

STATISTIQUE THÉORIQUE ET APPLIQUÉE

Tome 1

Statistique descriptive
et bases de l'inférence statistique

Pierre Dagnelie

INTRODUCTIONS DES DIFFÉRENTS CHAPITRES

Bruxelles, De Boeck et Larcier, 2007, 511 p.

Distributeur : Accès+, Fond Jean-Pâques 4, B-1348 Louvain-la-Neuve (Belgique).

Tél. : 32 (0)10 48 25 00 – Fax : 32 (0)10 48 25 19

E-mail : acces+cde@deboeck.be – Site web : www.deboeck.com

ISBN 978-2-8041-5312-0

Chapitre 1

Introduction générale

Sommaire ¹

- ⊕ 1.1 Définition
 - ⊕ 1.2 Historique
 - ⊕ 1.3 Cadre général
 - 1.4 Documentation complémentaire
- Principaux mots-clés

¹ Nous rappelons que, dans les sommaires des différents chapitres, le signe ⊕ indique les paragraphes considérés comme *entièrement ou partiellement* de première importance, au sens du « mode d'emploi » qui suit la table des matières. Les signes ⊕ qui apparaissent en marge dans la suite de ce chapitre montrent par exemple que les paragraphes 1.1 et 1.3 devraient être pris en considération dans leur ensemble (à l'exclusion, le cas échéant, des alinéas marqués par les symboles [et]), tandis qu'en ce qui concerne le paragraphe 1.2, seules les sections 1.2.3 et 1.2.4 devraient retenir l'attention au cours d'une première lecture.

⊕ 1.1 Définition

Dérivé du substantif latin *status* (État), le mot statistique possède, en français comme dans d'autres langues, plusieurs significations distinctes.

D'une part, utilisé le plus souvent au pluriel, le terme *statistiques* désigne tout ensemble cohérent de *données*, généralement numériques, relatives à un groupe d'individus ou d'objets. On parle par exemple de la ou des statistiques de la production agricole ou industrielle (quantités produites, prix de vente, coûts de production, etc.), des statistiques démographiques (natalité, mortalité, etc.), des statistiques du chômage, des statistiques des accidents de la circulation routière, etc. Il convient toutefois de remarquer que, contrairement à une opinion communément admise, cette acception du terme statistique ne concerne pas seulement des volumes importants de données.

D'autre part, le mot *statistique* désigne l'ensemble des *méthodes* qui permettent de recueillir et d'analyser les données dont il vient d'être question. C'est à cette signification que nous nous référons dans le présent ouvrage.

Enfin, le terme *statistique* est aussi utilisé parfois pour désigner l'un ou l'autre *paramètre*, tel qu'une moyenne, calculé à partir d'un ensemble de données².

Dans la première définition que nous avons présentée, le qualificatif « numériques » doit être considéré dans un sens très large. Il peut en effet concerner aussi bien des données quantitatives (résultats de comptages ou de mesures), que des données qualitatives (couleurs, appréciations gustatives, etc.), voire même des textes, codés sous forme numérique en vue d'un traitement informatique.

Informations complémentaires : BARTHOLOMEW [1995], DODGE [2004a], DUMAS [1955], RASCH *et al.* [1994], WILLCOX [1935].

² Les traductions anglaises sont d'une part *statistics*, à la fois pour des ensembles de données et pour l'ensemble des méthodes, et d'autre part *statistic*, pour des paramètres.

Chapitre 2

La collecte des données

Sommaire

- ⊕ **2.1** Introduction
- ⊕ **2.2** L'étude par enquête
- ⊕ **2.3** L'expérimentation
- ⊕ **2.4** La nature, l'enregistrement et le traitement des données
Principaux mots-clés

⊕ 2.1 Introduction

1° Comme nous l'avons signalé antérieurement (§ 1.3.2), nous consacrons ce chapitre 2 à la présentation, en termes très simples, de notions de base relatives à la collecte des données, c'est-à-dire à ce qui constitue normalement la première phase de toute étude statistique.

Nous envisagerons successivement les questions qui concernent les études par *enquête* (§ 2.2), les problèmes d'*expérimentation* (§ 2.3), et les questions relatives à la *nature*, à l'*enregistrement* et au *traitement des données* (§ 2.4). Nous reviendrons ultérieurement de façon plus détaillée sur certains de ces sujets, lorsque nous aurons présenté diverses notions de calcul des probabilités et de statistique théorique.

2° L'étude par enquête et l'expérimentation doivent normalement être organisées, l'une et l'autre, dans des conditions telles que de nombreux éléments (choix des unités ou des individus observés, affectation aux différentes unités expérimentales des différents traitements qui sont comparés, etc.) soient parfaitement maîtrisés. Dans certains cas, et notamment dans certaines enquêtes rétrospectives, les circonstances ne permettent pas de maîtriser de tels éléments. L'étude est alors basée sur une simple accumulation d'observations, sans structure ou sans ordre préétabli.

On parle dans ce cas d'*étude par observation*¹. Nous ne traitons pas ce sujet dans le présent ouvrage.

Informations complémentaires : en ce qui concerne l'observation par enquête, ARDILLY [2006], BARNETT [2002], DUSSAIX et GROSBAS [1993], THOMPSON [2002], TILLÉ [2001]; en ce qui concerne l'expérimentation, DAGNELIE [2003], FLEISS [1999], GOUPY et CREIGHTON [2006], KUEHL [2000], MONTGOMERY [2005]; en ce qui concerne l'étude par observation : KISH [2004], ROSENBAUM [2002], SMITH et SUGDEN [1988].

¹ En anglais : *observational study, uncontrolled observational study*.

Chapitre 3

La statistique descriptive à une dimension

Sommaire

- ⊕ 3.1 Introduction
 - ⊕ 3.2 Les distributions de fréquences
 - ⊕ 3.3 Les représentations graphiques
 - ⊕ 3.4 La réduction des données : généralités
 - ⊕ 3.5 Les paramètres de position
 - ⊕ 3.6 Les paramètres de dispersion
 - 3.7 Les moments et les paramètres de dissymétrie et d'aplatissement
 - ⊕ 3.8 Le calcul de la moyenne, de la variance et des moments d'ordre 3 et 4
 - ⊕ 3.9 Quelques informations relatives à l'exécution des calculs
 - 3.10 Les nombres-indices
- Principaux mots-clés
- Exercices

⊕ 3.1 Introduction

1° La *statistique descriptive*¹ a essentiellement pour but de présenter les données observées sous une forme telle qu'on puisse en prendre connaissance facilement. Elle peut concerner une variable ou une caractéristique à la fois, deux variables ou deux caractéristiques à la fois, ou encore plus de deux variables ou plus de deux caractéristiques simultanément. Selon les cas, on parle de statistique descriptive à *une variable* ou à *une dimension*², de statistique descriptive à *deux variables* ou à *deux dimensions*³, et de statistique descriptive à *plusieurs variables* ou à *plusieurs dimensions*⁴.

2° À une dimension, le but de simplification de la statistique descriptive peut être atteint en condensant les observations sous trois formes distinctes.

Des tableaux statistiques permettent de présenter les données sous la forme numérique de *distributions de fréquences* (§ 3.2). Différents types de *diagrammes* permettent de présenter graphiquement ces distributions, ou les données initiales elles-mêmes (§ 3.3). Et enfin, les données peuvent également être condensées sous la forme de quelques *paramètres* ou *valeurs typiques* : le calcul de ces paramètres constitue la *réduction des données*⁵ (§ 3.4 et suivants)⁶.

La présentation des données sous forme de tableaux et de graphiques concerne plus particulièrement les cas où les observations sont assez nombreuses, tandis que la réduction des données s'applique indifféremment à tous les cas.

Informations complémentaires : ALBARELLO *et al.* [2007], ALONZO [2006], MAZEROLLE [2005].

¹ En anglais : *descriptive statistics*.

² En anglais : *univariate, one-dimensional*.

³ En anglais : *bivariate, two-dimensional*.

⁴ En anglais : *multivariate, multidimensional*.

⁵ En anglais : *data reduction*.

⁶ L'expression « réduction des données » est parfois utilisée pour désigner l'ensemble de la statistique descriptive, y compris la préparation de tableaux et de graphiques.

Chapitre 4

La statistique descriptive à deux dimensions

Sommaire

- ⊕ 4.1 Introduction
- ⊕ 4.2 Les distributions de fréquences
- ⊕ 4.3 Les représentations graphiques
- ⊕ 4.4 La réduction des données : généralités
- ⊕ 4.5 Les moments et la covariance
- ⊕ 4.6 Le coefficient de corrélation et le coefficient de détermination
- ⊕ 4.7 La régression linéaire au sens des moindres carrés
- 4.8 La régression linéaire au sens des moindres rectangles
- ⊕ 4.9 Le calcul de la covariance et des paramètres dérivés
- 4.10 La régression curvilinéaire
- 4.11 Quelques notions de statistique descriptive à plusieurs dimensions

Principaux mots-clés

Exercices

⊕ 4.1 Introduction

1° La statistique descriptive à deux dimensions a pour objet de mettre en évidence les relations qui existent entre deux séries d'observations, considérées simultanément. Ces observations peuvent être de nature qualitative ou quantitative, continue ou discontinue, etc., et il n'est nullement exclu de considérer simultanément deux séries d'observations de natures différentes (§ 2.4.1).

2° Comme en statistique descriptive à une dimension, trois aspects doivent être envisagés : l'élaboration de tableaux, permettant de condenser les données sous la forme de *distributions de fréquences* (§ 4.2), la *représentation graphique* des observations (§ 4.3), et la *réduction des données*, c'est-à-dire le calcul de paramètres servant à caractériser numériquement les relations existant entre les deux séries d'observations (§ 4.4 à 4.10).

À ces notions de statistique descriptive à deux dimensions, nous ajouterons un paragraphe consacré à la présentation de quelques éléments de *statistique descriptive à plus de deux dimensions* (§ 4.11).

Informations complémentaires : ALONZO [2006], DODGE [2004b], MAZEROLLE [2005], TOMASSONE *et al.* [1992]¹.

¹ Certains des livres mentionnés ne concernent pas spécifiquement la statistique descriptive, mais traitent également de l'inférence statistique, principalement en matière de régression.

Chapitre 5

La probabilité mathématique et les distributions théoriques : généralités

Sommaire

- ⊕ 5.1 Introduction
- ⊕ 5.2 La notion de probabilité
- ⊕ 5.3 Quelques propriétés de la probabilité mathématique
- ⊕ 5.4 La probabilité conditionnelle et l'indépendance stochastique
- ⊕ 5.5 Les notions de variable aléatoire et de distribution théorique
- ⊕ 5.6 Quelques propriétés des variables aléatoires
- ⊕ 5.7 L'espérance mathématique et ses propriétés
- ⊕ 5.8 Les paramètres des distributions théoriques à une dimension

5.9 Les fonctions génératrices et la fonction caractéristique

Principaux mots-clés

Exercices

⊕ 5.1 Introduction

1° Après avoir exposé les notions essentielles de statistique descriptive à une et à deux dimensions, au cours des deux chapitres précédents, nous consacrons cette troisième partie à des concepts plus théoriques, de probabilité et de distribution de probabilité notamment. Nous nous efforcerons de présenter ces concepts d'une manière aussi intuitive que possible, par analogie avec les éléments correspondants de la statistique descriptive.

2° Au cours de ce chapitre 5, la notion de *probabilité mathématique* est tout d'abord introduite (§ 5.2), par comparaison avec celle de fréquence relative, puis caractérisée par certaines de ses *propriétés* (§ 5.3). Cette notion nous permettra alors de définir la *probabilité conditionnelle* et l'*indépendance stochastique* (§ 5.4).

Nous présenterons ensuite les notions de *variable aléatoire* et de *distribution de probabilité* ou *distribution théorique* (§ 5.5), par analogie avec celles de variable observée et de distribution de fréquences ou distribution observée, et nous en donnerons quelques *propriétés* (§ 5.6).

Nous définirons aussi le concept d'*espérance mathématique* (§ 5.7) et les *paramètres* des distributions théoriques à une dimension, en en donnant également diverses propriétés (§ 5.8).

Enfin, nous introduirons les notions de *fonctions génératrices* et de *fonction caractéristique* (§ 5.9).

Informations complémentaires : ROSS [1998, 2004], TASSI et LEGAIT [1990].

Chapitre 6

Les principales distributions théoriques à une dimension

Sommaire

- ⊕ **6.1** Introduction
- ⊕ **6.2** Les distributions binomiales et polynomiales
- 6.3** Les distributions hypergéométriques et hypergéométriques généralisées
- ⊕ **6.4** Les distributions de POISSON
- 6.5** Quelques autres distributions discontinues
- ⊕ **6.6** Les distributions normales et log-normales
- ⊕ **6.7** Les distributions t de STUDENT
- ⊕ **6.8** Les distributions χ^2 de PEARSON
- ⊕ **6.9** Les distributions F de FISHER-SNEDECOR
- 6.10** Schéma récapitulatif et notions complémentaires

Principaux mots-clés

Exercices

⊕ 6.1 Introduction

1° Au cours de ce chapitre, nous définirons et nous caractériserons les principales distributions théoriques à une dimension.

En ce qui concerne les distributions discontinues, nous étudierons pour commencer les *distributions binomiales* (§ 6.2), les *distributions hypergéométriques* (§ 6.3) et les *distributions de POISSON* (§ 6.4). Nous présenterons aussi des généralisations de ces distributions, à savoir les *distributions polynomiales* (§ 6.2.3) et les *distributions hypergéométriques généralisées* (§ 6.3.2), et nous dirons quelques mots de diverses *distributions discontinues* (§ 6.5).

En ce qui concerne les distributions continues, nous envisagerons ensuite successivement les *distributions normales et log-normales* (§ 6.6), les *distributions t de STUDENT* (§ 6.7), les *distributions χ^2 de PEARSON* (§ 6.8) et les *distributions F de FISHER-SNEDECOR* (§ 6.9).

Enfin, nous terminerons par la présentation d'un schéma récapitulatif et de quelques *notions complémentaires* (§ 6.10).

2° Au fur et à mesure de l'étude de ces distributions, nous indiquerons les relations qui existent entre elles, en soulignant notamment le rôle central des distributions normales. Pour ne pas être trop long, nous ne donnerons cependant pas la démonstration de toutes les propriétés énoncées.

D'autre part, pour les différentes distributions, nous fournirons soit des références bibliographiques, soit des formules de calcul qui permettent d'obtenir des *valeurs numériques* particulières. Pour la distribution normale réduite, nous incluons en outre deux tables (tables I.a et I.b), tandis que des tables relatives aux distributions t , χ^2 et F sont présentées dans le deuxième tome de cette série [STAT2, tables II, III et IV]. D'autres tables peuvent être trouvées dans les recueils cités dans l'introduction générale (§ 1.4.1.2°).

Informations complémentaires : BALAKRISHNAN et NEVZOROV [2003], EVANS *et al.* [2000], JOHNSON *et al.* [1994-1995, 2005], PATIL *et al.* [1984a, 1984b].

Chapitre 7

Les distributions théoriques à deux dimensions

Sommaire

7.1 Introduction

7.2 Quelques définitions et quelques propriétés relatives aux distributions théoriques à deux dimensions

7.3 Les paramètres des distributions théoriques à deux dimensions

7.4 Les distributions normales à deux dimensions

Principaux mots-clés

Exercices

7.1 Introduction

Pour pouvoir définir l'indépendance stochastique de deux ou plusieurs variables, nous avons été amené à introduire précédemment les notions de variable aléatoire et de distribution théorique à deux dimensions (§ 5.5.3 et 5.5.4). Nous y revenons ici de façon plus approfondie.

Nous donnerons tout d'abord quelques *définitions* et quelques *propriétés* générales, relatives notamment aux distributions marginales et aux distributions conditionnelles (§ 7.2). Nous définirons ensuite les *paramètres* des distributions théoriques à deux dimensions (§ 7.3), toujours par analogie avec les distributions observées. Enfin, nous étudierons en détail une famille de distributions particulièrement importante : les *distributions normales* à deux dimensions (§ 7.4).

Informations complémentaires : HUTCHINSON et LAI [1990, 1991], KOCHERLAKOTA et KOCHERLAKOTA [1992], MARDIA [1970].

Chapitre 8

Les distributions d'échantillonnage

Sommaire

- ⊕ 8.1 Introduction
 - ⊕ 8.2 L'échantillonnage : quelques notions complémentaires
 - ⊕ 8.3 Quelques distributions d'échantillonnage
 - ⊕ 8.4 Principes généraux relatifs aux distributions d'échantillonnage
 - 8.5 Deux théorèmes de convergence
- Principaux mots-clés
- Exercices

⊕ 8.1 Introduction

1° Au cours de l'introduction générale, nous avons vu qu'après la collecte des données, l'analyse des résultats obtenus se décompose le plus souvent en une phase de statistique descriptive et une phase d'inférence statistique. La première concerne uniquement les individus réellement observés, tandis que la deuxième a pour but d'étendre autant que possible les conclusions relatives à ces individus, considérés comme constituant un échantillon, à un ensemble plus vaste, appelé population.

Nous avons ensuite consacré un chapitre à la collecte des données (par enquête et par expérimentation), deux chapitres à la statistique descriptive (à une et à deux dimensions), et trois chapitres aux distributions théoriques (à une et à deux dimensions également).

2° Notre objectif est maintenant de jeter un pont entre la statistique descriptive et la statistique théorique. Au cours de ce chapitre, nous nous efforcerons tout d'abord de caractériser l'échantillon à partir de la population, supposée connue, en présentant la notion de distribution d'échantillonnage.

Nous donnerons pour commencer quelques compléments d'informations relatifs à l'*échantillonnage*, et plus particulièrement à l'échantillonnage aléatoire et simple (§ 8.2). Après quoi, nous introduirons le concept de *distribution d'échantillonnage*, en premier lieu par quelques exemples (§ 8.3), puis d'une manière générale (§ 8.4). Enfin, nous terminerons par la présentation de deux *théorèmes de convergence* (§ 8.5)¹.

¹ Nous ne donnons pas ici de références complémentaires relatives à l'ensemble du chapitre, en raison du fait qu'il est possible de consulter à ce sujet, non seulement les ouvrages déjà cités dans l'introduction générale (§ 1.4.1.1°), mais aussi ceux qui ont été mentionnés en matière d'enquêtes au début du chapitre 2 (§ 2.1).

Chapitre 9

Les problèmes d'estimation

Sommaire

- ⊕ **9.1** Introduction
 - ⊕ **9.2** L'estimation de la moyenne et de la variance
 - ⊕ **9.3** Principes généraux de l'estimation
 - ⊕ **9.4** Les intervalles de confiance
- Principaux mots-clés
Exercices

⊕ 9.1 Introduction

Les premiers problèmes d'inférence statistique auxquels s'applique la théorie des distributions d'échantillonnage sont les problèmes d'estimation. Le but poursuivi est d'estimer, à partir d'un échantillon, la ou les valeurs numériques d'un ou plusieurs paramètres de la population considérée, et de déterminer la précision de cette ou de ces estimations.

Nous aborderons l'étude de ces problèmes par l'examen de *deux cas particuliers* : l'estimation de la moyenne et l'estimation de la variance d'une population possédant une distribution normale (§ 9.2). Nous envisagerons ensuite la question en termes de *principes généraux* (§ 9.3). Enfin, nous verrons comment on peut chiffrer, en probabilité, la précision des estimations obtenues, grâce à la notion d'*intervalle de confiance*, et nous appliquerons cette notion au cas de la moyenne d'une population normale (§ 9.4).

Informations complémentaires : CASELLA et BERGER [2002], DAUDIN *et al.* [1999], LEHMANN et CASELLA [1998].

Chapitre 10

Les tests d'hypothèses

Sommaire

- ⊕ 10.1 Introduction
- ⊕ 10.2 Les différents buts poursuivis
- ⊕ 10.3 Les principes et la réalisation des tests
- 10.4 La fonction de puissance
- Principaux mots-clés
- Exercices

⊕ 10.1 Introduction

Un second volet de l'inférence statistique est constitué par les tests d'hypothèses. Nous en précisons tout d'abord les *différents buts* (§ 10.2). Nous considérerons ensuite les *principes* et la *réalisation* pratique de tels tests, en traitant notamment en détail le problème de la comparaison des moyennes de deux populations normales (§ 10.3). Enfin, nous introduirons la notion de *fonction de puissance* et nous l'appliquerons au même cas particulier (§ 10.4).

Informations complémentaires : CASELLA et BERGER [2002], DAUDIN *et al.* [1999], FOU-CART [1991], LEHMANN et ROMANO [2005].