INTRODUCTION GENERALE

« Après l'abbé Tuet, je maudissais Bezout ;
Car, outre les pensums où l'esprit se dissout,
J'étais alors en proie à la mathématique.
Temps sombre ! enfant ému du frisson poétique,
Pauvre oiseau qui heurtais du crâne mes barreaux,
On me livrait tout vif aux chiffres, noirs bourreaux ;
On me faisait de force ingurgiter l'algèbre ;
On me liait au fond d'un Boisbertrand funèbre ;
On me tordait, depuis les ailes jusqu'au bec,
Sur l'affreux chevalet des X et des Y ;
Hélas, on me fourrait sous les os maxillaires
Le théorème orné de tous ses corollaires ;
Et je me débattais, lugubre patient
Du diviseur prêtant main-forte au quotient.
De là mes cris ».

- Victor Hugo / *Les contemplations,* Livre premier, « A propos d'Horace » (1856)

Si l'on en croit ces quelques vers tirés des *Contemplations,* Victor Hugo n'a pas gardé un souvenir très heureux de son apprentissage des mathématiques. Il n'est pas le seul. A peu près à la même époque, Flaubert indiquait, à l'entrée « Mathématiques » de ce qui allait devenir son *Dictionnaire des idées reçues* : « Dessèchent le cœur ». C'est laconique mais clair. Et bon nombre d'étudiants seraient de nos jours tentés de dire la même chose. Pourtant, la discipline n'est pas plus inaccessible qu'une autre ; c'est particulièrement vrai pour sa branche statistique. Moins abstraite, plus appliquée, celle-ci est plus abordable.

Ceci car elle étudie de nombreux phénomènes rencontrés par les managers, experts et autres décideurs. Si les managers d'une compagnie de taxis ou de VTC[1] ont besoin de connaître la longueur et la fréquence des trajets de ses clients, ils auront recours à la statistique descriptive. Connaissant l'historique d'appel des voitures, s'ils souhaitent, par exemple, connaître la probabilité que huit véhicules soient sollicités pendant la journée, ils auront recours aux lois de probabilité. S'ils cherchent à évaluer la dépense moyenne de chaque client, ils devront construire des intervalles de confiance, etc. On pourrait multiplier les exemples. Les statistiques font partie de notre quotidien et du quotidien des entreprises. Pour les appréhender, il existe diverses méthodes et techniques. Ce sont ces dernières qui rebutent un certain nombre d'étudiants et que nous nous proposons d'exposer dans cet ouvrage.

Nombreux sont ceux en effet qui se plaignent de ne pas avoir « la bosse des maths ». Même si celle-ci n'existe pas, il n'en demeure pas moins que certains sont plus à l'aise avec les chiffres que d'autres. Or, est-il vraiment plus difficile de comprendre les lois de probabilité que la philosophie de Hegel? Est-il vraiment plus difficile de comprendre les statistiques inférentielles que le fonctionnement du système bancaire? Est-il vraiment

[1] Véhicules de tourisme avec chauffeur.

plus difficile de comprendre le test du Khi-deux que d'apprendre le chinois ? Dans cet ouvrage, nous montrons qu'il n'en est rien. Pour peu que la statistique soit exposée de façon claire, sans jargonnage, on peut aisément résoudre de nombreux problèmes et, pourquoi pas, y trouver du plaisir.

L'ouvrage s'adresse donc à toutes celles et tous ceux qui, à l'université, en école de commerce ou en classes préparatoires, veulent ou ont besoin de s'y retrouver dans cette branche particulière des mathématiques qu'est la statistique. Il leur permettra de parfaire leur formation de futur manager, expert et décideur. Nous exposons les diverses facettes de la discipline de la façon la plus pédagogique possible, en donnant, autant que faire se peut, des exemples tirés de l'actualité, en privilégiant l'usage de termes simples et en limitant au maximum l'usage de termes techniques.

Contrairement au docteur Knock ou à Diafoirus, votre médecin ne vous dirait pas que vous avez contracté une affection virale contagieuse causée par un virus à acide ribonucléique de la famille des Orthomyxoviridae. Il vous dirait tout simplement que vous avez la grippe. C'est ainsi que nous présentons la statistique dans ce livre (pas comme une maladie infectieuse, non!).

Pour cela, nous procédons en cinq étapes qui constituent autant de parties. La première évoque la statistique univariée. Le premier chapitre est consacré aux définitions de base et à la façon dont l'information est synthétisée ; le second se concentre sur les représentations graphiques ; le troisième expose les principaux paramètres permettant de caractériser une série statistique. La seconde partie est quant à elle consacrée aux lois de probabilités, tant les lois expliquant les phénomènes liés à des caractères discrets (chapitre 4) que ceux liés aux caractères continus (chapitre 5). La façon dont les individus prennent des décisions fait l'objet de la troisième partie, que ce soit une décision unique sur une période (chapitre 6) ou un ensemble de décisions sur plusieurs périodes (chapitre 7), à chaque fois en situation d'incertitude, probabilisable ou pas. La statistique inférentielle est au cœur de la quatrième partie. Celle-ci est composée de trois chapitres : estimation et intervalles de confiance (chapitre 8), les tests d'hypothèse avec l'exemple des tests de comparaison (chapitre 9) et le test du Khi-deux (chapitre 10). Enfin, les deux derniers chapitres de la cinquième et dernière partie s'attachent aux relations causales entre les variables : le chapitre 11 aborde la prévision et l'ultime chapitre de l'ouvrage la régression linéaire.

PARTIE I

STATISTIQUE DESCRIPTIVE UNIVARIEE

D'après le *Global Sports Salaries Survey* 2015, les basketteurs de la National Basketball Association (ou NBA, ligue de basket-ball nord-américaine) sont les sportifs les mieux payés au monde : 4 575 918$ (environ 4 093 131€) en moyenne et par an, même s'il existe une différence entre les joueurs des Brooklyn Nets (6 249 418$ [5 590 066€]) et ceux des Philadelphia 76ers qui sont les moins bien payés de la NBA avec -toujours en moyenne annuelle- une rémunération de « seulement » 2 205 831$ (1 973 102€). A cet égard, les rémunérations accordées aux joueurs de notre ligue 1 de football semblent bien modestes : 1 492 741$ (1 335 247€) annuels en moyenne. Mais la ligue 1 française est bien plus inégalitaire. Il y existe en effet un écart de 1 à 20 entre les rémunérations des joueurs du Paris Saint-Germain (9 083 993$ soit 8 125 577€) et celles de l'équipe de Guingamp (451 450$ ou 403 819€). Avec un ratio de 2,83, les inégalités sont donc moins grandes dans la NBA. Si on ne considère que les équipes (et pas l'ensemble des championnats), les joueurs du PSG sont d'ailleurs les sportifs les mieux payés au monde.

Il existe pléthore de tels chiffres. Pour ne citer qu'eux, des organismes comme l'INSEE ou le CSA disposent dans leurs bases de données de chiffres aussi divers que les dépenses des Français en livres, journaux et périodiques ou le nombre d'écrans par foyer. Tous ces chiffres sont communément qualifiés par les statisticiens de statistiques. Et pour traiter ces statistiques, nous avons impérativement besoin de statistique. La « statistique » (au singulier) diffère en effet des « statistiques » (au pluriel). Cette dernière expression est utilisée pour désigner les données alors que la statistique représente un ensemble de méthodes pour les traiter. Dans cette partie, nous ne nous intéresserons qu'au traitement des données. « Nous procéderons en trois temps.

Nous commencerons par un aperçu du langage statistique car, comme toute discipline, la statistique a son jargon (dont, par ailleurs, nous n'abuserons pas). Pour saisir la pertinence des données étudiées, il est en outre indispensable de pouvoir les agréger, c'est-à-dire en présenter un résumé, une synthèse. Dans cette optique, nous montrerons comment les tableaux et graphiques comment rendre intelligibles les informations collectées. C'est à cela que servent les tableaux et graphiques. Définitions et tableaux font l'objet du premier chapitre. Quant aux représentations graphiques, elles seront présentées dans le second. Mais rendre intelligibles les données ne suffit pas. Il faut aussi pouvoir les traiter, caractériser leur distribution, faire ressortir ce qu'elles nous apprennent. Nous devons élaborer des paramètres qui permettent de caractériser les spécificités de chaque série de données et/ou de les comparer (par exemple, comparer les salaires des hommes et des femmes). C'est pourquoi le troisième chapitre traitera successivement des indicateurs de position, de ceux de dispersion, du diagramme en boîte et des indicateurs de forme.

CHAPITRE 1.
STATISTIQUE DESCRIPTIVE UNIVARIEE (I)
TERMINOLOGIE DE BASE ET PRESENTATION DE DISTRIBUTIONS DE DONNEES (TABLEAUX STATISTIQUES)

> *« Les statistiques, c'est comme le bikini. Ce qu'elles révèlent est suggestif. Ce qu'elles dissimulent est essentiel »* - Aaron Levenstein */ Leonard Lyons' syndicated newspaper column* (1951).

A l'issue de ce chapitre, vous devrez être capable de :

1/ Caractériser, organiser et résumer l'information pertinente contenue dans une distribution statistique

2/ Déterminer les variables statistiques et leur nature

3/ Construire des tableaux statistiques

La *statistique* est née il y a plusieurs millénaires pour décrire l'ensemble des « données » indispensables aux chefs d'États (ou de ce qui en tient lieu à l'époque) pour gouverner : population, potentiel militaire, richesse. La statistique est aujourd'hui utilisée dans des domaines aussi divers et variés que l'économie, la gestion, le marketing, la finance, la sociologie, l'assurance, la santé, le sport, etc.

Ce large éventail de domaines d'application s'explique par le fait que dès lors que l'on dispose de beaucoup d'observations sur le phénomène que l'on souhaite étudier (ensemble des abonnés d'un opérateur téléphonique, notes de satisfaction des clients d'un restaurant, etc.), il devient vite impossible de toutes les présenter alors qu'il est fondamental pour saisir la pertinence de ces dernières de pouvoir les agréger, c'est-à-dire en exposer un résumé, une synthèse. « Trop d'informations tue l'information » a-t-on coutume de dire. Il est en effet illusoire d'inspecter des centaines, des milliers voire des millions d'observations et de prétendre donner un sens, une expression à l'information recueillie. La statistique *descriptive* (*déductive*) est précisément l'instrument statistique qui permet de résumer et de présenter l'information contenue par les données collectées (appelées *statistiques* au pluriel).

L'objectif de la statistique descriptive est ainsi de rendre plus compréhensible une série d'observations en permettant de dégager les caractéristiques essentielles qui se dissimulent au sein d'une masse de données. Autrement dit, elle nous fournit l'image la plus concise et synthétique possible de la réalité (et des tendances) en mettant en exergue des caractéristiques qui ne sont pas discernables de prime abord. Ainsi, grâce à la statistique descriptive, nous obtenons un nouvel éclairage sur les données collectées : un résumé statistique qui caractérise l'essentiel.

Lorsque l'on dispose d'une base de données, la première étape consiste à dresser un portrait général de ces données : identifier les variables, ce qu'on mesure, dans quelle unité, quelle allure elles ont, etc. Il faut pour cela caractériser les données en les représentants sous forme de tableaux et de graphiques. Par exemple, le TABLEAU 1.1 ci-après donne les pourcentages de femmes et d'hommes ayant utilisé un ordinateur, en France de 2007 à 2013, durant les trois derniers mois. Le GRAPHIQUE 1 représente ces mêmes données. La courbe en pointillés décrit l'évolution du pourcentage de femmes ayant utilisé un ordinateur au cours des trois derniers mois

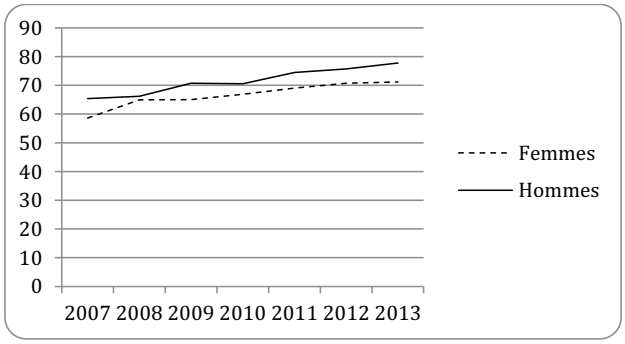
sur la période 2007-2013, la courbe en trait plein, celle des hommes. On constate globalement une tendance haussière avec un niveau plus élevé pour les hommes que pour les femmes.

TABLEAU 1.1 - Utilisation d'un ordinateur en France selon le sexe (2007-2013)

	Sexe	
Années	Femmes (%)	Hommes (%)
2007	56,8	65,4
2008	64,9	66,2
2009	65	70,7
2010	66,9	70,5
2011	69	74,5
2012	70,7	75,7
2013	71,2	77,8

Source : INSEE

GRAPHIQUE 1.1 - Utilisation d'un ordinateur en France selon le sexe (2007-2013)



De tels tableaux (objet de ce chapitre) et graphiques (objet du chapitre 2) permettent de résumer et rendent plus lisible l'information contenue dans les données étudiées. Ils doivent être complétés par le calcul de divers indicateurs statistiques dits de position (ou tendance centrale), de forme et de dispersion (qui feront l'objet du chapitre 3). Dans cette optique, afin de mettre en évidence comment la statistique descriptive permet de rendre intelligible les informations collectées, nous allons tout d'abord exposer dans ce premier chapitre la terminologie et les notations usuelles qu'elle utilise. Nous expliquerons ensuite comment organiser et résumer les données d'un phénomène observé au travers de tableaux statistiques afin de pouvoir les exploiter pour en extraire un certain nombre d'informations synthétiques pertinentes permettant de nous faire une idée plus précise du phénomène considéré.

1). TERMINOLOGIE ET NOTATIONS DE BASE

De la même façon que les médecins utilisent certains termes particuliers pour désigner ce qu'ils observent chez leurs patients (adénite, colite, extrasystole, etc.), les statisticiens ont un langage spécifique qu'il convient de maîtriser. Comme toute science, la statistique a son vocabulaire qu'il est primordial de définir de façon

rigoureuse afin d'indiquer le groupe sur lequel porte l'étude, les caractères ou variables relevés sur chacun des individus et les différents types de caractères.

1.1). Population, individu, échantillon

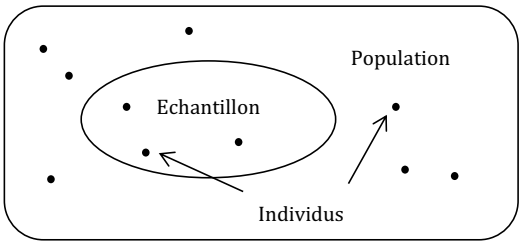
Lors de la construction du problème, il convient avant toute chose de préciser clairement sur quelle population porte le questionnement et quels individus composent cette population.

Une *population statistique*, notée *P*, est l'ensemble des éléments homogènes, c'est-à-dire disposant de certaines caractéristiques plus ou moins communes que l'on souhaite étudier (salariés d'une entreprise, étudiants d'une école de commerce, camemberts de Livry-sur Eure, etc.). Le nombre d'éléments de la population constitue l'*effectif total* (la *taille*) de la population usuellement noté *N*. Chacun des éléments de la population est qualifié d'*individu* ou d'*unité statistique* et est noté I. Il peut s'agir d'êtres humains (enfants, étudiants, salariés, etc.), d'objets (entreprises, logements, accidents, etc.) ou encore d'éléments non concrets (comme les intentions de vote), l'essentiel étant que l'on puisse dire clairement de tout élément qu'il appartient ou non à la population. Ces éléments sont indicés par i, i étant un entier variant de 1 à N. Alors, I_i constitue la i-ème unité statistique observée.

Il existe deux grandes façons de recueillir les données : de façon exhaustive (on parle alors de recensement) ou par échantillon (le recueil des données est alors partiel et aléatoire).

Hormis le cas très particulier des *recensements*, qui est une méthode très ancienne utilisée dès l'Antiquité par les Grecs pour mieux répartir la récolte de grains au sein de la communauté, et qui consiste en une étude sur population complète, une étude statistique ne peut que rarement être menée sur la totalité de la population, que ce soit pour des raisons de temps, de coût de collecte ou simplement de faisabilité (lorsqu'il faut, par exemple, user ou détruire des éléments d'une fabrication pour en mesurer la qualité). Bien que le recensement présente l'avantage de permettre une connaissance parfaite de la population, du fait des raisons évoquées plus avant, il est d'usage de se restreindre à l'étude d'un échantillon.

FIGURE 1.1 - Population, échantillon et individus



Un *échantillon* de taille n est ainsi un sous-ensemble d'individus réellement accessible de la population de taille N considérée qui doit posséder les mêmes caractéristiques que la population dont il est issu avec $n \leq N$. Par exemple, lorsqu'un magazine souhaite connaître les intentions de votes pour le prochain Président de la République, il interroge un échantillon de Français (environ 1000 individus), et non

toute la population résidant en France, soit près de 70 millions d'individus. A partir d'un échantillon dit *représentatif*, il est possible d'effectuer des analyses et d'en déduire des conclusions valables pour la population dans son ensemble[1]. Dit autrement, seul un échantillon reflétant fidèlement la complexité et la composition de la population est étudié et les résultats obtenus sont extrapolés de la population, le calcul d'une marge d'erreur étant possible[2].

1.2). Caractère, modalité et variable statistique

Chaque individu d'une population peut être décrit relativement à un ou plusieurs caractères ou variables statistiques.

a). Caractère et modalité

Afin d'étudier les individus composant une population selon certaines propriétés, on les classe en un certain nombre de sous-ensembles, appelés *caractères* ou *variables statistiques* notés X, Y, ou Z. Par exemple, si on étudie le personnel d'une entreprise, on pourra retenir comme caractères le sexe, la qualification, l'ancienneté, etc. Pour un smartphone, on retiendra le modèle, la marque, la puissance du microprocesseur, etc. Si on considère les notes d'un groupe d'étudiants à l'examen final, la population sera l'ensemble des étudiants constituant le groupe, un individu sera un des étudiants du groupe, et la variable sera la note obtenue à l'examen final. Pour chaque individu, une valeur de caractère sera observée et éventuellement mesurée. Si on étudie une variable X sur une population de N individus notés I_i, avec i variant de 1 à N, alors le résultat de l'observation de l'individu I_i sera noté x_i.

Les valeurs possibles prises par le caractère ou la variable sont dénommées modalités. La variable « sexe » a ainsi deux modalités (homme, femme) mais les caractères peuvent avoir un nombre élevé de modalités. Les modalités sont exclusives dans le sens où un individu ne peut appartenir simultanément à plusieurs modalités (un fromage ne peut pas à la fois être rond et carré) et exhaustives puisque toutes les situations doivent pouvoir être recensées[3]. Si le nombre total de modalités est noté k, l'ensemble des modalités de la variable X sera noté : $M = x_1 ; x_2 ; \ldots ; x_k$ avec $i = 1,\ldots, k$. Considérons les données ci-après concernant le nombre de femmes et d'hommes dans un festival de musique donné au printemps 2016.

---

[1] Souvent longue et fastidieuse, la collecte des données constitue le point de départ de toute étude statistique. Au-delà de l'importance de bien définir la population objet de l'étude, il convient de construire un échantillon fiable lorsqu'il n'est pas possible d'interroger l'ensemble de la population. Pour que l'échantillon soit représentatif de la population, le tirage doit être aléatoire. C'est une méthode peu coûteuse et qui, pour peu que le tirage soit aléatoire, empêche les manipulations. Il ne permet cependant pas de connaître parfaitement la population dans la mesure où les valeurs des caractéristiques de la population sont induites à partir de l'échantillon. Pour appréhender les principales méthodes de recueil de l'information et d'échantillonnage que nous ne traitons pas dans le présent ouvrage afin de nous concentrer sur le traitement et l'interprétation de l'information, nous renvoyons le lecteur à M.B. Miles et A.M. Huberman (2003), *Analyse de données qualitatives*, éd. De Boeck, coll. Méthodes en sciences humaines, et F. Lebaron (2006), *L'enquête quantitative en sciences sociales : recueil et analyse de données*, éd. Dunod, coll. Psycho sup.

[2] Voir le chapitre 8 sur les intervalles de confiance.

[3] Comme on peut être à la fois français et allemand, certaines conventions doivent être adoptées. De la même façon, pour respecter la condition d'exhaustivité lorsqu'il existe des ambiguïtés, on a quelquefois recours aux modalités « divers », « autres », « non déclaré ».

TABLEAU 1.2 - Population étudiante d'une école de commerce (rentrée 2016-2017)

Femmes	Hommes
5984	7671

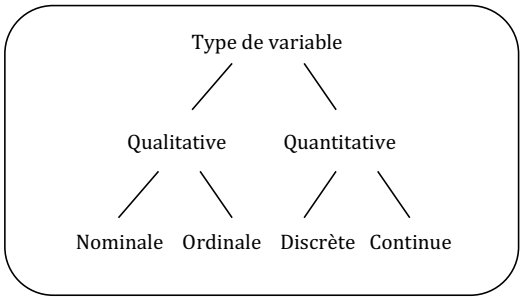
La population P étudiée est la population de N spectateurs au festival de musique et la variable X étudiée est le sexe. Cette variable a deux modalités : M = féminin ou masculin. Ces modalités sont en général numérotées : si la variable étudiée, ici le sexe, est notée X, les deux modalités seront respectivement notées x_1 (pour féminin) et x_2 (pour masculin).

b). Variables statistiques qualitatives et quantitatives

Une variable statistique peut être de nature *qualitative* ou *quantitative*[4]. Une variable statistique est dite *qualitative* (ou *catégorielle*) si ses modalités ne sont pas des valeurs chiffrées mais descriptives. Le sexe, la profession, la couleur des yeux, la ville de naissance, la nationalité sont quelques exemples de variables dont les modalités sont des catégories observables non mesurables par un nombre[5].

Une variable statistique est dite de nature *quantitative* si ses modalités sont mesurables. A chaque modalité est associée une valeur chiffrée (un nombre lié à l'unité choisie, qui doit toujours être précisée) représentant la mesure du caractère. Ainsi, l'âge, la taille, le chiffre d'affaires d'une entreprise, la puissance d'un ordinateur, le nombre de places assises d'un stade de football sont autant de variables statistiques dont les modalités sont des nombres.

FIGURE 1.2 - Nature de la variable statistique



c). Variables statistiques qualitatives nominales ou ordinales

Les modalités d'une variable qualitative peuvent être classées suivant deux types d'échelles de mesure : *nominale* (on nomme des catégories en utilisant des nombres nominaux) ou *ordinale* (on peut marquer le rang, c'est-à-dire l'ordre, et ce grâce à des

[4] Il convient de ne pas confondre les *variables statistiques*, objet d'étude des trois premiers chapitres, qui sont des entités pouvant prendre toutes les valeurs possibles au sein d'un ensemble de définition donné avec les *variables aléatoires*, introduites à partir du chapitre 4, dont les valeurs prises sont soumises au hasard (par exemple « pile » ou « face » dans le cas du lancer d'une pièce).

[5] Il convient de noter que qui compte ici est bien la nature de la modalité et pas le chiffre qui la représente. En effet, la variable « département de résidence » peut prendre plusieurs modalités chiffrées (par exemple 76, 95, 25, etc.) mais est malgré tout une variable qualitative.