(t + dt)$.

Le nombre d'arrivées pendant cet intervalle de temps suit une loi de Poisson de paramètre λdt . Comme cet intervalle de temps dt est petit, la probabilité qu'il n'y ait pas d'arrivée, qui est égale à $e^{-\lambda dt}$, est peu différente de $1 - \lambda dt$ et la probabilité d'une arrivée est égale à λdt .

2. Il y a un utilisateur dans le centre à l'instant t , la probabilité de cet événement est $\text{Pr}_1(t)$:

- d'une part, cet utilisateur quitte le centre pendant l'intervalle de temps $(t + dt)$, la probabilité de cet événement est μdt ;
- d'autre part, il n'y a pas d'arrivée pendant l'intervalle de temps $(t + dt)$, la probabilité de cet événement est $1 - \lambda dt$.

D'où l'expression de la probabilité $\text{Pr}_0(t + dt)$:

$$\text{Pr}_0(t + dt) = (1 - \lambda dt) \text{Pr}_0(t) + (1 - \lambda dt) \mu dt \text{Pr}_1(t)$$

L'intervalle de temps dt étant supposé petit, le terme $\lambda\mu(dt)^2$ est négligeable. Après simplification, on obtient la première équation d'état :

$$\frac{d \Pr_0(t)}{dt} = -\lambda \Pr_0(t) + \mu \Pr_1(t) \quad (1)$$

□ Équations d'état pour $n < k$

On obtient ces équations comme précédemment en envisageant toutes les éventualités possibles. Comme $n < k$, tous les usagers sont servis simultanément, tout nouvel usager sera servi et aucune file d'attente ne se formera.

La probabilité pour qu'un usager soit servi et donc quitte le centre est égale, s'il y a n usagers dans le centre, à $n\mu dt$, et la probabilité pour qu'il ne sorte aucun usager est égale à $1 - n\mu dt$.

$$\Pr_n(t + dt) = (\lambda dt)[1 - (n-1)\mu dt] \Pr_{n-1}(t) + (1 - \lambda dt)(1 - n\mu dt) \Pr_n(t) + (\lambda dt)(n\mu dt) \Pr_n(t) + (1 - \lambda dt)[(n+1)\mu dt] \Pr_{n+1}(t)$$

Après simplification, on obtient :

$$\frac{d \Pr_n(t)}{dt} = \lambda \Pr_{n-1}(t) - (\lambda + n\mu) \Pr_n(t) + (n+1)\mu \Pr_{n+1}(t) \quad (2)$$

□ Équations d'état pour $n \geq k$

Quand le nombre d'usagers est égal ou supérieur au nombre de guichets, la probabilité pour qu'un usager quitte le centre est égale $k\mu dt$ quelle que soit la valeur de n . L'équation d'état est de la forme précédente, après avoir remplacé n et $(n+1)$ par k :

$$\frac{d \Pr_n(t)}{dt} = \lambda \Pr_{n-1}(t) - (\lambda + k\mu) \Pr_n(t) + k\mu \Pr_{n+1}(t) \quad (3)$$

En résumé, l'évolution du système est définie par les équations (1), (2) et (3) auxquelles il faut ajouter la condition :

$$\sum_{n=0}^{\infty} \Pr_n(t) = 1$$

et les conditions initiales :

$$t = 0 \quad \begin{cases} n = 0 & \text{Pr}_0(0) = 1 \\ n \geq 1 & \text{Pr}_n(0) = 0 \end{cases}$$

La résolution d'un tel système fait appel à la transformation de Laplace et la solution fait intervenir les fonctions de Bessel incomplètes.

Il y a une période transitoire pendant laquelle les fonctions $\text{Pr}_n(t)$ varient très vite en fonction du temps.

Si le facteur de charge $\psi = \lambda/\mu$ est inférieur à k , les probabilités $\text{Pr}_n(t)$ tendent vers des probabilités Pr_n indépendantes du temps qui caractérisent le *régime stationnaire*, ce régime est atteint au bout d'un temps de l'ordre de $3/\mu k$.

□ Étude du régime stationnaire

On est dans l'hypothèse où le facteur de charge est inférieur au nombre de guichets. Le système à résoudre est obtenu à partir des équations (1), (2) et (3), en supposant les probabilités indépendantes du temps :

$$\begin{aligned} \sum_{n=0}^{\infty} \text{Pr}_n &= 1 \\ \mu \text{Pr}_1 &= \lambda \text{Pr}_0 \\ (n+1)\mu \text{Pr}_{n+1} &= (\lambda + n\mu) \text{Pr}_n - \lambda \text{Pr}_{n-1} \quad 1 \leq n \leq k \\ k\mu \text{Pr}_{n+1} &= (\lambda + k\mu) \text{Pr}_n - \lambda \text{Pr}_{n-1} \quad n \geq k \end{aligned}$$

Ce système a pour solution :

$$\begin{aligned} \text{Pr}_n &= \frac{\psi^n}{n!} \text{Pr}_0 \quad 1 \leq n < k \\ \text{Pr}_n &= \frac{\psi^n}{k! k^{n-1}} \text{Pr}_0 \quad n \geq k \end{aligned}$$

On calcule la probabilité Pr_0 en utilisant la première relation :

$$\text{Pr}_0 = \frac{1}{1 + \sum_{i=1}^{k-1} \frac{\psi^i}{i!} + \frac{\psi^k}{k! (1 - \psi/k)}}$$

□ Étude de deux cas extrêmes

1. Centre possédant un seul guichet ($k = 1$) :

$$\Pr_0 = 1 - \psi \quad \Pr_n = (1 - \psi)\psi^n$$

La distribution correspondante est une loi géométrique (voir chapitre 5, paragraphe 5.4.4) qui a pour fonction de répartition : $F(n) = 1 - \psi^{n+1}$

2. Centre possédant un nombre surabondant de guichets : le nombre n d'utilisateurs dans le centre dépasse rarement le nombre de guichets, le risque de voir se former une file d'attente est donc très faible. Ce nombre n est pratiquement distribué suivant une loi de Poisson :

$$\Pr_n = e^{-\psi} \frac{\psi^n}{n!}$$

9.12.3 Caractéristiques moyennes du phénomène

Soient :

- N le nombre d'utilisateurs dans le système,
- F la longueur de la file d'attente avec $F = N - k$ si $N > k$ et $F = 0$ si $N \leq k$,
- I le nombre de postes inoccupés,
- A le temps d'attente avant d'être servi.

Des calculs simples, mais un peu longs, conduisent aux résultats suivants :

$$\begin{aligned} E(N) &= \sum_{n=0}^{\infty} n \Pr_n = \psi + \frac{k}{(k - \psi)^2} \times \frac{\psi^{k+1}}{k!} \Pr_0 \\ E(F) &= \sum_{n=k+1}^{\infty} (n - k) \Pr_n = \frac{k}{(k - \psi)^2} \times \frac{\psi^{k+1}}{k!} \Pr_0 \\ E(I) &= \sum_{n=0}^{\infty} (k - n) \Pr_n = k - \psi \end{aligned}$$

En régime stationnaire, $E(F) = \lambda E(A)$. D'où :

$$E(A) = \frac{k}{(k - \psi)^2} \times \frac{\psi^k}{k!} \times \frac{1}{\mu} \Pr_0$$

Ces expressions se simplifient dans le cas où le nombre k de guichets est égal à 1. On obtient alors :

$$\begin{aligned} \Pr_0 &= 1 - \psi & E(N) &= \frac{\psi}{1 - \psi} & E(F) &= \frac{\psi^2}{1 - \psi} \\ E(I) &= 1 - \psi & E(A) &= \frac{\psi}{\mu(1 - \psi)} \end{aligned}$$



Statistique inférentielle

10 • CARACTÉRISTIQUES D'UN ÉCHANTILLON APPLICATION AUX ÉCHANTILLONS GAUSSIENS

10.1 Introduction

Le calcul des probabilités apporte les outils nécessaires aux techniques de la statistique mathématique, c'est-à-dire les modèles qui vont être utilisés pour décrire des phénomènes réels où le hasard intervient. La *statistique* est un ensemble de méthodes permettant de prendre de bonnes décisions en présence de l'incertain.

Tableau 10.1 – Quelques exemples.

Statistique descriptive	Statistique mathématique
Étude du débit d'une rivière pendant 50 ans.	Prévisions sur la hauteur maximale des crues en vue de la construction d'un barrage.
Étude des caractéristiques d'un ensemble de pièces fabriquées en série.	Contrôle d'une fabrication en série.
Étude du nombre de vacanciers pendant une période déterminée dans une station de sports d'hiver.	Prévoir le nombre de lits nécessaires pour l'hébergement.
Étude de données économiques sur les dépenses des ménages.	Prévoir l'évolution de la vente d'un produit.

En résumé :

- la mise en ordre des données relève des techniques de la statistique descriptive (caractéristiques numériques ou graphiques),

- la prévision de l'évolution d'un phénomène réel, à partir des données numériques et des lois de probabilité théoriques, relève de la statistique mathématique.

Une étude statistique portant sur tous les éléments d'une population étant, soit impossible à réaliser (trop grand nombre d'individus à étudier), soit trop onéreuse, il faut obtenir des résultats fiables sur les caractéristiques d'une population en se limitant à l'étude des éléments ou unités d'un échantillon. Cet échantillon doit non seulement donner des estimations non *biaisées* des paramètres mais permettre, de plus, d'évaluer la marge d'erreurs dues aux fluctuations d'échantillonnage.

L'échantillon doit être *représentatif* de la population ; il en résulte, en particulier, que chaque unité doit avoir une probabilité non nulle d'être tirée, un tel échantillon est qualifié d'*aléatoire*.

En conclusion, toute démarche statistique consiste :

- à prélever un *échantillon représentatif* de la population (échantillon aléatoire) par des techniques appropriées. Les différentes méthodes utilisées pour obtenir un tel échantillon relèvent de la théorie de l'échantillonnage ; quelques méthodes seront expliquées dans le chapitre 11,
- à étudier les principales caractéristiques d'un échantillon, issu d'une population dont on connaît la loi de probabilité ; le cas particulier des échantillons issus d'une population normale est traité en détail paragraphe 10.7. Les lois du *chi-deux*, de *Fisher-Snedecor*, de *Student*, lois dérivées de la loi normale, ayant de nombreuses applications dans la théorie de l'estimation et des tests, sont étudiées dans les paragraphes 10.4, 10.5 et 10.6,
- à savoir réaliser des échantillons de variables aléatoires pour vérifier des conclusions en utilisant des techniques de simulation (chapitre 11, paragraphe 11.5).

10.2 Définition d'un échantillon aléatoire

On étudie une caractéristique mesurable X d'une population de taille finie ou infinie. La composition de la population, vis à vis du caractère X , est entièrement définie par la connaissance des quantités $F(x)$:

$F(x)$ = Proportion des individus tels que $X < x$, pour toutes les valeurs de $x \in \mathbb{R}$

Soit E l'expérience consistant à choisir *au hasard* un élément de la population. Avant le tirage, on se propose de prévoir la valeur du caractère X que l'on obtiendra. Ce caractère est une variable aléatoire X telle que $\Pr(X < x) = F(x)$ pour toute valeur $x \in \mathbb{R}$. À l'expérience E , est associée une variable aléatoire X dont la fonction de répartition est $F(x)$.

On réalise n fois la même expérience E , dans des conditions indépendantes, par exemple en remettant dans la population l'élément tiré. À ces n expériences, on associe n variables aléatoires indépendantes X_i , suivant la même loi que la variable aléatoire X .

Par définition, l'ensemble $(X_1, X_2, \dots, X_n)$, de n variables aléatoires indépendantes suivant la même loi qu'une variable aléatoire X , appelée *variable aléatoire parente*, est un échantillon aléatoire, noté en abrégé $(X_i) \ i \in (1, n)$ ou $\underline{X}$.

Une réalisation de l'échantillon sera notée $(x_1, x_2, \dots, x_n)$, ce n'est plus une variable aléatoire.

10.3 Caractéristiques d'un échantillon aléatoire

Une *statistique* définie sur un échantillon aléatoire $\underline{X}$ de taille n est une fonction mesurable des variables X_i . Les principales caractéristiques d'un échantillon sont les statistiques $\bar{X}$ et S^2 .

On suppose que les moments d'ordre 1 et 2 de la variable aléatoire parente X existent et on pose $E(X) = m$ et $\text{Var}(X) = \sigma^2$.

10.3.1 Étude de la statistique $\bar{X}$

La *statistique* $\bar{X}$ est la fonction mesurable des variables X_i , définie par :

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

■ Cas d'une population infinie

Un calcul rapide donne les résultats suivants :

$$E(\bar{X}) = m \quad \text{Var}(\bar{X}) = \frac{\sigma^2}{n}$$

■ Cas d'une population finie

L'échantillon de taille n est extrait d'une population finie de taille N par un tirage sans remise ; on obtient les résultats suivants :

$$E(\bar{X}) = m \quad \text{Var}(\bar{X}) = \frac{N-n}{N-1} \frac{\sigma^2}{n} \cong \left(1 - \frac{n}{N}\right) \frac{\sigma^2}{n}$$

Le coefficient $\frac{n}{N}$ représente la fraction de la population constituant l'échantillon.

■ Comportement asymptotique

□ Loi faible des grands nombres

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

□ Loi forte des grands nombres

$\bar{X}$ converge presque sûrement vers m quand n tend vers l'infini,

car la série :

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

est une série convergente (chapitre 7, paragraphe 7.4).

Appliquons le théorème central limite à la variable aléatoire Y_n :

$$Y_n = \frac{\sum_{i=1}^n X_i - nm}{\sigma\sqrt{n}} = \frac{\bar{X} - m}{\sigma/\sqrt{n}}$$

La variable Y_n converge en loi vers une variable suivant la loi normale $N(0 ; 1)$ quand n tend vers l'infini.

Remarque

Pour appliquer le théorème central limite, il faut que les moments d'ordre 1 et 2 existent. La loi de Cauchy, loi qui ne possède aucun moment, est un contre-exemple. La densité de probabilité de cette variable (chapitre 4, paragraphe 4.7.1, exemple 4.4) et sa fonction caractéristique (chapitre 7, paragraphe 7.2.4) ont pour

expressions :

$$f(x) = \frac{1}{\pi} \frac{1}{1+x^2} \quad \varphi_X(t) = e^{-|t|}$$

La fonction caractéristique de la statistique $\bar{X}$, somme de n variables aléatoires indépendantes de Cauchy, est égale à :

$$\varphi_{\bar{X}}(t) = \left[\varphi_X\left(\frac{t}{n}\right) \right]^n = e^{-|t|}$$

La statistique $\bar{X}$ suit la même loi que la variable X , c'est-à-dire une loi de Cauchy, la loi de variable $\bar{X}$ ne converge donc pas vers la loi $N(0; 1)$ quand n tend vers l'infini.

■ Application : loi d'un pourcentage

Soit K la variable aléatoire représentant le nombre de succès au cours d'une suite de n épreuves indépendantes, la probabilité de succès au cours d'une épreuve étant égale à p . La loi de la variable aléatoire K est la loi binomiale $B(n; p)$.

Posons $F = \frac{K}{n}$, fréquence empirique du nombre de succès :

$$E(F) = p \quad \text{Var}(F) = \frac{pq}{n} \quad q = 1 - p$$

En appliquant le théorème central limite pour n grand, on démontre que la variable aléatoire F converge en loi vers une variable aléatoire suivant une loi normale :

$$N\left(p; \sqrt{\frac{pq}{n}}\right)$$

Ce résultat, connu sous le nom de théorème de De Moivre-Laplace, est à la base de l'approche fréquentiste de la théorie des probabilités.

10.3.2 Étude de la statistique S^2 ou variance empirique

La variance empirique S^2 d'un échantillon de taille n est définie par :

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

expression qui peut s'écrire :

$$S^2 = \frac{1}{n} \left[\sum_{i=1}^n (X_i - m)^2 \right] - \left(m - \bar{X} \right)^2$$

■ Propriétés caractéristiques de la statistique S^2

□ Espérance mathématique de S^2

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

Si la taille n de l'échantillon est grande, l'espérance de S^2 a pour valeur limite σ^2 .

□ Variance de S^2

(Calcul un peu plus long mais non difficile.)

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

μ_4 étant le moment centré d'ordre 4 de la variable X .

Si la taille n de l'échantillon est grande, la variance de S^2 a pour valeur limite :

$$\frac{\mu_4 - \sigma^4}{n}$$

■ Théorème central limite pour S^2

La variable aléatoire :

$$\frac{S^2 - \left(1 - \frac{1}{n} \right) \sigma^2}{\sqrt{\text{Var } S^2}}$$

converge en loi vers une variable suivant la loi normale $N(0 ; 1)$ quand n tend vers l'infini.

En prenant les limites de l'espérance et de la variance pour n grand, on obtient le résultat suivant. La variable aléatoire :

$$\frac{S^2 - \sigma^2}{\sqrt{\frac{\mu_4 - \sigma^4}{n}}}$$

converge en loi vers la loi $N(0 ; 1)$.

■ Corrélation entre $\bar{X}$ et S^2

Pour définir la corrélation entre $\bar{X}$ et S^2 , on calcule la covariance entre ces deux variables aléatoires :

$$\text{Cov}(\bar{X}, S^2) = \frac{n-1}{n^2} \mu_3$$

μ_3 étant le moment centré d'ordre 3 de la variable X .

- Si n tend vers l'infini, la covariance entre ces variables tend vers 0, les statistiques $\bar{X}$ et S^2 sont donc asymptotiquement non corrélées.
- Si la distribution de la variable X est symétrique, le moment centré μ_3 est égal à 0, les statistiques $\bar{X}$ et S^2 sont donc non corrélées quelle que soit la valeur de n .
- Si, de plus, X suit une loi normale, les statistiques $\bar{X}$ et S^2 sont indépendantes quelle que soit la valeur de n (paragraphe 10.7).

10.4 Distribution du chi-deux

La variable aléatoire, égale à la somme des carrés de ν variables aléatoires indépendantes, centrées, réduites, gaussiennes, suit la loi du chi-deux, χ^2 , à ν degrés de liberté :

$$\chi^2(\nu) = \sum_{i=1}^{\nu} U_i^2 = \sum_{i=1}^{\nu} \left(\frac{X_i - m}{\sigma} \right)^2$$

Cette distribution a été introduite par Karl Pearson en 1900.

10.4.1 Propriétés

- La variable aléatoire $\chi^2(\nu)$ varie de 0 à $+\infty$.
- Le paramètre ν est le nombre de degrés de liberté de la variable, il représente la dimension de l'espace dans lequel se trouve le point représentatif de l'échantillon $\underline{X}$. Si les variables aléatoires X_i vérifient k relations linéaires, le nombre de degrés de liberté diminue de k .
- La loi suivie par la somme de variables aléatoires indépendantes, suivant chacune des lois du chi-deux, est une loi du chi-deux dont le degré de liberté est la somme des degrés de liberté de chaque loi.

- Moments :

$$E \left[\chi^2(v) \right] = v \quad \text{Var} \left[\chi^2(v) \right] = 2v$$

- Mode :

$$M_0 = v - 2 \quad \text{si } v > 2$$

- Selon les valeurs de v , la distribution du chi-deux a des formes différentes :

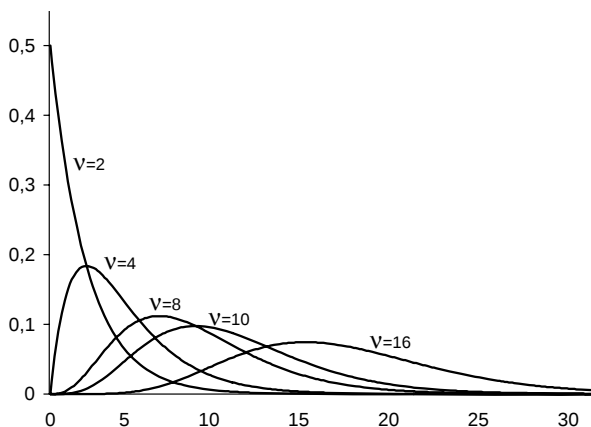


Figure 10.1 – Densité de la loi du chi-deux pour différentes valeurs du paramètre v .

La distribution de la variable aléatoire chi-deux tend à se rapprocher d'une distribution normale quand v augmente.

- La densité de probabilité de la variable aléatoire $\chi^2(v)$ a une expression mathématique compliquée et peu maniable :

$$f \left(\chi^2(v) \right) = \frac{1}{2^{v/2} \Gamma(v/2)} e^{-\chi^2/2} \left(\chi^2 \right)^{v/2-1}$$

Γ est la fonction eulérienne (annexe 2).

Ce résultat peut être obtenu en utilisant les propriétés des fonctions caractéristiques. Cette densité est dissymétrique.

La table 6 donne les fractiles d'ordre α de la loi du chi-deux pour différentes valeurs du paramètre v :

$$\alpha = \Pr \left(\chi^2(v) \leq \chi_{\alpha}^2(v) \right) = \int_0^{\chi_{\alpha}^2} f \left(\chi^2(v) \right) d \left(\chi^2(v) \right)$$

- La fonction caractéristique de la variable aléatoire $\chi^2(\nu)$ a pour expression :

$$\varphi_{\chi^2(\nu)}(t) = \frac{1}{(1 - 2it)^{\nu/2}}$$

- Pour les grandes valeurs de ν , il existe plusieurs formes limites :
- une première forme est obtenue en appliquant le théorème central limite. La loi de la variable aléatoire :

$$\frac{\chi^2(\nu) - \nu}{\sqrt{2\nu}}$$

converge, quand ν tend vers l'infini, vers la loi normale centrée réduite ;

- la deuxième forme est due à Fisher. Pour $\nu \geq 30$, la loi de la variable aléatoire :

$$\sqrt{2\chi^2(\nu)} - \sqrt{2\nu - 1}$$

est la loi normale centrée réduite.

- Il existe d'autres formules d'approximation, comme celle de Wilson-Hilferty, pour $\nu \geq 30$:

$$\chi^2_\alpha(\nu) \cong \nu \left(1 - \frac{2}{9\nu} + U_\alpha \sqrt{\frac{2}{9\nu}} \right)^3$$

où U_α désigne le fractile d'ordre α de la loi normale centrée réduite.

10.4.2 Relation entre la loi du chi-deux et les lois gamma

Si la variable aléatoire U est une variable gaussienne centrée réduite, la densité de la variable aléatoire $\chi^2(1)$ définie par $T = U^2$ est :

$$g(t) = \frac{1}{\sqrt{2\pi}} e^{-\frac{t}{2}} t^{-\frac{1}{2}}$$

On en déduit que la variable aléatoire $U^2/2 = \chi^2(1)/2$ suit une loi $\gamma(1/2; 1)$. La variable $\chi^2(n)/2$ suit donc une loi $\gamma(n/2; 1)$. Cette propriété permet de retrouver la densité de la variable $\chi^2(n)$, ainsi que les différents moments. Ce résultat s'énonce sous la forme suivante :

Si X est une variable aléatoire suivant la loi $\gamma(n/2; 1)$, la variable aléatoire $2X$ suit la loi $\chi^2(n)$.

10.5 Distribution de Fisher-Snedecor

La distribution F de Fisher-Snedecor (ou plus simplement distribution F de Fisher) a été étudiée en 1924 par Fisher (statisticien anglais né en 1890) et calculée en 1934 par Snedecor ; elle joue un rôle important en analyse de la variance (chapitre 16, paragraphe 16.3) et en analyse de la régression (chapitres 20 et 21).

On considère deux variables aléatoires indépendantes suivant des lois du chi-deux à ν_1 et ν_2 degrés de liberté respectivement.

La variable aléatoire F de Fisher est définie par :

$$F(\nu_1 ; \nu_2) = \frac{\chi^2(\nu_1) / \nu_1}{\chi^2(\nu_2) / \nu_2}$$

10.5.1 Propriétés

– Propriété évidente :

$$F(\nu_1 ; \nu_2) = \frac{1}{F(\nu_2 ; \nu_1)}$$

– La variable $F(\nu_1 ; \nu_2)$ varie de 0 à $+\infty$.

– La loi de probabilité de la variable F dépend de deux paramètres, les degrés de liberté, ν_1 et ν_2 .

– Moments :

$$E(F) = \frac{\nu_2}{\nu_2 - 2} \quad \text{si } \nu_2 > 2$$

$$\text{Var}(F) = \left(\frac{\nu_2}{\nu_2 - 2} \right)^2 \frac{2(\nu_1 + \nu_2 - 2)}{\nu_1(\nu_2 - 4)} \quad \text{si } \nu_2 > 4$$

– Mode :

$$M_0 = \frac{\nu_2(\nu_1 - 2)}{\nu_1(\nu_2 + 2)} \quad \text{si } \nu_1 > 2$$

– Densité : elle a une forme mathématique compliquée :

$$f \leq 0 \quad g(f) = 0$$

$$f > 0 \quad g(f) = \frac{\left(\frac{\nu_1}{\nu_2}\right)^{\nu_1/2}}{B\left(\frac{\nu_1}{2}; \frac{\nu_2}{2}\right)} \times \frac{f^{\nu_1/2-1}}{\left(1 + \frac{\nu_1}{\nu_2}f\right)^{\frac{\nu_1+\nu_2}{2}}}$$

- La figure 10.2 montre la forme de la densité de F , pour différentes valeurs des paramètres.

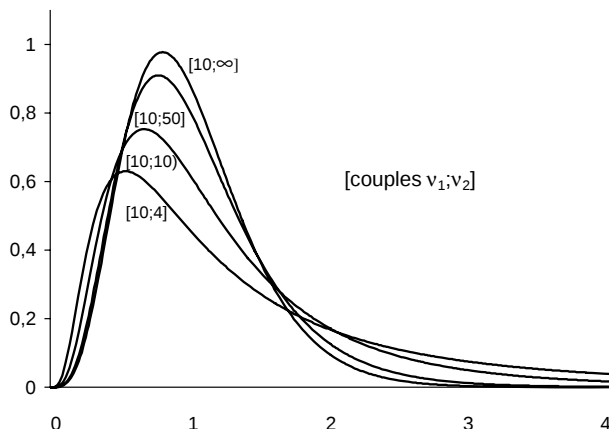


Figure 10.2 – Densité de la loi de Fisher-Snedecor.

- Les tables 7.1, 7.2, 7.3 et 7.4 donnent les fractiles d'ordre $\alpha = 0,95$; $\alpha = 0,975$; $\alpha = 0,99$ et $\alpha = 0,995$ respectivement de la loi de Fisher :

$$\alpha = \Pr(F(v_1; v_2) \leq F_\alpha(v_1; v_2)) = \int_0^{F_\alpha} g(f) df$$

- Ces fractiles correspondent à $\alpha > 0,50$, ils sont supérieurs à l'unité. Pour les fractiles inférieurs à l'unité, correspondant aux valeurs $\alpha < 0,50$, on utilise la relation suivante (facile à démontrer) :

$$F_\alpha(v_1; v_2) = \frac{1}{F_{1-\alpha}(v_2; v_1)}$$

10.5.2 Relation entre la loi de Fisher et les lois bêta

- Si une variable aléatoire X suit une loi bêta de type I, la variable $\frac{v_1}{v_2} \frac{X}{1-X}$ est une variable de Fisher $F(2v_1; 2v_2)$.

- De même, si une variable aléatoire Y suit une loi bêta de type II, la variable $\frac{v_1}{v_2} Y$ est une variable de Fisher $F(2v_1; 2v_2)$.

10.6 Distribution de Student

La distribution T de Student, pseudonyme du statisticien anglais W.S. Gosset (1876-1937), joue un rôle important dans l'étude de la statistique $\bar{X}$ pour une distribution normale dont on ne connaît pas la variance.

La loi de Student est la loi de la variable aléatoire T définie par :

$$T^2(v) = \frac{U^2}{\chi^2(v)/v} = F(1; v) \quad \text{ou} \quad T(v) = \frac{U}{\sqrt{\chi^2(v)/v}}$$

où U est une variable aléatoire centrée réduite normale.

10.6.1 Propriétés

- La variable $T(v)$ varie de $-\infty$ à $+\infty$.
- La loi de probabilité de la variable $T(v)$ dépend d'un paramètre, le degré de liberté v de la variable $\chi^2(v)$.
- Moments :

$$E[T(v)] = 0 \quad \text{Var}[T(v)] = \frac{v}{v-2} \quad \text{si } v > 2$$

- Mode :

$$M_0 = 0$$

- La densité de probabilité de la variable T a une expression mathématique compliquée et peu utilisée :

$$\begin{aligned} f(t) &= \frac{1}{\sqrt{\pi v}} \frac{\Gamma\left(\frac{v+1}{2}\right)}{\Gamma(v/2)} \left(1 + \frac{t^2}{v}\right)^{-(v+1)/2} \\ &= \frac{1}{\sqrt{v}} \frac{1}{B(1/2, v/2)} \left(1 + \frac{t^2}{v}\right)^{-(v+1)/2} \end{aligned}$$

B est la fonction eulérienne (annexe 2).

- La table 8 donne les fractiles d'ordre α de la loi de Student :

$$\alpha = \Pr(T(\nu) \leq T_\alpha(\nu)) = \int_{-\infty}^{t_\alpha(\nu)} f(t(\nu)) dt$$

Comme la fonction est symétrique, il suffit de prendre $\alpha > 50$.

- Selon les valeurs de ν , la distribution T de Student a des formes différentes (figure 10.3). La courbe admet un axe de symétrie (ressemblance avec la « courbe en cloche » de la distribution normale).

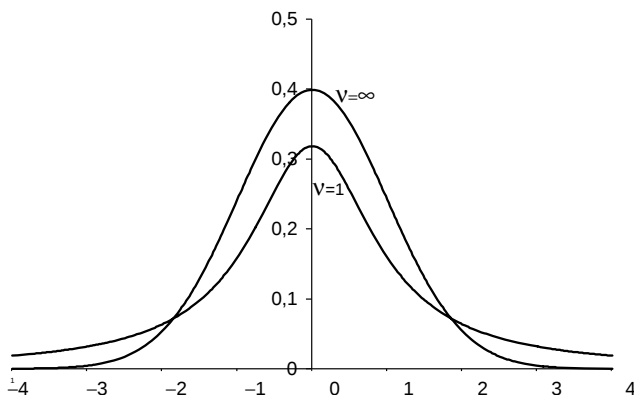


Figure 10.3 – Densité de la loi de Student.

- Pour les grandes valeurs de ν ($\nu > 100$), la loi de Student peut être remplacée par la loi normale réduite.

10.6.2 Relation entre la loi de Student et la loi de Cauchy

Pour $\nu = 1$, la densité de la variable $T(1)$:

$$f(t) = \frac{1}{\pi} \frac{1}{(1 + t^2)}$$

est celle d'une variable aléatoire suivant une loi de Cauchy (chapitre 4, paragraphe 4.7, exemple 4.4).

10.7 Cas particulier des échantillons gaussiens

Les échantillons considérés dans ce paragraphe sont issus d'une population suivant la loi normale $N(m; \sigma)$ et les propriétés démontrées ne sont valables que sous cette hypothèse.

10.7.1 Étude de la statistique $\bar{X}$

La variable $\bar{X}$, combinaison linéaire de n variables aléatoires indépendantes gaussiennes, est une variable gaussienne. Donc, quelle que soit la valeur de n :

la loi de la variable $\bar{X}$ est la loi $N\left(m; \frac{\sigma}{\sqrt{n}}\right)$

10.7.2 Étude de la statistique S^2

La décomposition de la statistique S^2 (paragraphe 10.3.2) :

$$n S^2 = \left[\sum_{i=1}^n (X_i - m)^2 \right] - n (m - \bar{X})^2$$

et le théorème de Cochran sur la décomposition d'une forme quadratique conduisent au résultat suivant :

$$\sum_{i=1}^n \left(\frac{X_i - m}{\sigma} \right)^2 = \frac{n S^2}{\sigma^2} + \left(\frac{\bar{X} - m}{\sigma/\sqrt{n}} \right)^2$$

- Le premier membre, somme de n carrés de variables aléatoires centrées réduites, indépendantes, gaussiennes est une variable $\chi^2(n)$.
- Le deuxième membre est une somme de deux formes quadratiques :
 - la première est de rang $(n - 1)$, car les variables vérifient la relation :

$$\sum_{i=1}^n (X_i - \bar{X}) = 0$$

- la deuxième est de rang 1.

On en déduit les deux résultats suivants :

$$\frac{n S^2}{\sigma^2} \quad \text{est une variable } \chi^2 (n - 1)$$

$\bar{X}$ et S^2 sont deux variables indépendantes

On démontre la réciproque suivante, qui est une propriété caractéristique des variables aléatoires gaussiennes :

Si les statistiques $\bar{X}$ et S^2 sont *indépendantes*, la variable aléatoire X est une *variable aléatoire gaussienne*.

10.7.3 Application : loi de la statistique $\bar{X}$

Des résultats démontrés précédemment :

$$\frac{\bar{X} - m}{\sigma/\sqrt{n}} \quad \text{est une variable suivant la loi } N(0 ; 1)$$

$$\frac{n S^2}{\sigma^2} \quad \text{est une variable } \chi^2 (n - 1)$$

on déduit que :

$$\frac{\bar{X} - m}{S/\sqrt{n-1}} \quad \text{est une variable suivant la loi de Student } T(n - 1)$$

Comme la variable aléatoire de Student ainsi définie ne dépend pas de σ , cette propriété sera utilisée dans la théorie de l'estimation (chapitre 14, paragraphe 14.2.1) quand l'écart-type σ est inconnu.

10.7.4 Comparaison des variances de deux populations indépendantes suivant des lois normales

Soient n_1 réalisations indépendantes d'une variable aléatoire X_1 suivant la loi normale $N(m_1 ; \sigma_1)$ et n_2 réalisations indépendantes d'une variable aléatoire X_2 suivant la loi normale $N(m_2 ; \sigma_2)$. Les variables X_1 et X_2 sont indépendantes.

De la propriété :

$$\frac{n_1 S_1^2}{\sigma_1^2} = \chi^2 (n_1 - 1) \quad \frac{n_2 S_2^2}{\sigma_2^2} = \chi^2 (n_2 - 1)$$

on déduit le résultat suivant qui sera utilisé dans la théorie de l'estimation (chapitre 14, paragraphe 14.2.4) :

$$\frac{n_1 S_1^2}{\sigma_1^2 (n_1 - 1)} \times \frac{\sigma_2^2 (n_2 - 1)}{n_2 S_2^2} = F(n_1 - 1; n_2 - 1)$$

10.7.5 Étude de la différence des moyennes de deux échantillons indépendants suivant des lois normales de variances inconnues mais égales

On utilise toujours les mêmes notations. La variable aléatoire :

$$\frac{(\bar{X}_1 - \bar{X}_2) - (m_1 - m_2)}{\sqrt{\sum_1 (X_i - \bar{X}_1)^2 + \sum_2 (X_i - \bar{X}_2)^2}} \times \frac{\sqrt{n_1 + n_2 - 2}}{\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

suit une *loi de Student* à $(n_1 + n_2 - 2)$ degrés de liberté. Les sommations sont faites sur les échantillons 1 et 2. Ce résultat, démontré et utilisé dans le chapitre 14, paragraphe 14.2.3, permet de construire un intervalle de confiance pour la différence des moyennes de deux échantillons indépendants gaussiens.

10.7.6 Autre application

Soient deux variables aléatoires indépendantes, X et Y , suivant la même loi de probabilité. On suppose de plus que *les variables aléatoires* $(X + Y)$ et $(X - Y)$ sont *indépendantes*. On démontre le résultat suivant :

Les variables aléatoires X et Y sont des *variables aléatoires gaussiennes*.

11 • LOIS DES VALEURS EXTRÊMES ÉCHANTILLONS ARTIFICIELS

11.1 Échantillons ordonnés et statistique d'ordre

On considère un échantillon de taille n d'une variable aléatoire X . F est la fonction de répartition et f la densité de cette variable.

Il est parfois nécessaire d'étudier le n -uplet ordonné de ces observations, c'est-à-dire la suite des valeurs observées, classées par valeurs croissantes ou décroissantes dans le but :

- soit de rechercher les valeurs aberrantes, trop grandes ou trop petites, d'une série d'observations,
- soit d'étudier la loi de la plus grande valeur d'une série d'observations (hauteur maximale des crues d'une rivière, intensité du plus fort tremblement de terre dans une région donnée...).

11.1.1 Définition d'une statistique d'ordre

Soit une suite finie d'observations indépendantes (X_i) , $i \in [1, n]$, classées par ordre croissant. On désigne par :

- $X_{(1)}$ la plus petite valeur observée, c'est-à-dire la plus petite des valeurs X_i ,
- $X_{(k)}$ la valeur de rang k ,
- et ainsi de suite jusqu'à la plus grande valeur observée $X_{(n)}$.

On écrit cette suite d'observations sous la forme :

$$X_{(1)} \leq X_{(2)} \leq \dots \leq X_{(n)}$$

La *suite ordonnée des $X_{(i)}$* est appelée *statistique d'ordre associée à la série des observations (X_i)* .

À un événement, ces variables font correspondre la suite obtenue en classant les observations par ordre croissant.

Remarques

- On aurait dû écrire $X_{(i,n)}$ car le rang d'une observation dépend du nombre n des observations. Cependant, quand il n'y a pas d'ambiguïté, on écrit simplement $X_{(i)}$.
- Si la loi de X est une loi continue, on peut se limiter à des inégalités strictes :

$$X_{(1)} < X_{(2)} < \dots < X_{(n)}$$

car l'événement $X = k$ est un événement de probabilité nulle, $\Pr(X = k) = 0$.

- La quantité $X_{(n)} - X_{(1)}$ est l'*étendue de l'échantillon*.

11.1.2 Fonction de répartition empirique d'un échantillon

■ Définition

Soit $F^*(x)$ la proportion des X_i inférieures à x . Pour toute valeur de x , $F^*(x)$ est une variable aléatoire définissant la *fonction de répartition empirique* de l'échantillon. C'est une fonction en escalier de sauts égaux à $1/n$ si toutes les valeurs observées sont distinctes.

Si les valeurs de l'échantillon sont ordonnées par valeurs croissantes, alors :

$$\begin{aligned} x < x_1 & \quad F^*(x) = 0 \\ x_{i-1} \leq x < x_i & \quad F^*(x) = \frac{i-1}{n} \\ x_n \leq x & \quad F^*(x) = 1 \end{aligned}$$

Pour calculer les valeurs de la fonction de répartition empirique d'un échantillon, différentes formules d'approximation peuvent être utilisées. Parmi les plus connues, on peut citer :

- l'approximation de Haazen (1930) $F^*(x) = \frac{i - 0,5}{n}$
- l'approximation de Weibull (1939) $F^*(x) = \frac{i}{n + 1}$
- l'approximation de Chegodayev (1955) $F_n^*(x) = \frac{i - 0,3}{n + 0,4}$
- l'approximation de Tukey (1962) $F^*(x) = \frac{i - 1/3}{n + 1/3}$

L'approximation de Chegodayev est la meilleure formule d'approximation, l'erreur maximale est inférieure à 1 % quelle que soit la taille n de l'échantillon et elle diminue lorsque le rang i se rapproche de $n/2$.

■ Convergence de la fonction de répartition empirique

- $\forall x \quad F^*(x) \rightarrow F(x)$ la convergence étant presque sûre.

En effet, le nombre Y de variables aléatoires X_i inférieures à x est une somme de variables aléatoires de Bernoulli de paramètre $F(x)$. Donc, $F^*(x)$, qui est égal à Y/n , converge presque sûrement vers la probabilité $F(x)$.

- La convergence de $F^*(x)$ vers $F(x)$ est presque sûrement uniforme (théorème de Glivenko-Cantelli) :

$$D_n = \text{Sup} |F^*(x) - F(x)| \rightarrow 0$$

(le Sup est pris sur toutes les valeurs de x).

- Théorème de Kolmogoroff :

$$n \rightarrow \infty \quad \lim \Pr (\sqrt{n} D_n < y) = K(y) = \sum_{-\infty}^{+\infty} (-1)^k e^{-2k^2 y^2}$$

La distribution asymptotique de D_n ne dépend pas de la variable aléatoire X ; elle permet donc de calculer des limites pour les valeurs de D_n . Des tables donnent les valeurs de la loi $K(y)$ de la variable aléatoire Y .

Les résultats précédents sont utilisés dans la théorie des tests d'ajustement (chapitre 16).

11.2 Loi de la variable $X_{(k)}$, réalisation de rang k

11.2.1 Fonction de répartition

Soit $R_n(x)$ le nombre de répétitions de l'événement $(X < x)$ au cours de n épreuves indépendantes. Par définition :

$$F(x) = \Pr(X < x)$$

Pour x fixé, cette probabilité est constante au cours des n épreuves. La variable aléatoire $R_n(x)$ suit donc la loi binomiale $B[n; F(x)]$. D'où :

$$\Pr[R_n(x) = h] = C_n^h [F(x)]^h [1 - F(x)]^{n-h}$$

La réalisation de l'événement $X_{(k)} < x$ implique que :

- k valeurs de la variable X , *au moins*, soient inférieures à x ,
- on peut, cependant, en avoir $k + 1$, $k + 2$... jusqu'à n .

On en déduit la fonction de répartition $H_{(k)}(x)$ de la variable aléatoire $X_{(k)}$:

$$H_{(k)}(x) = \Pr(X_{(k)} < x) = \sum_{h=k}^n C_n^h [F(x)]^h [1 - F(x)]^{n-h}$$

11.2.2 Densité

La densité de $X_{(k)}$ peut être obtenue à partir de la définition :

$$h_{(k)}(x) dx = \Pr(x \leq X_{(k)} < x + dx)$$

La réalisation de cet événement implique que :

- au moins une des valeurs x_i appartienne à l'intervalle $[x, x + dx]$;
la probabilité de réalisation de cet événement est $n f(x) dx$ car il y a n choix possibles pour la valeur x_i ;
- $(k - 1)$ valeurs des x_i soient inférieures à x ;
la probabilité de réalisation de cet événement est $[F(x)]^{k-1}$;
- $(n - k)$ valeurs des x_i soient supérieures à x ;
la probabilité de réalisation de cet événement est $[1 - F(x)]^{n-k}$;
le nombre de réalisations possibles de cet événement est $C_{n-1}^{n-k} = C_{n-1}^{k-1}$.

La *densité de probabilité* de la variable aléatoire $X_{(k)}$ est donc égale à :

$$h_{(k)}(x) = n C_{n-1}^{k-1} [F(x)]^{k-1} [1 - F(x)]^{n-k} f(x)$$

11.2.3 Remarques

- La fonction de répartition $H_{(k)}(x)$ ne dépend que de $F(x)$, fonction de répartition de la variable X , mais non de la nature de cette variable.
- Si X est une variable continue, la densité de la loi de probabilité de la variable $X_{(k)}$ peut être obtenue en dérivant la fonction de répartition $H_{(k)}(x)$.
- Si X est une variable discrète, la densité de la loi de probabilité de $X_{(k)}$ est égale à :

$$h_{(k)}(x) = H_{(k)}(x+1) - H_{(k)}(x)$$

- Il existe une relation mathématique simple entre $H_{(k)}(x)$ et la fonction bêta incomplète :

$$H_{(k)}(x) = I_{F(x)}(k, n - k + 1)$$

où la fonction bêta incomplète est définie par :

$$I_u(p; q) = \int_0^u t^{p-1} (1-t)^{q-1} dt$$

La *fonction de répartition* $H_{(k)}(x)$ est égale à l'intégrale bêta incomplète, tronquée en $F(x)$.

11.3 Loi de la variable $X_{(n)}$, plus grande valeur observée

On peut obtenir sa loi, soit directement, soit en donnant à k la valeur n dans l'expression de la loi de $X_{(k)}$.

11.3.1 Fonction de répartition et densité de $X_{(n)}$

En remplaçant k par n , dans les expressions donnant la densité et la fonction de répartition de la variable aléatoire $X_{(k)}$, on obtient :

$$h_{(n)}(x) = n [F(x)]^{n-1} f(x)$$

$$H_{(n)}(x) = \Pr(X_{(n)} < x) = [F(x)]^n$$

Le raisonnement direct consiste à écrire :

$$\Pr(X_{(n)} < x) = \prod_{i=1}^n \Pr(X_i < x)$$

11.3.2 Limites de cette loi quand n tend vers l'infini

Quand n tend vers l'infini :

$$H_{(n)}(x) \text{ tend vers } 0 \text{ si } F(x) < 1$$

$$H_{(n)}(x) \text{ tend vers } 1 \text{ si } F(x) = 1$$

mais ces deux cas limites présentent peu d'intérêt.

En écrivant $H_{(n)}(x)$, sous la forme, $H_{(n)}(x) = \{1 - [1 - F(x)]\}^n$, on montre que, quelle que soit la loi de la distribution initiale, il n'existe que trois types de lois asymptotiques des valeurs extrêmes, non dégénérées :

- La loi de Weibull et la loi de Fréchet qui sont dites à décroissance algébrique et sont obtenues si $[1 - F(x)]$ tend vers 0 comme x^{-k} quand x tend vers l'infini.
- La loi de Gumbel obtenue si $[1 - F(x)]$ tend vers 0 comme e^{-x} quand x tend vers l'infini, cette loi est dite à décroissance exponentielle.

Ces trois lois sont utilisées pour représenter des phénomènes aléatoires telles que la magnitude des tremblements de terre, les crues des rivières...

On peut remarquer le caractère arbitraire de ces lois limites en l'absence de la connaissance de la fonction F .

■ Loi de Weibull

Une variable aléatoire réelle X suit une loi de Weibull si sa fonction de répartition F et sa densité de probabilité f sont données par :

$$F(t) = 0 \quad \forall t < \gamma$$

$$F(t) = 1 - \exp \left[- \left(\frac{t - \gamma}{\eta} \right)^\beta \right] \quad \forall t \geq \gamma$$

$$f(t) = 0 \quad \forall t < \gamma$$

$$f(t) = \frac{\beta}{\eta} \left(\frac{t - \gamma}{\eta} \right)^{\beta-1} \exp \left[- \left(\frac{t - \gamma}{\eta} \right)^\beta \right] \quad \forall t \geq \gamma$$

Les propriétés de cette loi sont étudiées dans le chapitre sur la fiabilité (chapitre 18), l'estimation des paramètres est donnée dans le chapitre 13 (exemple 13.10) et l'ajustement graphique dans le chapitre 16, paragraphe 16.1.1.

■ Loi de Fréchet

Une variable aléatoire réelle X suit une *loi de Fréchet* si sa *fonction de répartition* G et sa *densité de probabilité* g sont données par :

$$G(x) = \exp\left(-\exp - \frac{\ln x - x_0}{a}\right)$$

$$g(x) = \frac{1}{ax} \times \left(\exp - \frac{\ln x - x_0}{a}\right) \times \exp\left(-\exp - \frac{\ln x - x_0}{a}\right)$$

■ Loi de Gumbel

Une variable aléatoire réelle X suit une *loi de Gumbel* ou loi des valeurs extrêmes si sa *fonction de répartition* F et sa *densité de probabilité* f sont données par :

$$F(x) = \exp\left[-\exp\left(-\frac{x-x_0}{a}\right)\right] \quad a \neq 0$$

$$f(x) = \frac{1}{a} \exp\left(-\frac{x-x_0}{a}\right) \times \exp\left[-\exp\left(-\frac{x-x_0}{a}\right)\right]$$

Espérance mathématique et variance :

$$E(X) = x_0 + \gamma a = x_0 + 0,577216 \times a \quad \text{Var}(X) = \frac{\pi^2}{6} \times a^2$$

où γ est la constante d'Euler (annexe 2).

Les probabilités correspondant à la moyenne ont pour valeurs :

$$F(\mu) = F(x_0 + 0,577216 \times a) = \begin{cases} 0,570 & \text{si } a > 0 \\ 0,430 & \text{si } a < 0 \end{cases}$$

Médiane : elle correspond à $F(M_e) = 1/2$

$$M_e = x_0 + 0,366513 \times a$$

Mode : c'est la valeur x_0

$$F(x_0) = \begin{cases} 0,368 & \text{si } a > 0 \\ 0,632 & \text{si } a < 0 \end{cases}$$

La variable réduite est définie par $u = \frac{x - x_0}{a}$. Elle a pour densité et fonction de répartition, les expressions simplifiées suivantes :

$$G(x) = e^{-e^{-x}}$$

$$g(x) = e^{-x} e^{-e^{-x}}$$

Son espérance mathématique et sa variance sont égales à :

$$E(X) = \gamma = 0,577216 \quad \text{Var}(X) = \frac{\pi^2}{6}$$

Remarque

On appelle parfois cette loi la loi de Gumbel standard et on donne le nom de loi de Gumbel à la loi de la variable aléatoire Y dont les caractéristiques sont les suivantes :

Densité : $h(y) = \exp(y - \exp y) \quad \forall y \in \mathbb{R}$

Fonction de répartition : $H(y) = 1 - \exp(-\exp y)$

$$E(Y) = -\gamma = -0,577216 \quad \text{Var}(Y) = \frac{\pi^2}{6}$$

11.4 Loi de la variable $X_{(1)}$, plus petite valeur observée

Comme pour la variable aléatoire $X_{(n)}$, on peut obtenir sa loi directement ou en donnant à k la valeur 1 dans l'expression de la loi de $X_{(k)}$.

On obtient, pour la fonction de répartition :

$$H_{(1)}(x) = 1 - [1 - F(x)]^n$$

et pour la densité :

$$h_{(1)}(x) = nf(x) [1 - F(x)]^{n-1}$$

Pour faire le calcul directement, on écrit :

$$\Pr(X_{(1)} < x) = 1 - \Pr(X_{(1)} > x) = 1 - \prod_{i=1}^n [1 - \Pr(X_i < x)]$$

11.5 Échantillons artificiels et simulation

11.5.1 Introduction

Un échantillon aléatoire peut être obtenu très simplement, soit par tirages de boules dans une urne, soit par jets de pièce de monnaie (pile ou face) ou loteries...

Cependant, ces procédés ne peuvent pas être utilisés dans des cas plus généraux ou plus complexes. Or actuellement, les échantillons aléatoires sont devenus un outil important utilisé dans de nombreux domaines. Citons :

- Les *techniques de sondage* : les résultats obtenus sur un échantillon doivent être extrapolés à la population entière, accompagnés d'une appréciation sur la précision de l'extrapolation ; le sondage doit être aléatoire pour donner, à chaque élément de la population, une probabilité non nulle d'être interrogée.
- Les *domaines de simulation* : on peut être conduit à réaliser des expériences *fictives* pour simuler à grande vitesse l'évolution d'un phénomène, c'est-à-dire pour rendre *visibles* les manifestations du hasard. Ces techniques sont appliquées, par exemple, à des problèmes de file d'attente, de gestion des stocks, de fiabilité...
- La *méthode de Monte-Carlo* utilisée en calcul numérique : à un problème difficile à résoudre, on associe un phénomène aléatoire dont une caractéristique, la moyenne par exemple, est liée à la grandeur que l'on veut calculer. Une réalisation du processus et le calcul de la caractéristique permettent d'approcher la solution du problème. Ces méthodes dites *de Monte-Carlo* ont été appliquées au calcul d'intégrales, d'inverse de matrices, à la résolution d'équations différentielles...

Une intégrale, par exemple, peut être considérée comme l'espérance mathématique d'une variable aléatoire.

En résumé, toutes ces applications reposent sur la réalisation d'un échantillon aléatoire d'une variable aléatoire X déterminée. Un choix « au hasard » s'appuie sur des règles très précises pour obtenir un échantillon représentatif de la population.

11.5.2 Principe de la construction d'un échantillon

On considère une population constituée de N unités statistiques dans laquelle on veut prélever un échantillon de taille n . On dispose d'une liste complète des N unités, cette liste constitue la *base de sondage*.

Pour construire un échantillon aléatoire, on peut attribuer un numéro unique et différent à chaque unité; ensuite tirer au sort n numéros constituant l'échantillon aléatoire (chaque unité a une probabilité non nulle d'être tirée). Si chaque unité a la même probabilité d'être tirée, on obtient un échantillon aléatoire simple.

■ Tirage sans remise ou tirage exhaustif

Les unités tirées n'étant pas remises dans la population, chaque unité figure au plus une fois dans la population et la composition de la base de sondage varie à chaque tirage.

■ Tirage avec remise ou tirage indépendant

Chaque unité tirée est examinée puis remise dans la population; chaque unité peut donc figurer plus d'une fois dans l'échantillon mais la composition de la base de sondage ne varie pas au cours de ce processus de tirage.

Remarque

C'est le premier mode de tirage qui est le plus utilisé. Il donne des estimations plus précises pour une même taille de l'échantillon. Cependant, si la taille de l'échantillon est petite par rapport à la taille de la population, les deux modes de tirage donnent des résultats comparables.

■ Construction d'un échantillon à l'aide d'une table de nombres au hasard

Une table de nombres au hasard est constituée des chiffres 0, 1... 9, chacune de ces valeurs ayant la même probabilité d'apparition. La table 9 est une page de nombres au hasard obtenue par la fonction *aléa* du logiciel Excel.

Un nombre quelconque de la table n'a aucun rapport avec le nombre qui est à sa droite, à sa gauche, au-dessus, au-dessous.

Pour rendre la lecture de ces tables plus facile, les nombres sont en général, regroupés par colonnes de 5 chiffres, chaque ligne comprenant 50 nombres (10×5). Pour obtenir les nombres qui seront utilisés pour constituer l'échantillon, on choisit un point de départ quelconque, puis on définit un itinéraire de parcours (on lit les nombres en lignes en sautant 2 nombres, ou bien on les lit en colonnes, ou en diagonales...).

Exemple 11.1

On veut tirer un échantillon de taille 30 dans une population de 300 unités ; on numérote ces unités de 001 à 300. On choisit comme point de départ, un nombre quelconque de 3 chiffres, compris entre 1 et 300. Puis, on décide de lire la table en colonnes, en sautant une ligne après chaque lecture, par exemple, ou toute autre règle de lecture. On ne garde que les nombres compris entre 1 et 300, et on élimine tous ceux qui sont déjà sortis.

11.5.3 Principe de la construction de nombres au hasard ou de nombres pseudo-aléatoires

La construction de nombres au hasard exige des procédés très précis et très complexes (boules de loterie, par exemple). De nombreux logiciels donnent des tables de nombres au hasard, plus ou moins exactes. Il existe de nombreuses tables de nombres au hasard. On peut citer une liste non exhaustive :

- Les tables de Fisher and Yates (Statistical Tables for Biological, Agricultural and Medical Research).
- Les tables de Kendall and Babington Smith, 100 000 chiffres obtenus à partir d'un disque tournant divisé en secteurs multiples de 10, éclairés de façon intermittente.

- Les tables de la Rand Corporation, 1 000 000 de chiffres obtenus à partir de l'écrêtement d'un bruit de fond...

L'utilisation de ces tables étant parfois lourde, on construit, par des procédés itératifs, non pas des nombres au hasard mais des nombres *pseudo-aléatoires*.

Ces procédés sont basés sur la construction de suites récurrentes qui donnent des suites périodiques, on essaie donc d'obtenir des périodes très grandes.

Comme les nombres au hasard, les nombres pseudo-aléatoires doivent vérifier certaines propriétés telles que l'équirépartition des 10 chiffres, l'indépendance des termes...

Exemple 11.2

La suite des décimales, de rang un nombre premier, du nombre π peut être considérée comme une suite de nombres pseudo-aléatoires.

Citons deux méthodes de construction de nombres pseudo-aléatoires :

- La méthode du milieu du carré de von Neumann. On choisit un nombre que l'on élève au carré, on en prend la partie médiane que l'on élève au carré et on recommence... Ces nombres ne sont pas aléatoires au sens strict, car ils dépendent du choix du premier nombre et de plus ils ont une faible période. Cette méthode est donc peu utilisée.
- La méthode de Lehmer. On définit une suite $\{x_n\}$ de la façon suivante :
 - x_0 est un entier arbitraire positif,
 - $x_{n+1} = kx_n \pmod{m}$ avec $m = 2^{31} - 1$ et $k = 23$.

La période est égale à $(m - 1)/2 = 1\,073\,741\,823$.

11.5.4 Tirage d'un échantillon artificiel de N valeurs d'une variable aléatoire continue

Soit X une variable aléatoire de fonction de répartition F continue et strictement croissante, de densité de probabilité f . À cette variable aléatoire X , on associe la variable aléatoire Y définie par $Y = F(X)$ qui a pour densité de probabilité :

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

La variable aléatoire Y est donc uniformément répartie sur $[0, 1]$.

Pour obtenir un échantillon de n valeurs de la variable X , il suffit de tirer n nombres uniformément répartis sur $[0, 1]$, soient $t_1, \dots, t_n$ ces nombres ; l'échantillon est constitué des valeurs $x_i = F^{-1}(t_i)$. On dit que l'on a simulé la variable aléatoire X .

Exemple 11.3 Construction d'un échantillon de 10 valeurs d'une loi exponentielle de paramètre $\lambda = 3$

La fonction de répartition d'une loi exponentielle est $F(x) = 1 - e^{-\lambda x}$.

Soient $t_1, \dots, t_{10}$ un échantillon de 10 nombres au hasard, relevés sur une table de nombres au hasard et compris entre 0 et 1. Les valeurs x_i sont données par la formule $x_i = -\frac{1}{3} \ln(1 - t_i)$.

t_i	$1 - t_i$	x_i
0,86593	0,13407	0,6698
0,84980	0,1502	0,6320
0,68645	0,31355	0,3866
0,36493	0,63507	0,1513
0,83679	0,16321	0,6042
0,57494	0,42506	0,2852
0,14499	0,85501	0,0522
0,42237	0,57763	0,1830
0,05764	0,94236	0,0198
0,22190	0,7781	0,0836

La moyenne des 10 valeurs de cet échantillon est égale à 0,3068, elle est peu différente de la valeur théorique égale à $1/3$.

11.5.5 Applications

■ Variable de Bernoulli et variable binomiale, de paramètre p

Pour simuler une variable aléatoire binomiale, on utilise la propriété suivante : une variable aléatoire binomiale $B(n; p)$ est une *somme* de n variables aléatoires de Bernoulli *indépendantes*.

On commence par simuler une variable aléatoire de Bernoulli X .

On tire un nombre au hasard entre 0 et 1, soit r ce nombre.

Si $r < p$, $X = 1$; si $r > p$, $X = 0$.

Puis, on fait la somme des n variables aléatoires indépendantes de Bernoulli ainsi obtenues pour simuler une variable aléatoire binomiale.

■ Loi gamma Γ_p avec p entier

Une variable aléatoire X suivant une loi γ_p est la *somme* de p variables aléatoires indépendantes Y_i , chaque variable Y_i suivant une loi γ_1 .

La fonction de répartition d'une variable aléatoire Y suivant une loi γ_1 est :

$$F(y) = 1 - e^{-y}$$

On commence par simuler la loi γ_1 .

Soit r une réalisation d'une variable aléatoire R uniformément distribuée sur $[0, 1]$. La variable aléatoire, $1 - R$, est aussi distribuée uniformément sur $[0, 1]$.

Pour simuler Y , il suffit donc de poser :

$$y = -\text{Ln } r$$

Puis, pour simuler X , il suffit de poser, si p est un entier :

$$x = -\text{Ln } r_1 - \text{Ln } r_2 \cdots - \text{Ln } r_p = -\text{Ln } \prod_{k=1}^p r_k$$

■ Loi de Poisson $P(\lambda)$

Pour simuler une loi de Poisson de paramètre λ , on simule un processus de Poisson de cadence 1 sur une période égale à λ .

Les intervalles successifs, $[E_i, E_{i+1}]$, sont des réalisations indépendantes de variables aléatoires suivant des lois γ_1 . On simule des variables γ_1 et on ajoute leurs valeurs jusqu'au moment où on dépasse la valeur λ . La réalisation n de la variable aléatoire de Poisson est le plus grand entier n tel que :

$$\sum_{i=1}^n (-\text{Ln } r_i) < \lambda$$

■ Loi normale

Une méthode est obtenue en utilisant le théorème central limite.

La variable aléatoire, $\frac{\bar{X} - m}{\sigma / \sqrt{n}}$, converge en loi vers une variable suivant la loi normale centrée réduite, $N(0 ; 1)$, quand n tend vers l'infini. On applique ce résultat à la somme de n variables aléatoires suivant une loi uniforme sur $[0, 1]$. L'espérance mathématique est égale à $1/2$ et la variance à $1/12$. Ce résultat est valable dès que n est supérieur à 12.

□ **Méthode de Box et Müller**

Cette méthode utilise le résultat suivant.

Soient U et V deux variables aléatoires, indépendantes, suivant des lois uniformes sur $[0, 1]$. Les variables aléatoires X et Y définies par :

$$X = (-2 \ln U)^{1/2} \cos 2\pi V$$

$$Y = (-2 \ln U)^{1/2} \sin 2\pi V$$

sont deux variables aléatoires, normales, centrées, réduites et indépendantes.

En effet, la densité du couple de variables indépendantes U, V , suivant des lois uniformes sur $[0, 1]$ est égale à $f(u, v) = 1$.

$$U = \exp \left[-\frac{1}{2} (X^2 + Y^2) \right]$$

$$V = \frac{1}{2\pi} \operatorname{Arctg} \frac{Y}{X}$$

Le jacobien de la transformation (calcul facile) est égal à :

$$-\frac{1}{2\pi} \exp \left[-\frac{1}{2} (x^2 + y^2) \right]$$

Comme la densité du couple (U, V) est égale à 1, il en résulte que la densité du couple de variables aléatoires (X, Y) est égale à :

$$g(x, y) = \frac{1}{2\pi} \exp \left[-\frac{1}{2} (x^2 + y^2) \right]$$

On reconnaît la densité d'un couple de variables aléatoires normales, centrées, réduites, indépendantes.

Il suffit de simuler deux variables U et V suivant une loi uniforme sur $[0, 1]$. Soient r_1 et r_2 , des valeurs au hasard, tirées entre 0 et 1 ; la valeur correspondante x de X est donnée par :

$$x = \sqrt{-2 \ln r_1} \cos (2\pi r_2)$$

12 • THÉORIE DE L'ESTIMATION

12.1 Exposé du problème et exemples

Un aspect important de l'inférence statistique consiste à obtenir des *estimations fiables* des caractéristiques d'une population à partir d'un échantillon extrait de cette population. C'est un *problème de décision* concernant des paramètres tels que :

- l'espérance mathématique notée m ou μ (pour un caractère mesurable),
- la variance ou l'écart-type notée σ ,
- la proportion p (pour un caractère dénombrable).

Comme un échantillon ne peut donner qu'une information partielle sur la population, les estimations ainsi obtenues seront inévitablement entachées d'*erreurs* que l'on doit minimiser autant que possible. En résumé :

Estimer un paramètre, c'est donner une valeur approchée de ce paramètre, à partir des résultats obtenus sur un échantillon aléatoire extrait de la population.

Exemple 12.1

On veut étudier une caractéristique X d'un phénomène économique, par exemple la proportion p des individus d'une population P d'effectif N , présentant un certain caractère : posséder un magnétoscope, avoir passé ses dernières vacances à l'étranger, etc.

Pour obtenir la valeur exacte de p , il suffirait d'interroger les N individus de la population ce qui, en général, est impossible ; on interroge donc les individus d'un échantillon aléatoire, de taille n , représentatif de la population. On obtient un ensemble de n valeurs x_i telles que :

$x_i = 1$ si l'individu i présente le caractère X , $x_i = 0$ sinon.

Soit f_n la proportion des individus ayant le caractère X dans l'échantillon. Cette proportion f_n converge vers p quand n tend vers l'infini (loi des grands nombres). On estimera donc p par f_n . C'est une *estimation ponctuelle* qui donne une seule valeur pour p .

Par un choix adéquat de n , on peut essayer de minimiser l'erreur due à cette approximation.

Exemple 12.2

Un autre type de problème relevant de la théorie de l'estimation consiste en l'estimation des paramètres de la loi suivie par une variable aléatoire X , loi dont on connaît la forme.

Ainsi, par exemple, le nombre d'accidents dans un atelier pendant une semaine suit probablement une loi de Poisson. Cette loi dépend d'un paramètre λ dont on ne connaît pas la valeur. Pour donner une valeur à ce paramètre, on note le nombre d'accidents survenus pendant n semaines, c'est l'échantillon aléatoire de taille n .

Le nombre moyen d'accidents est une estimation ponctuelle du paramètre λ (propriété de la loi de Poisson).

Sous une forme générale, le problème à résoudre peut se formuler ainsi :

Soit X une caractéristique d'un phénomène dont les réalisations dépendent du « hasard ». La variable X est donc une variable aléatoire dont les caractéristiques (moment, variance...) sont inconnues. On observe les réalisations d'un échantillon aléatoire issu de la population étudiée. Cette réalisation doit permettre d'induire des valeurs ou *estimations* des paramètres de la loi suivie par la variable X . Ces estimations peuvent revêtir deux formes :

- soit une valeur unique, l'*estimation ponctuelle*, ou valeur la plus probable que prendra le paramètre,
- soit un ensemble de valeurs appartenant à un intervalle, l'*estimation par intervalle de confiance*. Un intervalle de confiance doit avoir de « grandes chances » de contenir la vraie valeur du paramètre, il est toujours associé à un risque d'erreur α .

La théorie de l'estimation fait intervenir des fonctions ou statistiques particulières, appelées *estimateurs*, dont nous allons donner les propriétés essentielles puis, nous étudierons principalement les statistiques exhaustives et la quantité d'information apportée par un échantillon de taille n .

12.2 Définition d'une statistique

X est une variable aléatoire dont la fonction de répartition $F(x; \theta)$ et la densité $f(x; \theta)$ dépendent du paramètre θ ; D_θ est l'ensemble des valeurs possibles de ce paramètre.

On considère un échantillon de taille n de cette variable $\underline{X} = (X_1, \dots, X_n)$.

Une *statistique* est une fonction mesurable T des variables aléatoires X_i :

$$T(X_1, \dots, X_n)$$

À un échantillon, on peut associer différentes statistiques. La théorie de l'estimation consiste à définir des statistiques particulières, appelées *estimateurs*.

Une fois l'échantillon effectivement réalisé, l'estimateur prend une valeur numérique, appelée *estimation* du paramètre θ .

On notera $\hat{\theta}$ l'estimateur du paramètre θ .

Exemple 12.3

Soit $\bar{X}$ la statistique :

$$\bar{X} = T(X_1, \dots, X_n) = \frac{1}{n} \sum_{i=1}^n X_i$$

c'est-à-dire la fonction *moyenne arithmétique* des n observations d'un échantillon.

Cette statistique peut être considérée comme un estimateur, *a priori* raisonnable, de l'espérance mathématique $E(X) = m$. En effet :

- cette statistique prend en compte toutes les observations,
- cette statistique possède les propriétés suivantes :

$$E(T) = E(X) \quad \text{Var}(T) = \frac{\text{Var}(X)}{n}$$

- si la loi de la variable aléatoire X est connue, on peut en déduire celle de $\bar{X}$.

L'estimation ponctuelle de l'espérance mathématique m de la variable X est obtenue en réalisant effectivement un échantillon de taille n et en calculant la moyenne arithmétique des n observations.

On montrera que les deux premières propriétés de cet estimateur sont, en fait, de « bonnes propriétés » pour un estimateur.

12.3 Statistique exhaustive

Une statistique T , dépendant d'un échantillon de taille n , apporte des informations sur un paramètre θ si sa loi de probabilité dépend de ce paramètre. Si la loi conditionnelle de l'échantillon, la statistique $T = t$ étant supposée connue, ne dépend plus du paramètre θ , cet échantillon ne peut plus donner d'informations sur θ . La statistique T a donc apporté toute l'information possible sur le paramètre. Une telle statistique est appelée *statistique exhaustive* ou *résumé exhaustif* pour le paramètre θ .

D'où, la définition d'une *statistique exhaustive* (les notations sont celles du paragraphe 12.2) :

Soit E_θ l'ensemble de définition :

$$E_\theta = \left\langle f(x; \theta) > 0 \quad \forall x \in \mathbb{R} \quad \text{et} \quad \forall \theta \in D_\theta \subset \mathbb{R} \right\rangle$$

que l'on notera E s'il ne dépend pas de θ .

Les variables aléatoires (X_i) étant indépendantes, la densité de l'échantillon $\underline{X}$ est :

$$L(\underline{x}; \theta) = \prod_i f(x_i; \theta) \quad \forall \theta \in D_\theta \quad \text{et} \quad \forall \underline{x} \in \mathbb{R}^n$$

où $\underline{x} = (x_1, \dots, x_n)$ est une réalisation de l'échantillon $\underline{X}$.

Cette densité $L(\underline{x}; \theta)$ est une fonction de θ appelée *vraisemblance de l'échantillon*. Elle peut se mettre sous la forme :

$$L(\underline{x}; \theta) = g(t; \theta) h(\underline{x}; \theta / T = t)$$

$g(t; \theta)$ est la densité de la statistique T .

$h(\underline{x}; \theta / T = t)$ est la densité conditionnelle de l'échantillon sachant $T = t$.

La statistique T est une *statistique exhaustive* si la densité conditionnelle de l'échantillon sachant T ne dépend pas de θ , c'est-à-dire si :

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

Quand la valeur t de la statistique est connue, l'échantillon n'apporte plus aucune information sur le paramètre θ .

12.3.1 Propriétés d'une statistique exhaustive

- La propriété d'exhaustivité pour une statistique est intéressante si elle ne dépend pas de la taille de l'échantillon.
- Soient T une statistique exhaustive pour le paramètre θ et Ψ une fonction strictement monotone de T . Alors :

La statistique $S = \Psi(T)$ est une *statistique exhaustive* pour le paramètre θ .

12.3.2 Exemple et contre-exemple

Exemple 12.4

X est une variable aléatoire suivant une loi uniforme sur $[0, \theta]$. Elle a pour densité :

$$f(x, \theta) = \begin{cases} \frac{1}{\theta} & \forall x \in [0, \theta] = E_\theta \quad \theta > 0 \\ 0 & \text{sinon} \end{cases}$$

La statistique :

$$R = \text{Sup } X_i \quad i \in [1, n]$$

est un résumé exhaustif de l'échantillon $\underline{X} = (X_1, \dots, X_n)$ pour le paramètre θ .

En effet :

$$\text{– Vraisemblance de l'échantillon : } \begin{cases} L(\underline{x}, \theta) = \frac{1}{\theta^n} & \forall x \in [0, \theta] \\ L(\underline{x}, \theta) = 0 & \text{sinon} \end{cases}$$

– Loi de la statistique R :

$$\text{Fonction de répartition : } \begin{cases} G(r, \theta) = 0 & r < 0 \\ G(r, \theta) = \Pr(R < r) = \left(\frac{r}{\theta}\right)^n & 0 \leq r \leq \theta \\ G(r, \theta) = 1 & r \geq \theta \end{cases}$$

$$\text{Densité : } \begin{cases} g(r, \theta) = \frac{n r^{n-1}}{\theta^n} & 0 \leq r \leq \theta \\ g(r, \theta) = 0 & \text{sinon} \end{cases}$$

$$\text{D'où la vraisemblance : } L(\underline{x}, \theta) = \frac{1}{\theta^n} = \frac{n r^{n-1}}{\theta^n} \times \frac{1}{n r^{n-1}}$$

On retrouve la factorisation d'une statistique exhaustive.

Exemple 12.5

Soit X une variable aléatoire de densité :

$$f(x, \theta) = \frac{\theta}{e^{\theta^2} - 1} e^{x\theta} \quad \forall x \in [0, \theta] = E_\theta \quad \theta > 0$$

$$f(x, \theta) = 0 \quad \text{sinon}$$

On montre que la statistique $T = \sum_{i=1}^n X_i$ n'est pas un résumé exhaustif pour un échantillon de taille n pour le paramètre θ .

En fait, cette variable ne permet pas de résumé exhaustif pour θ .

12.3.3 Forme canonique des lois de probabilité admettant une statistique exhaustive et théorème de Darmois

On garde les notations des paragraphes 12.2 et 12.3, et on suppose que l'ensemble de définition E ne dépend pas de θ . Le *théorème de Darmois* donne les conditions d'existence d'une statistique exhaustive.

S'il existe un entier $n > 1$ tel que l'échantillon $\underline{X}$ admette une statistique exhaustive pour le paramètre θ , la fonction $f(x; \theta)$ est de la forme :

$$f(x; \theta) = \exp [a(x) \alpha(\theta) + b(x) + \beta(\theta)] \quad \forall x \in E \quad \forall \theta \in D_\theta$$

ou de la forme équivalente :

$$\ln [f(x; \theta)] = [a(x) \alpha(\theta) + b(x) + \beta(\theta)] \quad \forall x \in E \quad \forall \theta \in D_\theta$$

Si f est de la forme exponentielle précédente et si l'application :

$$x_j \rightarrow t = \sum_i a(x_i)$$

est bijective et continûment différentiable pour tout x_j , alors la statistique T :

$$T = \sum_i a(X_i)$$

est une statistique exhaustive particulière pour le paramètre θ .

Remarque

Si l'ensemble E dépend de θ , la première partie du théorème de Darrois est vraie, mais pas la deuxième partie. Cependant, il peut exister une statistique exhaustive que l'on trouve par d'autres méthodes.

Exemple 12.6

Reprenons l'exemple de la loi uniforme de densité $f(x; \theta) = 1/\theta$ sur $[0, \theta]$.

La densité de la variable X est de la « forme exponentielle » :

$$f(x; \theta) = e^{-\text{Ln } \theta} \quad \text{ou} \quad \text{Ln } f(x; \theta) = -\text{Ln } \theta$$

On remarque que $a(x) = 1$. Cependant, on vérifie que la statistique T :

$$T = \sum_i a(X_i) = n$$

n'est pas une statistique exhaustive pour le paramètre θ . En effet, le domaine de définition de la variable X dépend de θ .

La statistique exhaustive pour le paramètre θ est la statistique $R = \text{Sup}(X_i)$.

Exemple 12.7

La variable aléatoire X suit une loi de Bernoulli de paramètre p . Son domaine de définition ne dépend pas de p . La densité est :

$$\begin{aligned} f(x; p) &= \text{Pr}(X = x) = C_n^x p^x (1-p)^{n-x} \\ \text{Ln } f(x; p) &= \text{Ln } C_n^x + x \text{Ln } p + (n-x) \text{Ln } (1-p) \\ &= \text{Ln } C_n^x + x \left[\text{Ln } \frac{p}{1-p} \right] + n \text{Ln } (1-p) \end{aligned}$$

La statistique $T = \sum_{i=1}^n X_i$ est une statistique exhaustive pour le paramètre p .

Exemple 12.8

La variable aléatoire X suit une loi de Poisson de paramètre λ . Le domaine de définition ne dépend pas du paramètre λ . La densité est :

$$f(x; \lambda) = e^{-\lambda} \frac{\lambda^x}{x!}$$

$$\text{Ln } f(x; \lambda) = x \text{Ln } \lambda - \text{Ln } (x!) - \lambda$$

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

Exemple 12.9

La variable aléatoire X suit une loi normale $N(m; \sigma)$. Cette variable vérifie toutes les hypothèses du théorème de Darmois. En effet, la densité est de la forme exponentielle :

$$\begin{aligned}\ln f(x; \theta) &= -\ln \sigma \sqrt{2\pi} - \frac{(x-m)^2}{2\sigma^2} \\ &= -\frac{x^2}{2\sigma^2} + \frac{mx}{\sigma^2} - \frac{m^2}{2\sigma^2} - \ln \sigma \sqrt{2\pi}\end{aligned}$$

et le domaine de définition ne dépend pas de la valeur des paramètres.

– Cas 1 : l'écart-type σ est connu, le paramètre inconnu à estimer est la moyenne m .

La statistique $T = \sum_{i=1}^n X_i$ est une statistique exhaustive pour la moyenne m .

– Cas 2 : la moyenne m est connue, le paramètre inconnu à estimer est l'écart-type σ .

La statistique $T = \sum_{i=1}^n (X_i - m)^2$ est une statistique exhaustive pour la variance.

Exemple 12.10

La variable aléatoire X suit une loi exponentielle de densité :

$$f(x; \theta) = \frac{1}{\theta} e^{-x/\theta}$$

Le domaine de définition ne dépend pas de la valeur du paramètre.

La statistique $T = \sum_{i=1}^n X_i$ est une statistique exhaustive pour le paramètre θ .

Exemple 12.11

Soit X une variable aléatoire suivant une loi gamma $\Gamma(\theta; 1)$, de densité :

$$f(x; \theta) = \frac{1}{\Gamma(\theta)} e^{-x} x^{\theta-1}$$

$$\ln f(x; \theta) = -x + (\theta - 1) \ln x - \ln \Gamma(\theta)$$

Le domaine de définition ne dépend pas de la valeur du paramètre.

La statistique $T = \sum_{i=1}^n \ln X_i$ est une statistique exhaustive pour le paramètre θ .

Exemple 12.12

La variable aléatoire X suit une loi de Cauchy de densité :

$$f(x; \theta) = \frac{1}{\pi} \frac{1}{1 + (x - \theta)^2} \quad \forall x \in \mathbb{R}$$

$$\text{Ln} f(x; \theta) = -\text{Ln} [1 + (x - \theta)^2] - \text{Ln} \pi$$

$\text{Ln} f(x; \theta)$ ne vérifie pas les conditions du théorème de Darmois, il n'existe pas de résumé exhaustif pour le paramètre θ .

12.4 Information de Fisher

La *quantité d'information apportée par un échantillon* est l'expression suivante, sous réserve de l'existence de l'intégrale :

$$\begin{aligned} I_n(\theta) &= I_{x_1, \dots, x_n}(\theta) = E \left[\left(\frac{d \text{Ln} L(\underline{x}; \theta)}{d\theta} \right)^2 \right] \\ &= \int_{E_\theta} \left[\frac{d \text{Ln} L(\underline{x}; \theta)}{d\theta} \right]^2 L(\underline{x}; \theta) d\underline{x} \end{aligned}$$

Les notations sont celles des paragraphes précédents.

12.4.1 Propriétés de la quantité d'information

Si l'ensemble E_θ ne dépend pas de θ et si la vraisemblance $L(\underline{x}; \theta)$ est dérivable au moins jusqu'à l'ordre deux, la quantité d'information de Fisher possède les propriétés suivantes :

$$\frac{d \text{Ln} L(\underline{x}; \theta)}{d\theta} \text{ est une variable aléatoire centrée}$$

$$I_n(\theta) = \text{Var} \left[\frac{d \text{Ln} L(\underline{x}; \theta)}{d\theta} \right]$$

$$I_n(\theta) = -E \left[\frac{d^2 \text{Ln} L(\underline{x}; \theta)}{d\theta^2} \right]$$

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

Remarque

Précision apportée par un échantillon : supposons que le paramètre θ à estimer soit la moyenne d'une loi normale. En remplaçant $f(x; \theta)$ par la densité de probabilité d'une loi normale dans les expressions précédentes, on obtient :

$$I_n(\theta) = \frac{n}{\sigma^2} = n I_1(\theta)$$

L'information $I_n(\theta)$ est donc d'autant plus grande que l'écart-type σ est petit (justification du mot « précision »).

Exemple 12.13

L'exemple suivant montre le rôle important des hypothèses formulées au début du paragraphe. Soit X la variable aléatoire suivant une loi uniforme sur $[0, \theta]$. Le domaine de définition dépend du paramètre θ . Un calcul facile donne :

$$I_1(\theta) = \frac{1}{\theta^2} \quad I_n(\theta) = \frac{n^2}{\theta^2} \neq n I_1(\theta)$$

Les propriétés précédentes (paragraphe 12.4.1) ne sont pas vérifiées.

12.4.2 Dégradation de l'information

Si l'ensemble E_θ ne dépend pas du paramètre θ , l'information apportée par un échantillon est supérieure ou égale à l'information apportée par une statistique.

Il y a égalité si la statistique est exhaustive (justification du qualificatif « exhaustive »). Ce résultat sera démontré dans le chapitre 13 (estimation ponctuelle).

13 • ESTIMATION PONCTUELLE

Le but de la théorie de l'estimation est de choisir, parmi toutes les statistiques possibles, le meilleur estimateur, c'est-à-dire celui qui donnera une *estimation ponctuelle* la plus proche possible du paramètre et ceci, quel que soit l'échantillon.

13.1 Définition d'un estimateur

Soit X une variable aléatoire dont la loi de probabilité $f(x; \theta)$ dépend d'un seul paramètre θ . Le cas de plusieurs paramètres sera traité dans le paragraphe 13.6.

$\underline{X} = (X_1, \dots, X_n)$ est un échantillon de taille n de cette variable (variable parente).

Une *statistique* est une fonction mesurable $T(\underline{X})$ des variables aléatoires X_i . On note, en général, $\hat{\theta}(\underline{X})$ ou $\hat{\theta}_n$, ou plus simplement, $\hat{\theta}$ l'estimateur du paramètre θ .

Exemple 13.1

La moyenne arithmétique $\bar{X}$ des n observations est un exemple de statistique :

$$\bar{X} = T(X_1, \dots, X_n) = \frac{1}{n} \sum_{i=1}^n X_i$$

Un *estimateur* est une statistique qui a des propriétés bien définies.

Une suite T_n de statistiques, $T_n = \varphi(\underline{X}_n)$, est appelée *estimateur du paramètre* θ si T_n tend θ quand n tend vers l'infini, la convergence étant une convergence en probabilité presque sûre ou en moyenne quadratique.

Une fois l'échantillon effectivement réalisé, l'estimateur prend une valeur numérique, appelée *estimation ponctuelle* du paramètre θ par la statistique T .

13.2 Principales qualités d'un estimateur

13.2.1 Estimateur convergent

La première condition imposée à un estimateur est d'être *convergent* :

Un estimateur est *convergent* si sa distribution tend à se concentrer autour de la valeur inconnue du paramètre θ quand la taille n de l'échantillon tend vers l'infini.

$$\forall \varepsilon, \forall \eta \quad \exists n_0 \in \mathbb{N}^* \text{ tel que } n > n_0 \Rightarrow \Pr \left\{ \left| \hat{\theta}_n - \theta \right| < \varepsilon \right\} > 1 - \eta$$

Cette condition implique en particulier que $E(T)$ soit égale à θ et $\text{Var}(T)$ faible. Il suffit d'appliquer l'inégalité de Bienaymé-Tchebyshev pour le démontrer.

Inégalité de Bienaymé-Tchebyshev (ses propriétés sont données chapitre 4, paragraphe 4.7.4) :

$$\Pr(|X - E(X)| < k\sigma_X) > 1 - \frac{1}{k^2}$$

La *statistique* $\bar{X}$ (exemple 13.1) est donc un *estimateur convergent* pour l'espérance mathématique.

Pour un paramètre donné, on peut trouver différents estimateurs convergents, mais en général, ils convergent avec des vitesses différentes. Les figures 13.1 et 13.2 illustrent cette notion et montrent les deux caractéristiques importantes d'un estimateur.

13.2.2 Estimateur sans biais

L'*erreur d'estimation* est mesurée par la quantité $T - \theta$ qui peut s'écrire :

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

- $T - E(T)$ représente les fluctuations de l'estimateur T autour de sa valeur moyenne $E(T)$ (espérance mathématique).

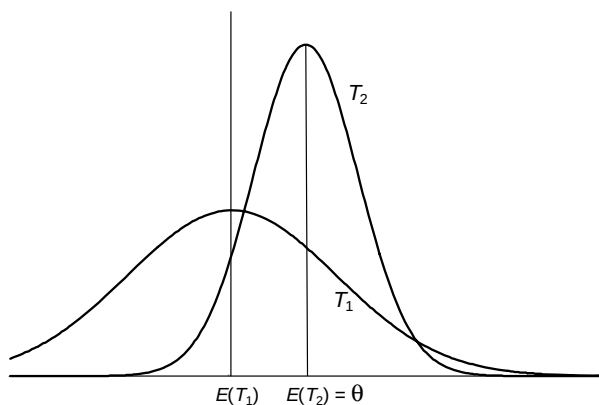


Figure 13.1 – Comparaison d'estimateur avec $E(T_1) \neq \theta$ et $E(T_2) = \theta$.

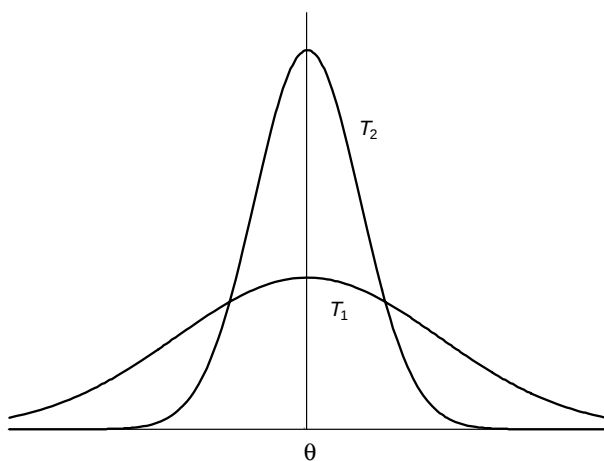


Figure 13.2 – Comparaison d'estimateur avec $E(T_1) = E(T_2)$ et $\text{Var}(T_1) > \text{Var}(T_2)$.

- $E(T) - \theta$ est une erreur systématique car l'estimateur T varie autour de son espérance mathématique $E(T)$ et non autour de la valeur θ du paramètre sauf si $E(T) = \theta$.

La quantité $E(T) - \theta$ est le *biais* de l'estimateur.

Un estimateur est *sans biais* si $E(T) = \theta$.

Un estimateur est *biaisé* si $E(T) \neq \theta$.

Un estimateur est *asymptotiquement sans biais* si $E(T) \rightarrow \theta$, quand la taille n de l'échantillon tend vers l'infini.

Un estimateur *biaisé* donne des estimations qui peuvent s'écarter systématiquement de la valeur à estimer ; il est donc moins satisfaisant qu'un estimateur sans biais (figure 13.1).

Cependant, l'absence de biais n'est pas une garantie absolue de « bon estimateur ». Il faut aussi tenir compte de sa variance (figure 13.2).

Exemple 13.2 Estimateur de l'espérance mathématique

La statistique $\bar{X}$ déjà étudiée (exemple 13.1) est un *estimateur sans biais* pour l'espérance mathématique $E(X)$. En effet, $E(\bar{X}) = E(X)$.

Exemple 13.3 Estimateur de la variance

La statistique $S^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2$ est un estimateur *biaisé* pour la variance :

$$E(S^2) = \frac{n-1}{n} \sigma^2 = \left(1 - \frac{1}{n}\right) \sigma^2, \text{ le biais est égal à } \frac{\sigma^2}{n}.$$

En revanche, la statistique S^{*2} définie par $S^{*2} = \frac{n}{n-1} S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$ est un estimateur *sans biais* pour la variance. En effet :

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

Exemple 13.4 Comparaison d'estimateurs de la variance

On suppose connue l'espérance mathématique $E(X) = m$ de la loi de probabilité de la variable aléatoire X et on cherche alors le meilleur estimateur de la variance.

Comparons les deux estimateurs sans biais T et S^{*2} :

$$T = \frac{1}{n} \sum_{i=1}^n (X_i - m)^2 \quad S^{*2} = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$

Un calcul rapide donne (μ_4 est le moment centré d'ordre 4) :

$$\text{Var}(T) = \frac{1}{n} \left(\mu_4 - \sigma^4 \right) \quad \text{Var}(S^{*2}) = \frac{1}{n} \left(\mu_4 - \frac{n-3}{n-1} \sigma^4 \right)$$

D'où : $\text{Var}(T) < \text{Var}(S^{*2})$.

Conclusion : si l'espérance mathématique m est connue, l'estimateur T

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

est un estimateur de la variance « meilleur » que la statistique S^{*2} .

13.2.3 Précision d'un estimateur

La précision d'un estimateur est mesurée par l'erreur quadratique moyenne :

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

En écrivant comme précédemment $T - \theta = T - E(T) + E(T) - \theta$ et en remarquant, après développement, que le terme $E \{ [T - E(T)] [E(T) - \theta] \}$ est nul ($E(T) - \theta$ est une constante et $T - E(T)$, une variable centrée), on obtient :

$$E \left[(T - \theta)^2 \right] = \text{Var}(T) + [E(T) - \theta]^2$$

Pour rendre l'erreur quadratique moyenne la plus petite possible, il faut que :

- $E(T) = \theta$, donc choisir un estimateur sans biais,
- $\text{Var}(T)$ soit petite.

Parmi les estimateurs sans biais, on choisira donc celui qui a la variance la plus petite, cette propriété traduit l'*efficacité* de l'estimateur.

Ainsi (exemple 13.4), si l'espérance mathématique m est connue, on choisira comme estimateur de la variance σ^2 , la statistique T et non la statistique S^{*2} .

13.2.4 Estimateur absolument correct

Un estimateur est *convergent* si sa distribution tend à se concentrer autour de la valeur inconnue du paramètre.

Un estimateur sans biais, dont la variance tend vers 0 quand n tend vers l'infini, est donc convergent. Il est *absolument correct*. Mais il *n'est pas nécessairement unique* comme le montrent les exemples ci-dessous.

Exemple 13.5 Comparaison d'estimateurs de l'espérance mathématique

– Comme estimateur de l'espérance mathématique m , on peut choisir la statistique T_1 :

$$T_1 = \frac{1}{n} \sum_{i=1}^n X_i \quad E(T_1) = m \quad \text{Var}(T_1) = \frac{\sigma^2}{n}$$

T_1 est un estimateur sans biais dont la variance tend vers 0 quand n tend vers l'infini, il est donc convergent.

– On peut aussi choisir la statistique T_2 , moyenne arithmétique des observations de rang impair. On obtient si on suppose n pair :

$$T_2 = \frac{2}{n} \sum_{i=0}^{p-1} X_{2i+1} \quad n = 2p \quad E(T_2) = m \quad \text{Var}(T_2) = \frac{2\sigma^2}{n}$$

L'estimateur T_2 est sans biais et sa variance tend vers 0 quand n tend vers l'infini, il est donc convergent.

Conclusion : les deux estimateurs T_1 et T_2 ont les mêmes propriétés. Il est évident cependant que T_1 est « meilleur » que T_2 . En effet :

- il tient compte de toute l'information apportée par l'échantillon,
- sa variance est la plus petite : $\text{Var}(T_1) < \text{Var}(T_2)$.

Exemple 13.6 Estimation du paramètre p d'une loi binomiale

On veut estimer la proportion p d'électeurs qui voteront pour le candidat A lors des élections municipales. On interroge un échantillon représentatif de taille n de l'ensemble des électeurs et soit k_n le nombre de réponses favorables ou f_n la fréquence des réponses.

$$E(k_n) = np \quad \text{Var}(k_n) = np(1-p)$$

$$E(f_n) = p \quad \text{Var}(f_n) = \frac{p(1-p)}{n}$$

La fréquence f_n est un estimateur sans biais du paramètre p . De plus, sa variance tend vers 0 quand n tend vers l'infini, il est donc convergent.

Résumé : estimateurs convergents et sans biais

- d'une moyenne $E(X)$: la statistique $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$
- d'une variance : la statistique $S^{*2} = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$

Attention

S^* n'est pas un estimateur sans biais de l'écart-type σ , en effet :

$$E(\sqrt{S^{*2}}) \neq \sqrt{E(S^{*2})}$$

- d'une proportion p : la fréquence $f_n = \frac{k_n}{n}$
(k_n nombre de réalisations de l'événement étudié au cours de n épreuves).

13.2.5 Recherche du meilleur estimateur

La recherche du meilleur estimateur d'un paramètre est un problème difficile à résoudre. En effet :

- la précision d'un estimateur T dépend de sa variance, c'est-à-dire de la loi de T qui dépend elle-même de la loi de la variable aléatoire X . Il faut donc connaître la forme de cette loi ;
- une statistique est un résumé apporté par un échantillon, il est donc très important de ne pas perdre d'information.

En tenant compte de ces deux impératifs, on peut aborder la recherche du meilleur estimateur suivant deux méthodes :

- soit en recherchant des *statistiques exhaustives* qui conduisent à des estimateurs sans biais, de variance minimale,
- soit en étudiant la *quantité d'information de Fisher* qui apporte des indications sur la précision d'un estimateur.

13.3 Estimateur sans biais de variance minimale

Les quatre résultats suivants résument les propriétés des estimateurs sans biais de variance minimale. Ils sont donnés sans démonstration.

■ Unicité

S'il existe un estimateur sans biais de variance minimale, il est unique presque sûrement (p.s.).

■ Théorème de Rao-Blackwell

Soient T un estimateur sans biais du paramètre θ et U une statistique exhaustive pour ce paramètre.

Alors $T^* = E(T/U)$ est un estimateur sans biais de θ au moins aussi bon que T .

■ Statistique exhaustive et estimateur de variance minimale

S'il existe une statistique exhaustive U , alors l'estimateur sans biais de variance minimale du paramètre θ ne dépend que de la statistique U .

Il peut exister plusieurs estimateurs sans biais, fonction d'une statistique exhaustive U . Pour obtenir l'unicité, il faut introduire la notion de statistique complète.

Une *statistique* U est dite *complète* pour une famille de lois $f(x; \theta)$ si :

$E[h(U)] = 0 \forall \theta$ entraîne $h = 0$ presque sûrement, h étant une fonction réelle.

On en déduit le théorème suivant.

■ Théorème de Lehman-Scheffe

Soit T^* un estimateur sans biais du paramètre θ dépendant d'une statistique exhaustive complète U .

T^* est l'unique estimateur sans biais de variance minimale. En particulier, si on connaît un estimateur T sans biais, T^* est donné par $T^* = E(T/U)$.

En conclusion, le meilleur estimateur d'un paramètre est un estimateur sans biais dépendant d'une statistique exhaustive complète.

Exemple 13.7 Statistique complète

La statistique exhaustive des familles de lois exponentielles est complète.

■ Dégradation de l'information

On reprend les notations du chapitre 12, paragraphe 12.4.

$I_n(\theta)$ est l'information apportée par l'échantillon.

$I_T(\theta)$ est l'information apportée par la statistique.

$I_{n/T}(\theta)$ est l'information conditionnelle apportée par l'échantillon sachant la statistique.

L'ensemble de définition E_θ ne dépend pas de θ .

Les quantités d'information vérifient les propriétés suivantes :

$$I_n(\theta) = I_T(\theta) + I_{n/T}(\theta)$$

$$I_n(\theta) \geq I_T(\theta)$$

Si l'ensemble E_θ ne dépend pas de θ , l'information apportée par un échantillon est supérieure ou égale à l'information apportée par une statistique. L'égalité a lieu si et seulement si la statistique est exhaustive.

Toute l'information apportée par un échantillon, concernant un paramètre, est donc contenue dans une statistique exhaustive.

13.4 Précision intrinsèque d'un estimateur et inégalité de Cramer-Rao

Avec les mêmes notations et en supposant de plus les hypothèses suivantes vérifiées :

- la densité $f(x; \theta)$ est telle que la quantité d'information de Fisher $I_n(\theta)$ existe et est finie, donc en particulier $E(T)$ et $\text{Var}(T)$ existent,
- l'ensemble E_θ ne dépend pas de θ ,
- les dérivées par rapport à θ de $L(x; \theta)$ existent et sont intégrables dans $\mathbb{R}^n$, on a alors l'inégalité de Cramer-Rao :

$$\text{Var}(T) \geq \frac{1}{I_n(\theta)} \left[\frac{dE(T)}{d\theta} \right]^2$$

■ Cas particuliers

- T est un estimateur sans biais d'une fonction $h(\theta)$, c'est-à-dire $E(T) = h(\theta)$. L'inégalité de Cramer-Rao s'écrit :

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

La variance d'un estimateur sans biais est *minorée* par une quantité indépendante de cet estimateur.

- Si T est un estimateur sans biais de θ , $E(T) = \theta$, on obtient :

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

c'est-à-dire le même résultat :

La variance d'un estimateur sans biais est *minorée* par une quantité indépendante de cet estimateur, elle ne peut donc pas être inférieure à une certaine borne.

13.5 Méthode du maximum de vraisemblance (MV)

Une méthode pour obtenir un estimateur consiste à choisir comme estimateur, une fonction $\hat{\theta}(\underline{X})$ qui réalise un maximum strict de la vraisemblance de l'échantillon, c'est-à-dire telle que :

$$L\left[\underline{X} ; \hat{\theta}(\underline{X})\right] \geq L(\underline{X} ; \theta) \quad \forall \theta$$

Remarques

- $L(\underline{x} ; \theta)$ étant une densité de probabilité, cette méthode revient à supposer que l'événement qui s'est produit était le plus probable.
- Dans la pratique, on prend comme estimation du maximum de vraisemblance, la solution de l'équation de la vraisemblance :

$$\frac{d \text{Ln } L(\underline{X} ; \theta)}{d\theta} = 0$$

On démontre facilement le résultat suivant, *propriété d'invariance fonctionnelle* :

Si $\hat{\theta}$ est l'estimateur de θ par la méthode du maximum de vraisemblance, $f(\hat{\theta})$ est l'estimateur de $f(\theta)$ par la méthode du maximum de vraisemblance.

Exemple 13.7 Loi uniforme

La variable aléatoire X suit une loi uniforme sur $[0, \theta]$, donc de densité :

$$f(x; \theta) = \begin{cases} \frac{1}{\theta} & x \in [0, \theta] \quad 0 < \theta \\ 0 & \text{sinon} \end{cases}$$

La vraisemblance d'un échantillon de taille n est : $L(\underline{x}; \theta) = \frac{1}{\theta^n} \quad \underline{x} \in [0, \theta]^n$

Pour rendre la vraisemblance maximum, l'échantillon $\underline{x} = (x_1, \dots, x_n)$ étant fixé, il faut rendre θ minimum. D'après la définition de θ , l'estimateur du maximum de vraisemblance de ce paramètre est :

$$\hat{\theta}(\underline{X}) = \sup (X_i) \quad i \in [1, \dots, n]$$

Exemple 13.8 Loi normale

On veut estimer le paramètre m de la loi normale $N(m; \sigma)$, à partir d'un échantillon de taille n . La méthode du maximum de vraisemblance conduit à :

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

$$\frac{d \ln L(\underline{x}; m)}{d m} = \frac{n}{\sigma^2} (\bar{x} - m)$$

Quand l'échantillon est donné, $\hat{m} = \bar{x}$ réalise un maximum strict de la vraisemblance (la dérivée seconde de la vraisemblance est strictement négative).

Conclusion : l'estimateur de l'espérance m par la méthode du maximum de vraisemblance est la statistique $\bar{X}$.

Exemple 13.9

Une population est caractérisée par la densité de probabilité :

$$f(x; \theta) = \begin{cases} e^{-(x-\theta)} & x \geq \theta \\ 0 & x < \theta \end{cases}$$

Quelle est l'estimation du maximum de vraisemblance du paramètre θ à partir d'un échantillon de taille n ?

Vraisemblance de l'échantillon :

$$L(\underline{x}; \theta) = \begin{cases} 0 & \text{si } \theta > \text{Inf } x_i \\ \prod_{i=1}^n f(x_i; \theta) = \exp\left(n\theta - \sum x_i\right) & \text{si } \theta < \text{Inf } x_i \end{cases}$$

Considérée comme une fonction de θ , la vraisemblance est positive et croissante pour $\theta < \text{Inf } x_i$, puis elle est nulle. L'estimateur du maximum de vraisemblance du paramètre θ est donc $\hat{\theta} = \text{Inf } x_i$.

13.5.1 Maximum de vraisemblance et exhaustivité

Soit T une statistique exhaustive : $L(\underline{x}; \theta) = g(t; \theta) h(\underline{x})$.

Toute fonction de $\underline{x}$ solution de l'équation de la vraisemblance :

$$\frac{d \text{Ln } L(\underline{x}; \theta)}{d\theta} = 0$$

est solution de l'équation :

$$\frac{d \text{Ln } g(t; \theta)}{d\theta} = 0$$

Pour réaliser un maximum de $L(\underline{x}; \theta)$, il suffit de réaliser un maximum de $g(t; \theta)$.

Donc, toute estimation de θ par la méthode du maximum de vraisemblance est une fonction de T mais n'est pas nécessairement une statistique exhaustive.

En résumé, s'il existe une statistique exhaustive T , l'estimateur du maximum de vraisemblance en dépend.

S'il n'existe pas de statistique exhaustive, on démontre le résultat suivant :

Il existe une suite θ_n de racines de l'équation de la vraisemblance qui converge presque sûrement vers θ quand n tend vers l'infini.

De plus, il existe N tel que, pour $n > N$, θ_n réalise un maximum pour $L(\underline{x}, \theta)$.

■ Propriété asymptotique

La variable aléatoire

$$\frac{\theta_n - \theta}{\sqrt{\frac{1}{I_n(\theta)}}}$$

converge en loi vers la loi $N(0; 1)$, quand n tend vers l'infini.

13.5.2 Estimateurs obtenus par la méthode du maximum de vraisemblance

Tableau 13.1 – Estimateurs obtenus par la méthode du maximum de vraisemblance.

Distribution	Paramètre à estimer	Estimateur
Loi uniforme sur $[0, a]$	a	$\text{Sup}(x_i)$
Loi binomiale k nombre de succès en n épreuves	p	$\hat{p} = \frac{k}{n}$
Loi de Poisson	λ	$\hat{\lambda} = \frac{1}{n} \sum_{i=1}^n x_i$
Loi normale	m	$\hat{m} = \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$
	σ^2	$\hat{\sigma}^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$

13.6 Extension au cas de plusieurs paramètres

Supposons que l'on ait à estimer un paramètre $\underline{\theta} \in \mathbb{R}^k$, c'est-à-dire k paramètres, $\theta_1 \dots \theta_k$.

La *matrice de l'information* est la matrice I_k symétrique définie positive qui a pour terme général :

$$I_{ij} = \text{Cov} \left[\frac{d \ln f(x; \underline{\theta})}{d\theta_i}, \frac{d \ln f(x; \underline{\theta})}{d\theta_j} \right]$$

Un *système exhaustif* est un système de s statistiques, (T_i) , fonctionnellement indépendantes, telles que :

$$L(\underline{x}; \underline{\theta}) = g(t_1, \dots, t_s; \underline{\theta}) h(\underline{x})$$

La notion de *dégradation de l'information* se généralise de la façon suivante. La matrice :

$$I_k(\underline{\theta}) - I_{T_1 \dots T_s}(\underline{\theta})$$

est une matrice définie positive.

Le *théorème de Darmais* s'énonce comme suit :

Une condition nécessaire et suffisante pour qu'un échantillon de taille n admette un résumé exhaustif est que :

$$\ln f(\underline{x} ; \underline{\theta}) = \sum_{i=1}^n a_i(\underline{x}) \alpha_i(\underline{\theta}) + b(\underline{x}) + \beta(\underline{\theta})$$

En particulier, le système de statistiques :

$$T_i = \sum_{j=1}^n a_j(x_j)$$

est un système exhaustif pour le paramètre $\underline{\theta}$.

La *méthode du maximum de vraisemblance* consiste à choisir comme estimateur de $\underline{\theta}$ une fonction appartenant à $\mathbb{R}^k$ réalisant un maximum strict de la vraisemblance :

$$L(\underline{x} ; \hat{\underline{\theta}}(\underline{x})) \geq L(\underline{x} ; \underline{\theta}) \quad \forall \underline{\theta}$$

On est donc amené à résoudre le système des k équations simultanées ou *système d'équations de la vraisemblance* :

$$\frac{d \ln L(\underline{x}, \theta_i)}{d \theta} = 0 \quad i \in [1, k]$$

Les propriétés de convergence, d'invariance fonctionnelle ainsi que les propriétés asymptotiques se généralisent sans difficulté.

Exemple 13.10 Loi de Weibull

La variable aléatoire X suit la loi de Weibull ; la densité de cette loi dépend de deux paramètres η et β .

– Densité : $f(x ; \beta ; \eta) = \frac{\beta}{\eta} \left(\frac{x}{\eta}\right)^{\beta-1} \exp\left(-\left(\frac{x}{\eta}\right)^{\beta}\right)$

– Vraisemblance d'un échantillon de taille n :

$$\begin{aligned} L(x_1, \dots, x_n ; \beta ; \eta) &= \prod_i f(x_i ; \beta ; \eta) \\ &= \frac{\beta^n}{\eta^n} \left(\prod_i x_i^{\beta-1}\right) \exp\left(-\sum_i \left(\frac{x_i}{\eta}\right)^{\beta}\right) \end{aligned}$$

$$\begin{aligned} \text{Ln } L(x_1, \dots, x_n ; \beta ; \eta) &= n \text{Ln } \beta - n\beta \text{Ln } \eta + (\beta - 1) \sum_i \text{Ln } x_i \\ &\quad - \sum_i \left(\frac{x_i}{\eta} \right)^\beta \end{aligned}$$

Les deux paramètres à estimer sont solution du système suivant :

$$\begin{aligned} \frac{d}{d\beta} \text{Ln } L(x_1, \dots, x_n ; \beta ; \eta) &= \frac{n}{\beta} - n \text{Ln } \eta + \sum_i \text{Ln } x_i - \sum_i \left(\frac{x_i}{\eta} \right)^\beta \text{Ln } \frac{x_i}{\eta} \\ &= 0 \end{aligned}$$

$$\frac{d}{d\eta} \text{Ln } L(x_1, \dots, x_n ; \beta ; \eta) = -\frac{n\beta}{\eta} + \frac{\beta}{\eta^{\beta+1}} \sum_i x_i^\beta = 0$$

La deuxième équation donne :

$$\hat{\eta} = \left(\frac{1}{n} \sum_i x_i^\beta \right)^{1/\beta}$$

En substituant cette valeur dans la première équation, on obtient :

$$\hat{\beta} = \left(\frac{1}{\sum_i x_i^\beta} \sum_i x_i^\beta \text{Ln } x_i - \frac{1}{n} \text{Ln } x_i \right)^{-1}$$

Pour trouver la solution de cette équation, on peut utiliser des méthodes numériques (méthodes des approximations successives, par exemple).

14 • ESTIMATION PAR INTERVALLE DE CONFIANCE

C

STATISTIQUE INFÉRENTIELLE

L'estimation ponctuelle d'un paramètre θ donne une valeur numérique unique à ce paramètre, mais n'apporte aucune information sur la précision des résultats, c'est-à-dire qu'elle ne tient pas compte des erreurs dues aux fluctuations d'échantillonnage, par exemple.

Pour évaluer la confiance que l'on peut avoir en une estimation, il est nécessaire de lui associer un intervalle qui contient, avec une certaine probabilité, la vraie valeur du paramètre, c'est l'*estimation par intervalle de confiance*.

14.1 Définition d'un intervalle de confiance

L'estimation par intervalle de confiance d'un paramètre θ consiste donc à associer à un échantillon, un intervalle aléatoire I , choisi de telle façon que la probabilité pour qu'il contienne la valeur inconnue du paramètre soit égale à un nombre fixé à l'avance, aussi grand que l'on veut. On écrit :

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

$(1 - \alpha)$ est la probabilité associée à l'intervalle d'encadrer la vraie valeur du paramètre, c'est le *seuil de confiance* ou la *quasi-certitude*.

14.1.1 Intervalle de probabilité : rappel

Soit X une variable aléatoire, f la densité de sa loi de probabilité. Étant donnée une probabilité α , on choisit deux nombres α_1 et α_2 ayant pour somme α

$(\alpha_1 + \alpha_2 = \alpha)$ et on définit deux valeurs x_1 et x_2 de la variable X telles que :

$$\Pr(X < x_1) = \alpha_1 \quad \Pr(X > x_2) = \alpha_2$$

L'intervalle $I = [x_1, x_2]$ a une probabilité égale à $(1 - \alpha)$ de contenir une valeur observée de la variable X . En négligeant la probabilité α , on résume la distribution de la variable X en ne considérant que les valeurs appartenant à l'intervalle I , on définit *un intervalle de probabilité au seuil* $(1 - \alpha)$ pour la variable X , la valeur α est le *seuil critique*.

Pour construire un intervalle de probabilité, deux questions se posent :

- quel est le seuil de probabilité α susceptible d'être valablement considéré comme négligeable ?
- pour une loi de probabilité et pour un seuil α donnés, il existe une infinité d'intervalles $[x_1, x_2]$ qui dépendent du choix de α_1 et α_2 . Comment choisir ces deux valeurs ?

Les réponses à ces deux questions dépendent des problèmes traités.

Exemple 14.1 Intervalle de probabilité

On suppose qu'un dosage sanguin est une variable aléatoire X suivant la loi normale $N(100 ; 20)$. On considère comme *normales* les valeurs de X comprises entre deux limites a et b telles que $\Pr(a < X < b) = 0,95$, les autres valeurs étant considérées comme *pathologiques*.

La donnée du seuil critique $\alpha = 0,05$ sans précision supplémentaire ne permet pas de calculer les limites a et b ; une infinité d'intervalles de probabilité répondent à la question. En revanche, la probabilité de mesurer une valeur *pathologique* est égale à $\alpha = 0,05$, quel que soit l'intervalle.

On suppose maintenant que les valeurs a et b sont symétriques par rapport à la moyenne $m = 100$. En introduisant la variable aléatoire centrée réduite

$$U = \frac{X - 100}{20}, \text{ on sait que :}$$

$$\Pr(-1,96 < U < 1,96) = 0,95$$

D'où : $a = 100 - 1,96 \times 20 = 60,80$ et $b = 100 + 1,96 \times 20 = 139,20$

et l'intervalle de probabilité correspondant est $\Pr(60,80 < X < 139,20) = 0,95$.

Cependant, les faibles valeurs de X ne présentant pas un caractère pathologique, on garde seulement la valeur supérieure $b = 139,80$.

La probabilité d'observer une valeur pathologique devient égale à 0,025 et un intervalle de probabilité unilatéral au seuil de confiance 0,975 est alors :

$$\Pr(X < 139,20) = 0,975$$

14.1.2 Construction d'un intervalle de confiance

X est une variable aléatoire dont la densité, $f(x; \theta)$, dépend du paramètre θ et $\underline{X} = (X_1, \dots, X_n)$ est un échantillon de taille n de cette variable.

Soit $T = \varphi(\underline{X})$ un estimateur du paramètre θ et $g(t; \theta)$ la loi de probabilité de cet estimateur.

Étant donnée une probabilité α , on peut, à partir de cette loi et si *on suppose le paramètre θ connu*, construire un intervalle de probabilité pour la variable aléatoire T :

$$\Pr(\theta - h_1 < T < \theta + h_2) = 1 - \alpha \quad (14.1)$$

Les bornes de l'intervalle sont définies par (avec $\alpha_1 + \alpha_2 = \alpha$) :

$$\Pr(T < \theta - h_1) = \int_{-\infty}^{\theta - h_1} g(t; \theta) dt = \alpha_1$$

$$\Pr(T > \theta + h_2) = \int_{\theta + h_2}^{+\infty} g(t; \theta) dt = \alpha_2$$

Si l'égalité (14.1) est vérifiée, l'égalité suivante :

$$\Pr(t - h_2 < \theta < t + h_1) = 1 - \alpha \quad (14.2)$$

où t est la valeur de la statistique T donnée par l'échantillon, est également vérifiée.

L'intervalle $I = [t - h_2, t + h_1]$ a une probabilité égale à $(1 - \alpha)$ de contenir le paramètre θ , c'est un *intervalle de confiance au seuil de confiance* ou *niveau de confiance* $(1 - \alpha)$.

14.1.3 Propriétés des intervalles de confiance

- Un intervalle de confiance est un *intervalle aléatoire* car les bornes de cet intervalle sont des variables aléatoires, fonctions des observations.

- Le seuil α étant donné, il faut définir les nombres α_1 et α_2 . Leur choix dépend des problèmes à traiter, des risques encourus à négliger les petites ou les grandes valeurs du paramètre. Si on choisit $\alpha_1 = \alpha_2 = \alpha/2$, on construit un *intervalle de confiance bilatéral* à risques symétriques. On peut construire des *intervalles de confiance unilatéraux*, soit avec $\alpha_1 = 0$, soit avec $\alpha_2 = 0$.
- Le seuil α , les nombres α_1 et α_2 et la taille n de l'échantillon étant fixés, on peut construire un intervalle de confiance associé à chaque échantillon. Cependant, parmi ces intervalles, une proportion égale à α % ne contiendra pas la valeur exacte du paramètre. Ce seuil α représente donc le risque que l'intervalle de confiance ne contienne pas la vraie valeur du paramètre. La situation la plus favorable correspond à choisir un risque α petit, associé à un intervalle de faible étendue.
- On peut diminuer la valeur du seuil α , et même à la limite, choisir $\alpha = 0$ pour avoir la certitude absolue. Dans ce cas, l'intervalle de confiance s'étend à tout le domaine de définition du paramètre, $]-\infty, +\infty[$ pour l'espérance mathématique ou $[0, +\infty[$ pour l'écart-type, par exemple ! Donc :
diminuer la valeur de $\alpha \Rightarrow$ augmenter l'étendue de l'intervalle
- Dans la pratique, on donne à α une valeur acceptable, de l'ordre de 5 % puis, quand cela est possible, on augmente la taille de l'échantillon.
- La probabilité $(1 - \alpha)$ représente le niveau de confiance de l'intervalle ; ce niveau de confiance est associé à l'intervalle et non à la valeur inconnue du paramètre.
- Pour définir un intervalle de confiance, il faut connaître un estimateur ponctuel du paramètre ainsi que sa loi de distribution.

14.2 Exemples d'intervalles de confiance

14.2.1 Intervalle de confiance pour les paramètres d'une loi normale

La variable aléatoire X suit une loi normale $N(m; \sigma)$. Les paramètres à estimer sont la moyenne m et l'écart-type σ .

■ Estimation et intervalle de confiance de la moyenne

L'estimateur sans biais de la moyenne m est la statistique $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ qui suit la loi normale, $N(m ; \sigma/\sqrt{n})$.

Deux cas sont à distinguer, selon que l'écart-type est connu ou estimé.

□ Cas 1 : l'écart-type σ est connu

Étant donné un seuil α , on construit, pour la moyenne $\bar{X}$ de l'échantillon, un intervalle de probabilité :

$$\Pr \left(m - u_{\alpha/2} \frac{\sigma}{\sqrt{n}} < \bar{X} < m + u_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right) = 1 - \alpha$$

la valeur $u_{\alpha/2}$ étant lue sur la table de la loi normale réduite.

On en déduit l'intervalle de confiance pour la moyenne m :

$$\Pr \left(\bar{x} - u_{\alpha/2} \frac{\sigma}{\sqrt{n}} < m < \bar{x} + u_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right) = 1 - \alpha$$

où $\bar{x}$ est la moyenne arithmétique de l'échantillon.

Exemple 14.2 Intervalle de confiance pour la moyenne, σ connu

Après des essais antérieurs, on peut supposer que la résistance à l'éclatement d'un certain type de réservoirs est une variable aléatoire suivant une loi normale de moyenne m inconnue et d'écart-type égal à 4 kg/cm^2 . Des essais sur un échantillon de 9 réservoirs donnent une résistance moyenne à l'éclatement égale à 215 kg/cm^2 .

- Estimation ponctuelle de la moyenne donnée par l'échantillon : 215 kg/cm^2 .
- Loi suivie par la moyenne d'un échantillon de taille $n = 9$ (avec l'hypothèse admise sur la loi suivie par la résistance) : la loi normale $N(215 ; 4/3)$.
- Niveau de confiance : $1 - \alpha = 0,95$.

$$\text{– Intervalle de confiance : } \Pr \left(-1,96 < \frac{\bar{X} - m}{4/3} < 1,96 \right) = 0,95 \quad \bar{x} = 215$$

$$\begin{aligned} \Pr(215 - 4/3 \times 1,96 < m < 215 + 4/3 \times 1,96) \\ = \Pr(212,386 < m < 217,613) = 0,95 \end{aligned}$$

L'intervalle $[212,386, 217,613]$ a une probabilité égale à 0,95 de contenir la vraie valeur de la résistance à l'éclatement de ce type de réservoirs.

Remarque

Cet exemple montre que si la taille n de l'échantillon augmente, α et σ restant constants, l'étendue de l'intervalle diminue ; en revanche, si le seuil α diminue l'étendue de l'intervalle augmente.

□ Cas 2 : l'écart-type σ n'est pas connu

L'estimateur sans biais de la variance est la statistique

$$S^{*2} = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2 = \widehat{\sigma}^2$$

La variable aléatoire :

$$T(n-1) = \frac{\bar{X} - m}{S^* / \sqrt{n}} = \frac{\bar{X} - m}{S / \sqrt{n-1}}$$

suit une loi de Student à $(n-1)$ degrés de liberté.

Le seuil α étant donné, on lit sur la table 8 (loi de Student) le nombre $t_{1-\alpha/2}(n-1)$ tel que :

$$\Pr \left(\bar{x} - \frac{s^*}{\sqrt{n}} t_{1-\alpha/2}(n-1) < m < \bar{x} + \frac{s^*}{\sqrt{n}} t_{1-\alpha/2}(n-1) \right) = 1 - \alpha$$

D'où l'intervalle de confiance pour la moyenne m :

$$\Pr \left(-t_{1-\alpha/2}(n-1) < t(n-1) < t_{1-\alpha/2}(n-1) \right) = 1 - \alpha$$

que l'on peut écrire en considérant la statistique S^2 :

$$\Pr \left(\bar{x} - \frac{s}{\sqrt{n-1}} t_{1-\alpha/2}(n-1) < m < \bar{x} + \frac{s}{\sqrt{n-1}} t_{1-\alpha/2}(n-1) \right) = 1 - \alpha$$

Exemple 14.3 Intervalle de confiance pour la moyenne, σ estimé

Afin d'étudier le salaire journalier, en euros, des ouvriers d'un secteur d'activité, on procède à un tirage non exhaustif, d'un échantillon de taille $n = 16$. On a obtenu les résultats suivants :

41	40	45	50	41	41	49	43
45	52	40	48	50	49	47	46

On suppose que la loi suivie par la variable aléatoire « salaire journalier » est une loi normale de moyenne m et d'écart-type σ inconnus.

– Estimation ponctuelle de la moyenne, la moyenne arithmétique : 45,4375.

– Estimation ponctuelle de la variance :

$$S^2 = 15,2460 = (3,9046)^2 \quad (\text{estimateur biaisé})$$

$$S^{*2} = 16,2625 = (4,0326)^2 \quad (\text{estimateur non biaisé})$$

– Intervalle de confiance pour la moyenne, seuil de confiance 0,95 (intervalle bilatéral à risques symétriques).

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

D'où la suite des calculs en tenant compte des résultats donnés par l'échantillon :

$$\Pr(-2,131 < T(15) < 2,131) = 0,95 \quad (T_{0,975}(15) = 2,131)$$

$$\Pr\left(-2,131 < \frac{45,4375 - m}{3,9046 / \sqrt{15}} < 2,131\right) = 0,95$$

$$\Pr\left(45,4375 - 2,131 \times 3,9046 / \sqrt{15} < m < 45,4375 + 2,131 \times 3,9046 / \sqrt{15}\right) = 0,95$$

$$\Pr(43,2895 < m < 47,5859) = 0,95$$

L'intervalle $[43,2895, 47,5859]$ a une probabilité égale à 0,95 de contenir la vraie valeur du salaire moyen journalier des ouvriers de ce secteur d'activité.

C

STATISTIQUE INFÉRENTIELLE

■ Estimation et intervalle de confiance pour la variance

Comme précédemment, deux cas sont à distinguer, selon que la moyenne m est connue ou estimée.

□ Cas 1 : la moyenne m est connue

Le meilleur estimateur de la variance est la statistique : $T = \frac{1}{n} \sum_{i=1}^n (X_i - m)^2$.

La variable aléatoire $\frac{nT}{\sigma^2}$ suit une loi du chi-deux à n degrés de liberté. Un intervalle de probabilité pour la variable aléatoire $\chi^2(n)$ est (les bornes de l'intervalle sont lues sur la table 6 loi du chi-deux) :

$$\Pr\left(\chi_{\alpha/2}^2(n) < \chi^2(n) < \chi_{1-\alpha/2}^2(n)\right) = 1 - \alpha$$

On en déduit un intervalle de confiance bilatéral pour σ^2 à risques symétriques (t est la valeur de la statistique T donnée par l'échantillon) :

$$\Pr \left(\frac{nt}{\chi_{1-\alpha/2}^2(n)} < \sigma^2 < \frac{nt}{\chi_{\alpha/2}^2(n)} \right) = 1 - \alpha$$

Exemple 14.4 Intervalle de confiance pour la variance, m connue

Soit X une variable aléatoire suivant la loi normale $N(40 ; \sigma)$. Pour estimer la variance, on prélève un échantillon de taille $n = 25$ et on calcule la valeur de la statistique T (définie précédemment) pour cet échantillon. On obtient $t = 12$.

– Intervalle de confiance bilatéral, à risques symétriques, pour la variance.

– Niveau de confiance : $1 - \alpha = 0,95$.

– La statistique $\frac{nT}{\sigma^2}$ suit une loi du chi-deux à $n = 25$ degrés de liberté. On obtient successivement :

$$\begin{aligned} \Pr(\chi_{0,025}^2(25) < \chi^2(25) < \chi_{0,975}^2(25)) &= \Pr(13,120 < \chi^2(25) < 40,644) \\ &= 0,95 \\ \Pr\left(13,120 < \frac{25 \times 12}{\sigma^2} < 40,644\right) &= \Pr\left(\frac{25 \times 12}{40,644} < \sigma^2 < \frac{25 \times 12}{13,120}\right) \\ &= \Pr(7,381 < \sigma^2 < 22,866) = 0,95 \end{aligned}$$

L'intervalle $[7,381, 22,866]$ a une probabilité égale à 0,95 de contenir la vraie valeur de la variance de la loi considérée.

De même, l'intervalle $[2,716 ; 4,782]$ a une probabilité égale à 0,95 de contenir la vraie valeur de l'écart-type de la loi considérée.

□ Cas 2 : la moyenne m n'est pas connue

La statistique $\frac{nS^2}{\sigma^2} = \frac{1}{\sigma^2} \sum_{i=1}^n (X_i - \bar{X})^2$ est une variable chi-deux à $(n - 1)$ degrés de liberté. La démarche est la même que dans le cas 1.

$$\Pr(\chi_{\alpha/2}^2(n-1) < \chi^2(n-1) < \chi_{1-\alpha/2}^2(n-1)) = 1 - \alpha$$

Les bornes de l'intervalle sont lues sur la table 6 (loi du chi-deux).

D'où l'intervalle de confiance, bilatéral, à risques symétriques :

– pour la variance (s^2 étant la valeur de la statistique S^2 donnée par l'échantillon) :

$$\Pr \left(\frac{ns^2}{\chi_{1-\alpha/2}^2(n-1)} < \sigma^2 < \frac{ns^2}{\chi_{\alpha/2}^2(n-1)} \right) = 1 - \alpha$$

– pour l'écart-type :

$$\Pr \left(\sqrt{\frac{ns^2}{\chi_{1-\alpha/2}^2(n-1)}} < \sigma < \sqrt{\frac{ns^2}{\chi_{\alpha/2}^2(n-1)}} \right) = 1 - \alpha$$

Exemple 14.5 Intervalle de confiance pour la variance, m estimée

(Suite de l'exemple 14.3.)

La variable $\frac{nS^2}{\sigma^2}$ suit une loi du chi-deux à $16 - 1 = 15$ degrés de liberté.

D'où la suite des calculs :

$$\begin{aligned} \Pr(\chi_{0,025}^2(15) < \chi^2(15) < \chi_{0,975}^2(15)) &= \Pr(6,262 < \chi^2(15) < 27,488) \\ &= 0,95 \\ \Pr\left(6,262 < \frac{16 \times 15,246}{\sigma^2} < 27,488\right) &= \Pr\left(\frac{16 \times 15,246}{27,488} < \sigma^2 < \frac{16 \times 15,246}{6,262}\right) \\ &= \Pr(8,874 < \sigma^2 < 38,955) = 0,95 \end{aligned}$$

L'intervalle $[8,874, 38,955]$ a une probabilité égale à 0,95 de contenir la vraie valeur de la variance du salaire moyen horaire des ouvriers de ce secteur d'activité et l'intervalle $[2,98 ; 6,24]$ a la même propriété pour l'écart-type.

D'une façon analogue, on peut déterminer des intervalles de confiance unilatéraux (à droite ou à gauche) :

$$\begin{aligned} \Pr(\chi_{\alpha}^2(n-1) < \chi^2(n-1)) &= 1 - \alpha \\ \Pr\left(0 < \sigma^2 < \frac{ns^2}{\chi_{\alpha}^2(n-1)}\right) &= 1 - \alpha \quad \Pr\left(0 < \sigma < \sqrt{\frac{ns^2}{\chi_{\alpha}^2(n-1)}}\right) = 1 - \alpha \\ \Pr(0 < \chi^2(n-1) < \chi_{1-\alpha}^2(n-1)) &= 1 - \alpha \\ \Pr\left(\frac{ns^2}{\chi_{1-\alpha}^2(n-1)} < \sigma^2\right) &= 1 - \alpha \quad \Pr\left(\sqrt{\frac{ns^2}{\chi_{1-\alpha}^2(n-1)}} < \sigma\right) = 1 - \alpha \end{aligned}$$

Exemple 14.6 Intervalle de confiance unilatéral pour la variance

Reprenons l'exemple 14.5 et déterminons une valeur a telle que (seuil de confiance 0,95) :

$$\Pr(0 < \sigma^2 < a) = 0,95$$

$$\begin{aligned}\Pr(\chi_{0,05}^2(15) < \chi^2(15)) &= \Pr(7,26 < \chi^2(15)) \\ &= \Pr\left(7,26 < \frac{16 \times 15,264}{\sigma^2}\right) = 0,95\end{aligned}$$

$$\Pr(\sigma^2 < 33,64) = \Pr(\sigma < 5,80) = 0,95$$

Exemple 14.7 Détermination de la taille d'un échantillon pour estimer un paramètre avec un niveau de confiance donné

On suppose que la durée de vie d'ampoules électriques suit une loi normale d'écart-type 100 heures. Quelle est la taille minimale de l'échantillon à prélever pour que l'intervalle de confiance, à 95 %, de la durée de vie moyenne de ces ampoules ait une longueur inférieure à 20 heures ?

Les hypothèses faites entraînent que la largeur de l'intervalle de confiance (bilatéral, à risques symétriques) pour la moyenne est égale à : $2 \times 1,96 \times \frac{\sigma}{\sqrt{n}}$.

D'où l'équation : $2 \times 1,96 \times \frac{100}{\sqrt{n}} = 20$ et $n = 385$.

Remarque

Si la taille n de l'échantillon est supérieure à 30, la variable aléatoire :

$$\sqrt{2\chi^2(n)} - \sqrt{2n-1}$$

suit une loi normale centrée réduite. On utilisera donc la table de la loi normale pour trouver les bornes de l'intervalle.

14.2.2 Intervalle de confiance pour la moyenne d'une loi quelconque

Quelle que soit la taille de l'échantillon, on prendra pour estimateurs sans biais de la moyenne, la statistique $\bar{X}$, et de la variance, la statistique S^{*2} .

Si la taille de l'échantillon est grande, en pratique $n > 30$, grâce au théorème central limite, on utilise les résultats précédents (paragraphe 14.2) pour calculer l'intervalle de confiance pour la moyenne.

Si la série prélevée est faible, il faut utiliser les lois suivies par les paramètres étudiés.

Exemple 14.8

On considère un échantillon de 40 paquets de biscuits provenant d'une production de 2 000 unités. Le poids moyen obtenu pour cet échantillon est égal à 336 g et l'écart-type empirique, c'est-à-dire la quantité s , est égal à 0,86 g.

Quelle est l'estimation, par intervalle de confiance, du poids moyen de ces paquets de biscuits, pour l'ensemble de la fabrication, avec les seuils de confiance 0,90, 0,98 et 0,99 ?

La distribution du poids de ces paquets est inconnue ainsi que la variance de la population. Cependant, pour un grand échantillon ($n = 40 > 30$), l'intervalle de confiance est de la forme :

$$\bar{x} - h \frac{s}{\sqrt{n-1}} < m < \bar{x} + h \frac{s}{\sqrt{n-1}} \quad \text{avec } \bar{x} = 336 \quad s = 0,86 \quad n = 40$$

Le fractile h est lu sur la table de Student, degré de liberté $n - 1 = 39$, il dépend du seuil choisi.

Les résultats sont regroupés dans le tableau suivant.

Tableau 14.1 – Limites de l'intervalle de confiance.

Niveaux de confiance	h	Limites inférieures	Limites supérieures
0,90	1,685	$336 - 0,232 = 335,768$	$336 + 0,232 = 336,232$
0,98	2,429	$336 - 0,334 = 335,666$	$336 + 0,334 = 336,334$
0,99	2,708	$336 - 0,373 = 335,627$	$336 + 0,373 = 336,373$

Ainsi, l'intervalle de confiance, qui a 98 chances sur 100 de contenir la vraie valeur du poids moyen de cette fabrication de biscuits, est : $[335,666, 336,334]$.

14.2.3 Intervalle de confiance de la différence des moyennes de deux lois normales indépendantes

Soient X_1 et X_2 deux variables aléatoires indépendantes, suivant les lois normales $N(m_1; \sigma_1)$ et $N(m_2; \sigma_2)$.

Le paramètre à estimer est la différence des moyennes $D = m_1 - m_2$.

On considère deux échantillons indépendants de taille n_1 pour la variable X_1 et de taille n_2 pour la variable X_2 .

La variable aléatoire $\bar{D} = \bar{X}_1 - \bar{X}_2$ suit une loi normale d'espérance $m = m_1 - m_2$ et d'écart-type :

$$\sigma_{\bar{D}} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Les écarts-types σ_1 et σ_2 étant inconnus, il faut chercher une estimation qui prend en compte toutes les données du problème. En utilisant la relation :

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

et les propriétés de la loi du chi-deux, on montre que la statistique :

$$\frac{1}{n_1 + n_2 - 2} \left[\sum_1 (X_i - \bar{X})^2 + \sum_2 (X_i - \bar{X})^2 \right]$$

est un estimateur sans biais de la variance commune des deux distributions prenant en compte toutes les observations (l'indice 1 indique que la sommation est faite sur le premier échantillon et l'indice 2 sur le deuxième).

Pour obtenir ce résultat, on a supposé que *les écarts-types inconnus σ_1 et σ_2 étaient égaux*. Un test de Fisher (paragraphe 14.2.4) permet de vérifier ce résultat.

Toutes les propriétés démontrées entraînent que la variable aléatoire :

$$\frac{(\bar{X}_1 - \bar{X}_2) - (m_1 - m_2)}{\sqrt{\sum_1 (X_i - \bar{X})^2 + \sum_2 (X_i - \bar{X})^2}} \times \frac{\sqrt{n_1 + n_2 - 2}}{\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

suit une loi de Student à $(n_1 + n_2 - 2)$ degrés de liberté.

Pour déterminer un intervalle de confiance pour la différence des moyennes de deux lois normales ou pour tester l'égalité de ces deux moyennes, il suffit d'utiliser la table statistique de Student.

Remarque

En « mélangeant » les échantillons, on a perdu un degré de liberté.

Exemple 14.9

Des tubes sont construits par deux procédés de fabrications A et B. On dispose de deux échantillons de tubes dont on mesure les diamètres (en mm) :

– Procédé A : échantillon 1 de taille $n_1 = 5$. Résultats : 63,12 ; 63,57 ; 62,81 ; 64,32 ; 63,76.

– Procédé B : échantillon 2 de taille $n_2 = 4$. Résultats : 62,51 ; 63,24 ; 62,31 ; 62,21.

On suppose que les diamètres des tubes de chaque fabrication sont distribués suivant des lois normales de variances inconnues mais égales.

– Intervalle de confiance bilatéral, à risques symétriques, pour la différence $m_1 - m_2$ des moyennes des deux lois (seuil de confiance 0,95) :

Échantillon 1 :

$$\bar{x}_1 = 63,516 \quad \sum_1 (x_i - \bar{x})^2 = 1,3638$$

Échantillon 2 :

$$\bar{x}_2 = 62,5675 \quad \sum_2 (x_i - \bar{x})^2 = 0,6498$$

– La variable $\frac{(\bar{x}_1 - \bar{x}_2) - (m_1 - m_2)}{\sqrt{\sum_1 (x_i - \bar{x})^2 + \sum_2 (x_i - \bar{x})^2}} \times \frac{\sqrt{n_1 + n_2 - 2}}{\sqrt{1/n_1 + 1/n_2}}$ suit une loi de Student à $n_1 + n_2 - 2 = 7$ degrés de liberté.

– Avec les valeurs numériques données, on obtient :

$$\frac{0,9485 - (m_1 - m_2)}{\sqrt{1,3638 + 0,6498}} \times \frac{\sqrt{5 + 4 - 2}}{\sqrt{1/5 + 1/4}} = \frac{0,9485 - (m_1 - m_2)}{0,3598}$$

$$\Pr(-2,365 < t(7) < 2,365) = 0,95$$

D'où l'intervalle de confiance :

$$\Pr(0,9485 - 2,365 \times 0,3598 < (m_1 - m_2) < 0,9485 + 2,365 \times 0,3598) = 0,95$$

L'intervalle $[0,10, 1,80]$ a une probabilité égale à 0,95 de contenir la vraie valeur de la différence des moyennes des deux lois normales.

14.2.4 Intervalle de confiance pour le rapport des variances de deux lois normales

Soient X_1 et X_2 deux variables aléatoires indépendantes suivant les lois normales $N(m_1; \sigma_1)$ et $N(m_2; \sigma_2)$.

On a démontré (théorie de l'échantillonnage, chapitre 10, paragraphe 10.7.4) le résultat suivant :

$$\frac{n_1 S_1^2}{(n_1 - 1) \sigma_1^2} \times \frac{(n_2 - 1) \sigma_2^2}{n_2 S_2^2} = F(n_1 - 1; n_2 - 1)$$

ou, en introduisant la statistique S^{*2} :

$$\frac{S_1^{*2}}{\sigma_1^2} \times \frac{\sigma_2^2}{S_2^{*2}} = F(n_1 - 1; n_2 - 1)$$

Ce résultat permet de déterminer un intervalle de confiance pour le rapport de deux variances ou pour tester l'égalité de deux variances. On utilise les méthodes exposées dans les paragraphes précédents. On obtient par exemple :

$$\Pr \left(\frac{s_1^{*2}}{s_2^{*2}} \times \frac{1}{F_{\alpha/2}(n_1 - 1; n_2 - 2)} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{s_1^{*2}}{s_2^{*2}} \times \frac{1}{F_{1-\alpha/2}(n_1 - 1; n_2 - 2)} \right) = 1 - \alpha$$

où s^* désigne la valeur de la statistique S^* donnée par l'échantillon considéré ; les valeurs de $F_{\alpha/2}(n_1 - 1, n_2 - 1)$ et de $F_{1-\alpha/2}(n_1 - 1, n_2 - 1)$ sont lues sur les tables 7.1, 7.2, 7.3 ou 7.4 selon la valeur du seuil α . On en déduit l'égalité ou non des variances.

Exemple 14.10

Une firme d'expertises en contrôle de matériaux a demandé à un laboratoire d'effectuer une vérification sur la résistance à la compression de tubes fabriqués par deux usines U_1 et U_2 .

On supposera que la résistance à la compression des tubes provenant de chaque usine peut être considérée comme la réalisation de variables aléatoires suivant des lois normales.

Les résultats à la compression en kg/cm^2 sont résumés dans le tableau 14.2 :

Tableau 14.2 – Résistance à la compression des tubes.

	Usine U_1	Usine U_2
Nombre de cylindres	$n_1 = 25$	$n_2 = 23$
Résistance moyenne	$\bar{x}_1 = 90,60$	$\bar{x}_2 = 94,40$
Variance empirique	$S_1^2 = 62,80$	$S_2^2 = 55,70$

Déterminer un intervalle de confiance bilatéral, à risques symétriques, pour le rapport σ_1^2 / σ_2^2 des deux variances (seuil de confiance 90 %).

Peut-on considérer comme vraisemblable au seuil $\alpha = 10\%$, l'hypothèse selon laquelle la variance de la résistance à la compression des tubes provenant de chaque usine est identique ?

– Intervalle de confiance pour le rapport σ_1^2 / σ_2^2 :

L'hypothèse de normalité entraîne :

$$\frac{n_1 s_1^2}{(n_1 - 1) \sigma_1^2} \times \frac{(n_2 - 1) \sigma_2^2}{n_2 s_2^2} = F(n_1 - 1 ; n_2 - 1) = F(24 ; 22)$$

$$\Pr\left(0,50 < \frac{n_1 s_1^2}{(n_1 - 1) \sigma_1^2} \times \frac{(n_2 - 1) \sigma_2^2}{n_2 s_2^2} < 2,03\right) = 0,90 \quad (\text{Tables statistiques})$$

$$\Pr\left(\frac{1,12}{2,03} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{1,12}{0,50}\right) = \Pr\left(0,55 < \frac{\sigma_1^2}{\sigma_2^2} < 2,24\right) = 0,90$$

L'intervalle $[0,55, 2,24]$ a 90 chances sur 100 de contenir la vraie valeur du rapport σ_1^2 / σ_2^2 .

Cet intervalle contient la valeur 1, on ne peut pas donc rejeter l'hypothèse d'égalité des variances, au seuil $\alpha = 10\%$.

– Deuxième démonstration de ce résultat : si les variances des deux populations sont égales, la variable $\frac{n_1 s_1^2}{(n_1 - 1)} \times \frac{(n_2 - 1)}{n_2 s_2^2}$ est la réalisation d'une variable de Fisher-Snedecor $F(24 ; 22)$. Cette variable prend la valeur 1,1233 pour les échantillons étudiés. Au seuil 10 %, on trouve comme intervalle de probabilité : $\Pr(0,50 < F(24 ; 22) < 2,03) = 0,90$

La valeur 1,1233 appartenant à cette intervalle, on retrouve la même conclusion que précédemment.

14.2.5 Intervalle de confiance pour une proportion

Ce problème apparaît dans de nombreuses situations quand on veut estimer par exemple (liste non exhaustive) :

- la proportion de pièces défectueuses dans une fabrication donnée,
- la proportion d'électeurs qui voteront pour un candidat déterminé,
- la proportion de ménagères qui achèteront une nouvelle marque de lessive.

Soit une population où une proportion p des individus possède un certain caractère, *cette population est supposée infinie (ou finie si le tirage s'effectue avec remise)*. Le problème consiste à déterminer un intervalle de confiance pour la proportion p à partir des résultats apportés par un échantillon de taille n .

À cet échantillon, on associe la variable aléatoire X qui « compte » le nombre de succès au cours de n essais indépendants, cette variable suit la loi binomiale $B(n; p)$. Le paramètre à estimer est la probabilité p de succès au cours d'une épreuve.

Un *estimateur sans biais* du paramètre p est la *fréquence* $F_n = K/n$ de succès à l'issue de n épreuves, K étant le nombre de succès obtenus au cours de ces n épreuves :

$$E(F_n) = p \quad \text{Var}(F_n) = \sqrt{\frac{p(1-p)}{n}}$$

Selon les valeurs de n et de p , cette loi admet différentes lois limites qui sont utilisées pour déterminer un intervalle de confiance. Dans la pratique, on peut :

- utiliser les tables statistiques qui donnent les limites inférieures et supérieures d'un intervalle de confiance calculées pour différents seuils et différentes valeurs de n et k ,
- utiliser et justifier l'approximation normale.

Intervalle de confiance d'une proportion calculée avec l'approximation normale : si $n \geq 50$, $np > 5$ et $n(1-p) > 5$, la loi de la variable aléatoire F_n (fréquence des succès) peut être approchée par la loi normale :

$$N\left(p; \sqrt{\frac{p(1-p)}{n}}\right)$$

La table de la loi normale centrée réduite permet de construire un intervalle de confiance pour la probabilité p .

Un intervalle bilatéral à risques symétriques (f_n est la fréquence observée sur l'échantillon) est donné par :

$$\Pr \left(-u_{1-\alpha/2} \sqrt{\frac{p(1-p)}{n}} < f_n - p < u_{1-\alpha/2} \sqrt{\frac{p(1-p)}{n}} \right) = 1 - \alpha$$

Les bornes de l'intervalle de confiance :

$$f_n \pm u_{1-\alpha/2} \sqrt{\frac{p(1-p)}{n}}$$

dépendent du paramètre à estimer dont on ne connaît pas la valeur. Pour résoudre ce problème, on peut :

- soit remplacer p par son estimation f_n ,
- soit remplacer p par $1/2$, car le produit de deux nombres p et $(1-p)$, dont la somme est constante (égale à 1 dans ce cas), est maximal quand ces nombres sont égaux ; on obtient un intervalle de confiance un peu plus grand que le précédent,
- utiliser soit une méthode analytique conduisant à la résolution d'une équation du second degré soit une méthode graphique (méthode de l'ellipse) :

1) La détermination de l'intervalle de confiance revient en fait à résoudre l'inéquation :

$$(f_n - p)^2 \leq u_{1-\alpha/2}^2 \times \frac{p(1-p)}{n}$$

qui s'écrit en posant, pour simplifier l'écriture, $k = u_{1-\alpha/2}$:

$$p^2 \left(1 + \frac{k^2}{n} \right) - p \left(\frac{k^2}{n} + 2f_n \right) + (f_n)^2 \leq 0$$

La résolution de cette inéquation donne comme bornes de l'intervalle de confiance :

$$\frac{\frac{k^2}{n} + 2f_n \pm \sqrt{\frac{k^2}{n} \left(\frac{k^2}{n} + 4f_n - 4f_n^2 \right)}}{2 \left(1 + \frac{k^2}{n} \right)}$$

2) On peut tracer l'ellipse d'équation :

$$p^2 \left(1 + \frac{k^2}{n} \right) - p \left(\frac{k^2}{n} + 2f_n \right) + (f_n)^2 = 0$$

et trouver graphiquement l'intervalle de confiance.

Attention

Il faut vérifier que l'approximation par la loi normale était légitime.

Application : on peut déterminer la taille de l'échantillon en fonction de la précision souhaitée.

Exemple 14.11

Dans un échantillon pris au hasard de 100 automobilistes, on constate que 25 d'entre eux possèdent une voiture de cylindrée supérieure à 1 600 cc.

Quel est l'intervalle de confiance pour la proportion d'automobilistes possédant une voiture de cylindrée supérieure à 1 600 cc (intervalle bilatéral, à risques symétriques, seuil de confiance 95 %) ?

– Estimation ponctuelle de p : le nombre d'automobilistes possédant une voiture de cylindrée supérieure à 1 600 cc dans un échantillon de taille $n = 100$ suit la loi binomiale $B(100; p)$.

Un estimateur non biaisé pour la proportion p est donné par la fréquence $\hat{p} = \frac{K}{n}$.

D'où l'estimation ponctuelle de p : $\hat{p} = \frac{25}{100} = 0,25$.

– Intervalle de confiance pour p , on utilise l'approximation normale :

$$\Pr \left(f_n - 1,96 \sqrt{p(1-p)/n} < p < f_n + 1,96 \sqrt{p(1-p)/n} \right) = 0,95$$

D'où les trois méthodes :

– remplacer p par son estimation (0,25). Intervalle de confiance :

$$\Pr(0,25 - 0,085 < p < 0,25 + 0,085) = \Pr(0,165 < p < 0,335) = 0,95$$

– remplacer p par 1/2. Intervalle de confiance :

$$\Pr(0,25 - 0,098 < p < 0,25 + 0,098) = \Pr(0,152 < p < 0,348) = 0,95$$

$$- \text{résoudre l'inéquation : } (p - 0,25)^2 \leq (1,96)^2 \times \frac{p(1-p)}{100}$$

$$1,038 p^2 - 0,538 p + 0,0625 \leq 0$$

On obtient pour intervalle de confiance :

$$\Pr(0,176 < p < 0,342) = 0,95$$

On vérifie facilement que l'approximation normale était valable.

– Comparaison des trois intervalles : le troisième intervalle n'est pas symétrique par rapport à la valeur estimée (0,25) ; il a une étendue, 0,166, comparable à celle du premier intervalle, 0,17, alors que l'étendue du deuxième intervalle, 0,196, est nettement plus grande.

14.3 Estimation et intervalle de confiance dans le cas d'une population d'effectif fini

La plupart des cas étudiés (estimation d'une proportion, d'une moyenne ou d'une variance) supposait implicitement que la population était d'effectif fini. Certains résultats ne sont plus valables si l'on suppose que la population est d'effectif fini.

14.3.1 Estimation de la moyenne m et de l'écart-type σ

Soit une variable aléatoire X distribuée sur une population d'effectif fini N . Cette variable ne prend qu'un nombre fini de valeurs, elle est donc discrète. On considère un échantillon de taille n .

■ Tirage avec remise

– Estimateur sans biais de m : la statistique $\bar{X}$

$$E(\bar{X}) = m \quad \text{Var}(\bar{X}) = \frac{\sigma^2}{n}$$

– Estimateur sans biais de σ^2 : la statistique

$$S^{*2} = \frac{n S^2}{n-1}$$

■ Tirage sans remise

– Estimateur sans biais de m : la statistique $\bar{X}$

$$E(\bar{X}) = m \quad \text{Var}(\bar{X}) = \frac{N-n}{N-1} \times \frac{\sigma^2}{n}$$

– Estimateur sans biais de σ^2 : la statistique

$$\frac{N-1}{N} \times \frac{n}{n-1} \times S^2$$

14.3.2 Estimation d'une proportion**■ Tirage avec remise**

Ce problème a été traité. La fréquence empirique f est un bon estimateur de p :

$$E(f) = p \quad \text{Var}(f) = \frac{p(1-p)}{n}$$

■ Tirage sans remise

La loi concernée est la loi hypergéométrique. La fréquence empirique f est un estimateur sans biais de p :

$$E(f) = p \quad \text{Var}(f) = \frac{N-n}{N-1} \times \frac{p(1-p)}{n}$$

Pour déterminer les intervalles de confiance, on utilise des tables spéciales ou un ordinateur.

15 • LES TESTS STATISTIQUES

La théorie des tests consiste à formuler des hypothèses particulières sur les paramètres ou sur les lois qui interviennent dans les problèmes étudiés, puis à apporter un jugement sur ces hypothèses. Ce jugement est basé d'une part, sur les résultats obtenus sur un ou plusieurs échantillons extraits de la population concernée et d'autre part, sur l'acceptation d'un certain *risque* dans la prise de décision.

Les tests peuvent être classés en différentes catégories :

- tests sur une hypothèse relative à la valeur particulière d'un ou plusieurs paramètre(s) ou *tests paramétriques* (chapitre 15),
- tests de conformité de deux distributions ou *tests d'ajustement* entre une distribution théorique et une distribution expérimentale (chapitre 16),
- *tests de comparaison* de deux populations (comparaison des moyennes, des variances...) (chapitre 16),
- *tests d'indépendance* de deux caractères quantitatifs ou qualitatifs (chapitre 17).

15.1 Notions générales sur les tests statistiques

Exemple 15.1

Considérons des lampes à incandescence dont la durée de vie est une variable aléatoire gaussienne de moyenne $m = 1\,000$ heures et d'écart-type $\sigma = 100$ heures (résultats établis expérimentalement). Un ingénieur propose un nouveau procédé

de fabrication qui doit améliorer, affirme-t-il, cette durée de vie moyenne et la rendre égale à 1 075 heures. Cependant, le procédé, étant plus onéreux, ne pourra être accepté que si l'amélioration est réelle.

Le résultat qui doit être vérifié est donc $m > 1\,000$ heures. Deux hypothèses sont en présence :

- soit $m = 1\,000$ heures est une hypothèse encore vraie et le nouveau procédé n'a pas modifié de façon significative la durée de vie des lampes,
- soit $m = 1\,075$ heures et le nouveau procédé a apporté une réelle amélioration.

La deuxième affirmation est aussi une hypothèse car aucune expérience ne permet d'affirmer que la moyenne des durées de vie augmentera réellement ($m > 1\,000$ heures). En effet, il est possible que le nouveau procédé de fabrication n'améliore pas la durée de vie des ampoules de façon appréciable, il pourrait même la diminuer.

En conclusion, rejeter l'hypothèse $m = 1\,000$ heures ne conduit pas nécessairement à accepter l'hypothèse $m > 1\,000$ heures.

Les hypothèses ayant été formulées, il faut les accepter ou les rejeter à l'aide des données apportées par un échantillon. Une incertitude est toujours associée au jugement apporté par le statisticien, mais ce dernier doit essayer de limiter ce risque.

15.1.1 Principe d'un test d'hypothèse

Soit une population dont les éléments possèdent un certain caractère, dénombrable ou mesurable. Ce caractère est une variable aléatoire X dont la loi de probabilité dépend d'un paramètre θ , dont la valeur exacte est inconnue. Cependant, grâce à des connaissances déduites des propriétés d'échantillons ou grâce à une certaine expérience, on est en mesure de formuler *une hypothèse* sur ce paramètre, $\theta = \theta_0$, par exemple.

Une hypothèse est un énoncé quantitatif sur les caractéristiques d'une population.

La statistique utilisée pour estimer ce paramètre θ lui donne une valeur différente de θ_0 , θ'_0 par exemple. La différence entre ces deux valeurs θ_0 et θ'_0 peut être due, soit à des fluctuations d'échantillonnage, soit à une mauvaise appréciation de la valeur de θ , ou encore à d'autres raisons.

Pour décider si l'hypothèse $\theta = \theta_0$, formulée à l'égard du paramètre, peut être gardée ou rejetée, par comparaison avec la valeur θ'_0 déduite de l'échantillon,

il faut élaborer une stratégie permettant de tester si l'écart observé $|\theta_0 - \theta'_0|$ est trop grand pour être dû aux erreurs d'échantillonnage, ou au contraire, n'est pas en contradiction avec la loi de la variable aléatoire X .

Dans le premier cas, on doit rejeter l'hypothèse $\theta = \theta_0$, on dit que *le test est significatif* ; en revanche, dans le deuxième cas, on doit garder l'hypothèse $\theta = \theta_0$.

15.1.2 Élaboration d'un test

Dans l'exemple 15.1, on teste une hypothèse, que l'on appelle *hypothèse nulle* ou *hypothèse* H_0 :

$$H_0 : m = m_0 = 1000 \text{ heures}$$

Les résultats apportés par un échantillon aléatoire de taille n doivent permettre de rejeter ou non cette hypothèse.

Si cette hypothèse ne peut pas être rejetée, on doit admettre que la différence observée entre la moyenne $\bar{x}$ de l'échantillon et la valeur m_0 est due au hasard.

Si, au contraire, le procédé a apporté un changement dans la fabrication, l'hypothèse proposée H_0 est erronée. Il existe alors une deuxième hypothèse appelée *hypothèse alternative* ou *hypothèse* H_1 . Cette hypothèse peut se formuler de différentes façons :

$$m \neq m_0 \quad m > m_0 \quad m < m_0 \quad m = m_1$$

En général, si on est amené à rejeter l'hypothèse H_0 , c'est-à-dire si on considère que l'hypothèse H_1 est vraie, on devra apporter des corrections à un procédé de fabrication. Mais ces modifications ne peuvent être envisagées que dans la mesure où les résultats observés sont convaincants, car cette décision repose sur l'information apportée par un échantillon.

Il faut donc admettre que toute décision prise comporte un certain risque qu'elle soit erronée. Ce risque est donné par le *seuil de signification du test*.

On peut, par exemple, accepter un risque $\alpha = 0,05$ de rejeter l'hypothèse H_0 , alors qu'elle est vraie, c'est donc le risque d'accepter l'hypothèse H_1 . Cette conclusion signifie que le résultat obtenu sur un échantillon n'avait que 5 chances sur 100 de se produire ou, en d'autres termes, les valeurs apportées par l'échantillon et la valeur m_0 sont considérées comme significativement différentes.

15.1.3 Principales définitions (Résumé)

- Une *hypothèse statistique* est une affirmation concernant certaines caractéristiques d'une population telles que la valeur d'un ou de plusieurs paramètres, la forme de la distribution...
- Un *test d'hypothèse* ou *test statistique* est une démarche conduisant à élaborer une règle de décision permettant de faire un choix entre deux hypothèses statistiques. Les hypothèses envisagées *a priori* s'appellent :
 - L'*hypothèse nulle* H_0 . C'est l'hypothèse selon laquelle on fixe *a priori* la valeur d'un paramètre.
 - L'*hypothèse alternative* H_1 . On peut choisir pour cette hypothèse n'importe quelle hypothèse compatible avec le problème étudié, mais *différente de* H_0 .

Avant toute démarche statistique, il faut définir à quelle condition l'une ou l'autre des hypothèses sera considérée comme vraisemblable. Les deux hypothèses ne jouent pas le même rôle. En effet, c'est l'hypothèse nulle H_0 qui est soumise au test et toute démarche statistique consiste à la considérer comme vraie. Si le test conduit à la rejeter, c'est l'hypothèse alternative H_1 qui sera considérée comme vraie.

Comme hypothèse nulle H_0 , on peut tester :

- une valeur particulière d'un paramètre : $\theta = \theta_0$,
- l'égalité des valeurs d'un paramètre défini sur deux populations différentes,
- l'ajustement d'une distribution théorique à une distribution expérimentale.

L'hypothèse alternative H_1 peut être :

$$\theta = \theta_1 \quad \theta > \theta_0 \quad \theta < \theta_0 \quad \theta \neq \theta_0$$

■ Risques et probabilités d'erreur

Les tests d'hypothèse font intervenir la loi de la distribution de la statistique utilisée comme estimateur pour le paramètre entrant en jeu dans l'hypothèse H_0 .

Pour établir la crédibilité de l'hypothèse H_0 , des règles très précises doivent être énoncées pour permettre de conclure au rejet ou à l'acceptation de H_0 .

Cependant, pour des événements dans lesquels le hasard intervient, il est impossible de prendre la bonne décision sans risque de se tromper. Il faut donc

mettre en œuvre une règle conduisant à rejeter H_0 , si elle est vraie, que dans une faible proportion des cas. Cette décision a un caractère probabiliste, toute décision comporte un risque qu'elle soit erronée. Ce risque, noté α , qui est le risque de rejeter à tort l'hypothèse H_0 alors qu'elle est vraie et qui favorise donc l'hypothèse H_1 s'appelle *seuil de signification* ou *risque de première espèce*.

$$\alpha = \Pr \{ \text{rejeter } H_0 / H_0 \text{ vraie} \} = \Pr \{ \text{choisir } H_1 / H_0 \text{ vraie} \}$$

Ce risque α définit la *région critique* W , d'aire α et de probabilité α , sous l'hypothèse H_0 . C'est l'ensemble des valeurs de la variable aléatoire de décision qui conduisent à écarter H_0 au profit de H_1 .

La région complémentaire, $\overline{W}$, d'aire $(1 - \alpha)$ et de probabilité $(1 - \alpha)$, représente la *région d'acceptation de l'hypothèse* H_0 .

La *règle de décision* peut se formuler de la façon suivante :

Si la valeur de la statistique considérée appartient à la région d'acceptation $\overline{W}$, on favorise l'hypothèse nulle H_0 , si elle appartient à la région critique W , on favorise l'hypothèse alternative H_1 .

La règle de décision comporte un *risque* β ou *risque de deuxième espèce*, c'est le risque de ne pas rejeter H_0 alors que H_1 est vraie :

$$\beta = \Pr \{ \text{ne pas rejeter } H_0 / H_1 \text{ vraie} \} = \Pr \{ \text{choisir } H_0 / H_1 \text{ vraie} \}$$

On peut aussi écrire, en introduisant la région critique W :

$$\Pr (\overline{W} / H_0) = 1 - \alpha \quad \Pr (W / H_1) = 1 - \beta$$

En résumé :

- Erreur de première espèce : on rejette H_0 alors que H_0 est vraie.
- Erreur de deuxième espèce : on ne rejette pas H_0 alors que H_1 est vraie.

Remarque

La probabilité α , appelée aussi *risque du client*, est choisie *a priori* par l'utilisateur. Les valeurs les plus utilisées pour α sont 0,05 et 0,01.

En revanche, la probabilité β , appelée aussi *risque du fournisseur* (celui de voir, par exemple, une bonne production refusée), dépend de l'hypothèse alternative H_1 . Pour calculer cette probabilité, on donne au paramètre la (ou les) valeur(s) figurant dans l'hypothèse H_1 .

La quantité $(1 - \beta)$ est la *puissance du test* à l'égard de la valeur du paramètre figurant dans l'hypothèse H_1 . Elle représente la probabilité d'accepter H_1 alors que celle-ci est vraie.

Ces différentes situations sont résumées dans le tableau suivant.

Tableau 15.1 – Décisions et probabilités.

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

■ Choix de l'hypothèse nulle H_0

L'hypothèse H_0 étant soumise au test, son choix est très important ; il peut être guidé par différentes raisons :

- l'hypothèse H_0 est une hypothèse de prudence, par exemple, pour tester l'efficacité d'un nouveau procédé ou d'un nouveau médicament, on part d'une hypothèse défavorable au nouveau produit,
- l'hypothèse H_0 est une hypothèse solidement établie, ou c'est la seule hypothèse facile à formuler.

Exemple 15.2 (suite de l'exemple 15.1)

On suppose que la durée de vie des lampes suit une loi normale de même écart-type σ , égal à 100, sous les deux hypothèses. Le meilleur estimateur de l'espérance mathématique (théorie de l'estimation) est la statistique $\bar{X}$, moyenne d'un échantillon de taille n . C'est la variable de décision utilisée pour construire le test.

On décide de contrôler un échantillon de taille $n = 25$ lampes fabriquées suivant le nouveau procédé. Les deux hypothèses en présence sont :

$$H_0 : m = m_0 = 1\,000 \text{ heures}$$

$$H_1 : m = m_1 = 1\,075 \text{ heures}$$

- La variable de décision, $\bar{X}$, suit une loi normale qui a pour paramètres :

$$m_0 = 1\,000 \text{ heures sous l'hypothèse } H_0$$

$$m_1 = 1\,075 \text{ heures sous l'hypothèse } H_1$$

L'écart-type est égal à 100 heures sous les deux hypothèses.

Si le risque de première espèce α est égal à 5 %, la région critique, de rejet de H_0 , est définie par :

$$\Pr(W/H_0) = 0,05 \quad \text{c'est-à-dire} \quad \Pr(\bar{X} > d) = 0,05$$

Soit U la variable aléatoire centrée réduite associée à $\bar{X}$:

$$\Pr(\bar{X} > d) = \Pr\left(U = \frac{\bar{X} - 1\,000}{20} > \frac{d - 1\,000}{20}\right) = 0,05$$

$$\frac{d - 1\,000}{20} = 1,6449 \quad \Rightarrow \quad d \cong 1\,033 \text{ heures}$$

La valeur 1,6449 est lue sur les tables 5.1 ou 5.2.

– Règles de décision :

$\bar{x} \geq 1\,033$ heures, on rejette H_0

$\bar{x} < 1\,033$ heures, on garde H_0 .

– L'échantillon a donné pour la statistique $\bar{X}$ la valeur 1 050 heures. On doit donc rejeter l'hypothèse H_0 , et accepter l'hypothèse H_1 .

– Le risque β de deuxième espèce est défini par :

$$\beta = \Pr(\bar{W}/H_1) = \Pr(\bar{X} < d/H_1)$$

$$\beta = \Pr(\bar{X} < 1\,033) = \Pr\left(U = \frac{\bar{X} - 1\,075}{20} < \frac{1\,033 - 1\,075}{20} = -2,10\right)$$

D'où $\beta = 0,0179$.

La probabilité de refuser H_1 alors que cette hypothèse est vraie est donc égale à 0,0179, elle est assez faible ; la puissance du test est égale à 0,9821.

L'hypothèse H_1 considérée dans cet exemple a conduit à un test unilatéral à droite. La figure 15.1 montre les régions critiques correspondants aux différents tests sur une moyenne.

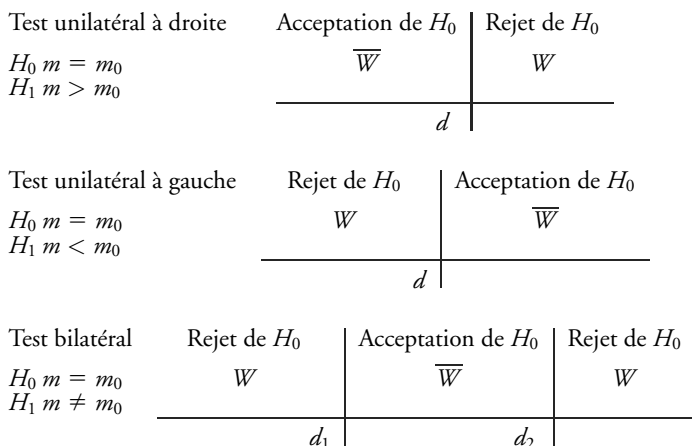


Figure 15.1 – Régions critiques correspondant à différents tests sur une moyenne.

15.1.4 Propriétés des risques de première et deuxième espèces

- La taille n de l'échantillon et le risque α sont fixés. Dans ces conditions, le risque β diminue si la différence entre les deux valeurs proposées, m_0 et m_1 , augmente.
- Si le risque α diminue, la zone de non-rejet de l'hypothèse H_0 augmente. À ne pas vouloir rejeter à tort l'hypothèse H_0 , on finit par la garder trop souvent. De plus, dans ces conditions, le risque β augmente, donc la région de refus de l'hypothèse H_1 augmente.

Les deux risques de première et deuxième espèces sont antagonistes.

- Si on fixe le risque α et si la taille n de l'échantillon augmente, la zone de non-rejet de l'hypothèse H_0 devient plus petite, d'où une diminution du risque β ; le test est donc plus puissant.

15.1.5 Élaboration d'un test et démarche à suivre

Pour élaborer un test statistique, il faut :

- formuler de façon précise l'hypothèse nulle H_0 et l'hypothèse alternative H_1 ,

- fixer, avant l'expérience, le risque α de première espèce, c'est-à-dire le risque de rejeter l'hypothèse H_0 alors qu'elle est vraie,
 - préciser les conditions d'application du test : forme de la loi de probabilité de la population étudiée, taille de l'échantillon, variance connue ou inconnue...
 - choisir la statistique la mieux adaptée en fonction des caractéristiques de la population étudiée et donner sa loi de probabilité sous les deux l'hypothèses, ces lois doivent être différentes,
 - déterminer la région critique ou région de rejet de l'hypothèse H_0 au profit de l'hypothèse H_1 et en déduire la règle de décision :
 - $\overline{W}$ région critique conduisant au rejet de H_0 : $\Pr\{\overline{W}/H_0\} = \alpha$,
 - W région de non-rejet donc d'acceptation de H_0 : $\Pr\{W/H_0\} = (1-\alpha)$.
- On en déduit la valeur du risque de deuxième espèce β :

$$\Pr\{W/H_1\} = (1 - \beta)$$

- calculer effectivement la valeur numérique t de la variable de décision en utilisant les résultats apportés par l'échantillon,
- donner les conclusions du test :
 - si $t \in \overline{W}$, on rejette l'hypothèse H_0 au profit de l'hypothèse H_1 sans conclure que l'hypothèse H_0 est fausse, mais elle a une forte probabilité de l'être, le test est significatif,
 - si $t \in W$, on ne peut pas rejeter l'hypothèse H_0 donc on garde cette hypothèse, le test n'est pas significatif.

15.2 Différentes catégories de tests statistiques

Les tests paramétriques sont des tests relatifs à un ou plusieurs paramètres d'une loi spécifiée. On distingue :

- les hypothèses simples du type $\theta = \theta_0$,
- les hypothèses composites du type $\theta \in \delta_0$ où δ_0 est un intervalle de $\mathbb{R}$; elles se ramènent, en général soit à $\theta > \theta_0$, soit à $\theta < \theta_0$ ou encore à $\theta \neq \theta_0$.

Ces tests supposent, en général, l'existence d'une variable aléatoire X suivant une loi normale. Si les résultats obtenus sont valables même si la variable

aléatoire X n'est pas une variable gaussienne, on dit que *le test est robuste*, les résultats restent valables après quelques modifications des données.

Parmi les tests robustes, les tests libres sont les plus intéressants, ils sont valables quelle que soit la forme de la loi de la variable aléatoire X . On peut donc les utiliser quand on ne connaît pas cette loi. Les tests de moyenne, de non-corrélation par exemple, sont des tests robustes. Les tests robustes sont souvent des tests paramétriques.

15.3 Test entre deux hypothèses simples et méthode de Neyman et Pearson

15.3.1 Énoncé du problème

X est une variable aléatoire dont la densité $f(x; \theta)$ dépend du paramètre réel θ et $\underline{X} = (X_1, \dots, X_n)$ est un échantillon aléatoire de taille n de cette variable. On veut tester :

$$H_0 : \theta = \theta_0$$

$$H_1 : \theta = \theta_1$$

Soit $L(\underline{x}; \theta)$ la densité de probabilité de l'échantillon ou vraisemblance. La région critique W , pour un risque de première espèce égal à α , est l'ensemble des points de $\mathbb{R}^n$ défini par :

$$\Pr(W/H_0) = \alpha = \int_W L(\underline{x}; \theta_0) d\underline{x}$$

La méthode de Neyman et Pearson permet de construire cette région critique.

15.3.2 Théorème de Neyman et Pearson

Avec les hypothèses précédentes sur la variable aléatoire X et sur le test, on démontre le résultat suivant :

$\forall \alpha \in [0, 1]$, il existe un test T , de puissance maximale, défini par la région critique W au seuil de signification α :

$$W = \{\underline{x} \in \mathbb{R}^n / L(\underline{x}; \theta_1) > k_\alpha L(\underline{x}; \theta_0) \text{ où } k_\alpha \geq 0\}$$

La méthode de Neyman et Pearson consiste à rendre maximale la puissance du test, c'est-à-dire la quantité :

$$\Pr(W/H_1) = 1 - \beta = \int_W L(\underline{x}; \theta_1) d\underline{x}$$

Les étapes de la démonstration sont les suivantes :

- s'il existe une constante k_α telle que l'ensemble W défini par :

$$W = \{\underline{x} \in \mathbb{R}^n / L(\underline{x}; \theta_1) > k_\alpha L(\underline{x}; \theta_0)\}$$

soit de probabilité α sous H_0 , alors cet ensemble W réalise le maximum de $(1 - \beta)$.

- pour démontrer l'existence de la constante k_α , on définit une région $A(K)$ de $\mathbb{R}^n$ telle que :

$$A(K) = \{\underline{x} \in \mathbb{R}^n / L(\underline{x}; \theta_1) > KL(\underline{x}; \theta_0)\}$$

K étant une constante positive donnée.

La probabilité $\Pr(A(K)/H_0)$ est une fonction de K , continue, monotone si la variable aléatoire X est à densité continue. En effet :

- si $K = 0$, $L(\underline{x}; \theta_1)$ étant positive, $\Pr(A(0)/H_0) = 1$,
- si $K \rightarrow \infty$, $L(\underline{x}; \theta_1)$ étant une densité est bornée, $\Pr(A(K)/H_0) \rightarrow 0$,
- il existe donc une valeur intermédiaire k_α telle que $\Pr(A(k_\alpha)/H_0) = \alpha$ et le théorème de Neyman et Pearson est démontré.

15.3.3 Étude de la puissance $(1 - \beta)$ du test

On démontre que $(1 - \beta) > \alpha$. Ce test est dit sans biais. Par définition :

$$1 - \beta = \int_W L(\underline{x}; \theta_1) d\underline{x} \quad \text{et} \quad \alpha = \int_W L(\underline{x}; \theta_0) d\underline{x}$$

Dans W , on a $L(\underline{x}; \theta_1) > k_\alpha L(\underline{x}; \theta_0)$, donc $(1 - \beta) > \alpha k_\alpha$.

- Si $k_\alpha > 1$, le résultat est trivial : $1 - \beta > \alpha$.
- Si $k_\alpha \leq 1$, on montre de la même façon que $1 - \alpha > \beta$ en intégrant dans $\overline{W}$. En effet, dans $\overline{W}$, $L(\underline{x}; \theta_1) \leq k_\alpha L(\underline{x}; \theta_0)$. D'où :

$$\beta = \int_W L(\underline{x}; \theta_1) d\underline{x} < k_\alpha \int_W L(\underline{x}; \theta_0) d\underline{x} < \int_{\overline{W}} L(\underline{x}; \theta_0) d\underline{x} = 1 - \alpha$$

15.3.4 Convergence du test

Si $n \rightarrow \infty$, $1 - \beta \rightarrow 1$.

15.3.5 Test et statistique exhaustive

Soit $T = \varphi(X_i)$ une statistique exhaustive ; la densité de l'échantillon se met alors sous la forme :

$$L(\underline{x}; \theta) = g(t; \theta) h(\underline{x}) \quad \text{où} \quad t = \varphi(x_i)$$

La région critique, selon la méthode de Neyman et Pearson, est définie par :

$$g(t; \theta_1) > k_\alpha g(t; \theta_0)$$

Elle dépend donc exclusivement de la statistique exhaustive.

Exemple 15.3

On veut définir la région critique, par la méthode de Neyman-Pearson, pour la moyenne m d'une loi normale dont l'écart-type σ est supposé connu. Les deux hypothèses à tester sont :

$$H_0 : m = m_0$$

$$H_1 : m = m_1$$

Une statistique exhaustive, pour la moyenne m , est la statistique $T = \bar{X}$ dont la loi de probabilité $N(m; \sigma/\sqrt{n})$ a pour densité :

$$g(\bar{x}; m) = \frac{1}{(\sigma/\sqrt{n}) \sqrt{2\pi}} \exp \left[-\frac{(\bar{x} - m)^2}{2\sigma^2/n} \right]$$

Le rapport des densités de l'échantillon pour les deux valeurs du paramètre m est égal à :

$$\frac{g(t; m_1)}{g(t; m_0)} = \exp \left[-\frac{n}{2\sigma^2} [(\bar{x} - m_1)^2 - (\bar{x} - m_0)^2] \right]$$

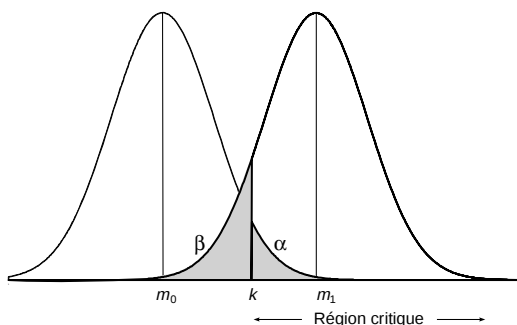
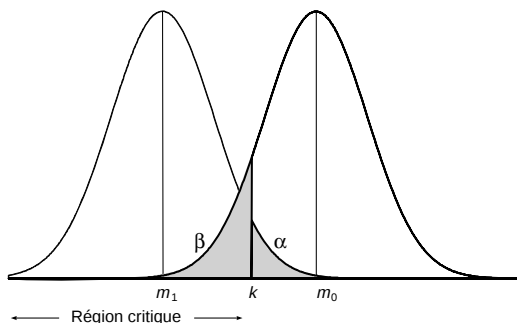
La condition de Neyman s'écrit, après simplification :

$$\begin{aligned} (\bar{x} - m_0)^2 - (\bar{x} - m_1)^2 &> k_\alpha \\ (m_0 - m_1)(m_0 + m_1 - 2\bar{x}) &> k_\alpha \end{aligned}$$

Pour définir la région critique (de rejet de H_0), on doit distinguer deux cas :

- $m_0 > m_1$, la condition précédente est équivalente à $\bar{X} < K_\alpha$,
- $m_0 < m_1$, la condition précédente est équivalente à $\bar{X} > K'_\alpha$.

Intuitivement, on aurait fait ces choix. En effet, si l'hypothèse H_1 est $m_1 < m_0$, on doit « refuser » les valeurs de la statistique $\bar{X}$ qui sont trop petites et inversement si l'hypothèse H_1 est $m_1 > m_0$.

Figure 15.2 – Région critique pour $m_0 < m_1$.Figure 15.3 – Région critique pour $m_0 > m_1$.

15.4 Tests entre deux hypothèses composites

15.4.1 Test d'une hypothèse simple contre une hypothèse composite

Différents cas peuvent être envisagés, comme par exemple :

$$H_0 : \theta = \theta_0$$

$$H_1 : \theta > \theta_0$$

$$H_0 : \theta = \theta_0$$

$$H_1 : \theta \neq \theta_0$$

L'hypothèse H_1 est composée d'un ensemble de valeurs. Le risque de deuxième espèce β doit être calculé pour chaque valeur du paramètre définie par cette hypothèse. On obtient ainsi une fonction $\beta(\theta)$, son graphe est *la courbe d'efficacité du test*, la fonction $1 - \beta(\theta)$ est la *puissance du test*, et son graphe *la courbe de puissance du test*.

Un test est appelé *uniformément le plus puissant* (UPP) si, quelle que soit la valeur du paramètre θ appartenant à H_1 , sa puissance est supérieure à la puissance de tout autre test. Il en est ainsi si, par exemple, la région critique ne dépend pas de la valeur θ du paramètre.

Exemple 15.4

On veut vérifier que le pourcentage p de pièces défectueuses dans un lot de plusieurs milliers de pièces n'excède pas 3 %. On prélève un échantillon de $n = 200$ pièces et on adopte la règle de décision suivante, en désignant par K le nombre de pièces défectueuses dans l'échantillon prélevé :

- si $K \leq 10$ le lot est accepté,
- si $K \geq 11$ le lot est refusé.
- *Risque de première espèce* associé à cette règle de décision :

$$\alpha = \Pr(\text{refuser } H_0 / H_0 \text{ vraie}) = \Pr(K \geq 11 / p_0 = 0,03)$$

La variable K suit la loi binomiale $B(n; p)$ avec $n = 200$ et $p = p_0 = 0,03$ sous H_0 et $p = p_1 > 0,03$ sous H_1 .

On peut utiliser l'approximation normale, en effet, $np = 200 \times 0,03 = 6 > 5$ et $n(1 - p) = 200 \times 0,97 = 194 > 5$.

Paramètres de la loi normale :

$$E(K) = 6 \quad \text{Var}(K) = 200 \times 0,03 \times 0,97 = 5,82 = (2,41)^2$$

$\alpha = \Pr(11 \leq K \leq 200) = \Pr(10,5 < K < 200,5)$ (avec la correction de continuité)

$$\Pr\left(\frac{10,5 - 6}{2,41} < \frac{K - 6}{2,41} = U < \frac{20,5 - 6}{2,41}\right) = 0,031 = \alpha$$

– *Risque de deuxième espèce* :

$$\beta = \Pr(\text{refuser } H_1 / H_1 \text{ vraie}) = \Pr(0 \leq K \leq 10 / p > 0,03)$$

La loi limite de la variable K est la loi normale $N(200p; \sqrt{200p(1-p)})$.

$$\beta = \Pr \left(\frac{0,5 - np}{\sqrt{np(1-p)}} < \frac{K - np}{\sqrt{np(1-p)}} < \frac{10,5 - np}{\sqrt{np(1-p)}} \right) \text{ (avec la correction de continuité).}$$

D'où les résultats où U est la variable centrée réduite normale $U = \frac{10,5 - np}{\sqrt{np(1-p)}}$:

Tableau 15.2 – Résultats numériques.

p	0,04	0,05	0,06	0,08	0,10
np	8	10	12	16	20
$\sqrt{np(1-p)}$	2,77	3,08	3,36	3,84	4,24
u	0,902	0,162	-0,446	-1,432	-2,24
β	0,81	0,56	0,328	0,076	0,012
$1 - \beta$	0,19	0,44	0,672	0,924	0,988

On remarque que le test est d'autant plus puissant que $p \gg 0,03$.

Si on observe un pourcentage de pièces défectueuses $p = 0,08$, on trouve $\beta = 0,076$ (risque d'accepter le lot) et donc $1 - \beta = 0,924$.

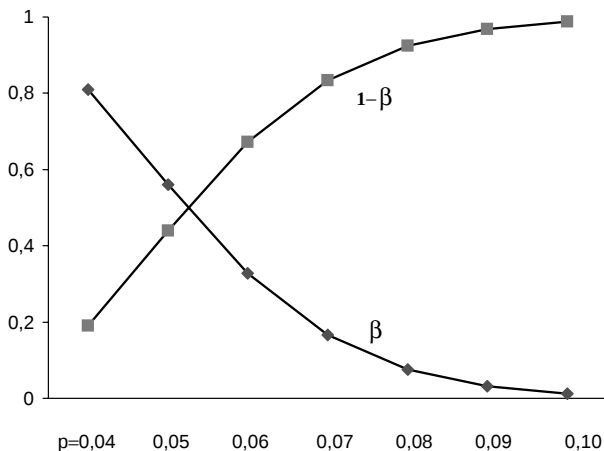


Figure 15.4 – Courbe puissance et courbe d'efficacité du test.

15.4.2 Test entre deux hypothèses composites

L'hypothèse H_0 est aussi une hypothèse composite et le risque α de première espèce dépend de la valeur du paramètre, on imposera la condition $\alpha(\theta) \leq \alpha$ valeur donnée. On démontre l'existence de tests UPP dans certains cas, comme par exemple :

$$H_0 : \theta < \theta_0$$

$$H_1 : \theta \geq \theta_0$$

15.5 Principaux tests paramétriques

15.5.1 Tests sur les paramètres m et σ d'une loi normale

■ Test sur la moyenne m

L'écart-type σ est connu. Selon l'hypothèse H_1 , différents cas sont à étudier :

$$H_0 : m = m_0 \quad \text{et} \quad H_1 : m = m_1 > m_0 \quad (1)$$

La variable de décision est la statistique $\bar{X}$ dont la loi est facile à établir :

$$\text{Loi de } \bar{X} \text{ sous } H_0 \quad N(m_0; \sigma/\sqrt{n})$$

$$\text{Loi de } \bar{X} \text{ sous } H_1 \quad N(m_1; \sigma/\sqrt{n})$$

En désignant par α le risque de première espèce, la région critique est définie par :

$$\alpha = \Pr(\bar{X} > k / H_0) = \Pr\left(U > \frac{k - m_0}{\sigma/\sqrt{n}}\right)$$

La valeur $u = \frac{k - m_0}{\sigma/\sqrt{n}}$ est lue sur la table 5.2, on en déduit la valeur de k et donc la région critique.

La forme de l'hypothèse H_1 conduit à rejeter les valeurs trop grandes de $\bar{X}$.

Le risque de deuxième espèce β est défini par :

$$\beta = \Pr(\bar{X} < k / H_1) = \Pr\left(U < \frac{k - m_1}{\sigma/\sqrt{n}}\right) = \Pr(U < U_\beta)$$

$$U_\beta = \frac{k - m_1}{\sigma/\sqrt{n}}$$

Sur la table 5.2, on lit la valeur de U_β . D'où la valeur de β .

$$H_0 : m = m_0 \quad \text{et} \quad H_1 : m = m_1 < m_0 \quad (2)$$

La variable de décision est encore la statistique $\bar{X}$ et la région critique est définie par :

$$\alpha = \Pr(\bar{X} < k / H_0) = \Pr\left(U < \frac{k - m_0}{\sigma/\sqrt{n}}\right) = \Pr(U < U_\alpha)$$

De façon analogue, on en déduit la valeur de β :

$$\beta = \Pr(\bar{X} > k / H_1) = \Pr\left(U > \frac{k - m_1}{\sigma/\sqrt{n}}\right)$$

$$H_0 : m = m_0 \quad \text{et} \quad H_1 : m = m_1 \neq m_0 \quad (3)$$

L'hypothèse H_1 implique $m_1 < m_0$ ou $m_1 > m_0$. La région critique est déterminée avec la même variable de décision $\bar{X}$ par :

$$\alpha = \Pr(|\bar{X}| > k / H_0) = \Pr(|U| > \frac{k - m_0}{\sigma/\sqrt{n}}) = \Pr(|U| > u)$$

La valeur u est lue sur la table 5.2. Puis, on calcule la valeur de β comme dans les cas précédents.

Exemple 15.5

On prélève, au hasard, dans une population suivant une loi normale de variance égale à 25, un échantillon de taille $n = 16$.

– En choisissant un risque de première espèce $\alpha = 0,05$ (risque bilatéral, symétrique), quelle est la règle de décision si l'on veut tester les hypothèses :

$H_0 : m = m_0 = 45$ et $H_1 : m = m_1 \neq 45$?

Soient k_1 et k_2 les seuils critiques.

La règle de décision est : on accepte l'hypothèse H_0 si $k_1 < \bar{x} < k_2$

$$\Pr(k_1 < \bar{x} < k_2 / H_0) = 0,95$$

$$\Pr\left(\frac{k_1 - 45}{5/4} < \frac{\bar{x} - 45}{5/4} < \frac{k_2 - 45}{5/4}\right) = 0,95$$

$$\text{D'où : } \frac{k_1 - 45}{5/4} = -1,96 \quad \frac{k_2 - 45}{5/4} = 1,96 \quad k_1 = 42,55 \quad k_2 = 47,45$$

– On observe une moyenne de l'échantillon égale à 49. Cette valeur est en contradiction avec l'hypothèse H_0 , on refuse donc l'hypothèse H_0 et on accepte l'hypothèse H_1 .

On peut calculer le risque de deuxième espèce β associé à cette valeur $m_1 = 49$:

$$\beta = \Pr(k_1 < \bar{x} < k_2 / H_1)$$

$$\beta = \Pr\left(\frac{42,55 - 49}{5/4} < \frac{\bar{x} - 49}{5/4} < \frac{47,45 - 49}{5/4}\right) = \Pr(-5,16 < U < -1,24)$$

D'où : $\beta = 0,1075$

La probabilité de refuser l'hypothèse H_1 , $m_1 = 49$, alors qu'elle est vraie, est égale à 0,1075, la puissance du test est égale à 0,8925.

Si on avait observé une moyenne égale à 48, on refusait également l'hypothèse H_0 ; un calcul analogue donnait pour le risque β la valeur 0,33 et pour la puissance du test la valeur 0,67.

Ce test est d'autant plus puissant que les valeurs m_0 et m_1 sont très différentes.

L'écart-type σ n'est pas connu mais estimé. La variable aléatoire $\frac{\bar{X} - m}{S/\sqrt{n-1}}$ suit une loi de Student à $(n-1)$ degrés de liberté. On procède alors de façon analogue en utilisant la table de la loi de Student (table 8).

Exemple 15.6

Un fabricant de téléviseurs achète un certain composant électronique à un fournisseur. Un accord entre le fournisseur et le fabricant stipule que la durée de vie de ces composants doit être égale à 600 heures au moins. Le fabricant qui vient de recevoir un lot important de ce composant veut en vérifier la qualité. Il tire au hasard un échantillon de 16 pièces. Le test de durée de vie pour cet échantillon donne les résultats suivants :

620	570	565	590	530	625	610	595
540	580	605	575	550	560	575	615

Le fabricant doit-il accepter le lot ? (On choisit un risque de première espèce égal à 5 %.) Quel est le risque de deuxième espèce ?

On admettra que la durée de vie de ces composants suit une loi normale.

– Caractéristiques de l'échantillon : $\bar{x} = 581,60$ et $s = 27,90$.

Les hypothèses à tester sont :

$$H_0 : m = 600 \text{ heures} \quad \text{et} \quad H_1 : m < 600 \text{ heures}$$

– La variable de décision est la moyenne de l'échantillon.

La variable $\frac{\bar{X} - m}{S/\sqrt{n-1}}$ suit une loi de Student à $(n-1) = 15$ degrés de liberté.

– Calcul du seuil critique d_c . La règle de décision est la suivante :

$$\bar{x} < d_c \quad \text{Refus de } H_0 \quad \bar{x} > d_c \quad \text{Acceptation de } H_0$$

$$\Pr(\bar{x} < d_c / H_0) = 0,05$$

$$\Pr\left(\frac{\bar{x} - 600}{27,90/\sqrt{15}} < \frac{d_c - 600}{27,90/\sqrt{15}}\right) = 0,05$$

$$\text{Or } \Pr[t(15) < -1,753] = 0,05.$$

$$\text{D'où : } \frac{d_c - 600}{27,90/\sqrt{15}} = -1,753 \quad \text{et} \quad d_c = 587,40.$$

La valeur de la moyenne arithmétique donnée par l'échantillon étant égale à 581,60, on doit rejeter l'hypothèse H_0 .

– Risque de deuxième espèce

La forme de l'hypothèse H_1 implique que l'on doit calculer le risque β pour toutes les valeurs de $m < 600$. On fait le calcul pour $m = 575$ par exemple :

$$\beta = \Pr(\bar{x} > d_c / H_1)$$

$$\beta = \Pr\left(\frac{\bar{x} - 580}{27,90/\sqrt{15}} > \frac{587,40 - 575}{27,90/\sqrt{15}}\right) = \Pr[t(15) > 1,721] \cong 0,05$$

Remarque

Si la loi suivie par la variable aléatoire X n'est pas une loi normale, mais si la taille de l'échantillon est supérieure à 30, on peut admettre que la loi limite est la loi normale. Si de plus l'écart-type n'est pas connu, on utilisera la loi de Student.

■ Test sur l'écart-type σ

La moyenne m est connue. Les hypothèses à tester sont, par exemple :

$$H_0 : \sigma = \sigma_0 \quad H_1 : \sigma = \sigma_1 > \sigma_0 \quad (1)$$

La variable aléatoire $\frac{1}{\sigma^2} \sum_{i=1}^n (X_i - m)^2$ suit une loi du chi-deux à n degrés de liberté.

La variable de décision est la statistique D qui est un estimateur sans biais de la variance :

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

La région critique (ou de rejet de H_0) est définie par :

$$\alpha = \Pr(D > k / H_0) = \Pr\left(\chi^2(n) > \frac{nk}{\sigma_0^2}\right)$$

En utilisant la table 6, on détermine k puis le risque β :

$$\beta = \Pr(D < k / H_1) = \Pr\left(\chi^2(n) < \frac{nk}{\sigma_1^2}\right)$$

$$H_0 : \sigma = \sigma_0 \quad H_1 : \sigma = \sigma_1 < \sigma_0 \quad (2)$$

La région critique est définie par :

$$\alpha = \Pr(D < k / H_0) = \Pr\left(\chi^2(n) < \frac{nk}{\sigma_0^2}\right)$$

En utilisant la table 6, on détermine k puis le risque β :

$$\beta = \Pr(D > k / H_1) = \Pr\left(\chi^2(n) > \frac{nk}{\sigma_1^2}\right)$$

La moyenne m n'est pas connue, elle doit donc être estimée. Dans ce cas, la variable de décision est la statistique :

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

La variable aléatoire $\frac{nS^2}{\sigma^2}$ suit une loi du chi-deux à $(n-1)$ degrés de liberté. Les différents tests s'étudient comme dans le cas précédent.

Exemple 15.7

On veut contrôler la précision d'une balance au bout d'un an de fonctionnement. Si on pèse un poids de 1 g, on peut considérer que l'observation est la réalisation d'une variable aléatoire suivant une loi normale d'espérance mathématique $m = 1$ g (la balance est juste) et d'écart-type $\sigma_0 = 1,2$ mg. Si au bout d'un an de fonctionnement, on constate que l'écart-type σ est supérieur à σ_0 , la précision de la balance a diminué.

– On veut tester (seuil critique $\alpha = 0,10$, taille de l'échantillon $n = 10$) :

$$H_0 \sigma_0 = 1,2 \text{ mg contre } H_1 \sigma_1 = 1,5 \text{ mg}$$

La moyenne m étant connue, la variable de décision est la statistique $D = \frac{1}{n} \sum_{i=1}^n (X_i - m)^2$ et la variable $\frac{nD}{\sigma^2}$ suit une loi du chi-deux à n degrés de liberté.

La région critique est définie par :

$$\Pr(D > k/H_0) = \Pr\left(\frac{nD}{\sigma_0^2} > \frac{nk}{\sigma_0^2}\right) = \Pr\left(\chi^2(n) > \frac{nk}{\sigma_0^2}\right) = 0,10$$

$$\frac{nk}{\sigma_0^2} = 15,987 \quad n = 10 \quad \sigma_0 = 1,2 \quad \Rightarrow \quad k = 2,3$$

k est exprimé en mg comme l'écart-type.

Conclusion : soit d la valeur de la statistique donnée D par l'échantillon :

$d > k$, on rejette l'hypothèse H_0 .

– Risque de deuxième espèce :

$$\beta = \Pr(D < k/H_1) = \Pr\left(\frac{nD}{\sigma_1^2} < \frac{10 \times 2,30}{(1,5)^2} = 10,222\right) \cong 0,55$$

$1 - \beta = 0,45$, le test est peu puissant.

– Les résultats de 10 pesées ont donné : $d = 3$. On rejette donc l'hypothèse H_0 .

Remarque

Les résultats précédents ne sont valables que dans le cas où la variable aléatoire X suit une loi normale.

15.5.2 Tests sur une proportion

Le but est de tester si la proportion p d'individus d'une population P , présentant un certain caractère qualitatif peut être considérée comme égale ou non à une valeur p_0 .

Un estimateur sans biais de p est la proportion F d'individus présentant ce caractère dans un échantillon aléatoire de taille n . La variable aléatoire $K = nF$ (nombre d'individus présentant ce caractère) suit la loi binomiale $B(n; p)$. Si np et $n(1 - p)$ sont supérieurs à 5, on peut remplacer la loi binomiale par une loi normale. On en déduit que F suit approximativement la loi normale :

$$N \left(p; \sqrt{\frac{p(1-p)}{n}} \right)$$

Un test sur une proportion a été traité dans l'exemple 15.4.

On termine le problème comme dans le paragraphe précédent.

16 • TESTS D'AJUSTEMENT ET DE COMPARAISON

Les tests d'ajustement permettent de juger l'adéquation entre une situation réelle et un modèle théorique, les tests de comparaison d'échantillons sont utilisés pour comparer deux ou plusieurs échantillons.

16.1 Tests d'ajustement

Deux problèmes différents peuvent se rencontrer en statistique :

- soit ajuster une loi de probabilité à un échantillon, la loi est inconnue, sa forme et les valeurs des paramètres sont obtenues à partir des caractéristiques de l'échantillon,
- soit ajuster un échantillon à une loi de probabilité donnée, la loi est connue (fonction de répartition ou densité entièrement spécifiée), on doit vérifier l'adéquation entre la loi théorique et l'échantillon.

Le choix d'une loi est lié :

- à la nature du phénomène étudié afin de choisir entre loi discrète et loi continue,
- à la forme de la distribution (histogramme),
- à la connaissance et à l'interprétation des principales caractéristiques de l'ensemble des données, espérance, médiane, variance ou écart-type, coefficients d'asymétrie et d'aplatissement...
- au nombre de paramètres des lois, une loi dépendant de plusieurs paramètres peut s'adapter plus facilement à une distribution donnée.

Une loi étant proposée, différents tests peuvent être utilisés pour juger de la concordance entre une distribution théorique et une distribution réelle :

- le test le plus utilisé est le test de Pearson, plus connu sous le nom de *test du chi-deux*. Il peut aussi être utilisé pour tester l'égalité de k proportions, l'indépendance de deux variables aléatoires étudiées suivant différentes modalités (tableau de contingence),
- le test de Kolmogorov-Smirnov,
- le test de Cramer-Von-Mises.

16.1.1 Méthodes empiriques

■ Forme de l'histogramme

La forme de l'histogramme permet de privilégier certains modèles si des conditions de symétrie sont respectées ou au contraire d'éliminer des modèles :

- une distribution symétrique peut suggérer une loi normale, une loi de Cauchy ou une loi de Student,
- une distribution fortement dissymétrique fait penser à une loi Log-normale, à une loi gamma, à une loi de Weibull ou à une loi bêta de type II.

Cependant, comme on étudie un phénomène réel, certains modèles devront être privilégiés alors que d'autres devront être systématiquement écartés (ainsi les lois utilisées en fiabilité sont surtout les lois exponentielles ou de Weibull).

■ Vérification de certaines propriétés mathématiques

L'échantillon permet de calculer $\bar{x}$ et s^2 , c'est-à-dire des estimations de l'espérance mathématique et de la variance σ^2 .

Pour une loi de Poisson de paramètre λ , $E(X) = \text{Var}(X) = \lambda$ et pour une loi exponentielle de paramètre λ , $E(X) = \sigma = 1/\lambda$.

Si la première propriété est approximativement vérifiée par l'échantillon, on peut penser à ajuster une loi de Poisson et si c'est la deuxième propriété, on ajustera une loi exponentielle.

Si une variable aléatoire X suit une loi normale centrée réduite, le coefficient d'asymétrie γ_1 est nul et le coefficient d'aplatissement γ_2 est égal à 3. Si ces propriétés sont approximativement vérifiées par l'échantillon (variable continue), on peut penser à ajuster une loi normale.

RappelCoefficient d'asymétrie : $\gamma_1 = \frac{\mu_3}{\sigma^3}$ Coefficient d'aplatissement : $\gamma_2 = \frac{\mu_4}{\sigma^4}$ **■ Ajustement graphique**

Soit L une loi de probabilité de fonction de répartition F . Cette fonction varie de 0 à 1 et est représentée dans un plan par une courbe Γ .

On considère une série classée, par ordre croissant, de n observations réparties en k classes d'effectifs n_i . La fonction de répartition empirique F^* de l'échantillon doit être peu différente de la fonction de répartition théorique F .

Soit z_i le centre de la classe $[x_{i-1}, x_i]$ et h_i l'étendue de cette classe ; le point P_i

$$\text{d'abscisse } x_{i-1} + \frac{h_i}{2} \text{ et d'ordonnée } \frac{1}{n} \sum_{j=1}^i n_j$$

est un point de la fonction de répartition empirique.

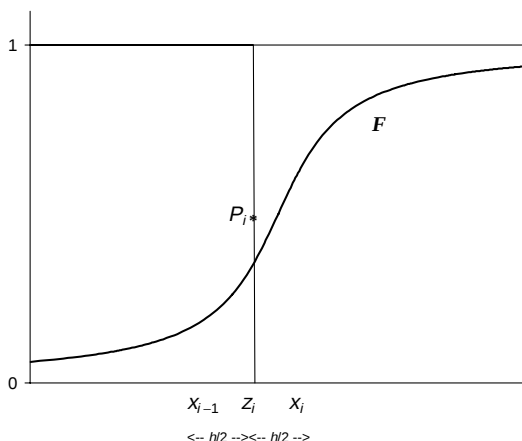


Figure 16.1 – Fonction de répartition.

Si les points P_i ne sont pas trop éloignés de la courbe Γ , on peut admettre que la loi suivie par les observations est voisine de la loi L .

Pour estimer la distance des points P_i à une courbe théorique Γ , on cherche une transformation mathématique simple, permettant de représenter la fonction de répartition par une droite. Cette transformation ou anamorphose existe pour la plupart des lois de probabilité.

■ Loi normale

Soit U la variable aléatoire centrée réduite associée à la variable normale X .

D'une part, il existe une bijection entre les valeurs de F comprises entre 0 et 1 et les valeurs de U comprises entre $-\infty$ et $+\infty$.

D'autre part, il existe une relation linéaire simple entre les variables X et U :

$$U = \frac{X - m}{\sigma}$$

La transformée de la fonction de répartition dans le plan (U, X) est une droite de pente $1/\sigma$, appelée *droite de Henry* (elle a été introduite par le Commandant P. Henry en 1894, dans les cours de l'École d'artillerie de Fontainebleau).

On utilise un papier spécial que l'on trouve dans le commerce dit *gausso-arithmétique* ou que l'on peut tracer facilement à l'ordinateur. Il suffit de graduer l'axe des ordonnées selon les valeurs de F mais proportionnellement aux valeurs de U , par exemple :

$U = 0$	$F(0) = 0,5$		
$U = 1$	$F(1) = 0,8417$	$U = -1$	$F(-1) = 0,1583$
$U = 2$	$F(2) = 0,9772$	$U = -2$	$F(-2) = 0,0228$

On répète ce procédé pour toutes les valeurs de la variable U . On peut, de la même façon choisir les valeurs de F et en déduire les valeurs de U .

Pour vérifier si un échantillon est extrait d'une population normale, on porte :

- en abscisses, les valeurs des observations, c'est-à-dire les *limites supérieures des classes*,
- en ordonnées, les fréquences cumulées correspondantes.

Si les points obtenus sont sensiblement alignés, on peut accepter comme distribution théorique une loi normale.

L'intersection avec la droite $U = 0$ ($F = 0,50$) donne la valeur de l'espérance mathématique $E(X) = m$.

Quant à la valeur de σ , elle peut être obtenue de deux façons :

- si $U = 1$, ($F = 0,8415$), $x_i - m = \sigma$.
- si $U = -1$, ($F = 0,1585$), $x_i - m = -\sigma$.

Ces deux valeurs sont indiquées sur le papier gaussio-arithmétique.

■ Loi exponentielle

On suppose que la durée de vie X d'un composant suit une loi exponentielle de fonction répartition F :

$$\Pr(X > x) = e^{-\lambda x} = 1 - F(x)$$

d'où : $\text{Ln}[1 - F(x)] = -\lambda x$

Si on dispose d'un échantillon de taille n , on porte :

- en abscisses, les temps x_i de fonctionnement,
- en ordonnées, les pourcentages de « survivants » au temps x_i , en utilisant une échelle logarithmique.

En pratique, on ordonne les temps x_i par valeurs croissantes, et on prend pour ordonnées correspondantes, les valeurs : $\text{Ln} \left(1 - \frac{1-i}{n} \right)$ $1 < i \leq n$

Si l'échantillon est représentatif de la population, les points sont pratiquement alignés. Pour estimer la valeur de λ , c'est-à-dire la pente de la droite, on remarque que si $x = 1$, $\text{Ln}[1 - F(x)] = -\lambda$. L'intersection avec la droite $x = 1$ donne alors une estimation graphique du paramètre.

■ Loi de Weibull

Une variable aléatoire réelle T suit une loi de Weibull si sa fonction de répartition F et sa densité f sont :

$$F(t) = 0 \quad \forall t < \gamma$$

$$F(t) = 1 - \exp \left[- \left(\frac{t - \gamma}{\eta} \right)^\beta \right] \quad \forall t \geq \gamma$$

$$f(t) = 0 \quad \forall t < \gamma$$

$$f(t) = \frac{\beta}{\eta} \left(\frac{t - \gamma}{\eta} \right)^{\beta-1} \exp \left[- \left(\frac{t - \gamma}{\eta} \right)^\beta \right] \quad \forall t \geq \gamma$$

Le paramètre de position γ peut être pris égal à 0 (simple translation sur t).
D'où la fonction de répartition simplifiée :

$$F(t) = 1 - \exp \left[- \left(\frac{t}{\eta} \right)^\beta \right]$$

Les constantes β et η peuvent être estimées par la méthode du maximum de vraisemblance (chapitre 13, exemple 13.10). Une estimation graphique de ces constantes est obtenue en utilisant un papier spécial à échelle fonctionnelle, le *papier d'Allan Plait*.

La transformation mathématique donnant pour représentation de la fonction de répartition une droite est la suivante :

$$X = \text{Ln } t \quad Y = \text{Ln} \left[\text{Ln} \left(\frac{1}{1 - F(t)} \right) \right]$$

$$X = \text{Ln } t \quad Y = \beta (X - \text{Ln } \eta)$$

La pente de la droite donne la valeur de β et l'intersection de la *droite empirique* avec la droite $Y = 0$ donne la valeur de η (la valeur $Y = 0$ correspond à $F(t) = 0,632$).

Sur le graphique, on porte les points de coordonnées :

$$\text{abscisse } \text{Ln } t_i \quad \text{ordonnée } \text{Ln} \left[-\text{Ln} \left(1 - \frac{i-1}{n} \right) \right]$$

Un rapporteur permet de lire la valeur de β , elle est comprise entre 0 et 4.

Une échelle verticale donne, pour les valeurs de β , les valeurs de $\Gamma(1 + 1/\beta)$; elle permet de calculer une estimation de l'espérance mathématique qui est égale à :

$$E(T) = \eta \Gamma \left(1 + \frac{1}{\beta} \right)$$

Remarques sur la méthode : ajustement graphique

- Dans chaque classe, y compris les classes extrêmes, on doit avoir au moins cinq observations ; si cette condition n'est pas remplie, on regroupe certaines classes.
- Pour calculer la fonction de répartition empirique, on peut utiliser les formules d'approximation données dans le chapitre 11, paragraphe 11.1.2.

- En fait, ce procédé n'est pas un test, mais une méthode rapide et simple pour voir si une distribution observée est compatible avec une loi que l'on s'est fixée à l'avance. Elle permet aussi de comparer les valeurs lues sur le graphique pour les paramètres, aux estimations calculées sur l'échantillon.

16.1.2 Ajustement analytique et principaux tests

Un test d'ajustement permet de juger si une hypothèse concernant une loi de probabilité, c'est-à-dire une loi théorique, telle que loi binomiale, exponentielle..., est compatible avec la réalisation d'un échantillon de taille n d'une variable aléatoire X .

Pour mettre en œuvre un test d'ajustement, il faut :

- prélever un échantillon suffisamment important de la population étudiée,
- classer les observations par ordre croissant dans le cas d'une variable aléatoire discrète, les répartir en classes (par ordre croissant) pour une variable aléatoire continue, d'égale amplitude ou d'égale probabilité,
- définir une variable de décision D permettant de mesurer les écarts entre la distribution théorique F et la distribution empirique F^* de l'échantillon.

Pour vérifier la concordance des deux distributions, on doit :

- définir les hypothèses H_0 et H_1 ,
 - H_0 : les observations suivent une distribution théorique spécifiée $F = F_0$,
 - H_1 : les observations ne suivent pas la distribution théorique spécifiée $F \neq F_0$.
- accepter un risque de première espèce α de refuser l'hypothèse H_0 alors qu'elle est vraie,
- calculer la valeur d de la variable de décision D (à partir des valeurs données par l'échantillon),
- énoncer une règle de décision :
 - on rejette l'hypothèse H_0 si la valeur calculée d est supérieure à une valeur d_0 n'ayant qu'une probabilité α d'être dépassée par la variable D ,
 - sinon, on garde l'hypothèse H_0 et on considère que la distribution théorique spécifiée peut décrire le phénomène étudié, c'est-à-dire $F = F_0$.

■ Test du chi-deux

Le test du chi-deux utilise des propriétés de la loi multinomiale (chapitre 5, paragraphe 5.5). Deux cas sont à distinguer :

- 1) la fonction de répartition F est entièrement spécifiée, en particulier, les paramètres sont connus,
- 2) on connaît seulement la forme de la loi de distribution, les paramètres de la fonction de répartition F sont estimés à partir d'un échantillon.

Soit X la variable aléatoire parente, de fonction de répartition F . On considère une partition du domaine de définition en r intervalles $I_1 \dots I_r$, d'égale étendue ou non.

Pour chaque intervalle I_i , on considère l'ensemble E_i tel que :

$$E_i = \{\omega : X(\omega) \in I_i\} \quad p_i = \Pr(E_i)$$

np_i est égal à la *fréquence (absolue) théorique* espérée dans la classe I_i que l'on compare à la *fréquence observée* N_i dans cette même classe I_i .

□ Variable de décision

C'est la statistique :

$$D^2 = \sum_{i=1}^r \frac{(N_i - np_i)^2}{np_i}$$

Si l'hypothèse H_0 est vraie (concordance acceptable entre la distribution théorique et les valeurs observées), cette quantité ne peut pas être trop grande. En fait, Pearson a montré que la statistique D^2 suit une loi du chi-deux, à ν degrés de liberté quelle que soit la loi considérée, quand le nombre n d'observations tend vers l'infini. Le nombre ν de degrés de liberté est égal à :

- $(r - 1)$ si la distribution théorique est entièrement déterminée, aucun paramètre n'ayant été estimé,
- $(r - 1 - k)$ si k paramètres ont été estimés à partir des observations, pour définir complètement la distribution.

□ Règle de décision

On rejette l'hypothèse H_0 si la valeur de la statistique D^2 obtenue à partir de l'échantillon est supérieure à une valeur n'ayant qu'une probabilité α d'être dépassée par la variable χ^2 considérée. Sinon, on garde l'hypothèse H_0 et on

considère que la distribution théorique spécifiée est acceptable, c'est-à-dire $F = F_0$.

Remarques

- La distribution limite de la statistique D^2 est indépendante de la loi F , ce test peut donc être utilisé dans de nombreuses situations.
- Les effectifs de chaque classe doivent être supérieurs à cinq. Si cette condition n'est pas vérifiée, on regroupe les classes d'effectifs trop faibles.

Exemple 16.1

Il existe de nombreux procédés de générations de nombres au hasard (aléatoires) sur ordinateur. Pour chacun de ces générateurs, on peut se poser la question de savoir si la suite générée est bien aléatoire. Un générateur a donné 1 000 nombres compris entre 0 et 1, leur répartition est la suivante :

Tableau 16.1 – Exemples de suite de nombres aléatoires.

x	0 à 0,09	0,10 à 0,19	0,20 à 0,29	0,30 à 0,39	0,40 à 0,49
n	113	73	125	115	90

x	0,50 à 0,59	0,60 à 0,69	0,70 à 0,79	0,80 à 0,89	0,90 à 0,99
n	101	95	93	110	85

On a obtenu, par exemple, 113 nombres entre 0 et 0,09. Si la répartition était uniforme, on aurait dû obtenir 100 nombres dans chaque classe. Un test du chi-deux permet de garder ou de rejeter cette hypothèse :

$$D^2 = \frac{(113 - 100)^2}{100} + \frac{(73 - 100)^2}{100} + \frac{(125 - 100)^2}{100} + \frac{(115 - 100)^2}{100} + \frac{(90 - 100)^2}{100} \\ + \frac{(101 - 100)^2}{100} + \frac{(95 - 100)^2}{100} + \frac{(93 - 100)^2}{100} + \frac{(110 - 100)^2}{100} + \frac{(85 - 100)^2}{100} \\ D^2 = 22,48$$

Sous l'hypothèse *loi uniforme*, la variable D^2 suit une loi du chi-deux à $(10 - 1) = 9$ degrés de liberté, aucun paramètre n'ayant été estimé :

$$\Pr(\chi^2(9) > 16,9) = 0,05 \quad \Pr(\chi^2(9) > 19) = 0,025$$

On doit donc rejeter l'hypothèse *loi uniforme*, car la valeur 22,48 a une probabilité inférieure à 0,025 de se réaliser.

Le générateur n'est pas à l'abri de toute critique.

■ Test de Kolmogorov-Smirnov (1933, 1939)

On suppose que la fonction de répartition F de la variable aléatoire X est continue et strictement croissante. Soit F^* la fonction de répartition empirique d'un échantillon de taille n issu de cette population.

□ Variable de décision

La variable de décision est la variable aléatoire D_n définie par :

$$D_n = \sup_{x \in \mathbb{R}} |F^*(x) - F(x)|$$

Glivenko et Kolmogorov ont démontré que la fonction $K_n(y)$ définie par :

$$K_n(y) = \Pr(\sqrt{n} D_n < y)$$

converge, quand n tend vers l'infini, vers une fonction $K(y)$:

$$K(y) = 0 \quad y \leq 0$$

$$K(y) = \sum_{k=-\infty}^{\infty} (-1)^k \exp\left(-2k^2 y^2\right) \quad y > 0$$

Des tables donnent les valeurs de cette fonction K .

□ Règle de décision

On rejette l'hypothèse H_0 si la valeur de la statistique D_n , obtenue à partir de l'échantillon, est supérieure à une valeur $d(n)$ n'ayant qu'une probabilité α d'être dépassée.

Sinon, on garde l'hypothèse H_0 et on considère que la distribution théorique spécifiée est acceptable, c'est-à-dire $F = F_0$.

Remarques

Le test de Kolmogorov-Smirnov est préférable au test du chi-deux pour des variables continues. En effet, la variable aléatoire de décision D_n utilise l'échantillon tel qu'il se présente, en revanche, le test du chi-deux appauvrit l'information en regroupant les données par classes et en assimilant les données d'une classe à la valeur centrale.

Exemple 16.2

On a relevé, dans une entreprise, le nombre de personnes qui ne se sont pas présentées au travail pendant une période de 200 jours.

Tableau 16.2 – Répartition des absences en fonction du nombre de jours.

Nombre d'absents	0	1	2	3	4	5	6 et plus
Nombre de jours	18	30	45	44	34	24	5

Peut-on admettre que le nombre de personnes absentes en une journée suit une loi de Poisson de paramètre $\lambda = 3$ (taux moyen d'absentéisme par jour) ?

Soit F^* la fonction de répartition empirique et F la fonction de répartition théorique (loi de Poisson de paramètre 3).

Tableau 16.3 – Calcul de la variable de décision.

N	0	1	2	3	4	5	≥ 6
200 $F^*(n)$	18	48	93	137	171	195	200
200 $F(n)$	9,96	39,83	84,64	129,45	163,05	183,22	200
200 $ F^* - F $	8,04	8,17	8,36	7,55	7,95	11,78	0

$$d_n = \text{Sup}_n |F^* - F| = 11,78/\sqrt{200} = 0,833.$$

Au risque 0,05 et pour un test bilatéral, le seuil critique est 1,358. On ne peut pas rejeter l'hypothèse d'une loi de Poisson de paramètre $\lambda = 3$.

■ Test de Cramer-Von-Mises

On considère la statistique $n\omega^2$ définie par :

$$n\omega^2 = \int_{-\infty}^{+\infty} [F(x) - F^*(x)] dF(x)$$

Il existe des tables pour la loi de cette variable aléatoire, loi indépendante de F . Elle est utilisée pour évaluer l'écart entre une distribution empirique et une distribution théorique. Si les valeurs x_i de l'échantillon sont ordonnées par valeurs croissantes, on démontre que :

$$n\omega^2 = \frac{1}{12n} + \sum_{i=1}^n \left[\frac{2i-1}{2n} - F(x_i) \right]^2$$

On rejette l'hypothèse H_0 si la valeur de la statistique $n\omega^2$ donnée par l'échantillon est supérieure à une valeur n'ayant qu'une probabilité α d'être dépassée. Pour $\alpha = 0,05$, on rejette l'hypothèse H_0 si $n\omega^2 > 0,46136$.

■ Tests de normalité et d'exponentialité

Si les paramètres des lois de distribution ne sont pas connus mais estimés, on peut utiliser les résultats empiriques suivants (Biometrika Tables).

□ Test de normalité

– Hypothèse H_0 . Loi normale $N(m; \sigma)$

m est estimée par $\bar{x}$ et l'écart-type σ par : $\sqrt{\frac{1}{n-1} \sum_i (x_i - \bar{x})^2}$

– Règle de décision. On rejette l'hypothèse H_0 :

$$\text{au seuil } \alpha = 0,05 \quad \text{si} \quad \left(\sqrt{n} + \frac{0,85}{\sqrt{n}} - 0,01 \right) D_n > 0,895$$

$$\text{ou si} \quad \left(1 + \frac{0,5}{n} \right) n \omega^2 > 0,126$$

$$\text{au seuil } \alpha = 0,01 \quad \text{si} \quad \left(\sqrt{n} + \frac{0,85}{\sqrt{n}} - 0,01 \right) D_n > 1,035$$

$$\text{ou si} \quad \left(1 + \frac{0,5}{n} \right) n \omega^2 > 0,178$$

□ Test d'exponentialité

– Hypothèse H_0 . Loi exponentielle de densité $f(x) = \frac{1}{\theta} \exp\left(-\frac{x}{\theta}\right)$

θ est estimée par $\bar{x}$.

– Règle de décision. On rejette l'hypothèse H_0 :

$$\text{au seuil } \alpha = 0,05 \quad \text{si} \quad \left(\sqrt{n} + \frac{0,5}{\sqrt{n}} + 0,26 \right) \left(D_n - \frac{0,2}{n} \right) > 1,094$$

$$\text{ou si} \quad \left(1 + \frac{0,16}{n} \right) n \omega^2 > 0,224$$

$$\text{au seuil } \alpha = 0,01 \quad \text{si} \quad \left(\sqrt{n} + \frac{0,5}{\sqrt{n}} + 0,26 \right) \left(D_n - \frac{0,2}{n} \right) > 1,308$$

$$\text{ou si} \quad \left(1 + \frac{0,16}{n} \right) n \omega^2 > 0,337$$

16.2 Tests de comparaison d'échantillons

16.2.1 Tests paramétriques de comparaison des moyennes de deux échantillons

On considère deux échantillons aléatoires de tailles n_1 et n_2 , prélevés indépendamment l'un de l'autre et on pose la question :

Sont-ils, ou non, issus de la même population ?

Soient X_1 la variable aléatoire parente et F_1 la fonction de répartition de la population dont est issu le premier échantillon, X_2 et F_2 , les mêmes caractéristiques pour le second. Le test correct est le suivant :

$$H_0 : F_1(x) = F_2(x) \quad \text{et} \quad H_1 : F_1(x) \neq F_2(x)$$

mais, il est beaucoup trop vague. Dans la pratique, on traite le problème plus général suivant : *comparaison des moyennes m_1 et m_2 de deux populations connaissant les estimations données par deux échantillons indépendants de tailles n_1 et n_2 .*

Pour caractériser la variable de décision $\bar{D} = \bar{X}_1 - \bar{X}_2$, il faut connaître la forme de la loi suivie par cette variable, son espérance mathématique et sa variance.

Selon les données, on distingue trois situations différentes.

■ Cas 1 : comparaison des moyennes de deux échantillons gaussiens indépendants, les variances étant connues

Soient $N(m_i ; \sigma_i)$, les lois suivies par les deux populations ($i = 1$ ou 2). La variable aléatoire $\bar{D} = \bar{X}_1 - \bar{X}_2$ suit alors la loi normale :

$$N(m_1 - m_2 ; \sigma_{\bar{D}}) \quad \text{où} \quad \sigma_{\bar{D}} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

On considère la variable aléatoire centrée réduite, U :

$$U = \frac{\bar{D} - (m_1 - m_2)}{\sigma_{\bar{D}}}$$

Hypothèses : $H_0 : m_1 = m_2$ et $H_1 : m_1 \neq m_2$

Sous l'hypothèse H_0 et compte tenu de l'hypothèse H_1 ($m_1 \neq m_2$), la région critique est de la forme $|U| > k$.

Exemple 16.3

On dispose de deux échantillons de tubes, construits suivant deux procédés de fabrication A et B. On a mesuré les diamètres de ces tubes et on a trouvé (en millimètres) :

Procédé A 52,80 52,90 51,90 50,90 53,40

Procédé B 52,10 51,30 51,50 51,10

On suppose que les diamètres sont distribués suivant une loi normale et que les écarts-types sont égaux à : $\sigma_A = 1$ mm et $\sigma_B = 0,45$ mm.

Peut-on affirmer au niveau 5 % qu'il y a une différence significative entre les procédés de fabrication A et B ?

Soient X_A et X_B , les variables aléatoires « diamètres des tubes fabriqués suivant les procédés A et B ».

– Loi de X_A : $N(m_A ; 1)$ Loi de $\bar{X}_A$: $N(m_A ; 1/\sqrt{5})$,

– Loi de X_B : $N(m_B ; 0,45)$ Loi de $\bar{X}_B$: $N(m_B ; 0,45/\sqrt{4})$,

$$-\bar{D} = \bar{X}_A - \bar{X}_B \quad m_{\bar{D}} = m_A - m_B \quad \sigma_{\bar{D}}^2 = \frac{1}{5} + \frac{(0,45)^2}{4} = 0,2506$$

Loi de $\bar{D}$: $N(m_{\bar{D}} ; \sqrt{0,2506} \cong 0,5)$

On veut tester $m_{\bar{D}} = m_A - m_B = 0$ avec un seuil critique égal à 5 %.

$$\Pr(-1,96 < U < 1,96) = 0,95$$

Avec les données apportées par les échantillons, on obtient pour la moyenne des deux échantillons $\bar{x}_A = 52,38$ et $\bar{x}_B = 51,50$.

$$\bar{d} = 52,38 - 51,50 = 0,88$$

$$\frac{\bar{d}}{\sigma_{\bar{d}}} = \frac{0,88}{0,50} = 1,76$$

On ne peut rejeter l'hypothèse d'égalité des moyennes.

■ **Cas 2 : comparaison des moyennes de deux échantillons gaussiens indépendants, les variances n'étant pas connues mais supposées égales**

Hypothèses :

$$H_0 : m_1 = m_2 \quad \sigma_1 = \sigma_2$$

$$H_1 : m_1 \neq m_2 \quad \sigma_1 \neq \sigma_2$$

On doit vérifier *dans l'ordre suivant* et à partir des estimations données par les échantillons :

- l'égalité des variances,
- l'égalité des moyennes, si les variances sont égales.

□ Test de l'égalité des variances ou test de Fisher-Snedecor

Hypothèses :

$$H_0 : \sigma_1 = \sigma_2$$

$$H_1 : \sigma_1 \neq \sigma_2$$

Comme les populations sont gaussiennes, on sait que (chapitre 10, paragraphe 10.7.4) :

$$\frac{n_1 S_1^2}{(n_1 - 1) \sigma_1^2} \times \frac{(n_2 - 1) \sigma_2^2}{n_2 S_2^2} = F(n_1 - 1 ; n_2 - 1)$$

Sous l'hypothèse H_0 , on obtient :

$$\frac{n_1 S_1^2}{(n_1 - 1)} \times \frac{(n_2 - 1)}{n_2 S_2^2} = F(n_1 - 1 ; n_2 - 1)$$

que l'on peut écrire, en introduisant les estimateurs sans biais des deux variances :

$$\frac{n_i S_i^2}{(n_i - 1)} = \hat{\sigma}_i^2 \quad i = 1 \text{ ou } 2 \quad \frac{\hat{\sigma}_1^2}{\hat{\sigma}_2^2} = F(n_1 - 1 ; n_2 - 1)$$

Donc, sous l'hypothèse H_0 , le rapport des estimateurs sans biais des deux variances est une variable aléatoire de Fisher.

Les conclusions du test sont obtenues en calculant le rapport :

$$F = \frac{n_1 S_1^2}{(n_1 - 1)} \times \frac{(n_2 - 1)}{n_2 S_2^2}$$

pour les valeurs données par les échantillons.

L'hypothèse alternative étant $H_1 : \sigma_1 \neq \sigma_2$, la règle de décision est :

Rejeter H_0 si $F < F_{1-\alpha/2}(n_1 - 1, n_2 - 1)$ ou $F > F_{\alpha/2}(n_1 - 1, n_2 - 1)$
 α étant le seuil critique.

Remarque

Pour déterminer la région critique, on doit toujours avoir une valeur du rapport F supérieure à 1.

□ **Test de comparaison de deux moyennes ou test des espérances de Student**

Si le test de Fisher-Snedecor a permis de conclure à l'égalité des variances des deux populations, la variable de décision $\overline{D} = \overline{X}_1 - \overline{X}_2$ suit la loi normale de paramètres :

$$E(\overline{D}) = m_1 - m_2 \quad \text{Var}(\overline{D}) = \sigma^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)$$

La variable aléatoire :

$$\frac{(\overline{X}_1 - \overline{X}_2) - (m_1 - m_2)}{\sqrt{\sum_1 (X_i - \overline{X})^2 + \sum_2 (X_i - \overline{X})^2}} \times \frac{\sqrt{n_1 + n_2 - 2}}{\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

suit une loi de Student à $(n_1 + n_2 - 2)$ degrés de liberté (chapitre 10, paragraphe 10.7.5).

Sous l'hypothèse H_1 ($m_1 \neq m_2$), la région critique est de la forme $|T| > k$.

Remarque

Il est indispensable de tester d'abord l'égalité des variances pour appliquer le test de Student.

Exemple 16.4

On reprend les données numériques de l'exemple 16.3, en supposant que les variances sont inconnues.

– Estimations non biaisées des variances $\hat{\sigma}_A^2 = 0,977$ $\hat{\sigma}_B^2 = 0,187$.

D'où : $\frac{\hat{\sigma}_A^2}{\hat{\sigma}_B^2} = 5,2246$

$\Pr(0,10 < F(4; 3) < 15,1) = 0,95$

Donc, au seuil critique égal à 5 %, on ne peut pas refuser l'hypothèse de l'égalité des variances des deux populations.

– Test de Student de comparaison des moyennes :

La variable aléatoire

$$Z = \frac{\bar{X}_A - \bar{X}_B - (m_A - m_B)}{\sqrt{\sum_A (x_i - \bar{x}_A)^2 + \sum_B (x_i - \bar{x}_B)^2}} \times \frac{\sqrt{n_A + n_B - 2}}{\sqrt{\frac{1}{n_A} + \frac{1}{n_B}}}$$

suit une loi de Student à $n_A + n_B - 2 = 5 + 4 - 2 = 7$ degrés de liberté.

$$\Pr(-2,365 < t(7) < 2,365) = 0,95$$

Si $m_A = m_B$, la variable Z prend la valeur 1,642. On ne peut donc pas rejeter l'hypothèse d'égalité des moyennes.

■ Cas 3 : comparaison des moyennes de deux échantillons non gaussiens indépendants

Si les populations ne sont pas gaussiennes, on ne peut pas appliquer le test des variances de Fisher. Cependant, si les effectifs des échantillons sont assez grands, supérieurs à 30 environ, on peut tester l'égalité des moyennes, que les variances soient égales ou non, avec la formule de Student. Le test de Student est un *test robuste*, il est insensible à une modification des hypothèses de base.

16.2.2 Tests non paramétriques de comparaison

Le problème consiste à décider si deux échantillons de tailles n_1 et n_2 sont issus ou non d'une même population de fonction de répartition F . Différents tests sont proposés.

■ Test de Smirnov (analogue au test d'ajustement de Kolmogorov-Smirnov)

Soit F^* et G^* les fonctions de répartition empiriques des deux échantillons.

Hypothèses :

$$H_0 : F(x) = G(x)$$

$$H_1 : F(x) \neq G(x)$$

On montre que :

$$\Pr \left(\sqrt{\frac{n_1 n_2}{n_1 + n_2}} \sup |F^*(x) - G^*(x)| < y \right) \rightarrow K(y)$$

On rejette l'hypothèse H_0 si la valeur de la statistique :

$$\sqrt{\frac{n_1 n_2}{n_1 + n_2}} \sup |F^*(x) - G^*(x)|$$

calculée à partir des échantillons est supérieure à une valeur qui a une faible probabilité d'être dépassée.

La fonction $K(y)$ a été définie dans le paragraphe « Test de Kolmogorov-Smirnov ».

■ Test de Wilcoxon (1945)

Si deux échantillons (x_i) $i \in [1, n]$ et (y_j) $j \in [1, m]$, sont issus de la même population, on doit obtenir, en mélangeant les observations et en les classant par valeurs croissantes, une population homogène.

Après avoir ordonné les suites, on désigne par U le nombre obtenu en comptant le nombre de couples (x_i, y_j) tel que :

- $x_i > y_j$ si les variables sont quantitatives,
- le rang de x_i est supérieur au rang de y_j si les variables sont qualitatives.

Le nombre U varie de 0 (si tous les x_i sont inférieurs à tous les y_j) à nm dans le cas contraire. Si les deux échantillons sont issus de la même population :

$$E(U) = \frac{nm}{2} \quad \text{Var}(U) = \frac{nm(n+m+1)}{12}$$

Si les effectifs n et m des échantillons sont supérieurs à 8, la loi de U tend asymptotiquement vers une loi normale ayant pour paramètres les valeurs $E(U)$ et $\text{Var}(U)$ définies précédemment.

On rejettera l'hypothèse H_0 (échantillons issus d'une même population) si la valeur observée de U est trop grande.

Nous proposons un calcul plus rapide. Après avoir classé les observations comme on l'a indiqué précédemment, on calcule la somme des rangs des individus d'un des groupes, le groupe X par exemple. Soit W_x cette somme.

$$W_x = U + \frac{n(n+1)}{2}$$

$$E(W_x) = \frac{n(n+m+1)}{2} \quad \text{Var}(W_x) = \frac{nm(n+m+1)}{12}$$

Si les effectifs n et m des échantillons sont supérieurs à 8, la loi de W_x tend asymptotiquement vers une loi normale dont les paramètres sont $E(W_x)$ et $\text{Var}(W_x)$.

On rejette l'hypothèse H_0 , « échantillons issus d'une même population », si la valeur calculée de W_x est trop grande pour le seuil de confiance choisi.

Remarque

On aurait pu appliquer le test de Fisher d'égalité des variances, puis le test de Student de comparaison des moyennes, si les conditions d'application de ces tests étaient vérifiées (population gaussienne ou échantillons de tailles suffisantes).

16.2.3 Test de comparaison de deux échantillons appariés

On considère un échantillon d'individus soumis à deux mesures successives d'une même variable. On pose la question :

Les deux séries de valeurs sont-elles semblables ?

Soient X et Y les variables parentes associées à chaque série, ces variables sont indépendantes et suivent des lois normales.

On teste seulement l'égalité des moyennes $m_x = m_y$ avec la variable aléatoire $X - Y$ qui suit une loi normale d'espérance $m_x - m_y$.

Hypothèses :

$$H_0 : m_x = m_y$$

$$H_1 : m_x \neq m_y$$

Comme on ne connaît pas, en général, la variance σ^2 , on fait un test de Student sur la moyenne des différences :

$$T(n-1) = \frac{\bar{D}}{s_d / \sqrt{(n-1)}}$$

On rejette H_0 si $|T| > k$, la valeur critique k dépend du seuil α choisi.

16.2.4 Test de comparaison de plusieurs échantillons

On dispose de k échantillons, décrits par une variable aléatoire qualitative prenant r modalités. Le tableau des observations est le suivant :

Tableau 16.4 – Tableau des observations.

	Modalité 1	Modalité j	Modalité r	Total
Échantillon 1	n_{11}	n_{1j}	n_{1r}	$n_{1.}$
Échantillon i	n_{i1}	n_{ij}	n_{ir}	$n_{i.}$
Échantillon k	n_{k1}	n_{kj}	n_{kr}	$n_{k.}$
Total	$n_{.1}$	$n_{.j}$	$n_{.r}$	N

n_{ij} effectif de l'échantillon i prenant la modalité j

$n_{i.}$ effectif total de l'échantillon i , $n_{i.} = \sum_{j=1}^r n_{ij}$

$n_{.j}$ nombre total des individus possédant le caractère j , $n_{.j} = \sum_{i=1}^k n_{ij}$

N nombre total d'observations, $N = \sum_{i=1}^k \sum_{j=1}^r n_{ij} = \sum_{j=1}^r n_{.j} = \sum_{i=1}^k n_{i.}$

Hypothèses :

H_0 : les échantillons sont issus de la même population,

H_1 : les échantillons sont issus de populations différentes.

Sous l'hypothèse H_0 , on désigne par p_j la probabilité théorique, mais inconnue, de posséder la modalité j . Si cette probabilité était connue, il serait possible de comparer les effectifs observés n_{ij} aux effectifs espérés $p_j n_{i.}$ pour toutes les valeurs des indices i et j . La statistique :

$$d^2 = \sum_{i=1}^k \sum_{j=1}^r \frac{(n_{ij} - p_j n_{i.})^2}{p_j n_{i.}}$$

est une mesure de la distance entre une distribution théorique et la distribution observée. Sous l'hypothèse H_0 , cette variable est la réalisation d'une variable aléatoire D^2 suivant une *loi du chi-deux* à v degrés de liberté. Le tableau 16.4 contient kr termes, liés par k relations. La variable aléatoire D^2 est donc une variable χ^2 à $(kr - k)$ degrés de liberté.

Cependant, les probabilités p_j ne sont pas connues, mais estimées par les quantités :

$$\hat{p}_j = \frac{n_{.j}}{N}$$

Il en résulte une nouvelle expression pour la statistique d^2 :

$$d^2 = N \left(\sum_{i=1}^k \sum_{j=1}^r \frac{n_{ij}^2}{n_{i.} n_{.j}} - 1 \right)$$

obtenue en remplaçant les probabilités p_j par leurs estimations.

Avec les fréquences relatives au lieu des fréquences absolues, on obtient :

$$d^2 = N \left(\sum_{i=1}^k \sum_{j=1}^r \frac{f_{ij}^2}{f_{i.} f_{.j}} - 1 \right)$$

Les r estimations des probabilités p_j sont liées par une relation (leur somme est égale à 1). En fait, on a estimé $(r-1)$ paramètres indépendants. La statistique D^2 est donc une variable aléatoire χ^2 à $(kr - k - r + 1) = (k-1)(r-1)$ degrés de liberté.

On rejette l'hypothèse H_0 si la valeur observée d^2 est trop grande pour un seuil α donné.

16.2.5 Test de comparaison de pourcentages

On considère des échantillons de *grandes tailles*.

Soient n_1 et n_2 les tailles de ces échantillons, f_1 et f_2 les pourcentages des individus présentant un certain caractère dans chaque échantillon et soient p_1 et p_2 les probabilités correspondantes.

On veut savoir si les probabilités p_1 et p_2 diffèrent significativement ou non à partir des pourcentages observés.

Hypothèses :

$$H_0 : p_1 = p_2 = p$$

$$H_1 : p_1 \neq p_2$$

Les échantillons étant de grande taille, les pourcentages observés f_1 et f_2 peuvent être considérés comme des réalisations de variables aléatoires F_1 et F_2 suivant des lois normales :

$$N \left(p; \sqrt{\frac{p(1-p)}{n_1}} \right) \quad N \left(p; \sqrt{\frac{p(1-p)}{n_2}} \right)$$

Leur différence $F_1 - F_2$ suit donc la loi normale :

$$N\left(0; \sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}\right)$$

On rejette l'hypothèse H_0 si (risque $\alpha = 0,05$ par exemple) :

$$|f_1 - f_2| > 1,96 \sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

Remarques

– La probabilité p étant inconnue, on prend comme valeur son estimation calculée à partir des deux échantillons :

$$\hat{p} = \frac{n_1 f_1 + n_2 f_2}{n_1 + n_2}$$

– On aurait pu utiliser le test précédent (paragraphe 16.2.4) du chi-deux dont le calcul est simple dans ce cas particulier. En notant a, b, c, d , les effectifs observés dans chaque échantillon et pour chaque modalité, le tableau 16.1 devient :

Tableau 16.5 – Comparaison de deux pourcentages, tableau des observations.

	Modalité 1	Modalité 2	Total
Échantillon 1	a	b	$a + b$
Échantillon 2	c	d	$c + d$
Total	$a + c$	$b + d$	$N = a + b + c + d$

Pour mettre en œuvre le test du chi-deux, on calcule la statistique D^2 qui est égale à :

$$D^2 = N \left[\frac{a^2}{(a+b)(a+c)} + \frac{b^2}{(b+a)(b+d)} + \frac{c^2}{(c+a)(c+d)} + \frac{d^2}{(d+b)(d+c)} - 1 \right]$$

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

La statistique D^2 suit une loi du chi-deux à un degré de liberté.

16.3 Analyse de la variance à simple entrée

16.3.1 Objet de cette étude

Le problème traité dans le paragraphe 16.2.4 est un cas particulier du problème plus général suivant : *comparaison des moyennes de plusieurs échantillons*.

Le procédé qui consiste à tester l'égalité des moyennes de chaque couple n'est pas satisfaisant. Il faut utiliser une procédure permettant de tester globalement l'ensemble de tous les échantillons : c'est la *théorie de l'analyse de la variance*.

Le but de cette théorie est d'étudier la variabilité d'un produit en fonction d'un ensemble de facteurs de production dont on peut contrôler systématiquement les modes d'intervention et dont on souhaite dissocier la part revenant à chaque facteur.

On distingue :

- l'*analyse de la variance à simple entrée* (étudiée dans ce paragraphe), un seul facteur est contrôlé, les autres facteurs étant regroupés sous le nom « facteurs non contrôlés »,
- l'analyse de la variance à double entrée, qui étudie l'action simultanée de deux facteurs contrôlés, chacun agissant individuellement avec une possibilité d'interaction entre les deux,
- l'analyse de la variance à entrées multiples, plusieurs facteurs contrôlés.

Ces deux derniers cas seront traités dans la partie « analyse multidimensionnelle ».

Le facteur contrôlé peut intervenir dans des conditions différant :

- soit par *leur nature* : variations quantitatives (par exemple, étude de la durée de vie d'ampoules électriques en fonction de la provenance des filaments),
- soit par *leur intensité* (étude de la durée de vie d'ampoules électriques en fonction de la pression du gaz de remplissage).

De plus, le facteur contrôlé peut être, soit à effets fixes, soit à effets aléatoires.

16.3.2 Série statistique des observations

On suppose que le facteur contrôlé prend k modalités A_j .

Tableau 16.6 – Tableau des observations.

Facteurs	A_1		A_i		A_k
	x_1^1		x_i^1		x_k^1
	x_1^j		x_i^j		x_k^j
	$x_1^{n_i}$		$x_i^{n_i}$		$x_k^{n_i}$
Moyenne	$\bar{x}_1$		$\bar{x}_i$		$\bar{x}_k$
Nombre d'observations	n_1		n_i		n_k

N nombre total d'observations.

Moyenne des observations pour la modalité i , $\bar{x}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} x_i^j$

Moyenne générale des observations, $\bar{x} = \frac{1}{N} \sum_{i=1}^k \sum_{j=1}^{n_i} x_i^j$

16.3.3 Mise en œuvre du test

On suppose que le facteur contrôlé agit sur les moyennes et n'agit pas sur les variances, ce qui en toute rigueur devrait être vérifié.

La loi de la variable aléatoire parente X_i est, pour toutes les valeurs de l'indice i , une loi normale $N(m_i; \sigma)$. Chaque observation s'écrit, en désignant par ξ une fluctuation aléatoire gaussienne :

$$x_i^j = m_i + \xi_i^j \quad \text{où} \quad E(\xi) = 0 \quad \text{Var}(\xi) = \sigma^2$$

Les hypothèses à tester sont :

$$H_0 \quad \forall i \quad m_i = m \quad H_1 \quad \exists i \text{ et } j \text{ tels que } m_i \neq m_j$$

Sous l'hypothèse H_0 , la population est homogène, le facteur contrôlé n'exerce donc aucune influence sur la production, on peut alors comparer toutes les observations à la moyenne générale $\bar{x}$.

16.3.4 Définition des variations

La statistique T ou *variation totale* est la somme des carrés des écarts par rapport à la moyenne générale, elle est définie par :

$$T = \sum_{i=1}^k \sum_{j=1}^{n_i} (x_i^j - \bar{x})^2$$

Le quotient $S^2 = \frac{T}{N}$ est la *variance totale*.

L'écart $X_i^j - \bar{X}$ peut s'écrire : $X_i^j - \bar{X} = (X_i^j - \bar{X}_i) + (\bar{X}_i - \bar{X})$

$(X_i^j - \bar{X}_i)$: écarts des observations par rapport à la moyenne pour chaque modalité du facteur contrôlé.

$(\bar{X}_i - \bar{X})$: écarts des différentes moyennes par rapport à la moyenne générale.

La statistique :

$$A = \sum_{i=1}^k n_i (\bar{X}_i - \bar{X})^2$$

est la *variation due au facteur contrôlé* (entre différents traitements, entre différents laboratoires...). Le quotient $S_A^2 = \frac{A}{N}$ est la *variance due au facteur contrôlé*.

La statistique :

$$R = \sum_{i=1}^k \sum_{j=1}^{n_i} (X_i^j - \bar{X}_i)^2$$

est la *variation résiduelle*. Le quotient $S_R^2 = \frac{R}{N}$ est la *variance résiduelle*.

Un calcul facile conduit au résultat suivant :

$$T = A + R \quad \text{ou} \quad S^2 = S_A^2 + S_R^2$$

La variance totale S^2 est égale à la somme de la variance des moyennes et de la moyenne des variances.

■ Étude de la statistique S_R^2

Par hypothèse, les variables aléatoires X_i suivent des lois normales $N(m_i; \sigma)$.
Donc, la statistique :

$$S_i^2 = \frac{1}{n_i} \sum_{j=1}^{n_i} (X_i^j - \bar{X}_i)^2$$

est telle que $\frac{n_i S_i^2}{\sigma^2}$ suit la loi du chi-deux $\chi^2 (n_i - 1)$.

La variance résiduelle qui est égale à :

$$S_R^2 = \frac{1}{N} \sum_{i=1}^k n_i S_i^2$$

est telle que $\frac{N S_R^2}{\sigma^2}$ suit la loi du chi-deux $\chi^2(N - k)$ (propriété d'additivité de la loi chi-deux). On en déduit que $\frac{N S_R^2}{N - k}$ est une estimation de la variance σ^2 à $(N - k)$ degrés de liberté.

■ Étude de la statistique S^2

Si l'hypothèse H_0 est vraie, les variables X_i suivent la même loi normale $N(m; \sigma)$. La statistique $\frac{N S^2}{\sigma^2}$ suit donc la loi du chi-deux $\chi^2(N - 1)$.

■ Étude de la statistique S_A^2

Cette statistique (voir sa définition) peut être considérée comme la variance de l'échantillon formé par les k moyennes $\bar{X}_i$ pondérées par les effectifs n_i . On en déduit que la statistique $\frac{N S_A^2}{k - 1}$ suit la loi du chi-deux $\chi^2(k - 1)$.

De l'équation de l'analyse de la variance et sous l'hypothèse H_0 , on déduit que les statistiques S_A^2 et S_R^2 sont indépendantes.

D'où le test : si l'hypothèse H_0 est vraie et d'après les définitions de S_A^2 et S_R^2 , on a :

$$\frac{S_A^2}{k - 1} \times \frac{N - k}{S_R^2} = F(k - 1; N - k)$$

On rejette l'hypothèse H_0 si :

$$\frac{S_A^2}{k - 1} \times \frac{N - k}{S_R^2} > f_\alpha$$

la valeur critique f_α , lue sur les tables de Fisher, dépend du seuil α choisi. Ce résultat signifie que le facteur contrôlé a une *influence significative*.

■ Calcul rapide des différentes statistiques

Les résultats suivants sont faciles à démontrer après développement des différents termes carrés :

$$N S^2 = T = \sum_{i=1}^k \sum_{j=1}^{n_i} (X_i^j)^2 - \frac{1}{N} \left(\sum_{i=1}^k \sum_{j=1}^{n_i} X_i^j \right)^2 \quad \text{degré de liberté } (N - 1)$$

– Le calcul de S^2 a été donné au chapitre 1 : moyenne des carrés moins carré de la moyenne.

– La quantité $\Delta = \frac{1}{N} \left(\sum_{i=1}^k \sum_{j=1}^{n_i} X_i^j \right)^2$ est un terme correctif.

$$N S_A^2 = \sum_{i=1}^k \frac{1}{n_i} \left(\sum_{j=1}^{n_i} X_i^j \right)^2 - \Delta \quad \text{degré de liberté } (k - 1)$$

$$N S_R^2 = N S^2 - N S_A^2 \quad \text{degré de liberté } (N - k)$$

Tous ces résultats sont résumés dans le tableau d'analyse de la variance.

Tableau 16.7 – Analyse de la variance.

Variation	Somme des carrés	Degré de liberté	Quotient
Variation due au facteur	$A = N S_A^2$	$k - 1$	$v_A = N S_A^2 / (k - 1)$
Variation résiduelle	$R = N S_R^2$	$N - k$	$v_R = N S_R^2 / (N - k)$
Variation totale	$T = N S^2$	$N - 1$	

■ Conclusion

On choisit un seuil de confiance α .

On garde l'hypothèse H_0 , le facteur contrôlé n'a pas d'influence, donc la population est homogène, si :

$$\frac{v_A}{v_R} < F_\alpha(k - 1; N - k)$$

Dans ce cas, on prend :

- comme estimation de m , la moyenne générale des observations,
- comme estimation de la variance, le quotient v_R .

On peut donner un intervalle de confiance pour m car la variable aléatoire :

$$\frac{m - \bar{x}}{\sqrt{N S_R^2}} \sqrt{N - k}$$

est une *variable de Student* à $(N - k)$ degrés de liberté.

On refuse l'hypothèse H_0 , le facteur exerce une influence et donc, la population n'est pas homogène si :

$$\frac{v_A}{v_R} > F_\alpha(k - 1, N - k)$$

Dans ce cas, les observations se mettent sous la forme :

$$x_i^j = \mu_i + \xi_i^j$$

où le terme μ_i est une correction correspondant au niveau i et le terme ξ_i^j est une fluctuation aléatoire suivant la loi normale $N(0 ; \sigma)$, la variance étant indépendante du niveau choisi. Une estimation de chaque terme μ_i est donnée par chaque moyenne $\bar{x}_i$.

Exemple 16.5

On veut comparer l'usure de quatre types de pneumatiques P_1 , P_2 , P_3 et P_4 . Sur chacun d'eux, on fait un certain nombre d'essais, 4 ou 5 ; les coefficients d'usure sont donnés dans le tableau 16.8 (en excès au-delà de la valeur 80).

Tableau 16.8 – Coefficients d'usure.

N° de l'essai	P ₁	P ₂	P ₃	P ₄
I	3	1	2	3
II	3	1	5	3
III	4	2	6	2
IV	5	4	4	1
V			4	4
Total	15	8	21	13

Peut-on considérer que les quatre types de pneumatiques sont équivalents ?

– Statistique des différents résultats :

$$\text{Pneumatique } P_1 : \bar{x}_1 = 3,75 \quad s_1^2 = 0,6875$$

$$\text{Pneumatique } P_2 : \bar{x}_2 = 2 \quad s_2^2 = 1,5$$

$$\text{Pneumatique } P_3 : \bar{x}_3 = 4,2 \quad s_3^2 = 1,76$$

$$\text{Pneumatique } P_4 : \bar{x}_4 = 2,6 \quad s_4^2 = 1,04$$

– Test sur l'égalité des variances : on compare les estimations des variances des échantillons I et III (la plus petite et la plus grande). Si ces variances peuvent être considérées comme égales, le rapport : $\frac{5 \times 1,76}{4} \times \frac{3}{4 \times 0,68756} = 2,40$ est la réalisation d'une variable de Fisher $F(4; 3)$.

$$\Pr(F(4; 3) > 9,12) = 0,05$$

On ne peut pas rejeter l'hypothèse d'égalité des variances des échantillons I et III. Les quatre variances peuvent être considérées comme égales.

– Analyse de la variance :

Nombre total d'observations, $N = 18$

Somme de tous les termes, $15 + 8 + 21 + 13 = 57$

$$\text{Variation totale, } N S^2 = 3^2 + 3^2 + \dots + 1^2 + 4^2 - \frac{57^2}{18} = 36,5$$

$$\text{Variation due au facteur, } N S_A^2 = \frac{15^2}{4} + \frac{8^2}{4} + \frac{21^2}{5} + \frac{13^2}{5} - \frac{57^2}{18} = 13,75$$

$$\text{Variation résiduelle, } N S_R^2 = 36,5 - 13,75 = 22,75$$

Tableau 16.9 – Analyse de la variance.

Variation	Somme des carrés	Degré de liberté	Quotient
due au facteur	13,75	3	$V_A = 4,58$
résiduelle	22,75	14	$V_R = 1,625$
totale	36,50	17	

$$V_A / V_R = 2,82$$

$$\Pr(F(3; 14) > 3,34) = 0,95$$

On peut admettre que la population est homogène, il n'y a pas de différence entre les quatre types de pneumatiques.

L'estimation de l'usure moyenne est égale à $57/18 = 3,17$ (moyenne générale) et celle de la variance au quotient $V_R = 1,625$.

17 • TESTS D'INDÉPENDANCE

Pour tester l'indépendance entre deux variables aléatoires X et Y , par l'examen d'échantillons de taille n , on étudie différentes mesures de liaison dépendant de la nature de ces variables quantitatives ou qualitatives.

17.1 Variables quantitatives

Soient X et Y deux variables aléatoires quantitatives de densités respectives f et g ; h est la densité du couple (X, Y) . Si ces variables aléatoires sont indépendantes, les densités doivent vérifier la propriété :

$$h(x, y) = f(x) g(y)$$

Cette propriété est donc l'hypothèse H_0 qu'il faut tester.

Pour vérifier cette hypothèse, on étudie les propriétés du coefficient de corrélation ρ , pour des populations gaussiennes, puis pour des populations quelconques.

17.1.1 Échantillons gaussiens

Une condition nécessaire et suffisante d'indépendance de *deux variables aléatoires gaussiennes* est que le coefficient de corrélation ρ soit nul (chapitre 8, paragraphe 8.4.3).

Coefficient de corrélation (rappel) :

$$\rho = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}$$

D'où le test :

$$\text{Hypothèses :} \quad H_0 \quad \rho = 0 \quad H_1 \quad \rho \neq 0$$

Le coefficient ρ est estimé par le coefficient de corrélation empirique r , calculé sur un échantillon de taille n , avec la formule :

$$r = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x}) (y_i - \bar{y})}{\sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}}$$

Ce coefficient r peut être considéré comme la réalisation d'une variable aléatoire R ; il existe des tables donnant les valeurs de la densité de la variable R . Les propriétés de la loi de cette variable dépendent des valeurs du coefficient ρ . Supposons $\rho = 0$. La distribution de R est symétrique seulement dans ce cas.

La variable aléatoire $\frac{R\sqrt{n-2}}{\sqrt{1-R^2}}$ est une variable aléatoire de Student à $(n-2)$ degrés de liberté. On en déduit la densité de la variable R ainsi que ses moments :

$$g(r) = \frac{1}{B\left(\frac{1}{2}, \frac{n-2}{2}\right)} (1-r^2)^{\frac{n-4}{2}}$$

$$E(R) = 0 \quad \text{Var}(R) = \frac{1}{n-1}$$

On en déduit les propriétés de ρ ; pour un test d'indépendance, on est amené à rejeter les grandes valeurs de $|r|$. La région critique est donc de la forme $|R| > k$:

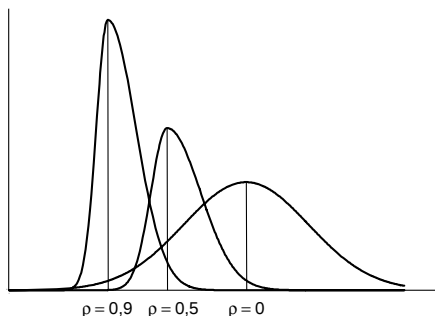


Figure 17.1 – Distribution de R pour différentes valeurs de ρ .

Remarques

- Si $n = 4$, la loi de la variable R est une loi uniforme sur $(-1, 1)$.
- Si $n > 100$, la loi de R peut être approchée par la loi normale $N\left(0; \sqrt{1/n-1}\right)$.

Pour ρ différent de 0, la loi de la variable R est connue mais difficile à utiliser. On trouve en particulier :

$$E(R) = \rho - \frac{\rho(1-\rho^2)}{2n} \quad \text{Var}(R) = \frac{(1-\rho^2)^2}{n-1}$$

On considère la transformée Z de Fisher de R :

$$Z = \frac{1}{2} \text{Ln} \frac{1+R}{1-R}$$

Quand n est supérieur à 25, la loi de Z tend vers la loi normale :

$$N\left(\frac{1}{2} \text{Ln} \frac{1+\rho}{1-\rho} ; \frac{1}{\sqrt{n-3}}\right)$$

17.1.2 Échantillons quelconques

Le test précédent est encore valable si n est grand, $n > 25$. Cependant, on ne teste pas l'indépendance, mais seulement la « non-corrélation linéaire ». Ce test est *robuste*.

17.2 Variables ordinales et corrélation des rangs

Soient X et Y deux variables ordinales. Ce sont, par exemple :

- les classements établis par deux critiques et par ordre de préférence, de n livres, ou de n films,
- la comparaison des notes obtenues par deux étudiants à une série de n épreuves (ce ne sont pas les notes que l'on étudie mais leurs classements relatifs).

On a donc un ensemble de n individus ou objets qui ont été soumis à deux classements :

Tableau 17.1 – Classements de n objets réalisés par deux individus.

Objet	1	2	3	...	n
1 ^{er} classement	u_1	u_2	u_3	...	u_n
2 ^e classement	v_1	v_2	v_3	...	v_n

Le problème posé est de comparer les deux classements, c'est-à-dire de répondre à la question :

Ces classements sont-ils identiques ou non ?

Les tests de Spearman et de Kendall ont été proposés pour résoudre ce problème.

17.2.1 Coefficient de corrélation des rangs de Spearman

Le coefficient de corrélation des rangs, proposé en 1904 par le psychologue Spearman, est en fait le coefficient de corrélation usuel calculé sur les rangs.

Soient U et V les variables aléatoires associées aux deux rangs.

Le coefficient r_s de Spearman est donné par (s_u et s_v étant les écarts-types empiriques des deux classements) :

$$r_s = \frac{\text{Cov}(u, v)}{s_u s_v}$$

Chaque classement étant une permutation des n premiers nombres entiers, on peut utiliser les résultats démontrés pour la loi uniforme discrète sur $[1, n]$ (chapitre 5, paragraphe 5.3).

$$\overline{U} = \overline{V} = \frac{n+1}{2} \quad \text{Var}(U) = \text{Var}(V) = \frac{n^2-1}{12}$$

D'où l'expression, déduite des valeurs observées, du coefficient de Spearman :

$$r_s = \frac{\frac{1}{n} \sum_{i=1}^n u_i v_i - \left(\frac{n+1}{2} \right)^2}{\frac{n^2-1}{12}}$$

Si on pose $d_i = u_i - v_i$, on obtient :

$$d_i^2 = u_i^2 + v_i^2 - 2 u_i v_i$$

u_i et v_i sont des nombres entiers compris entre 1 et n , donc :

$$\sum_{i=1}^n u_i^2 = \sum_{i=1}^n v_i^2 = \frac{n(n+1)(2n+1)}{6}$$

D'où l'expression suivante du coefficient de Spearman (utilisée pour le calcul) :

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

Propriétés du coefficient r_s :

- $r_s = 1$, $d_i = 0 \forall i$, les classements sont identiques,
- $r_s = -1$, les classements sont inverses l'un de l'autre,
- $r_s = 0$, les classements sont indépendants.

Sous l'hypothèse H_0 , indépendance des classements, la distribution exacte de R_s s'obtient en considérant les $n!$ permutations équiprobables des rangs.

Il existe des tables pour les petites valeurs de n .

Pour les grandes valeurs de n , $n > 30$, la distribution de R_s peut être approchée par la loi normale dont les moments sont : $E(R) = 0$, $\text{Var}(R) = 1/(n-1)$.

Ayant choisi un risque de première espèce α , la table (ou la loi normale) donne la valeur critique k de R_s . La région critique, refus de l'hypothèse H_0 d'indépendance, est définie par $|R_s| > k$. Plus précisément :

- si $r_s > k$, il y a concordance des rangs,
- si $r_s < -k$, il y a discordance des rangs.

17.2.2 Coefficient de corrélation des rangs de Kendall

Pour tester l'indépendance de deux classements, Kendall a proposé, en 1938, d'étudier la statistique τ définie de la façon suivante.

Soient deux classements, donc deux séries de valeurs (x_i) et (y_j) , permutations des entiers de 1 à n . On considère tous les couples de résultats et on leur attribue :

- La note + 1 si les classements correspondants des variables X et Y sont dans le même ordre, c'est-à-dire, si on a à la fois : $x_i < x_j$ et $y_i < y_j$.

Il y a *concordance des classements*.

- La note -1 si les classements ne sont pas dans le même ordre, c'est-à-dire, si on a à la fois : $x_i < x_j$ et $y_i > y_j$.

Il y a *discordance des classements*.

Soit S la somme obtenue en considérant les $n(n-1)/2$ couples distincts (x_i, x_j)

On vérifie que :

$$S_{\max} = -S_{\min} = \frac{n(n-1)}{2}$$

La valeur $S_{\max}$ correspond à la concordance parfaite et la valeur $S_{\min}$ à la discordance complète.

Le coefficient τ de Kendall est défini par :

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

- $\tau = 1$, les classements sont identiques,
- $\tau = -1$, les classements sont inverses,
- s'il n'y a pas de dépendance monotone, on peut s'attendre à une valeur de τ voisine de 0.

Si l'hypothèse H_0 d'indépendance des deux classements est vraie, la loi de la variable aléatoire τ est approximativement la loi normale :

$$N\left(0; \sqrt{\frac{2(2n+5)}{9n(n-1)}}\right)$$

Cette approximation est valable si $n \geq 8$.

Pour calculer rapidement le coefficient τ de Kendall, on ordonne une des séries, la série des (x_i) par exemple, de 1 à n , la série des (y_j) en découle.

Tableau 17.2 – Tableau des résultats, la série X étant ordonnée.

Série (x_i)	1	2	3	...	n
Série (y_i)	y_{i1}	y_{i2}	y_{i3}	...	y_{in}

Pour chaque x_i , on compte le nombre de y_j tel que $y_j > y_i$. Si R est la valeur de la somme, le coefficient de Kendall est égal à :

$$S = 2R - \frac{n(n-1)}{2} \quad \tau = \frac{4R}{n(n-1)} - 1$$

Exemple 17.1 Coefficients de Spearman et de Kendall

Les classements de douze élèves en mathématiques et en musique sont les suivants :

mathématiques (x_i)	3	6	2	1	4	9	11	12	5	7	8	10
musique (y_i)	6	1	3	4	7	9	2	12	5	11	10	8

Les aptitudes de ces douze élèves en ces deux matières sont-elles indépendantes ?

– *Coefficient de Spearman* :

Posons $d_i = x_i - y_i$

$$\sum_i d_i^2 = 9 + 25 + 1 + 9 + 9 + 81 + 16 + 4 + 4 = 158$$

$$r_s = 1 - \frac{6 \times 158}{12 \times 143} = 0,45$$

Pour $n = 12$, pour un seuil critique égal à 5 % et un test bilatéral, la valeur critique est égale à $0,587 > 0,45$. On accepte l'hypothèse d'indépendance des classements.

– *Coefficient de Kendall* :

Pour calculer ce coefficient, on range les élèves dans l'ordre de leur classement en mathématiques :

Classement en mathématiques	1	2	3	4	5	6	7	8	9	10	11	12
Classement correspondant en musique	4	3	6	7	5	1	11	10	9	8	2	12
$\sum_j j/y_j > y_i$	8	8	6	5	5	6	1	1	1	1	1	0

Dans le classement des notes obtenues en musique, il y a, par exemple, 8 places supérieures à 4 (les places 6, 7, 5, 11, 10, 9, 8 et 12), 6 supérieures à 1 (les places 11, 10, 9, 8, 2 et 12)...

$$R = \sum_j j/y_j > y_i = 8 + 8 + 6 + 5 + 5 + 6 + 1 + 1 + 1 + 1 + 1 + 1 = 43$$

D'où la valeur du coefficient de Kendall :

$$\tau = \frac{4R}{n(n-1)} - 1 = 0,30$$

Si l'hypothèse d'indépendance est vraie, la variable X suit la loi normale :

$$N\left(0; \sqrt{\frac{2(2n+5)}{9n(n-1)}}\right) = N(0; 0,23)$$

$\tau = 0,30$ d'où $\tau/0,23 = 1,30$ (variable aléatoire centrée réduite).

Or $\Pr(-1,96 < U < 1,96) = 0,95$

On ne peut donc pas rejeter l'hypothèse d'indépendance entre les deux séries de notes.

17.2.3 Conclusion

Les coefficients de corrélation des rangs sont très utiles pour tester l'indépendance de deux variables non gaussiennes car le test du coefficient de corrélation linéaire ne s'applique pas dans ce cas. De plus, ils sont invariants par toute transformation monotone croissante des variables.

17.3 Concordance de p classements

C'est une généralisation du problème étudié dans le paragraphe 17.2.

n individus ont été classés selon p critères, comme, par exemple, le classement de n livres par p critiques.

Les résultats se présentent sous la forme suivante :

Tableau 17.3 – Classement de n individus selon p critères.

Critères \ Individus	1	2	...	n
1	r_{11}	r_{21}		r_{n1}
2	r_{12}	r_{22}		r_{n2}
p	r_{1p}	r_{2p}		r_{np}
Total	$r_{1.}$	$r_{2.}$		$r_{n.}$

Chaque ligne est une permutation des entiers de 1 à n , la somme des termes de n'importe quelle ligne est égale à $n(n+1)/2$. La somme des termes du tableau est donc égale à $N = pn(n+1)/2$.

Si les classements étaient rigoureusement identiques, une des colonnes aurait pour somme p , une autre $2p$, une autre $3p$, etc.

Pour étudier la concordance entre ces classements, on considère la statistique :

$$S = \sum_{i=1}^n \left(r_{i.} - \frac{N}{n} \right)^2$$

Cette statistique est une mesure de la dispersion des sommes des colonnes par rapport à leur moyenne.

Si la concordance est parfaite, la statistique S est maximale et vaut :

$$S_{\max} = \frac{n p^2 (n^2 - 1)}{12}$$

Kendall a proposé, pour étudier la concordance de p classements, le coefficient W :

$$W = \frac{12 S}{n p^2 (n^2 - 1)}$$

Ce coefficient est compris entre 0 et 1.

$W = 0$ si les sommes de toutes les colonnes sont égales.

Une faible valeur de W indique l'indépendance entre les classements.

L'hypothèse H_0 d'indépendance des classements est rejetée si W est trop grand ; des tables donnent les valeurs critiques de W pour différentes valeurs de n et de p .

Pour $n \geq 15$ et $p < 7$, la variable $\frac{(p-1) W}{1-W}$ est une variable de Fisher $F[n-1-2/p; (n-1)(n-1-2/p)]$

Pour $p \geq 7$, la variable $p(n-1)W$ suit une loi du chi-deux à $(n-1)$ degrés de liberté.

17.4 Liaison entre une variable quantitative et une variable qualitative

Pour étudier la liaison entre une variable quantitative Y et une variable qualitative X , définies sur un ensemble de n individus (Y est par exemple le salaire proposé au premier emploi et X le niveau des études), on utilise le rapport de corrélation.

17.4.1 Rapport de corrélation théorique

Le rapport de corrélation de la variable Y en la variable X est donné par :

$$\eta_{Y/X}^2 = \frac{\text{Var}[E(Y/X)]}{\text{Var}(Y)}$$

Ses propriétés ont été données dans le chapitre 8, paragraphe 8.2.5.

17.4.2 Rapport de corrélation empirique

La variable qualitative X prend k modalités. On note :

- n_i l'effectif observé pour la variable Y quand la variable X prend la modalité i ,
- $\bar{Y}_i$ la moyenne des n_i valeurs prises par la variable Y pour la modalité i de la variable X ,
- $\bar{Y}$ la moyenne générale des valeurs prises par la variable Y .

Le coefficient empirique de corrélation e^2 est donné par :

$$e^2 = \frac{\frac{1}{n} \sum_{i=1}^k n_i (\bar{y}_i - \bar{y})^2}{s_Y^2} \quad s_Y^2 = \frac{1}{n} \sum_{i=1}^k \sum_{j=1}^{n_i} (y_i^j - \bar{y})^2$$

C

STATISTIQUE INFÉRENTIELLE

■ Propriétés du coefficient de corrélation e^2

Ce coefficient est compris entre 0 et 1 :

- $e^2 = 0$. Pour toutes les valeurs de l'indice i , on a $\bar{y}_i = \bar{y}$. Il n'y a donc pas de dépendance en moyenne.
- $e^2 = 1$. Pour une modalité i de la variable X , tous les individus ont la même valeur et ceci pour toutes les valeurs de l'indice i . En effet :

$$e^2 = 1 \quad \Rightarrow \quad \frac{1}{n} \sum_{i=1}^k n_i (\bar{y}_i - \bar{y})^2 = s_Y^2 = \frac{1}{n} \sum_{i=1}^k \sum_{j=1}^{n_i} (y_i^j - \bar{y})^2$$

- Dans le cas où $e^2 \neq 0$, on utilise les résultats de l'analyse de la variance à simple entrée (chapitre 16, paragraphe 16.3).

La quantité :

$$\frac{1}{n} \sum_{i=1}^k n_i (\bar{y}_i - \bar{y})^2 = e^2 s_Y^2 = S_A^2$$

représente les variations entre les différentes modalités, c'est-à-dire la variation expliquée par le *facteur contrôlé*. C'est la réalisation d'une variable χ^2 à $(k-1)$ degrés de liberté.

La variation totale est représentée par la quantité s_Y^2 , et la variation résiduelle par la différence :

$$s_Y^2 - e^2 s_Y^2 = (1 - e^2) s_Y^2 = S_R^2$$

Le rapport :

$$\frac{S_A^2 / (k-1)}{S_R^2 / (n-k)} = \frac{e^2 / (k-1)}{(1 - e^2) / (n-k)}$$

suit une loi de Fisher à $(k-1 ; n-k)$ degrés de liberté sous l'hypothèse H_0 , $\eta^2 = 0$ (test de l'analyse de la variance).

Si le rapport $\frac{S_A^2 / (k-1)}{S_R^2 / (n-k)}$ est supérieur à la valeur critique, pour un seuil donné α , d'une variable de Fisher $F(k-1 ; n-k)$, on rejette l'hypothèse H_0 .

Remarque

Pour appliquer ces résultats, il faut supposer que, pour chaque modalité du facteur contrôlé, les distributions de Y suivent des lois normales de même espérance et de même variance.

17.5 Liaison entre deux variables qualitatives

Soient X et Y deux variables qualitatives prenant respectivement p et q modalités, notées x_i et y_j .

Les résultats des observations sont donnés sous forme d'un tableau à double entrée, ou tableau de contingence :

Tableau 17.4 – Tableau des résultats dans l'hypothèse de deux variables qualitatives.

$X \backslash Y$	y_1	y_2		y_j		y_q	Total
x_1	n_{11}	n_{12}		n_{1j}		n_{1q}	$n_{1.}$
x_2	n_{21}	n_{22}		n_{2j}		n_{2q}	$n_{2.}$
x_i	n_{i1}	n_{i2}		n_{ij}		n_{iq}	$n_{i.}$
x_p	n_{p1}	n_{p2}		n_{pj}		n_{pq}	$n_{p.}$
Total	$n_{.2}$	$n_{.1}$		$n_{.j}$		$n_{.q}$	N

n_{ij} nombre d'individus présentant le caractère x_i (pour le facteur X) et le caractère y_j (pour le facteur Y).

$n_{i.}$ nombre total d'individus ayant le caractère x_i .

$n_{.j}$ nombre total d'individus ayant le caractère y_j .

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

N nombre total d'observations.

C

STATISTIQUE INFÉRENTIELLE

17.5.1 Mesure de l'indépendance entre les variables X et Y

La variable Y est statistiquement indépendante de la variable X si les distributions conditionnelles de Y à X fixé sont identiques, c'est-à-dire si on a :

$$\forall j \quad \frac{n_{1j}}{n_{1.}} = \frac{n_{2j}}{n_{2.}} = \dots = \frac{n_{pj}}{n_{p.}} = \frac{\sum_{k=1}^p n_{kj}}{\sum_{i=1}^p n_{i.}} = \frac{n_{.j}}{N} \Rightarrow \frac{n_{ij}}{n_{i.}} = \frac{n_{.j}}{N}$$

L'indépendance de la variable Y par rapport à la variable X se traduit donc par :

$$\forall i, \forall j \quad n_{ij} = \frac{n_{i.} \cdot n_{.j}}{N}$$

La variable X est donc indépendante de la variable Y (symétrie en X et Y du résultat précédent), c'est-à-dire *les variables X et Y sont statistiquement indépendantes*.

Une mesure de l'écart à l'indépendance est donnée par la valeur de la quantité d^2 :

$$d^2 = \sum_{i=1}^p \sum_{j=1}^q \frac{\left(n_{ij} - \frac{n_{i.} n_{.j}}{N} \right)^2}{\frac{n_{i.} n_{.j}}{N}} = N \left[\sum_{i=1}^p \sum_{j=1}^q \frac{n_{ij}^2}{n_{i.} n_{.j}} - 1 \right]$$

L'indépendance parfaite implique que la valeur de d^2 soit nulle. Pour mesurer l'écart à l'indépendance, il faut trouver la borne supérieure de d^2 .

Comme :

$$\frac{n_{ij}}{n_{i.}} \leq 1 \quad \text{et} \quad \frac{n_{ij}}{n_{.j}} \leq 1$$

on obtient facilement :

$$\sum_{i=1}^p \sum_{j=1}^q \frac{n_{ij}^2}{n_{i.} n_{.j}} \leq \sum_{i=1}^p \sum_{j=1}^q \frac{n_{ij}}{n_{.j}} \leq \sum_{j=1}^q \sum_{i=1}^p \frac{n_{ij}}{n_{.j}} = q$$

Donc $d^2 \leq N(q-1)$ et de la même façon $d^2 \leq N(p-1)$ d'où :

$$d^2 < N \inf(p-1; q-1)$$

Si la borne est atteinte, il existe une relation fonctionnelle entre les variables X et Y . En effet, en supposant par exemple $q < p$, on obtient :

$$d^2 = n(q-1) \quad \text{si} \quad \frac{n_{ij}}{n_{i.}} = 1$$

ou, en d'autres termes, il n'y a aucune case nulle.

17.5.2 Étude de la distribution d^2

Soient p_{ij} , $p_{i.}$ et $p_{.j}$ les probabilités conjointes et marginales.

L'hypothèse d'indépendance H_0 est :

$$p_{ij} = p_{i.} p_{.j}$$

La statistique d^2 :

$$d^2 = \sum_{i=1}^p \sum_{j=1}^q \frac{(n_{ij} - N p_{i.} p_{.j})^2}{N p_{i.} p_{.j}}$$

est la réalisation d'une variable aléatoire suivant une loi du χ^2 à $(pq - 1)$ degrés de liberté.

Les probabilités $p_{i.}$ et $p_{.j}$ ne sont pas connues mais estimées :

$$p_{i.} \text{ est estimée par } \frac{n_{i.}}{N} \text{ et } p_{.j} \text{ par } \frac{n_{.j}}{N}$$

On calcule donc la quantité :

$$d^2 = N \left(\sum_{i=1}^p \sum_{j=1}^q \frac{n_{ij}^2}{n_{i.} n_{.j}} - 1 \right)$$

La statistique d^2 est la réalisation d'une variable aléatoire suivant une loi du χ^2 dont le nombre de degrés de liberté est égal à :

$$pq - 1 - (p - 1) - (q - 1) = (p - 1)(q - 1)$$

car le nombre de paramètres estimés est égal à $(p - 1)$ et $(q - 1)$.

On choisit un risque de première espèce α et on en déduit le seuil critique d_α .

On rejette l'hypothèse H_0 si d^2 est supérieure à cette valeur d_α .

Remarque

On a vu comment le calcul se simplifie lorsque $p = q = 2$ (chapitre 16, paragraphe 16.2.5).

Exemple 17.2 Test de l'indépendance de deux critères de classification

On a interrogé 1 205 ménages choisis au hasard et on les a classés suivant :

- 1) la catégorie socio-professionnelle du chef de famille (ces catégories ont été notées A, B, C et D),
- 2) le nombre d'enfants.

On a obtenu les résultats suivants regroupés dans le tableau 17.5.

Tableau 17.5 – Classification des ménages.

Catégories	Nombre d'enfants			Total
	0 ou 1	2 ou 3	Plus de 3	
A	155	100	5	260
B	200	95	5	300
C	205	105	5	315
D	190	125	15	330
Total	750	425	30	1 205

$$d^2 = 1\,205 \left(\frac{155^2}{750 \times 260} + \frac{100^2}{425 \times 260} + \dots + \frac{15^2}{30 \times 330} - 1 \right) = 13,4041$$

Cette valeur, 13,4041, est sous l'hypothèse d'indépendance des deux critères de classification, la réalisation d'une variable chi-deux à $2 \times 3 = 6$ degrés de liberté.

$$\Pr(\chi^2(6) > 12,6) = 0,95$$

L'hypothèse d'indépendance doit être rejetée.

18 • FIABILITÉ

18.1 Généralités et principales définitions

La fiabilité est une science relativement récente dont on peut situer approximativement les débuts vers les années 1960. Elle s'est développée très rapidement, elle a des applications dans de nombreux domaines.

En effet, pour des raisons de sécurité, il est absolument nécessaire que certains matériels assurent un fonctionnement sans défaillance ; c'est le cas, par exemple, des systèmes de défense, des grands réseaux de distribution, des systèmes centralisés des informations...

D'autres raisons sont plutôt d'ordre économique ; au prix de revient d'une première installation s'ajoutent tous les frais d'exploitation et ceux-ci comportent généralement une large part due aux coûts de défaillance ou aux entretiens préventifs, il est donc impératif de les minimiser.

En général, on considère qu'un matériel est constitué de composants ou pièces alors qu'un système est un ensemble de composants ou de matériels interconnectés ou en interaction.

18.1.1 Fiabilité

La fiabilité est la caractéristique d'une « chose » à laquelle on peut se fier. Plus précisément, *la fiabilité est l'aptitude d'un dispositif à accomplir une fonction requise, dans des conditions données, pour une période de temps donné.*

(Définition de la norme de la Commission internationale d'électrotechnique.)

18.1.2 Défaillance

La défaillance est la fin de l'aptitude d'un dispositif ou d'un système à accomplir la fonction que l'on attendait de ce matériel.

On distingue :

- les *défaillances graves* ou *totales* entraînant la fin de la fonction,
- les *défaillances partielles* réduisant les performances mais non la fonction.

18.1.3 Intervalle de temps entre défaillances

L'intervalle de temps entre défaillances ou *temps de bon fonctionnement* est la durée de fonctionnement d'un dispositif réparable entre deux défaillances successives. On lui associe le MTBF ou *temps moyen entre défaillance*.

Pour les dispositifs non réparables, on introduit la notion de *durée de vie*, c'est la durée de fonctionnement jusqu'à la défaillance totale.

18.2 Définition mathématique de la fiabilité

Exemple 18.1

À partir de cet exemple simple, on introduit les principales notions intervenant dans la théorie de la fiabilité.

On met en service, au temps $t = 0$, 200 matériels identiques fonctionnant dans les mêmes conditions. On relève à intervalles réguliers, toutes les 50 heures par exemple, le nombre $N(t)$ de matériels survivants à cette date. On obtient les résultats suivants :

Dates	0	1	2	3	4	5	6	7	8
$N(t)$	200	195	175	150	110	75	50	20	0

À partir de ces observations, on peut déterminer :

- le nombre de matériels défaillants à chaque date t : $\Delta N(t) = N(t-1) - N(t)$,
- le pourcentage de survivants à chaque date t ou fréquence relative des survivants : $R(t) = N(t) / N(0)$,
- la proportion de défaillants dans l'intervalle $(t-1, t)$ ou fréquence relative des défaillants :

$$f(t) = \frac{N(t-1) - N(t)}{N(0)} = R(t-1) - R(t)$$

– le taux moyen de défaillance :

$$\lambda(t) = \frac{N(t - \Delta t) - N(t)}{N(t - \Delta t)} = \frac{R(t - \Delta t) - R(t)}{R(t - \Delta t)}$$

Le tableau suivant donne les résultats :

Dates	$N(t)$	$\Delta N(t)$	$R(t)$	$f(t)$	$\lambda(t)$
0	200	0	1	0	
1	195	5	0,975	0,025	0,025
2	175	20	0,875	0,10	0,102
3	150	25	0,75	0,125	0,143
4	110	40	0,55	0,20	0,266
5	75	35	0,375	0,175	0,318
6	50	25	0,25	0,125	0,333
7	20	30	0,10	0,15	0,60
8	0	20	0	0,10	1

La *fiabilité* d'un matériel au temps t est la probabilité pour que la variable aléatoire T , non négative, représentant la durée de vie de ce matériel, soit supérieure à une valeur t :

$$R(t) = \Pr(T > t) = 1 - \Pr(T < t) = 1 - F(t)$$

$R(t)$ est la *fonction de survie* ou *fiabilité du matériel* ou *reliability*.

$F(t)$ est la *distribution cumulée des défaillances*, c'est-à-dire la probabilité de mourir au temps t dans le cas d'une défaillance totale.

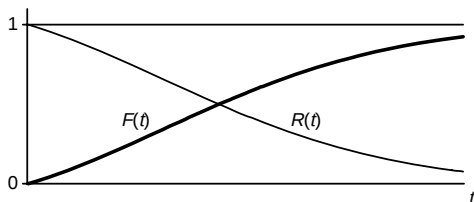


Figure 18.1 – Relation entre la fonction de répartition F et la fiabilité R d'un groupe d'objets identiques.

Remarques

- La fiabilité $R(t)$ n'est définie que pour des matériels identiques fonctionnant dans les mêmes conditions.
- Pour certains types de matériels, on doit distinguer la fiabilité en cours de stockage et la fiabilité en service.
- La variable aléatoire T ne désigne pas toujours un temps mais peut représenter le nombre de kilomètres parcourus, le nombre de manœuvres effectuées...

18.3 Taux de défaillance

18.3.1 Probabilité de défaillance après un temps de fonctionnement

On suppose que la fonction de répartition $F(t)$ admet une dérivée, c'est-à-dire que la densité de probabilité de la variable aléatoire T existe. Dans ces conditions, on écrit :

$$f(t)dt = \frac{dF(t)}{dt}dt = \frac{d[1 - R(t)]}{dt}dt$$

$f(t)dt$ est la *probabilité de défaillance* pendant l'intervalle de temps dt sachant que le matériel a déjà fonctionné pendant un temps égal à t . D'où la définition du taux de défaillance.

18.3.2 Définition du taux de défaillance

On considère un matériel ayant fonctionné sans incident pendant un temps t . $dF(t) = f(t)dt$ est la probabilité que ce matériel cesse d'être utilisable pendant l'intervalle $(t, t + dt)$ après avoir fonctionné sans incident pendant le temps t .

$R(t)$ est la probabilité que le matériel soit encore en service à l'instant t .

Soit $\lambda(t)dt$ le risque ou le taux de défaillance immédiate.

Le théorème de la probabilité conditionnelle donne :

$$dF(t) = f(t)dt = R(t)\lambda(t)dt = [1 - F(t)]\lambda(t)dt$$

D'où la définition du *taux de défaillance immédiate* :

$$\lambda(t) = \frac{f(t)}{1 - F(t)} = -\frac{1}{R(t)} \frac{dR(t)}{dt} = -\frac{d \text{Ln}[1 - F(t)]}{dt} = -\frac{d \text{Ln}R(t)}{dt}$$

Inversement, la connaissance du taux de défaillance donne, par intégration, la fonction de survie :

$$R(t) = 1 - F(t) = \exp - \int_0^t \lambda(u) du$$

Dans la plupart des cas, la courbe représentant le taux de défaillance d'un matériel ou d'un composant, en fonction de son âge, a une forme caractéristique, c'est la *courbe en baignoire* :

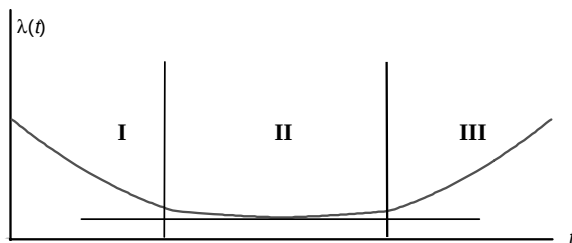


Figure 18.2 – Taux de défaillance en fonction du temps. Courbe en baignoire.

On distingue les trois périodes suivantes :

- une première période dite « de jeunesse » ou période de mortalité infantile due à des erreurs de fabrication ou de conception. Le taux de défaillance d'abord très élevé décroît rapidement. La durée de cette période peut varier de 10 heures pour les composants mécaniques à 1 000 heures pour les composants électroniques,
- une deuxième période où le taux de défaillance est pratiquement constant et peu élevé. Cette période correspond à des défaillances apparaissant aléatoirement, sans cause systématique. Elle peut atteindre plus de 100 000 heures pour les composants électroniques mais est beaucoup plus courte pour les composants mécaniques,
- la troisième période est caractérisée par un taux de défaillance croissant très rapidement, les défaillances sont dues à l'usure, elle marque la fin de la vie utile du matériel.

18.3.3 Estimation du taux de défaillance

On considère des matériels identiques fonctionnant dans les mêmes conditions. Soit $N(t)$ le nombre de matériels ayant fonctionné sans incident pendant le temps t et dN le nombre de matériels ayant une défaillance pendant l'intervalle de temps $(t - dt, t)$.

Une estimation du taux de défaillance $\lambda(t)$ est donnée par :

$$\lambda(t) = \frac{dN}{N(t - dt) - N(t)}$$

On montre que :

$$E\left(\frac{dN}{N}\right) = \lambda(t)$$

18.4 Fiabilité d'un matériel usagé

Soit un matériel neuf mis en service au temps $t = 0$ et $R_0(t)$ sa *fiabilité à l'état neuf*.

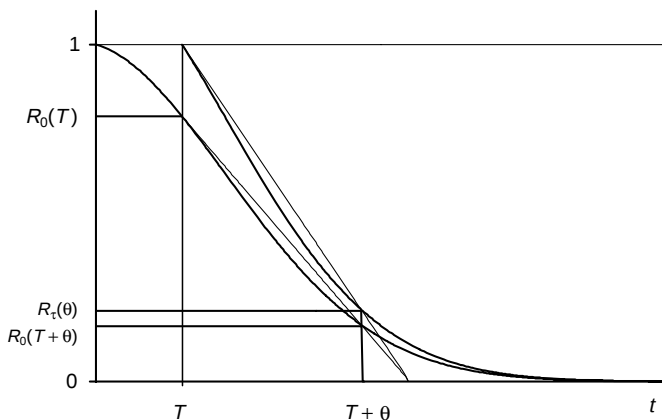
Au temps τ , on suppose qu'il est encore en service. On veut calculer la probabilité conditionnelle $R_\tau(\theta)$ pour que ce matériel usagé puisse encore fournir sans défaillance un service de durée θ sachant qu'il a déjà fonctionné sans défaillance pendant le temps τ .

Le théorème des probabilités conditionnelles donne :

$$R_0(\tau + \theta) = R_0(\tau) R_\tau(\theta)$$

$$\frac{R_\tau(\theta)}{R_0(\tau + \theta)} = \frac{1}{R_0(\tau)} = \frac{R_\tau(0)}{R_0(\tau)} = \text{cste}$$

On en déduit que la courbe de survie d'un matériel usagé $R_\tau(\theta)$ et sa courbe de survie à l'état neuf $R_0(\tau + \theta)$ se déduisent l'une de l'autre par une affinité de rapport $1/R_0(\tau)$.

Figure 18.3 – Courbes de survie $R_\tau(\theta)$ et $R_0(\tau)$.

18.5 Fiabilité en cas de remplacement préventif

Pour des raisons économiques ou de sûreté, on peut procéder au remplacement préventif d'un matériel ayant fonctionné sans défaillance pendant un temps T_m .

Soit $R(t)$ sa fiabilité intrinsèque, c'est-à-dire la fiabilité définie sans limitation du temps de fonctionnement. La probabilité Pr de remplacement pendant l'intervalle de temps $(t + dt, t)$ est égale à :

$$\text{Pr} = f(t) dt = -\frac{dR(t)}{dt} dt \quad \forall t < T_m$$

$$\text{Pr} = R(T_m) \quad t = T_m$$

$$\text{Pr} = 0 \quad t > T_m$$

La courbe représentant la fiabilité de ce matériel a une discontinuité au point $t = T_m$.

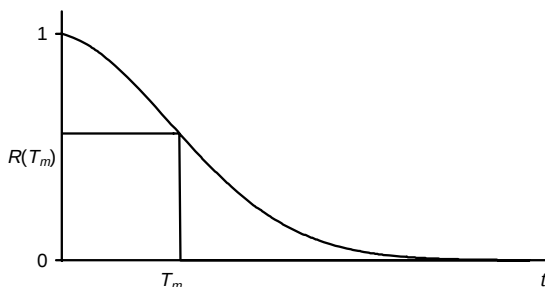


Figure 18.4 – Fiabilité d'un matériel ayant subi un remplacement préventif au temps T_m .

18.6 Espérance de vie

L'espérance de vie, souvent désignée sous le sigle MTBF (*Mean Time Between Failure*), est donnée par l'intégrale :

$$E(T) = \int_0^{\infty} t f(t) dt = \int_0^{\infty} t dF = - \int_0^{\infty} t dR$$

Une intégration par parties conduit au résultat suivant :

$$E(T) = \int_0^{\infty} R(t) dt \quad \text{si } \tau R(\tau) \rightarrow 0 \text{ quand } \tau \rightarrow \infty$$

Remarques

- La condition imposée à la fonction R est vérifiée dans de nombreux cas.
- Dans le cas d'un remplacement préventif, il faut faire attention à l'expression de la fiabilité.

18.7 Exemples de lois de fiabilité

18.7.1 Loi de fiabilité exponentielle

On considère des matériels fonctionnant sans usure comme par exemple la majorité des composants électroniques, des circuits intégrés, des semi-conducteurs...

Ces matériels sont caractérisés par un taux de défaillance λ constant. Ce cas correspond à la région II de la courbe représentant les variations du taux de défaillance en fonction du temps t .

■ Forme mathématique de la loi de fiabilité

Le taux de défaillance étant égal à une constante λ , on obtient facilement la fonction de survie :

$$R(t) = \exp - \int_0^t \lambda \, d u = e^{-\lambda t}$$

$$F(t) = 1 - e^{-\lambda t}$$

La distribution F des durées de vie suit une loi exponentielle de paramètre λ (loi étudiée chapitre 6, paragraphe 6.3).

■ Moments de cette distribution

$$E(T) = \frac{1}{\lambda} \quad \text{Var}(T) = \frac{1}{\lambda^2}$$

On utilise souvent le paramètre $\theta = \frac{1}{\lambda}$. Dans ces conditions, l'espérance ou le MTBF est égale au paramètre θ .

■ Nombre de remplacements à effectuer pour assurer un service de durée τ

Ce nombre est distribué suivant une loi de Poisson de paramètre $\lambda\tau$. La probabilité d'avoir à effectuer x remplacements pour assurer un service d'une durée totale égale à τ est donnée par l'expression :

$$\Pr(X = x) = e^{-\lambda\tau} \frac{(\lambda\tau)^x}{x!}$$

■ Fiabilité d'un matériel usagé

La formule donnant la fiabilité d'un matériel usagé (paragraphe 18.4) s'écrit dans le cas d'une loi exponentielle :

$$R_\tau(\theta) = \frac{R(\tau + \theta)}{R(\tau)} = \frac{e^{-\lambda(\tau + \theta)}}{e^{-\lambda\tau}} = e^{-\lambda\theta}$$

On retrouve la même loi exponentielle. La fiabilité d'un matériel usagé, non soumis au vieillissement, est égale à sa fiabilité à l'état neuf. Donc, tout remplacement préventif d'un tel matériel est inutile. On rappelle que la loi exponentielle est qualifiée de *loi sans mémoire*, car elle ne souvient pas de son passé.

■ Estimation du paramètre λ ou θ ($\theta = \frac{1}{\lambda}$)

Soit T_f le temps cumulé des essais et n le nombre cumulé des défaillances pendant ce temps. L'estimation de θ est la quantité : $\hat{\theta} = \frac{T_f}{n}$.

■ Ajustement graphique

Cet ajustement a été exposé chapitre 16, paragraphe 16.1.1.

■ Test du chi-deux

Ce test, d'un emploi très fréquent, a été expliqué chapitre 16, paragraphe 16.1.1.

■ Test des temps cumulés entre défaillances

Ce test non paramétrique s'applique spécialement à la loi exponentielle.

On considère un dispositif réparable en essai pendant un temps T ayant eu C défaillances. On note $t_1, t_2, \dots, t_c$ les temps aux termes desquels sont apparues les C défaillances et $T(c)$ la somme : $T(c) = t_1 + t_2 + \dots + t_c$

Si le nombre C de défaillances constatées est grand, supérieur à 10, la somme

$T(c)$ suit la loi normale $N\left(\frac{CT}{2}; \sqrt{\frac{CT^2}{12}}\right)$.

On rejette l'hypothèse d'une loi exponentielle, test significatif, si la valeur calculée, $T(c)$, n'appartient à l'intervalle :

$$\left[\frac{CT}{2} - u \sqrt{\frac{CT^2}{12}} ; \frac{CT}{2} + u \sqrt{\frac{CT^2}{12}} \right]$$

Le coefficient u est lu sur les tables de la loi normale centrée réduite, il dépend de la probabilité α de rejet à tort choisi, $u = 1,6449$ si $\alpha = 10\%$, $u = 1,96$ si $\alpha = 5\%$.

18.7.2 Loi de fiabilité de Weibull

La loi exponentielle s'applique aux matériels dont le taux de défaillance peut être considéré comme constant, au moins pendant une longue période. Cette hypothèse n'est pas toujours vérifiée. Il est alors possible de représenter la fiabilité des matériels dont le taux de défaillance évolue avec le temps, par la loi proposée et étudiée par le mathématicien suédois Weibull, en 1951.

Une variable aléatoire réelle T , strictement positive, suit une loi de Weibull, si sa fonction de répartition est :

$$F(t) = 0 \quad \forall t < \gamma$$

$$F(t) = 1 - \exp - \left(\frac{t - \gamma}{\eta} \right)^\beta \quad \forall t \geq \gamma$$

Sa densité de probabilité est :

$$f(t) = 0 \quad \forall t < \gamma$$

$$f(t) = \frac{\beta}{\eta} \left(\frac{t - \gamma}{\eta} \right)^{\beta-1} \exp - \left(\frac{t - \gamma}{\eta} \right)^\beta \quad \forall t \geq \gamma$$

La fonction de répartition de cette loi dépend de trois paramètres positifs β , γ et η .

γ est un paramètre de position ayant la même dimension que la variable T . Il peut être pris égal à 0 (simple translation sur t).

η est un paramètre d'échelle, appelé parfois *caractéristique de vie* de la distribution, ayant la même dimension que la variable T .

β est un paramètre de forme, sans dimension.

On supposera que le paramètre γ est égal à 0, dans ces conditions, la loi de Weibull a pour fonction de répartition :

$$F(t) = 0 \quad \forall t < 0$$

$$F(t) = 1 - \exp - \left(\frac{t}{\eta} \right)^\beta \quad \forall t \geq 0$$

et pour densité :

$$f(t) = 0 \quad \forall t < 0$$

$$f(t) = \frac{\beta}{\eta} \left(\frac{t}{\eta}\right)^{\beta-1} \exp - \left(\frac{t}{\eta}\right)^{\beta} \quad \forall t \geq 0$$

Le taux de défaillance immédiate $\lambda(t)$, donné par la formule :

$$\lambda(t) dt = \frac{f(t) dt}{1 - F(t)}$$

est égal pour la loi de Weibull à :

$$\lambda(t) = \frac{\beta}{\eta} \times \left(\frac{t}{\eta}\right)^{\beta-1}$$

Le taux de défaillance est :

- décroissant si $\beta < 1$, période de défaillances précoces, période de rodage...
- constant si $\beta = 1$, loi exponentielle,
- croissant si $\beta > 1$, matériels soumis à l'usure.

L'espérance mathématique de cette loi est :

$$E(X) = \eta \Gamma \left(1 + \frac{1}{\beta}\right)$$

Ajustement graphique : voir chapitre 16, paragraphe 16.1.1.

Estimation des paramètres : voir chapitre 13, exemple 13.10.

18.8 Fiabilité d'un système en fonction de celle de ses composants

Un système est un ensemble de composants assemblés dans un but spécifique.

18.8.1 Système à structure série au sens de la fiabilité

Un système à structure série est caractérisé par la propriété que la défaillance d'un seul de ses composants entraîne la défaillance du système.

Soient R la fiabilité du système, R_i celle du composant i . Si les défaillances des différents composants sont *indépendantes* entre elles, on obtient pour la fiabilité R du système :

$$R = R_1 \times R_2 \times \dots \times R_n$$

La fiabilité d'un système à structure série est plus faible que celle de son composant le plus faible.

Application : système à structure série dont tous les composants obéissent à des lois de fiabilité exponentielles.

La fiabilité du système est elle-même de forme exponentielle, elle a pour expression :

$$R_S(t) = \exp -t \sum_{i=1}^n \lambda_i$$

Le taux de défaillance du système est égal à la somme des taux de défaillance de chaque composant.

Le MTBF du système est égal à :

$$\theta = \frac{1}{\sum_{i=1}^n \lambda_i} = \frac{1}{\sum_{i=1}^n \frac{1}{\theta_i}}$$

c'est-à-dire à la moyenne harmonique des MTBF des composants du système.

18.8.2 Système à structure parallèle au sens de la fiabilité

Un système à structure parallèle est défaillant si tous ses composants sont défaillants. Cette propriété se traduit par la relation (même notation que dans le paragraphe précédent) :

$$1 - R_S(t) = \prod_{i=1}^n [1 - R_i(t)]$$

$$R_S(t) = 1 - \prod_{i=1}^n [1 - R_i(t)]$$

La fiabilité d'un système à structure parallèle est toujours plus élevée que celle de son composant le plus sûr.

Application : système à structure parallèle à deux composants obéissant à des lois de fiabilité exponentielles :

$$R_S(t) = R_1(t) + R_2(t) - R_1(t) R_2(t)$$

$$R_S(t) = e^{-\lambda_1 t} + e^{-\lambda_2 t} - e^{-(\lambda_1 + \lambda_2) t}$$

Le calcul de l'espérance mathématique avec la formule du paragraphe 18.6 conduit au résultat suivant :

$$E(T) = \frac{1}{\lambda_1} + \frac{1}{\lambda_2} - \frac{1}{\lambda_1 + \lambda_2}$$

18.8.3 Système à structure mixte

On peut concevoir différents types de systèmes selon les situations, par exemple une structure en parallèle avec des systèmes en série sur chaque déviation.

On peut aussi avoir des composants n'ayant pas les mêmes lois de fiabilité.

D

Analyse des données

19 • INTRODUCTION À L'ANALYSE DES DONNÉES

D

ANALYSE DES DONNÉES

L'analyse des données s'est développée depuis près d'un siècle. Les premières méthodes ont été élaborées au début du vingtième siècle par Spearman et Pearson (analyse canonique) ; puis, grâce à leurs travaux, Hotelling a donné, vers 1930, les bases de l'analyse en composantes principales ; vers 1973, Benzécri a proposé et étudié l'analyse des correspondances, mais c'est surtout grâce au développement extraordinaire des ordinateurs que ces techniques ont connu un grand essor.

La statistique classique étudie un nombre restreint de caractères mesurés sur un petit nombre d'individus (théories de l'estimation et des tests fondées sur des hypothèses probabilistes restreintes). En revanche, l'analyse des données multidimensionnelles permet de traiter un ensemble de variables observées sur un ensemble d'individus, la notion de données multidimensionnelles se référant surtout au nombre important de variables et non au nombre d'individus concernés.

Contrairement à la démarche utilisée en statistique inférentielle, on ne cherche pas à induire des lois valables pour la population entière à partir des résultats obtenus sur les individus observés. On décrit ou on analyse les données recueillies, les méthodes de l'analyse des données s'apparentent donc plus à la statistique descriptive.

19.1 Échantillon d'une variable aléatoire

Soit un ensemble I de n individus décrits par une variable quantitative X (ou caractère). Cette variable est une application de I dans $\mathbb{R}$ telle que :

$$I \xrightarrow{X} \mathbb{R}$$

$$i \xrightarrow{X} X(i) = x_i$$

L'ensemble E des n valeurs x_i est un échantillon de la variable X . Dans certains cas, la variable X ne prend que des valeurs entières, ou seulement les valeurs 0 ou 1.

À chaque individu i , est attaché un poids p_i tel que :

$$\forall i \in I \quad p_i > 0 \quad \sum_i p_i = 1$$

Remarques

- Les individus ont souvent le même poids, $p_i = 1/n$.
- L'ensemble des poids p_i définit une loi de probabilité sur $(I, p(I))$.

19.1.1 Représentation de l'ensemble E

On peut donner de l'ensemble E deux représentations ou interprétations :

- l'ensemble E peut être considéré comme un sous-ensemble de $\mathbb{R}$ (ensemble des nombres réels) : $E = (x_1, \dots, x_n)$,
- l'ensemble E peut être considéré comme un point ou vecteur de $\mathbb{R}^n$. Si e_i , $i \in [1, n]$, est la base canonique de $\mathbb{R}^n$, le point $\underline{X} = \sum_{i \in I} x_i e_i$ est le vecteur de $\mathbb{R}^n$ associé à l'échantillon E .

Remarques

- La représentation $\underline{X}$ de l'ensemble E dépend de l'ordre des valeurs x_i .
- Le vecteur $\underline{X}$ peut n'occuper que certaines régions de $\mathbb{R}^n$.
- L'échantillon E sera noté $\underline{X}$ quel que soit le mode de représentation choisi.

19.1.2 Valeurs caractéristiques de l'ensemble E

Certaines définitions ont déjà été données dans le chapitre 1, elles sont rappelées pour mémoire et éventuellement complétées dans l'optique « analyse des données ».

■ Caractéristiques de tendance centrale

□ Moyenne

Elle est définie par : $\bar{x} = \sum_{i \in I} p_i x_i$

- Son image, sur la droite $\mathbb{R}$, est le centre de gravité des points x_i munis des poids p_i .
- La moyenne ne suffit pas à caractériser un échantillon.
- On a toujours :

$$\text{Min}(x_1, \dots, x_n) \leq \bar{x} \leq \text{Max}(x_1, \dots, x_n)$$

□ Médiane $M(\underline{X})$

Elle est définie comme le nombre, tel que la somme des poids des valeurs x_i qui lui sont inférieures ou égales soit égale à la somme des poids des valeurs x_i qui lui sont supérieures.

- La médiane n'existe pas toujours.
- Il peut exister un intervalle « médian ».

□ Moyenne des valeurs extrêmes $ME(\underline{X})$

Sa définition est simple :

$$ME(\underline{X}) = \frac{1}{2} [\text{Min}(x_1, \dots, x_n) + \text{Max}(x_1, \dots, x_n)]$$

$$\text{Min}(x_1, \dots, x_n) \leq ME(\underline{X}) \leq \text{Max}(x_1, \dots, x_n)$$

■ Caractéristiques de dispersion

□ Variance et écart-type

$$\text{Variance : Var}(\underline{X}) = \sigma_{\underline{X}}^2 = \sum_{i=1}^n p_i (x_i - \bar{x})^2 = \sum_{i=1}^n p_i x_i^2 - \bar{x}^2$$

Écart-type : $\sigma_{\underline{X}} = \sqrt{\text{Var}(\underline{X})}$

- $\text{Var}(\underline{X}) = 0 \Rightarrow$ l'ensemble E est constitué de n valeurs égales.
- L'écart-type est homogène aux valeurs x_i .

□ Écart moyen $EC(\underline{X})$

$$EC(\underline{X}) = \sum_{i \in I} p_i |x_i - M(\underline{X})|$$

Si la médiane n'est pas unique, c'est-à-dire s'il existe un intervalle médian, on prendra, pour $M(\underline{X})$, dans la définition de $EC(\underline{X})$, n'importe quelle valeur de cet intervalle.

$EC(\underline{X}) = 0 \Rightarrow$ l'ensemble E est constitué de n valeurs égales.

□ Étendue $W(\underline{X})$

$$W(\underline{X}) = \text{Max}(x_1, \dots, x_n) - \text{Min}(x_1, \dots, x_n)$$

$W(\underline{X}) = 0 \Rightarrow$ l'ensemble E est constitué de n valeurs égales.

19.1.3 Valeurs caractéristiques de l'ensemble E associées au choix d'une distance dans $\mathbb{R}^n$

« Réduire » un échantillon consiste à le résumer par une valeur unique T la plus proche possible de l'ensemble des valeurs x_i de l'échantillon, c'est donc chercher un point $T(t, \dots, t)$ de $\mathbb{R}^n$ le plus proche de l'échantillon $\underline{X}$.

D'où la démarche :

- choisir une distance d dans $\mathbb{R}^n$,
- chercher le point T qui minimise $d(T, \underline{X})$.

T est la *caractéristique de tendance centrale* associée à une distance d ; $d(T, \underline{X})$ caractérise la *dispersion* associée à cette distance.

- Distance euclidienne d_1 définie par :

$$d_1^2(T, \underline{X}) = \sum_{i \in I} p_i (t - x_i)^2$$

La distance $d_1(T, \underline{X})$ est minimale pour $t = \bar{x}$ et

$$d_1^2(T, \underline{X}) = \sum_{i \in I} p_i (\bar{x} - x_i)^2 = \text{Var}(\underline{X}) = \sigma_{\underline{X}}^2$$

La *moyenne* est la *caractéristique de valeur centrale* et l'*écart-type* est la *dispersion* associés à la distance euclidienne.

– Distance d_2 définie par :

$$d_2(T, \underline{X}) = \text{Max} [|t - x_i| \text{ } / i = 1, \dots, n]$$

La distance $d_2(T, \underline{X})$ est minimale pour $T = ME(\underline{X})$ et

$$d_2(T, \underline{X}) = \text{Max} [|ME(\underline{X}) - x_i| \text{ } / i = 1, \dots, n] = \frac{W}{2}$$

La *moyenne des valeurs extrêmes* est la *caractéristique de valeur centrale* et la *moitié de l'étendue* est la *dispersion* associées à la distance d_2 .

– Distance d_3 définie par :

$$d_3(T, \underline{X}) = \sum_{i \in I} p_i |t - x_i|$$

La distance d_3 est définie et continue sur $\mathbb{R}$, elle atteint donc un minimum en un point au moins qui est, par définition, la médiane. D'où la définition rigoureuse de la médiane.

On appelle *médiane d'un échantillon* toute valeur qui rend minimale la quantité :

$$\sum_{i \in I} p_i |t - x_i|$$

Si $T = M(\underline{X})$ alors

$$d_3(T, \underline{X}) = \sum_{i \in I} p_i |M(\underline{X}) - x_i| = EC(\underline{X})$$

La *médiane* est la *caractéristique de valeur centrale* et l'*écart moyen* est la *dispersion* associés à la distance d_3 .

■ Généralisation

On peut construire d'autres caractéristiques de valeur centrale et de dispersion en définissant différentes distances dans $\mathbb{R}^n$. Les trois distances étudiées précédemment sont des cas particuliers de la distance de Minkowski définie, pour $r \geq 1$, par :

$$d_r(\underline{x}, \underline{y}) = \left[\sum_{i=1}^n p_i |x_i - y_i|^r \right]^{1/r}$$

$r = 2$: distance d_1 $r = \infty$: distance d_2 $r = 1$: distance d_3

■ Interprétation géométrique de la moyenne et de la variance dans $\mathbb{R}^n$

Soit Δ l'axe de $\mathbb{R}^n$ engendré par le vecteur $\underline{e}$ de composantes $(1, \dots, 1)$. Supposons l'espace $\mathbb{R}^n$ muni de la métrique des poids définie par la matrice diagonale D_p (les éléments non écrits sont des zéros) :

$$D_p = \begin{pmatrix} p_1 & . & . & . & . \\ . & p_2 & . & . & . \\ . & . & p_3 & . & . \\ . & . & . & . & . \\ . & . & . & . & p_n \end{pmatrix}$$

La *moyenne* $\overline{X}$ est la *projection* D_p -orthogonale de $\underline{X}$ sur l'axe Δ ; l'*écart-type* σ_x est la *distance* de $\underline{X}$ à cet axe.

19.1.4 Diagrammes et histogrammes

Pour décrire au mieux un échantillon, il faut compléter les caractéristiques de valeurs centrales et de dispersion par des représentations graphiques. Tous les logiciels de statistique permettent de les construire.

- La variable X est entière : diagramme en bâtons.
- La variable X est quelconque : histogramme.
- Différents types d'histogrammes : symétrique et unimodal, dissymétrique et unimodal, bimodal (échantillon non homogène), uniforme.

19.2 Échantillon d'un couple de variables aléatoires

Soit un ensemble I de n individus décrits par deux variables quantitatives X et Y (ou caractères) ; les variables X et Y sont deux applications de I dans $\mathbb{R}$ telles que :

$$\begin{aligned} I &\xrightarrow{X} \mathbb{R} & I &\xrightarrow{Y} \mathbb{R} \\ i &\xrightarrow{X} X(i) = x_i & i &\xrightarrow{Y} Y(i) = y_i \end{aligned}$$

Un échantillon E du couple (X, Y) est donc l'ensemble des valeurs :

$$E = \{(x_i, y_i) \mid i \in I\}$$

Chaque individu est muni d'un poids p_i tel que :

$$\forall i \in I \quad p_i > 0 \quad \sum_{i \in I} p_i = 1$$

Si les individus ont tous le même poids, alors $p_i = 1/n$.

À cet échantillon E , on peut associer deux échantillons d'une seule variable :

– un échantillon E_x de la variable X :

$$E_x = (x_i, i \in I)$$

– un échantillon E_y de la variable Y :

$$E_y = (y_i, i \in I)$$

L'étude des échantillons E_x et E_y ne suffit pas pour étudier l'échantillon E ; il faut mettre en évidence les liens ou l'absence de liens entre les deux variables X et Y .

19.2.1 Représentation de l'échantillon dans $\mathbb{R}^2$

- Nuage de points dans $\mathbb{R}^2$.
- Centre de gravité $\underline{G}$; en général, il n'appartient pas au nuage.

19.2.2 Covariance et coefficient de corrélation

– Covariance des variables X et Y :

$$\text{Cov}(X, Y) = \sum_{i \in I} p_i (x_i - \bar{x}) (y_i - \bar{y})$$

La covariance est sensible aux changements d'échelle sur les variables X et Y .

– Coefficient de corrélation linéaire (ou plus simplement coefficient de corrélation) entre les variables X et Y :

$$r(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}$$

Le coefficient de corrélation est insensible aux changements d'origine et d'échelle sur les variables X et Y .

– Matrice de variance-covariance associée à l'échantillon E :

$$V = \begin{pmatrix} \text{Var}(X) & \text{Cov}(X, Y) \\ \text{Cov}(X, Y) & \text{Var}(Y) \end{pmatrix}$$

– Matrice de corrélation associée à l'échantillon E :

$$R = \begin{pmatrix} 1 & r(X, Y) \\ r(X, Y) & 1 \end{pmatrix}$$

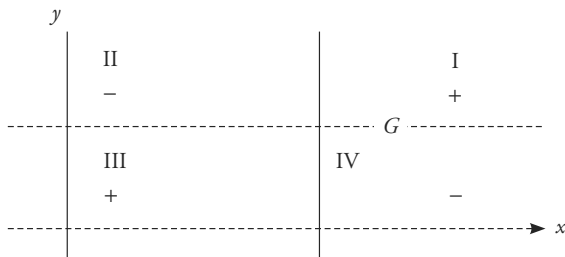
Si $D_{1/\sigma}$ est la matrice diagonale de l'inverse des écarts-types :

$$D_{1/\sigma} = \begin{pmatrix} 1/\sigma_X & 0 \\ 0 & 1/\sigma_Y \end{pmatrix}$$

la matrice R est égale à : $R = D_{1/\sigma} V D_{1/\sigma}$

– Signe de la covariance et du coefficient de corrélation : on représente le nuage de points de l'échantillon dans $\mathbb{R}^2$, ainsi que les parallèles aux axes passant par le centre de gravité du nuage.

On partage le plan en quatre régions I, ... , IV.



- Si le nuage de points a une allure croissante, c'est-à-dire si les points appartiennent aux régions I et III : $\text{Cov}(X, Y) > 0$ et $r(X, Y) > 0$.
- Si le nuage de points a une allure décroissante, c'est-à-dire si les points appartiennent aux régions II et IV : $\text{Cov}(X, Y) < 0$ et $r(X, Y) < 0$.

19.2.3 Interprétation du coefficient de corrélation

Il faut être très prudent dans l'interprétation d'un coefficient de corrélation.

Si $r = 1$ exactement, $y = ax + b$ (relation linéaire)

Dans tous les autres cas, il faut commencer par examiner le nuage de points. En effet, selon la forme de ce nuage, ce coefficient n'a aucune signification : le nuage peut être très hétérogène ou bien la liaison n'est pas linéaire...

19.2.4 Représentation de l'échantillon dans $\mathbb{R}^n$

Dans l'espace vectoriel $\mathbb{R}^n$, muni de la métrique des poids D_p , l'échantillon E a pour image le couple de vecteurs colonnes :

$$\underline{X} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \quad \underline{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

D

ANALYSE DES DONNÉES

19.3 Échantillon de p variables aléatoires

19.3.1 Description d'un tableau : Individus $\times$ Caractères

On généralise facilement les résultats précédents au cas où n individus sont décrits par p caractères quantitatifs. À chaque caractère k , on associe une application $\underline{X}^k$, $k \in [1, p]$ telle que :

$$k \rightarrow X^k(i) = x_i^k$$

x_i^k est la mesure du caractère k pour l'individu i . L'échantillon E des n individus décrits par p caractères est l'ensemble :

$$E = \left\{ (x_i^1, \dots, x_i^p) / i = 1, \dots, n \right\}$$

L'indice inférieur caractérise l'individu, l'indice supérieur le caractère.

■ Types de données

Les *individus* (un client, une région géographique, un animal...) sont les entités de base sur lesquelles on relève un certain nombre de caractéristiques. Les individus peuvent représenter la population entière ou provenir d'un échantillon aléatoire, tiré au hasard dans la population.

Les *variables caractères* peuvent être :

- des *caractéristiques quantitatives* (âge, salaire, dépenses des ménages, teneur en carbone d'un acier...),
- des *caractéristiques qualitatives* (niveau des études, origine socio-professionnelle, lieu d'habitation...); dans ces conditions, les caractères prennent les valeurs 1 si l'individu possède la modalité, 0 sinon.

On suppose, dans de nombreux cas, que les variables quantitatives suivent une loi normale multidimensionnelle; cependant, cette hypothèse, rarement vérifiée, est inutile pour les études géométriques.

■ Représentation des données

Les données quantitatives sont représentées sous la forme d'un tableau à n lignes (individus) et p colonnes (caractères) ou *tableau de description* ou *tableau de contingence* :

$$X = \begin{pmatrix} x_1^1 & \cdots & x_1^k & \cdots & x_1^p \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_i^1 & \cdots & x_i^k & \cdots & x_i^p \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_n^1 & \cdots & x_n^k & \cdots & x_n^p \end{pmatrix}$$

Un tableau de contingence contient les fréquences d'association de deux caractéristiques quantitatives; à l'intersection de la ligne i et de la colonne k , on trouve la valeur prise par la k^e variable sur le i^e individu.

Les caractéristiques qualitatives, ou variables indicatrices, prenant les valeurs 0 ou 1, sont représentées sous forme disjonctive complète.

■ Espaces de représentation des données

Le statisticien a le choix entre les deux espaces de représentation suivants.

□ Espace des individus ou observations

Les éléments d'une ligne du tableau des données sont des nombres représentant les différentes mesures (variables ou caractères) effectuées sur une même *unité statistique*.

La ligne n° i caractérise l'individu i ; un individu $\underline{X}_i$ est un point d'un espace $\mathbb{R}^p$ de dimension p (p nombre de caractères étudiés).

Les individus sont munis de poids p_i , éventuellement différents d'un individu à l'autre. Soit D_p la matrice diagonale des poids (paragraphe 19.1.3).

Le point moyen $\underline{G}$, ou centre de gravité, résume, dans l'espace $\mathbb{R}^p$, l'échantillon E ; la coordonnée n° k ($k = 1, \dots, p$) de ce point est le nombre :

$$g^k = \sum_{i=1}^n p_i x_i^k$$

Si on désigne par I_n le vecteur (colonne) de l'espace dont tous les éléments sont égaux à 1, on obtient :

$$\underline{G} = {}^t X D_p I_n$$

Si, dans l'espace $\mathbb{R}^p$, l'origine est choisie au point $\underline{G}$, les caractères sont appelés *caractères centrés*; le tableau des données est un *tableau centré*.

□ Espace des variables ou caractères

Une colonne est associée à une variable, c'est-à-dire à une mesure effectuée sur les n individus, la colonne n° k caractérise le caractère k .

Une variable $\underline{X}^k$ ($k = 1, \dots, p$) est une liste de n valeurs numériques, c'est un vecteur de l'espace $\mathbb{R}^n$ de dimension n .

En résumé :

- $\mathbb{R}^p$ est l'espace des individus, l'échantillon E est représenté par un nuage de n points.
- $\mathbb{R}^n$ est l'espace des caractères, l'échantillon E est représenté par un nuage de p points.

19.3.2 Matrice de variance-covariance

L'ensemble des variances $V_{ii} = \text{Var}(\underline{X}^i)$ et des covariances $V_{ij} = \text{Cov}(\underline{X}^i, \underline{X}^j)$ est regroupé dans une matrice V , symétrique, de rang égal à p en général, si

$n > p$, sauf s'il existe des relations exactes entre les p variables :

$$V = \begin{pmatrix} V_{11} & V_{12} & \cdot & V_{1p} \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ V_{p1} & \cdot & \cdot & V_{pp} \end{pmatrix}$$

$$V = {}^tX D_p X - \underline{G} {}^tG$$

(tX est la matrice transposée de X , tG le vecteur transposé du vecteur colonne G).

La *matrice R de corrélation* est la matrice suivante où r_{ij} est le coefficient de corrélation linéaire entre les variables $\underline{X}^i$ et $\underline{X}^j$:

$$R = \begin{pmatrix} 1 & r_{12} & \cdot & r_{1p} \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ r_{p1} & \cdot & \cdot & 1 \end{pmatrix}$$

Si $D_{1/\sigma}$ est la matrice des inverses des écarts-types, la matrice R s'écrit :

$$R = D_{1/\sigma} V D_{1/\sigma}$$

19.4 Présentation des principales méthodes

Les différentes méthodes de l'analyse des données ont pour but :

- soit d'expliquer un phénomène, c'est-à-dire de trouver un lien, fonctionnel ou non, entre une ou plusieurs variables expliquées et une ou plusieurs variables explicatives, ce sont les *méthodes explicatives* ou *méthodes d'analyse de dépendance* ;
- soit de décrire et de résumer les informations obtenues, ce sont les *méthodes descriptives* ;
- soit d'analyser et de résumer les informations obtenues, ce sont les *méthodes de prévision*.

Comme exemples, trois méthodes sont décrites dans les chapitres 20, 21 et 22.

19.4.1 Méthodes explicatives

■ Méthodes de régression

Les méthodes de régression consistent à trouver une relation, linéaire ou non, entre une variable expliquée et une ou plusieurs variables explicatives, l'ajustement étant obtenu par la méthode des moindres carrés. Elles sont utilisées quand toutes les variables explicatives sont quantitatives (voir chapitre 20 pour la régression à une seule variable explicative et chapitre 21 pour la régression multiple).

■ Analyse de la variance

Cette méthode consiste à tester l'influence d'une ou plusieurs variables qualitatives sur une variable quantitative (durée de vie d'ampoules électriques en fonction de la nature du filament, par exemple). On cherche à contrôler si une variation des modalités prises par les variables explicatives, seules ou combinées, entraîne une variation de la variable expliquée Y (voir chapitre 16, paragraphe 16.3 pour l'analyse à simple entrée et chapitre 22, pour l'analyse à entrées multiples).

■ Analyse de la covariance

Cette méthode généralise les méthodes de régression et de l'analyse de la variance. Dans un modèle d'analyse de la variance, la valeur prise par la variable expliquée est déterminée, au terme résiduel ε près, par la classe dans laquelle est faite l'observation. On peut imaginer un modèle où il intervient des variables discrètes et des variables continues. La variable expliquée Y dépend :

- d'une variable qualitative prenant q modalités ;
- pour chaque modalité d'une variable continue X .

Dans la classe i , $i \in [1, p]$, l'observation n° k est déterminée par le modèle général suivant :

$$y_{ik} = (\mu + \alpha_i) + (\eta + \beta_i) x_{ik} + \varepsilon_{ik}$$

Supposons que l'on étudie les dépenses pour l'habillement d'un échantillon de n individus ; on peut étudier ces dépenses en fonction de la classe socio-professionnelle de l'individu (variable discrète prenant dix valeurs) et pour chaque valeur de l'indice i en fonction du revenu de l'individu (variable continue).

Si les deux droites correspondant à $i = 1$ et à $i = 2$ par exemple, sont parallèles, c'est-à-dire si $\beta_1 = \beta_2$, les dépenses pour l'habillement ne dépendent pas de ces deux classes. La distance $\alpha_1 - \alpha_2$ mesure l'effet du facteur classe socio-professionnelle pour ces deux classes considérées. En revanche, si $\beta_1 \neq \beta_2$, on peut conclure à un effet différent pour les dépenses en habillement, à revenus égaux, pour ces deux catégories socio-professionnelles.

■ Analyse canonique

L'analyse canonique développée par Hotelling généralise la méthode de régression multiple, mais présente un intérêt théorique assez limité car elle conduit à de grandes difficultés d'interprétation. Cette méthode cherche à synthétiser les relations pouvant exister entre deux groupes de variables, en déterminant les combinaisons linéaires des variables du premier groupe les plus corrélées à des combinaisons linéaires des variables du second groupe. Si le second groupe est constitué d'une seule variable, on retrouve la régression multiple.

19.4.2 Méthodes descriptives

■ Méthodes de classification

Ces méthodes ont pour but de regrouper des individus, décrits par un certain nombre de variables ou de caractères, en un nombre restreint de classes de sorte que :

- les individus appartenant à une même classe sont le plus semblable possible ;
- les classes sont bien séparées.

■ Analyse en composantes principales

L'analyse en composantes principales (ACP en abrégé), due en particulier à Pearson et Hotelling, a pour but d'étudier les liens existant entre p variables mesurées sur n individus, d'éliminer les redondances (deux variables corrélées apportant à peu près la même information) et de remplacer les variables initiales par un petit nombre de variables, 1, 2 ou 3, appelées composantes principales. Ces variables sont des combinaisons linéaires des variables initiales non corrélées entre elles.

■ Analyse factorielle des correspondances

Cette méthode, proposée par Benzécri, vers 1973, pour l'étude des tableaux de contingence est devenue la méthode privilégiée pour la description des données qualitatives et un outil puissant pour le dépouillement des enquêtes. Le tableau des données contient les fréquences observées des modalités de deux phénomènes. Le test du chi-deux permet de déterminer s'il existe une liaison entre ces deux phénomènes, l'analyse factorielle des correspondances décrit cette liaison. Une analyse en composantes principales effectuée sur un tableau de contingence peut mettre en évidence des ressemblances entre les colonnes du tableau, entre les lignes ou des proximités entre les lignes et les colonnes.

■ Analyse factorielle discriminante

Sur l'ensemble des individus d'une population P , on étudie p caractères quantitatifs et un caractère qualitatif prenant un nombre fini k de modalités. La population est répartie en k classes.

Le but de l'analyse factorielle discriminante est de rechercher si ce caractère qualitatif a une influence sur les p variables mesurées et de déterminer, éventuellement, des caractères discriminants, définissant sur l'ensemble des individus, une partition aussi proche que possible de la partition induite par la variable qualitative initiale.

L'analyse factorielle discriminante se ramène à une analyse en composantes principales, effectuée sur l'ensemble des centres de gravité des individus d'une même classe, chaque classe correspondant à une des k modalités de la variable qualitative initiale.

19.4.3 Méthodes de prévision

Ces méthodes concernent principalement *l'analyse et la prévision des séries chronologiques*.

Elles ont principalement pour but de mettre en évidence une tendance, une saisonnalité et un résidu à l'aide d'un modèle multiplicatif, le plus utilisé en gestion, ou d'un modèle additif.

20 • RÉGRESSION LINÉAIRE SIMPLE

20.1 Introduction

Dans le domaine des sciences appliquées, on observe fréquemment des phénomènes tels qu'il est possible de supposer l'existence d'une liaison entre deux variables. Par exemple :

- les dépenses annuelles d'un ménage sont fonction des revenus de la famille,
- la durée de vie d'une ampoule électrique peut être liée à son rendement énergétique.

Dans une étude statistique, on mesure, sur chaque unité d'un échantillon, différentes variables et on cherche s'il existe une certaine forme d'association entre elles. Le cas le plus simple est celui d'une dépendance statistique ou corrélation.

Il y a *corrélation* entre deux variables observées sur les éléments d'une population si les *variations* de ces deux variables *se produisent dans le même sens* (*corrélation positive*) ou *en sens contraires* (*corrélation négative*).

Dans le chapitre 17, relatif aux tests d'indépendance, des méthodes permettant de mettre en évidence et de mesurer l'intensité de la liaison pouvant exister entre deux variables aléatoires X et Y ont été présentées.

En *régression*, le problème est de nature différente. On dispose de n couples (x_i, y_i) constituant un échantillon d'observations indépendantes du couple de variables X et Y . On cherche une relation statistique pouvant exister entre la variable expliquée Y et la variable explicative X . Cette relation doit permettre de prévoir la valeur de Y pour une valeur donnée de X . Le problème est de

nature dissymétrique. Au moins, trois questions se posent :

- 1) Quel est le modèle statistique le mieux adapté pour décrire la liaison entre les variables X et Y ? Doit-on utiliser un modèle linéaire, parabolique, exponentiel, etc. ?
- 2) En admettant comme plausible un modèle particulier, comment estimer les paramètres figurant dans ce modèle ?
- 3) Comment définir les outils permettant de calculer les valeurs prévisionnelles de la variable Y en fonction de la variable X ?

20.2 Mesures de liaison

Les mesures de liaison entre des variables qualitatives ou quantitatives, étudiées dans le chapitre 17, sont le coefficient de corrélation linéaire et le rapport de corrélation.

20.2.1 Coefficient de corrélation linéaire

Dans le cas de variables numériques, le coefficient de corrélation linéaire ρ :

$$\rho = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}$$

donne une bonne indication sur l'intensité de la liaison selon sa valeur, et son signe permet de voir si les deux variables varient dans le même sens ou non.

Si la corrélation linéaire se révèle significative, on peut trouver, à l'aide d'une méthode d'ajustement appropriée, la fonction décrivant la liaison. Si on se limite à une liaison linéaire, on a obtenu la *droite de régression*.

Remarques

– Le coefficient de corrélation linéaire ne mesure pas une relation de cause à effet entre deux variables. En effet, deux variables peuvent être corrélées sans que les variations d'une variable entraînent les variations de l'autre ; cela signifie seulement que les variations des deux variables sont dues à une même cause commune extérieure. Un été particulièrement chaud peut entraîner une augmentation de la vente des crèmes solaires et des crèmes glacées sans qu'il existe une relation entre la vente de ces deux produits. Il faut toujours être très prudent dans l'interprétation des résultats d'une analyse de corrélation ou de régression.

– La nullité du coefficient de corrélation linéaire n'entraîne pas l'indépendance, sauf pour des variables aléatoires gaussiennes.

20.2.2 Dépendance entre deux variables

Les variables aléatoires X et Y ne sont pas indépendantes si la loi conditionnelle Y/X est différente de la loi marginale de Y . Cette notion étant très restrictive, on utilise la notion de dépendance en moyenne. Par définition :

La variable Y est *corrélée en moyenne* avec la variable X si l'espérance conditionnelle de Y sachant X dépend des valeurs prises par la variable X , donc si $E(Y/X = x) = \varphi(x)$.

X et Y sont non corrélées réciproquement si :

$$E(Y/X) = E(Y) \quad \text{et} \quad E(X/Y) = E(X)$$

En général, ces deux propriétés ne sont pas vraies simultanément et de plus, la non-corrélation n'entraîne pas l'indépendance.

20.2.3 Rapport de corrélation

Le rapport de corrélation est un coefficient dissymétrique entre deux variables, défini par :

$$\eta_{Y/X}^2 = \frac{\text{Var}[E(Y/X)]}{\text{Var} Y}$$

$$\eta_{Y/X}^2 = 0 \Rightarrow \text{absence de dépendance en moyenne et } E(Y/X) = \text{cste p.s.}$$

$$\eta_{Y/X}^2 = 1 \Rightarrow E[\text{Var}(Y/X)] = 0 \Rightarrow \text{Var}(Y/X) = 0 \text{ p.s.}$$

à une valeur de X , correspond une seule valeur de Y , d'où $Y = \varphi(x)$.

Si la dépendance entre les variables X et Y est linéaire, donc si

$$E(Y/X) = \alpha + \beta X$$

ou si $E(X/Y) = \gamma + \delta Y$, le rapport de corrélation est égal au carré du coefficient de corrélation linéaire $\eta_{Y/X}^2 = \rho^2$.

20.3 Choix des variables

Si la corrélation linéaire entre les variables X et Y est significative (relation mise en évidence soit par des tests, soit par une méthode graphique), on cherche l'équation de la droite traduisant au mieux cette relation.

Si les variables aléatoires X et Y ne sont pas indépendantes et s'il est logique d'expliquer Y en fonction de X , il n'en sera pas de même de X en fonction de Y ; ceci peut d'ailleurs n'avoir aucun sens. Il faut donc définir la variable qui sera expliquée en fonction de l'autre. On note, en général :

- Y la *variable expliquée* ou *critère* qui est toujours une variable aléatoire,
- X la *variable explicative* ou *prédicteur* qui peut être une variable aléatoire ou non.

Si les variables X et Y sont des variables aléatoires, on peut étudier l'intensité de la liaison existant entre ces deux variables, l'analyse de la corrélation permet d'évaluer le degré de dépendance.

20.4 Modèle théorique de la régression simple

20.4.1 Approximation conditionnelle

X et Y étant deux variables aléatoires, on cherche une fonction f telle que $f(X)$ soit aussi proche que possible de Y en moyenne quadratique. Au sens des moindres carrés :

- la meilleure approximation de Y par une constante est l'espérance mathématique $E(Y)$,
- la meilleure approximation de Y par une fonction $\varphi(X)$ est l'espérance conditionnelle $E(Y/X)$:

$$E \left[(Y - f(X))^2 \right] \text{ est minimale si } f(X) = E(Y/X)$$

Le *rapport de corrélation* mesure la qualité de l'approximation :

$$\eta_{Y/X}^2 = \frac{\text{Var} [E(Y/X)]}{\text{Var}(Y)} = \frac{\text{Variation expliquée}}{\text{Variation totale}}$$

La fonction $E(Y/X)$ qui, à chaque valeur x de X , associe $E(Y/X = x)$ est la *fonction de régression* de Y en X et son graphe est la *courbe de régression* de Y en X .

Si ε désigne un résidu aléatoire, on pose :

$$Y = E(Y/X) + \varepsilon$$

Propriétés du résidu ε :

- $E(\varepsilon) = 0$ car $E(Y) = E[E(Y/X)]$.
- ε n'est pas corrélé avec X ni avec $E(Y/X)$ car on montre que ε est orthogonal à l'espace des variables aléatoires fonction de X .
- $\text{Var}(\varepsilon) = (1 - \eta_{Y/X}^2) \text{Var}(Y)$

20.4.2 Cas où la régression est linéaire

Le cas le plus important en pratique est celui où $E(Y/X) = \alpha + \beta X$.

Cette propriété est vérifiée si le couple de variables (X, Y) est un couple gaussien. L'équation de la droite de régression est :

$$Y = \alpha + \beta X + \varepsilon$$

Calcul des coefficients α et β :

- l'espérance mathématique des deux membres de l'équation précédente donne après simplification :

$$Y - E(Y) = \beta [X - E(X)] + \varepsilon$$

La droite de régression passe donc par le point de coordonnées $E(X), E(Y)$.

- on multiplie chaque membre de la relation précédente par $[X - E(X)]$ et on calcule l'espérance mathématique :

$$\text{Cov}(X, Y) = \beta \text{Var}(X) + \text{Cov}(\varepsilon, X)$$

Le terme $\text{Cov}(\varepsilon, X)$ est nul car le résidu ε n'est pas corrélé avec la variable X . D'où :

$$\beta = \frac{\text{Cov}(X, Y)}{\text{Var}(X)} = \rho \frac{\sigma_Y}{\sigma_X}$$

et l'équation de la droite de régression est :

$$Y - E(Y) = \rho \frac{\sigma_Y}{\sigma_X} [X - E(X)] + \varepsilon$$

Le résidu ε n'étant pas corrélé avec X , on obtient, en calculant la variance des deux membres :

$$\rho^2 = \eta_{Y/X}^2$$

20.5 Ajustement du modèle de régression linéaire sur des données expérimentales

On dispose d'un échantillon de n observations indépendantes (x_i, y_i) du couple de variables (X, Y) .

20.5.1 Approche descriptive

Une représentation graphique du nuage de points dans un plan donne une première indication sur la nature de la liaison pouvant exister entre ces variables.

On calcule le *coefficient de corrélation empirique* :

$$\begin{aligned}
 r &= \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x}) (y_i - \bar{y})}{\sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 \times \frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2}} \\
 &= \frac{\sum_{i=1}^n (x_i - \bar{x}) (y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \times \sum_{i=1}^n (y_i - \bar{y})^2}}
 \end{aligned}$$

Ce nombre, sans dimension, compris entre -1 et $+1$, symétrique par rapport aux variables X et Y , donne une indication sur l'intensité de la relation linéaire entre ces variables.

- Les valeurs extrêmes traduisent la *corrélation parfaite*, positive si $r = +1$, négative si $r = -1$. Ces cas extrêmes sont rares ; cependant, si r est voisin de $+1$ ou de -1 , les points de coordonnées (x_i, y_i) sont sensiblement alignés.
- Si les variables aléatoires X et Y sont indépendantes, $r = 0$, mais la réciproque n'est pas vraie (sauf pour des variables aléatoires gaussiennes). Une

valeur du coefficient r voisine de 0 ne traduit pas l'indépendance mais l'absence de relation linéaire entre X et Y .

20.5.2 Recherche du modèle de régression

■ Modèle de régression linéaire

Après examen du nuage de points ou après le calcul du coefficient de corrélation, on suppose que la corrélation entre les variables X et Y est significative. Le modèle le plus simple que l'on étudie est le *modèle de régression linéaire* :

$$Y = \alpha + \beta X + \varepsilon$$

Les coefficients α et β sont les paramètres du modèle.

Le terme ε est un résidu aléatoire.

Pour appliquer le modèle de régression linéaire, il faut faire les hypothèses suivantes :

- la variable expliquée Y et la variable explicative X sont des variables aléatoires,
- la courbe de régression (courbe joignant les diverses moyennes de Y pour les différentes valeurs de X) est une droite, c'est-à-dire $E(Y/X) = \alpha + \beta X$,
- la variance de la distribution du résidu ε ne dépend pas des valeurs prises par la variable X .

■ Modèle linéaire simple

La variable Y est une variable *aléatoire*, mais la variable X *n'est pas* une variable *aléatoire* ; c'est une variable mesurée sans erreur ou à niveaux fixés. Dans ces conditions, on pose pour chaque valeur x_i de X :

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

Chaque valeur y_i est la somme de deux termes :

- un terme $(\alpha + \beta x_i)$ non aléatoire,
- une composante aléatoire ε_i tenant compte du caractère aléatoire de Y . L'ensemble de ces composantes représente la fluctuation des valeurs y_i pour chaque valeur de x_i , autour des valeurs $(\alpha + \beta x_i)$. Cette fluctuation est due soit à des facteurs non contrôlables, soit à des variables indépendantes non

prises en compte dans le modèle. La variable aléatoire ε est une variable non observable.

Remarques

- L'adjectif « simple » indique la présence d'une seule variable explicative, l'adjectif « linéaire » s'applique aux paramètres du modèle.
- Un modèle non linéaire peut être rendu linéaire par des transformations appropriées. C'est le cas des modèles $y = \alpha x^\beta$ et $y = \alpha e^{\beta x}$.
- Le modèle $Y = \alpha + \beta X + \varepsilon$ est un modèle linéaire d'ordre 1 à une seule variable explicative.

La méthode utilisée pour obtenir une droite qui s'ajuste le mieux possible au diagramme de dispersion, dans le cas de la régression linéaire ou dans le cas du modèle linéaire, est la méthode des moindres carrés qui consiste à *rendre minimale la somme des carrés des écarts des valeurs observées y_i à la droite*.

Cette méthode, ne dépendant que des lois conditionnelles pour la variable X fixée, utilise les mêmes techniques pour les deux modèles. Cependant, on parle de corrélation seulement dans le cas où la variable X est aléatoire.

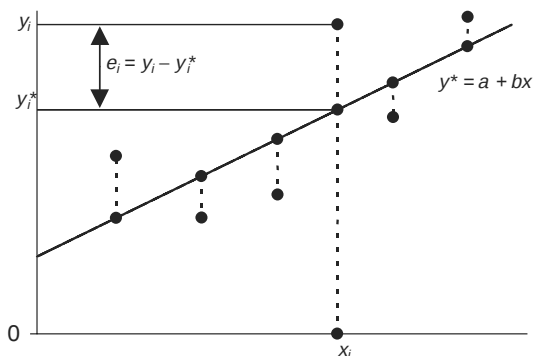
La droite de régression donne une bonne approximation de la relation fonctionnelle qui peut exister entre les variables X et Y , mais ne permet pas d'établir avec exactitude cette relation fonctionnelle.

20.6 Étude de la régression linéaire (aspects descriptifs)

Soit $y^* = a + bx$ l'équation de la droite des moindres carrés, $y_i^* = a + bx_i$ la valeur calculée et $e_i = y_i - y_i^*$ la valeur résiduelle, ou écart.

Les coefficients a et b de la droite des moindres carrés vérifient la propriété :

$$\text{Rendre } \sum_{i=1}^n (y_i - y_i^*)^2 \text{ minimum}$$

Figure 20.1 – Nuage de points (x_i, y_i) et droite des moindres carrés.

20.6.1 Calcul des coefficients a et b

$$\sum_{i=1}^n (y_i - a - bx_i)^2 = F(a, b)$$

Le minimum de $F(a, b)$ est obtenu pour :

$$\frac{dF(a, b)}{da} = 0 \Rightarrow \sum_{i=1}^n (y_i - a - bx_i) = 0$$

$$\frac{dF(a, b)}{db} = 0 \Rightarrow \sum_{i=1}^n x_i (y_i - a - bx_i) = 0$$

La première équation a pour solution :

$$\bar{y} = a + b\bar{x}$$

et la deuxième, compte tenu de la première solution :

$$b = \frac{\sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} = r \frac{s_Y}{s_X}$$

D'où, l'équation de la droite des moindres carrés :

$$y^* = \bar{y} + r \frac{s_Y}{s_X} (x - \bar{x})$$

Cette droite passe par le centre de gravité $(\bar{x}, \bar{y})$ du nuage de points. Elle a pour pente *un coefficient empirique* analogue à la pente de la droite de régression.

Les valeurs y_i et, dans le cas de la régression linéaire, les valeurs x_i étant des réalisations de variables aléatoires, il en est de même pour tous les coefficients de la droite des moindres carrés r , s_X , s_Y ou a et b .

20.6.2 Contribution de chaque observation à la droite des moindres carrés

La pente b de la droite des moindres carrés peut s'écrire :

$$b = \sum_{i=1}^n \left[\frac{(x_i - \bar{x})^2}{\sum_i (x_i - \bar{x})^2} \right] \times \frac{y_i - \bar{y}}{x_i - \bar{x}}$$

Cette pente est la moyenne pondérée des pentes des droites passant par le centre de gravité $(\bar{x}, \bar{y})$ et chaque observation.

La pondération de l'observation n° i est égale à $\frac{(x_i - \bar{x})^2}{\sum_i (x_i - \bar{x})^2}$. Elle mesure

la contribution de cette observation dans le calcul de la pente. Elle donne l'impact d'un point éloigné de la variable X .

En fait, cet impact est donné par le *levier* défini pour chaque observation par :

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_i (x_i - \bar{x})^2}$$

La moyenne des leviers est égale à $\bar{h} = \frac{2}{n}$.

On considère qu'un *levier* est *important* s'il est *supérieur* à $4/n$.

20.6.3 Décomposition de la variation totale

Comme dans l'analyse de la variance, on décompose la variation totale

$\sum_{i=1}^n (y_i - \bar{y})^2$ en une somme de deux carrés faisant intervenir :

- une somme de carrés due à la régression $\sum_{i=1}^n (y_i^* - \bar{y})^2$ ou variation expliquée par la régression ; cette variation est égale à $b^2 \sum_{i=1}^n (x_i - \bar{x})^2$,
- une somme de carrés résiduelle $\sum_{i=1}^n (y_i - y_i^*)^2$ ou variation résiduelle :

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (y_i^* - \bar{y})^2 + \sum_{i=1}^n (y_i - y_i^*)^2$$

La variation totale représente la variance des valeurs y_i autour de leur moyenne, la variation expliquée par la régression dépend de la pente et des valeurs de la variable X .

20.6.4 Coefficient de détermination et coefficient de corrélation

Le coefficient de détermination R^2 est défini par le rapport :

$$R^2 = \frac{\text{Variation expliquée par la régression}}{\text{Variation totale}} = \frac{\sum_{i=1}^n (y_i^* - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

Ce coefficient donne la part de la variance de la variable Y expliquée par la régression. Il est compris entre 0 et 1.

$$R^2 = 1 \Rightarrow \sum_{i=1}^n (y_i - y_i^*)^2 = 0 \quad y_i = y_i^* = a + bx_i$$

La liaison est parfaite, les points sont alignés.

$$R^2 = 0 \Rightarrow \sum_{i=1}^n (y_i^* - \bar{y})^2 = 0 \quad y_i^* = \bar{y}$$

Les points sont alignés sur une droite *horizontale* ($b = 0$), il n'y a aucune liaison entre les variables X et Y .

Un calcul rapide conduit à $r^2 = R^2$. En effet :

$$r^2 = b^2 \frac{s_X^2}{s_Y^2} = \frac{\sum_{i=1}^n (y_i^* - \bar{y})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \times \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = R^2$$

$r > 0 \Rightarrow b > 0 \Rightarrow$ corrélation positive, liaison croissante.

$r < 0 \Rightarrow b < 0 \Rightarrow$ corrélation négative, liaison décroissante.

Le coefficient de corrélation mesure, à la fois, la force et le sens de la liaison.

20.7 Étude de la régression linéaire (aspects inférentiels)

20.7.1 Estimation des coefficients α et β

Les coefficients a et b et la valeur y^* sont des *estimations sans biais* de α , β et de $E(Y/X)$ (régression) ou de $\alpha + \beta x$ (modèle linéaire).

- b est une réalisation de la variable aléatoire B :

$$B = \frac{\sum_{i=1}^n (Y_i - \bar{Y}) (X_i - \bar{X})}{\sum_{i=1}^n (X_i - \bar{X})^2}$$

Pour démontrer que $E(B) = \beta$, on calcule l'espérance conditionnelle $E(B/X_j = x_j)$, (les valeurs de la variable X sont fixées), puis on applique le théorème de l'espérance totale.

- a est une réalisation de la variable aléatoire A :

$$A = \bar{Y} - B\bar{X}$$

La même méthode conduit à $E(A) = \alpha$.

Comme $E(Y/X = x) = \alpha + \beta x$, on en déduit que $y^* = a + bx$ est une estimation sans biais de $\alpha + \beta X$.

Pour démontrer que la variable B n'est pas corrélée avec la variable $\bar{Y}$, on calcule $\text{Cov}(B, \bar{Y})$ pour les valeurs x_i fixées de X :

$$\text{Cov}(B, \bar{Y}) = \frac{\sigma^2}{n \sum_{i=1}^n (x_i - \bar{x})^2} \sum_{i=1}^n (x_i - \bar{x}) = 0$$

Les variables B et $\bar{Y}$ ne sont pas corrélées conditionnellement aux valeurs x_i , elles ne sont donc pas corrélées marginalement.

La qualité des estimateurs est donnée par le théorème de Gauss-Markov :

Parmi les estimateurs sans biais de α et β , fonctions linéaires des Y_i , A et B sont les *estimateurs de variance minimale*.

Avec les mêmes méthodes, on calcule les variances conditionnelles de A et B .

On pose $s_X^2 = \frac{1}{n} \sum_i (x_i - \bar{x})^2$:

$$\text{Var}(A) = \frac{\sigma^2}{n} \left(1 + \frac{\bar{x}^2}{s_X^2} \right) \quad \text{Var}(B) = \frac{\sigma^2}{n s_X^2}$$

Une estimation sans biais de la variance σ^2 est donnée par :

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^n (y_i - y_i^*)^2$$

20.7.2 Propriétés des écarts résiduels

Les écarts résiduels sont définis par : $e_i = y_i - y_i^* = y_i - [\bar{y} + b(x_i - \bar{x})]$.

La moyenne des écarts e_i est nulle (calcul facile).

La variance des écarts e_i ou *variance résiduelle*, notée $s_{Y/X}^2$, est égale à :

$$\begin{aligned} s_{Y/X}^2 &= \frac{1}{n} \sum_{i=1}^n e_i^2 \\ &= \frac{1}{n} \left[\sum_{i=1}^n (y_i - \bar{y})^2 - 2b \sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x}) + b^2 \sum_{i=1}^n (x_i - \bar{x})^2 \right] \\ s_{Y/X}^2 &= s_Y^2 - 2b \operatorname{Cov}(x, y) + b^2 s_X^2 = (1 - r^2) s_Y^2 \end{aligned}$$

20.7.3 Cas particulier : le résidu ε suit une loi normale

Si le résidu ε suit une loi normale $N(0; \sigma)$, on obtient les résultats suivants :

- les lois conditionnelles $Y/X = x$ sont des lois normales $N(\alpha + \beta x; \sigma)$,
- les variables aléatoires A , B et Y^* sont des combinaisons linéaires, pour x fixé, de variables aléatoires gaussiennes, elles suivent donc des lois normales :

$$\text{Loi de la variable } A \quad N\left(\alpha; \sigma \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{n s_X^2}}\right)$$

$$\text{Loi de la variable } B \quad N\left(\beta; \frac{\sigma}{\sqrt{n s_X^2}}\right)$$

$$\text{Loi de la variable } Y^* \quad N\left(\alpha + \beta x; \sigma \sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{n s_X^2}}\right)$$

- les variables aléatoires A , B et Y^* sont les estimateurs de variance minimale de α , β et σ^2 ,
- la variable aléatoire

$$\frac{(n-2) \hat{\sigma}^2}{\sigma^2} = \frac{1}{\sigma^2} \sum_{i=1}^n (y_i - y_i^*)^2 = \frac{n s_{Y/X}^2}{\sigma^2}$$

est une réalisation d'une variable aléatoire, indépendante de A , B et $\bar{Y}$; elle suit une loi du chi-deux à $(n-2)$ degrés de liberté.

Remarques

– Les lois suivies par les variables aléatoires A et B supposent la variance σ^2 connue. Cette variance étant inconnue, on utilise son estimateur. Donc, les variables

$$\frac{A - \alpha}{s_{Y/X} \sqrt{1 + \frac{\bar{x}^2}{s_X^2}}} \sqrt{n-2} \quad \text{et} \quad \frac{(B - \beta) s_X}{s_{Y/X}} \sqrt{n-2}$$

suivent des lois de Student $T(n-2)$. On peut en déduire des intervalles de confiance pour les coefficients α et β .

– Supposons $\rho = 0$, alors $\beta = 0$. En remplaçant B et $s_{Y/X}$ par leurs expressions en fonction de R et des écarts-types s_X et s_Y , dans la loi de β , on trouve que la variable aléatoire $\frac{R}{\sqrt{1-R^2}} \sqrt{n-2}$ suit une loi de Student $T(n-2)$ (résultat donné sans démonstration dans le chapitre 17, paragraphe 17.1.1).

20.7.4 Tests dans le modèle linéaire

On suppose que le résidu ε suit la loi normale $N(0; 1)$.

■ Analyse de variance de la régression

Les hypothèses à tester sont :

$$H_0 : \beta = 0$$

$$H_1 : \beta \neq 0$$

L'hypothèse H_0 est une hypothèse de non-régression, équivalente dans le cas où les variables X et Y sont des variables aléatoires, au test :

$$H_0 : \rho = 0$$

$$H_1 : \rho \neq 0$$

On décompose la variation totale $\sum_{i=1}^n (y_i - \bar{y})^2$ en une somme de deux carrés (paragraphe 20.6.3) :

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (y_i^* - \bar{y})^2 + \sum_{i=1}^n (y_i - y_i^*)^2$$

- La variable aléatoire $\frac{1}{\sigma^2} \sum_{i=1}^n (Y_i - Y_i^*)^2$ est une variable chi-deux à $(n-2)$ degrés de liberté.
- Sous l'hypothèse H_0 de non-régression, la variable aléatoire $\frac{1}{\sigma^2} \sum_{i=1}^n (Y_i - \bar{Y})^2$ est une variable chi-deux à $(n-1)$ degrés de liberté.
- La variable aléatoire $\frac{1}{\sigma^2} \sum_{i=1}^n (Y_i^* - \bar{Y})^2 = \frac{1}{\sigma^2} \sum_{i=1}^n B^2 (X_i - \bar{X})^2$ est une variable chi-deux à 1 degré de liberté.

D'où le résultat : sous l'hypothèse H_0 , $\beta = 0$, la variable aléatoire

$$\frac{\sum_{i=1}^n (Y_i^* - \bar{Y})^2}{\sum_{i=1}^n (Y_i^* - Y_i)^2} \times (n-2)$$

est une variable de Fisher $F(1; n-2)$. Le test en découle immédiatement.

Ce rapport étant égal à $\frac{R^2}{1-R^2} (n-2)$, les deux tests donnés au début du paragraphe sont équivalents.

D

ANALYSE DES DONNÉES

■ Test d'une équation de régression spécifiée

Les hypothèses à tester sont :

$$H_0 : \alpha = \alpha_0 \quad \beta = \beta_0$$

$$H_1 : \alpha \neq \alpha_0 \quad \beta \neq \beta_0$$

Les variables aléatoires A et B n'étant pas indépendantes, on ne peut pas tester α , puis tester β .

Si l'hypothèse H_0 est vraie, la quantité

$$\frac{1}{2\hat{\sigma}^2} \left[n(a - \alpha_0)^2 + 2n\bar{x}(a - \alpha_0)(b - \beta_0) + (b - \beta_0)^2 \sum_{i=1}^n x_i^2 \right]$$

est une variable aléatoire suivant une loi de Fisher $F(2 ; n - 2)$. On rejette l'hypothèse H_0 si cette quantité, calculée sur les données observées, est trop grande.

■ Test de linéarité de la régression

L'hypothèse à tester consiste à vérifier la validité du modèle linéaire :

$$E(Y/X) = \alpha + \beta X$$

Ce test nécessite d'avoir des observations répétées y_{ij} de la variable Y pour chaque valeur x_i de la variable X . Dans l'hypothèse de linéarité, le coefficient de corrélation linéaire ρ^2 est égal au rapport de corrélation. Il faut donc comparer le coefficient de corrélation empirique r^2 au rapport de corrélation empirique e^2 défini par :

$$e^2 = \frac{1}{s_Y^2} \times \frac{1}{n} \sum_{i=1}^n n_i (\bar{y}_i - \bar{y})^2 \quad \bar{y}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij}$$

Si l'hypothèse H_0 est vraie, $\eta_{Y/X}^2 = \rho^2$ ou $E(Y/X) = \alpha + \beta X$, la quantité

$$\frac{(e^2 - r^2) / (k - 2)}{(1 - e^2) / (n - k)}$$

est une variable de Fisher $F(k - 2 ; n - k)$, k étant le nombre de valeurs distinctes de X .

De la même façon, on peut tester $H_0 : \eta_{Y/X}^2 = 0$ et $H_1 : \eta_{Y/X}^2 \neq 0$.

Si l'hypothèse H_0 est vraie, la quantité

$$\frac{e^2 / (k - 1)}{(1 - e^2) / (n - k)}$$

est une variable de Fisher $F(k - 1 ; n - k)$.

Remarque

Les propriétés de la méthode des moindres carrés et les tests mis en œuvre supposent que le résidu ε a une variance constante, quelle que soit la valeur x de X , et qu'il n'y a pas auto-corrélation entre les diverses réalisations de ε . Ces hypothèses devraient être toujours vérifiées.

20.7.5 Étude des résidus

L'étude des résidus fait apparaître différentes situations (figure 20.2).

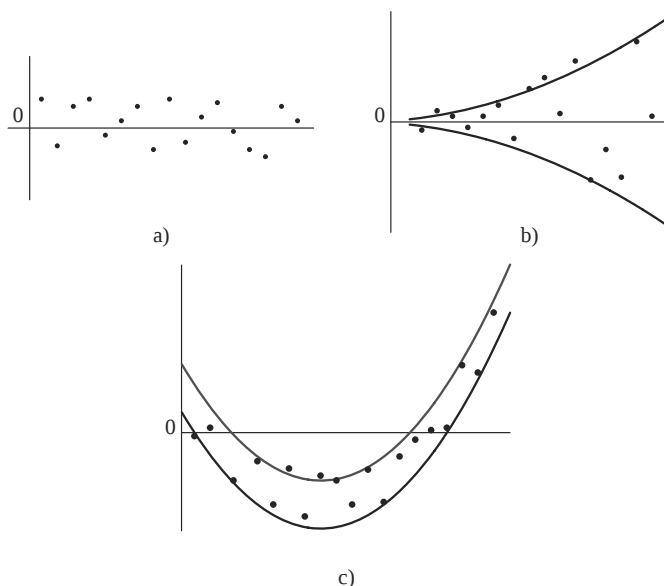


Figure 20.2 – Étude des résidus. (a) La situation est correcte. (b) La variance des résidus ne semble pas constante. La variable $\ln Y$ devrait stabiliser les variances. (c) Une liaison quadratique du type $Y = aX^2 + bX + c$ serait mieux adaptée.

L'erreur observée e_i suit la loi normale $N\left(0 ; \sigma\sqrt{1-h_i}\right)$ où h_i est le levier relatif à l'observation n° i (paragraphe 20.6.2). Pour étudier l'importance de l'erreur, on peut utiliser deux indices.

■ Résidu studentisé

Il est défini par :

$$t_i = \frac{e_i}{\hat{\sigma}\sqrt{1-h_i}}$$

Une observation est mal reconstituée si la quantité $|t_i|$ est grande. Comme on peut identifier la loi des résidus t_i à une loi normale, on considérera alors qu'une observation est mal reconstituée si $|t_i|$ est supérieure à 2 (on peut admettre 2,5 ou 3).

■ R Student

On étudie la régression de la variable Y sur la variable X sans l'observation n° i .

Pour mesurer l'importance de l'erreur e_i , on calcule l'estimation $\hat{\sigma}(i)$ de l'écart-type σ sans tenir compte de l'observation n° i . Cette estimation est égale à :

$$\hat{\sigma}^2(i) = \frac{\hat{\sigma}^2}{n-3} (n-2-t_i^2)$$

Donc, si on enlève l'observation qui a le plus fort résidu studentisé t_i , on diminue le plus fortement l'estimation $\hat{\sigma}$. Pour l'observation n° i , l'indice R Student est défini par :

$$t_i^* = \frac{e_i}{\hat{\sigma}(i) \sqrt{1-h_i}}$$

Une observation est mal reconstituée si $|t_i^*| \geq t_{0,975}(n-3)$.

■ Mesure de l'influence d'une observation

Pour mesurer l'influence de l'observation n° i , on étudie la régression de la variable Y sur la variable X sans cette observation. On obtient une nouvelle droite de régression. On désigne par :

- y_i la valeur observée de la variable Y pour la valeur x_i de la variable X ,
- y_i^* la valeur donnée par la régression avec toutes les observations,
- $y^*(i)$ la valeur donnée par la régression sans l'observation n° i .

On étudie la quantité

$$k_i = \frac{y_i^* - y^*(i)}{\hat{\sigma}(i) \sqrt{h_i}} = t_i^* \sqrt{\frac{h_i}{1-h_i}}$$

Cette quantité est significative si $k_i > 2\sqrt{2/n}$.

Si l'écart entre les valeurs y_i^* et $y^*(i)$ est grand, la droite des moindres carrés a été « attirée » par l'observation n° i .

20.8 Étude d'une valeur prévisionnelle

À l'aide du modèle de régression linéaire, il est possible de prévoir la valeur Y_0 de la variable Y pour une valeur non observée x_0 de la variable X .

La prévision de Y est égale à $y_0^* = a + bx_0$.

Pour encadrer cette valeur, on définit un *intervalle de prévision*. La loi de Y^* est la loi normale :

$$N\left(\alpha + \beta x ; \sigma \sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{n s_X^2}}\right)$$

La loi conditionnelle de $Y/X = x_0$ est la loi normale $N(\alpha + \beta x_0 ; \sigma)$.

Les variables aléatoires Y_0 et Y_0^* sont indépendantes ; en effet, Y_0 ne dépend que de la valeur x_0 et Y_0^* dépend des valeurs déjà observées. Il en résulte que $Y_0 - Y_0^*$ suit la loi normale :

$$N\left(0 ; \sigma \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{n s_X^2}}\right)$$

L'écart-type σ n'étant pas connu mais estimé, la variable aléatoire

$$\frac{Y_0 - Y_0^*}{\hat{\sigma} \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{n s_X^2}}}$$

suit une loi de Student à $(n - 2)$ degrés de liberté. On en déduit l'intervalle de prévision pour la valeur Y_0 , cet intervalle est d'autant plus grand que la valeur x_0 est éloignée de la moyenne $\bar{x}$.

Remarque

La quantité $\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{n s_X^2}$ est le levier de l'observation correspondant à la valeur x_0 de la variable X .

20.8.1 Élimination des valeurs atypiques

Une valeur est considérée comme *atypique* si la valeur observée y_k n'appartient pas à l'intervalle de prévision calculé pour les valeurs x_k et la valeur prévue y_k^* .

Si on peut expliquer cette anomalie par des causes externes, on élimine cette valeur mais on réduit le champ d'application.

Exemple 20.1

On veut étudier l'influence du nombre d'immatriculations de voitures sur le nombre d'accidents pendant une période de 12 ans. Les données et les résultats ont été regroupés dans un même tableau (tableau 20.1).

Tableau 20.1 – Tableau des données et des résultats.

X	$X - \bar{X}$	Y	Y_i^*	e_i	e_i^2	$(x_i - \bar{X})^2$	h_i	t_i
150	-121,25	84	78,2295	5,7705	33,2986	14 701,562	0,2606	0,7527
160	-111,25	75	81,6401	-6,6401	44,0909	12 376,562	0,2326	-0,8502
210	-61,25	90	98,6932	-8,6932	75,5717	3 751,5625	0,1286	-1,0445
215	-56,25	100	100,3985	-0,3985	0,1588	3 164,0625	0,1215	-0,0477
230	-41,25	104	105,5145	-1,5145	2,2937	1 701,5625	0,1038	-0,1794
250	-21,25	112	112,3357	-0,3357	0,1127	451,5625	0,0888	-0,0394
260	-11,25	130	115,7464	14,2536	203,165	126,5625	0,0848	1,6711
300	28,75	140	129,3889	10,6111	112,595	826,5625	0,0933	1,2499
320	48,75	120	136,2102	-16,2102	262,771	2 376,5625	0,112	-1,9294
340	68,75	150	143,0314	6,9686	48,5614	4 726,5625	0,1403	0,8430
400	128,75	160	163,4952	-3,4952	12,2164	16 576,562	0,2833	-0,4630
420	148,75	170	170,3164	-0,3164	0,1001	22 126,562	0,3502	-0,044
500			197,6015					

La plupart des résultats ont été obtenus par le logiciel de statistique Stat View.

La variable expliquée est le nombre d'accidents, $Y \times 1\,000$:

$$\sum_i y_i = 1435 \quad \sum_i y_i^2 = 182\,041 \quad \bar{y} = 119,5833$$

$$s^2 = 869,9097 \quad s = 29,494232$$

La variable explicative est le nombre d'immatriculations, $X \times 1\,000$:

$$\sum_i x_i = 3255 \quad \sum_i x_i^2 = 965\,825 \quad \bar{x} = 271,25$$

$$s^2 = 6908,8542 \quad s = 83,119517$$

$$\text{Cov}(x, y) = 2\,356,351$$

Tableau 20.2 – Analyse de variance de la régression.

Variation	Somme des carrés	Degré de liberté	Quotient
Variation due à la régression	9 643,9812	1	9 643,9812
Variation résiduelle	794,9355	10	79,4935
Variation totale	10 438,9167	11	

Coefficient de corrélation :

$$r = 0,96118$$

$$r^2 = 0,9238$$

$$R^2_{\text{ajusté}} = 0,9162$$

Équation de la droite de régression : $y^* = 0,3411x + 27,07$

Le quotient $9\,643,9812/79,4935 = 121,3178$.

Or $\Pr(F(1; 10) > 121,32) = 0,0001$.

Le modèle linéaire convient donc.

– Variance résiduelle $\hat{\sigma}^2 = 79,4935$ $\hat{\sigma} = 8,916$.

– Étude de la pente $b = 0,341$.

Estimation de l'erreur : $\hat{\sigma}_B = \frac{\hat{\sigma}}{\sqrt{\sum_i (x_i - \bar{x})^2}} = 0,031$

Intervalle de confiance : à 95 % [0,272 ; 0,410] ; à 90 % [0,285 ; 0,397].

– Étude du coefficient $a = 27,07$.

Estimation de l'erreur : $\hat{\sigma}_A = \hat{\sigma} \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum_i (x_i - \bar{x})^2}} = 8,785$

Intervalle de confiance : à 95 % [7,50 ; 46,64] ; à 90 % [11,15 ; 42,99].

Les seuils critiques pour la variable de Student $T(10)$ sont 1,812 (à 90 %) et 2,228 (à 95 %).

– Moyenne des valeurs ajustées 119,58333

Estimation de l'écart-type :

$$\hat{\sigma} \sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_i (x_i - \bar{x})^2}} = 8,916 \times 0,2886 = 2,574$$

Intervalle de confiance : à 95 % [113,8485 ; 125,3181] ;
à 90 % [114,9184 ; 124,2483].

– Étude des résidus : 4 résidus sont positifs et 8 sont négatifs.

$$\sum_i (e_{i-1} - e_i)^2 = 1832,0224$$

– Étude d'une valeur prévisionnelle $x_0 = 500$: $y_0^* = 197,60$

$$\text{Écart-type : } \hat{\sigma} \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_i (x_i - \bar{x})^2}} = 8,916 \times 1,309 = 11,6745$$

Intervalle de prévision à 95 % : $T(10) = 2,228$

D'où l'intervalle : $197,60 \pm 26,01$.

– Recherche des valeurs atypiques :

Une valeur est atypique si la valeur observée n'appartient pas à son intervalle de prévision. Cet intervalle est défini, pour un seuil de confiance égal à 95 %, par :

$$y_i^* \pm t_{0,975}(n-2) \hat{\sigma} \sqrt{1 + \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_i (x_i - \bar{x})^2}} \quad t_{0,975}(10) = 2,228$$

Le tableau 20.3 donne les résultats pour les douze valeurs de la variable X .

Tableau 20.3 – Recherche des valeurs atypiques.

X	Y	Y*	Intervalle de prévision
150	84	78,2295	78,2295 ± 22,30
160	75	81,6401	81,6401 ± 22,055
210	90	98,6932	98,6932 ± 21,103
215	100	100,3985	100,3985 ± 21,037
230	104	105,5145	105,5145 ± 20,871
250	112	112,3357	112,3357 ± 20,728
260	130	115,7464	115,7464 ± 20,69
300	140	129,3889	129,3889 ± 20,771
320	120	136,2102	136,2102 ± 20,947
340	150	143,0314	143,0314 ± 21,213
400	160	163,4952	163,4952 ± 22,503
420	170	170,3164	170,3164 ± 23,083

Il n'y a pas de valeurs atypiques.

20.9 Conclusions

Les principaux problèmes traités en inférence statistique consistent principalement en l'estimation de paramètres ou en la vérification de certaines hypothèses. Ainsi, on a analysé des modèles du type :

$$Y_i = \alpha + \varepsilon_i E(Y_i) = \alpha$$

$$Y_i = \alpha + \beta_i X_i + \varepsilon_i E(Y_i) = \alpha + \beta_i X_i$$

Si une seule variable aléatoire Y est en cause, la meilleure estimation de la moyenne m de la population est la moyenne de l'échantillon.

Si une variable aléatoire Y est liée à une autre variable X , appelée *variable explicative*, la meilleure estimation de la valeur moyenne de Y est donnée par : $Y^* = \alpha + \beta X = \bar{Y} + \beta(X - \bar{X})$ (modèle linéaire simple). La connaissance de la variable explicative améliore donc l'estimation de la valeur de Y par rapport à celle obtenue avec la moyenne de l'échantillon.

21 • RÉGRESSION MULTIPLE

MODÈLE LINÉAIRE GÉNÉRAL

21.1 Introduction

Dans le chapitre précédent, on a étudié le modèle le plus simple de régression, la régression linéaire simple, qui consiste à expliquer une variable quantitative, notée Y , à l'aide d'une autre variable quantitative, notée X . Mais il peut arriver que cette variable X ne suffise pas à expliquer Y , car certaines variables explicatives peuvent avoir été omises, soit volontairement dans un but de simplification, soit à cause d'une mauvaise planification, soit parce que ces variables n'étaient pas mesurées avec une précision suffisante ou parce que leur introduction pouvait entraîner des coûts d'exploitation prohibitifs...

Dans le cas de la régression multiple, on introduit alors un ensemble de p variables explicatives et on cherche à estimer Y sous la forme d'une fonction affine des variables explicatives. L'introduction de ces variables explicatives, aléatoires ou contrôlées par l'expérimentateur, améliore l'estimation de la valeur moyenne de la variable à expliquer Y en réduisant le pourcentage de variation non expliquée dans cette variable et permettant ainsi d'expliquer la variabilité existant dans le comportement de toute variable aléatoire.

Les domaines d'application de la régression multiple sont très nombreux, domaines relevant de la technologie, de la finance, de la gestion, de la médecine, de la biologie, de l'agriculture, etc.

21.2 Régression entre variables aléatoires

21.2.1 Recherche d'un ajustement linéaire

Sur un échantillon de n individus, on a effectué une série de mesures concernant $(p + 1)$ variables représentées par $(p + 1)$ vecteurs de l'espace $\mathbb{R}^n$. On suppose $n > p$. Parmi ces variables, on distingue :

- une variable aléatoire à expliquer $\underline{Y}$ de composantes $(y_1, \dots, y_n)$;
- un ensemble de p variables explicatives $\underline{X}_i$. Ces p variables explicatives sont linéairement indépendantes, mais peuvent éventuellement être corrélées.

On cherche à estimer la variable aléatoire $\underline{Y}$ au moyen des p variables explicatives par une relation de la forme :

$$\underline{Y}^* = b_0 I + \sum_{i=1}^p b_i \underline{X}_i$$

en utilisant la méthode des moindres carrés qui consiste à rendre minimale la norme :

$$\|\underline{Y} - \underline{Y}^*\|^2$$

où $\underline{Y}^*$ est la valeur estimée de $\underline{Y}$, I le vecteur unité de $\mathbb{R}^n$ (toutes ses composantes sont égales à 1).

Le coefficient b_0 et les p coefficients b_i sont les *paramètres de la régression*.

On note :

- X la matrice à n lignes et $(p + 1)$ colonnes (le vecteur I et les p vecteurs $\underline{X}_i$),
- tX la matrice transposée de X ,
- $\underline{b}$ le vecteur colonne dont les composantes sont les paramètres b_0 et b_i :

$$X = \begin{bmatrix} 1 & x_{11} & \dots & x_{1p} \\ 1 & x_{21} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \dots & x_{np} \end{bmatrix} \quad \underline{b} = \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_p \end{bmatrix}$$

La variable $\underline{X}_i$ prend la valeur x_{ki} pour l'individu i .

En notation matricielle, la valeur estimée de $\underline{Y}$ s'écrit : $\underline{Y}^* = X\underline{b}$. En utilisant les résultats de la statistique multidimensionnelle, on en déduit que $\underline{Y}^*$ est

la projection de $\underline{Y}$ sur le sous-espace affine $\mathcal{W}$ de dimension $(p + 1)$ de $\mathbb{R}^n$ engendré par le vecteur I et les p vecteurs $\underline{X}_i$, donc $\underline{Y}^* = P \times \underline{Y}$ où P est la matrice de la projection sur l'espace vectoriel $\mathcal{W}$. Cette matrice P est égale à : $P = X({}^tXX)^{-1}{}^tX$.

Démonstration

La matrice tXX est symétrique et définie positive. Elle est donc inversible.

$P^2 = X({}^tXX)^{-1}{}^tX \times X({}^tXX)^{-1}{}^tX = X({}^tXX){}^tX = P$ donc P est bien la matrice d'un opérateur de projection.

Les vecteurs $\underline{X}_i$ étant linéairement indépendants, $\text{rang}(X) = p + 1$, donc $\text{rang}(P) = p + 1$. La matrice P laisse invariant un sous-espace de dimension $p + 1$.

De plus, $PX = X$, donc $P\underline{X}_i = \underline{X}_i$, l'espace laissé invariant par P est bien l'espace $\mathcal{W}$ défini précédemment.

D'où la valeur de $\underline{Y}^*$:

$$\underline{Y}^* = X({}^tXX)^{-1}{}^tX\underline{Y}$$

Et comme $\underline{Y}^* = X\underline{b}$, on obtient :

$$\underline{b} = ({}^tXX)^{-1}{}^tX\underline{Y}$$

Remarque

Ce calcul suppose que la métrique choisie dans l'espace $\mathbb{R}^n$ est la métrique $D_p = 1/nI_n$, I_n étant la matrice identité de $\mathbb{R}^n$ (toutes les observations ont donc le même poids). S'il n'en est pas ainsi, les formules précédentes deviennent, D_p définissant une métrique quelconque :

$$\underline{Y}^* = X({}^tXD_pX)^{-1}{}^tXD_p\underline{Y} \quad \underline{b} = ({}^tXD_pX)^{-1}{}^tXD_p\underline{Y}$$

21.2.2 Hypothèses de régression linéaire multiple

Pour mettre en œuvre un modèle de régression linéaire multiple, on doit supposer que les variables $\underline{Y}$ et $\underline{X}_i$ constituent un échantillon de n observations indépendantes de $(p + 1)$ variables aléatoires désignées respectivement par ψ et φ_i .

La recherche de la meilleure approximation de la variable ψ par une fonction des variables φ_i est l'espérance conditionnelle $E(\psi/\varphi_1, \dots, \varphi_p)$. L'hypothèse de régression multiple est alors :

$$E(\psi/\varphi_1, \dots, \varphi_p) = \beta_0 + \sum_{i=1}^p \beta_i \varphi_i$$

et conduit au modèle :

$$\psi = \beta_0 + \sum_{i=1}^p \beta_i \varphi_i + \varepsilon$$

Le terme ε est une variable aléatoire d'espérance mathématique nulle et de variance σ^2 , non corrélée aux variables φ_i .

21.2.3 Estimation des paramètres d'une régression linéaire multiple

En admettant l'hypothèse de régression linéaire multiple, on peut écrire les relations suivantes entre les valeurs observées et les paramètres :

$$y_i = \beta_0 + \sum_{j=1}^p \beta_j x_{ij} + e_i \quad \forall i \in (1, n)$$

En utilisant la notation matricielle, ces relations sont équivalentes à :

$$\underline{Y} = X \underline{\beta} + \underline{e}$$

$$\underline{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} \quad \underline{e} = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}$$

■ Estimation du paramètre $\underline{\beta}$

Le vecteur $\underline{b}$ obtenu par la méthode des moindres carrés est la meilleure approximation du paramètre $\underline{\beta}$.

Le vecteur $\underline{b}$ est un estimateur sans biais du paramètre $\underline{\beta}$. En effet, la matrice X étant à coefficients constants et les erreurs étant distribuées suivant une loi normale d'espérance nulle :

$$E(\underline{b}) = ({}^tXX)^{-1} {}^tXE(\underline{Y})$$

$$E(\underline{Y}) = X\underline{\beta} + E(\underline{e}) = X\underline{\beta}$$

D'où : $E(\underline{b}) = \underline{\beta}$.

De tous les estimateurs sans biais de $\underline{\beta}$ de la forme $\underline{B}\underline{Y}$, $\underline{b}$ est celui qui a la variance minimale :

$$\text{Var}(\underline{b}) = \text{Var} \left[({}^tXX)^{-1} {}^tX\underline{Y} \right] = ({}^tXX)^{-1} {}^tX \text{Var}(\underline{Y}) X ({}^tXX)^{-1}$$

$$\text{Var}(\underline{Y}) = \text{Var}(\underline{e}) = \sigma^2 I_n \Rightarrow \text{Var}(\underline{b}) = \sigma^2 ({}^tXX)^{-1}$$

Les estimateurs $\underline{b}$ et $\underline{B}\underline{Y}$ sont des estimateurs sans biais du paramètre $\underline{\beta}$, on a donc :

$$E(\underline{b}) = \underline{\beta} = E(\underline{B}\underline{Y}) = \underline{B}X\underline{\beta}$$

Cette égalité est vraie quelle que soit la valeur de $\underline{\beta}$, on en déduit que $\underline{B}X = I_{p+1}$.

Posons $\underline{B} = ({}^tXX)^{-1} {}^tX + \underline{C}$, la constante $\underline{C}$ est telle que $\underline{C}X = 0$ et ${}^tX\underline{C} = 0$.

La matrice de variance de $\underline{B}\underline{Y}$ est donc :

$$\text{Var}(\underline{B}\underline{Y}) = \underline{B} \text{Var}(\underline{Y}) {}^t\underline{B} = \sigma^2 \underline{B} ({}^tXX)^{-1} {}^t\underline{B}$$

$$\text{Var}(\underline{B}\underline{Y}) = \sigma^2 \left[({}^tXX)^{-1} {}^tX + \underline{C} \right] ({}^tXX)^{-1} \left[X ({}^tXX)^{-1} + \underline{C}^t \right]$$

Après simplification :

$$\text{Var}(\underline{B}\underline{Y}) = \sigma^2 (X^tX)^{-1} + \sigma^2 C^t C = \text{Var}(\underline{b}) + \sigma^2 C^t C$$

Or $\sigma^2 C^t C$ est une matrice symétrique, définie positive. Tous les estimateurs sans biais de la forme $\underline{B}\underline{Y}$ ont donc une variance supérieure ou égale à celle de $\underline{b}$ et $\underline{b}$ est l'estimateur de la forme $\underline{B}\underline{Y}$ de variance minimale.

■ Estimation de la variance σ^2

La meilleure estimation sans biais de la variance σ^2 est la quantité :

$$\hat{\sigma}^2 = \frac{\|\underline{Y} - \underline{Y}^*\|^2}{n - p - 1}$$

Pour démontrer ce résultat, on remarque que l'approximation de $\underline{Y}$ par $\underline{Y}^*$ est la projection orthogonale de $\underline{Y}$ sur le sous-espace vectoriel W . On peut donc écrire l'erreur d'approximation $\underline{e}$ sous la forme :

$$\underline{e} = (\underline{Y} - \underline{X}\underline{b}) + (\underline{X}\underline{b} - \underline{X}\underline{\beta})$$

Le deuxième terme appartient à l'espace W alors que le premier appartient au sous-espace W' orthogonal à W . Si P est la matrice de projection sur l'espace W , $I - P$ est la matrice de projection sur l'espace W' . Il en résulte que $(\underline{Y} - \underline{X}\underline{b}) = (I - P)\underline{e}$. D'où, en désignant par α le terme générique de la matrice $I - A$:

$$\|\underline{Y} - \underline{X}\underline{b}\|^2 = {}^t \underline{e} (I - A) \underline{e} = \sum_{i,j} \alpha_{i,j} e_i e_j$$

De cette relation, on déduit :

$$\begin{aligned} E \left[\|\underline{Y} - \underline{X}\underline{b}\|^2 \right] &= \sum_{i,j} \alpha_{i,j} E(e_i e_j) = \sum_{i,j} \delta_{i,j} \alpha_{i,j} \sigma^2 \\ &= \sigma^2 \text{Trace}(I - A) = \sigma^2 (n - p - 1) \end{aligned}$$

En effet :

- L'espérance du produit $e_i e_j$ est égale à la variance σ^2 si $i = j$ et à 0 sinon, car les erreurs ne sont pas corrélées.
- La trace du projecteur $(I - A)$ est égale à son rang, c'est-à-dire à la dimension du sous-espace sur lequel on projette, or $\text{rang}(A) = (p + 1)$ d'où $\text{rang}(I - A) = n - p - 1$.

D'où le résultat : $E \left[\|\underline{Y} - \underline{X}\underline{b}\|^2 \right] = \sigma^2 (n - p - 1)$

On en déduit que $\hat{\sigma}^2 = \frac{\|\underline{Y} - \underline{X}\underline{b}\|^2}{n - p - 1}$ est un estimateur sans biais de la variance σ^2 .

21.3 Modèle linéaire général

On suppose désormais qu'à l'ensemble des n valeurs des p variables explicatives, on ne fait plus correspondre une seule valeur $\underline{Y}$ mais un ensemble de k valeurs. Ainsi, par exemple, ayant fixé la valeur de la température et de la pression, on mesure plusieurs fois le résultat d'une expérience dans les mêmes conditions expérimentales. Les valeurs des variables explicatives peuvent être fixées expérimentalement ou observées sans erreur ; elles sont donc à caractère non aléatoire.

21.3.1 Énoncé du problème

La forme d'un modèle linéaire général est la suivante :

- $\underline{Y}$ est la variable aléatoire à expliquer, $\underline{Y}_i$ est l'observation n° i avec $i \in \{1, \dots, k\}$,
- $X_1, \dots, X_p$ sont p variables explicatives. L'observation n° i s'écrit donc :

$$X_{i1}, \dots, X_{ip}, \underline{Y}_i$$

- $b_0, b_1, \dots, b_p$ sont les paramètres du modèle,
- ε_i est l'erreur aléatoire représentant l'écart entre les valeurs observées Y_i de la variable à expliquer et les valeurs espérées $E(Y_i)$ de cette variable : $\varepsilon_i = Y_i - E(Y_i)$.

Exemple 21.1

Dans une entreprise, le responsable des ressources humaines souhaite établir un modèle de prévision permettant d'évaluer la quantité moyenne de pièces assemblées par des employés sur une chaîne de montage de pièces complexes.

On a retenu les *variables* suivantes :

(X_1) : résultat d'un test de dextérité manuelle.

(X_2) : conception visuelle.

(X_3) : connaissances techniques.

(X_4) : nombre d'années d'expérience.

Identification des variables :

Y_i : variable aléatoire à expliquer « nombre de pièces assemblées par l'employé i pendant un temps déterminé ».

X_{i1} : variable explicative « résultat du test de dextérité manuelle pour l'employé i ».

X_{i2} : variable explicative « vitesse de perception visuelle pour l'employé i ».

X_{i3} : variable explicative « résultat du test subi par l'employé i portant sur ses connaissances techniques ».

X_{i4} : variable explicative « nombre d'années passées dans un atelier de montage par l'employé i ».

Énoncé du modèle : c'est un modèle à quatre variables explicatives de la forme

$$Y_i = b_0 + b_1 X_{i1} + b_2 X_{i2} + b_3 X_{i3} + b_4 X_{i4} + \varepsilon_i$$

21.3.2 Hypothèses du modèle linéaire multiple

On suppose que le centre de gravité $\underline{g}$ du nuage de points $\underline{Y}_i$ appartient à l'espace W :

$$\underline{g} = X\beta$$

En pratique cependant, on ne connaît souvent qu'une seule observation ; il faut donc approcher au mieux le vecteur $\underline{g}$ à l'aide d'une seule observation $\underline{g}^*$. L'approximation de $\underline{g}$ est obtenue en projetant l'unique observation de $\underline{Y}$ sur l'espace W selon une métrique M :

$$\underline{g}_i^* = X ({}^t X M X)^{-1} {}^t X M \underline{Y}_i$$

Cette métrique doit être telle que, si on a k observations $\underline{Y}_i$ de la variable $\underline{Y}$, les k vecteurs projetés $\underline{g}_i^*$ ne doivent pas être trop dispersés autour de $\underline{g}$. La métrique rendant minimale l'inertie du nuage de points $\underline{g}_i^*$ est la métrique définie par la matrice V^{-1} , où V est la matrice de variance-covariance des vecteurs $\underline{g}_i^*$. D'après la définition de ces vecteurs, cette propriété entraîne que le nuage des vecteurs est le moins dispersé possible dans l'espace $\mathbb{R}^{p+1}$.

Avec une seule observation, on obtient :

$$\underline{g}^* = X ({}^t X V^{-1} X)^{-1} {}^t X V^{-1} \underline{Y} \quad \underline{b} = ({}^t X V^{-1} X)^{-1} {}^t X V^{-1} \underline{Y}$$

On généralise ces résultats dans le cas où on dispose de k observations de la variable aléatoire $\underline{Y}$. $\underline{Y}$ est la réalisation d'un vecteur aléatoire d'espérance $X\beta$ et de matrice de variance Σ . On considère le modèle :

$$\underline{Y} = X\beta + \varepsilon$$

Le résidu $\underline{e}$ est un vecteur aléatoire d'espérance nulle et de matrice de variance Σ .

On note ce modèle $(\underline{Y}, X\underline{\beta}, \tilde{\Sigma})$.

Estimation du vecteur $\underline{\beta}$: on démontre que

$$\underline{b} = \left({}^tX \Sigma^{-1} X \right)^{-1} {}^tX \Sigma^{-1} \underline{Y}$$

est l'estimation du vecteur $\underline{\beta}$ de variance minimale.

21.3.3 Comparaison des deux modèles

Dans l'hypothèse de la régression linéaire et du modèle linéaire général, on a posé le même modèle : $\underline{Y} = X\underline{\beta} + \underline{e}$. Cependant, les conditions d'application de ces deux modèles sont différentes. En effet :

- En régression, X est un élément aléatoire ; dans le modèle linéaire, X est un tableau de données certaines.
- En régression, le résidu $\underline{e}$ a pour matrice de variance-covariance la matrice $\sigma^2 I_n$, car les données sont supposées être indépendantes ; en revanche, dans le modèle linéaire général, cette matrice est une matrice Σ quelconque.
- Les objectifs sont différents. En effet, en régression, on cherche à estimer au mieux $\underline{Y}$. Dans le modèle linéaire général, on cherche à estimer l'effet moyen des variables explicatives.

Cependant, si dans le modèle de régression on supposait les variables explicatives constantes ou bien si on travaillait conditionnellement aux variables φ_i , cela reviendrait à poser le modèle linéaire suivant $(\underline{Y}, X\underline{\beta}, \sigma^2 I_n)$, en donnant le même poids à toutes les observations. Comme dans la plupart des cas, on travaille conditionnellement aux variables φ_i , on peut utiliser le même modèle $(\underline{Y}, X\underline{\beta}, \sigma^2 I_n)$.

Enfin, l'utilisation du modèle linéaire général suppose connue la matrice Σ , ce qui n'est pas vrai en pratique. On est donc conduit à faire les hypothèses simplificatrices suivantes :

Absence de corrélation des erreurs :

$$\text{Cov}(\varepsilon_i, \varepsilon_j) = E(\varepsilon_i, \varepsilon_j) = 0 \quad i \neq j$$

La matrice Σ est donc une matrice diagonale.

Homocédasticité : on suppose que la variance des variables ε_i est constante quelles que soient les valeurs des variables explicatives $X_{i1}, \dots, X_{ip}$. La matrice Σ est donc égale à la matrice $\sigma^2 I_n$.

Ces deux hypothèses devraient être vérifiées *a posteriori*.

On peut être amené dans certains cas à supposer que les erreurs sont distribuées suivant une loi normale $N(0, \sigma)$.

Cependant, il faut dans tous les cas étudier la répartition des erreurs avant d'accepter un modèle de régression linéaire. En effet, il faut vérifier que l'espérance est proche de 0 en veillant à ce qu'il y ait autant de termes de résidus positifs que de résidus négatifs. On vérifiera aussi que les termes d'erreurs sont dans une bande centrée en 0.

En conclusion, on étudie donc le modèle simplifié ($\underline{Y}, X \underline{\beta}, \sigma^2 I_n$). Les problèmes à résoudre sont les suivants :

- Estimation des paramètres du modèle.
- Tests d'hypothèses.
- Corrélations multiples et partielles.
- Intervalle de prévision.
- Choix du meilleur ensemble de variables explicatives.

D

ANALYSE DES DONNÉES

21.4 Estimations des paramètres du modèle de régression ($\underline{Y}, X \underline{\beta}, \sigma^2 I_n$)

Le modèle théorique de régression est le modèle $\underline{Y} = X \underline{\beta} + \underline{\varepsilon}$. On pose : $\underline{Y}^* = X \underline{b}$. Le paramètre $\underline{b}$, obtenu par la méthode des moindres carrés, est donné par :

$$\underline{b} = \left({}^t X V^{-1} X \right)^{-1} {}^t X V^{-1} \underline{Y} = \left({}^t X X \right)^{-1} {}^t X \underline{Y}$$

car la matrice V est égale à la matrice $\sigma^2 I_n$.

21.4.1 Estimation du paramètre $\underline{\beta}$

■ Espérance de $\underline{b}$

X étant une matrice à coefficients constants, on obtient :

$$E(\underline{b}) = ({}^t X X)^{-1} {}^t X E(\underline{Y})$$

L'hypothèse du modèle linéaire général, $\underline{Y} = X\underline{\beta} + \underline{\varepsilon}$, entraîne $E(\underline{Y}) = X\underline{\beta}$.
D'où :

$$E(\underline{b}) = \underline{\beta}$$

$\underline{b}$ est donc un estimateur sans biais de $\underline{\beta}$.

■ Variance de $\underline{b}$

On démontre le résultat suivant (théorème de Gauss-Markov) :

$\underline{b}$ est de tous les estimateurs sans biais de $\underline{\beta}$ de la forme $B\underline{Y}$ celui qui a la variance minimale.

La démonstration consiste à calculer la variance de $\underline{b}$, puis à la comparer à celle d'un autre estimateur sans biais de la forme $B\underline{Y}$. La variance de $\underline{b}$ est égale à :

$$\text{Var}(\underline{b}) = \sigma^2 ({}^t X X)^{-1}$$

21.4.2 Estimation de la variance σ^2

Un estimateur sans biais de la variance σ^2 est donné par la statistique :

$$\hat{\sigma}^2 = \frac{\|\underline{Y} - \underline{Y}^*\|^2}{n-p-1} = \frac{\|\underline{Y} - X\underline{b}\|^2}{n-p-1}$$

On suppose que le résidu $\underline{\varepsilon}$ suit une loi normale. Dans ces conditions, la loi de la variable aléatoire $\underline{\varepsilon}_i$ est la loi normale $N(0; \sigma)$ quelle que soit la valeur de l'indice i . Il en résulte que le vecteur aléatoire $\underline{Y}$ suit une loi normale multidimensionnelle $N(X\underline{\beta}; \sigma^2 I_n)$ de densité :

$$\frac{1}{(\sigma\sqrt{2\pi})^n} \exp \left[-\frac{1}{2\sigma^2} {}^t (\underline{Y} - X\underline{\beta}) (\underline{Y} - X\underline{\beta}) \right]$$

D'où les estimateurs du maximum de vraisemblance :

- du paramètre β : $\hat{\underline{\beta}} = (\text{}^t X X)^{-1} \text{}^t X \underline{Y}$
- de la variance σ^2 : $\hat{\sigma}^2 = \frac{1}{n} \|\underline{Y} - X \hat{\underline{\beta}}\|^2$

Le premier estimateur est sans biais et le deuxième est biaisé. Le vecteur $\underline{b}$ suit une loi normale de dimension $(p + 1)$, d'espérance $\underline{\beta}$ et de variance $(\text{}^t X X)^{-1} \sigma^2$. L'étude de la décomposition (résultat facile à démontrer géométriquement) :

$$\|\underline{e}\|^2 = \|\underline{Y} - X \underline{\beta}\|^2 + \|\underline{X \underline{b}} - X \underline{\beta}\|^2$$

conduit aux résultats suivants :

$$\frac{\|\underline{e}\|^2}{\sigma^2} = \chi^2(n) \quad \frac{\|\underline{Y} - X \underline{b}\|^2}{\sigma^2} = \chi^2(n - p - 1) \quad \frac{\|\underline{X \underline{b}} - X \underline{\beta}\|^2}{\sigma^2} = \chi^2(p + 1)$$

Ces propriétés permettent de construire des intervalles de confiance pour les différents paramètres.

21.5 Estimation du paramètre β du modèle linéaire

En utilisant les méthodes mises en œuvre au paragraphe 21.2.3, on obtient les résultats suivants :

- parmi les estimateurs sans biais de β , fonction linéaire de $\underline{Y}$, l'estimateur de variance minimale est :

$$\underline{b} = (\text{}^t X \Sigma^{-1} X)^{-1} \text{}^t X \Sigma^{-1} \underline{Y}$$

- si de plus, l'hypothèse de normalité est vérifiée, $\underline{b}$ est l'estimateur du maximum de vraisemblance et de variance minimale.

21.6 Tests dans le modèle linéaire

Différents tests permettent :

- soit de tester tous les coefficients par rapport à des valeurs spécifiées,
- soit de tester la nullité d'un ou plusieurs coefficients,

- soit de tester simultanément la nullité de q coefficients, car, les coefficients b_i étant corrélés, il n'est pas possible de tester successivement la nullité de q coefficients.

Les résultats obtenus à la fin du paragraphe 21.4.2 permettent de construire ces tests.

21.6.1 Test simultané de tous les coefficients de régression

Les hypothèses à tester sont les suivantes :

$$H_0 : \beta = \beta_0 \quad H_1 : \beta \neq \beta_0$$

On utilise la propriété :

$$\frac{\|X\underline{b} - X\underline{\beta}\|^2}{\|\underline{Y} - X\underline{b}\|^2} \times \frac{n-p-1}{p+1} = F(p+1; n-p-1)$$

Le test consiste donc en la mesure de la quantité, où on a remplacé β par β_0 :

$$\frac{\|X\underline{b} - X\underline{\beta}_0\|^2}{\|\underline{Y} - X\underline{b}\|^2} \times \frac{n-p-1}{p+1} = K(\beta_0)$$

Si cette quantité est supérieure à une valeur qu'une variable de Fisher $F(p+1; n-p-1)$ n'a qu'une probabilité α (donnée) de dépasser, c'est-à-dire si $K(\beta_0) > F_\alpha(p+1; n-p-1)$, alors on rejette l'hypothèse H_0 et les coefficients calculés ont une probabilité inférieure à α d'être satisfaisants. Sinon, les coefficients ont une probabilité α d'être satisfaisants.

21.6.2 Test du caractère significatif d'un coefficient de régression

Les hypothèses à tester sont les suivantes :

$$H_0 : \beta_i = 0 \quad H_1 : \beta_i \neq 0$$

On sait que $E(b_i) = \beta_i$ et que la matrice de variance de $\underline{b}$ est la matrice $\sigma^2 (^tXX)^{-1}$ (paragraphe 21.4.1). La variance du coefficient empirique b_i est le terme diagonal (i, i) de cette matrice, c'est-à-dire le terme $\sigma^2 \left[(^tXX)^{-1} \right]_{i,i}$.

De plus :

$$\frac{\|\underline{Y} - X\underline{b}\|^2}{\sigma^2} = \sum_i \left(\frac{y_i - y_i^*}{\sigma} \right)^2 = \chi^2 (n - p - 1)$$

On en déduit que la variable aléatoire

$$\frac{b_i - \beta_i}{\sqrt{\sum_i (y_i - y_i^*)^2 [({}^t X X)^{-1}]_{ii}}} \times \sqrt{n - p - 1}$$

suit une loi de Student à $(n - p - 1)$ degrés de liberté. On accepte l'hypothèse $\beta_i = 0$ si la quantité ci-dessus est inférieure à une valeur qu'une variable de Student à $(n - p - 1)$ degrés de liberté n'a qu'une probabilité α de dépasser.

21.6.3 Test simultané de q coefficients de régression

Tester les hypothèses simultanées

$$\beta_1 = \beta_{10} \quad \beta_2 = \beta_{20} \quad \beta_\theta = \beta_{\theta 0}$$

revient à tester l'hypothèse $M\underline{\beta} = \underline{\theta}$ où M est une matrice de rang q . Les hypothèses à tester sont donc les suivantes :

$$H_0 : M\underline{\beta} = \underline{\theta} \quad H_1 : M\underline{\beta} \neq \underline{\theta}$$

Comme précédemment, on considère la solution $\underline{Y}^*$ des moindres carrés :

$$\underline{Y}^* = X ({}^t X X)^{-1} {}^t X \underline{Y}$$

puis la solution $\underline{Y}_0^*$ des moindres carrés sous la contrainte $M\underline{\beta} = \underline{\theta}$.

Si l'hypothèse H_0 est vraie, la variable aléatoire

$$\frac{\|\underline{Y} - \underline{Y}_0^*\|^2 - \|\underline{Y} - \underline{Y}^*\|^2}{\|\underline{Y} - \underline{Y}^*\|^2} \times \frac{n - p - 1}{q}$$

est une variable aléatoire de Fisher $F(q; n - p - 1)$. La mise en œuvre du test est analogue aux cas précédents.

21.7 Intervalle de prévision

On donne aux p variables explicatives des valeurs $x_{i,0}$ que l'on écrit sous la forme du vecteur $\underline{X}_0$:

$$\underline{X}_0 = \begin{bmatrix} 1 \\ x_{1,0} \\ \vdots \\ x_{p,0} \end{bmatrix}$$

La valeur prévue pour un individu supplémentaire est donnée par $\underline{Y}_0^* = {}^t \underline{X}_0 \underline{b}$.

On cherche à encadrer la valeur obtenue $\underline{Y}_0^*$. Cette valeur est la réalisation d'une variable aléatoire suivant la loi normale :

$$N \left({}^t \underline{X}_0 \underline{\beta}_0 ; \sigma \sqrt{{}^t \underline{X}_0 ({}^t X X)^{-1} \underline{X}_0} \right).$$

Comme l'écart-type σ n'est pas connu mais seulement estimé, il faut utiliser la loi de Student pour calculer un intervalle de confiance ; la variable aléatoire

$$\frac{\underline{Y}_0 - \underline{Y}_0^*}{\hat{\sigma} \sqrt{1 + {}^t \underline{X}_0 ({}^t X X)^{-1} \underline{X}_0}} \times \sqrt{n - p - 1}$$

suit une loi de Student à $(n - p - 1)$ degrés de liberté.

21.8 Corrélations

21.8.1 Corrélation multiple

Pour un modèle de régression multiple, écrivons la décomposition classique :

$$\sum_i \left(\underline{Y}_i - \overline{\underline{Y}} \right)^2 = \sum_i \left(\underline{Y}_i - \underline{Y}_i^* \right)^2 + \sum_i \left(\underline{Y}_i^* - \overline{\underline{Y}} \right)^2$$

Cette égalité exprime que la variation totale $\sum_i (\underline{Y}_i - \overline{Y})^2$ est égale à la somme de la variation résiduelle $\sum_i (\underline{Y}_i - \underline{Y}_i^*)^2$ et de la variation expliquée par la régression $\sum_i (\underline{Y}_i^* - \overline{Y})^2$.

Pour mesurer la qualité de la régression, on définit le *coefficient de détermination* ou *coefficient de corrélation multiple* R^2 qui permet de mesurer, dans la variation de $\underline{Y}$, la proportion expliquée par la régression. Ce coefficient qui est compris entre 0 et 1 est défini par la formule :

$$R^2 = \frac{\sum_i (\underline{Y}_i^* - \overline{Y})^2}{\sum_i (\underline{Y}_i - \overline{Y})^2}$$

Supposons vraie l'hypothèse $H_0 : \beta_i = 0 \forall i \in (1, p)$ et β_0 quelconque (hypothèse de non-régression). On montre alors les résultats suivants :

La variable aléatoire $\frac{\sum_i (\underline{Y}_i^* - \overline{Y})^2}{\sigma^2}$ suit la loi $\chi^2(p)$.

La variable aléatoire $\frac{\sum_i (\underline{Y}_i - \overline{Y})^2}{\sigma^2}$ suit la loi $\chi^2(n-1)$.

On en déduit que sous l'hypothèse H_0 :

La variable aléatoire $\frac{R^2}{1-R^2} \times \frac{n-p-1}{p}$ suit la loi $F(p; n-p-1)$.

On retrouve le test classique. L'hypothèse de non-régression correspond à la nullité du coefficient de corrélation multiple théorique entre variables aléatoires. Sous cette hypothèse, la loi du coefficient R est une *loi bêta de type 1* et on obtient :

$$E(R^2) = \frac{p}{2} \quad \text{Var}(R^2) = \frac{2p(n-p-1)}{(n^2-1)(n-1)}$$

Remarques

Si $p = 1$, on retrouve le test du coefficient de corrélation linéaire simple.

Si $p = q$, le test est le même que le test de nullité de q coefficients.

Si l'hypothèse de non-régression n'est pas satisfaite, la loi de R^2 n'a pas une forme simple. Il est conseillé d'utiliser le coefficient de détermination ajusté :

$$\widehat{R}^2 = \frac{(n-1)R^2 - p}{n-p-1}$$

On en déduit :

$$\widehat{\sigma}^2 = \frac{n}{n-1} (1 - \widehat{R}^2) s_Y^2$$

s_Y^2 est la variation totale de $\underline{Y}$.

21.8.2 Coefficients de corrélation partielle

Les coefficients de corrélation partielle ne peuvent être calculés que si les variables explicatives sont aléatoires (cas de la régression linéaire). Ces coefficients permettent de tester l'influence de chaque variable ou d'une combinaison d'un certain nombre de variables explicatives.

On donne la forme de ces coefficients en supposant que la variable à expliquer et les p variables explicatives suivent une loi normale de dimension $(p+1)$.

Dans le cas où $p = 2$, on obtient :

$$\rho_{y, x_1, x_2} = \frac{\rho_{y, x_1} - \rho_{y, x_2} \rho_{x_1, x_2}}{\sqrt{(1 - \rho_{y, x_2}^2)(1 - \rho_{x_1, x_2}^2)}}$$

On calcule les autres coefficients de proche en proche.

Soient ε_p le résidu obtenu dans l'ajustement de $\underline{Y}$ par les p variables explicatives $x_1, \dots, x_p$, et ε_{p-1} le résidu que l'on aurait obtenu si on avait calculé l'ajustement avec $(p-1)$ variables explicatives $x_1, \dots, x_{p-1}$. Par définition :

$$1 - R^2 = \frac{\text{Variation résiduelle}}{\text{Variation totale}} = \frac{\text{Var}(\varepsilon_p)}{\text{Var}(\underline{Y})}$$

Or, $\text{Var}(\varepsilon_{p-1}) > \text{Var}(\varepsilon_p)$, donc :

$$r_{y, x_p, x_1, \dots, x_{p-1}}^2 = \frac{\text{Var}(\varepsilon_{p-1}) - \text{Var}(\varepsilon_p)}{\text{Var}(\varepsilon_{p-1})} \Rightarrow 1 - r_{y, x_p, x_1, \dots, x_{p-1}}^2 = \frac{\text{Var}(\varepsilon_p)}{\text{Var}(\varepsilon_{p-1})}$$

D'où la relation :

$$1 - R_{y, x_1, x_2, \dots, x_p}^2 = (1 - r_{y, x_1}^2) (1 - r_{y, x_2, x_1}^2) \dots (1 - r_{y, x_p, x_1, x_2, \dots, x_{p-1}}^2)$$

21.8.3 Tests de liaison partielle

Si toutes les variables suivent des lois normales, la loi suivie par un coefficient de corrélation partielle est la même que celle d'un coefficient de corrélation simple mais avec un degré de liberté égal à $(n - d - 2)$ où d est le nombre de variables fixées.

$\frac{r}{\sqrt{1 - r^2}} \sqrt{n - d - 2}$ suit une loi de Student à $(n - d - 2)$ degrés de liberté.

21.9 Fiabilité de la régression

La régression linéaire multiple a pour but d'expliquer une variable à partir d'un ensemble de variables explicatives, l'équation de la régression doit donc retenir le plus grand nombre de variables explicatives s'avérant significatives. Cette équation de la régression doit avoir :

- le meilleur coefficient de détermination possible, R^2 doit être le plus grand possible ;
- la meilleure précision, l'erreur résiduelle doit être aussi petite que possible.

Ces exigences interviennent dans l'acceptation ou le rejet d'une équation de régression. Différentes méthodes peuvent être utilisées :

- soit *étudier toutes les régressions possibles* en définissant certains critères pour sélectionner le meilleur ensemble de variables explicatives ;
- soit introduire progressivement les variables explicatives : *Forward selection method* ;
- soit utiliser une méthode de régression pas à pas : *Stepwise regression method*.

Cependant, toutes ces méthodes ne sont pas équivalentes et ne conduisent pas forcément aux mêmes équations de régression ; le jugement personnel du statisticien intervient dans l'application de ces méthodes de sélection et peut se révéler important dans le choix d'une méthode.

21.9.1 Étudier toutes les régressions possibles

Si le nombre p de variables explicatives n'est pas trop élevé, on peut envisager d'étudier *toutes les régressions possibles*, d'abord avec une seule variable explicative, puis avec toutes les combinaisons 2 à 2, puis 3 à 3... Au total, $(2^p - 1)$ équations de régression sont à étudier.

Pour chaque régression, on teste les coefficients à l'aide du test de Fisher pour mettre en évidence les variables ou les combinaisons de variables qui sont significatives puis on conserve la ou les variables sélectionnées. Si l'adjonction d'une variable supplémentaire n'apporte qu'une contribution marginale très faible, il n'est peut-être pas utile de garder cette variable.

On peut déterminer le meilleur ensemble de variables significatives à l'aide du coefficient de détermination R^2 , ce coefficient étant maximal pour la combinaison qui comprend toutes les variables. Il arrive cependant que l'adjonction de variables supplémentaires ne contribue pas à accroître de façon sensible ce coefficient et il n'est donc pas utile à un certain stade de continuer à ajouter des variables explicatives.

Remarques

Le coefficient de détermination R^2 ne tient pas compte des degrés de liberté. On utilise de préférence le coefficient « ajusté » (voir les remarques en fin du paragraphe 21.8.1).

On peut déterminer le meilleur ensemble de variables explicatives à l'aide du carré moyen résiduel qui tient compte des degrés de liberté ; le nombre de degrés de liberté diminue quand on introduit une variable supplémentaire.

On cherche donc une combinaison des variables explicatives qui minimise le carré moyen ou plutôt telle que l'adjonction d'une variable supplémentaire modifie très peu ce carré résiduel.

21.9.2 Introduire les variables explicatives progressivement

Cette méthode de régression se fait en plusieurs étapes en respectant certains critères :

- On choisit comme *première variable* à introduire, celle qui est le plus fortement corrélée (en valeur absolue) avec la variable à expliquer $\underline{Y}$. On teste les coefficients de l'équation de régression (test de Fisher) et on calcule ensuite les coefficients de corrélation partielle en tenant compte de la variable déjà introduite.
- On introduit une *deuxième variable*, on choisit la variable qui est le plus fortement corrélée avec la variable déjà introduite et on teste la contribution marginale apportée par cette nouvelle variable :
 - si cette contribution n'est pas significative, on s'arrête au stade précédent ;

- si cette contribution est significative, on garde la variable ainsi introduite.
- On continue la sélection en calculant les coefficients de corrélation partielle avec les variables déjà introduites et la contribution marginale de la nouvelle variable introduite. On s'arrête quand la contribution d'une variable introduite n'est pas significative.

Remarques

Par rapport à la méthode « *toutes les régressions possibles* », cette méthode présente l'avantage d'imposer moins de calculs. En revanche, elle ne permet pas de réexaminer la contribution marginale d'une variable introduite à une étape précédente.

Une méthode semblable à cette méthode consiste en l'élimination progressive des variables explicatives (*Backward elimination procedure*). On commence par introduire toutes les variables explicatives et on calcule la contribution marginale de chaque variable. On élimine la variable dont la contribution n'est pas significative. On calcule une nouvelle équation de régression avec $(p - 1)$ variables explicatives et on recommence la même procédure. On s'arrête quand aucune variable ne peut plus être retranchée.

21.9.3 Méthode de régression pas à pas

On commence par effectuer toutes les régressions simples. On choisit de garder une variable explicative selon un critère bien défini, soit la valeur du rapport F de Fisher, soit la valeur du coefficient de détermination.

On sélectionne une deuxième variable en prenant celle dont la contribution marginale est la plus importante.

On continue jusqu'à ce que le processus ait fini la sélection des variables.

Cette méthode de régression pas à pas permet de calculer une série d'équations de régression où à chaque pas une variable est ajoutée ou retranchée selon un critère défini. De plus, elle permet de tester la contribution des variables déjà introduites et de voir si elle est significative compte tenu des nouvelles variables introduites. Cependant, si le nombre p de variables explicatives est grand, elle conduit à de nombreux calculs. Pour remédier à ce trop grand nombre de calculs, on peut :

- soit éliminer la variable qui entraîne la plus faible diminution du coefficient R^2 ; on fait une régression avec $(p - 1)$ variables et on recommence ;

- soit chercher la variable la plus fortement corrélée avec $\underline{Y}$ puis introduire successivement les variables qui provoquent le plus fort accroissement du coefficient R^2 .

21.9.4 Comparaison des différentes méthodes

Ces méthodes ne conduisent en général pas aux mêmes résultats. Il peut arriver que la première variable introduite dans une méthode ascendante soit la première variable éliminée dans une méthode descendante.

Exemple 21.2

On veut étudier la relation qui existe entre le prix d'une voiture et les six critères suivants : cylindrée (en cm^3), puissance maximale (en CV), longueur (en mm), poids (en kg), vitesse maximale (en km/h), consommation (en l/100 km). Les données ont été recueillies dans l'*AutoJournal*.

Le tableau 21.1 donne ces caractéristiques pour 9 voitures berlines essence et le tableau 21.2 pour 13 voitures berlines diesel. Il est évident que tout le parc automobile ne figure pas dans ces tableaux et on ne tirera aucune conclusion d'ordre général ou économique de cette étude. Ces données ont été soumises aux programmes *Forward selection method*, *Backward elimination procedure* et *Stepwise regression method* du logiciel SAS (Statistical Analyse System).

Les résultats obtenus sont résumés dans les différents tableaux suivants.

Tableau 21.1 – Caractéristiques des voitures essence.

Marques/Types	Cylindrée	Puissance	Poids	Longueur	Vitesse	Consommation	Prix
AlfaRomeo 159	3 195	260	1 815	4 660	240	8,5	41 800
BMW Série1	1 596	115	1 280	4 227	200	5,9	21 950
Citroën C2	1 360	75	991	3 666	169	4,9	14 650
Citroën C6	2 946	215	1 816	4 908	230	8,2	44 800
Ford Mondeo	1 999	145	1 419	4 731	190	7,2	25 200
Mercedes S350	3 498	272	1 880	5 079	250	7,7	82 900
Renault Laguna	1 998	135	1 280	4 598	207	6,1	26 400
Volkswagen	4 172	335	2 249	5 175	250	9,8	95 880
Volvo S40	2 435	170	1 358	4 468	220	6,6	27 500

Tableau 21.2 – Caractéristiques des voitures diesel.

Marques/Types	Cylindrée	Puissance	Poids	Longueur	Vitesse	Consommation	Prix
BMW Série 7	4 423	300	2 040	5 039	250	7,2	95 550
Citroën C2	1 398	70	996	3 666	166	3,7	14 600
Fiat Idea	1 910	100	1 275	3 930	179	4,7	17 900
Jaguar X-Type	2 198	155	1 502	4 672	220	4,7	35 800
Mercedes E	2 148	150	1 610	4 818	215	5,3	44 100
Opel Vectra	2 958	184	1 575	4 611	230	5,2	31 200
Peugeot 207	1 398	70	1 251	4 030	166	3,8	13 400
Peugeot 407	2 179	170	1 505	4 676	222	5,1	30 700
Peugeot 607	2 179	170	1 723	4 818	222	5,3	44 650
Renault Clio	1 461	85	1 165	3 986	174	4	15 950
Renault VelSatis	2 958	180	2 320	4 860	210	6,8	44 250
Rover 45	1 994	115	1 230	4 390	190	4,3	19 945
Toyota Aygo	1 398	54	880	3 430	154	3,4	11 900

Exemple 21.2.1

On étudie le prix des 9 voitures essence en fonction des six critères retenus.

Tableau 21.3 – Caractéristiques statistiques des 6 critères.

	Cylindrée	Puissance	Poids	Longueur	Vitesse	Consommation	Prix
Moyenne	2 577,6667	191,3333	1 565,3333	4 612,4444	217,3333	7,2111	42 342,222
Dev. Std	938,3021	84,8219	395,0747	461,4123	28,0758	1,5103	28 435,377
Minimum	1 360	75	991	3 666	169	4,9	14 650
Maximum	4 172	335	2 249	5 175	250	9,8	95 880

Dev. Std ou *standard deviation* est la racine carrée de l'estimation sans biais de la variance, c'est-à-dire de la quantité

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

Tableau 21.4 – Matrice de corrélation.

	Cylindrée	Puissance	Poids	Longueur	Vitesse	Consommation	Prix
Cylindrée	1,0000	0,9950	0,9729	0,8534	0,9400	0,9329	0,9289
Puissance	0,9950	1,0000	0,9793	0,8516	0,9407	0,9449	0,9210
Poids	0,9729	0,9793	1,0000	0,8830	0,9077	0,9721	0,9091
Longueur	0,8534	0,8516	0,8830	1,0000	0,8432	0,8670	0,8017
Vitesse	0,9400	0,9407	0,9077	0,8432	1,0000	0,8456	0,8421
Consommation	0,9329	0,9449	0,9721	0,8670	0,8456	1,0000	0,8088
Prix	0,9289	0,9210	0,9091	0,8017	0,8421	0,8088	1,0000

Tableau 21.5 – Coefficients de corrélation et coefficients de détermination entre le prix et les différentes variables.

Variable Y	Variable X	Coefficient ρ	Coefficient R^2
Prix	Cylindre	0,9289	0,8628
Prix	Puissance	0,9210	0,8482
Prix	Poids	0,9091	0,8264
Prix	Longueur	0,8017	0,6467
Prix	Vitesse	0,8421	0,7091
Prix	Consommation	0,8088	0,6541

Tableau 21.6 – Analyse de la variance.

Source	DDL	Somme des carrés	Carré moyen	Valeur F	Pr > F
Régression	6	6 418 265 115	1 069 710 853	42,53	0,0231
Résiduelle	2	50 300 241	25 150 120		
Totale	8	6 468 565 356			

Conclusions :

 $F_{0,95}(6; 2) = 19,353$, on rejette l'hypothèse $\beta_i = 0$

$$R^2 = \frac{6\,418\,265\,115}{6\,468\,565\,356} = 0,9922 \quad \widehat{R}^2 = 0,9689$$

$$(\widehat{\sigma})^2 = 25\,150\,120 \quad \widehat{\sigma} = 5\,014,98956$$

Tableau 21.7 – Estimation des paramètres.

Variable	DDL	Coefficient estimé	Écart-type	Valeur de t	Pr > t
Constante	1	88 540	42 579	2,08	0,1731
Cylindrée	1	12,26191	19,84536	0,62	0,5996
Puissance	1	378,03692	271,11424	1,39	0,2979
Poids	1	109,04896	33,01624	3,30	0,0807
Longueur	1	17,82579	9,34736	1,91	0,1968
Vitesse	1	-797,7852	225,65462	-3,54	0,0715
Consommation	1	-31 849	5 753,59359	-5,54	0,0311

Prix = 88 540 + 12,26191 Cylindrée + 378,03692 Puissance + 109,04896 Poids
+ 17,82579 Longueur - 797,7852 Vitesse - 31 849 Consommation

Tableau 21.8 – Comparaison des prix ajustés et des prix réels et calcul des résidus.

Marques/Types	Prix ajustés	Prix	Résidus
AlfaRomeo 159	44 813	41 800	-3 013
BMW Série 1	19 050	21 950	2 900
Citroën C2	16 100	14 650	-1 450
Citroën C6	46 811	44 800	-2 011
Ford Mondeo	26 049	25 200	-849
Mercedes S350	85 124	82 900	-2 224
Renault Laguna	26 199	26 400	201
Volkswagen	92 272	95 880	3 608
Volvo S40	24 682	27 500	2 818
	42 345	42 342	-2,4

La dernière ligne du tableau donne les moyennes des prix réels, des prix ajustés et des résidus. On remarque que la moyenne des résidus est très proche de 0. Cependant, les résidus sont dans l'ensemble assez élevés.

Forward selection method

Cette méthode consiste à introduire successivement les variables qui font augmenter le plus la valeur du coefficient de détermination ou qui font diminuer le plus la somme des carrés résiduelles.

À chaque pas, le premier tableau est une analyse de la variance, le deuxième tableau donne les coefficients de la régression, puis on écrit l'équation de la régression.

Pas 1 : variable « cylindrée » sélectionnée $R^2 = 0,8628$

Source	DDL	Somme des carrés	Carré moyen	Valeur F	Pr > F
Régression	1	5 581 390 368	5 581 390 368	44,04	0,0003
Résiduelle	7	887 174 988	126 739 284		
Totale	8	6 468 565 356			

Variable	Coefficient estimé	Écart-type	Type II SS	Valeur F	Pr > F
Constante	-30 220	11 560	866 068 471	6,83	0,0347
Cylindrée	28,15033	4,24197	5 581 390 368	44,04	0,0003

« Type II SS » représente l'accroissement que subirait la somme des carrés résiduels si on éliminait la variable en question. F est le quotient usuel d'analyse de la variance de la régression ; il permet de tester le caractère globalement significatif des variables explicatives. $\text{Pr} > F$ signifie : $\text{Pr}(F_{p;n-p-1} > F)$. Cette variable est le carré d'une variable de Student.

$$\text{Prix} = -30\,220 + 28,15033 \text{Cylindrée}$$

Pas 2 : variable « consommation » sélectionnée $R^2 = 0,8886$

Source	DDL	Somme des carrés	Carré moyen	Valeur F	Pr > F
Régression	2	5 747 895 915	287 394 957	23,93	0,0014
Résiduelle	7	720 669 441	120 111 573		
Totale	8	6 468 565 356			

Variable	Coefficient estimé	Écart-type	Type II SS	Valeur F	Pr > F
Constante	-2 204,712	26 322	842 689	0,01	0,9360
Cylindrée	40,74396	11,46568	1 516 743 166	12,63	0,0120
Consommation	-8 386,698	7 123,10077	166 505 547	1,39	0,2836

$$\text{Prix} = -2\,204,712 + 40,74396 \text{Cylindrée} - 8\,386,698 \text{Consommation}$$

Pas 3 : variable « poids » sélectionnée $R^2 = 0,9429$

Source	DDL	Somme des carrés	Carré moyen	Valeur F	Pr > F
Régression	3	6 099 010 816	2 033 003 605	27,51	0,0016
Résiduelle	5	369 554 539	73 910 908		
Totale	8	6 468 565 356			

Variable	Coefficient estimé	Écart-type	Type II SS	Valeur F	Pr > F
Constante	-10 793	21 020	19 485 326	0,26	0,6295
Cylindrée	16,20276	14,41094	93 433 316	1,26	0,3119
Poids	114,49491	52,53102	351 114 902	4,75	0,0812
Consommation	-23 277	8 825,81357	514 105 343	6,96	0,0461

$$\text{Prix} = -10\,793 + 16,20276 \text{ Cylindrée} + 114,49491 \text{ Poids} \\ - 23\,277 \text{ Consommation}$$

Pas 4 : variable « vitesse » sélectionnée $R^2 = 9\,758$

Source	DDL	Somme des carrés	Carré moyen	Valeur F	Pr > F
Régression	4	6 312 269 362	1 578 067 341	40,39	0,0017
Résiduelle	4	156 295 993	39 073 998		
Totale	8	6 468 565 356			

Variable	Coefficient estimé	Écart-type	Type II SS	Valeur F	Pr > F
Constante	78 524	41 174	142 120 559	3,64	0,1292
Cylindrée	32,42417	12,56991	259 992 638	6,65	0,0614
Poids	130,95663	38,83944	444 218 664	11,37	0,0280
Vitesse	-566,6694	242,56073	213 258 546	5,46	0,0797
Consommation	-27 956	6 722,49934	675 746 174	17,29	0,0142

$$\text{Prix} = 78\,524 + 32,42417 \text{ Cylindrée} + 130,95663 \text{ Poids} \\ - 566,6694 \text{ Vitesse} - 27\,956 \text{ Consommation}$$

Pas 5 : variable « longueur » sélectionnée $R^2 = 0,9847$

Source	DDL	Somme des carrés	Carré moyen	Valeur F	Pr > F
Régression	5	6 369 365 682	1 273 873 136	38,52	0,0064
Résiduelle	3	99 199 674	33 066 558		
Totale	8	6 468 565 356			

Variable	Coefficient estimé	Écart-type	Type II SS	Valeur F	Pr > F
Constante	56 881	41 303	62 715 442	1,90	0,2622
Cylindrée	35,87933	11,85851	302 703 255	9,15	0,0565
Poids	121,49836	36,44703	367 456 100	11,11	0,0446
Longueur	13,14389	10,00264	57 096 319	1,73	0,2803
Vitesse	-671,3122	236,92110	265 479 258	8,03	0,0660
Consommation	-29 390	6 279,72493	724 294 847	21,90	0,0184

$$\begin{aligned}\text{Prix} = & 56\,881 + 35,87933 \text{ Cylindrée} + 121,49836 \text{ Poids} \\ & + 13,14389 \text{ Longueur} - 671,3122 \text{ Vitesse} \\ & - 29\,390 \text{ Consommation}\end{aligned}$$

Pas 6 : variable « puissance » sélectionnée $R^2 = 0,9922$

Toutes les variables ont été sélectionnées.

L'analyse de la variance et la valeur des paramètres sont données dans les tableaux 21.6 et 21.7.

$$\begin{aligned}\text{Prix} = & 88\,540 + 12,26191 \text{ Cylindrée} + 378,03692 \text{ Puissance} \\ & + 109,04896 \text{ Poids} + 17,82579 \text{ Longueur} - 797,7852 \text{ Vitesse} \\ & - 31\,849 \text{ Consommation}\end{aligned}$$

Tableau 21.9 – Résumé de quelques caractéristiques.

Pas	Variable sélectionnée	Nombre de variables introduites	R ² partiel	R ² du modèle	Valeur F	Pr > F
1	Cylindrée	1	0,8628	0,8628	44,04	0,0003
2	Consommation	2	0,0257	0,8886	1,39	0,2836
3	Poids	3	0,0543	0,9429	4,75	0,0812
4	Vitesse	4	0,0330	0,9758	5,46	0,0797
5	Longueur	5	0,0088	0,9847	1,73	0,2803
6	Puissance	6	0,0076	0,9922	1,94	0,2979

Backward elimination procedure

Au pas 0, toutes les variables sont sélectionnées ; les résultats sont ceux qui ont été obtenus au pas 6 de la procédure précédente. La première variable éliminée est la variable « cylindrée » qui avait été introduite au pas 1 précédemment. Ce phénomène très classique se produit assez souvent, les deux procédures ne fournissent pas toujours les mêmes sous-ensembles ni les meilleurs.

Pas 1 : variable « cylindrée » éliminée $R^2 = 0,9702$

Source	DDL	Somme des carrés	Carré moyen	Valeur F	Pr > F
Régression	5	6 408 663 636	1 281 732 727	64,19	0,0030
Résiduelle	3	59 901 720	19 967 240		
Totale	8	6 468 565 356			

Variable	Coefficient estimé	Écart-type	Type II SS	Valeur F	Pr > F
Constante	94 376	36 994	129 951 392	6,51	0,0839
Puissance	521,0063	125,88911	342 001 209	17,13	0,0256
Poids	110,07476	29,38102	280 259 135	14,04	0,0332
Longueur	18,97360	8,16258	107 885 454	5,40	0,1027
Vitesse	-806,9060	200,63282	322 968 393	16,17	0,0276
Consommation	-32 751	4 958,98498	870 919 497	43,62	0,0071

$$\text{Prix} = 94\,376 + 521,0063 \text{ Puissance} + 110,07476 \text{ Poids} \\ - 806,9060 \text{ Vitesse} - 32\,751 \text{ Consommation}$$

Pas 2 : variable « longueur » éliminée $R^2 = 0,9741$

Source	DDL	Somme des carrés	Carré moyen	Valeur F	Pr > F
Régression	5	6 300 778 182	1 575 194 545	37,55	0,0020
Résiduelle	3	167 787 174	41 946 793		
Totale	8	6 468 565 356			

Variable	Coefficient estimé	Écart-type	Type II SS	Valeur F	Pr > F
Constante	109 437	52 790	180 269 755	4,30	0,1069
Puissance	412,38019	169,42703	248 501 458	5,92	0,0717
Poids	129,93026	40,74559	426 538 516	10,17	0,0333
Vitesse	-607,2779	262,81035	223 969 497	5,34	0,0820
Consommation	-30 148	7 001,92566	777 635 996	18,54	0,0126

$$\text{Prix} = 109\,437 + 412,38019 \text{ Puissance} + 129,93026 \text{ Poids} \\ - 607,2779 \text{ Vitesse} - 30\,148 \text{ Consommation}$$

Toutes les autres variables sont significatives au seuil 0,01.

Tableau 21.10 – Résumé de la méthode *Backward elimination*.

Pas	Variable éliminée	Nombre de variables introduites	R^2 partiel	R^2 du modèle	Valeur F	$Pr > F$
1	Cylindrée	5	0,0015	0,9907	0,38	0,5996
2	Longueur	4	0,0167	0,9741	5,40	0,1027

*Stepwise selection method***Pas 1 : variable « cylindrée » sélectionnée**

On introduit la variable « cylindrée » qui donne la plus grande valeur à la variable F (44,04). On retrouve les résultats obtenus au pas 1 du procédé *Forward selection method*. Aucune autre variable ne peut être ajoutée au modèle car elles ont un niveau de signification inférieur à 0,1500.

Remarque : il est en général difficile d'obtenir le meilleur modèle. Le critère de sélection R^2 n'est pas nécessairement le plus intéressant, en effet c'est le modèle complet donc il donne à R^2 la valeur maximale. Un autre critère intéressant est d'étudier la valeur de la variance résiduelle qui tient compte des degrés de liberté.

Exemple 21.2.2

On étudie le prix des 13 voitures diesel en fonction des six critères retenus.

Tableau 21.11 – Caractéristiques statistiques des 6 critères.

	Cylindrée	Puissance	Poids	Longueur	Vitesse	Consommation	Prix
Moyenne	2 200,1538	138,6923	1 467,0769	4 378,9231	199,8462	4,8846	32 303,4615
Dev. Std	851,3086	66,8860	403,4806	513,7466	29,9133	1,1364	22 666,3037
Minimum	1 398	54	880	3 430	154	3,4	11 900
Maximum	4 423	300	2 320	5 039	250	7,2	95 550

Tableau 21.12 – Matrice de corrélation.

	Cylindrée	Puissance	Poids	Longueur	Vitesse	Consommation	Prix
Cylindre	1,0000	0,9603	0,8098	0,7621	0,8460	0,9167	0,9177
Puissance	0,9603	1,0000	0,8341	0,8759	0,9464	0,9192	0,9423
Poids	0,8098	0,8341	1,0000	0,8767	0,7895	0,9572	0,7859
Longueur	0,7621	0,8759	0,8767	1,0000	0,9365	0,8498	0,7852
Vitesse	0,8460	0,9464	0,7895	0,9365	1,0000	0,8310	0,8316
Consommation	0,9167	0,9192	0,9572	0,8498	0,8310	1,0000	0,8869
Prix	0,9177	0,9423	0,7859	0,7852	0,8316	0,8869	1,0000

Tableau 21.13 – Coefficients de corrélation et coefficients de détermination entre le prix et les différentes variables.

Variable Y	Variable X	Coefficient ρ	Coefficient R^2
Prix	Cylindrée	0,9177	0,8422
Prix	Puissance	0,9423	0,8880
Prix	Poids	0,7859	0,6176
Prix	Longueur	0,7242	0,5245
Prix	Vitesse	0,8316	0,7915
Prix	Consommation	0,8869	0,7666

Tableau 21.14 – Analyse de la variance.

Source	DDL	Somme des carrés	Carré moyen	Valeur F	Pr > F
Régression	6	5 794 078 993	965 679 832	15,62	0,020
Résiduelle	6	371 056 876	61 842 813		
Totale	12	6 165 135 869			

Conclusions :

 $F_{0,95}(6 ; 6) = 4,28$, on rejette l'hypothèse $\beta_i = 0$

$$R^2 = \frac{5\,794\,078\,993}{6\,165\,135\,869} = 0,9398 \quad \widehat{R}^2 = 0,8796$$

$$(\widehat{\sigma})^2 = 61\,842\,813 \quad \widehat{\sigma} = 7\,864,0202$$

Tableau 21.15 – Estimation des paramètres.

Variable	DDL	Coefficient estimé	Écart-type	Valeur de t	Pr > t
Constante	1	74 564	73 850	1,02	0,3516
Cylindrée	1	-12,95047	14,74654	0,77	0,4136
Puissance	1	771,50397	317,71317	5,90	0,0513
Poids	1	-9,99168	31,55958	0,10	0,7623
Longueur	1	11,68799	21,47129	0,30	0,6058
Vitesse	1	-837,6290	445,12997	3,54	0,1089
Consommation	1	2 068,76672	14104	0,02	0,8882

Prix = 74 564 - 12,95047 Cylindrée + 771,50397 Puissance - 9,99168 Poids
 + 11,68799 Longueur - 837,6290 Vitesse + 2 068,76672 Consommation

*Forward selection procedure***Pas 1 : variable « puissance » sélectionnée** $R^2 = 0,8880$

Source	DDL	Somme des carrés	Carré moyen	Valeur F	Pr > F
Régression	1	5 474 341 356	5 474 341 356	87,17	< 0,0001
Résiduelle	11	690 794 513	62 799 501		
Totale	12	6 165 135 869			

Variable	Coefficient estimé	Écart-type	Type II SS	Valeur F	Pr > F
Constante	-11 985	5 228,01559	330 046 077	5,26	0,0426
Puissance	319,33049	34,20208	5 474 341 356	87,17	< 0,0001

$$\text{Prix} = -11\,985 + 319,33049 \text{ Puissance}$$

Pas 2 : variable « vitesse » sélectionnée $R^2 = 0,9227$

Source	DDL	Somme des carrés	Carré moyen	Valeur F	Pr > F
Régression	2	5 688 742 013	2 844 371 007	59,71	< 0,0001
Résiduelle	10	476 393 856	4 763 938		
Totale	12	6 165 135 869			

Variable	Coefficient estimé	Écart-type	Type II SS	Valeur F	Pr > F
Constante	49 782	29470	135 944 648	2,85	0,1221
Puissance	504,56818	92,25871	1 424 922 125	29,91	0,0003
Vitesse	-437,6303	206,28965	214 400 657	4,50	0,0599

$$\text{Prix} = 49\,782 + 504,56818 \text{ Puissance} - 437,6303 \text{ Vitesse}$$

Pas 3 : variable « cylindre » sélectionnée $R^2 = 0,9367$

Source	DDL	Somme des carrés	Carré moyen	Valeur F	Pr > F
Régression	3	5 774 898 618	1 924 966 206	44,40	< 0,0001
Résiduelle	9	390 237 251	43 359 695		
Totale	12	6 165 135 869			

Variable	Coefficient estimé	Écart-type	Type II SS	Valeur F	Pr > F
Constante	96 030	43 207	214 184 771	4,94	0,0533
Cylindrée	-15,78688	11,19941	86 156 605	1,99	0,1923
Puissance	812,32178	235,39860	516 338 704	11,91	0,0073
Vitesse	-708,82514	275,21985	287 610 628	6,63	0,0299

Aucune autre variable ne peut être sélectionnée.

$$\text{Prix} = 96\,030 - 15,78688 \text{ Cylindrée} + 812,32178 \text{ Puissance} \\ - 708,82514 \text{ Vitesse}$$

Tableau 21.16 – Résumé de quelques caractéristiques.

Pas	Variable sélectionnée	Nombre de variables introduites	R ² partiel	R ² du modèle	Valeur F	Pr > F
1	Puissance	1	0,8880	0,8880	87,17	< 0,0001
2	Vitesse	2	0,0348	0,9227	4,50	0,0599
3	Cylindre	3	0,0140	0,9367	1,99	0,1923

Backward elimination procedure

On rappelle que, au pas 0, toutes les variables sont sélectionnées.

Pas 1 : variable « consommation » éliminée $R^2 = 0,9396$

Source	DDL	Somme des carrés	Carré moyen	Valeur F	Pr > F
Régression	5	5 792 748 503	1 158 549 701	21,78	0,0004
Résiduelle	7	372 387 366	53 198 195		
Totale	12	6 165 135 869			

Variable	Coefficient estimé	Écart-type	Type II SS	Valeur F	Pr > F
Constante	81 166	54 305	118 839 857	2,23	0,1787
Cylindrée	-12,74452	13,61496	46 613 463	0,88	0,3804
Puissance	792,21065	263,98461	479 094 333	9,01	0,0199
Poids	-6,07481	15,59979	8 067 231	0,15	0,7085
Longueur	11,22868	19,70123	17 280 952	0,32	0,5865
Vitesse	-855,42401	397,21847	246 718 037	4,64	0,0683

$$\text{Prix} = 81\,166 - 12,74452 \text{ Cylindrée} + 792,21065 \text{ Puissance} \\ - 6,077481 \text{ Poids} + 11,22868 \text{ Longueur} - 855,42401 \text{ Vitesse}$$

Pas 2 : variable « poids » éliminée $R^2 = 0,9383$

Source	DDL	Somme des carrés	Carré moyen	Valeur F	Pr > F
Régression	4	5 784 681 272	1 446 170 318	30,41	< 0,0001
Résiduelle	8	3 804 554 597	47 556 825		
Totale	12	6 165 135 869			

Variable	Coefficient estimé	Écart-type	Type II SS	Valeur F	Pr > F
Constante	87 954	48 628	155 579 599	3,27	0,1081
Cylindrée	-14,68847	12,97635	71 534 664	1,50	0,2549
Puissance	794,87530	249,51154	482 646 966	10,15	0,0129
Longueur	5,13991	11,33271	9 782 654	0,21	0,6622
Vitesse	-781,02029	329,26587	267 574 090	5,63	0,0451

$$\text{Prix} = 87\,954 - 14,68847 \text{ Cylindrée} + 794,87530 \text{ Puissance} \\ + 5,13991 \text{ Longueur} - 781,02029 \text{ Vitesse}$$

Pas 3 : variable « longueur » éliminée $R^2 = 0,9367$

Source	DDL	Somme des carrés	Carré moyen	Valeur F	Pr > F
Régression	3	5 774 898 618	192 466 206	44,40	< 0,0001
Résiduelle	9	390 237 251	43 359 695		
Totale	12	6 165 135 869			

Variable	Coefficient estimé	Écart-type	Type II SS	Valeur F	Pr > F
Constante	96 030	43 207	214 184 771	4,94	0,0533
Cylindrée	-15,78688	11,19941	86 156 605	1,99	0,1923
Puissance	812,32178	235,39860	516 338 704	11,91	0,0073
Vitesse	-708,82514	275,21985	287 610 628	6,63	0,0299

$$\text{Prix} = 96\,030 - 15,78688 \text{ Cylindrée} + 812,32178 \text{ Puissance} - 708,82514 \text{ Vitesse}$$

Pas 4 : variable « cylindrée » éliminée $R^2 = 0,9227$

Source	DDL	Somme des carrés	Carré moyen	Valeur F	Pr > F
Régression	2	5 688 742 013	2 844 371 007	59,71	< 0,0001
Résiduelle	10	476 393 856	47 639 386		
Totale	12	6 165 135 869			

Variable	Coefficient estimé	Écart-type	Type II SS	Valeur F	Pr > F
Constante	49 782	29 470	135 944 648	2,85	0,1221
Puissance	504,56818	92,25871	1 424 922 125	29,91	0,0003
Vitesse	-437,63033	206,28965	214 400 657	4,50	0,0599

Prix = 49 782 + 504,56818 Puissance - 437,63033 Vitesse

Toutes les autres variables sont significatives au seuil 0,01.

Tableau 21.17 – Résumé de la méthode *Backward elimination*.

Pas	Variable éliminée	Nombre de variables introduites	R ² partiel	R ² du modèle	Valeur F	Pr > F
1	Consommation	5	0,0002	0,9396	0,02	0,8882
2	Poids	4	0,0013	0,9383	0,15	0,7085
3	Longueur	3	0,0016	0,9367	0,21	0,6622
4	Cylindrée	2	0,0140	0,9227	1,99	0,0599

Stepwise selection method

Pas 1 : variable « puissance » sélectionnée

Pas 2 : variable « vitesse » sélectionnée

On introduit la variable « cylindrée » qui donne la plus grande valeur à la variable F (44,04), puis la variable « vitesse ». Ce sont les deux variables introduites aux pas 1 et 2 du procédé *Forward selection method*. Aucune autre variable ne peut être ajoutée au modèle car elles ont un niveau de signification inférieur à 0,1500.

Équation de la droite de régression avec une variable explicative :

$$\text{Prix} = -11\,985 + 319,33049 \text{ Puissance}$$

Équation de la droite de régression avec deux variables explicatives :

$$\text{Prix} = 49\,782 + 504,56818 \text{ Puissance} - 437,63033 \text{ Vitesse}$$

Remarque : ces deux modèles sont les meilleurs modèles respectivement à une seule variable explicative ($R^2 = 0,8880$) et à deux variables explicatives ($R^2 = 0,9227$). De même, avec trois variables explicatives, on ajoute la variable « cylindrée » ($R^2 = 0,9367$)... Ces résultats ont été donnés par la *Forward selection method*.

Conclusion : si on compare les catégories de voitures étudiées et compte tenu du nombre restreint d'individus, on s'aperçoit que ce ne sont pas les mêmes critères qui ont une influence sur le prix d'un véhicule.

22 • ANALYSE DE LA VARIANCE

22.1 Généralités et but de la théorie

Le but de la théorie de l'analyse de la variance est d'étudier la variabilité d'un produit en fonction d'un ensemble de facteurs de production dont on peut contrôler systématiquement les modes d'intervention et dont on souhaite dissocier la part revenant à chaque facteur. On distingue :

- l'analyse de la variance à simple entrée (étudiée dans le chapitre 16, paragraphe 16.3), un seul facteur est contrôlé, tous les autres facteurs étant regroupés sous le nom de « facteurs non contrôlés » ;
- l'analyse de la variance à double entrée qui étudie l'action simultanée de deux facteurs contrôlés, chacun agissant individuellement avec une possibilité d'interaction entre les deux ;
- l'analyse de la variance à entrées multiples qui étudie l'action simultanée de plusieurs facteurs contrôlés, chacun agissant individuellement avec une possibilité d'interaction à deux, trois facteurs... Différents plans d'expérience peuvent être mis en œuvre dans les cas où interviennent plusieurs facteurs contrôlés.

Les méthodes utilisées dans cette théorie supposent que les résultats d'une mesure, influencés en général par un certain nombre de facteurs, sont distribués suivant une loi normale. Cette hypothèse, difficile à vérifier si le nombre de mesures est petit, peut être admise sur la base de considérations physiques.

22.2 Analyse de la variance à double entrée

Soient A et B les deux facteurs contrôlés, le facteur A intervenant à k_A niveaux différents et le facteur B à k_B niveaux différents. On est amené à expérimenter, puis à analyser les $k_A k_B$ combinaisons de type $A_i B_j$. Pour chacune de ces combinaisons, on a effectué, pour examen, ν mesures. Si le nombre ν de mesures est le même pour chaque combinaison, l'analyse est *orthogonale*.

Pour une analyse orthogonale à double entrée, avec répétitions, le nombre total de mesures est :

$$N = \nu k_A k_B$$

Les résultats sont regroupés sous la forme d'un tableau à double entrée (tableau 22.1).

Tableau 22.1 – Analyse de la variance à double entrée (tableau des données).

	B_1	B_j	B_{k_B}	Moyenne
A_1				$x_{1..}$
A_i		$x_{ij\alpha}$ $x_{ij\beta}$. $x_{ij\nu}$		$x_{i..}$
A_{k_A}				$x_{k_A..}$
Moyenne	$x_{.1.}$	$x_{.j.}$	$x_{.k_B.}$	$\bar{x}$

Les différentes moyennes figurant dans le tableau sont :

- pour chaque case, case $A_i B_j$ par exemple : $x_{ij} = \frac{1}{\nu} \sum_{\alpha=1}^{\nu} x_{ij\alpha}$
- pour chaque ligne, ligne A_i par exemple : $x_{i..} = \frac{1}{\nu k_B} \sum_{j=1}^{k_B} \sum_{\alpha=1}^{\nu} x_{ij\alpha}$
- pour chaque colonne, colonne j par exemple : $x_{.j.} = \frac{1}{\nu k_A} \sum_{i=1}^{k_A} \sum_{\alpha=1}^{\nu} x_{ij\alpha}$

$$\text{Moyenne générale : } \bar{x} = \frac{1}{\nu k_A k_B} \sum_{i=1}^{k_A} \sum_{j=1}^{k_B} \sum_{\alpha=1}^{\nu} x_{ij\alpha}$$

22.2.1 Variations

Toutes les variations, et par conséquent tous les calculs, se ramènent, en fait, à une succession d'analyses de la variance à simple entrée.

On peut analyser l'effet global (AB) des facteurs A et B , sous la forme d'une analyse de la variance à simple entrée, où les différentes modalités du facteur contrôlé sont toutes les combinaisons des facteurs A et B .

On décompose la variation totale :

$$S^2 = \sum_{i=1}^{k_A} \sum_{j=1}^{k_B} \sum_{\alpha=1}^{\nu} (x_{ij\alpha} - \bar{x})^2$$

à l'aide de la variation globale calculée pour chacune des $k_A k_B$ combinaisons des facteurs A et B :

$$S_{(AB)}^2 = \nu \sum_{i=1}^{k_A} \sum_{j=1}^{k_B} (x_{ij.} - \bar{x})^2$$

et de la variation résiduelle par rapport à l'effet global (AB) :

$$S_R^2 = \sum_{i=1}^{k_A} \sum_{j=1}^{k_B} \sum_{\alpha=1}^{\nu} (x_{ij\alpha} - x_{ij.})^2$$

D'où une première décomposition de la variation totale sous la forme :

$$S^2 = S_{(AB)}^2 + S_R^2$$

Cette décomposition est insuffisante car elle ne fait pas apparaître l'action des facteurs A et B pris individuellement, d'une part, et leur interaction éventuelle, d'autre part. En écrivant la différence qui intervient dans la variation globale (AB) sous la forme :

$$x_{ij.} - \bar{x} = (x_{i..} - \bar{x}) + (x_{.j.} - \bar{x}) + [x_{ij.} - (x_{i..} + x_{.j.}) + \bar{x}]$$

on fait apparaître, après avoir élevé au carré et fait la sommation sur tous les indices, trois termes que l'on appelle :

– effet principal du facteur A :

$$S_A^2 = v k_B \sum_{i=1}^{k_A} (x_{i..} - \bar{x})^2$$

– effet principal du facteur B :

$$S_B^2 = v k_A \sum_{j=1}^{k_B} (x_{.j.} - \bar{x})^2$$

– interaction AB :

$$S_{AB}^2 = v \sum_{i=1}^{k_A} \sum_{j=1}^{k_B} [x_{ij.} - (x_{i..} + x_{.j.}) + \bar{x}]^2$$

L'effet principal du facteur A correspond à une analyse à simple entrée où les différentes modalités du facteur contrôlé sont les k_A lignes du tableau 22.1, comprenant chacune vk_B termes. Même interprétation pour l'effet principal du facteur B . L'interaction correspond à la variation résiduelle dans cette optique.

D'où la décomposition complète de la variation totale :

$$S^2 = S_A^2 + S_B^2 + S_{AB}^2 + S_R^2$$

22.2.2 Calcul rapide des différentes variations

Terme correctif :

$$\Delta = \frac{1}{v k_A k_B} \left(\sum_{i=1}^{k_A} \sum_{j=1}^{k_B} \sum_{\alpha=1}^v x_{ij\alpha} \right)^2$$

Variation totale :

$$S^2 = \sum_{i=1}^{k_A} \sum_{j=1}^{k_B} \sum_{\alpha=1}^v x_{ij\alpha}^2 - \Delta$$

Effet global (AB) :

$$S_{(AB)}^2 = \frac{1}{v} \sum_{i=1}^{k_A} \sum_{j=1}^{k_B} \left(\sum_{\alpha=1}^v x_{ij\alpha} \right)^2 - \Delta$$

Variation résiduelle :

$$S_R^2 = S^2 - S_{(AB)}^2$$

Effet principal A :

$$S_A^2 = \frac{1}{v k_B} \sum_{i=1}^{k_A} \left(\sum_{j=1}^{k_B} \sum_{\alpha=1}^v x_{ij\alpha} \right)^2 - \Delta$$

Effet principal B :

$$S_B^2 = \frac{1}{v k_A} \sum_{j=1}^{k_B} \left(\sum_{i=1}^{k_A} \sum_{\alpha=1}^v x_{ij\alpha} \right)^2 - \Delta$$

Interaction AB :

$$S_{AB}^2 = S_{(AB)}^2 - S_A^2 - S_B^2$$

22.2.3 Analyse de la variance et tests d'homogénéité

Le but de cette analyse est de mettre en évidence les effets significatifs.

Pour chaque niveau (A_i , B_j), on suppose que les résultats des mesures sont distribués selon une loi normale de même écart-type σ .

Si la *population* est *homogène*, c'est-à-dire si les facteurs A et B n'exercent aucune influence sur le résultat des mesures, l'ensemble des résultats peut être considéré comme un échantillon unique de $N = vk_A k_B$ valeurs extraites au hasard d'une population normale, et réparties également au hasard, avec v valeurs par case, dans les $k_A k_B$ cases du tableau des données (tableau 22.1).

Le résultat d'une mesure se met sous la forme :

$$x_{ij\alpha} = \lambda_0 + \xi_{ij\alpha}$$

λ_0 est une constante et $\xi_{ij\alpha}$ est une fluctuation aléatoire suivant une loi normale d'espérance nulle et d'écart-type σ .

Sous l'hypothèse d'homogénéité, les quotients des variations S_A^2 , S_B^2 , S_{AB}^2 et S_R^2 par la variance σ^2 suivent des lois du chi-deux dont les degrés de liberté sont donnés dans le tableau 22.2 :

Tableau 22.2 – Analyse de la variance à double entrée.

Variations	Somme des carrés	Degrés de liberté	Quotients
Effet principal A	S_A^2	$k_A - 1$	$V_A = \frac{S_A^2}{k_A - 1}$
Effet principal B	S_B^2	$k_B - 1$	$V_B = \frac{S_B^2}{k_B - 1}$
Interaction	S_{AB}^2	$(k_A - 1)(k_B - 1)$	$V_{AB} = \frac{S_{AB}^2}{(k_A - 1)(k_B - 1)}$
Variation résiduelle	S_R^2	$k_A k_B (v - 1)$	$V_R = \frac{S_R^2}{(v - 1) k_A k_B}$
Variation totale	S^2	$vk_A k_B - 1$	

Les différents quotients V_A , V_B , V_{AB} et V_R sont des estimations de la variance σ^2 basées sur leurs différents degrés de liberté.

Les *tests d'homogénéité* se font dans l'ordre suivant.

On étudie d'abord l'interaction AB que l'on compare à la variation résiduelle S_R^2 :

1) L'interaction AB n'est pas significative au seuil α si :

$$\frac{V_{AB}}{V_R} \leq F_{1-\alpha} [(k_A - 1)(k_B - 1) ; k_A k_B (v - 1)]$$

Dans cette hypothèse, on étudie les effets principaux A et B en les comparant à la variation résiduelle S_R^2 :

– l'effet A n'est pas significatif au seuil α si :

$$\frac{V_A}{V_R} \leq F_{1-\alpha} [k_A - 1 ; k_A k_B (v - 1)]$$

– l'effet B n'est pas significatif au seuil α si :

$$\frac{V_B}{V_R} \leq F_{1-\alpha} [k_B - 1 ; k_A k_B (v - 1)]$$

La population est homogène au seuil α si les trois effets A , B et AB ne sont pas significatifs. Dans ces conditions, le modèle adopté est le suivant :

$$x_{ij\alpha} = \lambda_0 + \xi_{ij\alpha}.$$

L'estimation de la constante λ_0 est la moyenne générale $\bar{x}$ et celle de la variance σ^2 , le quotient V_R .

2) L'interaction AB est significative au seuil α si :

$$\frac{V_{AB}}{V_R} > F_{1-\alpha} [(k_A - 1)(k_B - 1); k_A k_B (v - 1)]$$

On étudie les effets principaux A et B en les comparant à l'interaction AB :

– l'effet principal A n'est pas significatif si :

$$\frac{V_A}{V_{AB}} \leq F_{1-\alpha} [k_A - 1; (k_A - 1)(k_B - 1)]$$

– il est significatif si :

$$\frac{V_A}{V_{AB}} > F_{1-\alpha} [k_A - 1; (k_A - 1)(k_B - 1)]$$

Mêmes conclusions pour l'effet B .

La formule générale pour le résultat d'une mesure est la suivante :

$$x_{ij\alpha} = \lambda_0 + \lambda_{A_i} + \lambda_{B_j} + I_{A_i B_j} + \xi_{AB\alpha}$$

- λ_0 est une constante.
- λ_{A_i} est une correction intéressant un même niveau du facteur A , donc toutes les cases d'une même ligne.
- λ_{B_j} est une correction intéressant un même niveau du facteur B , donc toutes les cases d'une même colonne.
- $I_{A_i B_j}$ est une correction intéressant toutes les mesures d'une même case, elle est donc caractéristique d'une combinaison de deux niveaux des facteurs A et B .

Selon les résultats des tests, seuls certains termes figureront dans la décomposition.

On montre que les estimations des termes significatifs sont :

$$\hat{\lambda}_0 = \bar{x}$$

$$\hat{\lambda}_{A_i} = x_{i..} - \bar{x}$$

$$\hat{\lambda}_{B_j} = x_{.j.} - \bar{x}$$

$$\hat{I}_{A_i B_j} = x_{ij.} - (x_{i..} + x_{.j.}) + \bar{x}$$

Remarque

Les calculs sont longs mais il existe des logiciels d'analyse de la variance. Il suffit alors d'interpréter correctement les sorties des programmes.

Exemple 22.1

Les données sont volontairement simples afin de ne pas alourdir ce chapitre.

Un fabricant de coussinets en bronze fritté se propose de déterminer si la résistance à la rupture du bronze dépend des lots de poudre de cuivre et d'étain utilisés pour son élaboration. On réalise à partir de trois lots différents de poudre de cuivre (facteur A) et de trois lots différents de poudre d'étain (facteur B), neuf mélanges de composition identiques (90 % de cuivre et 10 % d'étain), correspondant aux neuf combinaisons deux à deux des lots de cuivre et d'étain utilisés.

À partir de chacun de ces neuf mélanges, on comprime, sous une même pression, quatre éprouvettes de flexion identique. Les trente-six éprouvettes obtenues sont ensuite frittées en une même opération dans un four à atmosphère réductrice. Ces éprouvettes sont enfin cassées sur une machine d'essai. Les charges de rupture ainsi déterminées arrondies à $0,1 \text{ kg/mm}^2$ près sont reportées dans le tableau 22.3 en hectogramme, en excès de la valeur de 2 kg/mm^2 choisie pour origine.

Tableau 22.3 – Charge de rupture de 36 éprouvettes en bronze fritté
(origine 2 kg/mm^2).

	Étain	B_1		B_2		B_3	
Cuivre							
A_1		6	7	1	1	0	5
		3	8	4	3	6	2
A_2		1	6	6	4	0	3
		7	4	4	10	2	2
A_3		6	10	8	3	2	4
		8	7	7	7	3	7

Une analyse de la variance à double entrée donne une réponse au fabricant.

Les étapes de calcul sont les suivantes :

- Somme par lignes : ligne A_1 : 46 ; ligne A_2 : 49 ; ligne A_3 : 72.
- Somme par colonnes : colonne B_1 : 73 ; colonne B_2 : 58 ; colonne B_3 : 36.
- Somme par cases : case $A_1 B_1$: 24 ; case $A_2 B_1$: 18 ; case $A_3 B_1$: 31 ; case $A_1 B_2$: 9 ; case $A_2 B_2$: 24 ; case $A_3 B_2$: 25 ; case $A_1 B_3$: 13 ; case $A_2 B_3$: 7 ; case $A_3 B_3$: 16.

– Variation totale :

Somme des termes : 167 ; terme correctif $\Delta = (167)^2/36 = 774,70$.

$$S^2 = (6^2 + 7^2 + \dots + 3^2 + 7^2) - 774,70 = 260,30$$

– Variation globale : $S_{(AB)}^2 = \frac{1}{4} (24^2 + 9^2 + \dots + 16^2) - \Delta = 129,60$

– Variation résiduelle : $S_R^2 = 260,30 - 129,60 = 130,70$

– Effet principal du facteur A : $S_A^2 = \frac{1}{12} (46^2 + 49^2 + 72^2) - \Delta = 33,70$

– Effet principal du facteur B : $S_B^2 = \frac{1}{12} (73^2 + 58^2 + 36^2) - \Delta = 57,70$

– Interaction AB (cuivre, étain) : $S_{AB}^2 = 129,60 - 33,70 - 57,70 = 38,20$

Tableau 22.4 – Analyse de la variance.

Variations	Somme des carrés	Degrés de liberté	Quotients
Effet principal du cuivre	$S_A^2 = 33,70$	2	$V_A = 16,85$
Effet principal de l'étain	$S_B^2 = 57,70$	2	$V_B = 28,85$
Interaction	$S_{AB}^2 = 38,20$	4	$V_{AB} = 9,55$
Résiduelle	$S_R^2 = 130,70$	27	$V_R = 4,85$
Totale	$S^2 = 260,30$	35	

On choisit un risque de première espèce $\alpha = 0,05$.

– L'interaction AB n'est pas significative car :

$$\frac{V_{AB}}{V_R} = \frac{9,55}{4,85} = 1,97 < F_{0,95}(4; 27) = 2,73$$

– Les effets principaux A et B sont significatifs car :

$$\frac{V_A}{V_R} = \frac{16,85}{4,85} = 3,48 > F_{0,95}(2; 27) = 3,35$$

$$\frac{V_B}{V_R} = \frac{28,85}{4,85} = 5,96 > F_{0,95}(2; 27) = 3,35$$

L'hypothèse d'homogénéité doit donc être rejetée. Les lots de cuivre et d'étain exercent une influence sur la résistance à la rupture de ces éprouvettes. Leurs effets sont additifs et il ne semble pas qu'il y ait une interaction dans les lots de cuivre et d'étain associés dans un même mélange.

22.3 Analyse de la variance orthogonale à entrées multiples

22.3.1 Généralités

Les méthodes d'analyse de la variance à double entrée peuvent être généralisées à l'étude d'un nombre quelconque de facteurs contrôlés. Les formules sont de plus en plus lourdes à écrire mais la théorie présente peu de difficultés.

Considérons, par exemple une analyse orthogonale à quatre facteurs contrôlés A , B , C et D . La décomposition de la variation totale doit tenir compte :

- des effets principaux A , B , C et D ,
- des interactions deux à deux AB , AC , AD , BC , BD , CD ,
- des interactions trois à trois ABC , ABD , ACD , BCD ,
- de l'interaction des quatre facteurs $ABCD$,
- et enfin de la variation résiduelle.

Pour calculer chaque effet ou chaque interaction, il suffit de considérer des analyses de la variance à simple entrée.

22.3.2 Analyse à triple entrée sans répétition

On considère une analyse à triple entrée sans répétition, le facteur contrôlé A intervient à k_A niveaux, le facteur contrôlé B à k_B niveaux et le facteur contrôlé C à k_C niveaux.

Pour chaque combinaison $A_i B_j C_l$ des trois facteurs, on a effectué une seule mesure (analyse sans répétition).

Dans ces conditions, on doit admettre que l'interaction du troisième ordre ABC n'est pas significative. Le quotient V_{ABC} donne l'estimation de la variance d'erreur σ^2 , il est utilisé dans les différents tests de Fisher pour trouver les actions significatives ou non. La variation totale est, en fait, l'effet global (ABC).

■ Décomposition de la variation totale

$$S^2 = S_A^2 + S_B^2 + S_C^2 + S_{AB}^2 + S_{AC}^2 + S_{BC}^2 + S_{ABC}^2$$

Tous les termes qui interviennent dans la décomposition de la variation totale, ainsi que la variation totale, se calculent de façon analogue par des formules

généralisant celles des paragraphes 22.2.1 et 22.2.2. Par exemple :

- effet global (AC) : $S_{(AC)}^2 = k_B \sum_{i=1}^{k_A} \sum_{l=1}^{k_C} (x_{i..l} - \bar{x})^2$
- effet principal C : $S_C^2 = k_A k_B \sum_{l=1}^{k_C} (x_{..l} - \bar{x})^2$
- interaction AC : $S_{AC}^2 = S_{(AC)}^2 - S_A^2 - S_C^2$

■ Calcul rapide des différents facteurs

Par exemple :

$$\Delta = \frac{1}{k_A k_B k_C} \left(\sum_{i=1}^{k_A} \sum_{j=1}^{k_B} \sum_{l=1}^{k_C} x_{ijl} \right)^2 \quad S_{(AC)}^2 = \frac{1}{k_B} \sum_{i=1}^{k_A} \sum_{l=1}^{k_C} \left(\sum_{j=1}^{k_B} x_{ijl} \right)^2 - \Delta$$

$$S_C^2 = \frac{1}{k_A k_B} \sum_{l=1}^{k_C} \left(\sum_{i=1}^{k_A} \sum_{j=1}^{k_B} x_{ijl} \right)^2 - \Delta \quad S_A^2 = \frac{1}{k_B k_C} \sum_{i=1}^{k_A} \left(\sum_{j=1}^{k_B} \sum_{l=1}^{k_C} x_{ijl} \right)^2 - \Delta$$

Tableau 22.5 – Tableau d'analyse de la variance à triple entrée sans répétition.

Variations	Somme des carrés	Degrés de liberté	Quotients
Effet principal A	S_A^2	$k_A - 1$	$V_A = \frac{S_A^2}{k_A - 1}$
Effet principal B	S_B^2	$k_B - 1$	$V_B = \frac{S_B^2}{k_B - 1}$
Effet principal C	S_C^2	$k_C - 1$	$V_C = \frac{S_C^2}{k_C - 1}$
Interaction AB	S_{AB}^2	$(k_A - 1)(k_B - 1)$	$V_{AB} = \frac{S_{AB}^2}{(k_A - 1)(k_B - 1)}$
Interaction AC	S_{AC}^2	$(k_C - 1)(k_A - 1)$	$V_{AC} = \frac{S_{AC}^2}{(k_A - 1)(k_C - 1)}$
Interaction BC	S_{BC}^2	$(k_C - 1)(k_B - 1)$	$V_{BC} = \frac{S_{BC}^2}{(k_B - 1)(k_C - 1)}$
Interaction ABC	S_{ABC}^2	$(k_C - 1)(k_A - 1)(k_B - 1)$	$V_{ABC} = \frac{S_{ABC}^2}{(k_A - 1)(k_B - 1)(k_C - 1)}$
Variation totale	S^2	$k_A k_B k_C - 1$	

22.3.3 Étude d'un cas simple

Pour illustrer cette méthode, on traite un cas simple obtenu en modifiant l'exemple 22.1 mais en gardant les mêmes valeurs numériques (pour utiliser les calculs précédents) :

- le facteur A correspond à trois températures différentes de frittage,
- le facteur B correspond à trois pressions différentes de compression,
- les observations placées à chaque angle des différentes cases du tableau correspondent à quatre durées de frittage : 15 min, 30 min, 45 min et 60 min (dans l'ordre angle supérieur gauche puis droit, angle inférieur gauche puis droit). Ces observations correspondent à un troisième facteur contrôlé, le facteur C .

Le tableau des données est donc le tableau 22.3.

Le nombre total d'observations est égal à 36.

Les variations qui n'ont pas été calculées sont les effets globaux (AC) et (BC), l'effet principal C et les interactions AC , BC et ABC ; la variation totale est l'effet global (ABC).

Les résultats sont donnés dans le tableau 22.6 d'analyse de la variance.

Tableau 22.6 – Analyse de la variance à triple entrée sans répétition (résultats numériques).

Variations	Somme des carrés	Degrés de liberté	Quotients
Effet principal A	$S_A^2 = 33,70$	2	$V_A = 16,85$
Effet principal B	$S_B^2 = 57,70$	2	$V_B = 28,85$
Effet principal C	$S_C^2 = 23,60$	3	$V_C = 7,87$
Interaction AB	$S_{AB}^2 = 38,20$	4	$V_{AB} = 9,55$
Interaction AC	$S_{AC}^2 = 4,30$	6	$V_{AC} = 0,716$
Interaction BC	$S_{BC}^2 = 39,70$	6	$V_{BC} = 6,61$
Interaction ABC	$S_{ABC}^2 = 63,10$	12	$V_{ABC} = 5,25$
Totale ou effet global (ABC)	$S^2 = 260,30$	35	

En prenant comme terme de comparaison le quotient V_{ABC} , on trouve que seul l'effet principal B (pression) est significatif car :

$$\frac{V_B}{V_{ABC}} = \frac{28,85}{5,25} = 5,5 > F_{0,95}(2; 12) = 3,88$$

22.4 Analyse de la variance emboîtée

Ce type d'analyse convient bien au cas où les facteurs sont *hiérarchisés* ou *emboîtés*.

22.4.1 Construction de l'arbre

Le plan d'expérience a la forme d'un arbre où l'observateur est le pied de l'arbre. Chaque sommet, ou nœud, représente une modalité particulière de l'un des facteurs susceptibles d'avoir une influence sur les mesures. Les branches extrêmes correspondent aux résultats des mesures (figure 22.1).

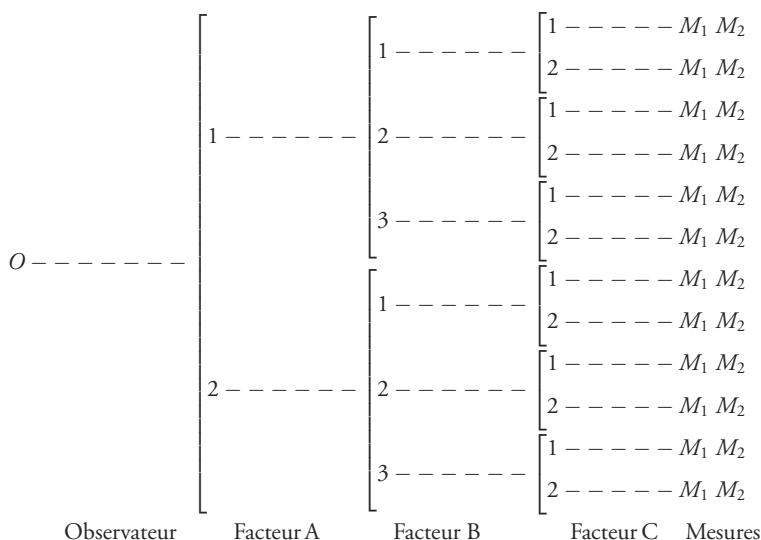


Figure 22.1 – Plan d'expérience : « analyse emboîtée ».

On suppose que l'on rencontre le même nombre de sommets dans tout trajet partant du pied de l'arbre et aboutissant à une branche extrême. L'ordre du plan est le nombre commun de sommets rencontrés. On suppose enfin qu'il part le même nombre de branches de chacun des sommets de même rang : le plan d'expérience est un plan orthogonal.

On note :

- k_A le nombre de branches partant du pied de l'arbre vers les sommets A ,
- k_B le nombre de branches de chaque sommet A vers les sommets B ,
- k_C le nombre de branches de chaque sommet B vers les sommets C ,
- v le nombre de branches de chaque sommet C (branches extrêmes).

Le nombre total N de mesures est :

$$N = vk_A k_B k_C$$

22.4.2 Tableau de l'analyse de la variance emboîtée

On part des sommets de rang le plus élevé, les sommets C dans le cas considéré. Les mesures recueillies $x_{ijl\alpha}$ peuvent être considérées comme les résultats d'une analyse de la variance à simple entrée dont les différents niveaux correspondraient aux k_C sommets de rang C .

La *variation résiduelle* ou *intraclasse* correspondant à cette analyse est :

$$M/ABC = \sum_{i=1}^{k_A} \sum_{j=1}^{k_B} \sum_{l=1}^{k_C} \sum_{\alpha=1}^v (x_{ijl\alpha} - x_{ijl.})^2$$

Cette variation caractérise la somme des dispersions des mesures provenant d'un même sommet de rang C et donc d'un même sommet de rang B et de rang A .

La moyenne $x_{ijl.}$ de toutes les mesures partant d'un même sommet C peut être prise comme mesure caractérisant ce sommet.

Partant ensuite des sommets de rang B , on peut considérer l'ensemble des moyennes $x_{ijl.}$ comme les données d'une analyse de la variance à simple entrée, dont les différents niveaux correspondraient aux sommets de rang B . On peut, comme précédemment, calculer la *variation résiduelle* ou *intraclasse* correspondant à cette analyse :

$$C/AB = \sum_{i=1}^{k_A} \sum_{j=1}^{k_B} \sum_{l=1}^{k_C} v (x_{ijl.} - x_{ij..})^2$$

Cette variation caractérise la somme, pour tous les sommets B , des dispersions des moyennes $x_{ijl.}$ caractérisant tous les sommets C reliés à un même sommet B .

Le même processus s'étend inclusivement jusqu'aux sommets A .

Partant enfin de l'observateur, on calcule la variation caractérisant la dispersion entre les sommets A ou *effet principal du facteur A* :

$$A = \sum_{i=1}^{k_A} v k_B k_C (x_{i...} - x_{....})^2$$

On a réalisé une suite d'analyses à simple entrée qui s'emboîtent les unes dans les autres, d'où le nom de ce type de plan d'expérience.

Tableau 22.7 – Tableau de l'analyse de la variance emboîtée.

Variations	Somme des carrés	Degrés de liberté	Quotients
Effet principal A	A	$\delta_A = k_A - 1$	$V_A = \frac{A}{\delta_A}$
Somme des effets principaux B pour un même A	B/A	$\delta_B = k_A(k_B - 1)$	$V_{B/A} = \frac{B/A}{\delta_B}$
Somme des effets principaux C pour un même A et un même B	C/AB	$\delta_C = k_A k_B (k_C - 1)$	$V_{C/AB} = \frac{C/AB}{\delta_C}$
Variations entre mesures pour un même C, un même B et un même A	M/ABC	$\delta_M = k_A k_B k_C (v - 1)$	$V_{M/ABC} = \frac{M/ABC}{\delta_M}$
Variation totale	$T = (ABCM)$	$\delta_T = v k_A k_B k_C - 1$	

On peut vérifier l'égalité suivante :

$$T = A + B/A + C/AB + M/ABC$$

22.4.3 Interprétation des résultats

La recherche des effets significatifs dépend du choix des modalités d'intervention des facteurs A , B et C .

■ Choix aléatoire

Le quotient $V_{M/ABC}$ sert de terme de comparaison pour déterminer si l'écart-type σ_C peut être considéré comme nul au seuil α .

Si l'on doit considérer l'écart-type σ_C comme différent de zéro, le quotient $V_{C/AB}$ est utilisé pour déterminer si l'écart-type σ_B peut être considéré comme nul au seuil α .

Si l'on doit considérer l'écart-type σ_B comme différent de zéro, le quotient $V_{B/A}$ est utilisé pour déterminer si l'écart-type σ_A peut être considéré comme nul au seuil α .

■ Choix systématique

Le quotient $V_{M/ABC}$ sert de terme de comparaison pour déterminer si l'un quelconque des trois effets A , B/A et C/AB est significatif au seuil α .

Exemple 22.2 Analyse de la variance emboîtée

Pour exécuter le revenu de petites pièces en acier, on dispose d'un four à circulation d'air forcé contenant un panier comportant 6 étages sur lesquels on place les pièces à traiter. Ces étages sont numérotés de 1 à 6 en sens inverse de la circulation d'air chaud.

On veut déterminer s'il existe une différence systématique de dureté entre les pièces traitées sur les différents étages. On devra tenir compte de la dispersion éventuelle de dureté entre pièces ainsi que de l'hétérogénéité de dureté au sein d'une même pièce.

On prélève sur chacun des six étages (facteur A , systématique), deux pièces (facteur B , aléatoire). Sur chaque pièce, on fait deux mesures de dureté.

Les résultats obtenus sont soumis à une analyse de la variance emboîtée.

Tableau 22.8 – Tableau des mesures (origine 30 Rockwell C).

Résultats des mesures de dureté												
Étages	1		2		3		4		5		6	
Pièces <i>B</i>	1	2	1	2	1	2	1	2	1	2	1	2
Mesures <i>M</i>	10	12	12	12	7	6	5	3	2	7	4	1
	8	12	11	10	9	8	6	5	2	4	4	5
Sommes par pièce	18	24	23	22	16	14	11	8	4	11	8	6
Sommes par étage	42		45		30		19		15		14	
Total général	165											

Tableau 22.9 – Analyse de la variance.

Variations	Somme des carrés	Degrés de liberté	Quotients
Entre étages	$A = 233,375$	5	46,675
Entre pièces d'un même étage	$B/A = 25,75$	6	4,292
Entre mesures d'une même pièce	$M/AB = 23,50$	12	1,958
Totale	$T = 282,625$	23	

L'hétérogénéité de dureté au sein d'une même pièce est caractérisée par une variance estimée à 1,958, soit un écart-type estimé à 1,4 point Rockwell, et approximativement une dispersion totale de la dureté estimée à $\pm 3,09 \times 1,4 = \pm 4,3$ points Rockwell.

Il n'a pas été possible de mettre en évidence de différences de dureté entre pièces.

En effet :

$$\frac{V_{B/A}}{V_{M/AB}} = \frac{4,292}{1,958} = 2,19 < F_{0,95}(6; 12) = 3$$

On peut admettre que l'écart-type σ_B est nul.

En revanche, il existe des différences significatives entre les étages, en effet :

$$\frac{V_A}{V_{A/B}} = \frac{46,675}{4,292} = 10,9 > F_{0,95}(5; 6) = 4,39$$

On peut déduire du tableau 22.8 des estimations des duretés moyennes des pièces traitées sur chaque étage, ainsi qu'un intervalle de confiance pour ces duretés. En prenant un seuil critique égal à 5 %, l'application de la formule de Student donne comme terme correctif :

$$\pm t_{0,975}(12) \frac{\hat{\sigma}}{\sqrt{4}} = \pm 2,179 \frac{1,4}{2} = \pm 1,5 \text{ point Rockwell}$$

D'où les duretés moyennes (exprimées en points Rockwell, en excès par rapport à la valeur 30 choisie comme origine) :

$$\text{Étage 1 : } 42/4 \pm 1,5 = 10,5 \pm 1,5$$

$$\text{Étage 2 : } 45/4 \pm 1,5 = 11,25 \pm 1,5$$

$$\text{Étage 3 : } 30/4 \pm 1,5 = 7,5 \pm 1,5$$

$$\text{Étage 4 : } 19/4 \pm 1,5 = 4,75 \pm 1,5$$

$$\text{Étage 5 : } 15/4 \pm 1,5 = 3,75 \pm 1,5$$

$$\text{Étage 6 : } 14/4 \pm 1,5 = 3,5 \pm 1,5$$

22.5 Carré latin

22.5.1 Présentation de la méthode

C'est un cas particulier de l'analyse à triple entrée très utilisé en pratique car il minimise le nombre d'essais.

Les trois facteurs contrôlés interviennent avec un même nombre de niveaux

$$k_A = k_B = k_C = k$$

Le plan d'expérience est représenté par un carré divisé en k lignes (les niveaux du facteur A) et en k colonnes (les niveaux du facteur B).

Chaque case du carré, correspondant à une combinaison du type $A_i B_j$, est associée à un niveau C_l du facteur C , de telle sorte que, sur chaque ligne et chaque colonne du carré, apparaisse une fois et une seule chacun des k niveaux du facteur C .

C'est un plan d'expérience limité ne permettant pas d'obtenir toutes les conclusions que l'on pourrait tirer d'une analyse classique.

22.5.2 Étude de cette méthode à partir d'un exemple simple

Les laboratoires de recherche d'un constructeur automobile souhaitent comparer la tenue en service de segments de « feu » (segment placé le plus près de la tête du piston). Les marques E , F et G sont des segments en fonte au nickel et la marque H en fonte ordinaire.

Quatre segments de chaque marque sont mis en service dans des conditions aussi identiques que possibles.

Les 16 segments ont été montés sur 4 moteurs d'un même type, comportant chacun 4 cylindres ; sur chaque moteur, un cylindre de chaque marque a été monté.

De plus, pour toutes les marques, les 4 segments d'une même marque ont été répartis à raison d'un segment par moteur entre les 4 positions possibles des cylindres dans le bloc moteur, les cylindres sont numérotés, à partir du côté embrayage dans l'ordre I, II, III et IV.

On cherche à déterminer si la différence de l'usure de ces segments provient de la marque, du moteur ou du cylindre.

Les résultats des 16 expériences (tableau 22.10) portent sur :

- 4 marques (repérées *E*, *F*, *G* et *H*) ;
- 4 moteurs (repérés 1, 2, 3, et 4) ;
- 4 positions de cylindres (repérées I, II, III et IV).

Tableau 22.10 – Conception du carré latin.

Segments	Moteurs			
	1	2	3	4
<i>E</i>	II	IV	I	III
<i>F</i>	I	III	II	IV
<i>G</i>	III	I	IV	II
<i>H</i>	IV	II	III	I

Ce plan d'expérience comporte trois facteurs. Pour effectuer une analyse de la variance, un tableau de données (tableau 22.11) et deux tableaux de calculs seront donc nécessaires avec les facteurs segments et moteurs (tableau 22.12), puis avec les facteurs moteurs et cylindres (tableau 22.13).

Les résultats de l'usure des segments sont donnés en excès de la valeur 50 mg.

Tests

Le quotient de référence est le quotient V_R pour les trois tests, le terme de comparaison est la variable de Fisher $F(3; 6)$ pour le seuil 0,95, soit la valeur 4,76.

$$\frac{V_S}{V_R} = \frac{19,16}{2,833} = 6,76 > F_{0,95}(3; 6) = 4,76$$

$$\frac{V_C}{V_R} = \frac{62,50}{2,833} = 22,11 > F_{0,95}(3; 6) = 4,76$$

$$\frac{V_M}{V_R} = \frac{14}{2,833} = 4,94 > F_{0,95}(3; 6) = 4,76$$

Il y a une différence significative entre les différentes marques de segments, entre les positions des cylindres, mais tout juste significative entre les moteurs.

Tableau 22.11 – Résultats des essais (segments, moteurs).

Segments	Moteurs			
	1	2	3	4
<i>E</i>	9	9	1	12
<i>F</i>	6	17	8	8
<i>G</i>	16	7	7	8
<i>H</i>	9	17	16	10

Tableau 22.12 – Résultats des calculs pour les facteurs segments et moteurs.

Segments	Moteurs				Somme par ligne	Moyenne par ligne	Somme des carrés
	1	2	3	4			
<i>E</i>	9	9	1	12	31	7,75	307
<i>F</i>	6	17	8	8	39	9,75	453
<i>G</i>	16	7	7	8	38	9,5	418
<i>H</i>	9	17	16	10	52	13	726
Somme par colonne	40	50	32	38	160		
Moyenne par colonne	10	12,5	8	9,5			
Somme des carrés	454	708	370	372			1 904

Tableau 22.13 – Résultats des calculs pour les facteurs moteurs et cylindres.

Cylindres	Moteurs				Somme par ligne	Moyenne par ligne	Somme des carrés
	1	2	3	4			
I	6	7	1	10	24	6	186
II	9	17	8	8	42	11,5	498
III	16	17	16	12	61	15,25	945
IV	9	9	7	8	33	8,25	275
Somme par colonne	40	50	32	38	160		1 904
Moyenne par colonne	10	12,5	8	9,5			
Somme des carrés	454	708	370	372			1 904

Tableau 22.14 – Analyse de la variance.

Variation	Somme des carrés	Degré de liberté	Quotients
Entre marques (segments)	57,5	3	$V_S = 19,16$
Entre moteurs	42	3	$V_M = 14$
Entre cylindres	187,5	3	$V_C = 62,5$
Résiduelle	17	6	$V_R = 2,833$
Totale	304	15	

Remarque

La variation résiduelle ainsi calculée intègre toutes les interactions que le carré latin ne peut pas évaluer.

Annexes

ANALYSE COMBINATOIRE

On considère une population S de n éléments et on définit dans cette population différents sous-ensembles ou sous-populations (tirage avec ou sans remise, échantillons ordonnés ou non).

Nombre de parties d'un ensemble

Le nombre de parties d'un ensemble S de n éléments a pour cardinal 2^n .

Exemple 1

Soit un ensemble de trois éléments (a, b, c) :

- une partie ne contient aucun élément, c'est l'ensemble vide $\emptyset$.
- trois parties ne contiennent qu'un élément, ce sont les singletons (a) , (b) et (c) .
- trois parties contiennent deux éléments (a, b) , (a, c) et (b, c) .
- une partie contient trois éléments, c'est l'ensemble entier (a, b, c) .

Au total : $1 + 3 + 3 + 1 = 8 = 2^3$.

Permutations sans répétition de n objets

L'espace Ω des permutations de n objets a pour cardinal $n!$.

C'est le nombre de bijections de l'ensemble $(1, 2, \dots, n)$ sur lui-même.

Exemple 2

Nombre de permutations de l'ensemble (a, b, c, d) de quatre objets.

$abcd, abdc, acbd, acdb, adbc, adcb, bacd, badc, bcad, bcda, bdac, bdca,$
 $cabd, cadb, cbad, cbda, cdab, cdba, dabc, dacb, dbac, dbca, dcab, dcba,$

soit $24 = 4!$ permutations.

Échantillon de taille r avec remise

Un échantillon de taille r , *avec remise*, est une application de l'ensemble $(1, \dots, r)$ dans l'ensemble S . Pour chaque élément de l'échantillon, n choix sont possibles, donc l'espace Ω des échantillons de taille $r \leq n$, avec remise, a pour cardinal n^r .

C'est le nombre d'arrangements avec répétition de n éléments pris r à r .

Exemple 3

Soit un ensemble de quatre lettres A, B, C et D. Écrire tous les mots possibles que l'on peut former, avec ou sans répétition, avec deux lettres prises parmi ces quatre lettres :

AA, AB, AC, AD, BA, BB, BC, BD, CA, CB, CC, CD, DA, DB, DC, DD

On obtient seize mots différents, on a quatre choix pour la première lettre, quatre choix pour la deuxième, soit $4 \times 4 = 4^2 = 16$.

Échantillon de taille $r \leq n$ sans remise

Un échantillon de taille $r \leq n$, *sans remise*, est une injection de l'ensemble $(1, \dots, r)$ dans l'ensemble S . L'espace Ω des échantillons ainsi définis a pour cardinal le nombre d'arrangements sans répétition de r éléments pris parmi n éléments :

$$A_n^r = n(n-1) \dots (n+1-r) = \frac{n!}{(n-r)!}$$

On a, en effet, n choix pour le premier élément, $(n-1)$ pour le deuxième, etc.

Exemple 4

Dans une course, dix chevaux prennent le départ. Il y a $10 \times 9 \times 8 = 720$ tiercés possibles.

10 choix pour le premier cheval, 9 pour le deuxième et 8 pour le troisième.

Sous-population de taille $r \leq n$

Une sous-population de taille r est un sous-ensemble de la population, de cardinal r . L'espace Ω des sous-populations de taille r a pour cardinal le

nombre de *combinaisons*, sans répétition de r éléments pris parmi n éléments ; ce nombre est égal à :

$$\binom{n}{r} = C_n^r = \frac{n!}{r! (n-r)!}$$

$$C_n^r = \frac{A_n^r}{r!}$$

Relations entre les combinaisons

$$C_n^r = C_n^{n-r} \quad C_n^r = C_{n-1}^r + C_{n-1}^{r-1}$$

Exemple 5

(Suite de l'exemple 4.) Quel est le nombre de tiercé possibles sans tenir compte de l'ordre ?

Avec trois chevaux appelés a , b et c , un tiercé sans ordre est (a, b, c) , il lui correspond $6 = 3!$ tiercés ordonnés :

(a, b, c) , (a, c, b) , (b, a, c) , (b, c, a) , (c, a, b) , (c, b, a)

Le nombre de tiercés sans ordre est égal à $720/6 = 120$.

Le nombre de tiercés avec ordre est supérieur au nombre de tiercés dans le désordre.

Sous-population de taille r avec répétition

Un élément peut apparaître plusieurs fois dans une telle sous-population. L'espace Ω peut aussi être considéré comme l'espace associé à l'expérience consistant à placer r éléments « indiscernables » dans n cases données :

$$\binom{n+r-1}{r} = C_{n+r-1}^r = \frac{(n+r-1)!}{r! (n-1)!}$$

Exemple 6

Soit trois éléments a , b et c .

Il y a $C_3^2 = C_3^1 = 3$ combinaisons de ces trois éléments pris deux à deux sans répétition : ab , ac et bc .

En revanche, il y a $C_{3+2-1}^2 = C_4^2 = \frac{4!}{2! 2!} = 6$ combinaisons de ces trois éléments pris deux à deux avec répétition : aa , ab , ac , bb , bc , cc .

RAPPELS MATHÉMATIQUES

Changement de variables et Jacobien

Soit f une densité de probabilité dans $\mathbb{R}^n$. On considère une application T , mesurable et bijective de $\mathbb{R}^n$ dans $\mathbb{R}^n$. À chaque élément $X = (x_1, x_2 \dots x_n)$, cette application associe l'élément $Y = (y_1, y_2 \dots y_n)$, défini par les formules :

$$y_j = T_j(x_1, \dots, x_n) \quad j = 1, \dots, n$$

Les formules définissant la transformation inverse sont :

$$x_i = h_i(y_1, \dots, y_n) \quad i = 1, \dots, n$$

On suppose que les fonctions h_i sont continûment dérivables par rapport à chaque variable y_j . On considère la matrice dont les coefficients sont les dérivées des n fonctions h_i par rapport aux n variables y_j . Le déterminant de cette matrice est le *Jacobien* J de la transformation.

La *densité* g de la variable aléatoire $Y = T(X)$ est donnée par :

$$g(y_1, \dots, y_n) = |J(y_1, \dots, y_n)| f[h_1(y_1, \dots, y_n), \dots, h_n(y_1, \dots, y_n)]$$

Exemple

Soit le couple de variables aléatoires X et Y , de loi conjointe $f(x, y)$.

On considère le changement de variables, $Z = X + Y$ et $W = X - Y$.

La transformation inverse est $X = 1/2(Z + W)$ et $Y = 1/2(Z - W)$.

Le Jacobien de la transformation est :

$$J = \begin{vmatrix} \frac{dx}{dz} & \frac{dx}{dw} \\ \frac{dy}{dz} & \frac{dy}{dw} \end{vmatrix} = \begin{vmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & -\frac{1}{2} \end{vmatrix} = -\frac{1}{2}$$

d'où la densité :

$$g(z, w) = f\left(\frac{z+w}{2}, \frac{z-w}{2}\right) \times \frac{1}{2}$$

Applications

1) Les variables X et Y sont indépendantes et suivent des lois uniformes sur $[0, 1]$:

$$g(z, w) = \begin{cases} \frac{1}{2} & 0 \leq z \leq 2 \quad \text{et} \quad -1 \leq w \leq 1 \\ 0 & \text{sinon} \end{cases}$$

2) Les variables X et Y sont indépendantes et suivent des lois normales, centrées, réduites. La formule précédente donne :

$$\begin{aligned} g(z, w) &= \frac{1}{4\pi} \exp \left[-\frac{1}{8} \{ (z+w)^2 + (z-w)^2 \} \right] \\ &= \frac{1}{4\pi} \exp \left[-\frac{1}{4} (z^2 + w^2) \right] \end{aligned}$$

Fonction Γ

La fonction Γ est définie pour $x > 0$ par l'intégrale :

$$\Gamma(x) = \int_0^{\infty} e^{-t} t^{x-1} dt$$

Elle vérifie la *relation fonctionnelle* :

$$\Gamma(x+1) = x\Gamma(x)$$

À partir de cette relation, on définit la fonction Γ pour les valeurs négatives de x . En effet, si :

$$0 < 1+x < 1 \quad \Rightarrow \quad -1 < x < 0$$

La relation fonctionnelle permet de définir la fonction Γ pour $-1 < x < 0$. Par un procédé analogue, on définit cette fonction pour $-2 < x < -1$, et donc pour toutes les valeurs négatives.

Les points de discontinuité de cette fonction sont les entiers négatifs et la valeur 0.

La fonction Γ généralise la notion de factorielle pour les nombres réels positifs :

$$\Gamma(n) = (n-1)!$$

Résultats remarquables

$$x \rightarrow 0 \quad \Gamma(x) \rightarrow \infty$$

$$x \rightarrow \infty \quad \Gamma(x) \rightarrow \infty$$

$$\Gamma(z) \Gamma(1-z) = \frac{\pi}{\sin \pi z}$$

$$\Gamma\left(k + \frac{1}{2}\right) = \frac{1 \times 3 \times 5 \times \dots \times (2k-1)}{2^k} \Gamma\left(\frac{1}{2}\right)$$

$$\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$$

$$n! \cong n^n e^{-n} \sqrt{2\pi n} \quad (\text{formule de Stirling})$$

Constante d'Euler

Elle est définie, soit par l'intégrale :

$$\gamma = - \int_0^{\infty} e^{-t} \operatorname{Ln} t \, dt = 0,57722$$

ou par la limite suivante :

$$\gamma = \lim \left(1 + \frac{1}{2} + \frac{1}{3} + \dots + \frac{1}{n} - \operatorname{Ln}(n) \right) \quad \text{quand } n \rightarrow \infty$$

Fonction bêta

Elle est définie par l'intégrale :

$$B(p, q) = \int_0^1 t^{p-1} (1-t)^{q-1} \, dt$$

et vérifie les relations :

$$B(p, q) = \frac{\Gamma(p) \Gamma(q)}{\Gamma(p+q)} = B(q, p)$$

Les fonctions Γ et bêta sont appelées fonctions eulériennes de première et deuxième espèces.

Fonction caractéristique

La *fonction caractéristique* ou *indicatrice* de la partie A d'un ensemble X est la fonction notée 1_A telle que :

$$1_A = \begin{cases} 1 & \text{si } x \in A \\ 0 & \text{si } x \notin A \end{cases}$$

Matrices M -symétrique

Soit V un espace vectoriel de dimension finie n , muni d'une métrique M (matrice symétrique définie positive). Le produit scalaire dans V est défini par :

$$\langle \underline{u}, \underline{v} \rangle = {}^t \underline{u} M \underline{v}$$

où ${}^t \underline{u}$ est le vecteur ligne, transposé du vecteur colonne $\underline{u}$.

A est une matrice carrée $n \times n$, ${}^t A$ la matrice transposée (permutation des lignes et des colonnes). L'adjointe A^* de la matrice A est définie par :

$$\langle A^* \underline{u}, \underline{v} \rangle = \langle \underline{u}, A \underline{v} \rangle \quad \forall \underline{u}, \underline{v}$$

La matrice A est M -symétrique si $A^* = A$ c'est-à-dire si :

$$\begin{aligned} \langle A \underline{u}, \underline{v} \rangle &= \langle \underline{u}, A \underline{v} \rangle \quad \forall \underline{u}, \underline{v} \quad \text{donc } {}^t u M A \underline{v} = {}^t u {}^t A M \underline{v} \\ M A &= {}^t A M \end{aligned}$$

Il en résulte que A est diagonalisable, que ses valeurs propres sont réelles et que ses vecteurs sont deux à deux M -orthogonaux (leur produit scalaire est nul).

Tribu

Soit X un ensemble quelconque. Un ensemble T de parties de X est une *tribu* définie sur X si cet ensemble vérifie les trois axiomes suivants :

- 1) l'ensemble X et l'ensemble vide $\emptyset$ appartiennent à T ,
- 2) si l'ensemble A appartient à la tribu, son complémentaire appartient aussi à T :

$$A \in T \Rightarrow \bar{A} \in T$$

- 3) toute réunion dénombrable d'éléments A_i de T appartient à T .

Une tribu est stable par passage au complémentaire (propriété 2) et stable par intersection dénombrable (propriété 3). De ces deux propriétés, on déduit qu'une tribu est stable par intersection dénombrable.

Exemples

- L'ensemble formé des deux parties X et $\emptyset$ est une tribu.
- L'ensemble P de toutes les parties de X est une tribu.

Produit de convolution

Soient f et g deux fonctions continues presque partout sur $] -\infty, +\infty [$. On appelle *produit de convolution* ou *de composition* de f et g la fonction $h = f * g$, définie par :

$$h(x) = \int_{-\infty}^{+\infty} f(t) g(x-t) dt$$

Ce produit est commutatif, en effet :

$$h(x) = \int_{-\infty}^{+\infty} f(t) g(x-t) dt = \int_{-\infty}^{+\infty} f(x-u) g(u) du = h_1(x)$$

où $h_1 = g * f$.

Tribu de Borel

Soit X un espace topologique. L'ensemble O des ouverts de X ne forme pas une tribu, le complémentaire d'un ouvert n'est pas un ouvert en général.

On considère la tribu B engendrée par l'ensemble O . Cette tribu contient :

- les ouverts et les fermés de X ,
- les réunions et les intersections dénombrables d'ouverts et de fermés,
- et, en général, beaucoup d'autres ensembles.

La tribu B est appelée *tribu de Borel*, ses éléments sont les *boréliens*.

Même dans le cas où $X = \mathbb{R}$, il est difficile de donner la description précise des ensembles de B . On montre que la tribu de Borel de $\mathbb{R}$ est engendrée par l'ensemble des demi-droites $] -\infty, a[$ ou par l'ensemble des demi-droites $[a, +\infty]$ ou par l'ensemble des intervalles.

TABLES STATISTIQUES

Loi binomiale

Tables 1.1 et 1.2 : probabilités individuelles pour certaines valeurs des paramètres n et p . Pour $p' = 1 - p$, les probabilités sont les mêmes que pour p à condition de changer k en $n - k$.

Pour $n > 50$ et $p < 0,10$ approximation par la loi de Poisson de paramètre np .
Pour $np > 5$ et $n(1 - p) > 5$, approximation par la loi normale de paramètres np pour la moyenne et $\sqrt{np(1 - p)}$ pour l'écart-type (chapitre 6, paragraphe 6.6.6).

Tables 2.1 et 2.2 : probabilités cumulées.

Pour les approximations, mêmes résultats que pour les probabilités individuelles.

Loi de Poisson

Tables 3.1 et 3.2 : probabilités individuelles pour certaines valeurs du paramètre λ .

Pour $\lambda > 18$, approximation par la loi normale de paramètres λ pour la moyenne et $\sqrt{\lambda}$ pour l'écart-type (chapitre 6, paragraphe 6.6.6).

Tables 4.1 et 4.2 : probabilités cumulées pour certaines valeurs du paramètre λ .

Approximation, même résultat que pour les probabilités individuelles.

Loi normale

Table 5.1 : fonction de répartition de la loi normale réduite.

La table donne les valeurs de $P = F(u)$ pour $u \geq 0$. Par exemple :

$$\Pr(U < 1,55) = 0,9394$$

Pour les valeurs de $u \leq 0$, on utilise la propriété $F(u) = 1 - F(-u)$.

Exemple

$$\Pr(U < -1,55) = 1 - \Pr(U < 1,55) = 1 - 0,9394 = 0,0606$$

Table 5.2 : fractile de la loi normale réduite.

Le fractile d'ordre p est défini par : $\Pr(U < u_p) = F(u_p) = P$.

Pour $P \leq 0,50$, on utilise la colonne de gauche et la ligne supérieure, les fractiles sont négatifs.

Exemple

$$\Pr(U < -0,6967) = 0,243$$

Pour $P \geq 0,50$, on utilise la colonne de droite et la ligne inférieure, les fractiles sont positifs.

Exemple

$$\Pr(U < 0,2715) = 0,607$$

Loi du chi-deux

Table 6 : le fractile d'ordre α pour le degré de liberté ν est défini par :

$$\Pr(\chi^2(\nu) < \chi_\alpha^2(\nu)) = \alpha$$

Exemple

$$\chi_{0,95}^2(15) = 25,0$$

$$\Pr(\chi^2(15) < 25,0) = 0,95$$

Loi de Fisher

Les tables 7.1a à 7.4b donnent les fractiles F_α d'ordre α pour certaines valeurs des degrés de liberté ν_1 et ν_2 .

Ces fractiles correspondent à $\alpha \geq 0,95$; pour les valeurs $\alpha \leq 0,05$, on utilise la relation :

$$F_\alpha(\nu_1 ; \nu_2) = \frac{1}{F_{1-\alpha}(\nu_2 ; \nu_1)}$$

Exemple

$$\Pr[F(8 ; 12) < 3,512] = 0,975$$

$$\Pr\left[F(8 ; 12) < \frac{1}{4,2} = 0,238\right] = 0,025$$

Loi de Student

La table 8 donne les fractiles t_α d'ordre α de la loi de Student pour certaines valeurs de $\alpha \geq 0,60$. Pour les valeurs de $\alpha \leq 0,60$, on utilise la relation $t_\alpha = -t_{1-\alpha}$.

$$\Pr(t(9) < 2,8214) = 0,99$$

$$\Pr(t(11) < -0,8755) = 1 - 0,80 = 0,20$$

Pour $\alpha > 100$, on calcule les fractiles en utilisant l'approximation par la loi normale centrée réduite.

Table de nombre au hasard

La table 9 donne 500 nombres au hasard. La présentation par colonnes de 5 nombres et par lignes de 5 nombres facilite la lecture.

Toutes les tables ont été obtenues en utilisant le logiciel Microsoft Excel 2000.

TABLE 1-1			LOI BINOMIALE									
Probabilités individuelles $\Pr(k) = C_n^k p^k (1-p)^{n-k}$												
n	p	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09	0,10	
	k											
5	0	0,95099	0,90392	0,85873	0,81537	0,77378	0,73390	0,69569	0,65908	0,62403	0,59049	
	1	0,04803	0,09224	0,13279	0,16987	0,20363	0,23422	0,26182	0,28656	0,30859	0,32805	
	2	0,00097	0,00376	0,00821	0,01416	0,02143	0,02990	0,03941	0,04984	0,06104	0,07290	
	3	0,00001	0,00008	0,00025	0,00059	0,00113	0,00191	0,00297	0,00433	0,00604	0,00810	
	4	0,00000	0,00000	0,00000	0,00001	0,00003	0,00006	0,00011	0,00019	0,00030	0,00045	
10	5				0,00000	0,00000	0,00000	0,00000	0,00000	0,00001	0,00001	
	0	0,90438	0,81707	0,73742	0,66483	0,59874	0,53862	0,48398	0,43439	0,38942	0,34868	
	1	0,09135	0,16675	0,22807	0,27701	0,31512	0,34380	0,36429	0,37773	0,38514	0,38742	
	2	0,00415	0,01531	0,03174	0,05194	0,07463	0,09875	0,12339	0,14781	0,17141	0,19371	
	3	0,00011	0,00083	0,00262	0,00577	0,01048	0,01681	0,02477	0,03427	0,04521	0,05740	
	4	0,00000	0,00003	0,00014	0,00042	0,00096	0,00188	0,00326	0,00522	0,00782	0,01116	
	5		0,00000	0,00001	0,00002	0,00006	0,00014	0,00029	0,00054	0,00093	0,00149	
15	6			0,00000	0,00000	0,00000	0,00001	0,00002	0,00004	0,00008	0,00014	
	7				0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00001	
	0	0,86006	0,73857	0,63325	0,54209	0,46329	0,39529	0,33670	0,28630	0,24301	0,20589	
	1	0,13031	0,22609	0,29378	0,33880	0,36576	0,37847	0,38015	0,37343	0,36051	0,34315	
	2	0,00921	0,03230	0,06360	0,09882	0,13475	0,16910	0,20029	0,22731	0,24958	0,26690	
	3	0,00040	0,00286	0,00852	0,01784	0,03073	0,04677	0,06533	0,08565	0,10696	0,12851	
	4	0,00001	0,00017	0,00079	0,00223	0,00485	0,00896	0,01475	0,02234	0,03174	0,04284	
	5	0,00000	0,00001	0,00005	0,00020	0,00056	0,00126	0,00244	0,00427	0,00691	0,01047	
20	6		0,00000	0,00000	0,00001	0,00005	0,00013	0,00031	0,00062	0,00114	0,00194	
	7				0,00000	0,00000	0,00001	0,00003	0,00007	0,00014	0,00028	
	8						0,00000	0,00000	0,00001	0,00001	0,00003	
	0	0,81791	0,66761	0,54379	0,44200	0,35849	0,29011	0,23424	0,18869	0,15164	0,12158	
	1	0,16523	0,27249	0,33637	0,36834	0,37735	0,37035	0,35262	0,32816	0,29996	0,27017	
	2	0,01586	0,05283	0,09883	0,14580	0,18868	0,22457	0,25214	0,27109	0,28183	0,28518	
	3	0,00096	0,00647	0,01834	0,03645	0,05958	0,08601	0,11387	0,14144	0,16724	0,19012	
	4	0,00004	0,00056	0,00241	0,00645	0,01333	0,02333	0,03643	0,05227	0,07030	0,08978	
	5	0,00000	0,00004	0,00024	0,00086	0,00224	0,00477	0,00877	0,01454	0,02225	0,03192	
	6		0,00000	0,00002	0,00009	0,00030	0,00076	0,00165	0,00316	0,00550	0,00887	
30	7			0,00000	0,00001	0,00003	0,00010	0,00025	0,00055	0,00109	0,00197	
	8				0,00000	0,00000	0,00001	0,00003	0,00008	0,00017	0,00036	
	9						0,00000	0,00000	0,00001	0,00002	0,00005	
	10							0,00000	0,00000	0,00000	0,00001	
	0	0,73970	0,54548	0,40101	0,29386	0,21464	0,15626	0,11337	0,08197	0,05905	0,04239	
	1	0,22415	0,33397	0,37207	0,36732	0,33890	0,29921	0,25599	0,21382	0,17521	0,14130	
	2	0,03283	0,09883	0,16686	0,22192	0,25864	0,27693	0,27939	0,26961	0,25127	0,22766	
	3	0,00310	0,01882	0,04816	0,08630	0,12705	0,16498	0,19627	0,21881	0,23194	0,23609	
	4	0,00021	0,00259	0,01005	0,02427	0,04514	0,07108	0,09972	0,12843	0,15484	0,17707	
	5	0,00001	0,00028	0,00162	0,00526	0,01235	0,02359	0,03903	0,05807	0,07963	0,10230	
30	6	0,00000	0,00002	0,00021	0,00091	0,00271	0,00627	0,01224	0,02104	0,03281	0,04736	
	7		0,00000	0,00002	0,00013	0,00049	0,00137	0,00316	0,00627	0,01113	0,01804	
	8			0,00000	0,00002	0,00007	0,00025	0,00068	0,00157	0,00316	0,00576	
	9				0,00000	0,00001	0,00004	0,00013	0,00033	0,00076	0,00157	
	10					0,00000	0,00001	0,00002	0,00006	0,00016	0,00037	
	11						0,00000	0,00000	0,00001	0,00003	0,00007	
	12							0,00000	0,00000	0,00000	0,00001	

TABLE 1-2		LOI BINOMIALE									
Probabilités individuelles $\Pr(k) = C_n^k p^k (1-p)^{n-k}$											
n	p	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09	0,10
	k										
40	0	0,66897	0,44570	0,29571	0,19537	0,12851	0,08416	0,05487	0,03561	0,02300	0,01478
	1	0,27029	0,36384	0,36583	0,32561	0,27055	0,21488	0,16519	0,12384	0,09097	0,06569
	2	0,05324	0,14479	0,22063	0,26456	0,27767	0,26746	0,24246	0,21000	0,17545	0,14233
	3	0,00681	0,03743	0,08643	0,13963	0,18511	0,21624	0,23116	0,23130	0,21979	0,20032
	4	0,00064	0,00707	0,02473	0,05381	0,09012	0,12768	0,16094	0,18605	0,20108	0,20589
	5	0,00005	0,00104	0,00551	0,01614	0,03415	0,05868	0,08722	0,11648	0,14318	0,16471
	6	0,00000	0,00012	0,00099	0,00392	0,01049	0,02185	0,03830	0,05908	0,08261	0,10676
	7		0,00001	0,00015	0,00079	0,00268	0,00677	0,01400	0,02495	0,03968	0,05761
	8		0,00000	0,00002	0,00014	0,00058	0,00178	0,00435	0,00895	0,01619	0,02641
	9			0,00000	0,00002	0,00011	0,00040	0,00116	0,00277	0,00569	0,01043
	10				0,00000	0,00002	0,00008	0,00027	0,00075	0,00175	0,00359
	11					0,00000	0,00001	0,00006	0,00018	0,00047	0,00109
	12						0,00000	0,00001	0,00004	0,00011	0,00029
	13							0,00000	0,00001	0,00002	0,00007
	14								0,00000	0,00000	0,00001
50	0	0,60501	0,36417	0,21807	0,12989	0,07694	0,04533	0,02656	0,01547	0,00896	0,00515
	1	0,30556	0,37160	0,33721	0,27060	0,20249	0,14467	0,09994	0,06725	0,04428	0,02863
	2	0,07562	0,18580	0,25552	0,27623	0,26110	0,22624	0,18430	0,14326	0,10730	0,07794
	3	0,01222	0,06067	0,12644	0,18416	0,21987	0,23106	0,22195	0,19932	0,16980	0,13857
	4	0,00145	0,01455	0,04595	0,09016	0,13598	0,17329	0,19629	0,20365	0,19732	0,18090
	5	0,00013	0,00273	0,01307	0,03456	0,06584	0,10176	0,13593	0,16292	0,17954	0,18492
	6	0,00001	0,00042	0,00303	0,01080	0,02599	0,04872	0,07673	0,10625	0,13317	0,15410
	7	0,00000	0,00005	0,00059	0,00283	0,00860	0,01955	0,03630	0,05808	0,08279	0,10763
	8		0,00001	0,00010	0,00063	0,00243	0,00671	0,01469	0,02714	0,04401	0,06428
	9		0,00000	0,00001	0,00012	0,00060	0,00200	0,00516	0,01102	0,02031	0,03333
	10			0,00000	0,00002	0,00013	0,00052	0,00159	0,00393	0,00824	0,01518
	11				0,00000	0,00002	0,00012	0,00044	0,00124	0,00296	0,00613
	12					0,00000	0,00001	0,00011	0,00035	0,00095	0,00222
	13						0,00000	0,00002	0,00009	0,00028	0,00072
	14							0,00000	0,00002	0,00007	0,00021
	15								0,00000	0,00002	0,00006
	16									0,00000	0,00001

TABLE 2-1

LOI BINOMIALE

Probabilités cumulées $\Pr(k) = \sum_{i \in 1, k} C_n^k p^i (1-p)^{n-i}$

n	p k	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09	0,10
5	0	0,95099	0,90392	0,85873	0,81537	0,77378	0,73390	0,69569	0,65908	0,62403	0,59049
	1	0,99902	0,99616	0,99153	0,98524	0,97741	0,96813	0,95751	0,94564	0,93262	0,91854
	2	0,99999	0,99992	0,99974	0,99940	0,99884	0,99803	0,99692	0,99547	0,99366	0,99144
	3	1,00000	1,00000	1,00000	0,99999	0,99997	0,99994	0,99989	0,99981	0,99970	0,99954
	4				1,00000	1,00000	1,00000	1,00000	1,00000	0,99999	0,99999
10	0	0,90438	0,81707	0,73742	0,66483	0,59874	0,53862	0,48398	0,43439	0,38942	0,34868
	1	0,99573	0,98382	0,96549	0,94185	0,91386	0,88241	0,84827	0,81212	0,77455	0,73610
	2	0,99989	0,99914	0,99724	0,99379	0,98850	0,98116	0,97166	0,95992	0,94596	0,92981
	3	1,00000	0,99997	0,99985	0,99956	0,99897	0,99797	0,99642	0,99420	0,99117	0,98720
	4		1,00000	0,99999	0,99998	0,99994	0,99985	0,99969	0,99941	0,99899	0,99837
	5			1,00000	1,00000	1,00000	0,99999	0,99998	0,99996	0,99992	0,99985
	6						1,00000	1,00000	1,00000	1,00000	0,99999
15	0	0,86006	0,73857	0,63325	0,54209	0,46329	0,39529	0,33670	0,28630	0,24301	0,20589
	1	0,99037	0,96466	0,92703	0,88089	0,82905	0,77376	0,71685	0,65973	0,60351	0,54904
	2	0,99958	0,99696	0,99063	0,97971	0,96380	0,94287	0,91714	0,88703	0,85310	0,81594
	3	0,99999	0,99982	0,99915	0,99755	0,99453	0,98964	0,98247	0,97269	0,96006	0,94444
	4	1,00000	0,99999	0,99994	0,99978	0,99939	0,99860	0,99722	0,99503	0,99180	0,98728
	5		1,00000	1,00000	0,99999	0,99995	0,99985	0,99966	0,99930	0,99870	0,99775
	6				1,00000	1,00000	0,99999	0,99997	0,99992	0,99984	0,99969
	7						1,00000	1,00000	0,99999	0,99998	0,99997
20	0	0,81791	0,66761	0,54379	0,44200	0,35849	0,29011	0,23424	0,18869	0,15164	0,12158
	1	0,98314	0,94010	0,88016	0,81034	0,73584	0,66045	0,58686	0,51686	0,45160	0,39175
	2	0,99900	0,99293	0,97899	0,95614	0,92452	0,88503	0,83900	0,78795	0,73343	0,67693
	3	0,99996	0,99940	0,99733	0,99259	0,98410	0,97103	0,95287	0,92938	0,90067	0,86705
	4	1,00000	0,99996	0,99974	0,99904	0,99743	0,99437	0,98929	0,98166	0,97096	0,95683
	5		1,00000	0,99998	0,99990	0,99967	0,99913	0,99807	0,99620	0,99321	0,98875
	6			1,00000	0,99999	0,99997	0,99989	0,99972	0,99936	0,99871	0,99761
	7				1,00000	1,00000	0,99999	0,99997	0,99991	0,99980	0,99958
	8						1,00000	1,00000	0,99999	0,99997	0,99994
	9							1,00000	1,00000	0,99999	0,99999
30	0	0,73970	0,54548	0,40101	0,29386	0,21464	0,15626	0,11337	0,08197	0,05905	0,04239
	1	0,96385	0,87945	0,77308	0,66118	0,55354	0,45547	0,36936	0,29579	0,23427	0,18370
	2	0,99668	0,97828	0,93993	0,88310	0,81218	0,73240	0,64875	0,56540	0,48553	0,41135
	3	0,99978	0,99711	0,98810	0,96941	0,93923	0,89738	0,84502	0,78421	0,71747	0,64744
	4	0,99999	0,99970	0,99815	0,99368	0,98436	0,96846	0,94474	0,91264	0,87231	0,82451
	5	1,00000	0,99997	0,99977	0,99894	0,99672	0,99205	0,98377	0,97071	0,95194	0,92681
	6		1,00000	0,99998	0,99985	0,99943	0,99833	0,99601	0,99175	0,98475	0,97417
	7			1,00000	0,99998	0,99992	0,99970	0,99917	0,99803	0,99588	0,99222
	8				1,00000	0,99999	0,99995	0,99985	0,99959	0,99904	0,99798
	9					1,00000	0,99999	0,99998	0,99993	0,99981	0,99955
	10						1,00000	1,00000	0,99999	0,99997	0,99991
	11							1,00000	1,00000	0,99999	0,99998
	12								1,00000	1,00000	1,00000

TABLE 2-2			LOI BINOMIALE								
Probabilités cumulées $\Pr(k) = \sum_{[i \in 1,k]} C_n^k p^i (1-p)^{n-i}$											
n	p	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09	0,10
	k										
40	0	0,66897	0,44570	0,29571	0,19537	0,12851	0,08416	0,05487	0,03561	0,02300	0,01478
	1	0,93926	0,80954	0,66154	0,52098	0,39906	0,29904	0,22006	0,15945	0,11397	0,08047
	2	0,99250	0,95433	0,88217	0,78553	0,67674	0,56650	0,46252	0,36945	0,28942	0,22281
	3	0,99931	0,99176	0,96860	0,92516	0,86185	0,78274	0,69369	0,60075	0,50921	0,42313
	4	0,99995	0,99882	0,99333	0,97898	0,95197	0,91042	0,85463	0,78679	0,71029	0,62902
	5	1,00000	0,99986	0,99884	0,99512	0,98612	0,96909	0,94185	0,90327	0,85347	0,79373
	6		0,99999	0,99983	0,99905	0,99661	0,99094	0,98015	0,96236	0,93608	0,90048
	7		1,00000	0,99998	0,99984	0,99929	0,99772	0,99415	0,98731	0,97576	0,95810
	8			1,00000	0,99998	0,99987	0,99950	0,99850	0,99626	0,99195	0,98450
	9				1,00000	0,99998	0,99990	0,99966	0,99903	0,99764	0,99494
	10					1,00000	0,99998	0,99993	0,99978	0,99939	0,99853
	11						1,00000	0,99999	0,99995	0,99986	0,99962
	12							1,00000	0,99999	0,99997	0,99991
	13								1,00000	0,99999	0,99998
	14									1,00000	1,00000
50	0	0,60501	0,36417	0,21807	0,12989	0,07694	0,04533	0,02656	0,01547	0,00896	0,00515
	1	0,91056	0,73577	0,55528	0,40048	0,27943	0,19000	0,12649	0,08271	0,05324	0,03379
	2	0,98618	0,92157	0,81080	0,67671	0,54053	0,41625	0,31079	0,22597	0,16054	0,11173
	3	0,99840	0,98224	0,93724	0,86087	0,76041	0,64730	0,53274	0,42530	0,33034	0,25029
	4	0,99985	0,99679	0,98319	0,95103	0,89638	0,82060	0,72903	0,62895	0,52766	0,43120
	5	0,99999	0,99952	0,99626	0,98559	0,96222	0,92236	0,86495	0,79187	0,70719	0,61612
	6	1,00000	0,99994	0,99930	0,99639	0,98821	0,97108	0,94169	0,89813	0,84037	0,77023
	7		0,99999	0,99989	0,99922	0,99681	0,99062	0,97799	0,95621	0,92316	0,87785
	8		1,00000	0,99998	0,99985	0,99924	0,99733	0,99268	0,98335	0,96717	0,94213
	9			1,00000	0,99998	0,99984	0,99933	0,99784	0,99437	0,98748	0,97546
	10				1,00000	0,99997	0,99985	0,99943	0,99829	0,99572	0,99065
	11					1,00000	0,99997	0,99986	0,99953	0,99868	0,99678
	12						0,99999	0,99997	0,99989	0,99963	0,99900
	13						1,00000	0,99999	0,99997	0,99991	0,99971
	14							1,00000	0,99999	0,99998	0,99993
	15								1,00000	1,00000	0,99998
	16										1,00000

TABLE 3-1		LOI DE POISSON								
Probabilités individuelles $\Pr(k) = \exp(-\lambda) \lambda^k / k!$										
k	λ	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9
0		0,90484	0,81873	0,74082	0,67032	0,60653	0,54881	0,49659	0,44933	0,40657
1		0,09048	0,16375	0,22225	0,26813	0,30327	0,32929	0,34761	0,35946	0,36591
2		0,00452	0,01637	0,03334	0,05363	0,07582	0,09879	0,12166	0,14379	0,16466
3		0,00015	0,00109	0,00333	0,00715	0,01264	0,01976	0,02839	0,03834	0,04940
4		0,00000	0,00005	0,00025	0,00072	0,00158	0,00296	0,00497	0,00767	0,01111
5			0,00000	0,00002	0,00006	0,00016	0,00036	0,00070	0,00123	0,00200
6				0,00000	0,00000	0,00001	0,00004	0,00008	0,00016	0,00030
7						0,00000	0,00000	0,00001	0,00002	0,00004
k	λ	1	1,5	2	2,5	3	3,5	4	4,5	5
0		0,36788	0,22313	0,13534	0,08208	0,04979	0,03020	0,01832	0,01111	0,00674
1		0,36788	0,33470	0,27067	0,20521	0,14936	0,10569	0,07326	0,04999	0,03369
2		0,18394	0,25102	0,27067	0,25652	0,22404	0,18496	0,14653	0,11248	0,08422
3		0,06131	0,12551	0,18045	0,21376	0,22404	0,21579	0,19537	0,16872	0,14037
4		0,01533	0,04707	0,09022	0,13360	0,16803	0,18881	0,19537	0,18981	0,17547
5		0,00307	0,01412	0,03609	0,06680	0,10082	0,13217	0,15629	0,17083	0,17547
6		0,00051	0,00353	0,01203	0,02783	0,05041	0,07710	0,10420	0,12812	0,14622
7		0,00007	0,00076	0,00344	0,00994	0,02160	0,03855	0,05954	0,08236	0,10444
8		0,00001	0,00014	0,00086	0,00311	0,00810	0,01687	0,02977	0,04633	0,06528
9		0,00000	0,00002	0,00019	0,00086	0,00270	0,00656	0,01323	0,02316	0,03627
10			0,00000	0,00004	0,00022	0,00081	0,00230	0,00529	0,01042	0,01813
11				0,00001	0,00005	0,00022	0,00073	0,00192	0,00426	0,00824
12				0,00000	0,00001	0,00006	0,00021	0,00064	0,00160	0,00343
13					0,00000	0,00001	0,00006	0,00020	0,00055	0,00132
14						0,00000	0,00001	0,00006	0,00018	0,00047
15							0,00000	0,00002	0,00005	0,00016
16								0,00000	0,00002	0,00005
17									0,00000	0,00001
18										0,00000
k	λ	5,5	6	6,5	7	7,5	8	8,5	9	9,5
0		0,00409	0,00248	0,00150	0,00091	0,00055	0,00034	0,00020	0,00012	0,00007
1		0,02248	0,01487	0,00977	0,00638	0,00415	0,00268	0,00173	0,00111	0,00071
2		0,06181	0,04462	0,03176	0,02234	0,01556	0,01073	0,00735	0,00500	0,00338
3		0,11332	0,08924	0,06881	0,05213	0,03889	0,02863	0,02083	0,01499	0,01070
4		0,15582	0,13385	0,11182	0,09123	0,07292	0,05725	0,04425	0,03374	0,02540
5		0,17140	0,16062	0,14537	0,12772	0,10937	0,09160	0,07523	0,06073	0,04827
6		0,15712	0,16062	0,15748	0,14900	0,13672	0,12214	0,10658	0,09109	0,07642
7		0,12345	0,13768	0,14623	0,14900	0,14648	0,13959	0,12942	0,11712	0,10371
8		0,08487	0,10326	0,11882	0,13038	0,13733	0,13959	0,13751	0,13176	0,12316
9		0,05187	0,06884	0,08581	0,10140	0,11444	0,12408	0,12987	0,13176	0,13000
10		0,02853	0,04130	0,05578	0,07098	0,08583	0,09926	0,11039	0,11858	0,12350
11		0,01426	0,02253	0,03296	0,04517	0,05852	0,07219	0,08530	0,09702	0,10666
12		0,00654	0,01126	0,01785	0,02635	0,03658	0,04813	0,06042	0,07277	0,08444
13		0,00277	0,00520	0,00893	0,01419	0,02110	0,02962	0,03951	0,05038	0,06171
14		0,00109	0,00223	0,00414	0,00709	0,01130	0,01692	0,02399	0,03238	0,04187
15		0,00040	0,00089	0,00180	0,00331	0,00565	0,00903	0,01359	0,01943	0,02652
16		0,00014	0,00033	0,00073	0,00145	0,00265	0,00451	0,00722	0,01093	0,01575
17		0,00004	0,00012	0,00028	0,00060	0,00117	0,00212	0,00361	0,00579	0,00880
18		0,00001	0,00004	0,00010	0,00023	0,00049	0,00094	0,00170	0,00289	0,00464
19		0,00000	0,00001	0,00003	0,00009	0,00019	0,00040	0,00076	0,00137	0,00232
20			0,00000	0,00001	0,00003	0,00007	0,00016	0,00032	0,00062	0,00110
21				0,00000	0,00001	0,00003	0,00006	0,00013	0,00026	0,00050
22					0,00000	0,00001	0,00002	0,00005	0,00011	0,00022
23						0,00000	0,00001	0,00002	0,00004	0,00009
24							0,00000	0,00001	0,00002	0,00004
25								0,00000	0,00001	0,00001
26									0,00000	0,00000

TABLE 3-2		LOI DE POISSON								
		Probabilités individuelles $\Pr(k) = \exp(-\lambda) \lambda^k / k!$								
k	λ	10	11	12	13	14	15	16	17	18
0		0,00005	0,00002	0,00001	0,00000	0,00000				
1		0,00045	0,00018	0,00007	0,00003	0,00001	0,00000	0,00000	0,00000	
2		0,00227	0,00101	0,00044	0,00019	0,00008	0,00003	0,00001	0,00001	0,00000
3		0,00757	0,00370	0,00177	0,00083	0,00038	0,00017	0,00008	0,00003	0,00001
4		0,01892	0,01019	0,00531	0,00269	0,00133	0,00065	0,00031	0,00014	0,00007
5		0,03783	0,02242	0,01274	0,00699	0,00373	0,00194	0,00098	0,00049	0,00024
6		0,06306	0,04109	0,02548	0,01515	0,00870	0,00484	0,00262	0,00139	0,00072
7		0,09008	0,06458	0,04368	0,02814	0,01739	0,01037	0,00599	0,00337	0,00185
8		0,11260	0,08879	0,06552	0,04573	0,03044	0,01944	0,01199	0,00716	0,00416
9		0,12511	0,10853	0,08736	0,06605	0,04734	0,03241	0,02131	0,01353	0,00833
10		0,12511	0,11938	0,10484	0,08587	0,06628	0,04861	0,03410	0,02300	0,01499
11		0,11374	0,11938	0,11437	0,10148	0,08436	0,06629	0,04960	0,03554	0,02452
12		0,09478	0,10943	0,11437	0,10994	0,09842	0,08286	0,06613	0,05036	0,03678
13		0,07291	0,09259	0,10557	0,10994	0,10599	0,09561	0,08139	0,06585	0,05093
14		0,05208	0,07275	0,09049	0,10209	0,10599	0,10244	0,09302	0,07996	0,06548
15		0,03472	0,05335	0,07239	0,08848	0,09892	0,10244	0,09922	0,09062	0,07858
16		0,02170	0,03668	0,05429	0,07189	0,08656	0,09603	0,09922	0,09628	0,08840
17		0,01276	0,02373	0,03832	0,05497	0,07128	0,08474	0,09338	0,09628	0,09360
18		0,00709	0,01450	0,02555	0,03970	0,05544	0,07061	0,08301	0,09094	0,09360
19		0,00373	0,00840	0,01614	0,02716	0,04085	0,05575	0,06990	0,08136	0,08867
20		0,00187	0,00462	0,00968	0,01766	0,02860	0,04181	0,05592	0,06916	0,07980
21		0,00089	0,00242	0,00553	0,01093	0,01906	0,02986	0,04261	0,05599	0,06840
22		0,00040	0,00121	0,00302	0,00646	0,01213	0,02036	0,03099	0,04326	0,05597
23		0,00018	0,00058	0,00157	0,00365	0,00738	0,01328	0,02156	0,03198	0,04380
24		0,00007	0,00027	0,00079	0,00198	0,00431	0,00830	0,01437	0,02265	0,03285
25		0,00003	0,00012	0,00038	0,00103	0,00241	0,00498	0,00920	0,01540	0,02365
26		0,00001	0,00005	0,00017	0,00051	0,00130	0,00287	0,00566	0,01007	0,01637
27	0,00000		0,00002	0,00008	0,00025	0,00067	0,00160	0,00335	0,00634	0,01092
28			0,00001	0,00003	0,00011	0,00034	0,00086	0,00192	0,00385	0,00702
29			0,00000	0,00001	0,00005	0,00016	0,00044	0,00106	0,00226	0,00436
30				0,00001	0,00002	0,00008	0,00022	0,00056	0,00128	0,00261
31				0,00000	0,00001	0,00003	0,00011	0,00029	0,00070	0,00152
32					0,00000	0,00001	0,00005	0,00015	0,00037	0,00085
33						0,00001	0,00002	0,00007	0,00019	0,00047
34						0,00000	0,00001	0,00003	0,00010	0,00025
35							0,00000	0,00002	0,00005	0,00013
36								0,00001	0,00002	0,00006
37								0,00000	0,00001	0,00003
38									0,00000	0,00001
39										0,00001
40										0,00000

TABLE 4-1		LOI DE POISSON								
Probabilités cumulées $\Pr(k) = \sum_{i \leq k} \exp(-\lambda) \lambda^i / i!$										
k	λ	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9
0		0,90484	0,81873	0,74082	0,67032	0,60653	0,54881	0,49659	0,44933	0,40657
1		0,99532	0,98248	0,96306	0,93845	0,90980	0,87810	0,84420	0,80879	0,77248
2		0,99985	0,99885	0,99640	0,99207	0,98561	0,97688	0,96586	0,95258	0,93714
3		1,00000	0,99994	0,99973	0,99922	0,99825	0,99664	0,99425	0,99092	0,98654
4			1,00000	0,99998	0,99994	0,99983	0,99961	0,99921	0,99859	0,99766
5				1,00000	1,00000	0,99999	0,99996	0,99991	0,99982	0,99966
6						1,00000	1,00000	0,99999	0,99998	0,99996
7								1,00000	1,00000	1,00000
k	λ	1	1,5	2	2,5	3	3,5	4	4,5	5
0		0,36788	0,22313	0,13534	0,08208	0,04979	0,03020	0,01832	0,01111	0,00674
1		0,73576	0,55783	0,40601	0,28730	0,19915	0,13589	0,09158	0,06110	0,04043
2		0,91970	0,80885	0,67668	0,54381	0,42319	0,32085	0,23810	0,17358	0,12465
3		0,98101	0,93436	0,85712	0,75758	0,64723	0,53663	0,43347	0,34230	0,26503
4		0,99634	0,98142	0,94735	0,89118	0,81526	0,72544	0,62884	0,53210	0,44049
5		0,99941	0,99554	0,98344	0,95798	0,91608	0,85761	0,78513	0,70293	0,61596
6		0,99992	0,99907	0,99547	0,98581	0,96649	0,93471	0,88933	0,83105	0,76218
7		0,99999	0,99983	0,99890	0,99575	0,98810	0,97326	0,94887	0,91341	0,86663
8		1,00000	0,99997	0,99976	0,99886	0,99620	0,99013	0,97864	0,95974	0,93191
9			1,00000	0,99995	0,99972	0,99890	0,99669	0,99187	0,98291	0,96817
10				0,99999	0,99994	0,99971	0,99898	0,99716	0,99333	0,98630
11				1,00000	0,99999	0,99993	0,99971	0,99908	0,99760	0,99545
12					1,00000	0,99998	0,99992	0,99973	0,99919	0,99798
13						1,00000	0,99998	0,99992	0,99975	0,99930
14							1,00000	0,99998	0,99993	0,99977
15								1,00000	0,99998	0,99993
16									0,99999	0,99998
17									1,00000	0,99999
18										1,00000
k	λ	5,5	6	6,5	7	7,5	8	8,5	9	9,5
0		0,00409	0,00248	0,00150	0,00091	0,00055	0,00034	0,00020	0,00012	0,00007
1		0,02656	0,01735	0,01128	0,00730	0,00470	0,00302	0,00193	0,00123	0,00079
2		0,08838	0,06197	0,04304	0,02964	0,02026	0,01375	0,00928	0,00623	0,00416
3		0,20170	0,15120	0,11185	0,08177	0,05915	0,04238	0,03011	0,02123	0,01486
4		0,35752	0,28506	0,22367	0,17299	0,13206	0,09963	0,07436	0,05496	0,04026
5		0,52892	0,44568	0,36904	0,30071	0,24144	0,19124	0,14960	0,11569	0,08853
6		0,68604	0,60630	0,52652	0,44971	0,37815	0,31337	0,25618	0,20678	0,16495
7		0,80949	0,74398	0,67276	0,59871	0,52464	0,45296	0,38560	0,32390	0,26866
8		0,89436	0,84724	0,79157	0,72909	0,66197	0,59255	0,52311	0,45565	0,39182
9		0,94622	0,91608	0,87738	0,83050	0,77641	0,71662	0,65297	0,58741	0,52183
10		0,97475	0,95738	0,93316	0,90148	0,86224	0,81589	0,76336	0,70599	0,64533
11		0,98901	0,97991	0,96612	0,94665	0,92076	0,88808	0,84866	0,80301	0,75199
12		0,99555	0,99117	0,98397	0,97300	0,95733	0,93620	0,90908	0,87577	0,83643
13		0,99831	0,99637	0,99290	0,98719	0,97844	0,96582	0,94859	0,92615	0,89814
14		0,99940	0,99860	0,99704	0,99428	0,98974	0,98274	0,97257	0,95853	0,94001
15		0,99980	0,99949	0,99884	0,99759	0,99539	0,99177	0,98617	0,97796	0,96653
16		0,99994	0,99983	0,99957	0,99904	0,99804	0,99628	0,99339	0,98889	0,98227
17		0,99998	0,99994	0,99985	0,99964	0,99921	0,99841	0,99700	0,99468	0,99107
18		0,99999	0,99998	0,99995	0,99987	0,99970	0,99935	0,99870	0,99757	0,99572
19		1,00000	0,99999	0,99998	0,99996	0,99989	0,99975	0,99947	0,99894	0,99804
20			1,00000	1,00000	0,99999	0,99996	0,99991	0,99979	0,99956	0,99914
21					1,00000	0,99999	0,99997	0,99992	0,99983	0,99964
22						1,00000	0,99999	0,99997	0,99993	0,99985
23							1,00000	0,99999	0,99998	0,99994
24								1,00000	0,99999	0,99998
25									1,00000	0,99999
26										1,00000

TABLE 4-2		LOI DE POISSON								
Probabilités cumulées $\Pr(k) = \sum_{i \in [1,k]} \exp(-\lambda) \lambda^i / i!$										
k	λ	10	11	12	13	14	15	16	17	18
0		0,00005	0,00002	0,00001	0,00000	0,00000				
1		0,00050	0,00020	0,00008	0,00003	0,00001	0,00000	0,00000	0,00000	
2		0,00277	0,00121	0,00052	0,00022	0,00009	0,00004	0,00002	0,00001	0,00000
3		0,01034	0,00492	0,00229	0,00105	0,00047	0,00021	0,00009	0,00004	0,00002
4		0,02925	0,01510	0,00760	0,00374	0,00181	0,00086	0,00040	0,00018	0,00008
5		0,06709	0,03752	0,02034	0,01073	0,00553	0,00279	0,00138	0,00067	0,00032
6		0,13014	0,07861	0,04582	0,02589	0,01423	0,00763	0,00401	0,00206	0,00104
7		0,22022	0,14319	0,08950	0,05403	0,03162	0,01800	0,01000	0,00543	0,00289
8		0,33282	0,23199	0,15503	0,09976	0,06206	0,03745	0,02199	0,01260	0,00706
9		0,45793	0,34051	0,24239	0,16581	0,10940	0,06985	0,04330	0,02612	0,01538
10		0,58304	0,45989	0,34723	0,25168	0,17568	0,11846	0,07740	0,04912	0,03037
11		0,69678	0,57927	0,46160	0,35316	0,26004	0,18475	0,12699	0,08467	0,05489
12		0,79156	0,68870	0,57597	0,46310	0,35846	0,26761	0,19312	0,13502	0,09167
13		0,86446	0,78129	0,68154	0,57304	0,46445	0,36322	0,27451	0,20087	0,14260
14		0,91654	0,85404	0,77202	0,67513	0,57044	0,46565	0,36753	0,28083	0,20808
15		0,95126	0,90740	0,84442	0,76361	0,66936	0,56809	0,46674	0,37145	0,28665
16		0,97296	0,94408	0,89871	0,83549	0,75592	0,66412	0,56596	0,46774	0,37505
17		0,98572	0,96781	0,93703	0,89046	0,82720	0,74886	0,65934	0,56402	0,46865
18		0,99281	0,98231	0,96258	0,93017	0,88264	0,81947	0,74235	0,65496	0,56224
19		0,99655	0,99071	0,97872	0,95733	0,92350	0,87522	0,81225	0,73632	0,65092
20		0,99841	0,99533	0,98840	0,97499	0,95209	0,91703	0,86817	0,80548	0,73072
21		0,99930	0,99775	0,99393	0,98592	0,97116	0,94689	0,91077	0,86147	0,79912
22		0,99970	0,99896	0,99695	0,99238	0,98329	0,96726	0,94176	0,90473	0,85509
23		0,99988	0,99954	0,99853	0,99603	0,99067	0,98054	0,96331	0,93670	0,89889
24		0,99995	0,99980	0,99931	0,99801	0,99498	0,98884	0,97768	0,95935	0,93174
25		0,99998	0,99992	0,99969	0,99903	0,99739	0,99382	0,98688	0,97476	0,95539
26		0,99999	0,99997	0,99987	0,99955	0,99869	0,99669	0,99254	0,98483	0,97177
27	1,00000		0,99999	0,99994	0,99980	0,99936	0,99828	0,99589	0,99117	0,98268
28			1,00000	0,99998	0,99991	0,99970	0,99914	0,99781	0,99502	0,98970
29				0,99999	0,99996	0,99986	0,99958	0,99887	0,99727	0,99406
30				1,00000	0,99998	0,99994	0,99980	0,99943	0,99855	0,99667
31					0,99999	0,99997	0,99991	0,99972	0,99925	0,99819
32					1,00000	0,99999	0,99996	0,99987	0,99963	0,99904
33						0,99998	0,99994	0,99982	0,99951	
34						1,00000	0,99999	0,99997	0,99991	0,99975
35							1,00000	0,99999	0,99996	0,99988
36								1,00000	0,99998	0,99994
37									0,99999	0,99997
38									1,00000	0,99999
39										0,99999
40										1,00000

TABLE 5-1		FONCTION de REPARTITION de la L O I N O R M A L E R E D U I T E								
u	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,00	0,500000	0,503989	0,507978	0,511967	0,515953	0,519939	0,523922	0,527903	0,531881	0,535856
0,10	0,539828	0,543795	0,547758	0,551717	0,555670	0,559618	0,563559	0,567495	0,571424	0,575345
0,20	0,579260	0,583166	0,587064	0,590954	0,594835	0,598706	0,602568	0,606420	0,610261	0,614092
0,30	0,617911	0,621719	0,625516	0,629300	0,633072	0,636831	0,640576	0,644309	0,648027	0,651732
0,40	0,655422	0,659097	0,662757	0,666402	0,670031	0,673645	0,677242	0,680822	0,684386	0,687933
0,50	0,691462	0,694974	0,698468	0,701944	0,705402	0,708840	0,712260	0,715661	0,719043	0,722405
0,60	0,725747	0,729069	0,732371	0,735653	0,738914	0,742154	0,745373	0,748571	0,751748	0,754903
0,70	0,758036	0,761148	0,764238	0,767305	0,770350	0,773373	0,776373	0,779350	0,782305	0,785236
0,80	0,788145	0,791030	0,793892	0,796731	0,799546	0,802338	0,805106	0,807850	0,810570	0,813267
0,90	0,815940	0,818589	0,821214	0,823814	0,826391	0,828944	0,831472	0,833977	0,836457	0,838913
1,00	0,841345	0,843752	0,846136	0,848495	0,850830	0,853141	0,855428	0,857690	0,859929	0,862143
1,10	0,864334	0,866500	0,868643	0,870762	0,872857	0,874928	0,876976	0,878999	0,881000	0,882977
1,20	0,884930	0,886860	0,888767	0,890651	0,892512	0,894350	0,896165	0,897958	0,899727	0,901475
1,30	0,903199	0,904902	0,906582	0,908241	0,909877	0,911492	0,913085	0,914656	0,916207	0,917736
1,40	0,919243	0,920730	0,922196	0,923641	0,925066	0,926471	0,927855	0,929219	0,930563	0,931888
1,50	0,933193	0,934478	0,935744	0,936992	0,938220	0,939429	0,940620	0,941792	0,942947	0,944083
1,60	0,945201	0,946301	0,947384	0,948449	0,949497	0,950529	0,951543	0,952540	0,953521	0,954486
1,70	0,955435	0,956367	0,957284	0,958185	0,959071	0,959941	0,960796	0,961636	0,962462	0,963273
1,80	0,964070	0,964852	0,965621	0,966375	0,967116	0,967843	0,968557	0,969258	0,969946	0,970621
1,90	0,971284	0,971933	0,972571	0,973197	0,973810	0,974412	0,975002	0,975581	0,976148	0,976705
2,00	0,977250	0,977784	0,978308	0,978822	0,979325	0,979818	0,980301	0,980774	0,981237	0,981691
2,10	0,982136	0,982571	0,982997	0,983414	0,983823	0,984222	0,984614	0,984997	0,985371	0,985738
2,20	0,986097	0,986447	0,986791	0,987126	0,987455	0,987776	0,988089	0,988396	0,988696	0,988989
2,30	0,989276	0,989556	0,989830	0,990097	0,990358	0,990613	0,990863	0,991106	0,991344	0,991576
2,40	0,991802	0,992024	0,992240	0,992451	0,992656	0,992857	0,993053	0,993244	0,993431	0,993613
2,50	0,993790	0,993963	0,994132	0,994297	0,994457	0,994614	0,994766	0,994915	0,995060	0,995201
2,60	0,995339	0,995473	0,995603	0,995731	0,995855	0,995975	0,996093	0,996207	0,996319	0,996427
2,70	0,996533	0,996636	0,996736	0,996833	0,996928	0,997020	0,997110	0,997197	0,997282	0,997365
2,80	0,997445	0,997523	0,997599	0,997673	0,997744	0,997814	0,997882	0,997948	0,998012	0,998074
2,90	0,998134	0,998193	0,998250	0,998305	0,998359	0,998411	0,998462	0,998511	0,998559	0,998605
3,00	0,998650	0,998694	0,998736	0,998777	0,998817	0,998856	0,998893	0,998930	0,998965	0,998999
3,10	0,999032	0,999064	0,999096	0,999126	0,999155	0,999184	0,999211	0,999238	0,999264	0,999289
3,20	0,999313	0,999336	0,999359	0,999381	0,999402	0,999423	0,999443	0,999462	0,999481	0,999499
3,30	0,999517	0,999533	0,999550	0,999566	0,999581	0,999596	0,999610	0,999624	0,999638	0,999650
3,40	0,999663	0,999675	0,999687	0,999698	0,999709	0,999720	0,999730	0,999740	0,999749	0,999758
3,50	0,999767	0,999776	0,999784	0,999792	0,999800	0,999807	0,999815	0,999821	0,999828	0,999835
3,60	0,999841	0,999847	0,999853	0,999858	0,999864	0,999869	0,999874	0,999879	0,999883	0,999888
3,70	0,999892	0,999896	0,999900	0,999904	0,999908	0,999912	0,999915	0,999918	0,999922	0,999925
3,80	0,999928	0,999930	0,999933	0,999936	0,999938	0,999941	0,999943	0,999946	0,999948	0,999950
3,90	0,999952	0,999954	0,999956	0,999958	0,999959	0,999961	0,999963	0,999964	0,999966	0,999967

TABLE 5-2		FRACTILE de la LOI NORMALE REDUITE											
P	0,000	0,001	0,002	0,003	0,004	0,005	0,006	0,007	0,008	0,009	0,010		
0,00	∞	3,0902	2,8782	2,7478	2,6521	2,5758	2,5121	2,4573	2,4089	2,3656	2,3263		0,99
0,01	3,2363	2,2904	2,2571	2,2262	2,1973	2,1701	2,1444	2,1201	2,0969	2,0748	2,0537		0,98
0,02	2,0537	2,0335	2,0141	1,9954	1,9774	1,9600	1,9431	1,9268	1,9110	1,8957	1,8808		0,97
0,03	1,8808	1,8663	1,8522	1,8384	1,8250	1,8119	1,7991	1,7866	1,7744	1,7624	1,7507		0,96
0,04	1,7507	1,7392	1,7279	1,7169	1,7060	1,6954	1,6849	1,6747	1,6646	1,6546	1,6449		0,95
0,05	1,6449	1,6352	1,6258	1,6164	1,6072	1,5982	1,5893	1,5805	1,5718	1,5632	1,5548		0,94
0,06	1,5548	1,5464	1,5382	1,5301	1,5220	1,5141	1,5063	1,4985	1,4909	1,4833	1,4758		0,93
0,07	1,4758	1,4684	1,4611	1,4538	1,4466	1,4395	1,4325	1,4255	1,4187	1,4118	1,4051		0,92
0,08	1,4051	1,3984	1,3917	1,3852	1,3787	1,3722	1,3658	1,3595	1,3532	1,3469	1,3408		0,91
0,09	1,3408	1,3346	1,3285	1,3225	1,3165	1,3106	1,3047	1,2988	1,2930	1,2873	1,2816		0,90
0,10	1,2816	1,2759	1,2702	1,2646	1,2591	1,2536	1,2481	1,2426	1,2372	1,2319	1,2265		0,89
0,11	1,2265	1,2212	1,2160	1,2107	1,2055	1,2004	1,1952	1,1901	1,1850	1,1800	1,1750		0,88
0,12	1,1750	1,1700	1,1650	1,1601	1,1552	1,1503	1,1455	1,1407	1,1359	1,1311	1,1264		0,87
0,13	1,1264	1,1217	1,1170	1,1123	1,1077	1,1031	1,0985	1,0939	1,0893	1,0848	1,0803		0,86
0,14	1,0803	1,0758	1,0714	1,0669	1,0625	1,0581	1,0537	1,0494	1,0451	1,0407	1,0364		0,85
0,15	1,0364	1,0322	1,0279	1,0237	1,0194	1,0152	1,0110	1,0069	1,0027	0,9986	0,9945		0,84
0,16	0,9945	0,9904	0,9863	0,9822	0,9782	0,9741	0,9701	0,9661	0,9621	0,9581	0,9542		0,83
0,17	0,9542	0,9502	0,9463	0,9424	0,9385	0,9346	0,9307	0,9269	0,9230	0,9192	0,9154		0,82
0,18	0,9154	0,9116	0,9078	0,9040	0,9002	0,8965	0,8927	0,8890	0,8853	0,8816	0,8779		0,81
0,19	0,8779	0,8742	0,8706	0,8669	0,8632	0,8596	0,8560	0,8524	0,8488	0,8452	0,8416		0,80
0,20	0,8416	0,8381	0,8345	0,8310	0,8274	0,8239	0,8204	0,8169	0,8134	0,8099	0,8064		0,79
0,21	0,8064	0,8030	0,7995	0,7961	0,7926	0,7892	0,7858	0,7824	0,7790	0,7756	0,7722		0,78
0,22	0,7722	0,7688	0,7655	0,7621	0,7588	0,7554	0,7521	0,7488	0,7454	0,7421	0,7388		0,77
0,23	0,7388	0,7356	0,7323	0,7290	0,7257	0,7225	0,7192	0,7160	0,7128	0,7095	0,7063		0,76
0,24	0,7063	0,7031	0,6999	0,6967	0,6935	0,6903	0,6871	0,6840	0,6808	0,6776	0,6745		0,75
0,25	0,6745	0,6713	0,6682	0,6651	0,6620	0,6588	0,6557	0,6526	0,6495	0,6464	0,6433		0,74
0,26	0,6433	0,6403	0,6372	0,6341	0,6311	0,6280	0,6250	0,6219	0,6189	0,6158	0,6128		0,73
0,27	0,6128	0,6098	0,6068	0,6038	0,6008	0,5978	0,5948	0,5918	0,5888	0,5858	0,5828		0,72
0,28	0,5828	0,5799	0,5769	0,5740	0,5710	0,5681	0,5651	0,5622	0,5592	0,5563	0,5534		0,71
0,29	0,5534	0,5505	0,5476	0,5446	0,5417	0,5388	0,5359	0,5330	0,5302	0,5273	0,5244		0,70
0,30	0,5244	0,5215	0,5187	0,5158	0,5129	0,5101	0,5072	0,5044	0,5015	0,4987	0,4958		0,69
0,31	0,4958	0,4930	0,4902	0,4874	0,4845	0,4817	0,4789	0,4761	0,4733	0,4705	0,4677		0,68
0,32	0,4677	0,4649	0,4621	0,4593	0,4565	0,4538	0,4510	0,4482	0,4454	0,4427	0,4399		0,67
0,33	0,4399	0,4372	0,4344	0,4316	0,4289	0,4261	0,4234	0,4207	0,4179	0,4152	0,4125		0,66
0,34	0,4125	0,4097	0,4070	0,4043	0,4016	0,3989	0,3961	0,3934	0,3907	0,3880	0,3853		0,65
0,35	0,3853	0,3826	0,3799	0,3772	0,3745	0,3719	0,3692	0,3665	0,3638	0,3611	0,3585		0,64
0,36	0,3585	0,3558	0,3531	0,3505	0,3478	0,3451	0,3425	0,3398	0,3372	0,3345	0,3319		0,63
0,37	0,3319	0,3292	0,3266	0,3239	0,3213	0,3186	0,3160	0,3134	0,3107	0,3081	0,3055		0,62
0,38	0,3055	0,3029	0,3002	0,2976	0,2950	0,2924	0,2898	0,2871	0,2845	0,2819	0,2793		0,61
0,39	0,2793	0,2767	0,2741	0,2715	0,2689	0,2663	0,2637	0,2611	0,2585	0,2559	0,2533		0,60
0,40	0,2533	0,2508	0,2482	0,2456	0,2430	0,2404	0,2378	0,2353	0,2327	0,2301	0,2275		0,59
0,41	0,2275	0,2250	0,2224	0,2198	0,2173	0,2147	0,2121	0,2096	0,2070	0,2045	0,2019		0,58
0,42	0,2019	0,1993	0,1968	0,1942	0,1917	0,1891	0,1866	0,1840	0,1815	0,1789	0,1764		0,57
0,43	0,1764	0,1738	0,1713	0,1687	0,1662	0,1637	0,1611	0,1586	0,1560	0,1535	0,1510		0,56
0,44	0,1510	0,1484	0,1459	0,1434	0,1408	0,1383	0,1358	0,1332	0,1307	0,1282	0,1257		0,55
0,45	0,1257	0,1231	0,1206	0,1181	0,1156	0,1130	0,1105	0,1080	0,1055	0,1030	0,1004		0,54
0,46	0,1004	0,0979	0,0954	0,0929	0,0904	0,0878	0,0853	0,0828	0,0803	0,0778	0,0753		0,53
0,47	0,0753	0,0728	0,0702	0,0677	0,0652	0,0627	0,0602	0,0577	0,0552	0,0527	0,0502		0,52
0,48	0,0502	0,0476	0,0451	0,0426	0,0401	0,0376	0,0351	0,0326	0,0301	0,0276	0,0251		0,51
0,49	0,0251	0,0226	0,0201	0,0175	0,0150	0,0125	0,0100	0,0075	0,0050	0,0025	0,0000		0,50
	0,010	0,009	0,008	0,007	0,006	0,005	0,004	0,003	0,002	0,001	0,000		P

TABLE 6			FRACTILES de la LOI χ^2 (v)											
α n	0,001	0,005	0,01	0,025	0,05	0,1	0,5	0,9	0,95	0,975	0,99	0,995	0,999	
1	0,00	0,00	0,00	0,00	0,00	0,02	0,45	2,71	3,84	5,02	6,63	7,88	10,83	
2	0,00	0,01	0,02	0,05	0,10	0,21	1,39	4,61	5,99	7,38	9,21	10,60	13,82	
3	0,02	0,07	0,11	0,22	0,35	0,58	2,37	6,25	7,81	9,35	11,34	12,84	16,27	
4	0,09	0,21	0,30	0,48	0,71	1,06	3,36	7,78	9,49	11,14	13,28	14,86	18,47	
5	0,21	0,41	0,55	0,83	1,15	1,61	4,35	9,24	11,07	12,83	15,09	16,75	20,51	
6	0,38	0,68	0,87	1,24	1,64	2,20	5,35	10,64	12,59	14,45	16,81	18,55	22,46	
7	0,60	0,99	1,24	1,69	2,17	2,83	6,35	12,02	14,07	16,01	18,48	20,28	24,32	
8	0,86	1,34	1,65	2,18	2,73	3,49	7,34	13,36	15,51	17,53	20,09	21,95	26,12	
9	1,15	1,73	2,09	2,70	3,33	4,17	8,34	14,68	16,92	19,02	21,67	23,59	27,88	
10	1,48	2,16	2,56	3,25	3,94	4,87	9,34	15,99	18,31	20,48	23,21	25,19	29,59	
11	1,83	2,60	3,05	3,82	4,57	5,58	10,34	17,28	19,68	21,92	24,73	26,76	31,26	
12	2,21	3,07	3,57	4,40	5,23	6,30	11,34	18,55	21,03	23,34	26,22	28,30	32,91	
13	2,62	3,57	4,11	5,01	5,89	7,04	12,34	19,81	22,36	24,74	27,69	29,82	34,53	
14	3,04	4,07	4,66	5,63	6,57	7,79	13,34	21,06	23,68	26,12	29,14	31,32	36,12	
15	3,48	4,60	5,23	6,26	7,26	8,55	14,34	22,31	25,00	27,49	30,58	32,80	37,70	
16	3,94	5,14	5,81	6,91	7,96	9,31	15,34	23,54	26,30	28,85	32,00	34,27	39,25	
17	4,42	5,70	6,41	7,56	8,67	10,09	16,34	24,77	27,59	30,19	33,41	35,72	40,79	
18	4,90	6,26	7,01	8,23	9,39	10,86	17,34	25,99	28,87	31,53	34,81	37,16	42,31	
19	5,41	6,84	7,63	8,91	10,12	11,65	18,34	27,20	30,14	32,85	36,19	38,58	43,82	
20	5,92	7,43	8,26	9,59	10,85	12,44	19,34	28,41	31,41	34,17	37,57	40,00	45,31	
21	6,45	8,03	8,90	10,28	11,59	13,24	20,34	29,62	32,67	35,48	38,93	41,40	46,80	
22	6,98	8,64	9,54	10,98	12,34	14,04	21,34	30,81	33,92	36,78	40,29	42,80	48,27	
23	7,53	9,26	10,20	11,69	13,09	14,85	22,34	32,01	35,17	38,08	41,64	44,18	49,73	
24	8,08	9,89	10,86	12,40	13,85	15,66	23,34	33,20	36,42	39,36	42,98	45,56	51,18	
25	8,65	10,52	11,52	13,12	14,61	16,47	24,34	34,38	37,65	40,65	44,31	46,93	52,62	
26	9,22	11,16	12,20	13,84	15,38	17,29	25,34	35,56	38,89	41,92	45,64	48,29	54,05	
27	9,80	11,81	12,88	14,57	16,15	18,11	26,34	36,74	40,11	43,19	46,96	49,65	55,48	
28	10,39	12,46	13,56	15,31	16,93	18,94	27,34	37,92	41,34	44,46	48,28	50,99	56,89	
29	10,99	13,12	14,26	16,05	17,71	19,77	28,34	39,09	42,56	45,72	49,59	52,34	58,30	
30	11,59	13,79	14,95	16,79	18,49	20,60	29,34	40,26	43,77	46,98	50,89	53,67	59,70	
31	12,20	14,46	15,66	17,54	19,28	21,43	30,34	41,42	44,99	48,23	52,19	55,00	61,10	
32	12,81	15,13	16,36	18,29	20,07	22,27	31,34	42,58	46,19	49,48	53,49	56,33	62,49	
33	13,43	15,82	17,07	19,05	20,87	23,11	32,34	43,75	47,40	50,73	54,78	57,65	63,87	
34	14,06	16,50	17,79	19,81	21,66	23,95	33,34	44,90	48,60	51,97	56,06	58,96	65,25	
35	14,69	17,19	18,51	20,57	22,47	24,80	34,34	46,06	49,80	53,20	57,34	60,27	66,62	
36	15,32	17,89	19,23	21,34	23,27	25,64	35,34	47,21	51,00	54,44	58,62	61,58	67,98	
37	15,97	18,59	19,96	22,11	24,07	26,49	36,34	48,36	52,19	55,67	59,89	62,88	69,35	
38	16,61	19,29	20,69	22,88	24,88	27,34	37,34	49,51	53,38	56,90	61,16	64,18	70,70	
39	17,26	20,00	21,43	23,65	25,70	28,20	38,34	50,66	54,57	58,12	62,43	65,48	72,06	
40	17,92	20,71	22,16	24,43	26,51	29,05	39,34	51,81	55,76	59,34	63,69	66,77	73,40	
50	24,67	27,99	29,71	32,36	34,76	37,69	49,33	63,17	67,50	71,42	76,15	79,49	86,66	
60	31,74	35,53	37,48	40,48	43,19	46,46	59,33	74,40	79,08	83,30	88,38	91,95	99,61	
70	39,04	43,28	45,44	48,76	51,74	55,33	69,33	85,53	90,53	95,02	100,43	104,21	112,32	
80	46,52	51,17	53,54	57,15	60,39	64,28	79,33	96,58	101,88	106,63	112,33	116,32	124,84	
90	54,16	59,20	61,75	65,65	69,13	73,29	89,33	107,57	113,15	118,14	124,12	128,30	137,21	
100	61,92	67,33	70,06	74,22	77,93	82,36	99,33	118,50	124,34	129,56	135,81	140,17	149,45	

TABLE 7-1a		FRACILES de la LOI de FISHER F(v1,v2)										$\alpha= 0,95$	
v ₁	1	2	3	4	5	6	7	8	9	10	12	14	
v ₂													
1	161,45	199,50	215,71	224,58	230,16	233,98	236,77	238,88	240,54	241,88	243,91	245,36	
2	18,513	19,000	19,164	19,247	19,296	19,329	19,353	19,371	19,385	19,396	19,412	19,424	
3	10,128	9,552	9,277	9,117	9,013	8,941	8,887	8,845	8,812	8,786	8,745	8,715	
4	7,709	6,944	6,591	6,388	6,256	6,163	6,094	6,041	5,999	5,964	5,912	5,873	
5	6,608	5,786	5,409	5,192	5,050	4,950	4,876	4,818	4,772	4,735	4,678	4,636	
6	5,987	5,143	4,757	4,534	4,387	4,284	4,207	4,147	4,099	4,060	4,000	3,956	
7	5,591	4,737	4,347	4,120	3,972	3,866	3,787	3,726	3,677	3,637	3,575	3,529	
8	5,318	4,459	4,066	3,838	3,687	3,581	3,500	3,438	3,388	3,347	3,284	3,237	
9	5,117	4,256	3,863	3,633	3,482	3,374	3,293	3,230	3,179	3,137	3,073	3,025	
10	4,965	4,103	3,708	3,478	3,326	3,217	3,135	3,072	3,020	2,978	2,913	2,865	
11	4,844	3,982	3,587	3,357	3,204	3,095	3,012	2,948	2,896	2,854	2,788	2,739	
12	4,747	3,885	3,490	3,259	3,106	2,996	2,913	2,849	2,796	2,753	2,687	2,637	
13	4,667	3,806	3,411	3,179	3,025	2,915	2,832	2,767	2,714	2,671	2,604	2,554	
14	4,600	3,739	3,344	3,112	2,958	2,848	2,764	2,699	2,646	2,602	2,534	2,484	
15	4,543	3,682	3,287	3,056	2,901	2,790	2,707	2,641	2,588	2,544	2,475	2,424	
16	4,494	3,634	3,239	3,007	2,852	2,741	2,657	2,591	2,538	2,494	2,425	2,373	
17	4,451	3,592	3,197	2,965	2,810	2,699	2,614	2,548	2,494	2,450	2,381	2,329	
18	4,414	3,555	3,160	2,928	2,773	2,661	2,577	2,510	2,456	2,412	2,342	2,290	
19	4,381	3,522	3,127	2,895	2,740	2,628	2,544	2,477	2,423	2,378	2,308	2,256	
20	4,351	3,493	3,098	2,866	2,711	2,599	2,514	2,447	2,393	2,348	2,278	2,225	
21	4,325	3,467	3,072	2,840	2,685	2,573	2,488	2,420	2,366	2,321	2,250	2,197	
22	4,301	3,443	3,049	2,817	2,661	2,549	2,464	2,397	2,342	2,297	2,226	2,173	
23	4,279	3,422	3,028	2,796	2,640	2,528	2,442	2,375	2,320	2,275	2,204	2,150	
24	4,260	3,403	3,009	2,776	2,621	2,508	2,423	2,355	2,300	2,255	2,183	2,130	
25	4,242	3,385	2,991	2,759	2,603	2,490	2,405	2,337	2,282	2,236	2,165	2,111	
26	4,225	3,369	2,975	2,743	2,587	2,474	2,388	2,321	2,265	2,220	2,148	2,094	
27	4,210	3,354	2,960	2,728	2,572	2,459	2,373	2,305	2,250	2,204	2,132	2,078	
28	4,196	3,340	2,947	2,714	2,558	2,445	2,359	2,291	2,236	2,190	2,118	2,064	
29	4,183	3,328	2,934	2,701	2,545	2,432	2,346	2,278	2,223	2,177	2,104	2,050	
30	4,171	3,316	2,922	2,690	2,534	2,421	2,334	2,266	2,211	2,165	2,092	2,037	
31	4,160	3,305	2,911	2,679	2,523	2,409	2,323	2,255	2,199	2,153	2,080	2,026	
32	4,149	3,295	2,901	2,668	2,512	2,399	2,313	2,244	2,189	2,142	2,070	2,015	
33	4,139	3,285	2,892	2,659	2,503	2,389	2,303	2,235	2,179	2,133	2,060	2,004	
34	4,130	3,276	2,883	2,650	2,494	2,380	2,294	2,225	2,170	2,123	2,050	1,995	
35	4,121	3,267	2,874	2,641	2,485	2,372	2,285	2,217	2,161	2,114	2,041	1,986	
36	4,113	3,259	2,866	2,634	2,477	2,364	2,277	2,209	2,153	2,106	2,033	1,977	
37	4,105	3,252	2,859	2,626	2,470	2,356	2,270	2,201	2,145	2,098	2,025	1,969	
38	4,098	3,245	2,852	2,619	2,463	2,349	2,262	2,194	2,138	2,091	2,017	1,962	
39	4,091	3,238	2,845	2,612	2,456	2,342	2,255	2,187	2,131	2,084	2,010	1,954	
40	4,085	3,232	2,839	2,606	2,449	2,336	2,249	2,180	2,124	2,077	2,003	1,948	
50	4,034	3,183	2,790	2,557	2,400	2,286	2,199	2,130	2,073	2,026	1,952	1,895	
60	4,001	3,150	2,758	2,525	2,368	2,254	2,167	2,097	2,040	1,993	1,917	1,860	
70	3,978	3,128	2,736	2,503	2,346	2,231	2,143	2,074	2,017	1,969	1,893	1,836	
80	3,960	3,111	2,719	2,486	2,329	2,214	2,126	2,056	1,999	1,951	1,875	1,817	
90	3,947	3,098	2,706	2,473	2,316	2,201	2,113	2,043	1,986	1,938	1,861	1,803	
100	3,936	3,087	2,696	2,463	2,305	2,191	2,103	2,032	1,975	1,927	1,850	1,792	
200	3,888	3,041	2,650	2,417	2,259	2,144	2,056	1,985	1,927	1,878	1,801	1,742	
500	3,860	3,014	2,623	2,390	2,232	2,117	2,028	1,957	1,899	1,850	1,772	1,712	
∞	3,841	2,996	2,605	2,372	2,214	2,099	2,010	1,938	1,880	1,831	1,752	1,692	

TABLE 7-1b		FRACTILES de la LOI de FISHER F(v1,v2)										$\alpha = 0,95$	
V ₁	16	18	20	22	24	26	28	30	40	60	80	1000	
1	246,46	247,32	248,02	248,58	249,05	249,45	249,80	250,10	251,14	252,20	252,72	254,19	
2	19,433	19,440	19,446	19,450	19,454	19,457	19,460	19,462	19,471	19,479	19,483	19,495	
3	8,692	8,675	8,660	8,648	8,639	8,630	8,623	8,617	8,594	8,572	8,561	8,529	
4	5,844	5,821	5,803	5,787	5,774	5,763	5,754	5,746	5,717	5,688	5,673	5,632	
5	4,604	4,579	4,558	4,541	4,527	4,515	4,505	4,496	4,464	4,431	4,415	4,369	
6	3,922	3,896	3,874	3,856	3,841	3,829	3,818	3,808	3,774	3,740	3,722	3,673	
7	3,494	3,467	3,445	3,426	3,410	3,397	3,386	3,376	3,340	3,304	3,286	3,234	
8	3,202	3,173	3,150	3,131	3,115	3,102	3,090	3,079	3,043	3,005	2,986	2,932	
9	2,989	2,960	2,936	2,917	2,900	2,886	2,874	2,864	2,826	2,787	2,768	2,712	
10	2,828	2,798	2,774	2,754	2,737	2,723	2,710	2,700	2,661	2,621	2,601	2,543	
11	2,701	2,671	2,646	2,626	2,609	2,594	2,582	2,570	2,531	2,490	2,469	2,410	
12	2,599	2,568	2,544	2,523	2,505	2,491	2,478	2,466	2,426	2,384	2,363	2,302	
13	2,515	2,484	2,459	2,438	2,420	2,405	2,392	2,380	2,339	2,297	2,275	2,212	
14	2,445	2,413	2,388	2,367	2,349	2,333	2,320	2,308	2,266	2,223	2,201	2,136	
15	2,385	2,353	2,328	2,306	2,288	2,272	2,259	2,247	2,204	2,160	2,137	2,072	
16	2,333	2,302	2,276	2,254	2,235	2,220	2,206	2,194	2,151	2,106	2,083	2,016	
17	2,289	2,257	2,230	2,208	2,190	2,174	2,160	2,148	2,104	2,058	2,035	1,967	
18	2,250	2,217	2,191	2,168	2,150	2,134	2,119	2,107	2,063	2,017	1,993	1,923	
19	2,215	2,182	2,155	2,133	2,114	2,098	2,084	2,071	2,026	1,980	1,955	1,884	
20	2,184	2,151	2,124	2,102	2,082	2,066	2,052	2,039	1,994	1,946	1,922	1,850	
21	2,156	2,123	2,096	2,073	2,054	2,037	2,023	2,010	1,965	1,916	1,891	1,818	
22	2,131	2,098	2,071	2,048	2,028	2,012	1,997	1,984	1,938	1,889	1,864	1,790	
23	2,109	2,075	2,048	2,025	2,005	1,988	1,973	1,961	1,914	1,865	1,839	1,764	
24	2,088	2,054	2,027	2,003	1,984	1,967	1,952	1,939	1,892	1,842	1,816	1,740	
25	2,069	2,035	2,007	1,984	1,964	1,947	1,932	1,919	1,872	1,822	1,796	1,718	
26	2,052	2,018	1,990	1,966	1,946	1,929	1,914	1,901	1,853	1,803	1,776	1,698	
27	2,036	2,002	1,974	1,950	1,930	1,913	1,898	1,884	1,836	1,785	1,758	1,679	
28	2,021	1,987	1,959	1,935	1,915	1,897	1,882	1,869	1,820	1,769	1,742	1,662	
29	2,007	1,973	1,945	1,921	1,901	1,883	1,868	1,854	1,806	1,754	1,726	1,645	
30	1,995	1,960	1,932	1,908	1,887	1,870	1,854	1,841	1,792	1,740	1,712	1,630	
31	1,983	1,948	1,920	1,896	1,875	1,857	1,842	1,828	1,779	1,726	1,699	1,616	
32	1,972	1,937	1,908	1,884	1,864	1,846	1,830	1,817	1,767	1,714	1,686	1,602	
33	1,961	1,926	1,898	1,873	1,853	1,835	1,819	1,806	1,756	1,702	1,674	1,589	
34	1,952	1,917	1,888	1,863	1,843	1,825	1,809	1,795	1,745	1,691	1,663	1,577	
35	1,942	1,907	1,878	1,854	1,833	1,815	1,799	1,786	1,735	1,681	1,652	1,566	
36	1,934	1,899	1,870	1,845	1,824	1,806	1,790	1,776	1,726	1,671	1,643	1,555	
37	1,926	1,890	1,861	1,837	1,816	1,798	1,782	1,768	1,717	1,662	1,633	1,545	
38	1,918	1,883	1,853	1,829	1,808	1,790	1,774	1,760	1,708	1,653	1,624	1,536	
39	1,911	1,875	1,846	1,821	1,800	1,782	1,766	1,752	1,700	1,645	1,616	1,526	
40	1,904	1,868	1,839	1,814	1,793	1,775	1,759	1,744	1,693	1,637	1,608	1,517	
50	1,850	1,814	1,784	1,759	1,737	1,718	1,702	1,687	1,634	1,576	1,544	1,448	
60	1,815	1,778	1,748	1,722	1,700	1,681	1,664	1,649	1,594	1,534	1,502	1,399	
70	1,790	1,753	1,722	1,696	1,674	1,654	1,637	1,622	1,566	1,505	1,471	1,364	
80	1,772	1,734	1,703	1,677	1,654	1,634	1,617	1,602	1,545	1,482	1,448	1,336	
90	1,757	1,720	1,688	1,662	1,639	1,619	1,601	1,586	1,528	1,465	1,429	1,314	
100	1,746	1,708	1,676	1,650	1,627	1,607	1,589	1,573	1,515	1,450	1,415	1,296	
200	1,694	1,656	1,623	1,596	1,572	1,551	1,533	1,516	1,455	1,386	1,346	1,205	
500	1,664	1,625	1,592	1,563	1,539	1,518	1,499	1,482	1,419	1,345	1,303	1,138	
∞	1,644	1,604	1,571	1,542	1,517	1,496	1,476	1,459	1,394	1,318	1,274	1,075	

TABLE 7-2a		FRACILES de la LOI de FISHER F(v ₁ ,v ₂)										$\alpha = 0,975$	
v ₂	v ₁	1	2	3	4	5	6	7	8	9	10	12	14
1	647,79	799,50	864,17	899,60	921,83	937,13	948,23	956,64	963,29	968,62	976,71	982,51	
2	38,507	39,000	39,166	39,248	39,298	39,331	39,355	39,373	39,386	39,398	39,415	39,426	
3	17,443	16,044	15,439	15,101	14,885	14,735	14,624	14,540	14,473	14,419	14,336	14,277	
4	12,218	10,649	9,979	9,605	9,364	9,197	9,074	8,980	8,905	8,844	8,751	8,684	
5	10,007	8,434	7,764	7,388	7,146	6,978	6,853	6,757	6,681	6,619	6,525	6,456	
6	8,813	7,260	6,599	6,227	5,988	5,820	5,695	5,600	5,523	5,461	5,366	5,297	
7	8,073	6,542	5,890	5,523	5,285	5,119	4,995	4,899	4,823	4,761	4,666	4,596	
8	7,571	6,059	5,416	5,053	4,817	4,652	4,529	4,433	4,357	4,295	4,200	4,130	
9	7,209	5,715	5,078	4,718	4,484	4,320	4,197	4,102	4,026	3,964	3,868	3,798	
10	6,937	5,456	4,826	4,468	4,236	4,072	3,950	3,855	3,779	3,717	3,621	3,550	
11	6,724	5,256	4,630	4,275	4,044	3,881	3,759	3,664	3,588	3,526	3,430	3,359	
12	6,554	5,096	4,474	4,121	3,891	3,728	3,607	3,512	3,436	3,374	3,277	3,206	
13	6,414	4,965	4,347	3,996	3,767	3,604	3,483	3,388	3,312	3,250	3,153	3,082	
14	6,298	4,857	4,242	3,892	3,663	3,501	3,380	3,285	3,209	3,147	3,050	2,979	
15	6,199	4,765	4,153	3,804	3,576	3,415	3,293	3,199	3,123	3,060	2,963	2,891	
16	6,115	4,687	4,077	3,729	3,502	3,341	3,219	3,125	3,049	2,986	2,889	2,817	
17	6,042	4,619	4,011	3,665	3,438	3,277	3,156	3,061	2,985	2,922	2,825	2,753	
18	5,978	4,560	3,954	3,608	3,382	3,221	3,100	3,005	2,929	2,866	2,769	2,696	
19	5,922	4,508	3,903	3,559	3,333	3,172	3,051	2,956	2,880	2,817	2,720	2,647	
20	5,871	4,461	3,859	3,515	3,289	3,128	3,007	2,913	2,837	2,774	2,676	2,603	
21	5,827	4,420	3,819	3,475	3,250	3,090	2,969	2,874	2,798	2,735	2,637	2,564	
22	5,786	4,383	3,783	3,440	3,215	3,055	2,934	2,839	2,763	2,700	2,602	2,528	
23	5,750	4,349	3,750	3,408	3,183	3,023	2,902	2,808	2,731	2,668	2,570	2,497	
24	5,717	4,319	3,721	3,379	3,155	2,995	2,874	2,779	2,703	2,640	2,541	2,468	
25	5,686	4,291	3,694	3,353	3,129	2,969	2,848	2,753	2,677	2,613	2,515	2,441	
26	5,659	4,265	3,670	3,329	3,105	2,945	2,824	2,729	2,653	2,590	2,491	2,417	
27	5,633	4,242	3,647	3,307	3,083	2,923	2,802	2,707	2,631	2,568	2,469	2,395	
28	5,610	4,221	3,626	3,286	3,063	2,903	2,782	2,687	2,611	2,547	2,448	2,374	
29	5,588	4,201	3,607	3,267	3,044	2,884	2,763	2,669	2,592	2,529	2,430	2,355	
30	5,568	4,182	3,589	3,250	3,026	2,867	2,746	2,651	2,575	2,511	2,412	2,338	
31	5,549	4,165	3,573	3,234	3,010	2,851	2,730	2,635	2,558	2,495	2,396	2,321	
32	5,531	4,149	3,557	3,218	2,995	2,836	2,715	2,620	2,543	2,480	2,381	2,306	
33	5,515	4,134	3,543	3,204	2,981	2,822	2,701	2,606	2,529	2,466	2,366	2,292	
34	5,499	4,120	3,529	3,191	2,968	2,808	2,688	2,593	2,516	2,453	2,353	2,278	
35	5,485	4,106	3,517	3,178	2,956	2,796	2,676	2,581	2,504	2,440	2,341	2,266	
36	5,471	4,094	3,505	3,167	2,944	2,785	2,664	2,569	2,492	2,429	2,329	2,254	
37	5,458	4,082	3,493	3,156	2,933	2,774	2,653	2,558	2,481	2,418	2,318	2,243	
38	5,446	4,071	3,483	3,145	2,923	2,763	2,643	2,548	2,471	2,407	2,307	2,232	
39	5,435	4,061	3,473	3,135	2,913	2,754	2,633	2,538	2,461	2,397	2,297	2,222	
40	5,424	4,051	3,463	3,126	2,904	2,744	2,624	2,529	2,452	2,388	2,288	2,213	
50	5,340	3,975	3,390	3,054	2,833	2,674	2,553	2,458	2,381	2,317	2,216	2,140	
60	5,286	3,925	3,343	3,008	2,786	2,627	2,507	2,412	2,334	2,270	2,169	2,093	
70	5,247	3,890	3,309	2,975	2,754	2,595	2,474	2,379	2,302	2,237	2,136	2,059	
80	5,218	3,864	3,284	2,950	2,730	2,571	2,450	2,355	2,277	2,213	2,111	2,035	
90	5,196	3,844	3,265	2,932	2,711	2,552	2,432	2,336	2,259	2,194	2,092	2,015	
100	5,179	3,828	3,250	2,917	2,696	2,537	2,417	2,321	2,244	2,179	2,077	2,000	
200	5,100	3,758	3,182	2,850	2,630	2,472	2,351	2,256	2,178	2,113	2,010	1,932	
500	5,054	3,716	3,142	2,811	2,592	2,434	2,313	2,217	2,139	2,074	1,971	1,892	
∞	5,024	3,689	3,116	2,786	2,567	2,408	2,288	2,192	2,114	2,048	1,945	1,866	

TABLE 7-2b		FRACILES de la LOI de FISHER F(v ₁ ,v ₂)									$\alpha = 0,975$	
V ₁	16	18	20	22	24	26	28	30	40	60	80	1000
1	986,93	990,35	993,10	995,38	997,25	998,86	1000,2	1001,4	1005,6	1009,8	1011,9	1017,8
2	39,435	39,442	39,448	39,453	39,456	39,460	39,462	39,464	39,473	39,481	39,486	39,497
3	14,232	14,196	14,167	14,144	14,124	14,107	14,093	14,080	14,037	13,992	13,970	13,908
4	8,633	8,592	8,560	8,533	8,511	8,492	8,475	8,461	8,411	8,360	8,335	8,264
5	6,403	6,362	6,329	6,301	6,278	6,258	6,242	6,227	6,175	6,123	6,096	6,022
6	5,244	5,202	5,168	5,141	5,117	5,097	5,080	5,065	5,012	4,959	4,932	4,856
7	4,543	4,501	4,467	4,439	4,415	4,395	4,378	4,362	4,309	4,254	4,227	4,149
8	4,076	4,034	3,999	3,971	3,947	3,927	3,909	3,894	3,840	3,784	3,756	3,677
9	3,744	3,701	3,667	3,638	3,614	3,594	3,576	3,560	3,505	3,449	3,421	3,340
10	3,496	3,453	3,419	3,390	3,365	3,345	3,327	3,311	3,255	3,198	3,169	3,087
11	3,304	3,261	3,226	3,197	3,173	3,152	3,133	3,118	3,061	3,004	2,974	2,890
12	3,152	3,108	3,073	3,043	3,019	2,998	2,979	2,963	2,906	2,848	2,818	2,733
13	3,027	2,983	2,948	2,918	2,893	2,872	2,853	2,837	2,780	2,720	2,690	2,603
14	2,923	2,879	2,844	2,814	2,789	2,767	2,749	2,732	2,674	2,614	2,583	2,495
15	2,836	2,792	2,756	2,726	2,701	2,679	2,660	2,644	2,585	2,524	2,493	2,403
16	2,761	2,717	2,681	2,651	2,625	2,603	2,584	2,568	2,509	2,447	2,415	2,324
17	2,697	2,652	2,616	2,585	2,560	2,538	2,519	2,502	2,442	2,380	2,348	2,256
18	2,640	2,596	2,559	2,529	2,503	2,481	2,461	2,445	2,384	2,321	2,289	2,195
19	2,591	2,546	2,509	2,478	2,452	2,430	2,411	2,394	2,333	2,270	2,237	2,142
20	2,547	2,501	2,464	2,434	2,408	2,385	2,366	2,349	2,287	2,223	2,190	2,094
21	2,507	2,462	2,425	2,394	2,368	2,345	2,325	2,308	2,246	2,182	2,148	2,051
22	2,472	2,426	2,389	2,358	2,331	2,309	2,289	2,272	2,210	2,145	2,111	2,012
23	2,440	2,394	2,357	2,325	2,299	2,276	2,256	2,239	2,176	2,111	2,077	1,977
24	2,411	2,365	2,327	2,296	2,269	2,246	2,226	2,209	2,146	2,080	2,045	1,945
25	2,384	2,338	2,300	2,269	2,242	2,219	2,199	2,182	2,118	2,052	2,017	1,915
26	2,360	2,314	2,276	2,244	2,217	2,194	2,174	2,157	2,093	2,026	1,991	1,888
27	2,337	2,291	2,253	2,222	2,195	2,171	2,151	2,133	2,069	2,002	1,966	1,862
28	2,317	2,270	2,232	2,201	2,174	2,150	2,130	2,112	2,048	1,980	1,944	1,839
29	2,298	2,251	2,213	2,181	2,154	2,131	2,110	2,092	2,028	1,959	1,923	1,817
30	2,280	2,233	2,195	2,163	2,136	2,112	2,092	2,074	2,009	1,940	1,904	1,797
31	2,263	2,217	2,178	2,146	2,119	2,095	2,075	2,057	1,991	1,922	1,886	1,778
32	2,248	2,201	2,163	2,131	2,103	2,080	2,059	2,041	1,975	1,905	1,869	1,760
33	2,234	2,187	2,148	2,116	2,088	2,065	2,044	2,026	1,960	1,890	1,853	1,743
34	2,220	2,173	2,135	2,102	2,075	2,051	2,030	2,012	1,946	1,875	1,838	1,727
35	2,207	2,160	2,122	2,089	2,062	2,038	2,017	1,999	1,932	1,861	1,824	1,712
36	2,196	2,148	2,110	2,077	2,049	2,025	2,005	1,986	1,919	1,848	1,811	1,698
37	2,184	2,137	2,098	2,066	2,038	2,014	1,993	1,974	1,907	1,836	1,798	1,684
38	2,174	2,126	2,088	2,055	2,027	2,003	1,982	1,963	1,896	1,824	1,786	1,672
39	2,164	2,116	2,077	2,045	2,017	1,993	1,971	1,953	1,885	1,813	1,775	1,660
40	2,154	2,107	2,068	2,035	2,007	1,983	1,962	1,943	1,875	1,803	1,764	1,648
50	2,081	2,033	1,993	1,960	1,931	1,907	1,885	1,866	1,796	1,721	1,681	1,557
60	2,033	1,985	1,944	1,911	1,882	1,857	1,835	1,815	1,744	1,667	1,625	1,495
70	1,999	1,950	1,910	1,876	1,847	1,821	1,799	1,779	1,707	1,628	1,585	1,449
80	1,974	1,925	1,884	1,850	1,820	1,795	1,772	1,752	1,679	1,599	1,555	1,414
90	1,955	1,905	1,864	1,830	1,800	1,774	1,752	1,731	1,657	1,576	1,531	1,386
100	1,939	1,890	1,849	1,814	1,784	1,758	1,735	1,715	1,640	1,558	1,512	1,363
200	1,870	1,820	1,778	1,742	1,712	1,685	1,661	1,640	1,562	1,474	1,425	1,250
500	1,830	1,779	1,736	1,700	1,669	1,641	1,617	1,596	1,515	1,423	1,370	1,166
∞	1,803	1,751	1,708	1,672	1,640	1,612	1,588	1,566	1,484	1,388	1,333	1,090

TABLE 7-3a		FRACILES de la LOI de FISHER F(v1,v2)										$\alpha=0,99$	
v ₂	v ₁	1	2	3	4	5	6	7	8	9	10	12	14
1	1	4052,3	4999,5	5403,2	5624,7	5763,7	5859,3	5928,2	5980,9	6022,4	6055,8	6106,6	6143,0
2	2	98,504	98,998	99,163	99,251	99,299	99,334	99,356	99,373	99,391	99,400	99,417	99,426
3	3	34,116	30,816	29,457	28,710	28,237	27,910	27,672	27,489	27,345	27,229	27,052	26,924
4	4	21,198	18,000	16,694	15,977	15,522	15,207	14,976	14,799	14,659	14,546	14,373	14,249
5	5	16,258	13,274	12,060	11,392	10,967	10,672	10,456	10,289	10,158	10,051	9,888	9,770
6	6	13,745	10,925	9,780	9,148	8,746	8,466	8,260	8,102	7,976	7,874	7,718	7,605
7	7	12,246	9,547	8,451	7,847	7,460	7,191	6,993	6,840	6,719	6,620	6,469	6,359
8	8	11,259	8,649	7,591	7,006	6,632	6,371	6,178	6,029	5,911	5,814	5,667	5,559
9	9	10,562	8,022	6,992	6,422	6,057	5,802	5,613	5,467	5,351	5,256	5,111	5,005
10	10	10,044	7,559	6,552	5,994	5,636	5,386	5,200	5,057	4,942	4,849	4,706	4,601
11	11	9,646	7,206	6,217	5,668	5,316	5,069	4,886	4,744	4,632	4,539	4,397	4,293
12	12	9,330	6,927	5,953	5,412	5,064	4,821	4,640	4,499	4,387	4,296	4,155	4,052
13	13	9,074	6,701	5,739	5,205	4,862	4,620	4,441	4,302	4,191	4,100	3,960	3,857
14	14	8,862	6,515	5,564	5,035	4,695	4,456	4,278	4,140	4,030	3,939	3,800	3,698
15	15	8,683	6,359	5,417	4,893	4,556	4,318	4,142	4,004	3,895	3,805	3,666	3,564
16	16	8,531	6,226	5,292	4,773	4,437	4,202	4,026	3,890	3,780	3,691	3,553	3,451
17	17	8,400	6,112	5,185	4,669	4,336	4,102	3,927	3,791	3,682	3,593	3,455	3,353
18	18	8,285	6,013	5,092	4,579	4,248	4,015	3,841	3,705	3,597	3,508	3,371	3,269
19	19	8,185	5,926	5,010	4,500	4,171	3,939	3,765	3,631	3,522	3,434	3,297	3,195
20	20	8,096	5,849	4,938	4,431	4,103	3,871	3,699	3,564	3,457	3,368	3,231	3,130
21	21	8,017	5,780	4,874	4,369	4,042	3,812	3,640	3,506	3,398	3,310	3,173	3,071
22	22	7,945	5,719	4,817	4,313	3,988	3,758	3,587	3,453	3,346	3,258	3,121	3,020
23	23	7,881	5,664	4,765	4,264	3,939	3,710	3,539	3,406	3,299	3,211	3,074	2,973
24	24	7,823	5,614	4,718	4,218	3,895	3,667	3,496	3,363	3,256	3,168	3,032	2,930
25	25	7,770	5,568	4,675	4,177	3,855	3,627	3,457	3,324	3,217	3,129	2,993	2,892
26	26	7,721	5,526	4,637	4,140	3,818	3,591	3,421	3,288	3,182	3,094	2,958	2,857
27	27	7,677	5,488	4,601	4,106	3,785	3,558	3,388	3,256	3,149	3,062	2,926	2,824
28	28	7,636	5,453	4,568	4,074	3,754	3,528	3,358	3,226	3,120	3,032	2,896	2,795
29	29	7,598	5,420	4,538	4,045	3,725	3,499	3,330	3,198	3,092	3,005	2,868	2,767
30	30	7,562	5,390	4,510	4,018	3,699	3,473	3,305	3,173	3,067	2,979	2,843	2,742
31	31	7,530	5,362	4,484	3,993	3,675	3,449	3,281	3,149	3,043	2,955	2,820	2,718
32	32	7,499	5,336	4,459	3,969	3,652	3,427	3,258	3,127	3,021	2,933	2,798	2,696
33	33	7,471	5,312	4,437	3,948	3,630	3,406	3,238	3,106	3,000	2,913	2,777	2,676
34	34	7,444	5,289	4,416	3,927	3,611	3,386	3,218	3,087	2,981	2,894	2,758	2,657
35	35	7,419	5,268	4,396	3,908	3,592	3,368	3,200	3,069	2,963	2,876	2,740	2,639
36	36	7,396	5,248	4,377	3,890	3,574	3,351	3,183	3,052	2,946	2,859	2,723	2,622
37	37	7,373	5,229	4,360	3,873	3,558	3,334	3,167	3,036	2,930	2,843	2,707	2,606
38	38	7,353	5,211	4,343	3,858	3,542	3,319	3,152	3,021	2,915	2,828	2,692	2,591
39	39	7,333	5,194	4,327	3,843	3,528	3,305	3,137	3,006	2,901	2,814	2,678	2,577
40	40	7,314	5,179	4,313	3,828	3,514	3,291	3,124	2,993	2,888	2,801	2,665	2,563
50	50	7,171	5,057	4,199	3,720	3,408	3,186	3,020	2,890	2,785	2,698	2,562	2,461
60	60	7,077	4,977	4,126	3,649	3,339	3,119	2,953	2,823	2,718	2,632	2,496	2,394
70	70	7,011	4,922	4,074	3,600	3,291	3,071	2,906	2,777	2,672	2,585	2,450	2,348
80	80	6,963	4,881	4,036	3,563	3,255	3,036	2,871	2,742	2,637	2,551	2,415	2,313
90	90	6,925	4,849	4,007	3,535	3,228	3,009	2,845	2,715	2,611	2,524	2,389	2,286
100	100	6,895	4,824	3,984	3,513	3,206	2,988	2,823	2,694	2,590	2,503	2,368	2,265
200	200	6,763	4,713	3,881	3,414	3,110	2,893	2,730	2,601	2,497	2,411	2,275	2,172
500	500	6,686	4,648	3,821	3,357	3,054	2,838	2,675	2,547	2,443	2,356	2,220	2,117
∞	∞	6,635	4,605	3,782	3,319	3,017	2,802	2,639	2,511	2,407	2,321	2,185	2,082

TABLE 7-3b		FRACTILES de la LOI de FISHER F(v1,v2)									$\alpha = 0,99$	
V_1	16	18	20	22	24	26	28	30	40	60	80	1000
1	6170,4	6191,6	6208,7	6222,6	6234,4	6244,7	6252,8	6260,4	6286,7	6313,1	6326,4	6362,7
2	99,435	99,444	99,448	99,452	99,457	99,461	99,461	99,466	99,474	99,483	99,487	99,496
3	26,827	26,751	26,690	26,640	26,598	26,562	26,532	26,504	26,411	26,316	26,269	26,136
4	14,154	14,079	14,020	13,970	13,929	13,894	13,864	13,838	13,745	13,652	13,605	13,474
5	9,680	9,610	9,553	9,506	9,466	9,433	9,404	9,379	9,291	9,202	9,157	9,031
6	7,519	7,451	7,396	7,351	7,313	7,280	7,253	7,229	7,143	7,057	7,013	6,891
7	6,275	6,209	6,155	6,111	6,074	6,043	6,016	5,992	5,908	5,824	5,781	5,660
8	5,477	5,412	5,359	5,316	5,279	5,248	5,221	5,198	5,116	5,032	4,989	4,869
9	4,924	4,860	4,808	4,765	4,729	4,698	4,672	4,649	4,567	4,483	4,441	4,321
10	4,520	4,457	4,405	4,363	4,327	4,296	4,270	4,247	4,165	4,082	4,039	3,920
11	4,213	4,150	4,099	4,057	4,021	3,990	3,964	3,941	3,860	3,776	3,734	3,613
12	3,972	3,909	3,858	3,816	3,780	3,750	3,724	3,701	3,619	3,535	3,493	3,372
13	3,778	3,716	3,665	3,622	3,587	3,556	3,530	3,507	3,425	3,341	3,298	3,176
14	3,619	3,556	3,505	3,463	3,427	3,397	3,371	3,348	3,266	3,181	3,138	3,015
15	3,485	3,423	3,372	3,330	3,294	3,264	3,237	3,214	3,132	3,047	3,004	2,880
16	3,372	3,310	3,259	3,216	3,181	3,150	3,124	3,101	3,018	2,933	2,889	2,764
17	3,275	3,212	3,162	3,119	3,084	3,053	3,026	3,003	2,920	2,835	2,791	2,664
18	3,190	3,128	3,077	3,035	2,999	2,968	2,942	2,919	2,835	2,749	2,705	2,577
19	3,116	3,054	3,003	2,961	2,925	2,894	2,868	2,844	2,761	2,674	2,630	2,501
20	3,051	2,989	2,938	2,895	2,859	2,829	2,802	2,778	2,695	2,608	2,563	2,433
21	2,993	2,931	2,880	2,837	2,801	2,770	2,743	2,720	2,636	2,548	2,503	2,372
22	2,941	2,879	2,827	2,785	2,749	2,718	2,691	2,667	2,583	2,495	2,450	2,317
23	2,894	2,832	2,780	2,738	2,702	2,671	2,644	2,620	2,535	2,447	2,401	2,268
24	2,852	2,789	2,738	2,695	2,659	2,628	2,601	2,577	2,492	2,403	2,357	2,223
25	2,813	2,751	2,699	2,657	2,620	2,589	2,562	2,538	2,453	2,364	2,317	2,182
26	2,778	2,715	2,664	2,621	2,585	2,554	2,526	2,503	2,417	2,327	2,281	2,144
27	2,746	2,683	2,632	2,589	2,552	2,521	2,494	2,470	2,384	2,294	2,247	2,109
28	2,716	2,653	2,602	2,559	2,522	2,491	2,464	2,440	2,354	2,263	2,216	2,077
29	2,689	2,626	2,574	2,531	2,495	2,463	2,436	2,412	2,325	2,234	2,187	2,047
30	2,663	2,600	2,549	2,506	2,469	2,437	2,410	2,386	2,299	2,208	2,160	2,019
31	2,640	2,577	2,525	2,482	2,445	2,414	2,386	2,362	2,275	2,183	2,135	1,993
32	2,618	2,555	2,503	2,460	2,423	2,391	2,364	2,340	2,252	2,160	2,112	1,969
33	2,597	2,534	2,482	2,439	2,402	2,370	2,343	2,319	2,231	2,139	2,090	1,946
34	2,578	2,515	2,463	2,420	2,383	2,351	2,323	2,299	2,211	2,118	2,070	1,925
35	2,560	2,497	2,445	2,401	2,364	2,333	2,305	2,281	2,193	2,099	2,050	1,905
36	2,543	2,480	2,428	2,384	2,347	2,315	2,288	2,263	2,175	2,082	2,032	1,886
37	2,527	2,464	2,412	2,368	2,331	2,299	2,271	2,247	2,159	2,065	2,015	1,868
38	2,512	2,449	2,397	2,353	2,316	2,284	2,256	2,232	2,143	2,049	1,999	1,850
39	2,498	2,434	2,382	2,339	2,302	2,270	2,242	2,217	2,128	2,034	1,984	1,834
40	2,484	2,421	2,369	2,325	2,288	2,256	2,228	2,203	2,114	2,019	1,969	1,819
50	2,382	2,318	2,265	2,221	2,183	2,151	2,123	2,098	2,007	1,909	1,857	1,698
60	2,315	2,251	2,198	2,153	2,115	2,083	2,054	2,028	1,936	1,836	1,783	1,617
70	2,268	2,204	2,150	2,106	2,067	2,034	2,005	1,980	1,886	1,785	1,730	1,558
80	2,233	2,169	2,115	2,070	2,032	1,999	1,969	1,944	1,849	1,746	1,690	1,512
90	2,206	2,142	2,088	2,043	2,004	1,971	1,942	1,916	1,820	1,716	1,659	1,476
100	2,185	2,120	2,067	2,021	1,983	1,949	1,919	1,893	1,797	1,692	1,634	1,447
200	2,091	2,026	1,971	1,925	1,886	1,851	1,821	1,794	1,694	1,583	1,521	1,304
500	2,036	1,970	1,915	1,869	1,829	1,794	1,763	1,735	1,633	1,517	1,452	1,201
∞	2,000	1,934	1,878	1,831	1,791	1,755	1,724	1,696	1,592	1,473	1,404	1,107

TABLE 7-4a		FRACILES de la LOI de FISHER F(v ₁ ,v ₂)										$\alpha=0,995$	
v ₂	v ₁	1	2	3	4	5	6	7	8	9	10	12	14
1	1	16211	19999	21613	22497	23058	23437	23715	23924	24091	24225	24426	24572
2	1	198,50	198,99	199,17	199,24	199,31	199,34	199,34	199,38	199,38	199,40	199,42	199,43
3	1	55,551	49,801	47,466	46,196	45,392	44,839	44,435	44,125	43,881	43,687	43,387	43,171
4	1	31,332	26,284	24,259	23,155	22,457	21,975	21,622	21,351	21,140	20,966	20,704	20,515
5	1	22,785	18,314	16,530	15,556	14,939	14,514	14,200	13,961	13,771	13,618	13,384	13,215
6	1	18,635	14,544	12,916	12,027	11,464	11,073	10,786	10,566	10,392	10,250	10,034	9,877
7	1	16,236	12,404	10,882	10,051	9,522	9,155	8,886	8,678	8,514	8,380	8,176	8,028
8	1	14,688	11,043	9,597	8,805	8,302	7,952	7,694	7,496	7,338	7,211	7,015	6,872
9	1	13,614	10,107	8,717	7,956	7,471	7,134	6,885	6,693	6,541	6,417	6,227	6,089
10	1	12,826	9,427	8,081	7,343	6,872	6,545	6,302	6,116	5,967	5,847	5,661	5,526
11	1	12,226	8,912	7,600	6,881	6,422	6,102	5,865	5,682	5,537	5,418	5,236	5,103
12	1	11,754	8,510	7,226	6,521	6,071	5,757	5,525	5,345	5,202	5,085	4,906	4,775
13	1	11,374	8,186	6,926	6,233	5,791	5,482	5,253	5,076	4,935	4,820	4,643	4,513
14	1	11,060	7,922	6,680	5,998	5,562	5,257	5,031	4,857	4,717	4,603	4,428	4,299
15	1	10,798	7,701	6,476	5,803	5,372	5,071	4,847	4,674	4,536	4,423	4,250	4,122
16	1	10,575	7,514	6,303	5,638	5,212	4,913	4,692	4,521	4,384	4,272	4,099	3,972
17	1	10,384	7,354	6,156	5,497	5,075	4,779	4,559	4,389	4,254	4,142	3,971	3,844
18	1	10,218	7,215	6,028	5,375	4,956	4,663	4,445	4,276	4,141	4,030	3,860	3,734
19	1	10,072	7,093	5,916	5,268	4,853	4,561	4,345	4,177	4,043	3,933	3,763	3,638
20	1	9,944	6,986	5,818	5,174	4,762	4,472	4,257	4,090	3,956	3,847	3,678	3,553
21	2	9,829	6,891	5,730	5,091	4,681	4,393	4,179	4,013	3,880	3,771	3,602	3,478
22	2	9,727	6,806	5,652	5,017	4,609	4,322	4,109	3,944	3,812	3,703	3,535	3,411
23	2	9,635	6,730	5,582	4,950	4,544	4,259	4,047	3,882	3,750	3,642	3,475	3,351
24	2	9,551	6,661	5,519	4,890	4,486	4,202	3,991	3,826	3,695	3,587	3,420	3,296
25	2	9,475	6,598	5,462	4,835	4,433	4,150	3,939	3,776	3,645	3,537	3,370	3,247
26	2	9,406	6,541	5,409	4,785	4,384	4,103	3,893	3,730	3,599	3,492	3,325	3,202
27	2	9,342	6,489	5,361	4,740	4,340	4,059	3,850	3,688	3,557	3,450	3,284	3,161
28	2	9,284	6,440	5,317	4,698	4,300	4,020	3,811	3,649	3,519	3,412	3,246	3,123
29	2	9,230	6,396	5,276	4,659	4,262	3,983	3,775	3,613	3,483	3,377	3,211	3,088
30	2	9,180	6,355	5,239	4,623	4,228	3,949	3,742	3,580	3,450	3,344	3,179	3,056
31	3	9,133	6,316	5,204	4,590	4,195	3,918	3,711	3,549	3,420	3,314	3,149	3,026
32	3	9,090	6,281	5,172	4,559	4,166	3,889	3,682	3,521	3,392	3,286	3,121	2,998
33	3	9,049	6,248	5,141	4,531	4,138	3,861	3,655	3,494	3,366	3,260	3,095	2,973
34	3	9,012	6,217	5,113	4,504	4,112	3,836	3,630	3,470	3,341	3,235	3,071	2,948
35	3	8,976	6,188	5,086	4,479	4,088	3,812	3,607	3,447	3,318	3,212	3,048	2,926
36	3	8,943	6,161	5,062	4,455	4,065	3,790	3,585	3,425	3,296	3,191	3,027	2,904
37	3	8,912	6,135	5,038	4,433	4,043	3,769	3,564	3,404	3,276	3,171	3,007	2,885
38	3	8,882	6,111	5,016	4,412	4,023	3,749	3,545	3,385	3,257	3,152	2,988	2,866
39	3	8,854	6,088	4,995	4,392	4,004	3,731	3,526	3,367	3,239	3,134	2,970	2,848
40	3	8,828	6,066	4,976	4,374	3,986	3,713	3,509	3,350	3,222	3,117	2,953	2,831
50	3	8,626	5,902	4,826	4,232	3,849	3,579	3,376	3,219	3,092	2,988	2,825	2,703
60	3	8,495	5,795	4,729	4,140	3,760	3,492	3,291	3,134	3,008	2,904	2,742	2,620
70	3	8,403	5,720	4,661	4,076	3,698	3,431	3,232	3,075	2,950	2,846	2,684	2,563
80	3	8,335	5,665	4,611	4,028	3,652	3,387	3,188	3,032	2,907	2,803	2,641	2,520
90	3	8,282	5,623	4,573	3,992	3,617	3,352	3,154	2,999	2,873	2,770	2,608	2,487
100	3	8,241	5,589	4,542	3,963	3,589	3,325	3,127	2,972	2,847	2,744	2,583	2,461
200	3	8,057	5,441	4,408	3,837	3,467	3,206	3,010	2,856	2,732	2,629	2,468	2,347
500	3	7,950	5,355	4,330	3,763	3,396	3,137	2,941	2,789	2,665	2,562	2,402	2,281
∞	3	7,879	5,298	4,279	3,715	3,350	3,091	2,897	2,744	2,621	2,519	2,358	2,237

TABLE 7-4b		FRACTILES de la LOI de FISHER F(v1,v2)									$\alpha = 0,995$	
V ₁	16	18	20	22	24	26	28	30	40	60	80	1000
V ₂												
1	24681	24766	24839	24890	24942	24977	25011	25046	25151	25252	25306	25453
2	199,43	199,45	199,45	199,45	199,45	199,45	199,47	199,47	199,49	199,49	199,49	199,50
3	43,009	42,882	42,777	42,693	42,622	42,564	42,509	42,464	42,308	42,150	42,071	41,849
4	20,371	20,259	20,167	20,092	20,030	19,977	19,932	19,892	19,751	19,611	19,540	19,342
5	13,086	12,985	12,903	12,836	12,780	12,732	12,692	12,656	12,530	12,403	12,338	12,159
6	9,758	9,664	9,589	9,526	9,474	9,430	9,392	9,358	9,241	9,122	9,062	8,894
7	7,915	7,826	7,754	7,695	7,645	7,603	7,566	7,535	7,422	7,309	7,251	7,090
8	6,763	6,678	6,608	6,551	6,503	6,462	6,427	6,396	6,288	6,177	6,121	5,964
9	5,983	5,899	5,832	5,776	5,729	5,689	5,655	5,625	5,519	5,410	5,356	5,201
10	5,422	5,340	5,274	5,219	5,173	5,134	5,100	5,071	4,966	4,859	4,805	4,652
11	5,001	4,921	4,855	4,801	4,756	4,717	4,684	4,654	4,551	4,445	4,391	4,239
12	4,674	4,594	4,530	4,476	4,431	4,393	4,360	4,331	4,228	4,123	4,069	3,917
13	4,413	4,334	4,270	4,217	4,173	4,134	4,101	4,073	3,970	3,866	3,812	3,660
14	4,200	4,122	4,059	4,006	3,961	3,923	3,891	3,862	3,760	3,655	3,602	3,449
15	4,024	3,946	3,883	3,830	3,786	3,748	3,715	3,687	3,585	3,480	3,427	3,274
16	3,875	3,797	3,734	3,682	3,638	3,600	3,567	3,539	3,437	3,332	3,279	3,125
17	3,747	3,670	3,607	3,555	3,511	3,473	3,441	3,412	3,311	3,206	3,152	2,998
18	3,637	3,560	3,498	3,446	3,402	3,364	3,332	3,303	3,201	3,096	3,042	2,887
19	3,541	3,465	3,402	3,350	3,306	3,269	3,236	3,208	3,106	3,000	2,946	2,790
20	3,457	3,380	3,318	3,266	3,222	3,184	3,152	3,123	3,022	2,916	2,861	2,705
21	3,382	3,305	3,243	3,191	3,147	3,110	3,077	3,049	2,947	2,841	2,786	2,628
22	3,315	3,239	3,176	3,125	3,081	3,043	3,011	2,982	2,880	2,774	2,719	2,560
23	3,255	3,179	3,116	3,065	3,021	2,983	2,951	2,922	2,820	2,713	2,658	2,498
24	3,201	3,125	3,062	3,011	2,967	2,929	2,896	2,868	2,765	2,658	2,603	2,442
25	3,152	3,075	3,013	2,961	2,918	2,880	2,847	2,819	2,716	2,609	2,553	2,391
26	3,107	3,031	2,969	2,917	2,873	2,835	2,802	2,774	2,671	2,563	2,507	2,345
27	3,066	2,990	2,928	2,876	2,832	2,794	2,761	2,733	2,630	2,522	2,466	2,302
28	3,028	2,952	2,890	2,838	2,794	2,756	2,724	2,695	2,592	2,483	2,427	2,262
29	2,993	2,917	2,855	2,803	2,759	2,722	2,689	2,660	2,557	2,448	2,391	2,225
30	2,961	2,885	2,823	2,771	2,727	2,689	2,657	2,628	2,524	2,415	2,358	2,191
31	2,931	2,855	2,793	2,741	2,697	2,660	2,627	2,598	2,494	2,385	2,328	2,160
32	2,904	2,828	2,766	2,714	2,670	2,632	2,599	2,570	2,466	2,356	2,299	2,130
33	2,878	2,802	2,740	2,688	2,644	2,606	2,573	2,544	2,440	2,330	2,272	2,102
34	2,854	2,778	2,716	2,664	2,620	2,582	2,549	2,520	2,415	2,305	2,247	2,076
35	2,831	2,755	2,693	2,641	2,597	2,559	2,526	2,497	2,392	2,282	2,224	2,052
36	2,810	2,734	2,672	2,620	2,576	2,538	2,505	2,475	2,371	2,260	2,202	2,029
37	2,790	2,714	2,652	2,600	2,556	2,518	2,484	2,455	2,350	2,239	2,181	2,007
38	2,771	2,695	2,633	2,581	2,537	2,499	2,465	2,436	2,331	2,220	2,161	1,986
39	2,753	2,677	2,615	2,563	2,519	2,481	2,448	2,418	2,313	2,201	2,143	1,967
40	2,737	2,661	2,598	2,546	2,502	2,464	2,431	2,401	2,296	2,184	2,125	1,948
50	2,609	2,533	2,470	2,418	2,373	2,335	2,301	2,272	2,164	2,050	1,989	1,804
60	2,526	2,450	2,387	2,335	2,290	2,251	2,217	2,187	2,079	1,962	1,900	1,707
70	2,468	2,392	2,329	2,276	2,231	2,192	2,158	2,128	2,019	1,900	1,837	1,637
80	2,425	2,349	2,286	2,233	2,188	2,149	2,115	2,084	1,974	1,854	1,789	1,584
90	2,393	2,316	2,253	2,200	2,155	2,115	2,081	2,051	1,939	1,818	1,752	1,542
100	2,367	2,290	2,227	2,174	2,128	2,089	2,054	2,024	1,912	1,790	1,723	1,508
200	2,252	2,175	2,112	2,058	2,012	1,972	1,936	1,905	1,790	1,661	1,590	1,343
500	2,185	2,108	2,044	1,990	1,943	1,903	1,867	1,835	1,717	1,584	1,509	1,225
∞	2,142	2,064	2,000	1,945	1,898	1,857	1,821	1,789	1,669	1,533	1,454	1,119

TABLE 8		FRACTILES de la LOI de STUDENT									
α v	0,6	0,7	0,8	0,9	0,95	0,975	0,99	0,995	0,999	0,9995	
1	0,3249	0,7265	1,3764	3,0777	6,3137	12,7062	31,8210	63,6559	318,2888	636,5776	
2	0,2887	0,6172	1,0607	1,8856	2,9200	4,3027	6,9645	9,9250	22,3285	31,5998	
3	0,2767	0,5844	0,9785	1,6377	2,3534	3,1824	4,5407	5,8408	10,2143	12,9244	
4	0,2707	0,5686	0,9410	1,5332	2,1318	2,7765	3,7469	4,6041	7,1729	8,6101	
5	0,2672	0,5594	0,9195	1,4759	2,0150	2,5706	3,3649	4,0321	5,8935	6,8685	
6	0,2648	0,5534	0,9057	1,4398	1,9432	2,4469	3,1427	3,7074	5,2075	5,9587	
7	0,2632	0,5491	0,8960	1,4149	1,8946	2,3646	2,9979	3,4995	4,7853	5,4081	
8	0,2619	0,5459	0,8889	1,3968	1,8595	2,3060	2,8965	3,3554	4,5008	5,0414	
9	0,2610	0,5435	0,8834	1,3830	1,8331	2,2622	2,8214	3,2498	4,2969	4,7809	
10	0,2602	0,5415	0,8791	1,3722	1,8125	2,2281	2,7638	3,1693	4,1437	4,5868	
11	0,2596	0,5399	0,8755	1,3634	1,7959	2,2010	2,7181	3,1058	4,0248	4,4369	
12	0,2590	0,5386	0,8726	1,3562	1,7823	2,1788	2,6810	3,0545	3,9296	4,3178	
13	0,2586	0,5375	0,8702	1,3502	1,7709	2,1604	2,6503	3,0123	3,8520	4,2209	
14	0,2582	0,5366	0,8681	1,3450	1,7613	2,1448	2,6245	2,9768	3,7874	4,1403	
15	0,2579	0,5357	0,8662	1,3406	1,7531	2,1315	2,6025	2,9467	3,7329	4,0728	
16	0,2576	0,5350	0,8647	1,3368	1,7459	2,1199	2,5835	2,9208	3,6861	4,0149	
17	0,2573	0,5344	0,8633	1,3334	1,7396	2,1098	2,5669	2,8982	3,6458	3,9651	
18	0,2571	0,5338	0,8620	1,3304	1,7341	2,1009	2,5524	2,8784	3,6105	3,9217	
19	0,2569	0,5333	0,8610	1,3277	1,7291	2,0930	2,5395	2,8609	3,5793	3,8833	
20	0,2567	0,5329	0,8600	1,3253	1,7247	2,0860	2,5280	2,8453	3,5518	3,8496	
21	0,2566	0,5325	0,8591	1,3232	1,7207	2,0796	2,5176	2,8314	3,5271	3,8193	
22	0,2564	0,5321	0,8583	1,3212	1,7171	2,0739	2,5083	2,8188	3,5050	3,7922	
23	0,2563	0,5317	0,8575	1,3195	1,7139	2,0687	2,4999	2,8073	3,4850	3,7676	
24	0,2562	0,5314	0,8569	1,3178	1,7109	2,0639	2,4922	2,7970	3,4668	3,7454	
25	0,2561	0,5312	0,8562	1,3163	1,7081	2,0595	2,4851	2,7874	3,4502	3,7251	
26	0,2560	0,5309	0,8557	1,3150	1,7056	2,0555	2,4786	2,7787	3,4350	3,7067	
27	0,2559	0,5306	0,8551	1,3137	1,7033	2,0518	2,4727	2,7707	3,4210	3,6895	
28	0,2558	0,5304	0,8546	1,3125	1,7011	2,0484	2,4671	2,7633	3,4082	3,6739	
29	0,2557	0,5302	0,8542	1,3114	1,6991	2,0452	2,4620	2,7564	3,3963	3,6595	
30	0,2556	0,5300	0,8538	1,3104	1,6973	2,0423	2,4573	2,7500	3,3852	3,6460	
31	0,2555	0,5298	0,8534	1,3095	1,6955	2,0395	2,4528	2,7440	3,3749	3,6335	
32	0,2555	0,5297	0,8530	1,3086	1,6939	2,0369	2,4487	2,7385	3,3653	3,6218	
33	0,2554	0,5295	0,8526	1,3077	1,6924	2,0345	2,4448	2,7333	3,3563	3,6109	
34	0,2553	0,5294	0,8523	1,3070	1,6909	2,0322	2,4411	2,7284	3,3480	3,6007	
35	0,2553	0,5292	0,8520	1,3062	1,6896	2,0301	2,4377	2,7238	3,3400	3,5911	
36	0,2552	0,5291	0,8517	1,3055	1,6883	2,0281	2,4345	2,7195	3,3326	3,5821	
37	0,2552	0,5289	0,8514	1,3049	1,6871	2,0262	2,4314	2,7154	3,3256	3,5737	
38	0,2551	0,5288	0,8512	1,3042	1,6860	2,0244	2,4286	2,7116	3,3190	3,5657	
39	0,2551	0,5287	0,8509	1,3036	1,6849	2,0227	2,4258	2,7079	3,3127	3,5581	
40	0,2550	0,5286	0,8507	1,3031	1,6839	2,0211	2,4233	2,7045	3,3069	3,5510	
50	0,2547	0,5278	0,8489	1,2987	1,6759	2,0086	2,4033	2,6778	3,2614	3,4960	
60	0,2545	0,5272	0,8477	1,2958	1,6706	2,0003	2,3901	2,6603	3,2317	3,4602	
70	0,2543	0,5268	0,8468	1,2938	1,6669	1,9944	2,3808	2,6479	3,2108	3,4350	
80	0,2542	0,5265	0,8461	1,2922	1,6641	1,9901	2,3739	2,6387	3,1952	3,4164	
90	0,2541	0,5263	0,8456	1,2910	1,6620	1,9867	2,3685	2,6316	3,1832	3,4019	
100	0,2540	0,5261	0,8452	1,2901	1,6602	1,9840	2,3642	2,6259	3,1738	3,3905	
200	0,2537	0,5252	0,8434	1,2858	1,6525	1,9719	2,3451	2,6006	3,1315	3,3398	
500	0,2535	0,5247	0,8423	1,2832	1,6479	1,9647	2,3338	2,5857	3,1066	3,3101	
∞	0,2533	0,5244	0,8416	1,2816	1,6449	1,9600	2,3264	2,5758	3,0902	3,2905	

TABLE 9		NOMBRES au HASARD compris entre 0 et 99 999								
44513	79304	45999	42931	89453	21766	38448	93139	42142	56585	
63769	64824	65680	87857	61487	75567	11573	29426	83282	91514	
23701	82446	7663	1159	2370	36369	44142	96184	71080	10198	
60706	23505	12922	24589	28002	39253	63763	43228	97956	63891	
62625	73388	76912	91768	51733	12049	37551	2527	95253	32041	
97741	52121	18280	12442	63394	59438	39491	96112	28915	13635	
11912	41105	14787	16148	12829	62742	8832	19681	20273	83739	
76048	48850	77643	70151	13843	76183	57912	11591	46125	11132	
82771	50001	9819	20424	46502	3984	67924	20023	96314	62812	
19147	73414	20411	67824	39797	74343	36203	47960	17027	81692	
55683	93456	7888	79094	48687	57981	63633	14901	75898	55000	
92530	96077	42427	6325	30555	48469	50089	41740	95583	46154	
29643	19245	83942	19292	34121	74902	86921	37803	84352	74561	
29829	77040	77299	83276	69806	77685	98886	23865	78292	45398	
77693	67158	66291	36386	29123	65473	17477	51502	90340	14130	
86528	17846	1771	80489	53147	66310	1491	85682	11024	52438	
33269	12678	42235	12282	36645	82254	7093	9091	6128	41878	
26069	36015	62195	98630	92280	80396	76414	90119	8468	92861	
69929	89008	14496	52947	63215	43829	89739	85355	64379	31549	
78048	35858	97792	54340	4443	41858	96550	21380	78322	9648	
73067	16714	407	56540	36975	82610	6228	70765	26112	60595	
68134	88407	31877	39692	83425	6706	61723	56149	72175	68990	
55743	50345	17516	97678	28040	6670	71407	65548	73184	98722	
43781	15540	16271	40006	80717	92074	23335	95520	8927	36360	
86928	18233	63476	65507	36589	51106	20581	21546	80369	77218	
69927	96034	85539	21522	90327	80243	44706	56758	21763	69586	
22426	94064	12664	22237	89003	83634	4527	23652	74830	57881	
19533	11659	83086	47629	88385	59553	65914	6460	16117	18928	
33499	76765	69306	51629	35631	225	23926	29804	77661	50744	
32301	56146	57065	14921	74890	63981	55111	22608	56471	66687	
38168	33664	41199	36709	78934	62800	35388	39916	43083	48399	
94199	91719	22948	31952	53957	23199	9327	85224	21287	29824	
32686	24194	40027	16426	81814	71160	57637	9361	38345	44814	
87010	954	68402	19830	34110	16069	91874	91067	19730	33286	
69964	30338	77968	49415	52706	15104	84197	25747	40943	94646	
13112	35902	90563	24719	72284	97551	27917	66124	70680	89609	
15535	98350	2319	73249	95260	62812	71158	39675	46606	13830	
7931	33651	30711	34824	72784	43952	31099	30670	51615	7141	
56785	12538	10558	68006	81789	17397	88575	84489	6411	40901	
56575	8394	39363	69211	45092	88313	88165	64746	56312	89404	
76272	44981	90635	17571	5009	64320	65212	15176	9866	90550	
79286	56843	5685	23448	12911	56674	73540	61282	20776	61267	
62696	70262	5850	17380	9194	22125	50708	8841	36258	41372	
58886	84028	42365	64802	37944	84474	64501	29126	61634	49918	
89617	85436	98571	57477	83142	44269	84159	94060	37499	59833	
28616	1253	23519	66328	25786	53294	14462	51929	38716	16965	
1313	14050	70980	38026	33053	68609	99343	63932	32201	63824	
90494	70102	36865	5398	3002	23774	9174	24049	45586	76044	
31632	74556	95566	15434	52000	22750	27044	82589	8519	82068	
50469	50212	59315	73225	49481	87183	97726	64482	49388	78416	

BIBLIOGRAPHIE

APMEP (Association des professeurs de mathématiques de l'enseignement public), *Analyse des données* (tomes 1 et 2), 1980.

LECOUTRE J.-P., *Statistique et probabilités*, Dunod, 3^e édition, 2006.

CERESTA, *Tables statistiques*.

BAILLARGEON G., *Méthodes statistiques de l'ingénieur*, Éditions SMG, 1990.

BAILLARGEON G., RAINVILLE J., *Statistique appliquée*, Tome 2 : *Tests statistiques, régression et corrélation*, et Tome 3 : *Régression multiple*, Éditions SMG, 1990.

BASS J., *Éléments de calcul des probabilités*, Masson, 1974.

BENZÉCRI J.-P. et coll., *L'Analyse des données* (tomes 2 et 3), 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, 1994.

BOREL E., *Les probabilités et la vie*, coll. « Que sais-je ? », PUF, 1950.

BOULEAU N., *Probabilités de l'ingénieur*, Hermann, 1986.

BOULEAU N., TALAY D., *Probabilités numériques*, INRIA, 1992.

BOUROCHE J.M., SAPORTA G., *L'Analyse des données*, coll. « Que sais-je ? », 1980.

BRODEAU F., ROMIER G., *Mathématiques pour l'informatique : probabilités*, Armand Colin, 1973.

BOWKER A.H., LIEBERMAN G.J., *Méthodes statistiques de l'ingénieur*, Dunod, 1965.

- BOX E.P., TIAO C., *Bayesian inference in statistical analysis*, Addison-Wesley, 1973.
- BUCKLEW J.A., *Large deviation techniques in decision, simulation and estimation*, John Wiley, 1990.
- CAILLEZ F., PAGES J.P., *Introduction à l'analyse des données*, SMASH, 1976.
- CARTON D., *Processus aléatoires utilisés en recherche opérationnelle*, Masson, 1975.
- CHAPOUILLE P., *La fiabilité*, coll. « Que sais-je ? », 1972.
- CROW E.L., SHIMIZU K., *Log-normal distributions : theory and applications*, 1988.
- DESROCHES A., *Concepts et méthodes probabilistes de base de la sécurité*, Lavoisier, 1995.
- DOOB J.L., *Stochastic processes*, John Wiley, 1953.
- DYNKIN E.B., *Théorie des processus markoviens*, Dunod, 1963.
- Dictionnaire des mathématiques*, *Encyclopedia Universalis*, Albin Michel, 1998.
- FELDMAN J., LAGNEAU G., MATALON B., *Moyenne, milieu, centre*, Éditions de l'EHESS (École des hautes études en sciences sociales).
- FOURGEAUD C., FUCHS A., *Statistique*, Dunod, 1967.
- FORTET R., *Éléments de la théorie des probabilités*, CNRS, 1965.
- FREEMAN, *Introduction aux statistiques*, Addison Wesley, 1963.
- GIRSCHIG R., *Analyse statistique*, École Centrale de Paris.
- GOOD Ph., *Permutation tests*, Springer Verlag, 1994.
- GOURIEROUX G., *Théorie des sondages*, Économica, 1981.
- GUMBEL E.J., *Statistics of extremes*, Columbia University Press, 1958.
- JACQUARD A., *Les probabilités*, coll. « Que sais-je ? », 1974.
- LANG-MICHAUT C., *Pratique des tests statistiques : interprétation des données*, Dunod, 1990.
- LEBART L., MORINEAU A., FENELON J.P., *Traitement des données statistiques*, Dunod, 1979.
- LEBART L., MORINEAU A., TABARD N., *Techniques de la description statistique*, Dunod, 1977.

LELOUCH J., LAZAR P., *Méthodes statistiques en expérimentation biologique*, Flammarion, 1988.

MÉTIVIER M., *Notions fondamentales de la théorie des probabilités*, Dunod, 1968.

MÉTIVIER M., NEVEU J., *Probabilités*, École polytechnique, 1979.

MONFORT A., *Cours de probabilités*, Économica, 1980.

MORISSON D.F., *Multivariate statistical analysis*, McGraw Hill, 1967.

NEVEU J., *Bases mathématiques du calcul des probabilités*, Masson, 1964.

PELLAUMAIL J., *Probabilités, statistique, files d'attente*, Dunod, 1986.

PILZ J., *Bayesian estimation and experimental design in linear regression models*, John Wiley, 1989.

PROACCIA H., PIEPSZOWNIK L., *Statistique fréquentielle et bayésienne*, Eyrolles, 1992.

ROSS M., *Initiation aux probabilités*, Presses polytechniques romandes, 1984.

ROSS G.J.S., *Non linear estimation*, Springer Verlag, 1990.

SAPORTA G., *Théories et méthodes de la statistique*, Technip, 1978.

SAPORTA G., *Probabilités et statistique*, École centrale de Paris, 1984.

SAPORTA G., *Probabilités : analyse des données et statistique*, Technip, 1990.

SAVILLE D.J., WOOD G.R., *Statistical methods : The geometric approach*, Springer Verlag, 1993.

SOIZE Ch., *Éléments mathématiques de la théorie déterministe et aléatoire du signal*, ENSTA, 1985.

STAUDTE G., SHEATHER S., *Robust estimation and testing*, John Wiley, 1990.

TASSI Ph., LEGAIT S., *Théorie des probabilités en vue des applications statistiques*, Technip, 1990.

TENENHAUS M., *Méthodes statistiques en gestion*, Dunod, 1994.

INDEX

A

- ajustement linéaire 377
- analyse
 - canonique 350
 - combinatoire 41
 - de la covariance 349
 - de la variance 349
 - à double entrée 411
 - à simple entrée 299
 - emboîtée 422
 - orthogonale à entrées multiples 419
 - en composantes principales 350
 - factorielle des correspondances 351
 - factorielle discriminante 351
- approximation conditionnelle 355

B

- biais de l'estimateur 223
- boîte à moustaches 19

C

- cadence 167
- caractère 4
- cardinal 51, 433, 434
- carré latin 427

chaîne

- apériodique 163
- de Markov 158
 - homogène 160
- irréductible 162
- classe 8, 9
- coefficient
 - d'aplatissement 22
 - d'asymétrie 22
 - de contingence 29
 - de corrélation des rangs de Kendall 310
 - de corrélation des rangs de Spearman 309
 - de corrélation linéaire 136, 353
 - de détermination 362
 - de Pearson 29
 - de Tschuprow 30
 - de variation 21
- convergence
 - en loi 122
 - en moyenne quadratique 124
 - en probabilité 120
 - presque sûre 121
- convolution 112, 440

- corrélation
 - multiple 390
 - partielle 392
- couple de variables aléatoires 127, 132
- courbe
 - d'efficacité du test 268
 - de concentration 23
 - de fréquences cumulées 12
 - de puissance du test 268
 - de régression 143, 355
- covariance 64, 134, 343
- D**
- défaillance 322
- densité de probabilité 54, 59, 89
 - conjointe 132
 - de transition 164
- distribution cumulée des défaillances 323
- droite
 - de Henry 280
 - de régression 353
 - des moindres carrés 361
- E**
- écart
 - moyen 340
 - résiduel 364
 - type 19, 62, 339
- échantillon aléatoire 5, 180
- effet global de deux facteurs 412
- efficacité de l'estimateur 224
- ensemble statistique 4
- équations d'état du système 172
- espace
 - des états 147
 - des individus 347
 - des variables 347
 - probabilisé 40
- espérance
 - conditionnelle 129
 - mathématique 59, 139
- estimateur 220
 - absolument correct 225
 - convergent 221
 - de variance minimale 227
 - sans biais 221
- estimation 210
 - par intervalle de confiance 235
 - ponctuelle 220
- état(s)
 - accessible 161
 - communicants 161
 - périodique 162
 - récurrent 163
 - transitoire 163
- étendue 22, 340
- événement(s)
 - aléatoire 37
 - incompatibles 38, 42, 44
 - indépendants 44
- expérience aléatoire 36
- F**
- facteur de charge du centre 172
- fiabilité 93, 321
 - d'un matériel usagé 326
- file d'attente 171
- fonction(s)
 - caractéristique
 - d'une variable aléatoire 116
 - du segment $[a, b]$ 90, 439
 - d'auto-corrélation 149

de régression 130
 de répartition 52, 89
 conjointe 132
 empirique 196
 de survie 323
 marginales 59
 fractiles 64
 fréquence
 absolue 5
 conditionnelle 27
 cumulée absolue 6
 cumulée relative 6

H

histogramme 11
 hypothèse
 alternative 257
 nulle 257

I

indépendance 129
 indice de Gini 23
 inégalité
 de Bienaymé-Tchebyshev 65
 de Cramer-Rao 228
 interaction 413
 intervalle
 de confiance 235
 bilatéral 238
 unilatéral 238
 de prévision 371, 390
 de probabilité 235

J

Jacobien 436

K

Kendall 314

L

levier 361
 loi
 bêta de type I 97, 189
 bêta de type II 98, 190
 binomiale 72, 85, 86, 106
 conditionnelle 128
 conjointe 127
 de Cauchy 60, 191
 de Dirac 69
 de Fisher-Snedecor 188, 194
 de Fréchet 201
 de Gumbel 201
 de Laplace-Gauss 100, 280
 de la somme 112
 de la variable $X_{(1)}$, plus petite valeur observée 202
 de la variable $X_{(n)}$, plus grande valeur observée 199
 de Pascal 76
 de Poisson 83, 86, 108, 113
 de probabilité 50
 de Student 190, 193, 194
 de Weibull 200, 281, 331
 des grands nombres 124
 du chi-deux 185, 319
 exponentielle 92, 281, 328
 gamma 95, 114, 187
 géométrique 75
 hypergéométrique 81, 85
 limite 84
 log-normale 109
 marginale 128, 132
 multidimensionnelle 77
 normale 100, 280

- à deux dimensions 142
 - multidimensionnelle 141
 - uniforme 70, 90
- M**
- marche au hasard 157
 - martingale 154
 - matrice
 - de corrélation 344
 - de transition 160
 - de variance-covariance 135, 139, 347
 - R de corrélation 348
 - stochastique 160
 - médiale 18
 - médiane 15, 339
 - méthode
 - de classification 350
 - de régression 349
 - du maximum de vraisemblance 229
 - modalité 4
 - mode 17
 - modèle linéaire simple 358
 - moment 59, 61, 133
 - mouvement brownien 156
 - moyenne 339
 - arithmétique 14
 - des valeurs extrêmes 339
 - MTBF 322
- N**
- nombres au hasard 205
 - pseudo-aléatoires 205
- P**
- papier d'Allan Plait 282
 - permutation 433
 - polygone de fréquences 12
 - population 4
 - précision d'un estimateur 224
 - probabilités 33, 40
 - composées 44
 - conditionnelles 42
 - totales 41, 46
 - processus
 - à accroissements indépendants du temps 153
 - à accroissements stationnaires 150
 - de Markov 158
 - de Wiener-Lévy 153
 - équivalents 148
 - ponctuels 166
 - stationnaires 150
 - stochastique 147
 - puissance d'un test 260
- Q**
- quantiles 18
 - quantité d'information 218
 - de Fisher 226
 - quasi-certitude 235
- R**
- rapport de corrélation 136, 315, 354
 - de Y en X 28
 - linéaire 28
 - régime stationnaire 174
 - règle de décision 259
 - régression linéaire 356, 358
 - multiple 378
 - résidu studentisé 369
 - risque
 - de deuxième espèce 259
 - de première espèce 259

S

séries chronologiques 148, 351
seuil

de confiance 235

de signification du test 257

simulation 203

statistique 179, 181, 212, 220

d'ordre 195

descriptive 3, 179

exhaustive 213, 226

complète 227

S^2 183, 192

$\bar{X}$ 181, 192

système

à structure parallèle 333

à structure série 332

T

tableau de contingence 26, 128

taux de défaillance 324

temps moyen entre défaillances 322

test

d'exponentialité 288

d'hypothèse 256

de comparaison de pourcentages 297

de Cramer-Von-Mises 287

de Fisher-Snedecor 291

de Kolmogorov-Smirnov 286

de normalité 288

de Smirnov 293

de Student 292

de Wilcoxon 294

du chi-deux 284

non paramétrique de comparaison
293

paramétrique de comparaison 289

théorème

central limite 106, 125

de Bayes 46

de Darmon 215

de l'espérance totale 130

de la variance totale 130

de Neyman et Pearson 264

de Rao-Blackwell 227

trajectoire 147

tribu 39, 51, 440

U

unité statistique 4

V

valeur

atypique 371

prévisionnelle 371

variable de Bernoulli 71

variable(s) aléatoire(s) 49, 59

continue 89

de décision 259, 284

discrète 56, 67

indépendantes 59, 133

variance 19, 62, 339

conditionnelle 130

vecteur ligne transposé 140

vraisemblance d'un échantillon 213

AIDE-MÉMOIRE DE L'INGÉNIEUR

Renée Veyseyre

STATISTIQUE ET PROBABILITÉS POUR L'INGÉNIEUR

2^e édition

Cet aide-mémoire rassemble toutes les définitions, lois et formules du calcul des probabilités et de la statistique utiles à l'ingénieur en activité aussi bien qu'à l'étudiant en formation.

- La première partie donne les principales définitions, et propose un résumé de tous les résultats que l'on peut obtenir à partir d'un tableau de données.
- La deuxième partie donne le vocabulaire du calcul des probabilités et étudie les principales lois discrètes et continues.
- La troisième partie traite des problèmes rencontrés par l'ingénieur dans le domaine de la décision : échantillonnage, estimation et tests d'hypothèse, tests de comparaison, tests d'ajustement, régression.
- La quatrième partie propose un résumé de l'analyse des données.

Cette nouvelle édition a été augmentée d'un chapitre sur la régression multiple.

RENEE VEYSSEYRE

est agrégée
de mathématiques
et professeur honoraire
à l'École centrale
de Paris.



ISBN 2 10 049994 7

L'USINE NOUVELLE

www.dunod.com

