Statistiques pour l'économie et la gestion

Théorie et applications en entreprise

Sous la direction de Claire **Borsenberger**

Frédéric **Bertrand** • Christian **Derquenne** • Gilles **Dufrénot**
Fredj **Jawadi** • Myriam **Maumy**

- Cours illustré
- Exemples concrets
- Applications pratiques
- Exercices et corrigés en ligne

OUVERTURES ÉCONOMIQUES

OFFERT

+ EN LIGNE

- Codes R et jeux de données
- Corrigés des exercices et approfondissements
- PPT des figures réservés aux enseignants

deboeck SUPÉRIEUR

Statistiques pour l'économie et la gestion

Statistiques pour l'économie et la gestion

Théorie et applications en entreprise

Sous la direction de Claire **Borsenberger**

Frédéric **Bertrand** • Christian **Derquenne** • Gilles **Dufrénot**
Fredj **Jawadi** • Myriam **Maumy**

deboeck SUPÉRIEUR

OFFERT : Ressources numériques :
Un *package* (sageR) contenant les codes R issus des exemples et exercices du livre, ainsi que les jeux de données est disponible sur le site du CRAN.

Des ressources complémentaires sont disponibles à l'adresse https://www.deboecksuperieur.com/site/319448 :

Etudiants :
- Le corrigé de tous les exercices.
- Des approfondissements et exercices complémentaires sur certaines parties de l'ouvrage.

Enseignants : Identifiez-vous à l'adresse https://www.deboecksuperieur.com/site/319448 et accédez aux PowerPoint des figures.

Pour plus de détails, cfr page 8.

Pour toute information sur notre fonds et les nouveautés dans votre domaine de spécialisation, consultez notre site web : **www.deboecksuperieur.com**

Copyright illustration de couverture : © BillionPhotos.com - stock.adobe.com

© De Boeck Supérieur s.a., 2021
Rue du Bosquet, 7 – B-1348 Louvain-la-Neuve

Tous droits réservés pour tous pays.
Il est interdit, sauf accord préalable et écrit de l'éditeur, de reproduire (notamment par photocopie) partiellement ou totalement le présent ouvrage, de le stocker dans une banque de données ou de le communiquer au public, sous quelque forme ou de quelque manière que ce soit.

Dépôt légal :
Bibliothèque nationale, Paris : août 2021 ISSN : 2030-501X
Bibliothèque royale de Belgique, Bruxelles : 2021/13647/121 ISBN : 978-2-8073-1944-8

SOMMAIRE

Avant-propos 7

Présentation des auteurs 9

PARTIE 1
Données et statistiques descriptives

CHAPITRE 1. La statistique et la collecte des données 13

CHAPITRE 2. Statistiques descriptives et visualisation des données univariables 37

CHAPITRE 3. Statistiques descriptives et visualisation des données bivariables ou multivariables 83

PARTIE 2
Analyse exploratoire des données

CHAPITRE 4. Analyse en composantes principales (ACP) 119

CHAPITRE 5. Analyse factorielle des correspondances (AFC) 163

CHAPITRE 6. Analyse des correspondances multiples (ACM) 233

CHAPITRE 7. Classification automatique 285

PARTIE 3
Probabilités et inférence statistique

CHAPITRE 8. Théorie des probabilités 357

CHAPITRE 9. Inférence statistique 421

PARTIE 4
Introduction à l'économétrie : modélisation et régression

Chapitre 10. Modèles gaussiens de régressions linéaires simple et multiple 491

Chapitre 11. Introduction aux séries temporelles 567

Chapitre 12. Introduction aux modèles logit et probit 605

Index 639

Bibliographie 649

Quelques tables 653

Table des matières 663

AVANT-PROPOS

Cet ouvrage original est né d'une étroite et riche collaboration entre des enseignants-chercheurs et des experts-chercheurs en entreprise. Leurs parcours et domaines d'expertise sont présentés dans la section suivante.

L'objectif affiché de *Statistiques pour l'économie et la gestion* est de montrer aux étudiants du premier cycle, notamment à ceux des filières économiques, commerciales et de gestion, comment les statistiques contribuent à la compréhension de comportements, d'actions mais aussi de phénomènes naturels observés au quotidien, éclairant les prises de décision de responsables politiques, de chefs d'entreprise, d'investisseurs, de consommateurs, de scientifiques, etc.

L'ouvrage comporte douze chapitres, répartis en quatre parties, traitant de façon simple, claire et progressive les sujets majeurs en statistiques, de la collecte à l'analyse des données et aux estimations et prévisions. Plus précisément, la première partie de l'ouvrage composée de trois chapitres fait le point sur les techniques de planification, de collecte des données et leur analyse « à plat » grâce aux statistiques descriptives et aux techniques de visualisation des données qui sont illustrées par des études sur des populations entières. La deuxième partie va plus loin dans l'analyse des données en les étudiant sous l'angle multivarié et présente quatre méthodes d'analyse exploratoire multidimensionnelle des données, qui ont pour but commun de résumer les données collectées et d'en extraire la substantifique moelle, en étudiant leur variabilité : l'analyse en composantes principales (ACP) de plusieurs variables quantitatives (au chapitre 4), l'analyse factorielle des correspondances (AFC) de deux variables catégorielles (au chapitre 5), l'analyse des correspondances multiples (ACM) applicables à plus de deux variables catégorielles (au chapitre 6) et les méthodes de classification automatique d'individus ou de variables (au chapitre 7). La troisième partie de l'ouvrage est consacrée à une introduction à la théorie des probabilités et à l'inférence statistique utilisée lorsque l'étude complète d'une population n'est pas possible et qu'il faut se contenter d'en étudier un échantillon. Une fois les données résumées et explorées, pour réussir à inférer des résultats au niveau des populations, il est important de pouvoir déterminer si ces données peuvent être considérées comme issues d'une « loi » de probabilité et si oui, laquelle. C'est l'objet du chapitre 8 qui, après avoir exposé des éléments de la théorie des probabilités, présente des lois de probabilités usuelles et leurs propriétés. Celles-ci sont indispensables pour « inférer » des informations supplémentaires sur les variables considérées dans l'étude (chapitre 9) : estimer la « vraie » valeur que peut prendre l'une des caractéristiques d'une variable de la population à partir des informations contenues dans un échantillon (estimation ponctuelle et par intervalle de confiance), effectuer des tests d'hypothèses et analyser la variance. Enfin, la dernière partie détaille les méthodes de régression linéaire (chapitre 10) qui permettent de modéliser un phénomène quantitatif continu à l'aide d'une ou plusieurs variables explicatives dans le but de le modéliser, de l'interpréter, de l'expliquer, de le simuler ou de le prévoir. Les chapitres 11 et 12 s'intéressent plus particulièrement à deux extensions des modèles précédents : les techniques de modélisation de variables observées à différents moments dans le temps (« séries temporelles ») et la modélisation d'un phénomène dont le résultat est une variable qualitative binaire, ordinale ou nominale (modèles dits « logit » ou « probit »).

Fondé sur des développements méthodologiques solides, cet ouvrage constitue une bonne base préparatoire à l'étude de sujets statistiques plus avancés. Certaines parties offrent au lecteur la possibilité d'aller un peu plus loin dans la compréhension des mécanismes à l'œuvre. Ces développements, clairement identifiés dans le texte, peuvent toutefois être facilement mis de côté par le lecteur qui souhaite se concentrer sur l'essentiel. Une bibliographie fournie en fin d'ouvrage permettra, là encore à ceux qui le désirent, d'approfondir les concepts vus dans chaque chapitre.

Enfin, soulignons que l'une des caractéristiques de cet ouvrage est son orientation volontairement marquée vers la pratique et l'application concrète des statistiques. Chaque outil statistique est illustré par des exemples d'application concrets, issus de divers domaines allant de l'économie et de la finance (risque de faillite d'entreprises, taux de rendement des capitaux employés, consommation de lait dans les ménages, prix des actifs, prix du pétrole, etc.) à la mycologie (l'étude des champignons), en passant par l'analyse des résultats du premier tour des élections présidentielles de 2017 en France, des taux de réussite des étudiants aux examens ou de l'évolution du nombre de patients hospitalisés atteints de la COVID-19.

LES RESSOURCES NUMÉRIQUES

Le praticien pourra reproduire toutes les analyses présentées dans l'ouvrage en utilisant le logiciel R et le package sageR en accès direct depuis le CRAN. De plus, il aura accès, sur le site https://fbertran.github.io/sageR/ non seulement aux jeux de données, mais également aux codes de programmation sous R lui permettant de retrouver par lui-même les résultats figurant dans l'ouvrage (issus des exemples ou des exercices). Ce faisant, il s'appropriera le langage R et pourra généraliser l'utilisation de ce logiciel à d'autres jeux de données ou s'entraîner grâce à des exercices supplémentaires.

Le corrigé des exercices est également disponible sur le site de De Boeck Supérieur, à l'adresse https://www.deboecksuperieur.com/site/319448, permettant ainsi au lecteur de vérifier sa bonne compréhension des concepts et des outils développés dans chacun des chapitres. Il y trouvera également des compléments (dont des développements supplémentaires) pour approfondir certaines parties de l'ouvrage.

Enfin, des PowerPoint reprenant les figures de l'ouvrage réservés uniquement aux enseignants sont également accessibles.

REMERCIEMENTS

Nous tenions à remercier très sincèrement l'équipe éditoriale de De Boeck Supérieur, et plus particulièrement Dominique De Raedt et Mireille Raskin sans qui cet ouvrage n'aurait pas pu voir le jour.

Merci à Alain Dessertaine ainsi qu'à tous les relecteurs anonymes pour leurs judicieux conseils qui ont permis d'améliorer la qualité de ce travail. Nous dédicaçons cet ouvrage à Antoine Gautier, professeur à l'Université de Laval, qui a également œuvré à la réalisation de cet ouvrage collectif.

PRÉSENTATION DES AUTEURS

Claire Borsenberger

Docteure en économie, diplômée de la Toulouse School of Economics, Claire Borsenberger a débuté sa carrière au sein du ministère de l'Économie et des Finances, en tant que chargée de mission à la Direction générale du trésor et de la politique économique de 2004 à 2007. En 2007, elle rejoint le groupe La Poste en tant qu'expert économiste à la Direction de la régulation européenne et nationale, poste qu'elle occupera jusqu'en 2011 avant d'être nommée responsable du département « Doctrine et modélisation », au sein de la Direction de la régulation et des études, rattachée à la Direction de la régulation, de la concurrence et des relations institutionnelles. Claire a également été professeure associée à l'Université François Rabelais à Tours, où elle a enseigné l'économie des réseaux et de la concurrence de 2011 à 2017, et continue à dispenser ponctuellement des cours d'économie de la régulation à l'Université de Lorraine. Claire est l'auteure de nombreux articles dans le domaine de la régulation du secteur postal et de l'économie des plateformes numériques.

Frédéric Bertrand

Professeur à l'Université de technologie de Troyes, Frédéric Bertrand est un ancien élève de l'École normale supérieure de Lyon, agrégé de mathématiques, docteur en statistique de l'Université Louis Pasteur de Strasbourg et habilité à diriger des recherches en statistique de l'Université de Strasbourg. Il travaille depuis 20 ans sur la statistique dans tous ses états, de la planification d'expériences à leur analyse à l'aide des modèles appropriés. Il est l'auteur de onze *packages* pour le langage statistique R, de nombreux articles de recherche en particulier sur les plans d'expérience, l'analyse de données ou la régression, plus particulièrement dans le domaine de la biologie, de la médecine ou de l'industrie. Il est également l'auteur de plus de quinze livres sur le langage R, les mathématiques, la statistique et leurs applications. Il enseigne depuis vingt ans la statistique à tous les niveaux de l'enseignement supérieur, licence, master et doctorat, à l'université ou dans des écoles d'ingénieurs, en formation initiale ou continue. Il a été membre du *board of directors* de l'*International Association for Statistical Computing* de 2016 à 2020. Il est membre de la commission *Data Science* de l'Institut de l'Intelligence Artificielle en Santé (IIAS), directeur de MaMOT, nœud du réseau européen Modéliser Simuler et Optimiser lié à EU-MATH-IN, ainsi que co-responsable du Mastère Spécialisé® Big Data : Analytics Avancées pour la Décision.

Christian Derquenne

Docteur en mathématiques appliquées, Christian Derquenne est chercheur senior à la Direction recherche et développement d'Électricité de France. Il travaille dans un département spécialisé en management de l'énergie. Il est l'auteur de nombreux articles dans le domaine de l'analyse de données, notamment en modélisation des données catégorielles, en classification de variables, en régression robuste, dans le traitement des données

manquantes, en fusion statistique des données et en segmentation de séries temporelles. Il enseigne également la statistique en licence et en master 2 à l'université ou dans des écoles d'ingénieurs. Enfin, il s'intéresse au domaine de la statistique et du sport.

Gilles Dufrénot

Professeur à l'Université d'Aix-Marseille, Gilles Dufrénot travaille depuis 20 ans sur la modélisation statistique des données macroéconomiques et financières. Il a enseigné l'économétrie dans les grandes écoles (Ecole Nationale des Ponts et Chaussées) et universités françaises et étrangères, auprès de publics variés (professionnels en formation continue, étudiants, responsables de politique économique). Il est l'auteur de nombreux articles et ouvrages en économie appliquée.

Fredj Jawadi

Docteur en sciences économiques de l'Université de Paris Ouest Nanterre et habilité à diriger des recherches en économétrie de l'Université d'Aix-Marseille, Fredj Jawadi est professeur d'économétrie et de finance à l'Université de Lille et chercheur au LUMEN. *Fellow* de la *Society for Economic Measurement,* il est l'auteur de nombreux articles et ouvrages dans le domaine de l'économétrie appliquée à la finance et la macroéconomie.

Myriam Maumy

Docteure en statistique de l'Université Pierre et Marie Curie et habilitée à diriger des recherches en statistique de l'Université de Strasbourg, Myriam Maumy est maîtresse de conférences hors classe à l'Université de technologie de Troyes. Elle travaille depuis plus de 20 ans sur la statistique et l'échantillonnage. Elle est l'auteure de onze *packages* pour le langage statistique R, de nombreux articles de recherche en particulier sur le sondage, l'analyse de données ou la régression, plus particulièrement dans le domaine de la biologie, de la médecine ou de l'industrie. Elle est également l'auteure de plus de seize livres sur le langage R, les mathématiques, la statistique et leurs applications. Elle enseigne depuis plus de vingt ans la statistique à tous les niveaux de l'enseignement supérieur, licence, master et doctorat, à l'université ou dans des écoles d'ingénieurs, en formation initiale ou continue. Elle est membre du bureau de l'AMIES depuis 2017, membre du *board of directors* de l'*International Association for Statistical Computing* depuis 2020, membre de la commission *Data Science* de l'Institut de l'Intelligence Artificielle en Santé (IIAS), ainsi que co-responsable du Mastère Spécialisé® Big Data : Analytics Avancées pour la Décision.

PARTIE 1

DONNÉES ET STATISTIQUES DESCRIPTIVES

1 La statistique et la collecte des données

2 Statistiques descriptives et visualisation des données univariables

3 Statistiques descriptives et visualisation des données bivariables ou multivariables

CHAPITRE 1

LA STATISTIQUE ET LA COLLECTE DES DONNÉES

1.1. Un peu d'histoire : des premiers recensements au XXIIIe siècle avant J.-C. en Chine à l'ère du *big data* 14

1.2. La terminologie employée dans le domaine des statistiques 16

1.3. Comment collecter des données et constituer un jeu de données ? 22

1.4. Plans d'expériences et de sondage 25

1.5. Traitement pratique des données manquantes 30

1.6. Pour un traitement éthique des données 33

Donner une définition de la statistique n'est pas simple : les définitions de la statistique évoluent en fonction de l'époque ou de son utilisation. Dérivé du substantif latin *status* (« état »), le mot statistique possède, en français comme dans d'autres langues, plusieurs significations distinctes. Nous ferons le point sur ce que sont la statistique et les statistiques dans la première section de ce chapitre.

Toute étude statistique peut être décomposée en trois phases au moins : la planification et l'organisation de la collecte des données, la collecte proprement dite des données, et leur analyse et/ou leur interprétation. Nous nous intéressons dans ce chapitre aux deux premières phases, qui ne doivent pas être négligées sous peine de fausser l'analyse et d'aboutir à une interprétation erronée de la réalité.

La collecte des données peut être réalisée soit par la simple observation des phénomènes auxquels on s'intéresse tels qu'ils se produisent naturellement, soit par l'expérimentation, c'est-à-dire en provoquant volontairement l'apparition de certains phénomènes contrôlés. Nous verrons qu'il existe différents types de données.

Selon le type de données collectées, les méthodes d'analyse applicables changent afin de mieux prendre en compte les propriétés statistiques de ces données. L'application de ces méthodes suppose également que les données respectent un certain nombre de critères. Il est donc essentiel de réfléchir au préalable à la manière dont les données sont collectées. La collecte doit ainsi être organisée. Elle nécessite la mise en œuvre d'un plan de sondage ou d'expériences rigoureux. Enfin, la collecte des données doit respecter des règles éthiques ou déontologiques.

Avant d'aborder ces différents points, revenons en arrière.

1.1. UN PEU D'HISTOIRE : DES PREMIERS RECENSEMENTS AU XXIII^e SIÈCLE AVANT J.-C. EN CHINE À L'ÈRE DU *BIG DATA*

Bien que le terme de statistique soit relativement récent (on attribue l'origine du mot aux travaux de G. Achenwall (1719-1772) relatifs à la connaissance de l'État, *Staatskunde*), cette activité qualifiée « d'arithmétique politique » semble exister dès la naissance des premières structures sociales. Les premières traces écrites de recensement du bétail ou de la population datent du XXIII^e siècle avant J.-C. en Chine et en Égypte.

Au XVIII^e siècle de notre ère, on voit apparaître le rôle prévisionnel des statistiques, avec la construction des premières tables de mortalité par A. Deparcieux en 1746[1] et les calculs rétrospectifs de la population gauloise, à partir des données chiffrées laissées par Jules César dans sa conquête des Gaules, par le Baron de Reiffenberg en 1835[2].

Parallèlement se développe, en France tout d'abord, le calcul des probabilités, dans le but de résoudre des problèmes relatifs aux jeux de hasard, avec notamment les travaux de B. Pascal (1623-1662), A. de Moivre (1667-1754), D. Bernoulli (1700-1782), P. de Laplace (1749-1827), C. Gauss (1777-1855) et S. Poisson (1781-1840).

L. Quetelet, mathématicien et physicien belge (1796-1874), réunit ces deux disciplines et consacre la naissance de la science statistique basée sur les probabilités.

Le XIX^e siècle voit l'activité statistique appliquée aux sciences politiques et sociales prendre son plein essor. Des règles précises sur la collecte et l'interprétation des données sont édictées. La première application industrielle des statistiques a lieu lors du recensement américain de 1890, qui met en œuvre la carte perforée inventée par le statisticien H. Hollerith.

1 Deparcieux, A. (1746), *Essai sur les probabilités de la durée de vie humaine*.

2 De Reiffenberg, F. (1835), *Essai sur la statistique ancienne de la Belgique. I. Population. – II. Architecture. – III. Mobilier, Costumes*.

La première moitié du XX^e siècle est essentiellement marquée, dans le domaine statistique, par le développement de méthodes de plus en plus nombreuses et par l'utilisation de ces méthodes dans des secteurs d'application de plus en plus diversifiés[3] : dans les sciences biologiques et psychologiques, donnant naissance respectivement à la biométrie et la psychométrie ; dans le secteur de la recherche pharmaceutique ou médicale ; en génétique et biologie moléculaire ; en agronomie ; en sciences de gestion ; en économie, donnant naissance à l'économétrie ; dans le domaine industriel, dans le but de maîtriser les processus de production ou de contrôler la qualité des produits ; en physique et sciences de l'ingénieur ; en météorologie et climatologie ; en géologie et géographie ; en écologie et environnement ; en chimie ; en sociologie ; en sciences de l'éducation, et aussi dans le secteur des assurances, de l'archéologie, de la linguistique, du droit, etc.

L'avènement de l'informatique, dans les années 1940 aux États-Unis, puis dans les années 1960 en Europe, a permis de traiter un plus grand nombre de données, d'effectuer plus rapidement et plus facilement des calculs parfois complexes. Comme le souligne P. Dagnelie (2013)[4], « l'analyse multidimensionnelle ou analyse statistique à plusieurs variables, c'est-à-dire l'étude simultanée de plus de deux caractéristiques des différents individus considérés, constitue un exemple typique de méthodes très largement antérieures à la venue de l'ordinateur, mais ayant connu une expansion considérable à partir des années 1960, du fait des nouveaux moyens de calcul disponibles. Les méthodes du *jackknife* et du *bootstrap* ou, d'une manière plus générale, de ré-échantillonnage se sont par contre entièrement développées en raison même de l'existence de l'ordinateur » (p. 21).

Aujourd'hui, les développements en informatique et la technologie numérique ouvrent de nouvelles perspectives aux statisticiens. La technologie numérique permet la « mise en données du monde » : la plupart de nos actions, déplacements, communications, traitements médicaux, démarches administratives, achats, génèrent une masse de données inédite. Les données produites par les objets connectés produiront des informations sur nos agissements, que nous en ayons conscience ou non. Les capacités de stockage de ces données ont explosé (et les coûts de stockage ont fortement diminué). De nouvelles méthodes d'analyse ont été créées, permettant de croiser des données (y compris non structurées).

C'est ainsi qu'est né le *big data* : la capacité de la société de l'information à collecter, stocker et traiter en temps quasi réel des masses de données plus ou moins structurées. À cette notion de *big data* étaient souvent initialement associés les cinq V : volume, variété, vélocité, véracité et valeur. Le volume fait naturellement référence à la quantité d'informations ou de données collectées ; la variété renvoie à cette même notion de quantité mais cette fois, sous un aspect plus qualitatif et fait référence à l'hétérogénéité des données collectées, que ce soit dans leur format ou dans leur nature. La vélocité a trait au traitement de cette masse colossale de données et plus précisément, elle caractérise la rapidité à laquelle ces données sont traitées, potentiellement en temps réel. La véracité est un enjeu majeur relatif à la fiabilité des données exploitées, comme nous le verrons plus loin. La valeur du *big data* est intrinsèquement liée aux quatre autres V : le *big data* génère de la valeur en transformant des données potentiellement hétérogènes en une information disponible instantanément et fiable. Aujourd'hui, cinq V supplémentaires viennent s'ajouter à ces caractéristiques : la variabilité (liée à la multiplicité des sources et des formats), la validité (qui fait référence à la précision et à la nécessaire phase de « nettoyage » des données), la vulnérabilité (en lien avec les failles de sécurité), la volatilité (en lien avec leur caractère « périssable ») et la visualisation des données[5].

Selon ses défenseurs, le *big data* offre des perspectives sans précédent pour les acteurs économiques, les pouvoirs publics, les consommateurs, bref, l'ensemble de la société : elle permet d'affiner la connaissance

3 Cf. les travaux de K. Pearson (1857-1936), C. Spearman (1863-1945), G. Yule (1871-1951), W. Sealy Gosset (1876-1937) qui publia ses travaux sous le pseudonyme de Student, ou Aylmer Fisher (1890-1962).

4 Dagnelie, P. (2013), *Statistique théorique et appliquée. Tome 1. Statistique descriptive et bases de l'inférence statistique*, De Boeck, Bruxelles.

5 Cf. https://analyticsinsights.io/les-10-v-du-big-data/.

des comportements des acteurs économiques, de leurs préférences, voire de leurs attentes, permettant ainsi de mieux adapter l'offre à la demande, d'optimiser les processus de production, de mieux gérer les flux (de marchandises, de personnes, d'énergie, etc.) et de définir des politiques publiques plus efficientes. La prolifération exceptionnelle des données est vue comme une opportunité pour améliorer et enrichir la production d'informations et donc, *in fine*, la prise de décisions.

À côté des nombreuses applications potentiellement bénéfiques que rend possible la mise en données du monde[6], elle n'en reste pas moins une source de risques. Certains dénoncent les conséquences négatives que peut générer la concentration de ce pouvoir entre les mains de quelques opérateurs dominants. D'autres avancent le risque de censure, de traitement inégalitaire ou de discrimination dont certaines personnes pourraient être victimes du fait du traitement de leurs données par des algorithmes biaisés. D'autres encore, le risque d'uniformisation de la pensée, de perte de son libre arbitre ou du principe d'auto-détermination, de dictature de la donnée et des algorithmes prédictifs (*fake news* et ingérence dans les campagnes électorales), de biais de sous- et sur-représentations de certaines catégories de personnes[7].

Tout ceci laisse entrevoir de nombreux et passionnants défis méthodologiques et technologiques à relever, pour vous, statisticiens de demain, afin d'exploiter au mieux cette masse de données croissante, pour inférer, aider à la décision et prédire !

1.2. LA TERMINOLOGIE EMPLOYÉE DANS LE DOMAINE DES STATISTIQUES

Avant d'aller plus loin, commençons par nous mettre d'accord sur les termes employés par les statisticiens et que nous retrouverons dans le reste de l'ouvrage.

1.2.1 *Les différentes définitions de la et des statistiques*

Le Larousse donne quatre définitions du mot « statistique ».

- En premier lieu, le terme « statistiques », utilisé le plus souvent au pluriel, désigne un « ensemble de données d'observation relatives à un groupe d'individus ou d'unités ». On parle par exemple des « statistiques de la production manufacturière » lorsque l'on évoque des données sur les quantités produites, les prix de vente, les coûts de production, la main-d'œuvre employée, etc. ; des « statistiques démographiques » lorsque l'on traite de données relatives à la natalité, à la pyramide des âges, etc. ; ou encore des « statistiques du chômage ».
- En second lieu, le mot « statistique » (utilisé généralement au singulier) désigne « l'ensemble des méthodes qui ont pour objet la collecte, le traitement et l'interprétation de ces données ». Notée avec une majuscule, la Statistique désigne la science statistique en tant que discipline qui vise à comprendre, prévoir et simuler en s'appuyant sur des règles et des méthodes communes en matière de collecte des données et d'interprétation des données afin que celles-ci puissent constituer une aide à la décision. La science statistique fait partie, pour certains, des mathématiques appliquées. Pour d'autres,

6 Par exemple, dans le domaine de la santé, le partage et l'accès aux données produites par divers acteurs, publics et privés, permettent de constituer et d'étudier plus rapidement des cohortes de patients, de croiser les données de santé avec des données environnementales et d'établir de nouveaux liens de causalité ouvrant la voie à de nouvelles politiques publiques en matière de prévention ou de nouveaux traitements. Dans le secteur agricole, l'analyse des données massives collectées grâce à des capteurs mis en place sur les tracteurs permet d'optimiser la culture des sols. La ville connectée a pour objectif de rendre la gestion de la ville plus efficace, moins énergivore.

7 Par exemple, les personnes en grande difficulté économique sont sous doute sous-représentées dans les bases de données issues des usages de la téléphonie mobile. *A contrario*, les jeunes sont certainement sur-représentés si on utilise des informations issues de Facebook.

il s'agit d'une science à part entière : la science des données (*data science* en anglais) qui possède une composante théorique et une composante appliquée. La composante théorique s'appuie sur la théorie des probabilités et forme avec cette dernière, les sciences de l'aléatoire. La statistique appliquée est utilisée dans presque tous les domaines de l'activité humaine : ingénierie, management, économie, biologie, informatique, etc.

- En troisième lieu, proche de la première définition, le Larousse définit les statistiques comme « l'ensemble des données numériques concernant un phénomène quelconque et dont on tire certaines conclusions ».
- En quatrième lieu, le terme « statistique » désigne une « variable aléatoire, fonction des observations, construite à partir d'un échantillon. Ainsi, la moyenne d'un échantillon est une statistique. »

Dans cet ouvrage, nous retiendrons les définitions suivantes :

Définition 1.1 La statistique

La statistique désigne à la fois un ensemble d'outils et l'ensemble des activités consistant à traiter ces données, à les interpréter et à les analyser, afin de les rendre compréhensibles par tous. C'est à la fois une science, une méthode et un ensemble de techniques.

Définition 1.2 Les statistiques

Les statistiques désignent un ensemble de données numériques qui interviennent pratiquement dans tous les domaines d'activité comme dans la gestion financière (des États, des banques, des compagnies d'assurances, des entreprises, etc.), dans la démographie, le contrôle de la qualité, les études de marché, les sciences expérimentales (biologie, psychologie, etc.), etc.

Définition 1.3 La statistique descriptive

La statistique descriptive correspond au traitement des données, pour en dégager un certain nombre de renseignements qualitatifs et/ou quantitatifs à des fins de comparaison par exemple.

Définition 1.4 La statistique inductive ou inférentielle

Un autre but de la statistique est d'extrapoler, à partir d'un échantillon (cf. définition 1.7) de la population à étudier (cf. définition 1.5), le comportement de la population dans son ensemble (sondages, contrôle de qualité comportant un test destructif, etc.). C'est ce que l'on appelle **la statistique inductive**, ou encore **la statistique inférentielle**.

1.2.2 *Population – Individus – Échantillon*

Définition 1.5 Population

L'ensemble sur lequel porte l'activité statistique s'appelle la **population**. La population est définie comme un ensemble, fini ou non, d'éléments à étudier. Ces éléments sont appelés **individus** ou **unités statistiques**. Elle est généralement notée Ω pour rappeler la notation des probabilités mais par exemple, dans la théorie des sondages, elle est notée U, U comme « Univers ».

Remarques 1.1

1) Le terme de « population » est employé aussi bien lorsqu'il s'agit d'un ensemble d'êtres vivants que d'un ensemble d'objets inanimés.
2) Il est recommandé de toujours bien définir la population étudiée.
3) Le nombre d'éléments contenus dans Ω ou le « cardinal » de la population est généralement noté N : $\text{Card}(\Omega) = N$ ou encore $|\Omega| = N$.

Exemple

Plusieurs caractères, définis par la suite, ont été recensés sur l'ensemble des pays européens. **L'ensemble des pays européens** constitue alors la **population étudiée**.

Définition 1.6 Individus ou unités statistiques

Les éléments qui constituent la population sont appelés les **individus** ou encore les **unités statistiques**. Un individu est noté ω lorsque la population est notée Ω ou u lorsque la population est notée U.

Remarque 1.2

Ces « individus » peuvent être de natures très diverses.

Exemples

Étudiants à l'université, mois d'une année, pièces produites par une usine, résultats d'expériences répétées un certain nombre de fois.

Définition 1.7 Échantillon

Un **échantillon**, noté généralement S, comme *sample* (« échantillon » en anglais), est une partie de la population prélevée soit de façon aléatoire, soit de façon déterministe. Ce sous-ensemble de la population considérée a normalement les mêmes caractéristiques statistiques que la population. Si l'échantillon est représentatif, il sera possible d'effectuer des analyses et de déduire des conclusions valables pour la population.

Les trois notions de population, d'individus et d'échantillon sont représentées sur la figure 1.1.

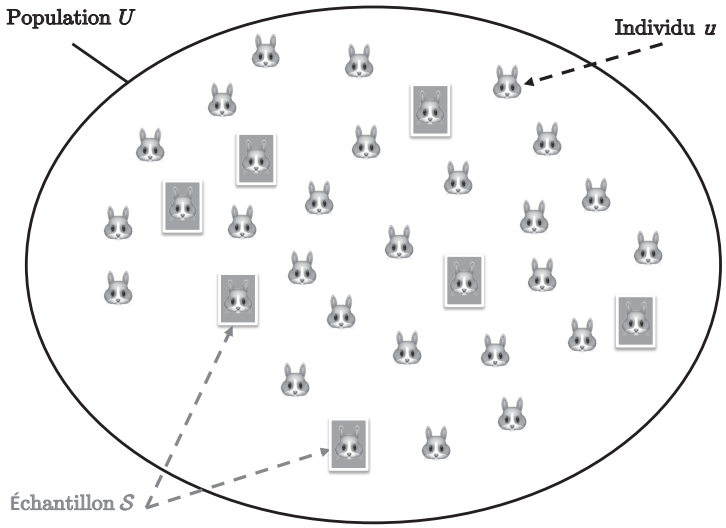

Figure 1.1 : *population, échantillon et individu*

1.2.3 *Observations – Caractères – Données brutes*

Définition 1.8 Observations

Le statisticien fait des relevés sur les unités statistiques : ce sont les **observations.**

Définition 1.9 Caractères

Les caractéristiques étudiées sur les individus d'une population sont appelées les **caractères.** Un caractère est une application χ d'un ensemble Ω fini de cardinal N (la population), dans un ensemble C (l'ensemble des valeurs possibles du caractère), qui associe à chaque individu ω de Ω la valeur $\chi(\omega)$ que prend ce caractère sur l'individu ω. Le caractère sert à classer un certain nombre de sous-ensembles pour définir des variables statistiques.

Exemples

Sexe, âge, profession, salaire, ancienneté. ■

Nous considérons deux types de caractères :

i. **Les caractères quantitatifs**. Leur détermination produit un nombre ou une suite de nombres.

   Nous distinguons à l'intérieur de ce type de caractères encore deux catégories.

- **Les caractères simples** ou **univariables**. Leur mesure sur un individu produit un seul nombre. Un caractère quantitatif simple est dit **continu** si l'ensemble C des valeurs possibles du caractère est un intervalle. Un caractère quantitatif simple est dit **discret** si l'ensemble C des valeurs possibles du caractère est discret.

Exemples

Taille d'un individu, masse d'un individu, salaire d'un individu, température extérieure, etc. ■

- **Les caractères multiples** ou **multivariables.** Leur mesure sur un individu produit une suite de cardinal fini de nombres. L'ensemble de leurs valeurs est donc $\mathbb{R}^n$ ou une partie de $\mathbb{R}^n$.

Exemples

Relevé de notes d'un étudiant à l'université, bulletin de salaire, etc. ■

ii. **Les caractères qualitatifs** simples ou multiples.

Exemples

Profession, adresse, sexe, numéro de téléphone, etc. ■

La distinction entre les deux types de caractères est très importante. En effet, les méthodes d'analyse d'une population diffèrent suivant la nature du caractère étudié. Par exemple, les représentations graphiques ne sont pas les mêmes.

Remarques 1.3

1) Un caractère qualitatif peut toujours être transformé en caractère quantitatif par codage. C'est ce qui se fait le plus généralement. Mais un tel codage est purement conventionnel et n'a pas de sens quantitatif. Par exemple, nous ne pouvons pas **calculer le sexe moyen.**
2) Certains caractères qualitatifs s'expriment à l'aide de nombres. Mais il n'en reste pas moins vrai qu'ils n'ont pas de sens quantitatif. Par exemple, **calculer un numéro de téléphone moyen** n'a pas de sens.

Définition 1.10 Données brutes

La suite des valeurs $\chi(\omega)$ prises par χ lorsque ω décrit toute la population Ω s'appelle les **données brutes**. C'est une suite finie $X_1, \ldots, X_N$ d'éléments de l'ensemble C des valeurs possibles du caractère. Ces valeurs X_i ne sont pas nécessairement distinctes : un caractère peut prendre la même valeur sur deux individus différents.

Exercice 1.1

Le tableau 1.1 représente le nombre de patients testés positifs à la Covid-19 et hospitalisés le 8 février 2021, par région et par classe d'âge, en valeur absolue.

Tableau 1.1 : *nombre de patients hospitalisés*

Région	Tous âges	0-9	10-19	20-29	30-39	40-49	50-59	60-69	70-79	80-89	90+
Ile-de-France	5 474	17	27	72	128	192	521	861	1 222	1 607	783
Auvergne-Rhône-Alpes	3 779	1	9	15	47	112	263	571	989	1 188	541
Provence-Alpes-Côte d'Azur	3 531	4	7	23	42	91	246	546	939	1 088	527
Grand-Est	2 782	2	3	13	51	69	192	384	679	939	426
Hauts-de-France	2 697	2	8	20	39	81	188	432	642	856	410
Occitanie	1 904	3	1	6	20	52	154	287	473	624	282
Bourgogne-Franche-Comté	1 757	1	7	19	19	34	94	237	432	596	300
Nouvelle-Aquitaine	1 570	6	3	10	16	36	102	214	375	533	265
Normandie	1 361	1	2	11	19	32	77	218	315	436	239
Centre-Val de Loire	1 074	0	0	8	10	17	85	155	277	318	184
Pays de la Loire	1 013	0	0	6	12	21	46	116	209	400	203
Bretagne	755	2	4	6	7	24	51	97	163	276	124
Mayotte	105	1	5	16	14	17	16	18	12	3	0
Corse	56	0	0	1	0	0	3	12	9	16	15
Guadeloupe	51	0	0	0	0	2	7	6	16	10	7
La Réunion	40	0	1	1	3	4	8	10	7	5	1
Guyane	34	0	0	2	3	4	3	15	5	2	0
Martinique	12	0	0	0	0	1	2	2	5	1	0
Total	27 995	40	77	229	430	789	2 058	4 181	6 769	8 898	4 307

Le tableau 1.2 représente le nombre de patients testés positifs à la Covid-19 et placés en réanimation au 8 février 2021, par région et par classe d'âge, en valeur absolue.

Tableau 1.2 : *nombre de patients en réanimation*

Région	Tous âges	0-9	10-19	20-29	30-39	40-49	50-59	60-69	70-79	80-89	90+
Ile-de-France	730	2	2	12	21	41	139	211	223	71	4
Provence-Alpes-Côte d'Azur	451	0	1	5	7	21	60	158	159	33	6
Auvergne-Rhône-Alpes	393	0	1	1	6	20	51	103	169	32	2
Grand-Est	327	0	0	2	4	22	36	115	109	35	2
Hauts-de-France	302	0	2	1	5	11	39	111	100	26	2
Occitanie	301	0	0	0	7	13	43	91	95	42	9
Nouvelle-Aquitaine	215	2	0	1	4	6	25	68	79	27	1

Région	Tous âges	0-9	10-19	20-29	30-39	40-49	50-59	60-69	70-79	80-89	90+
Bourgogne-Franche-Comté	159	0	1	2	3	5	15	42	69	16	5
Centre-Val de Loire	135	0	0	1	0	6	26	26	54	12	3
Normandie	133	0	0	3	8	4	16	41	49	10	1
Pays de la Loire	99	0	0	1	2	6	14	34	34	8	0
Bretagne	52	1	1	0	2	3	7	15	19	4	0
Mayotte	21	0	1	6	1	5	4	3	0	0	0
Guyane	12	0	0	0	1	1	1	5	3	1	0
La Réunion	9	0	0	0	0	2	3	1	3	0	0
Guadeloupe	6	0	0	0	0	0	2	2	2	0	0
Corse	5	0	0	0	0	0	0	1	2	1	1
Martinique	3	0	0	0	0	1	0	0	1	0	0
Total	3 353	5	9	35	71	167	481	1 027	1 170	318	36

Pour chaque tableau, répondre aux questions suivantes :

1) Quelle est la population étudiée ?
2) Quels sont les individus ?
3) Quels sont les caractères étudiés ?
4) Quelle est la nature des caractères ?

1.3. COMMENT COLLECTER DES DONNÉES ET CONSTITUER UN JEU DE DONNÉES ?

1.3.1 *Enquête, recensement, sondage*

Définition 1.11 Enquête

Dans le domaine statistique, on appelle **enquête** ou, parfois, **inventaire**, l'ensemble des opérations qui ont pour but de collecter de façon organisée des informations relatives à un groupe d'individus ou d'éléments, observés dans leur milieu ou dans leur cadre habituel.

Les individus ou les éléments en question, également appelés **unités de base** ou **unités statistiques**, peuvent être aussi bien des personnes humaines que des animaux, des plantes, des groupes de personnes (familles, ménages, etc.), des groupes d'animaux, des groupes de plantes, ou des éléments de toute autre nature (entreprises industrielles ou commerciales, exploitations agricoles, machines d'un type donné, etc.). L'ensemble des unités auxquelles nous nous intéressons est appelé **population**, **univers** ou **ensemble statistique**.

Définition 1.12 Recensement

Quand toutes les unités statistiques de la population considérée sont effectivement observées individuellement, l'**enquête** est dite **complète** ou **exhaustive**. Elle est alors appelée **recensement**.

Le recensement dresse un panorama d'une réalité sociale, par dénombrement : nous comptons le nombre d'individus possédant telle ou telle caractéristique. Il s'agit donc d'une méthode quantitative. Néanmoins, c'est une approche relativement lourde dans la mesure où pour obtenir un recensement valide, il importe de prendre en compte toute la population de référence.

Définition 1.13 Sondage

Quand, pour réduire l'importance du travail de collecte des données, une partie seulement des individus ou des éléments de la population sont réellement observés, l'**enquête** est dite **partielle** ou **par échantillonnage**. Elle est également appelée parfois **sondage**.

La partie de la population qui est réellement observée constitue l'**échantillon**, et la constitution de cet échantillon relève de ce qui est appelé l'**échantillonnage**. Dans ce cas, la représentativité de la population étudiée joue un rôle déterminant dans la constitution de l'échantillon. Ce point sera abordé dans les sections suivantes.

1.3.2 *L'expérimentation*

Définition 1.14 Expérimentation

Contrairement au cas de l'observation par enquête, l'**expérimentation**, c'est-à-dire la réalisation d'une ou plusieurs expériences ou d'un ou plusieurs essais, suppose que l'apparition des faits que nous désirons étudier est **volontairement provoquée**, dans des conditions que nous maîtrisons au moins partiellement.

L'expérimentation peut, dans certains cas, être plus efficace que la simple observation par enquête. C'est notamment le cas dans le domaine biologique.

Comme toute enquête, exhaustive ou par échantillonnage, toute expérience doit faire l'objet d'une préparation et d'une planification minutieuse. D'une façon générale, les questions qu'il faut examiner en élaborant un plan d'expérience ou un protocole expérimental sont la définition du but et des conditions de l'expérience, la définition des facteurs que nous désirons étudier, la définition des unités expérimentales, la définition des observations à réaliser, et la définition du dispositif expérimental. Nous examinerons ces différents points un peu plus loin.

1.3.3 *Les méta-études et études théoriques*

Définition 1.15 Méta-études

Les **études « méta »** correspondent à une **analyse de l'ensemble des études existantes** dans un domaine bien précis.

Il s'agit non seulement de recenser les études existantes, d'en synthétiser le contenu mais aussi de procéder à une critique épistémologique poussée, de mettre en perspective l'ensemble des données en fonction du bien-fondé des dispositifs méthodologiques qui sont à leur origine. C'est donc un examen synthétique des contenus et des méthodes qui remet les observations dans leurs contextes. Les méta-études sont d'autant plus pertinentes qu'il existe déjà beaucoup d'études et qu'elles semblent aboutir à des observations ou conclusions différentes.

1.3.4 *Les nouvelles sources de données issues de la numérisation du monde : approche participative, « moissonnage du Web », internet des objets*

Certains instituts de statistique nationaux ont développé de nouvelles manières de collecter des données à des fins statistiques.

L'une d'elles, l'**approche participative**, consiste à faire appel à une grande communauté d'utilisateurs pour recueillir des renseignements. Elle repose sur le principe selon lequel chaque citoyen est un expert dans son milieu. L'approche participative permet aux instituts de statistique nationaux de valider les données avec des sources de données complémentaires afin d'assurer des résultats de bonne qualité.

Par exemple, Statistique Canada, l'institut de statistiques canadien, a lancé en juillet 2016 un projet pilote (qui a pris fin en mars 2018) afin de mieux appréhender le potentiel de cette approche participative à la collecte des données à des fins statistiques, en utilisant *OpenStreetMap* comme plateforme. Les contributeurs étaient invités à fournir des renseignements géographiques sur des immeubles non résidentiels dans la région d'Ottawa (Ontario) et de Gatineau (Québec). Le réseau qui a été établi et l'expérience acquise dans le cadre de ce projet pilote ont contribué à lancer l'initiative « Bâtir le Canada 2020 », qui vise à cartographier, sur *OpenStreetMap,* tous les immeubles du Canada d'ici 2020. Plus récemment, Statistique Canada a lancé un projet novateur d'approche participative sur le prix du cannabis avant sa légalisation. L'institut demande aux particuliers, contributeurs volontaires, de fournir le prix de leur dernier achat de cannabis séché ; ces renseignements sont recueillis de façon anonyme et diffusés au public au moyen d'une application spécialisée et d'une page Web. Cette initiative contribue à recueillir des renseignements sur un marché relativement nouveau et permet de surveiller les prix de façon confidentielle et non intrusive.

Statistique Canada, comme d'autres instituts nationaux de statistique, étudient également la possibilité d'automatiser la collecte de renseignements à partir de sites Web (« **moissonnage du Web** »), à des fins statistiques et de recherche. En particulier, de plus en plus d'instituts nationaux de statistique utilisent des données moissonnées sur le Web pour produire des indices des prix à la consommation, comme le souligne J. Mehrhoff (2019)[8].

À titre d'exemple, nous pouvons citer l'initiative du *Centraal Bureau voor de Statistiek* (CBS) néerlandais qui a commencé à moissonner les prix de vente et les méta-données d'un grand magasin néerlandais de vente en ligne de vêtements et de chaussures en octobre 2012 (cf. travaux d'A. Chessa et R. Griffioen (2019)[9] réalisés à partir de ces données).

L'**internet des objets** et les différents **capteurs connectés à Internet** qui se répandent dans les villes, dans les transports en commun, dans les véhicules, les logements, etc. sont également de nouvelles sources de données massives que les statisticiens cherchent à intégrer aux systèmes de production statistique pour pouvoir produire des statistiques en « temps réel » et « automatiquement »[10]. Par exemple, entre 2012 et 2014, une étude a été réalisée par Eurostat pour évaluer la faisabilité d'utiliser les données mobiles de localisation pour établir les statistiques du tourisme[11].

8 Mehrhoff, J. (2019), « La chaîne de valeur des données de caisse et des données moissonnées sur le Web », *Économie et Statistique*, n° 509.

9 Chessa A. et R. Griffioen (2019), « Comparaison des indices de prix des vêtements et des chaussures à partir de données de caisse et de données moissonnées sur le Web », *Économie et Statistique*, n° 509.

10 « Statistiques intelligentes », groupe de travail d'Eurostat sur le *big data*. Projet de document, octobre 2016.

11 Cette étude, qui préfigurait les travaux actuels d'Eurostat dans le domaine des mégadonnées (*big data*), visait à répondre à la plupart des questions techniques, juridiques et qualitatives liées à l'utilisation, pour les statistiques officielles, des données conservées par les opérateurs de réseau mobile.

1.4. PLANS D'EXPÉRIENCES ET DE SONDAGE

La statistique est apparue en premier lieu comme la science de la description des populations. Malheureusement, cette statistique descriptive ne peut être utilisée que de manière exceptionnelle dans le monde réel puisqu'une connaissance complète des populations étudiées n'est que très rarement atteinte. Les plans d'expériences et de sondage servent à organiser le prélèvement d'individus au sein d'une population lorsqu'il n'est pas possible d'observer exhaustivement celle-ci.

Comme nous le verrons dans les chapitres suivants, et en particulier au chapitre 9, la nature et la qualité des traitements statistiques qui peuvent être appliqués à l'échantillon (notamment la précision des estimateurs et des estimations, la forme des modèles qu'il sera possible d'utiliser, ou encore les caractéristiques des populations qu'il sera possible d'étudier), dépendent de la manière dont ces données auront été collectées.

Ainsi, pourquoi ne pas choisir la méthode de collecte des données qui apportera la meilleure précision possible compte tenu des contraintes de mise en œuvre d'une expérience ? De même, pourquoi ne pas s'assurer que les modèles envisagés pour décrire les phénomènes à modéliser pourront être ajustés sans ambiguïté aux données ?

Ainsi la mise au point de la méthodologie de recueil des données est une étape cruciale qui doit être réalisée préalablement à l'analyse des données. En effet, une fois que l'expérience aura été réalisée, il sera trop tard pour pouvoir intervenir, comme l'a fait remarquer R.A. Fisher, l'un des fondateurs de l'utilisation des statistiques en recherche expérimentale, en 1938 :

« To consult the statistician after an experiment is finished is often merely to ask him to conduct a post mortem examination. He can perhaps say what the experiment died of »[12].

Il existe deux grandes méthodologies pour planifier la collecte des données :

- la **planification des expériences** ;
- la **planification des sondages**.

La différence entre ces deux méthodologies réside dans la manière dont l'aléatoire est mis en œuvre. Un plan d'expériences précise comment affecter aléatoirement un échantillon d'unités connues et déjà prélevées dans une population à des situations contrôlées alors qu'un plan de sondage précise comment choisir, généralement au hasard, des unités au sein d'une population.

1.4.1 *La planification des expériences*

Définition 1.16 Planification des expériences

La **planification des expériences** est l'art d'organiser des expériences contrôlées, c'est-à-dire des expériences pour lesquelles il est possible de choisir à l'avance, et exactement, les conditions de l'expérience pour chaque unité statistique ou individu.

12 Allocution lors du premier congrès de statistique en Inde. « Demander l'avis d'un statisticien une fois l'expérience terminée équivaut à un examen post-mortem. Il peut peut-être dire de quoi l'expérience est morte ».

Les premiers travaux de R.A. Fisher, consacrés à l'organisation des expériences en agronomie, remontent aux années 1920 et les 14 années de recherche qu'il a consacrées à ce sujet sont reprises dans son ouvrage de 1935[13] qui est mondialement connu. En 1931, R.A. Fisher préconise le respect de trois principes :

- **réplication** ;
- **randomisation** ;
- **contrôle local de la variabilité**.

Leurs relations sont détaillées dans la figure 1.2.

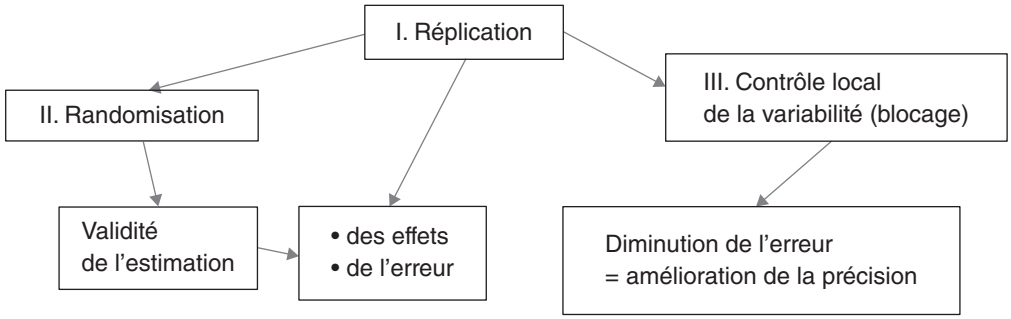

Figure 1.2 : *relations entre les principes applicables aux expériences*

I. La **réplication** de certaines des conditions de l'expérience a un objectif double :

- permettre une estimation de la variabilité résiduelle, la variabilité qui n'est pas liée à la variation des conditions de l'expérience ;
- augmenter la précision de l'expérience.

II. La **randomisation** des expériences, c'est-à-dire l'attribution au hasard des conditions de l'expérience aux différentes unités expérimentales, permet d'obtenir une estimation non biaisée de

- la variabilité résiduelle et
- des effets estimés.

III. Le **contrôle local de la variabilité**, encore appelé **blocage**, sert à augmenter la précision de l'expérience en ayant recours à des plans d'expériences particuliers comme les **plans en blocs complets** ou **incomplets**, les **carrés latins**, les **carrés gréco-latins**, les ***split-plots***, les ***cross-over,*** les ***carry-over,*** etc.

13 Fisher, R.A. (1935), *The design of experiments.* Edinburgh, Oliver and Boyd.

A. Exemples de plans d'expériences

a) Peser dix hommes.

b) Peser dix hommes qui mesurent 1,80 m.

c) Une entreprise de fabrication de gâteaux souhaite tester quatre emballages différents pour de nouveaux gâteaux pour les enfants (cf. tableau 1.3). 20 magasins, avec des volumes de vente à peu près égaux, ont été sélectionnés comme unités expérimentales. Chaque magasin a reçu au hasard un des modèles d'emballage. Il s'agit d'**un plan d'analyse de la variance à un facteur**.

Tableau 1.3 : *exemple de plan d'analyse de la variance à un facteur*

Emballage (*i*)	Magasin (*j*)				
	1	2	3	4	5
1	11	17	16	14	15
2	12	10	15	19	11
3	23	20	18	17	0
4	27	33	22	26	28

d) **Exemple de carré latin** pouvant servir à étudier des ventes de matériel. Un fabricant a mené une petite étude pilote sur l'effet du prix de l'un de ses produits sur les ventes de ce produit dans les quincailleries. Étant donné qu'un changement de prix répété au sein d'un même magasin pourrait être source de confusion pour les clients, un seul prix a été utilisé pour chaque magasin pendant les six mois de l'étude. Afin de réduire la variabilité des erreurs expérimentales, les magasins ont été choisis de manière à ce qu'il y en ait un et un seul pour chaque combinaison de classe de volume de vente (lignes *i*) et de localisation géographique (colonnes *j*). Les quatre niveaux de prix (A : 1,25 € ; B : 1,35 € ; C : 1,45 € ; D : 1,55 €) ont été attribués aux magasins selon le modèle du carré latin illustré par le tableau 1.4. Cette attribution correspond aux lettres entre parenthèses dans les cellules du tableau. Les données sur les ventes au cours de la période de six mois (en milliers d'euros) figurent dans le tableau 1.4.

Tableau 1.4 : *exemple de plan d'expériences en carré latin*

Classe de volume de vente (*i*)	Classe de localisation géographique (*j*)			
	Nord-Est	Nord-Ouest	Sud-Est	Sud-Ouest
1 (plus petit)	2,4 (B)	3,0 (C)	2,0 (A)	3,4 (D)
2	2,8 (A)	3,8 (D)	3,2 (B)	3,0 (C)
3	5,6 (C)	4,2 (B)	5,4 (D)	4,0 (A)
4 (plus large)	6,8 (D)	5,0 (A)	5,8 (C)	5,4 (B)

B. Le vocabulaire des plans d'expériences

Comme les plans d'expériences sont utilisés dans un large éventail de domaines d'application, une terminologie unifiée et générique est utilisée.

Unité expérimentale : ensemble de conditions/matériaux de base sur lesquels un seul élément de l'expérience est réalisé (par exemple, une parcelle de terrain, un pot de fleurs, un patient, un échantillon de sang, un récipient de réaction, un morceau de matériau, etc.)

Traitement : ce que nous faisons à l'unité expérimentale. Dans diverses situations, un traitement peut être un type de plante, un régime d'arrosage, un médicament, un catalyseur, un type d'engrais, une méthodologie, etc.

Essai : application unique et indépendante d'un traitement à une unité expérimentale.

Réponse : le résultat d'un essai – également appelé observation, ou collectivement données.

Bloc : un ensemble d'unités expérimentales ayant certaines propriétés communes, par exemple des parcelles voisines, des patients ayant des caractéristiques similaires, l'heure de la journée, des échantillons provenant d'une source commune.

Facteur : aspect commun et sous-jacent d'un ensemble de traitements, par exemple dans une expérience visant à déterminer comment obtenir le meilleur rendement dans une réaction chimique, nous pourrions faire varier la température, la pression, la durée de la réaction, la qualité du catalyseur, etc. Chacun de ces éléments est un facteur. Un **traitement** (ou une combinaison de traitements) serait un ensemble de valeurs particulières pour les facteurs. Nous utilisons parfois le terme « facteur » pour désigner une variable explicative catégorielle plutôt qu'une variable explicative quantitative.

Facteur de nuisance : comme ci-dessus mais hors du contrôle de l'expérimentateur, par exemple les variables environnementales.

Réplication : une réplication unique est l'application de chaque traitement (ou combinaison de traitements) une fois. La plupart des expériences ont plusieurs réplications. Pour être une véritable réplication, chaque réplication doit être effectuée indépendamment sur des ensembles distincts d'unités expérimentales.

Pseudo-réplication : si les réplications ne sont pas effectuées sur des unités expérimentales indépendantes, il n'y a pas de véritable réplication et les estimations d'erreurs peuvent être trompeuses. Cela peut se produire si, par exemple, des observations sont faites sur la même unité pendant une certaine période ou lorsque des mesures répétées sont effectuées pour améliorer la précision. Attention, la pseudo-réplication peut conduire à une fausse sensibilité et à une augmentation de la probabilité de rejeter l'hypothèse nulle lorsqu'elle est vraie (erreur de type I)[14].

C. Pourquoi planifier des expériences ?

Les plans d'expériences permettent de gagner du temps et de l'argent et aident à utiliser au mieux des ressources limitées.

- Expérience trop importante : gaspillage de ressources.
- Expérience trop petite : perte de temps et de ressources.

Dans une expérience bien planifiée, la significativité pratique sera très probablement à l'origine d'une significativité statistique.

14 Ces notions seront définies dans le chapitre 9.

1.4.2 *La planification des sondages*

A. Qu'est-ce qu'un plan de sondage ?

Une grande partie de notre vie est désormais régie par des résultats d'enquêtes menées par sondage : enquêtes de satisfaction, enquêtes d'opinions politiques, etc.

Il est ainsi de plus en plus important que la conception et l'analyse des enquêtes par sondage soient comprises par le plus grand nombre afin, entre autres, de garantir une bonne interprétation des résultats de ces enquêtes.

L'échantillon utilisé pour mettre en œuvre le sondage doit être représentatif de la population. Le meilleur moyen d'y parvenir est souvent l'échantillonnage aléatoire même si d'autres méthodes non-aléatoires existent comme la méthode des quotas. En outre, avant de collecter cet échantillon, il est important que le chercheur définisse soigneusement et complètement la population, jusqu'à donner une description des membres à inclure.

Définition 1.17 Plan de sondage

Un **plan de sondage** consiste en la donnée, pour chacun des individus de la population, de la probabilité qu'il puisse être prélevé lors de la mise en œuvre du sondage ainsi que du mode de prélèvement qui sera utilisé : aléatoire ou non, avec ou sans remise, etc.

Les plans de sondage les plus fréquemment utilisés sont les suivants :

- **échantillonnage aléatoire simple** ;
- **échantillonnage aléatoire stratifié** ;
- **échantillonnage en grappes** ;
- **échantillonnage systématique** ;
- échantillonnage d'un continuum de temps, de surface ou de volume.

Les problématiques les plus couramment rencontrées dans l'application des sondages sont les suivantes :

- estimation d'un ratio ;
- estimation par régression ;
- estimation de la taille d'une population ;
- conception d'un questionnaire ;
- évaluation de l'erreur dans les enquêtes.

B. Exemples de plans de sondage

1) Une société bancaire compte 50 000 clients répartis en :
   a) 40 000 « petits » clients
   b) 10 000 « gros » clients.

Nous nous intéressons au montant moyen μ des comptes au moment de l'enquête et à la proportion π des clients prêts à souscrire au nouveau produit financier. Un plan de sondage stratifié portant sur 200 clients pourra être défini ainsi :

c) 160 « petits » clients

d) 40 « gros » clients.

2) Dans une commune donnée, il y a $N = 500$ propriétés immobilières. Elles occupent une surface d'environ 36,5 ha de terrain sur la commune. Cinq d'entre elles ont été vendues le mois dernier (cf. tableau 1.5).

Avec une estimation par régression, il sera possible d'estimer la valeur totale de la propriété immobilière dans la commune.

Tableau 1.5 : *prix de vente de cinq propriétés*

Propriétés	Surface du terrain en m²	Prix de vente en milliers d'euros
1	130	935
2	255	1 170
3	510	1 920
4	340	1 500
5	450	1 900

1.5. TRAITEMENT PRATIQUE DES DONNÉES MANQUANTES

Une valeur manquante est une valeur absente de l'une des caractéristiques de l'un des individus de l'échantillon alors que nous aurions souhaité l'observer. Il est très important de se demander pourquoi certaines des valeurs sont manquantes.

1.5.1 *Plusieurs types de valeurs manquantes*

A. Données manquantes complètement aléatoires

Définition 1.18 Données manquantes complètement aléatoires

Si les sujets pour lesquels des données manquent sont un sous-ensemble aléatoire de l'échantillon complet de sujets, les données manquantes sont appelées **données « manquantes complètement aléatoires »** (*MCAR – missing completely at random).*

Des exemples typiques de données *MCAR* sont un tube contenant un échantillon de sang d'un sujet d'étude cassé par accident (de sorte que les paramètres sanguins ne peuvent pas être mesurés) ou un questionnaire d'un sujet d'étude accidentellement perdu. La raison de la perte est totalement aléatoire, c'est-à-dire que la probabilité qu'une observation soit manquante n'est pas liée à d'autres caractéristiques du sujet. Cette situation est très rarement observée dans la pratique courante.

Remarques 1.4

1) La probabilité qu'une observation soit incomplète est constante.
2) Une absence de valeur pour une variable X_i est indépendante d'une autre ou d'autres valeurs des autres variables $X_{j\neq i}$.

B. Données manquantes non aléatoires

Définition 1.19 Données manquantes non aléatoires

Si la probabilité qu'une observation soit manquante dépend d'informations qui ne sont pas observées, comme la valeur de l'observation elle-même, les **données** manquantes sont dites « **manquantes non aléatoires** » (*MNAR – missing not at random*).

Par exemple, lorsque l'on demande à un sujet son niveau de revenu, la probabilité de données manquantes est plus forte lorsque le niveau de revenu est relativement élevé. Ici, la raison de l'absence de données n'est évidemment pas totalement aléatoire, mais liée à des caractéristiques non observées du sujet.

Remarques 1.5

1) La probabilité qu'une observation soit incomplète dépend de valeurs non observées.
2) Elle n'est pas aléatoire : une absence de valeur pour une variable X_i observée est dépendante d'une autre ou d'autres valeurs non observées des variables $X_{j\neq i}$.

C. Données manquantes au hasard

La plupart du temps, les données manquantes ne sont ni des *MCAR* ni des *MNAR*. Au contraire, la probabilité qu'une observation soit manquante dépend des informations présentes pour ce sujet, c'est-à-dire que la raison de cette absence est basée sur d'autres caractéristiques observées chez le patient. Ce type de données manquantes est appelé de manière confuse « **données manquantes au hasard** » (*MAR*), car les données manquantes peuvent être considérées comme aléatoires à condition que leur absence soit conditionnée aux autres caractéristiques du sujet et que ces caractéristiques soient disponibles au moment de l'analyse.

Définition 1.20 Données manquantes au hasard

Les **données manquantes au hasard** sont des données manquantes (*MAR*) qui ne sont ni des *MCAR* ni des *MNAR*.

Par exemple, supposons que nous voulions évaluer la valeur prédictive d'un test de diagnostic particulier, et que les résultats du test soient connus pour tous les sujets malades mais inconnus pour un échantillon aléatoire de sujets non malades. Dans ce cas, les données manquantes seraient *MAR* : sous réserve d'une caractéristique du patient qui est observée (ici la présence ou l'absence de la maladie), les données manquantes sont aléatoires. Bien sûr, il n'est pas nécessaire que l'absence de données dépende uniquement de la variable de résultat. Mais c'est la situation la plus simple, avec une variable dépendante ou de résultat et une seule variable indépendante ou prédictive. En outre, des situations où les valeurs prédictives manquantes dépendent (directement ou indirectement) de l'état du résultat se produisent régulièrement dans la pratique.

Remarques 1.6

1) La probabilité qu'une observation soit incomplète ne dépend que de valeurs observées (pas de valeurs manquantes).
2) Une absence de valeur pour une variable X_i est dépendante d'une autre ou d'autres variables $X_{j\neq i}$ observées

1.5.2 *Traitements usuels des valeurs manquantes*

A. Analyse de données complètes

L'analyse des cas complets consiste à faire porter l'analyse au sous-ensemble du jeu de données formé par les observations qui sont complètes, c'est-à-dire pour lesquelles toutes les valeurs des variables sont présentes. Ceci peut conduire à une réduction massive des données sur lesquelles pourra porter l'analyse. Une solution intermédiaire est d'utiliser, pour chaque analyse, toutes les observations pour lesquelles toutes les valeurs requises par l'analyse sont présentes. Par exemple pour l'estimation de la variable X_1, nous utiliserons toutes les observations pour lesquelles une valeur est présente pour cette variable X_1, ou pour estimer le coefficient de corrélation linéaire entre X_2 et X_3, nous utiliserons toutes les observations pour lesquelles deux valeurs sont présentes pour les variables X_2 et X_3.

B. Méthode de l'indicatrice

Une méthode parfois utilisée pour traiter les valeurs manquantes est la méthode dite de l'**indicatrice d'absence**. Pour chaque variable indépendante comportant des valeurs manquantes, une nouvelle variable fictive, indicatrice binaire (0/1), est créée, « 1 » indiquant une absence dans la variable d'origine et « 0 » indiquant une valeur observée. Pour la variable d'origine, les valeurs manquantes sont recodées avec la valeur « 1 ». Pour les variables originales qui seraient catégorielles, cela correspond en fait à la création d'une catégorie supplémentaire de valeurs pour les valeurs manquantes. Lors de l'estimation de l'association entre une variable explicative et une variable à expliquer dans une analyse multi-variables, l'indicateur est toujours inclus avec la variable originale, même si celle-ci a été recodée. Le principal avantage de la méthode de l'indicatrice est que tous les sujets peuvent être utilisés dans l'analyse multi-variables.

C. Imputation déterministe

L'**imputation par la médiane** ou **par la moyenne** consiste à remplacer toutes les occurrences de valeurs manquantes dans une variable par la médiane ou la moyenne. Il s'agit des deux exemples les plus courants d'**imputation déterministe**.

D. Imputation aléatoire

Les techniques d'imputation aléatoire sont basées sur l'idée que tout sujet d'un échantillon d'étude peut être remplacé par un nouveau sujet choisi au hasard dans la même population source. L'imputation de données manquantes sur une variable consiste à remplacer celles qui manquent par une valeur tirée d'une estimation de la distribution des valeurs possibles de cette variable.

a) Imputation aléatoire simple

Dans l'imputation aléatoire simple, une seule estimation est utilisée.

b) Imputation aléatoire multiple

Dans l'imputation aléatoire multiple, plusieurs estimations sont utilisées, ce qui permet de refléter l'incertitude de l'estimation de cette distribution.

1.5.3 *Comment traiter les valeurs manquantes en fonction de leur nature ?*

A. Données manquantes complètement aléatoires

Lorsque les données manquantes sont des *MCAR*, l'ensemble des sujets pour lesquels il n'y a pas de données manquantes est également un échantillon aléatoire de la population source. Par conséquent, la plupart des techniques simples de traitement des données manquantes, y compris les analyses des cas complets[15] ou disponibles, donnent des résultats non biaisés. L'estimation des associations à l'aide d'une analyse des cas complets ou disponibles reste moins efficace (c'est-à-dire plus imprécise), car une partie des données n'est pas utilisée.

B. Données manquantes non aléatoires

Si les données manquantes sont des *MNAR*, des informations précieuses sont perdues et il n'existe pas de méthode universelle pour traiter correctement les données manquantes.

C. Données manquantes aléatoires

Lorsque les données manquantes sont *MAR*, une analyse des cas complets ou disponibles n'est plus basée sur un échantillon aléatoire de la population source et un biais de sélection se produit probablement. En général, lorsque les données manquantes sont *MAR*, toutes les techniques simples de traitement des données manquantes, la méthode de l'indicatrice et l'imputation de la moyenne donnent des résultats biaisés. Toutefois, des techniques plus sophistiquées, telles que les imputations simples et multiples, donnent des résultats non biaisés.

1.5.4 *Conclusion*

Dans la plupart des situations, des techniques simples de traitement des données manquantes, telles que l'analyse des cas complets ou disponibles, l'imputation par la moyenne ou la médiane et la méthode de l'indicatrice, produisent des résultats biaisés, tandis que les techniques d'imputation donnent des résultats valables sans compliquer particulièrement l'analyse une fois les imputations effectuées. Dans les conditions générales de « manquant au hasard » et de « manquant complètement au hasard », les imputations simples et multiples donnent toutes deux des estimations non biaisées des associations d'études. Mais l'imputation simple se traduit par des erreurs types estimées trop faibles, tandis que l'imputation multiple se traduit par des erreurs types et des intervalles de confiance correctement estimés.

1.6. POUR UN TRAITEMENT ÉTHIQUE DES DONNÉES

L'éthique est depuis longtemps au cœur des principes fondamentaux de la statistique. Le code déontologique de l'Institut international de statistique, révisé en 2010, reconnaît que les statisticiens « peuvent être confrontés à des situations et contraintes variées qui soulèvent des questions éthiques » (p. 4). Il fixe donc les principes éthiques suivants dans le but d'aider les statisticiens « à travers le monde à accomplir leurs objectifs professionnels et à remplir leurs responsabilités » (p. 9) :

15 Parfois également appelée analyse des cas concrets.

1) s'attacher à respecter l'objectivité ;
2) expliciter les obligations et les rôles ;
3) évaluer avec impartialité les diverses options ;
4) prendre en considération les intérêts contradictoires ;
5) éviter tout résultat prédéterminé ;
6) protéger les informations d'accès privilégié ;
7) faire preuve de compétence professionnelle ;
8) assurer la confiance dans la statistique ;
9) exposer et évaluer les méthodes et les résultats ;
10) faire connaître les principes éthiques ;
11) porter la responsabilité de l'intégrité de la discipline ;
12) protéger les intérêts des sujets.

En Europe, le Code de bonnes pratiques de la statistique européenne, révisé en 2017, fixe les normes pour le développement, la production et la diffusion de statistiques européennes pertinentes, disponibles à temps, précises et respectueuses des principes d'indépendance professionnelle, d'impartialité et d'objectivité. Pour cela, il définit seize principes liés à l'environnement institutionnel, aux processus de statistiques et à la production statistique.

Face à la montée en puissance de l'intelligence artificielle (IA) et au développement d'outils de décision automatique (*machine learning*) de plus en plus performants, de nouvelles questions se posent et les statisticiens sont confrontés à une certaine défiance comme le montre W. Davies (2018)[16] en raison de l'utilisation de la statistique par les autorités publiques pour influer sur les émotions des citoyens ou l'exploitation des mégadonnées par de nouveaux acteurs économiques (les GAFA – Google, Apple, Facebook, Amazon).

À l'ère du *big data*, les principales questions éthiques sont liées à des problèmes de consentement, de confidentialité, de transparence ou d'explicabilité des décisions[17], de responsabilités et de loyauté des algorithmes, des risques de ré-identification des personnes, d'entrave à la concurrence, de biais discriminatoires envers des individus ou groupes sensibles, des décisions arbitraires, etc. Pour P. Besse *et al.* (2019)[18], « aborder sérieusement ces questions nécessite à la fois de sérieuses compétences techniques, afin de comprendre finement le fonctionnement des algorithmes et de garder un regard critique sur le discours qui les entoure, et une expertise juridique, sociétale ou sociologique, voire politique ou philosophique » (p. 24).

Pour E. Ruppert *et al.* (2019)[19], « qui dit sources de big data et nouvelles sources de données dit nouvelles interrogations en matière de respect de la vie privée, de consentement et de confidentialité, auxquelles les cadres réglementaires existants ne répondent pas toujours en détail » (p. 188).

16 Davies, W. (2018), *Nervous States – Democracy and the Decline of Reason*, W. W. Norton & Company.

17 « Une décision algorithmique est dite explicable s'il est possible d'en rendre compte explicitement à partir de données et caractéristiques connues de la situation. Autrement dit, s'il est possible de mettre en relation les valeurs prises par certaines variables (les caractéristiques) et leurs conséquences sur la prévision, par exemple d'un score, et ainsi sur la décision », in Besse, P., C. Castets-Renard et A. Garivier (2017), « Loyauté des Décisions Algorithmiques », Contribution au débat public initié par la CNIL « Éthique et Numérique », p. 16.

18 Besse, P., C. Castets-Renard, A. Garivier et J-M. Loubes (2019), « L'IA du quotidien peut-elle être éthique ? », *Statistique et société*, Vol. 6, n° 3, Décembre.

19 Ruppert, E., F. Grommé, F. Ustek-Spilda et B. Cakici (2019), « Données numériques de masse, "données citoyennes" et confiance dans la statistique publique », *Économie et Statistique*, n° 509.

C'est dans ce contexte que le gouvernement britannique a publié en 2018 un « cadre éthique des données » fondé sur sept principes :

- identifier un besoin clair auquel peuvent répondre les données et un bénéfice à l'utilisation des données ;
- connaître la législation et les codes de bonnes pratiques pertinents ;
- utiliser de manière proportionnée les données pour répondre aux besoins identifiés (le minimum de données pour atteindre le résultat souhaité doit être utilisé) ;
- identifier des limites des données afin d'évaluer l'opportunité de les utiliser pour répondre aux besoins des utilisateurs ;
- se limiter à son domaine de compétence et reconnaître son manque de compétences ou d'expérience pour utiliser une approche ou un outil particulier ;
- être transparent sur les outils, les données et les algorithmes utilisés ;
- utiliser les données de manière responsable.

Les risques liés à une exploitation abusive des données personnelles ont également amené les législateurs et régulateurs à instaurer de nouvelles règles pour encadrer leur usage. On peut notamment citer le règlement européen de protection des données personnelles (RGPD), entré en vigueur en mai 2018, qui poursuit un triple objectif :

- renforcer les droits des personnes pour développer la confiance et contribuer à l'essor de l'économie numérique ;
- assurer une plus grande harmonisation des règles de protection des données tout en renforçant la responsabilité des entreprises (et en réduisant fortement les formalités) ;
- développer le rôle des autorités de protection des données en renforçant leur pouvoir de sanction et en instaurant un mécanisme de coopération au plan européen.

S'agissant de la recherche, le règlement précise le régime juridique applicable aux traitements de données réalisés à des fins de recherche scientifique ou historique ou à des fins statistiques. Reconnaissant les spécificités de ces types de traitements, il introduit des possibilités de dérogations à certaines dispositions du règlement[20] sous réserve que des garanties appropriées pour les droits et libertés de la personne concernée soient apportées (notamment en matière d'anonymisation) et que le principe de minimisation des données utilisées soit respecté.

20 Par exemple, dérogation à l'obligation d'information, dès lors que cette dernière serait susceptible de rendre impossible ou de compromettre gravement la réalisation des objectifs dudit traitement.

CHAPITRE 2

STATISTIQUES DESCRIPTIVES ET VISUALISATION DES DONNÉES UNIVARIABLES

2.1. Statistique descriptive univariable 38

2.2. Représentations graphiques pour les séries statistiques quantitatives 44

2.3. Représentations graphiques pour les séries statistiques qualitatives 49

2.4. Indices synthétiques et courbe de Lorenz 57

2.5. Caractéristiques de position 63

2.6. Caractéristiques de dispersion 68

2.7. Boîte à moustaches et caractéristiques de forme 72

La compréhension de la problématique posée, la connaissance des objectifs et la « prise en main » des données sont essentielles pour réussir une étude. Le chapitre précédent a montré l'importance de la collecte des données, notamment la façon de les constituer et d'étudier leur qualité. Les résultats issus des méthodes de traitement statistique des données et surtout leur interprétation sont par conséquent étroitement liés à cette première phase. En d'autres termes, les conclusions d'une étude dépendent autant, voire plus de cette première phase que de la sophistication des méthodes d'analyse utilisées.

2.1. STATISTIQUE DESCRIPTIVE UNIVARIABLE

2.1.1 *Modalités et classes*

Le choix d'un caractère détermine le critère qui sert à comparer les individus entre eux.

Définition 2.1 Modalité

Dans le cas d'un caractère qualitatif, cela revient à classer les individus en plusieurs groupes. Ils correspondent aux différentes valeurs que peut prendre le caractère qualitatif étudié. Ces dernières sont appelées **modalités**.

Exemples

Pour le caractère genre, il y a deux modalités.
Pour la profession des individus, la répartition peut se faire suivant plusieurs critères. En effet, elle peut, par exemple, se faire soit selon le nombre d'années d'étude après le baccalauréat, soit selon la classe socio-professionnelle des individus.

Remarques 2.1

1) Le nombre de modalités peut être fixé plus ou moins conventionnellement suivant la nature du caractère qualitatif étudié.
2) Les modalités du caractère étudié doivent être simultanément exhaustives (chaque individu appartient à une modalité, car tous les cas ont été prévus) et incompatibles (un individu ne peut appartenir à deux modalités ou plus).

Exemple

Un élève en école d'ingénieur sera diplômé ou ne le sera pas.

Définition 2.2 Classe

Dans le cas d'un caractère quantitatif simple, discret ou continu, les individus sont souvent regroupés dans des **classes**, c'est-à-dire au sein d'intervalles de valeurs que peut prendre le caractère. L'utilisation de classes est plus fréquente lorsque le caractère est un caractère quantitatif simple continu.

Exemple

Lors d'une étude sur la « taille » sur la population française, qui est un caractère quantitatif simple continu, il est possible de regrouper les observations en classes de 10 cm de largeur de 0 cm à 230 cm. Ce découpage aboutirait aux classes suivantes :]0 cm ; 10 cm],]10 cm ; 20 cm], …,]220 cm ; 230 cm].

2.1.2 *Tableaux statistiques*

En fonction du contexte, il est possible de représenter les données statistiques d'un échantillon à l'aide de différents types de tableaux.

La solution qui peut être employée dans tous les cas, et qui est le plus souvent utilisée par les logiciels statistiques, est de présenter les tableaux de données sous une **forme empilée**. Dans cette forme, un **tableau statistique** est constitué :

- d'une ligne par individu ;
- d'une colonne par caractère.

Concrètement, dans le cas de N observations d'un seul caractère X, nous pouvons avoir le **tableau de données brutes** suivant (cf. tableau 2.1) :

Tableau 2.1 : *données brutes pour un caractère et N individus*

	X
Individu 1	X_1
Individu 2	X_2
…	
Individu *N*	X_N

Dans le cas de N observations de deux caractères, nous pouvons avoir le **tableau de données brutes** suivant (cf. tableau 2.2) :

Tableau 2.2 : *données brutes pour deux caractères et N individus*

	X	Y
Individu 1	X_1	Y_1
Individu 2	X_2	Y_2
…		
Individu *N*	X_N	Y_N

Pour diminuer la place occupée par les tableaux, certaines personnes inversent malheureusement le rôle traditionnel des lignes et des colonnes dans les tableaux empilés. Si nous utilisons un logiciel pour réaliser les analyses statistiques, il faudra penser à rétablir la convention précédente.

2.1.3 *Série statistique*

Définition 2.3 Série statistique

Les observations, d'un ou plusieurs caractères, récoltées sur toute la population forment une **série statistique**.

Les séries statistiques les plus élémentaires sont les séries à un seul caractère (**série statistique simple**) univariable (**série statistique simple univariable**) quantitatif ou qualitatif. Lorsque le caractère de la série statistique simple univariable est quantitatif discret (respectivement continu), la série est dite **série statistique simple discrète** (respectivement **continue**).

Remarques 2.2

1) En général, ces séries sont représentées sous la forme d'un tableau statistique à une seule entrée puisqu'il n'y a qu'un seul caractère simple étudié.
2) Parfois, la lecture de ces tableaux se révèle difficile.
3) C'est pourquoi les représentations graphiques, qui sont présentées dans les sections 2.3, 2.4 et 2.8, sont utilisées très fréquemment pour faciliter l'analyse des informations qui sont contenues dans ces tableaux.

2.1.4 *Description d'une série statistique quantitative*

Soit X un caractère quantitatif simple discret ou continu. L'ensemble des valeurs atteintes par le caractère $X(\Omega) = \{X(\omega), \omega \in \Omega\} = \{X_1, \ldots, X_N\}$ est un ensemble fini de valeurs distinctes $\{x_1, \ldots, x_p\}$.

Le fait que telle valeur soit relative à tel individu est un renseignement qui n'intéresse pas le statisticien. Seules les informations relatives à l'ensemble des valeurs atteintes et le nombre de fois où chacune d'elles est atteinte lui sont utiles. Nous supposons que les valeurs $x_1, \ldots, x_p$ sont ordonnées : $x_1 < \ldots < x_p$.

A. Distribution statistique non groupée

Définition 2.4 Effectif et fréquence

Soit une série statistique simple, discrète ou continue, $(X_1, \ldots, X_N)$ qui prend pour valeurs $x_1, \ldots, x_p$ avec $x_1 < \ldots < x_p$.

Nous appelons :

- **effectif de la valeur x_i** : le nombre n_i de fois où la valeur x_i est prise, c'est-à-dire le cardinal de l'ensemble $X^{-1}(x_i)$;
- **effectif cumulé en x_i** : la somme $\sum_{j=1}^{i} n_j$;
- **fréquence de la valeur x_i** : le rapport $f_i = \frac{n_i}{N}$ de l'effectif de x_i à l'effectif total N de la population, c'est-à-dire le cardinal de Ω ou encore la somme des n_i ;
- **fréquence cumulée en x_i** : la somme $\sum_{j=1}^{i} f_j$.

La suite de couples $((x_i, n_i))_{i=1,\ldots,p}$ ou $((x_i, f_i))_{i=1,\ldots,p}$ est appelée **distribution statistique** (simple, discrète ou continue) **non groupée**.

Remarques 2.3

Notons que, par définition :

1) la somme des effectifs n_i est égale à N.
2) la somme des fréquences f_i est égale à un.

Il est évident que tous ces indicateurs peuvent figurer dans un même tableau statistique. Leur lecture est alors plus rapide, car les informations sont concentrées dans un seul élément statistique.

B. Distribution statistique groupée

Définition 2.5 Classe

Lorsque le caractère X quantitatif discret ou continu comprend un grand nombre de valeurs, il est préférable de regrouper ces valeurs en intervalles appelés **classes** pour rendre la statistique plus lisible.

L'ensemble C des valeurs du caractère est alors partagé en classes $]a_i\ ;\ a_{i+1}]$ avec $a_i < a_{i+1}$.

Définition 2.6 Effectif et fréquence

Soit une série statistique $(]a_i\ ;\ a_{i+1}],\ n_i)_{i=1,\ldots,p}$.

Nous appelons :

- **effectif de la classe $]a_i\ ;\ a_{i+1}]$** : le nombre n_i de valeurs prises par le caractère dans l'intervalle $]a_i\ ;\ a_{i+1}]$, c'est-à-dire $X^{-1}(]a_i\ ;\ a_{i+1}])$;
- **effectif cumulé en a_i** : le nombre de valeurs prises dans l'intervalle $]-\infty\ ;\ a_i]$;
- **fréquence de la classe $]a_i\ ;\ a_{i+1}]$** : le rapport $f_i = \dfrac{n_i}{N}$;
- **fréquence cumulée en a_i** : la somme $\displaystyle\sum_{j=1}^{i} f_j$.

La série statistique $(]a_i\ ;\ a_{i+1}],\ n_i)_{i=1,\ldots,p}$ ou $(]a_i\ ;\ a_{i+1}],\ f_i)_{i=1,\ldots,p}$ est appelée **distribution statistique** (simple, discrète ou continue) **groupée**.

2.1.5 *Description d'une série statistique qualitative*

Pour un caractère qualitatif simple, les notions d'effectif et de fréquence s'appliquent également en considérant les différentes modalités que peut prendre le caractère qualitatif. Si les modalités du caractère qualitatif peuvent être ordonnées, il est également naturel de se servir des effectifs cumulés et des fréquences cumulées. Il est aussi vivement recommandé de résumer toutes les informations concernant le caractère sous forme d'un tableau statistique.

Exemples

Distribution statistique non groupée

Le tableau 2.3 contient la population des dix provinces et des trois territoires du Canada au quatrième trimestre 2020.
L'effectif de cette série statistique en x_7 = 1 379 584 est 1 et l'effectif cumulé en x_7 = 1 379 584 est 9.
La fréquence en x_{10} = 5 145 851 est égale à 1/13 et la fréquence cumulée en x_{10} = 5 145 851 à 11/13.

Tableau 2.3 : *population des 10 provinces et 3 territoires du Canada au quatrième trimestre 2020*

x_1	Terre-Neuve-et-Labrador	520 998
x_2	Île-du-Prince-Édouard	159 713
x_3	Nouvelle-Écosse	979 115
x_4	Nouveau-Brunswick	781 315
x_5	Québec	8 575 779
x_6	Ontario	14 733 119
x_7	Manitoba	1 379 584
x_8	Saskatchewan	1 177 884
x_9	Alberta	4 428 112
x_{10}	Colombie-Britannique	5 145 851
x_{11}	Yukon	42 176
x_{12}	Territoires du Nord-Ouest	45 074
x_{13}	Nunavut	39 285

Source : Statistique Canada. Tableau 17-10-0009-01.
Estimations de la population, trimestrielles.
DOI : https://doi.org/10.25318/1710000901-fra.

Distribution statistique groupée

Le tableau 2.4 présente la répartition en classes d'âge de la population du Canada en 2020.

Tableau 2.4 : *répartition en classes d'âge de la population du Canada en 2020*

0 à 4 ans	1 921 944
5 à 9 ans	2 044 603
10 à 14 ans	2 072 100
15 à 19 ans	2 100 865
20 à 24 ans	2 482 802
25 à 29 ans	2 645 240
30 à 34 ans	2 661 723
35 à 39 ans	2 630 680
40 à 44 ans	2 464 247
45 à 49 ans	2 390 116
50 à 54 ans	2 449 915
55 à 59 ans	2 744 896
60 à 64 ans	2 560 241
65 à 69 ans	2 167 275
70 à 74 ans	1 786 622
75 à 79 ans	1 218 303
80 à 84 ans	811 370
85 à 89 ans	517 710
90 à 94 ans	248 593
95 à 99 ans	74 476
100 ans et plus	11 517

Source Statistique Canada. Tableau 17-10-0005-01.
Estimations de la population au 1er juillet, par âge et sexe.
DOI: https://doi.org/10.25318/1710000501-fra.

Le second élément, 2 464 247, du couple ([40 ; 45[, 2 464 247) correspond à l'effectif de la classe qui apparaît comme le premier élément, [40 ; 45[, du couple. Ce couple figure à la neuvième ligne du tableau.

L'effectif de la classe [45 ; 50[est 2 390 116. L'effectif cumulé en 60 est égal à 28 609 131.

La fréquence de la classe [20 ; 25[est égale à 6,53 % et la fréquence cumulée en 50 à 61,61 %.

Exercice 2.1

Pour chacune des séries statistiques présentées dans les sections 2.3 à 3.3, répondre aux questions suivantes.

1) Déterminer les effectifs et effectifs cumulés.
2) Déterminer les fréquences et fréquences cumulées.

2.2. REPRÉSENTATIONS GRAPHIQUES POUR LES SÉRIES STATISTIQUES QUANTITATIVES

2.2.1 *Représentations graphiques pour les séries statistiques non groupées*

Pour les séries statistiques non groupées, il existe deux types de représentation graphique :

1) le diagramme en bâtons ;

2) le polygone.

A. Diagramme en bâtons des effectifs ou des fréquences

Définition 2.7 Diagramme en bâtons des effectifs ou des fréquences

Le **diagramme en bâtons des effectifs** (cf. figure 2.2), (respectivement **des fréquences**, cf. figure 2.3) d'une distribution statistique non groupée est constitué d'une suite de segments verticaux d'abscisse x_i dont la longueur est proportionnelle à l'effectif (respectivement la fréquence) de x_i, pour i variant de 1 à p.

B. Diagramme en bâtons des effectifs cumulés ou des fréquences cumulées

Définition 2.8 Diagramme en bâtons des effectifs cumulés ou des fréquences cumulées

En remplaçant dans la définition précédente le mot effectifs (respectivement fréquences) par effectifs cumulés (respectivement fréquences cumulées), nous obtenons le **diagramme en bâtons des effectifs cumulés** (cf. figure 2.4), (respectivement **des fréquences cumulées**).

C. Polygone des effectifs ou des fréquences

Définition 2.9 Polygone des effectifs et des fréquences

Le **polygone des effectifs** (respectivement **des fréquences**) d'une distribution statistique non groupée est obtenu à partir du diagramme en bâtons des effectifs (respectivement des fréquences, cf. figure 2.3), en joignant par un segment les sommets des bâtons.

D. Polygone des effectifs cumulés ou des fréquences cumulées

Définition 2.10 Polygone des effectifs cumulés ou des fréquences cumulées

En remplaçant dans la définition précédente le mot effectifs (respectivement fréquences) par effectifs cumulés (respectivement fréquences cumulées), nous obtenons le **polygone des effectifs cumulés**, (cf. figure 2.4), (respectivement **des fréquences cumulées**).

2.2.2 *Représentations graphiques pour les séries statistiques groupées*

Pour les séries statistiques groupées, il existe deux types de représentation graphique :

1) l'histogramme ;

2) le polygone.

A. Histogramme

Définition 2.11 Histogramme

L'**histogramme** est une représentation graphique d'une distribution statistique groupée.

Il peut être utilisé pour estimer une densité ; il doit alors délimiter une aire totale égale à 1.

Premier cas : les amplitudes des classes sont égales

Cet histogramme, dont un exemple est représenté à gauche de la figure 2.1, est constitué d'un ensemble de rectangles dont

1) la largeur est égale à a, l'amplitude de la classe,

2) la hauteur à $K \times n_j$ où n_j est l'effectif de la classe j variant de 1 à p, et K est un coefficient arbitraire (choix d'une échelle).

L'aire totale sous l'histogramme est égale à $K \times N \times a$ où N est l'effectif total (cf. figure 2.5).

Second cas : les amplitudes des classes sont inégales

Cet histogramme, dont un exemple est représenté à droite de la figure 2.1, est constitué d'un ensemble de rectangles dont les largeurs sont des multiples entiers $k_j \times a$ de l'amplitude a de l'une d'entre elles.

Nous convenons, pour conserver le résultat précédent, de prendre pour hauteur du rectangle de la classe numéro j le quotient $\dfrac{(K \times n_j)}{k_j}$, pour j variant de 1 à p.

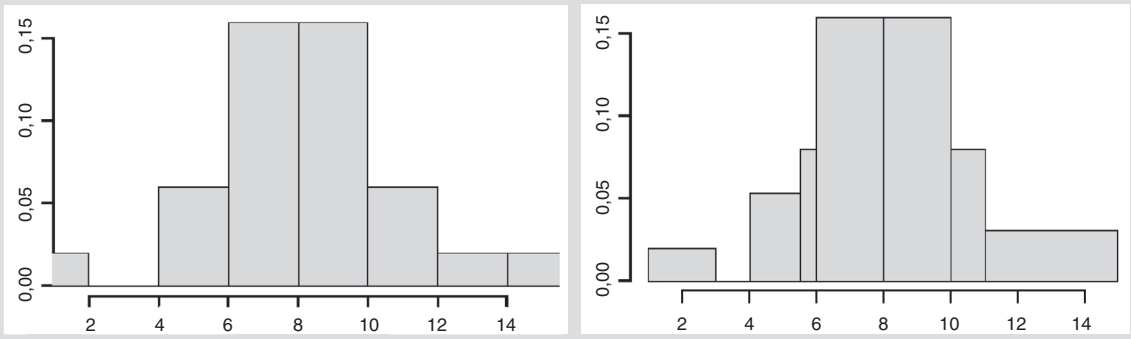

Figure 2.1 : *histogrammes avec amplitude des classes égales (à gauche) et inégales (à droite)*

B. Polygone des effectifs ou des fréquences

Définition 2.12 Polygone des effectifs ou des fréquences

Le **polygone des effectifs** ou **des fréquences** (cf. figure 2.6) d'une distribution statistique groupée est obtenu en joignant dans l'histogramme de cette distribution les milieux des côtés horizontaux supérieurs.

C. Polygone des effectifs cumulés ou des fréquences cumulées

Définition 2.13 Polygone des effectifs cumulés

Le **polygone des effectifs cumulés** d'une distribution statistique groupée est la représentation graphique de la fonction définie sur chaque intervalle $]a_i\,;\,a_{i+1}]$, $1 \leq i \leq p$ par :

$$\forall\ x \in]a_i\,;\,a_{i+1}],\ f(x) = \sum_{j=1}^{i-1} n_j + \frac{x - a_i}{a_{i+1} - a_i} n_i.$$

Définition 2.14 Polygone des fréquences cumulées

Le **polygone des fréquences cumulées** (cf. figure 2.7) d'une distribution statistique groupée est la représentation graphique de la fonction définie sur chaque intervalle $]a_i\,;\,a_{i+1}]$, $1 \leq i \leq p$ par :

$$\forall\ x \in]a_i\,;\,a_{i+1}],\ f(x) = \sum_{j=1}^{i-1} f_j + \frac{x - a_i}{a_{i+1} - a_i} f_i.$$

Exemple

Voici la répartition des ménages, de moins de 6 personnes, en France métropolitaine selon le nombre de personnes en 2017, d'après le recensement de la population réalisé par l'INSEE :

Tableau 2.5 : *répartition des ménages selon le nombre de personnes en 2017*

Nombre de personnes	1	2	3	4	5	6
Nombre de ménages	10 450 926	9 403 897	3 912 206	3 307 292	1 199 238	459 463

Source : INSEE.

Les figures 2.2, 2.3 et 2.4 correspondent respectivement au diagramme en bâtons des effectifs de cette série statistique, à la superposition du diagramme en bâtons et du polygone des fréquences, à la superposition du diagramme en bâtons et du polygone des effectifs cumulés.

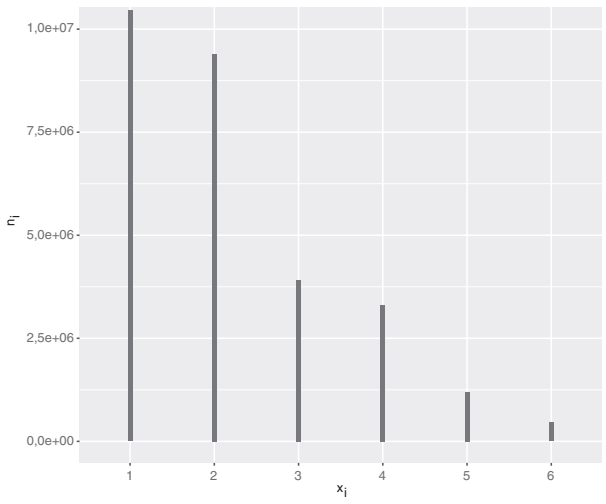

Figure 2.2 : *diagramme en bâtons des effectifs*

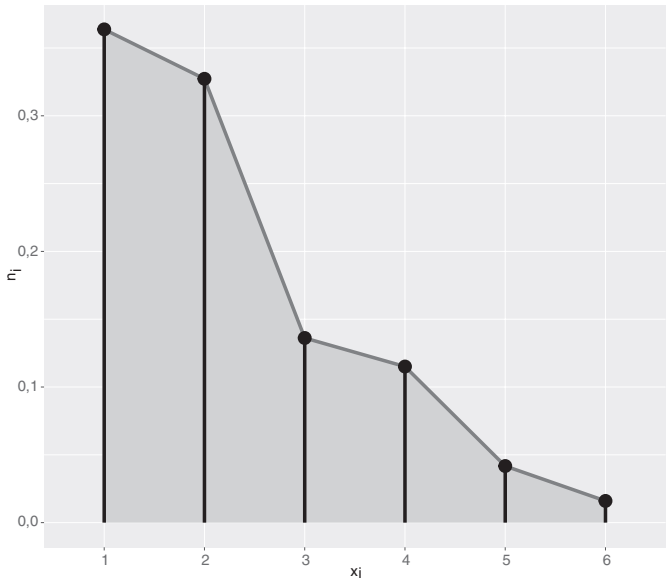

Figure 2.3 : *superposition du diagramme en bâtons et du polygone des fréquences*

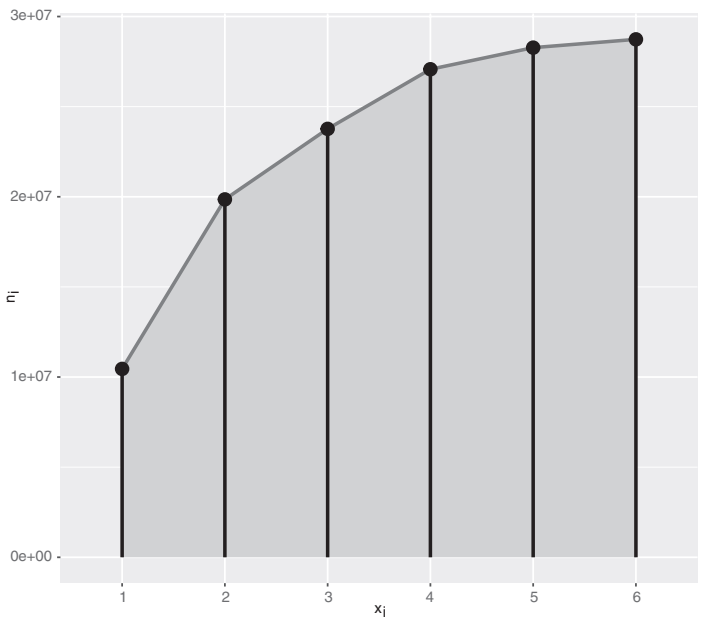

Figure 2.4 : *superposition du diagramme en bâtons et du polygone des effectifs cumulés*

Exemple

Considérons la série statistique groupée donnée par les lignes du tableau 2.4, correspondant à la répartition en classes d'âge de la population du Canada en 2020, en se restreignant aux âges compris entre 0 et 100 ans.

Les figures 2.5, 2.6 et 2.7 correspondent respectivement à l'histogramme des effectifs, à la superposition de l'histogramme des fréquences et du polygone des fréquences, à la superposition de l'histogramme des fréquences cumulées et du polygone des fréquences cumulées.

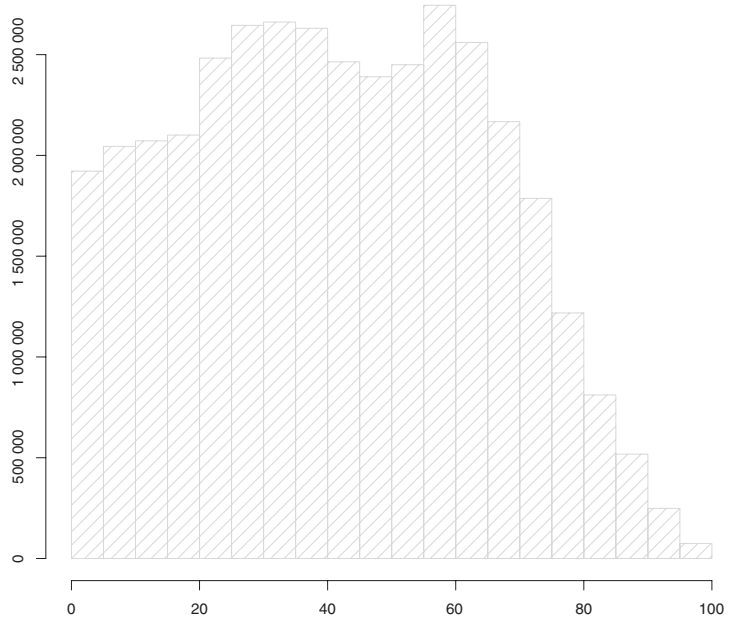

Figure 2.5 : *histogramme des effectifs*

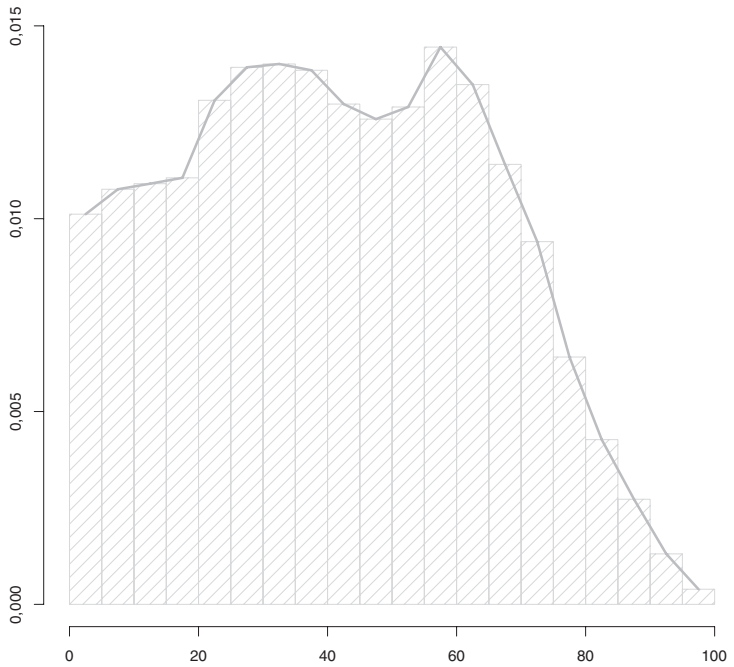

Figure 2.6 : *superposition de l'histogramme des fréquences et du polygone des fréquences*

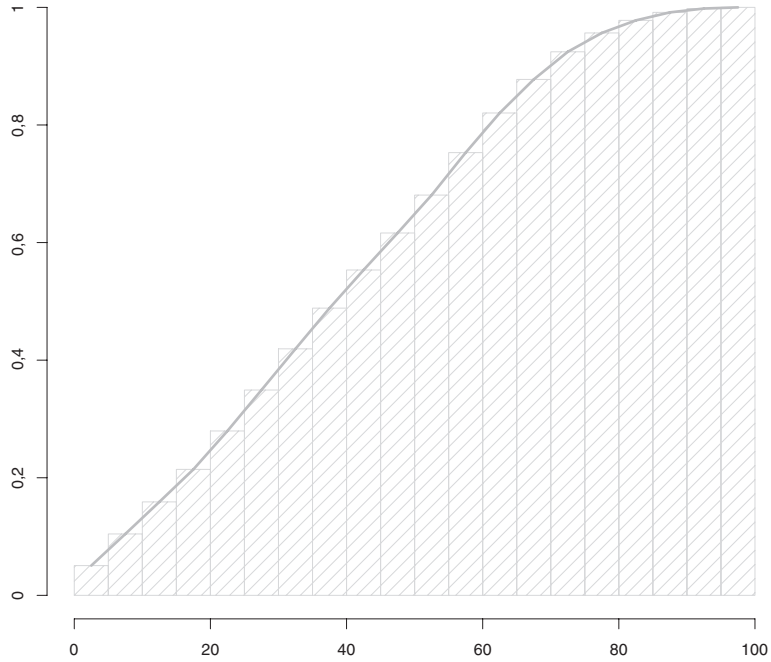

Figure 2.7 : *superposition de l'histogramme des fréquences cumulées et du polygone des fréquences cumulées*

Exercice 2.2

Pour chacune des séries statistiques quantitatives présentées dans les sections 2.5 à 3.3, construire les représentations graphiques suivantes.

1) Si la distribution n'est pas groupée, le diagramme en bâtons des effectifs, le polygone des fréquences superposé au diagramme en bâtons, le polygone des effectifs cumulés superposé au diagramme en bâtons.
2) Si la distribution est groupée, l'histogramme des fréquences, le polygone des fréquences superposé à l'histogramme et le polygone des effectifs cumulés superposé à l'histogramme.

2.3. REPRÉSENTATIONS GRAPHIQUES POUR LES SÉRIES STATISTIQUES QUALITATIVES

2.3.1 *Diagramme à barres*

Définition 2.15 Diagramme à barres verticales

Le **diagramme à barres verticales des effectifs** ou **des fréquences** (cf. figure 2.8) est une représentation graphique de la distribution, groupée ou non groupée, d'une série statistique qualitative simple par un ensemble de rectangles. L'axe des abscisses correspond aux différentes modalités du caractère de la série étudiée et l'axe des ordonnées aux effectifs ou fréquences associés.

Définition 2.16 Diagramme à barres horizontales

Le **diagramme à barres horizontales des effectifs** ou **des fréquences** (cf. figure 2.11) est une représentation graphique de la distribution, groupée ou non groupée, d'une série statistique qualitative simple par un ensemble de rectangles. L'axe des ordonnées correspond aux différentes modalités du caractère de la série étudiée et l'axe des abscisses aux effectifs ou fréquences associés.

2.3.2 *Diagramme à points*

Définition 2.17 Diagramme à points vertical

Le **diagramme à points vertical des effectifs** ou **des fréquences** est une représentation graphique de la distribution, groupée ou non groupée, d'une série statistique qualitative simple par un ensemble de points. L'axe des abscisses correspond aux différentes modalités du caractère de la série étudiée et l'axe des ordonnées aux effectifs ou fréquences associés.

Définition 2.18 Diagramme à points horizontal

Le **diagramme à points horizontal des effectifs** ou **des fréquences** (cf. figures 2.9 et 2.12) est une représentation graphique de la distribution, groupée ou non groupée, d'une série statistique qualitative simple par un ensemble de points. L'axe des ordonnées correspond aux différentes modalités du caractère de la série étudiée et l'axe des abscisses aux effectifs ou fréquences associés.

2.3.3 *Diagramme circulaire*

Définition 2.19 Diagramme circulaire

Le **diagramme circulaire** est une représentation graphique (cf. figures 2.10 et 2.13) de la distribution, groupée ou non groupée, d'une série statistique qualitative simple. À chacune des modalités du caractère est associée une portion circulaire du diagramme proportionnelle à sa fréquence.

Remarque 2.4

Les diagrammes circulaires ne sont pas des outils efficaces pour représenter les séries statistiques, car l'œil humain distingue difficilement des variations d'angles ou de surfaces. De plus, il est toujours possible d'utiliser à leur place un diagramme à points (cf. section 2.4.2).

Exemple

Considérons la répartition des nombres d'emplois par secteur d'activité, tout pays confondu, dans les pays de l'OCDE au troisième trimestre 2020. L'emploi par secteur est ventilé entre l'agriculture, la construction, l'industrie y compris la construction, les activités manufacturières et les services. Cet indicateur est désaisonnalisé (voir chapitre 12 pour comprendre la saisonnalité).

Tableau 2.6 : *emploi par activité*

Secteur	Nombre d'emplois
Agriculture	21 143 585
Construction	35 197 834
Industrie (sans la construction)	69 941 779
Fabrication	64 298 386
Services	368 931 820

Source OCDE (2021). doi: 10.1787/6b2fff89-fr
(Consulté le 11 février 2021).

Les figures 2.8, 2.9 et 2.10 correspondent aux représentations graphiques de cette série statistique qualitative simple, à l'aide respectivement d'un diagramme à barres verticales des effectifs, d'un diagramme à points horizontal des fréquences et d'un diagramme circulaire.

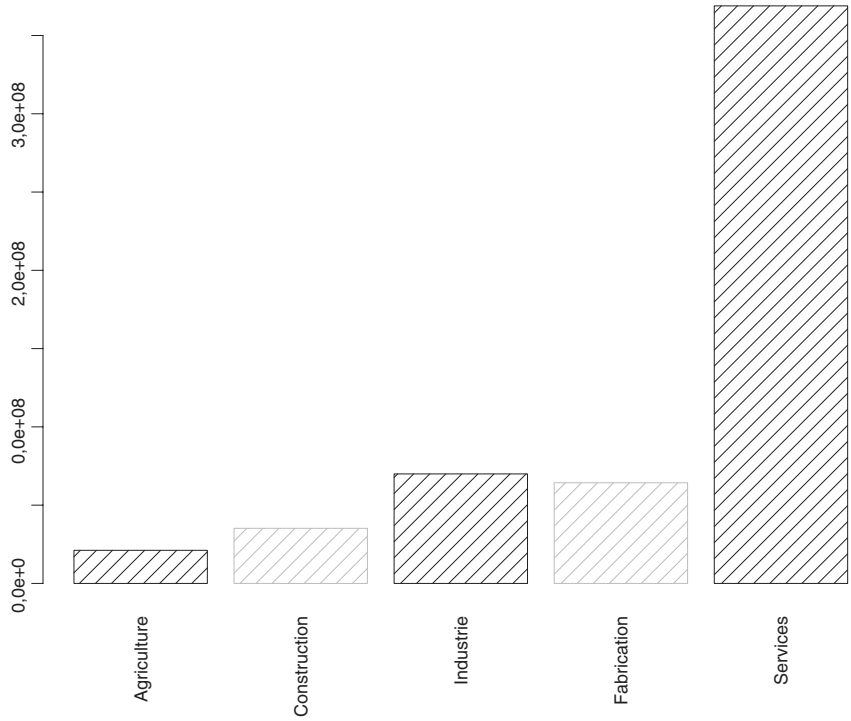

Figure 2.8 : *diagramme à barres verticales des effectifs*

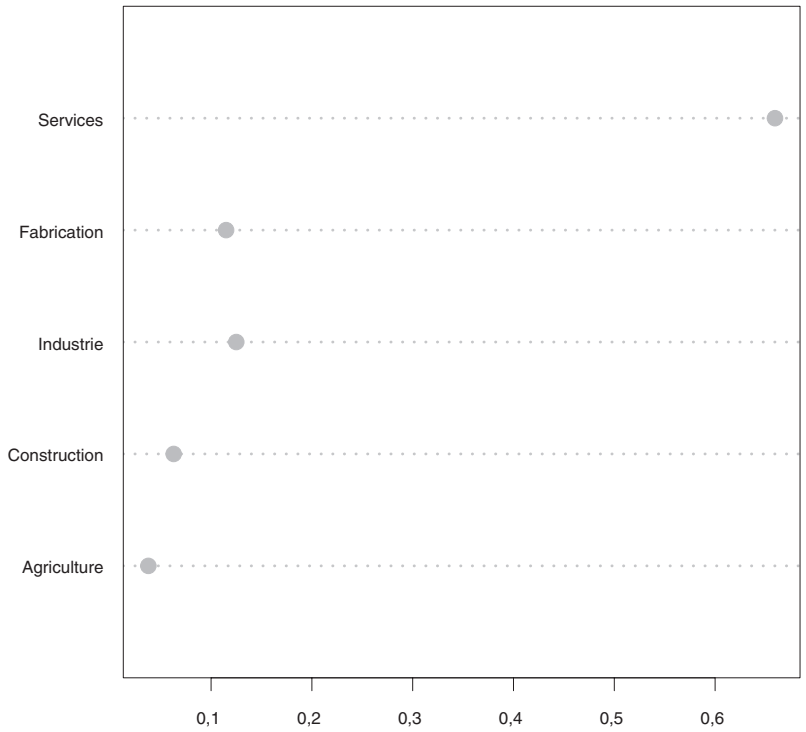

Figure 2.9 : *diagramme à points horizontal des fréquences*

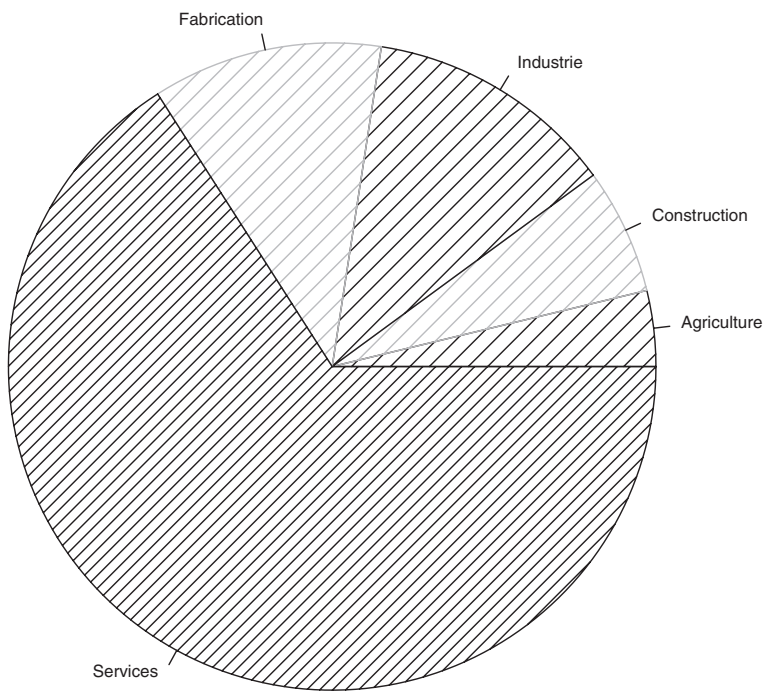

Figure 2.10 : *diagramme circulaire*

Exemple

Considérons à présent la répartition des nombres d'emplois par pays, tous secteurs d'activité confondus, dans les pays de l'OCDE au troisième trimestre 2020. Cet indicateur est désaisonnalisé.

Tableau 2.7 : *emploi par pays*

Pays	Nombre d'emplois	Pays	Nombre d'emplois
AUT	5 167 648	ISR	4 366 751
BEL	5 721 200	ITA	28 409 214
CAN	21 122 321	JPN	81 977 909
CHE	5 568 900	KOR	33 262 656
CHL	8 732 089	LTU	1 647 481
COL	22 791 085	LUX	316 889
CZE	7 015 674	LVA	253 516
DNK	3 343 269	NLD	10 143 660
ESP	22 701 902	NOR	3 114 348
EST	200 241	NZL	3 242 630
FIN	1 154 495	POL	21 006 175
FRA	31 775 373	PRT	5 902 090
GBR	37 353 540	SVK	3 385 346
GRC	4 382 010	SVN	1 275 598
HUN	5 787 638	SWE	5 872 257
IRL	2 692 289	TUR	33 604 952
ISL	40 222	USA	171 209 618

Source OCDE (2021). doi: 10.1787/6b2fff89-fr
(Consulté le 11 février 2021).

Les figures 2.11, 2.12 et 2.13 correspondent aux représentations graphiques de cette série statistique qualitative simple, à l'aide respectivement d'un diagramme à barres horizontales des fréquences, d'un diagramme à points horizontal des effectifs et d'un diagramme circulaire.

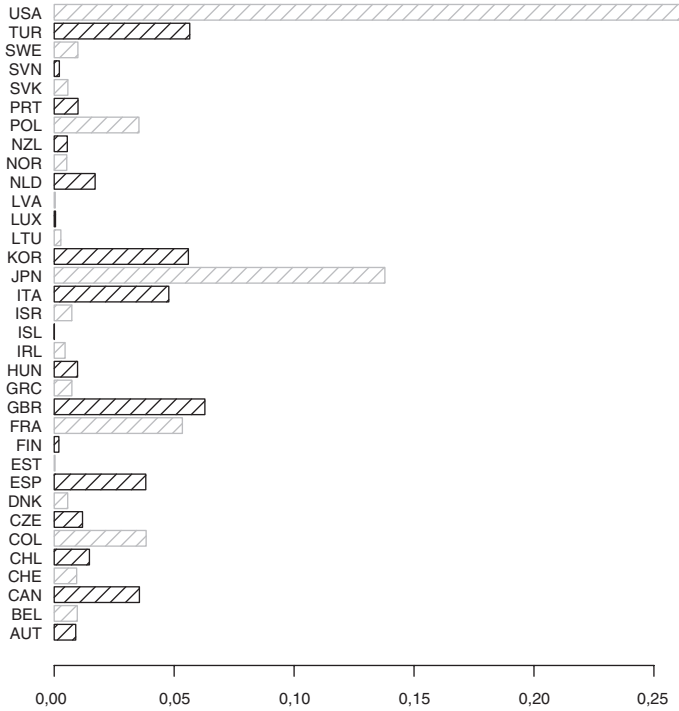

Figure 2.11 : *diagramme à barres horizontales des fréquences*

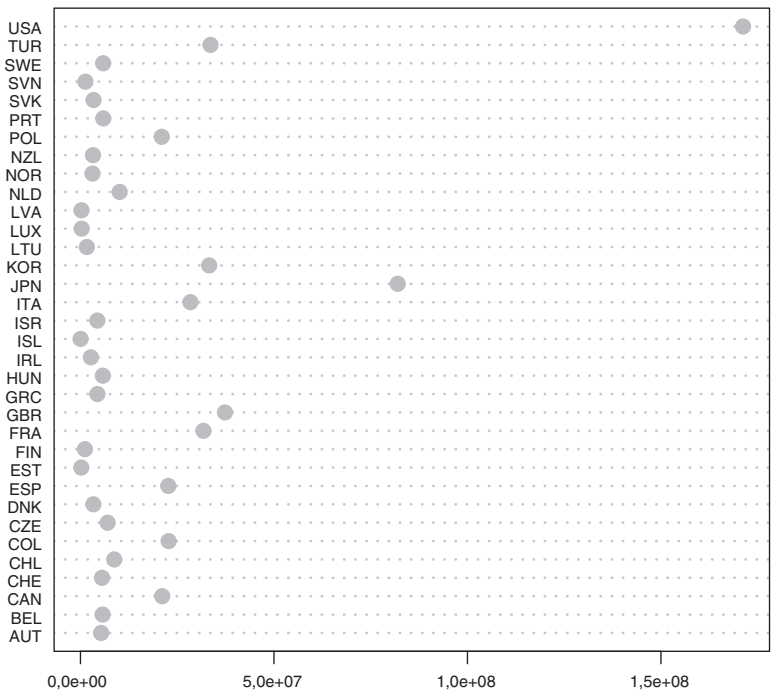

Figure 2.12 : *diagramme à points horizontal des effectifs*

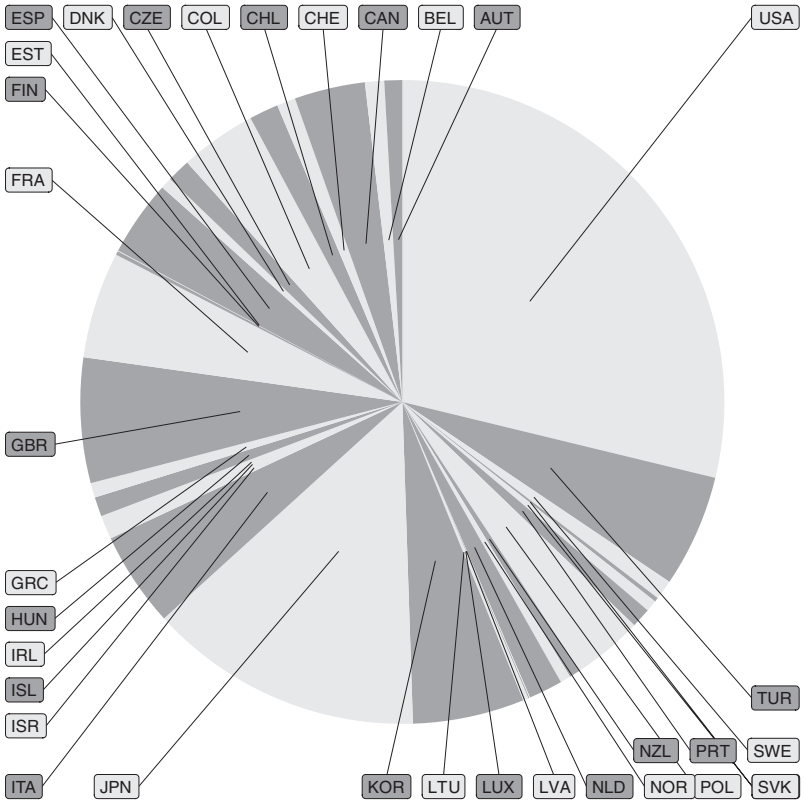

Figure 2.13 : *diagramme circulaire*

Exercice 2.3

Les colonnes 2, 3 et 4 du tableau 2.8 contiennent les taux d'emploi, au troisième trimestre 2020, selon le niveau d'études : premier cycle du second degré (PCSD), deuxième cycle du second degré (DCSD), supérieur (SUP). Le taux d'emploi est le pourcentage d'actifs occupés dans la population en âge de travailler. Les actifs occupés sont les personnes qui travaillent au moins une heure par semaine en tant que salarié ou à titre lucratif, ou qui ont un emploi, mais sont temporairement absentes de leur travail pour maladie, congé ou conflit social. Cet indicateur donne le pourcentage des actifs occupés âgés de 25 à 64 ans dans la population des individus âgés de 25 à 64 ans. Les trois dernières colonnes du tableau 2.8 contiennent le taux d'emploi, au troisième trimestre 2020, d'une classe d'âge. Cet indicateur se mesure en fonction du nombre des actifs occupés d'un âge donné rapporté à l'effectif total de cette classe d'âge. Les actifs occupés sont les personnes de 15 ans et plus qui, durant la semaine de référence, déclarent avoir effectué un travail rémunéré pendant une heure au moins ou avoir occupé un emploi dont elles étaient absentes. Les taux d'emploi sont présentés pour trois classes d'âge : les personnes âgées de 15 à 24 ans sont celles qui font leur entrée sur le marché du travail à l'issue de leur scolarité, les personnes âgées de 25 à 54 ans sont celles qui sont au plus fort de leur activité professionnelle, et les personnes âgées de 55 à 64 ans sont celles qui ont dépassé le pic de leur carrière professionnelle et approchent de l'âge de la retraite. Cet indicateur est désaisonnalisé et est mesuré en pourcentage de l'effectif total de la classe d'âge.

Tableau 2.8 : *emploi par niveau d'études, (doi: 10.1787/6e3d44f3-fr) et taux d'emploi par groupe d'âge. (doi: 10.1787/b01db125-fr)*

Pays	DCSD (en %)	PCSD (en %)	SUP (en %)	15 à 24 (en %)	25 à 54 (en %)	55 à 64 (en %)
AUS	79,01	61,92	84,57	56,21	79,34	63,38
AUT	78,07	55,56	86,46	50,69	84,23	54,90
BEL	74,59	47,36	86,44	24,87	80,33	53,45
CAN	74,55	56,45	83,11	50,13	79,77	60,32
CHE	82,11	68,79	89,37	58,42	87,01	73,39
CHL				16,52	63,72	53,50
COL	73,82	69,47	81,36	33,79	66,39	52,94
CZE	83,99	55,28	86,61	24,80	86,35	68,24
DEU	82,79	61,94	89,35	47,87	85,56	71,38
DNK	82,05	60,75	87,61	52,79	81,55	71,39
ESP	71,92	58,54	81,93	17,60	72,74	54,67
EST	80,63	62,39	86,62	33,09	82,24	72,71
FIN	76,59	54,95	86,41	40,58	82,01	67,36
FRA	73,19	52,53	85,56	28,69	80,38	53,55
GBR	81,38	65,40	86,63	47,01	84,67	65,50
GRC	62,34	52,41	75,97	14,16	70,23	45,32
HUN	79,94	56,99	85,80	27,62	82,99	60,05
IRL	74,73	53,41	85,59	37,27	78,43	61,68
ISL	86,34	76,03	91,63	68,53	86,55	80,92
ISR	73,46	51,08	87,69	37,62	77,89	67,62
ITA	71,45	52,82	81,40	16,71	69,21	53,81
JPN			86,02	46,42	85,05	76,74
KOR	72,21	63,63	77,67	25,40	74,58	66,55
LTU	74,77	55,21	91,26	27,56	82,52	67,62
LUX	74,74	61,63	85,57	25,80	83,23	44,07
LVA	75,15	64,92	89,26	30,07	82,15	68,62
MEX	71,53	65,60	79,70			
NLD	82,32	63,33	89,67	61,26	84,79	71,03

Pays	DCSD (en %)	PCSD (en %)	SUP (en %)	15 à 24 (en %)	25 à 54 (en %)	55 à 64 (en %)
NOR	80,79	61,54	89,50	49,21	82,28	72,42
NZL	82,95	71,74	88,31	53,48	83,48	76,07
POL	70,70	44,96	88,98	27,45	83,18	52,22
PRT	84,75	70,32	88,55	21,56	83,68	61,27
SVK	78,39	38,04	83,84	22,44	80,38	59,00
SVN	76,83	50,62	90,31	26,45	87,81	49,52
SWE	85,59	67,00	90,27	38,46	84,45	77,84
TUR	60,03	50,08	73,62	28,74	57,73	31,11
USA	71,05	56,06	82,68	45,00	74,75	60,15

Sources OCDE (2021) (Consultés le 11 février 2021).

Construire le diagramme à barres verticales des fréquences, le diagramme à points horizontal des fréquences et le diagramme circulaire pour chacune de ces six séries statistiques.

2.4. INDICES SYNTHÉTIQUES ET COURBE DE LORENZ

2.4.1 *Indices synthétiques*

A. Problématique

Considérons n biens ($i = 1,\ldots, n$) pour lesquels nous souhaitons étudier l'évolution en prix, en quantité et en valeur entre deux dates (0) et (t).

Notons encore $p_i(t)$ et $q_i(t)$ le prix et la quantité du bien i à une date (t).

Supposons alors que nous souhaitions déterminer un indice global capable de mesurer une variation de valeur (prix × quantité) des n biens entre une date (t) et une date de référence (0). Un indice synthétique naturel serait :

$$I_{t/0}(pq) = \frac{\sum_{i=1}^{n} p_i(t)\, q_i(t)}{\sum_{i=1}^{n} p_i(0)\, q_i(0)} \times 100.$$

Mais une telle formule soulève un problème : elle ne permet pas de savoir si la variation de valeur observée a pour origine une variation des prix, des quantités, ou des deux grandeurs simultanément. Pour lever l'ambiguïté, nous fixons arbitrairement une des deux grandeurs, quantité ou prix, pendant que nous laissons l'autre évoluer ; nous obtenons de cette façon respectivement les indices synthétiques de Laspeyres et de Paasche.

B. Indice de Laspeyres

Dans la formule précédente, fixons les quantités à la date de référence (0) et laissons évoluer les prix. Nous obtenons la formule suivante, appelée **indice de Laspeyres des prix** :

$$L_{t/0}(p) = \frac{\sum_{i=1}^{n} p_i(t)\, q_i(0)}{\sum_{i=1}^{n} p_i(0)\, q_i(0)} \times 100.$$

Selon un même raisonnement, nous pouvons définir l'**indice de Laspeyres des quantités** en fixant les prix à la date de référence (0) et en laissant évoluer les quantités :

$$L_{t/0}(q) = \frac{\sum_{i=1}^{n} p_i(0)\, q_i(t)}{\sum_{i=1}^{n} p_i(0)\, q_i(0)} \times 100.$$

Notons que l'indice de Laspeyres des prix peut également s'exprimer de la façon suivante :

$$L_{t/0}(p) = \frac{\sum_{i=1}^{n} p_i(t)\, q_i(0)}{\sum_{i=1}^{n} p_i(0)\, q_i(0)} \times 100 = \frac{\sum_{i=1}^{n} p_i(0)\, q_i(0) \times \dfrac{p_i(t)}{p_i(0)}}{\sum_{i=1}^{n} p_i(0)\, q_i(0)} \times 100$$

qui s'interprète comme la moyenne arithmétique de la série des indices élémentaires des prix, pondérée par les valeurs des biens à la date de référence. De même, l'indice de Laspeyres des quantités s'interprétera comme la moyenne arithmétique de la série des indices élémentaires des quantités, pondérée par les valeurs des biens à la date de référence. **Il prend comme référence les données de base, c'est-à-dire du passé.**

C. Indice de Paasche

Si nous fixons maintenant les quantités à la date courante (t) et laissons évoluer les prix, nous obtenons **l'indice de Paasche des prix** :

$$P_{t/0}(p) = \frac{\sum_{i=1}^{n} p_i(t)\, q_i(t)}{\sum_{i=1}^{n} p_i(0)\, q_i(t)} \times 100.$$

Quant à **l'indice de Paasche des quantités**, il s'obtient en fixant les prix à la date courante (t) :

$$P_{t/0}(q) = \frac{\sum_{i=1}^{n} p_i(t)\, q_i(t)}{\sum_{i=1}^{n} p_i(t)\, q_i(0)} \times 100.$$

Il prend comme référence les données actuelles, c'est-à-dire du présent.

D. Indices de Fisher

Le choix entre les deux indices de Laspeyres et Paasche présente un certain arbitraire. Dans les années 1920, un statisticien américain Fisher a proposé un indice combiné qui est la moyenne géométrique des indices Laspeyres et Paasche. Les **indices de Fisher** mesurant respectivement l'évolution **des prix** et **des quantités** sont par conséquent définis par les formules suivantes :

$$F_{t/0}(p) = \sqrt{L_{t/0}(p) \times P_{t/0}(p)}$$
$$F_{t/0}(q) = \sqrt{L_{t/0}(q) \times P_{t/0}(q)}.$$

Exemple

Le tableau 2.9 donne pour les années 2001 et 2020 les valeurs du SMIC (Salaire Minimum Interprofessionnel de Croissance) horaire brut en euros ainsi que les heures supplémentaires effectuées par un salarié qui relevait, en 2001, du régime général de bonification des heures supplémentaires : les huit premières heures supplémentaires (de la 36e à la 43e incluse) étaient majorées de 25 %. En 2020, ce même salarié relevait d'un système bonifié de 30 % pour les huit premières heures supplémentaires. Les durées légales de travail prennent en compte le passage aux 35 heures. Comment définir un « bon » indice de salaire en 2020, base 100 en 2001 ?

Tableau 2.9 : *évolution des durées légales du travail et du SMIC horaire*

Heures	SMIC 2001 (p_0)	Quantité 2001 (q_0)	SMIC 2020 (p_1)	Quantité 2020 (q_1)
Légales	6,6700 €	169 h	10,150 €	151,67 h
Supplémentaires	8,3375 €	3 h	13,195 €	4,00 h

Une personne travaille les quantités horaires indiquées dans le tableau 2.9. Calculons son salaire total en 2001 et en 2020 :

$$IS_{2001} = 169 * 6{,}67 + 3 * 8{,}3375 = 1152{,}243 \text{ €}$$
$$IS_{2020} = 151{,}67 * 10{,}15 + 4 * 13{,}195 = 1605{,}425 \text{ €}$$

Un indice de salaire global serait alors :

$$\frac{IS_{2020}}{IS_{2001}} = 1{,}3933$$

et indiquerait une augmentation de 39,33 %. Toutefois, cet indice ne tient pas compte de la modification simultanée des conditions de rémunération et des quantités de temps travaillées. Fixer ces quantités à leur niveau pris l'année de base, revient à privilégier le mode de travail du salarié en 2001 et à former l'indice de Laspeyres des salaires horaires :

$$L_{2020/2001}(p) = \frac{10{,}150 * 169 + 13{,}195 * 3}{6{,}67 * 169 + 8{,}3375 * 3} \times 100 = 1{,}5231.$$

Fixer ces quantités à leur niveau pris l'année courante, revient à privilégier le mode de travail du salarié en 2020 et à former l'indice Paasche des salaires horaires :

$$P_{2020/2001}(p) = \frac{10{,}150 * 151{,}67 + 13{,}195 * 4}{6{,}67 * 151{,}67 + 8{,}3375 * 4} \times 100 = 1{,}5241.$$

Pour aller au-delà de l'arbitraire du choix entre ces deux indices, nous calculons l'indice de Fisher :

$$F_{t/0}(p) = \sqrt{L_{t/0}(p) \times P_{t/0}(p)} = 1{,}5236.$$

L'indice de Fisher indique une augmentation de 52,36 %. Pour cet exemple, il est très proche des indices de Laspeyres et de Paasche, mais ce n'est pas nécessairement le cas.

2.4.2 *Courbe de Lorenz*

A. Qu'est-ce que la courbe de Lorenz ?

Définition 2.20 Courbe de Lorenz

La **courbe de Lorenz** est une représentation graphique permettant d'apprécier visuellement l'inégalité de la répartition des revenus (ou de toute autre richesse) dans une population.

Son principe est d'associer :

1) à une part donnée des ménages dont les revenus sont les plus faibles (en abscisse),

2) la part totale des revenus qu'ils possèdent (en ordonnée).

Exemple

Le tableau 2.10 indique la répartition des revenus en Belgique et en France en fraction du revenu total de chaque pays.

Tableau 2.10 : *déciles de la répartition des revenus en Belgique et en France*

Décile\Pays	Belgique (revenus en fraction du total)		France (revenus en fraction du total)	
	Brut	Cumulé	Brut	Cumulé
Premier	3,9	3,9	3,7	3,7
Deuxième	5,6	9,5	5,3	9,0
Troisième	6,7	16,2	6,4	15,4
Quatrième	7,8	24,0	7,3	22,7
Cinquième	8,8	32,8	8,2	30,9
Sixième	9,8	42,6	9,0	39,9
Septième	10,8	53,4	10,0	49,9
Huitième	12,1	65,5	11,4	61,3
Neuvième	13,9	79,4	13,7	75,0
Dixième	20,6	100,0	25,0	100,0

Source : Eurostat (2019).

La figure 2.14 représente les courbes de Lorenz pour le revenu des ménages belges et français en 2019.

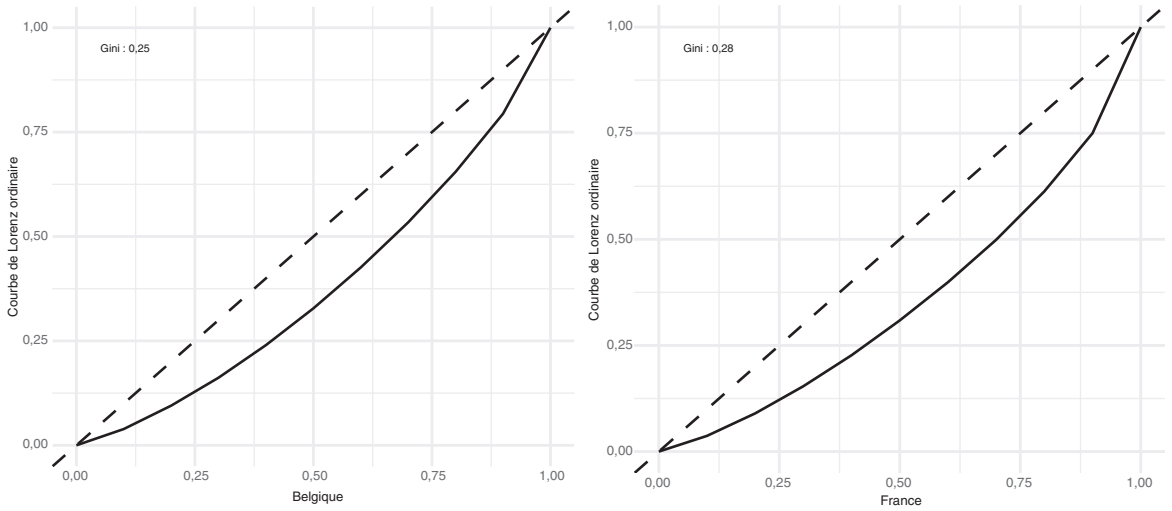

Figure 2.14 : *courbes de Lorenz pour la répartition des revenus en Belgique et en France en 2019*

Sur ce graphique, nous observons par exemple que 50 % des ménages belges ne perçoivent que 32,8 % des revenus et que 50 % des ménages français ne perçoivent 30,9 % des revenus. La bissectrice représente la situation d'égalité parfaite : chaque ménage perçoit exactement le même revenu. L'aire A située entre la courbe de Lorenz et la bissectrice est une mesure de l'importance de l'inégalité de la répartition des revenus (cf. le coefficient de Gini ci-dessous).

B. Construction de la courbe de Lorenz

La courbe est construite selon les étapes suivantes :

1) les ménages sont tout d'abord ordonnés par salaires croissants S_i, $i = 1,\ldots, n$. En pratique, il s'agit souvent de tranches de salaire $[S_{i-1}; S_i[$ (obtenues par exemple par la méthode des déciles) ;
2) nous calculons la fréquence ou part de la population f_i associée à un niveau de salaire donné (ou à une tranche de salaire) ;
3) puis enfin, la fréquence croissante (ou part cumulée) des salariés $F_i = \sum_{k \leq i} f_k$ associée à chaque niveau ou tranche de salaire.

D'autre part, nous calculons pour chaque tranche de salaire :

- la masse salariale $f_i \times S_i$. Lorsque les salaires sont donnés sous forme d'intervalle de valeurs $[S_{i-1}; S_i[$, nous retenons comme niveau de salaire le centre de classe $\frac{S_{i-1} + S_i}{2}$;
- le pourcentage de masse salariale dans la population totale : $g_i = \frac{f_i \times S_i}{\sum_{i=1}^{n} (f_i \times S_i)}$;
- le pourcentage croissant (ou part cumulée) de masse salariale dans la population : $G_i = \sum_{k \leq i} g_k$.

La courbe de Lorenz est alors obtenue en reliant les points de coordonnées (F_i, G_i) par des segments de droite.

C. Le coefficient de Gini

L'écart de la courbe de Lorenz à la bissectrice traduit l'importance des inégalités salariales et se mesure usuellement à l'aide du **coefficient de Gini**, noté G.

Définition 2.21 Coefficient de Gini

Le **coefficient de Gini**, noté G, est défini comme le rapport de l'aire A comprise entre la courbe de Lorenz et la bissectrice (cf. figure 2.14) à l'aire totale du triangle située sous la bissectrice (égale à 1/2) D'où $G = 2A$. Par ailleurs, nous avons bien sûr $G \in [0\,;\,1]$.

En pratique, la valeur de G est déterminée en retranchant à $\frac{1}{2}$ la somme des aires des trapèzes successifs situés sous la courbe de Lorenz. Il est possible alors de montrer que :

$$G = 1 - \sum_{i=1}^{n} (F_i - F_{i-1})(G_{i-1} + G_i)$$

où F_i et G_i sont les parts cumulées de salariés et de masse salariale définies plus haut (avec la convention $F_0 = G_0 = 0$).

Exemple

Les indices de Gini pour les revenus belges et français sont de 0,25 et de 0,28.

D. La médiale

Définition 2.22 Médiale

La **médiale**, notée Ml, est le niveau de salaire qui partage en deux parties égales la masse salariale. La médiale est donc telle que :

$$\sum_{s_k < Ml} f_k S_k = \sum_{s_k > Ml} f_k S_k .$$

Comme pour la médiane, lorsque le salaire est réparti en classes, la médiale est déterminée usuellement par interpolation linéaire.

Propriété

La médiale est toujours supérieure à la médiane.

La médiale est également à l'origine de la mesure de concentration des salaires suivante :

$$\frac{Ml - Me}{\max\{S_i\} - \min\{S_i\}}$$

où Me est le salaire médian. Exprimée en pourcentage, cette mesure est d'autant plus élevée que les inégalités salariales sont importantes.

2.5. CARACTÉRISTIQUES DE POSITION

2.5.1 *Moyenne arithmétique*

Définition 2.23 Moyenne arithmétique

Notons N l'effectif de la population étudiée. La **moyenne arithmétique**, notée $\mu(X)$, d'une série statistique simple, discrète ou continue, dont la distribution statistique est non groupée, $(x_i,n_i)_{i=1,\ldots,p}$, est le nombre réel défini par :

$$\mu(X) = \sum_{i=1}^{p} x_i f_i = \frac{1}{N}\sum_{i=1}^{p} x_i n_i.$$

Si la distribution est groupée, $(]a_i\ ;\ a_{i+1}], f_i)_{i=1,\ \ldots,\ p}$, la **moyenne arithmétique** se calcule par :

$$\mu(X) = \sum_{i=1}^{p} \frac{a_i + a_{i+1}}{2} f_i.$$

Remarque 2.5

Il existe d'autres moyennes comme la moyenne géométrique, la moyenne harmonique, la moyenne quadratique, la moyenne interquartile ou bien les moyennes tronquées.

2.5.2 *Minimum et maximum*

Définition 2.24 Minimum

Le **minimum**, noté $\min(X)$, d'une série statistique est sa plus petite valeur.

Définition 2.25 Maximum

Le **maximum**, noté $\max(X)$, d'une série statistique est sa plus grande valeur.

2.5.3 *Statistique d'ordre, médiane et classe médiane*

A. Cas d'une distribution non groupée

Considérons X une série statistique simple, discrète ou continue, **non groupée**.

Définition 2.26 Statistique d'ordre

$x_{(m+1)}$ signifie la $(m + 1)^e$ valeur lorsque la série statistique est classée par ordre croissant. Elle est appelée la $(m + 1)^e$ **statistique d'ordre** de la série statistique.

Définition 2.27 Médiane

1. Premier cas : N est un entier impair et de ce fait $N = 2m + 1$, où m est un entier.

La **médiane**, notée $Q_2(X)$ ou $Q_{0,5}(X)$, d'une série statistique est définie par :

$$Q_2(X) = x_{(m+1)}$$

où $x_{(m+1)}$ est la $(m + 1)^e$ statistique d'ordre de la série statistique.

2. Deuxième cas : N est un entier pair et de ce fait $N = 2m$, où m est un entier.

La **médiane**, notée $Q_2(X)$ ou $Q_{0,5}(X)$, d'une série statistique peut être prise égale à toute valeur supérieure ou égale à $x_{(m)}$ et inférieure ou égale à $x_{(m+1)}$. Toutefois, il est d'usage, bien qu'arbitraire, de prendre la moyenne arithmétique de ces deux valeurs comme médiane :

$$Q_2(X) = \frac{(x_{(m)} + x_{(m+1)})}{2}.$$

B. Cas d'une distribution groupée

Considérons X une distribution statistique simple, discrète ou continue, **groupée**.

Définition 2.28 Classe médiane

Nous déterminons d'abord, à partir des fréquences cumulées, la classe contenant la médiane, *i.e.* la classe $]a_i ; a_{i+1}]$ telle que $f_{i-1} < 0,5$ et $f_i \geq 0,5$, appelée **classe médiane**, puis la valeur de la médiane par interpolation linéaire dans cette classe, ce qui revient à supposer que les valeurs sont équiréparties au sein de cette classe.

2.5.4 *Quartiles*

Considérons X une distribution statistique simple, discrète ou continue, **non groupée**.

Définition 2.29 Premier quartile

1. Premier cas : l'entier N n'est pas un multiple de 4. Soit m la partie entière de $\frac{N}{4}$.

Le **premier quartile**, noté $Q_1(X)$ ou $Q_{0,25}(X)$, d'une série statistique est défini par :

$$Q_1(X) = x_{(m+1)}.$$

2. Deuxième cas : l'entier N est un multiple de 4 et de ce fait, $N = 4m$.

Le **premier quartile**, noté $Q_1(X)$ ou $Q_{0.25}(X)$, d'une série statistique peut être pris égal à toute valeur supérieure ou égale à $x_{(m)}$ et inférieure ou égale à $x_{(m+1)}$. Toutefois, il est d'usage, bien qu'arbitraire, de prendre la moyenne arithmétique de ces deux valeurs comme premier quartile :

$$Q_1(X) = \frac{x_{(m)} + x_{(m+1)}}{2}.$$

Définition 2.30 Troisième quartile

1. Premier cas : l'entier N n'est pas un multiple de 4. Soit m' la partie entière de $\frac{3N}{4}$.
Le **troisième quartile**, noté $Q_3(X)$ ou $Q_{0,75}(X)$, d'une série statistique est défini par :

$$Q_3(X) = x_{(m'+1)}.$$

2. Deuxième cas : l'entier N est un multiple de 4 et de ce fait, $N = 4m$. Posons $m' = \frac{3N}{4}$.
Le **troisième quartile**, noté $Q_3(X)$ ou $Q_{0.75}(X)$, d'une série statistique peut être pris égal à toute valeur supérieure ou égale à $x_{(m')}$ et inférieure ou égale à $x_{(m'+1)}$. Toutefois, il est d'usage, bien qu'arbitraire, de prendre la moyenne arithmétique de ces deux valeurs comme troisième quartile :

$$Q_3(X) = \frac{x_{(m')} + x_{(m'+1)}}{2}.$$

Considérons une distribution statistique simple, discrète ou continue, groupée.

Définition 2.31 Classes des quartiles

Nous déterminons d'abord, à partir des fréquences cumulées, la classe contenant le premier (respectivement le troisième) quartile, *i.e.* la classe $]a_i\ ;\ a_{i+1}]$ telle que $f_{i-1} < \frac{1}{4}$ et $f_i \geq \frac{1}{4}$ (resp. la classe $]a_i\ ;\ a_{i+1}]$ telle que $f_{i-1} < \frac{3}{4}$ et $f_i \geq \frac{3}{4}$), appelée **classe du premier** (respectivement **du troisième**) **quartile**, puis la valeur du premier (respectivement du troisième) quartile par interpolation linéaire dans cette classe, ce qui revient à supposer que les valeurs sont équiréparties au sein de cette classe.

Remarque 2.6

Le deuxième quartile est la médiane.

2.5.5 *Mode et classe modale*

Définition 2.32 Mode et classe modale

Un **mode**, noté Mo(X), d'une distribution statistique simple discrète *non groupée* est l'une des valeurs $x_1, x_2, \ldots, x_p$ dont la fréquence est maximale.

Pour une distribution statistique simple discrète *groupée*, une **classe modale**, Mo(X), est une classe de densité, c'est-à-dire de rapport fréquence/longueur, maximale. Un mode est alors le centre de l'une de ces classes, ce qui revient à supposer que les valeurs sont équiréparties au sein de chaque classe.

Remarque 2.7

1) Une distribution statistique simple, discrète ou continue, est **unimodale** si elle a un seul mode.
2) Si elle en a plusieurs, elle est **plurimodale** (**bimodale**, **trimodale**, etc.).

Exemple

Le tableau 2.11 contient des données relatives au temps de travail pour 35 pays, membres de l'Union européenne ou proches de l'Europe, en 2019

Tableau 2.11 : *temps de travail en Europe et dans des pays proches en 2019*

États	Travailleurs à temps partiel (en % de l'emploi total)			Durée de travail par semaine des travailleurs à temps plein (en heures)	
	Ensemble	Hommes	Femmes	Salariés	Non salariés
Allemagne	27,2	9,9	46,7	40,2	47,6
Autriche	27,2	9,5	47,1	41,1	51,2
Belgique	24,9	10,5	41,0	39,1	52,5
Bulgarie	1,9	1,7	2,1	40,7	43,0
Chypre	10,2	6,3	14,6	41,2	44,9
Croatie	4,8	3,1	6,7	40,4	44,6
Danemark	24,2	15,3	33,9	37,6	46,3
Espagne	14,5	6,8	23,7	39,6	47,0
Estonie	11,3	7,1	15,9	40,2	43,2
Finlande	15,5	10,1	21,3	39,4	46,4
France	17,5	7,5	28,0	39,1	50,1
Grèce	9,1	5,9	13,5	40,7	50,4
Hongrie	4,4	2,5	6,8	40,4	41,5
Irlande	19,7	10,1	30,6	39,4	49,4
Islande	21,5	10,3	34,1	43,4	48,5
Italie	18,7	8,2	32,9	39,0	46,0
Lettonie	8,4	5,8	10,9	40,2	41,5
Lituanie	6,4	4,7	8,0	39,9	40,6
Luxembourg	17,0	5,6	30,4	40,2	46,9
Macédoine du Nord	4,1	4,1	4,3	41,6	45,3
Malte	12,2	5,9	21,4	41,3	47,4
Monténégro	4,5	4,7	4,1	44,6	48,3
Norvège	25,8	15,2	37,7	38,4	45,0
Pays-Bas	50,2	27,9	75,2	38,9	47,9

États	Travailleurs à temps partiel (en % de l'emploi total)			Durée de travail par semaine des travailleurs à temps plein (en heures)	
	Ensemble	Hommes	Femmes	Salariés	Non salariés
Pologne	6,1	3,5	9,3	40,8	45,7
Portugal	8,1	5,4	10,9	40,8	47,6
Roumanie	6,1	6,0	6,2	40,6	38,9
Royaume-Uni	24,4	10,8	39,4	42,0	45,2
Serbie	9,7	8,9	10,6	42,8	51,2
Slovaquie	4,5	2,9	6,5	40,5	45,1
Slovénie	8,4	4,8	12,7	40,7	45,7
Suède	22,5	13,4	32,5	39,9	47,9
Suisse	38,0	17,1	61,7	41,8	49,4
Tchéquie	6,3	2,8	10,6	40,7	45,9
Turquie	9,9	6,6	17,0	48,1	49,6

Source : Eurostat.

Les caractéristiques de position de la variable « Durée du travail par semaine des travailleurs à temps plein », sont les suivantes :
Le minimum est $\min(X) = 37{,}60$ h ; le maximum est $\max(X) = 48{,}10$ h ; la moyenne arithmétique est $\mu(X) = 40{,}72$ h ; la médiane est $Q_{0,5}(X) = 40{,}50$ h ; le premier quartile est $Q_{0,25}(X) = 39{,}60$ h et le troisième quartile est $Q_{0,75}(X) = 41{,}20$ h.

Exemple

Considérons la série statistique groupée donnée par les lignes du tableau 2.4, correspondant à la répartition en classes d'âge de la population du Canada en 2020, en se restreignant aux âges compris entre 0 et 100 ans. Les caractéristiques de position de la variable introduite dans le tableau 2.4 de la section 2.3, sont les suivantes :

Le minimum est $\min(X) = 0$ ans ; le maximum est $\max(X) = 100$ ans ; la moyenne arithmétique est $\mu(X) = 41{,}43$ ans.

La classe médiane peut être déterminée à partir de la distribution des fréquences cumulées :

0 à 4	5 à 9	10 à 14	15 à 19	20 à 24	25 à 29	30 à 34	35 à 39	40 à 44	45 à 49
0,0506	0,1044	0,1589	0,2142	0,2796	0,3492	0,4193	0,4885	0,5534	0,6163
50 à 54	55 à 59	60 à 64	65 à 69	70 à 74	75 à 79	80 à 84	85 à 89	90 à 94	95 à 99
0,6808	0,7530	0,8204	0,8774	0,9244	0,9565	0,9779	0,9915	0,9980	1,0000

« Une présentation synthétique et rigoureuse de l'ensemble des connaissances en statistiques ! »

Claire Borsenberger
Docteur en économie, diplômée de la Toulouse School of Economics, Claire est responsable du département « Doctrine et modélisation », au sein de la Direction de la Régulation et des Études du groupe La Poste.

Frédéric Bertrand
Docteur en mathématiques appliquées de l'Université Louis Pasteur et habilité à diriger des recherches de l'Université de Strasbourg, Frédéric est professeur à l'Université de technologie de Troyes et travaille depuis 20 ans sur la statistique.

Christian Derquenne
Docteur en mathématiques appliquées, Christian est chercheur senior à la Direction Recherche et Développement d'Électricité de France. Il travaille dans un département spécialisé en management de l'énergie.

Gilles Dufrénot
Professeur à l'Université d'Aix-Marseille, Gilles travaille depuis 20 ans sur la modélisation statistique des données macroéconomiques et financières.

Fredj Jawadi
Docteur en sciences économiques de l'Université de Paris Ouest Nanterre et habilité à diriger des recherches en économétrie de l'Université d'Aix-Marseille, Fredj est Professeur d'économétrie et de finance à l'Université de Lille.

Myriam Maumy
Docteur en statistique de l'Université Pierre et Marie Curie et habilitée à diriger des recherches en statistiques de l'Université de Strasbourg, Myriam est Maîtresse de Conférences hors classe à l'Université de technologie de Troyes. Elle travaille depuis plus de 20 ans sur la statistique et l'échantillonnage.

Ce livre offre aux **étudiants de 1er cycle** une **introduction aux concepts indispensables en statistiques et à leurs applications**. Il comporte douze chapitres traitant de **façon simple et claire les sujets majeurs en statistiques** :

- statistiques descriptives et visualisation des données
- analyses factorielles et classification
- théorie des sondages
- théorie des probabilités
- tests d'hypothèses
- analyse de la régression
- modèles logit et probit
- séries temporelles.

La compréhension des concepts statistiques présentés ne requiert que la connaissance de l'algèbre. L'un des atouts de ce livre est son caractère appliqué : des **exemples concrets illustrent les concepts statistiques** présentés dans chaque chapitre.

L'objectif est de montrer aux étudiants comment les statistiques participent à la prise de décision des autorités publiques et des chefs d'entreprises. L'accent est mis sur **l'utilisation pratique des différents outils statistiques** grâce à la présentation des **techniques de programmation** et des **exemples sous R**.

OFFERT : RESSOURCES NUMÉRIQUES

Un *package* contenant les codes R des exemples du livre ainsi que les jeux de données est disponible en ligne.

Étudiants :
- Corrigés des exercices.
- Approfondissements et exercices complémentaires.

Enseignants : PPT reprenant les figures du manuel.

Dans la même collection :

ISBN 978-2-8073-1944-8
ISSN 2030-501X

deboeck SUPÉRIEUR

www.deboecksuperieur.com

Conception graphique : Primo&Primo®