CHAPITRE 1

Les notions fondamentales

Sommaire du chapitre

1.1 Introduction

Avant d'aborder le rôle des méthodes statistiques dans le processus de collecte des informations, de l'acquisition de nouvelles connaissances et de la prise de décision, notons d'abord que « les statistiques » au pluriel font référence aux *données statistiques*. Il s'agit donc de chiffres et de valeurs numériques – moyennes, pourcentages, fréquence d'occurrence d'un événement au cours d'une période donnée, etc. – relatifs à un ensemble d'unités ou d'individus (personnes, objets, entreprises, situations) qui partagent certaines caractéristiques et qu'on appelle une « population ».
Voici quelques exemples de données statistiques : le salaire moyen des employés d'une entreprise, les taux de chômage et d'emploi dans un pays, la production d'huile d'olive des grands pays producteurs en Europe durant les cinq dernières années, etc.

Nous associons souvent les méthodes statistiques (ou la statistique au singulier) au fait de produire des données statistiques. C'est exact, mais le domaine de la statistique englobe bien plus encore. En effet, que cela concerne l'économie ou notre quotidien, nos décisions, choix et comportements reposent quasiment toujours, directement ou indirectement, sur une forme d'activité statistique.

Par exemple, tout médicament prescrit par un médecin a été soumis à une évaluation complexe faite par des chercheurs, où la statistique occupe un rôle majeur. L'efficacité et l'innocuité du médicament sont démontrées à travers des essais cliniques durant lesquels les chercheurs s'appuient sur des méthodes statistiques pour planifier l'expérience, analyser les données et généraliser les résultats obtenus grâce aux données étudiées, ce qui permet de se prononcer sur l'efficacité et l'innocuité du médicament.

Prenons un autre exemple : le contrôle qualité effectué par un ingénieur dans la production industrielle. Lorsqu'il déclare aux responsables que la production est « sous contrôle », cela signifie que le pourcentage de composants non conformes présents dans l'échantillon prélevé sur la ligne de production ne dépasse pas la limite de tolérance. Ici, nous employons des méthodes statistiques pour fixer une limite de tolérance, sélectionner un échantillon et généraliser les résultats obtenus.

À la lumière de ces exemples, nous pouvons définir la *statistique* comme suit :

Définition 1.1

La **statistique** est une discipline scientifique qui fournit les principes et les méthodes en vue de collecter, d'organiser, de présenter, d'analyser et d'interpréter des données, voire, sous certaines conditions, de généraliser des conclusions à partir des données observées.

Une fois que vous serez familiarisé avec tout ce que cette discipline renferme réellement, la définition ci-dessus prendra alors tout son sens. L'expression « sous certaines conditions » fait écho à la nature des données observées : établir une inférence à partir d'un échantillon présuppose le recours à un mécanisme aléatoire visant à produire lesdites données. Nous approfondirons ce sujet au moment opportun.

1.2 La statistique dans l'histoire : les débuts de la discipline

La statistique est une discipline relativement récente : elle fut développée en grande partie aux XIX^e^ et XX^e^ siècles, et ses origines en tant que discipline indépendante datent du XVII^e^ siècle en Angleterre. John Graunt (1620-1674) et William Petty (1623-1687) y firent la promotion d'un axe de recherche : l'« **arithmétique politique** ». À cet effet, ils adoptèrent une approche empirico-inductive, alors commune dans le domaine des sciences naturelles, pour l'étude des phénomènes démographiques et sociaux[1]. Le mérite de ces auteurs repose plus sur le fait d'avoir montré à leurs contemporains l'importance de la collecte de données et de leur utilisation appropriée, que sur l'introduction de méthodes d'analyse particulières.

Ainsi, une nouvelle discipline vit le jour. Au départ, elle n'était cependant pas connue sous le nom de « statistique ». Il est communément admis que le terme « statistique » fut initialement employé pour désigner la recherche scientifique amorcée en Allemagne par l'intellectuel Hermann Conring (1606-1681), qui dispensa un cours en sciences politiques visant à fournir une « description de l'État ». Ce n'est que deux siècles plus tard que la dimension d'investigation soulevée par l'arithmétique politique commença

1 Un exemple qui illustre cette nouvelle approche dans l'étude des phénomènes démographiques est l'estimation du nombre de foyers à Londres en 1660 (rapportée dans l'ouvrage de Graunt intitulé *Observations naturelles et politiques sur les bulletins de mortalité…* publié en 1662, qui se fonde sur les archives des certificats de baptême, de mariage et d'enterrement des paroisses de l'Église d'Angleterre) : ayant observé que 3 décès étaient survenus dans 11 foyers au cours d'une année, Graunt en déduisit que Londres devait compter environ 84 000 foyers, en raison des 23 000 décès enregistrés cette année-là. L'étude des données selon un prisme scientifique permit à Graunt de découvrir les lois réelles sous-jacentes aux phénomènes démographiques et sociaux, tels que le surplus de naissances de garçons, l'urbanisation des populations rurales, etc.

à ressembler à la statistique moderne : il fallut en effet deux cents ans pour développer la branche des mathématiques qui se consacre à la logique de l'incertitude, à savoir la théorie des probabilités, et se pencher sur le problème de la « mesure de l'incertitude ».

Ici, mesurer l'incertitude signifie à quel point nous jugeons fiables les résultats d'une étude empirique : en analysant les registres paroissiaux (voir note 1), Graunt était en mesure d'établir que le taux de survie des hommes âgés de 50 à 70 ans était de 40 %, mais il était incapable d'évaluer la marge d'erreur ou la validité de cette estimation, nécessaire pour tirer des conclusions plus générales à partir des résultats.

La théorie des probabilités est l'outil visant à résoudre le **problème direct** : par exemple, elle permet de calculer la probabilité de prélever une balle d'une certaine couleur lorsqu'on connaît le contenu de l'urne. La statistique, en revanche, traite d'un phénomène appelé **problème inverse**. Pour ce faire, elle répond à des questions du type : « quel est le pourcentage de balles blanches dans l'urne, après que l'on a observé x balles blanches dans n tirages ? »

Le développement de la théorie des probabilités est attribué à de grands mathématiciens, parmi lesquels se sont particulièrement distingués : Blaise Pascal (1623-1662), Pierre Fermat (1601-1665), Jacob Bernoulli (1654-1705), Abraham de Moivre (1667-1754), Thomas Bayes (1702-1761), Pierre-Simon Laplace (1749-1827), Adrien-Marie Legendre (1752-1833) et Carl Friedrich Gauss (1777-1855). Les contributions scientifiques des quatre derniers savants ne se rapportent pas aux probabilités à proprement parler : leur intérêt s'est porté sur différents aspects du problème inverse, ce qui leur a permis d'obtenir par là même des résultats essentiels pour la mise au point de la statistique.

Un grand mérite revient à Bayes pour s'être consacré au problème suivant, malgré son échec (les calculs étant trop complexes pour l'époque). Soit θ la probabilité qu'un événement E se produise lors d'une expérience aléatoire : quelle est la probabilité que θ appartienne à l'intervalle $[a, b]$, sachant que l'événement E s'est produit x fois lors de n expériences aléatoires ? L'importance de cette question est évidente : être en mesure de dire, au vu des observations empiriques, que la quantité inconnue θ (par exemple, la probabilité mentionnée plus haut, qu'un individu de 50 ans vive jusqu'à ses 70 ans) a de fortes chances de se trouver dans tel ou tel intervalle est l'exemple d'un raisonnement parfaitement inductif.

L'étendue des recherches de Laplace était impressionnante. Parmi les sujets particulièrement pertinents pour la statistique, il convient de mentionner : le raisonnement inductif préalablement abordé par Bayes[2], le choix de la moyenne arithmétique pour résumer un ensemble de mesures répétées de la même quantité afin d'en faire la meilleure estimation, et le théorème central limite[3].

La mise au point de la méthode des moindres carrés fut une avancée majeure, d'une grande influence sur le développement des statistiques. Sous sa forme la plus simple, cette méthode permet de résoudre le problème suivant : étant donné n mesures de la même quantité, quel résumé statistique se rapproche le plus de la quantité inconnue ? Selon la méthode des moindres carrés, la moyenne arithmétique fournit le meilleur résumé de n

2 Appliquant sa méthode, Laplace a étudié le phénomène du surplus de naissances masculines (déjà observé par Graunt, voir note 1). Notant θ la probabilité d'une naissance masculine, il a trouvé que la probabilité que $\theta \leq 0,5$ était de $1,1521 \times 10^{-42}$, en s'appuyant sur l'observation empirique que 251 527 garçons et 241 945 filles sont nés entre 1745 et 1770 à Paris. Ainsi, il a conclu que $\theta > 0,5$.

3 Selon ce théorème, dans certaines conditions, la distribution de probabilité de la moyenne d'un grand nombre de quantités homogènes (par exemple, le poids moyen de composants venant d'un processus de production donné) peut être approximée par une courbe normale.

mesures. La mise au point de cette méthode est attribuée à la fois à Legendre et à Gauss[4]. Au moment de son introduction, l'émergence de plusieurs problèmes scientifiques en astronomie, tels que la détermination et la modélisation mathématiques de l'orbite de la Lune, témoignait de la pertinence d'une telle méthodologie.

Bien que les travaux de Bernoulli et Laplace aient pu présager l'application des statistiques aux phénomènes sociaux, une étape cruciale fut franchie dans ce sens grâce aux contributions d'Adolphe Quételet (1796-1874) et de Francis Galton (1822-1911) notamment.

Quételet fut un homme éclectique : mathématicien de formation, ses connaissances profitèrent à l'astronomie, la physique, la météorologie et la sociologie. Gestionnaire doué, il mit en place le système officiel de la statistique de Belgique et fonda plusieurs associations scientifiques, à la fois dans son pays et à l'étranger. Sa contribution la plus décisive à l'analyse des données sociales fut l'élaboration du concept de l'**homme moyen**.

Lors de son analyse des données de la population, Quételet examina tout un éventail de liens possibles entre des phénomènes en élaborant des tableaux et des graphiques. Il se pencha sur les taux de natalité et de mortalité, prenant en compte le mois, la ville, la température et l'heure. Il étudia les caractéristiques anthropométriques (le poids, la taille, le tour de poitrine, etc.) et psychologiques des individus, grâce aux statistiques notamment sur l'alcoolisme, le suicide ou les maladies mentales. À l'instar des travaux des astronomes sur les lois de l'univers un siècle auparavant, ses recherches visaient à découvrir des lois régissant la société humaine.

L'idée de l'homme moyen est probablement née pour répondre au besoin de résumer des données anthropométriques et permettre ainsi de comparer différents groupes de personnes. Par exemple, grâce à l'accès aux données sur la taille et le poids d'un grand nombre de conscrits français, la taille et le poids moyens servaient à définir « le conscrit français moyen ». La même analyse a été effectuée pour un nombre considérable de conscrits belges, de sorte à pouvoir ensuite comparer les deux. Pouvant s'appliquer à n'importe quelle caractéristique mesurable, cette approche permit de mettre en parallèle différentes catégories de personnes à travers l'espace et le temps : l'homme moyen était un concept pratique pour neutraliser les variations aléatoires des individus et pour révéler des régularités, c'est-à-dire des lois qui régissent la société.

Une deuxième branche de recherche se servit de la courbe normale pour analyser les données anthropométriques et en faire un outil pour évaluer l'homogénéité de la population étudiée – une condition nécessaire pour comparer des moyennes. Le raisonnement est le suivant : si l'ensemble des mesures d'une variable donnée est homogène (c'est-à-dire que les mesures sont influencées par des facteurs communs dominants, les différences ne résultant que de causes accidentelles), les données suivent forcément la courbe des erreurs accidentelles. On peut noter que l'importance des travaux de Quételet repose moins sur les innovations méthodologiques que sur l'essor qu'ont connu les modèles probabilistes dans l'étude des phénomènes sociaux.

Pourtant, c'est le nom de Galton qui sera à jamais associé au véritable tournant décisif pour la mise au point d'une méthodologie empirique et conceptuelle dans l'étude des phénomènes sociaux. Galton était étudiant en médecine à Cambridge. Or, un héritage considérable lui permit de renoncer à cette carrière au profit de ses innombrables centres d'intérêt : il explora l'Afrique pendant deux ans, pour se consacrer ensuite à la météorologie, la psychologie, l'anthropologie et la sociologie. Parmi les nombreux

4 Gauss a affirmé dans ses écrits avoir utilisé la méthode avant 1805, année où Legendre a publié une description de la méthode. Quoi qu'il en soit, Gauss a abordé le problème selon des termes probabilistes qui l'ont emmené vers la redécouverte de la courbe normale en tant que distribution de la probabilité des erreurs accidentelles.

accomplissements de Quételet, Galton se servit notamment de la représentation des données des cas homogènes (tels que les individus appartenant à une même ethnie) en utilisant la distribution normale. Il poussa toutefois ses analyses bien plus loin, déterminé à expliquer les mécanismes qui sous-tendent souvent la représentation graphique en forme de cloche des données. Il se demanda notamment comment l'hérédité des traits d'une génération à l'autre pouvait être compatible avec cette spécificité des données. Et ce fut précisément dans le cadre de l'analyse de l'hérédité des traits que Galton apporta sa plus grande contribution, à savoir le concept et la méthode de la régression. En analysant un ensemble de données très fourni sur la taille des pères et des fils, il observa que les pères plus grands que la moyenne avaient des fils plus grands que la moyenne et que les pères plus petits que la moyenne avaient des fils plus petits que la moyenne. Il observa en outre que la taille des fils se rapprochait plus de la moyenne que celle des pères. Ce phénomène porta dès lors le nom de « régression ». Il désigne une évolution de retour à la moyenne.

Après Galton, nous abordons les années 1900. C'est durant les quatre premières décennies du XX^{e} siècle que l'on a plus ou moins défini la statistique telle que nous la connaissons aujourd'hui. En témoigne l'émergence de deux figures de proue : Karl Pearson (1857-1936) et Ronald Aylmer Fisher (1890-1962).

Bien qu'anglais de naissance, Pearson était un fervent admirateur de la culture allemande et de Marx. Il alla même jusqu'à changer de prénom en l'honneur de ce dernier : de Charles, il devint Karl. Il introduisit l'utilisation des probabilités en vue de tirer des conclusions générales en partant de résultats empiriques. Par ailleurs, il proposa une méthode très importante (connue sous le nom de test du Khi-deux) qui permettait de tester si les données observées étaient en adéquation avec un modèle théorique. Il définit un large éventail de courbes asymétriques, utiles dans la description des données lorsque les conditions pour employer la courbe normale ne sont pas réunies. Pearson fonda non seulement des journaux scientifiques ayant pour objet la statistique, mais également une école de pensée à l'influence considérable.

Le plus éminent parmi les statisticiens du siècle dernier était peut-être Fisher : aujourd'hui, la branche de la statistique dédiée à l'analyse des données empiriques (données provenant d'expériences supervisées réalisées dans les essais cliniques, la biométrie, la psychologie, etc.) est largement imprégnée de ses contributions. Il joua également un rôle prépondérant dans la définition et l'élaboration des modèles expérimentaux.

Il ne s'agit là que d'un bref aperçu des figures les plus emblématiques du développement de la méthode statistique. Pour aller plus loin, voir Stigler (1986).

1.3 Les statistiques et le processus de découverte de nouvelles connaissances

Comme nous l'avons vu ci-dessus, la statistique avait d'abord pour objectif de déterminer si les résultats empiriques et les données observées étaient en adéquation avec une hypothèse ou une théorie scientifique dans une situation donnée. Astronomes, Laplace, Gauss et d'autres scientifiques procédaient de cette façon dans leur domaine. Ce même paradigme a ensuite été appliqué dans les sciences sociales. Souvenons-nous dans ce contexte de l'exemple du surplus de naissances de garçons étudié par Laplace : les données empiriques faisaient apparaître un surplus de naissances de garçons (le phénomène a été découvert par Graunt en 1662). Dans l'analyse de Laplace, cette observation fut considérée comme une hypothèse à vérifier. Pour ce faire, l'hypothèse fut comparée aux données réelles du registre des naissances à Paris pour une période donnée et n'a pas été rejetée.

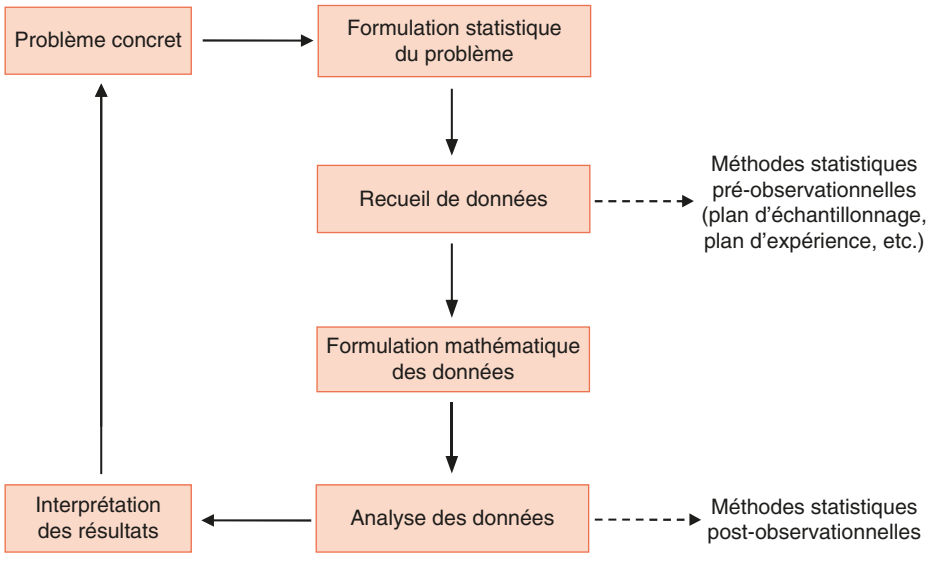
Ainsi, la recherche scientifique au sein des sciences empiriques, regroupant à la fois les sciences naturelles (astronomie, physique, biologie, etc.) et les sciences humaines et sociales (économie, psychologie, sociologie, etc.), se déroule idéalement comme suit :

a. Formuler le problème, possiblement comme une hypothèse.
b. Identifier des données pertinentes.
c. Concevoir la méthode de recueil des données.
d. Recueillir les données.
e. Utiliser des méthodes d'inférence statistique pour tirer des conclusions générales à partir des données observées, en procédant à des tests d'hypothèses ou à l'estimation de quantités inconnues inhérentes au phénomène étudié.
f. Interpréter les résultats, c'est-à-dire utiliser les preuves empiriques obtenues par les données de façon à pouvoir répondre aux questions qui sont à l'origine de l'étude.

La statistique joue un rôle majeur dans les étapes (c), (d) et (e). Lors de l'étape (c), elle permet de planifier judicieusement l'enquête statistique, l'expérience ou l'étude observationnelle (voir section 1.8). Au cours de l'étape (e), on utilise des méthodes d'inférence statistique pour étendre les conclusions obtenues par les données observées avec des tests d'hypothèses ou des estimations de quantités inconnues inhérentes au phénomène étudié.

Lors de l'étape (b), la statistique fournit uniquement les critères et la démarche du recueil des données : sélectionner des données pertinentes pour l'étude dépend du domaine sur lequel portera cette dernière (astronomie, économie…). L'interprétation des résultats se fera également en fonction du domaine en question.

Figure 1.1 Statistique et processus de découverte de nouvelles connaissances



La figure 1.1 permet de visualiser comment le fait de poser la problématique aboutit à la découverte de nouvelles connaissances en passant par des méthodes statistiques qui précèdent l'observation (type d'échantillonnage, conception des expériences, enquêtes de terrain), le recueil de données, et par les méthodes statistiques post-observationnelles.

1.4 La statistique dans les activités opérationnelles et au quotidien

Le terme « activités opérationnelles » englobe toutes les activités qui visent à atteindre un but. Lorsque la réalisation d'une activité opérationnelle s'accompagne d'un certain niveau de complexité, le choix se fait idéalement en s'appuyant sur des étapes déterminées. Prenons l'exemple d'un magasin qui cherche à réaliser une étude de marché afin d'analyser chacune d'entre elles :

a. Identifier et décrire le but de l'activité.
b. S'appuyer sur les statistiques – alors nécessaires – pour récupérer et analyser des données sur les magasins concurrents dans la zone sélectionnée, pour collecter et analyser des informations concernant les clients potentiels, dans le but de répondre notamment à leurs besoins et à leurs attentes.
c. Étudier les moyens et les outils qui seront employés.
d. Prévoir les résultats réalistes.
e. Opérer un choix final.

L'utilité des statistiques est évidente lors des étapes (b) et (d), par exemple lorsque les dirigeants d'une enseigne de la grande distribution doivent décider d'ouvrir ou non un nouveau magasin.

Pendant la phase (b), les statistiques permettent de récupérer et d'analyser des données sur les magasins concurrents sur le territoire en question, en vue de collecter et d'analyser des informations concernant les clients potentiels, dans le but de répondre entre autres à leurs besoins et à leurs attentes.

Lors de l'étape (d), les statistiques constituent un atout supplémentaire lors d'une analyse du marché pour prédire le comportement des clients. Prévoir une enquête de terrain pour sonder des clients potentiels à cette fin pourrait alors s'avérer utile.

Prenons pour autre exemple une aide publique fournie dans le but de stimuler le marché des véhicules électriques. Une démarche rationnelle et rigoureuse nécessite l'utilisation des statistiques au cours des deux étapes mentionnées plus haut. Durant l'étape (b), elles servent à analyser les caractéristiques de l'ensemble des véhicules potentiellement concernés par cette aide. Inutile de faire une enquête pour récupérer ces données, car on peut utiliser celles du service de l'immatriculation des véhicules. Lors de l'étape (d), le recours aux statistiques permet d'évaluer le coût de l'opération en s'appuyant sur des modèles visant à prévoir le nombre et les spécificités des personnes qui bénéficieront de cette aide.

En guise de dernier exemple, imaginons un opérateur de Bourse chargé de choisir une transaction financière. Dans cette optique, il va s'appuyer lui aussi sur des statistiques : d'abord pour évaluer l'état du marché boursier, ensuite pour prévoir l'évolution des titres financiers en question.

Le rôle des statistiques est donc de fournir un soutien à toute activité commerciale ou entrepreneuriale importante ou complexe, de sorte à prendre une décision éclairée fondée sur l'observation et l'anticipation des faits.

Lorsque l'on évoque le rôle de la statistique dans la recherche scientifique ou dans les activités économiques, nous pensons à son utilisation dans le cadre professionnel. Pour autant, il serait souhaitable que chacun dispose de connaissances en statistiques afin de comprendre et d'interpréter les informations quantitatives dont les médias regorgent aujourd'hui : graphiques, tableaux, résumés statistiques sur l'inflation, l'emploi, les accidents de la route, etc. De plus, tout le monde devrait être en mesure d'exprimer et de transmettre des informations quantitatives en se servant des outils statistiques de base.

Véritable langage doté de sa propre syntaxe, la statistique démultiplie notre capacité à communiquer et à interpréter, tout comme la littératie (aptitude à lire et à écrire) – que la plupart des personnes maîtrisent aujourd'hui. Une bonne compréhension des statistiques est désormais une compétence clé pour le citoyen moderne, tout comme savoir lire et écrire.

1.5 Les sources des données statistiques nationales et internationales

Chaque pays dispose évidemment de son propre organisme officiel, un institut national de la statistique, chargé de collecter, traiter et diffuser les chiffres officiels du pays concernant la population, l'économie, l'emploi, etc. On peut citer par exemple le Census Bureau aux États-Unis, Statistics Canada au Canada, l'Insee en France, l'Istat en Italie. Ces instituts recueillent eux-mêmes des données statistiques *via* des sondages ou des recensements et reçoivent également des données de la part des bureaux administratifs centraux ou locaux.

Au niveau de l'Union européenne, les instituts nationaux de statistique des pays membres sont coordonnés par Eurostat, une organisation supranationale. Celle-ci collecte et traite leurs données en promouvant l'harmonisation des instruments de mesure employés par les instituts nationaux de statistique, afin de produire des données statistiques d'une qualité comparable entre les pays et les régions de l'Union européenne. L'une des activités principales d'Eurostat réside dans la définition des données macroéconomiques visant à valider les décisions sur lesquelles reposent les politiques monétaires de la Banque centrale européenne. Eurostat travaille conjointement avec d'autres organisations internationales, telles que les Nations unies et l'OCDE pour définir des standards statistiques à l'échelle internationale. Elle joue également un rôle majeur dans l'amélioration des capacités statistiques des pays en développement.

De manière générale, tous les pays disposent d'un grand nombre d'organismes et d'institutions qui produisent, traitent et diffusent des données statistiques dans leurs domaines d'activité respectifs. Les banques centrales, les chambres de commerce, les organisations professionnelles, etc. sont toutes des sources de données statistiques. Quant aux données internationales, elles parviennent des grandes organisations supranationales citées ci-dessous :

- **La Banque centrale européenne, la BCE** (https://www.ecb.europa.eu/). Son rôle est de mettre en place des politiques monétaires pour les pays de l'Union européenne ayant adopté la monnaie unique et faisant partie de la zone euro.
- **La Commission de statistique de l'ONU** (https://unstats.un.org/). Commission du Département des affaires économiques et sociales des Nations unies, elle génère les différentes statistiques démographiques, économiques et commerciales dont les agences des Nations unies ont besoin.
- **Le Fonds monétaire international, le FMI** (https://www.imf.org/). Il s'agit d'une institution internationale créée par les gouvernements de 189 pays. Son but est

de promouvoir la coopération monétaire internationale et la stabilité des taux de change, de faciliter l'expansion et une croissance harmonieuse du commerce mondial, de fournir de l'aide aux pays membres de sorte à corriger des déséquilibres temporaires dans leur balance commerciale.

- **La Banque mondiale** (https://www.banquemondiale.org/). Cette institution internationale comprend la Banque internationale pour la reconstruction et le développement et l'Association internationale de développement. Leurs engagements incluent la lutte contre la pauvreté et l'organisation de l'aide et des fonds pour des pays en difficulté.
- **L'Organisation de coopération et de développement économiques, l'OCDE** (https://www.oecd.org/). Il s'agit d'une organisation internationale d'études économiques pour les pays membres – des pays développés ayant en commun un système démocratique de gouvernance et une économie de marché. Plus précisément, les objectifs de l'OCDE sont la promotion des politiques permettant d'atteindre des niveaux plus élevés en matière de croissance économique durable et d'emploi au sein des pays membres, d'encourager les investissements et la compétitivité, de maintenir la stabilité financière, de contribuer au développement des pays non membres et de favoriser l'expansion du commerce mondial sur une base non discriminatoire en suivant les obligations internationales.
- **L'Organisation mondiale de la santé, l'OMS** (https://www.who.int/). C'est l'agence spécialisée de l'Organisation des Nations unies pour la santé publique.
- **L'Organisation des Nations unies pour l'alimentation et l'agriculture, la FAO** (https://www.fao.org/). Il s'agit de l'agence spécialisée de l'ONU. Elle vise à atteindre un meilleur niveau de nutrition, accroître la productivité agricole, améliorer la vie des populations rurales et contribuer à une croissance économique globale.

1.6 Les termes de base en statistique

Afin d'être exploités comme des données statistiques, les nombres doivent provenir de l'observation d'une multitude de cas individuels (personnes, objets, lieux, etc.) à travers lesquels il est possible d'observer le phénomène à l'étude. En ce sens, la température mesurée dans une station météorologique à 13 heures, le 21 juin, ne constitue pas une donnée statistique. Elle le devient lorsqu'elle fait partie d'un ensemble de températures mesurées à la même heure dans les stations météorologiques d'une région donnée, d'un pays donné, parce que, dans ce cas, les données nous permettent de faire des évaluations et des comparaisons potentiellement pertinentes. Prenons un autre exemple : le taux de cholestérol d'une personne, selon l'analyse clinique, ne constitue pas une donnée statistique. En revanche, il le devient lorsque ce taux fait partie d'un ensemble d'observations faites sur plusieurs individus – avec une santé comparable – qui suivent la même thérapie. L'effet de la thérapie sur le cholestérol pourrait constituer un sujet intéressant pour une étude.

Définition 1.2

Une **population** est l'ensemble de tous les individus, au sens statistique, que nous souhaitons étudier. Il peut s'agir d'un ensemble d'objets, de transactions, d'événements, de personnes, etc.

Définition 1.3

Un **échantillon** est un sous-ensemble d'éléments d'une population, obtenu par un processus de sélection dans le but d'enquêter sur les caractéristiques de la population.

Voici quelques exemples :

a. Une enquête portant sur toutes les entreprises de l'industrie mécanique dans une zone géographique donnée en vue d'étudier leurs caractéristiques et, plus précisément, de déterminer le nombre total de salariés.
b. Une enquête visant à connaître l'opinion des Italiens sur l'euro, menée à partir d'interviews sur un échantillon de 1 500 personnes.
c. L'évaluation du niveau de défectuosité de composants manufacturés par une société industrielle, grâce à l'observation de 500 composants prélevés sur la ligne de production à intervalles réguliers au cours d'une journée de travail.

Dans l'exemple (a), la population est composée de l'ensemble de toutes les entreprises de l'industrie mécanique présentes sur ce territoire. L'intégralité des Italiens adultes compose la population dans l'exemple (b). Il est difficile de déterminer la population dans l'exemple (c). Les 500 composants étudiés constituent sans doute un échantillon. La population faisant défaut dans cet exemple, nous pouvons imaginer ceci : un ensemble virtuel (illimité en théorie) de toutes les observations possibles, en partant du principe que cette ligne de production continue de fonctionner dans les mêmes conditions, sans interruption. Ce sujet sera abordé plus amplement dans le chapitre 17.

Définition 1.4

Les **unités statistiques** sont les membres individuels d'une population. Elles sont les éléments pour lesquels les données sont collectées.

Définition 1.5

Une **variable** est une caractéristique ou particularité des unités statistiques que l'on souhaite analyser. Elle peut « varier », car sa valeur peut changer d'une unité à l'autre.

Dans l'exemple (a), une des variables examinées est le nombre d'employés dans une entreprise. Dans l'exemple (b), l'opinion du citoyen italien adulte concernant l'euro constitue la variable. La variable de l'exemple (c) est la caractéristique que l'on étudie afin d'évaluer si un composant est défectueux : il peut s'agir du poids, de la longueur, etc. En pratique, toutes les études concernent toujours une ou plusieurs variables qu'il est possible d'étudier individuellement ou conjointement.

On distingue deux types de variables : quantitatives et qualitatives. Une **variable qualitative** s'exprime en différentes catégories (attributs, noms, étiquettes) qui doivent

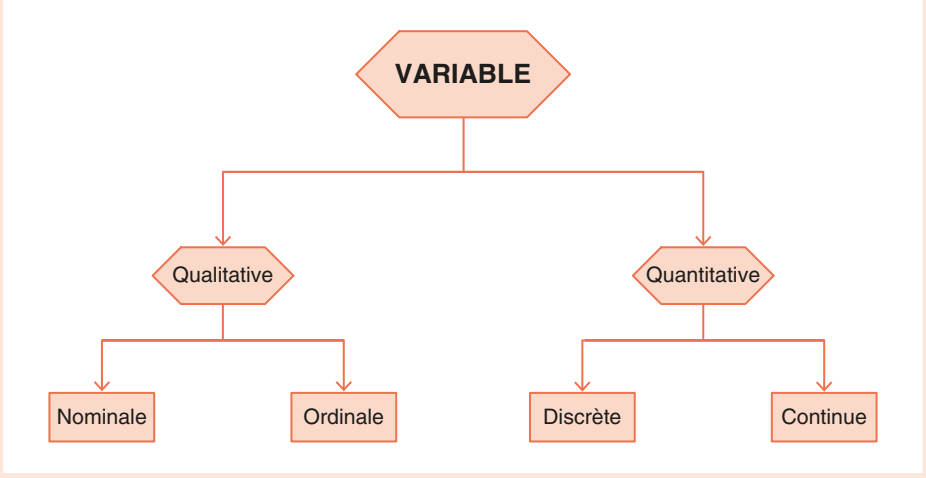
s'exclure mutuellement (aucune unité statistique ne peut faire partie de plus d'une catégorie) et qui doivent être exhaustives (il existe une catégorie pour chaque unité statistique). Ces catégories peuvent être classées ou non selon un ordre particulier (par exemple, par ordre croissant). Dans le premier cas, nous parlons d'une **variable ordinale** ; dans le second, il s'agit d'une **variable nominale**. Les grades militaires (caporal, sergent, lieutenant, commandant, colonel, etc.) sont un exemple de variable ordinale. La préférence religieuse, le sexe, le genre, l'affiliation à un parti politique sont des variables nominales.

Une **variable quantitative** peut être discrète ou continue. On parle de **variable discrète** lorsque les valeurs possibles sont dénombrables. Une **variable continue**, à l'inverse, prend un nombre infini de valeurs réelles possibles dans un intervalle.

La taille des ménages, qui peut être de 1, 2, 3…, est un exemple de variable discrète. Il en est de même pour le nombre d'étudiants dans l'amphithéâtre lors d'un cours à l'université. Si 152 étudiants sont inscrits dans le cursus, le nombre d'étudiants présents pendant un cours peut être de 131, ou 140, ou 96, ou tout autre nombre entier entre 1 et 152.

La taille, l'âge et la température sont des exemples de variables continues. En raison de la précision limitée des instruments de mesure, les valeurs sont forcément discrètes, même pour les variables continues. Par exemple, on peut mesurer l'âge d'une personne à une certaine fraction de seconde près, sans toutefois être infiniment précis.

Figure 1.2 Schéma de classification des variables



1.7 Les échelles de mesure et les types de variables

Mesurer est le processus par lequel sont attribués des chiffres ou des étiquettes à des variables. Les valeurs réelles que prennent les variables dans un ensemble d'unités statistiques (une population ou un échantillon) constituent les **observations**, c'est-à-dire les données brutes qu'il faut analyser avec des méthodes statistiques. Nous allons voir maintenant ce que « mesurer » signifie pour chacun des quatre types de variables mentionnés précédemment.

1.7.1 Les variables nominales

Mesurer une variable nominale revient à grouper les unités statistiques en deux catégories ou plus que l'on ne peut ordonner. Par exemple, le sexe est une variable nominale qui comprend deux catégories (femme et homme) sans ordre intrinsèque. Ici, observer cette variable sur une unité statistique nous permet uniquement de placer ladite unité dans une des deux catégories : femme ou homme.

1.7.2 Les variables ordinales

Mesurer une variable ordinale revient à grouper les unités statistiques en deux catégories ou plus qui sont classées en fonction de leur valeur. On peut attribuer des chiffres ou d'autres symboles (par exemple, des lettres de l'alphabet) à ces catégories. Chaque catégorie peut être considérée comme étant supérieure à (>) ou inférieure à (<) la catégorie voisine. Dans l'exemple des rangs militaires, on ne se limite pas à mettre les unités statistiques dans des groupes bien distincts, à savoir caporal, sergent, lieutenant, commandant, colonel, etc. On peut également établir un ordre, à savoir que le grade de caporal est inférieur à celui de sergent, lui-même inférieur à celui de lieutenant, etc.

1.7.3 Les variables discrètes

La mesure des variables discrètes est le résultat du dénombrement des unités statistiques à l'étude (population ou échantillon). Par exemple, dans une enquête sur les tailles des ménages, la variable « taille des ménages » correspond au nombre de personnes par ménage. On peut ainsi attribuer un nombre à chaque unité statistique. Il est ensuite possible de recourir à des méthodes arithmétiques (addition, soustraction, multiplication ou division) et à des méthodes logiques (« égal à » ou « supérieur à ») dans le traitement des données numériques ainsi obtenues.

1.7.4 Les variables continues

À l'inverse des variables discrètes, la mesure ne se fait pas par un dénombrement, mais par l'observation de ce qu'indique un instrument de mesure prévu à cet effet. Mesurer la variable pour l'intégralité des unités à l'étude produit des données numériques sur lesquelles on peut faire toutes les opérations arithmétiques et logiques précédemment mentionnées pour les variables discrètes[5]. En pratique, lorsque nous mesurons une variable continue, le résultat est toujours approximatif par rapport à la véritable valeur. Le niveau d'approximation dépend de plusieurs facteurs : précision de l'instrument utilisé, minutie de la personne qui procède au relevé, fait d'arrondir les mesures relevées à une décimale raisonnable, etc.

5 Clairement, il est possible d'établir une hiérarchie dans les échelles de mesure, grâce aux opérations logiques et arithmétiques que nous pouvons réaliser. Se trouve au dernier rang l'échelle nominale, avec laquelle les unités statistiques peuvent uniquement être classées en deux catégories. Vient ensuite en deuxième position l'échelle ordinale, selon laquelle nous pouvons regrouper les unités statistiques en catégories puis les classer dans un ordre. En haut du classement se trouve l'échelle de mesure utilisée pour des variables discrètes et pour quasiment toutes les variables continues. Ici, nous pouvons procéder aux quatre opérations arithmétiques (–, +, ×, /) et aux opérations logiques (« égal à » ou « supérieur à ») sur les données. De nombreux manuels de statistique mentionnent que ces variables sont mesurées à l'**échelle de rapport** afin de les distinguer de ces variables quantitatives pour lesquelles la mesure 0 n'indique pas l'absence de la quantité que nous sommes en train de mesurer. Pour ces variables, le rapport entre deux mesures n'est pas significatif à des fins de comparaison. Par exemple, si la température (en degrés Celsius) est de 20 à Rome et de 10 à Paris, nous ne pouvons pas dire qu'à Rome il fait deux fois plus chaud qu'à Paris. En effet, le 0 sur l'échelle Celsius ne signifie pas « absence » de température ; il s'agit simplement d'une valeur de cette échelle (le point de congélation de l'eau).

Pour commencer, il faut déterminer le degré de précision que l'on souhaite atteindre. Nous pouvons mesurer la taille d'une personne en centimètres, avec ou sans décimales. Avec une décimale, la mesure se fait au millimètre près. Avec deux décimales, elle est de l'ordre d'un dixième de millimètre. Le degré de précision dépend également de la grandeur de la variable étudiée. Ainsi, pour la taille d'une personne, une approximation au centimètre près suffit généralement. En revanche, il en est autrement pour la longueur des pétales d'un certain type de fleur, où la précision recherchée est de l'ordre du millimètre. Il est également évident que le choix de l'instrument le plus adapté dépend du degré de précision que l'on vise (un mètre ruban pour la taille d'une personne, un pied à coulisse pour les pétales de fleur). Pour toute mesure approximative, le dernier chiffre (souvent appelé le « chiffre incertain ») se trouve typiquement à la limite de la sensibilité de l'instrument de mesure. Pour les tailles par exemple, relever 174 cm signifie que l'on a observé une valeur qui se situe entre 173,5 et 174,5 cm. Par conséquent, le fait d'avoir relevé 174 cm signifie qu'il existe une erreur par rapport à la taille réelle. L'erreur maximum est alors d'un demi-centimètre, et cela se produit quand la mesure réelle est de 173,5 ou 174,5 cm. Si le degré de précision est d'un millimètre, une valeur de 174,6 cm indique que le relevé du mètre ruban a donné un nombre entre 174,55 et 174,65 cm. Dans ce cas, l'arrondi comprend une erreur d'un demi-millimètre maximum.

Les réflexions ci-dessus s'appliquent à chaque fois que nous mesurons une variable continue : la mesure prise se réfère toujours à une étendue de valeurs possibles, appelée **étendue implicite**, et comprend un écart de la mesure réelle ne dépassant pas la moitié de l'unité exprimant le degré de précision sélectionné. Il existe pourtant une exception notable à cette règle : prenons par exemple l'âge des personnes exprimé en nombre entier d'années vécues. Quand une personne a 52 ans, cela signifie que l'âge observé se situe dans l'intervalle fermé à gauche [52, 53) et non pas dans l'intervalle (51,5, 52,5), comme indiqué précédemment. La raison est que la mesure « nombre entier d'années vécues » contient uniquement une approximation d'arrondi vers le bas : en effet, nous considérons le nombre qui exprime l'âge réel du sujet dans son entièreté (en années et en fractions). Par conséquent, les âges réels de 52 ans et 2 mois et 52 ans et 11 mois correspondent tous les deux au même nombre entier d'années vécues, à savoir 52. Donc, l'erreur maximum que l'on peut commettre est de presque un an. Nous devrions évidemment étendre cette observation à toutes les situations où le temps constitue la variable et où les durées exprimées en nombre de périodes temporelles (années, mois…) entières constituent les valeurs.

1.8 La provenance des données statistiques

Les données statistiques sont le fruit d'une collecte ciblée d'informations sur un sujet ou un phénomène digne d'intérêt. Une grande partie des données que traitent les bureaux de statistiques, ou d'autres organismes à qui incombent la production et la diffusion des données, est en lien avec les activités administratives ou institutionnelles des organisations qui opèrent dans des domaines divers et variés. Le registre d'état civil par exemple, dont les archives sont utilisées à des fins civiles et administratives, produit des données statistiques qui sont la source des statistiques démographiques concernant les résidents d'un pays. Les chambres de commerce, qui tiennent des registres des entreprises dans le secteur sous leur juridiction, fournissent des services pour les opérateurs économiques, mais elles produisent également des données statistiques d'un intérêt plus général. Il en est de même, par exemple, pour les banques centrales – qui produisent des données sur le secteur du crédit du fait de la nature de leur activité – ou pour les services d'immatriculation – source de données concernant les véhicules.

Il est possible de classer les sources de données en deux catégories, **primaire** et **secondaire**, selon qu'elles ont été collectées par l'usager lui-même ou non. Par exemple, les données sur la création ou la faillite d'une affaire commerciale sont une source primaire pour les chambres de commerce et une source secondaire pour les chercheurs en économie qui s'en servent pour leur recherche. Dans ce chapitre, nous découvrons la statistique du point de vue de ceux qui souhaitent réaliser une étude ou une enquête, par une brève description des différentes méthodes par lesquelles il est possible d'obtenir des données statistiques : l'**enquête statistique**, l'**expérience** et l'**étude observationnelle** ou **de terrain**.

1.8.1 L'enquête statistique

Le terme **enquête statistique** décrit l'étude d'un ensemble d'unités identifiables et observables (personnes, entreprises, maisons, etc.), que nous appelons une population finie. L'enquête sur le marché du travail en est un exemple. Son but est de faire une estimation sur la population active occupée et le nombre de personnes en recherche d'emploi, de sorte à recueillir des informations sur l'offre de travail au niveau agrégé, comme le type de poste, le secteur d'activité et ainsi de suite. Un autre exemple est l'enquête sur les ménages, dont l'objectif est de collecter des données sociodémographiques sur les conditions de vie des personnes.

Lorsque l'enquête est réalisée à partir d'un sous-ensemble de la population, nous parlons d'**enquête-échantillon**. En revanche, lorsque la population entière est sondée, nous parlons de **recensement**, lequel constitue un cas particulier d'enquête. L'enquête-échantillon et le recensement nécessitent l'accès à une **liste** des unités statistiques composant la population, y compris des informations pour l'identification et la localisation de ces unités. Parfois, la police ou des institutions privées, telles que le service d'immatriculation des véhicules pour les véhicules en circulation, peuvent mettre ce genre de répertoire à disposition. Dans d'autres cas, il est nécessaire de dresser une nouvelle liste, avec toutes les difficultés et tous les problèmes quant à la qualité des résultats : cet inventaire pourrait se révéler incomplet (absence d'une partie des unités de la population) et certaines unités pourraient y figurer en double (unités de population enregistrées plusieurs fois).

Les instituts nationaux de la statistique, ou toute autre institution publique ou privée, effectuent presque exclusivement des enquêtes-échantillons, bien moins coûteuses et plus rapides comparées au recensement. Rappelons que toute analyse statistique effectuée avec des données d'échantillon (par exemple, pour le calcul de la moyenne) constitue une estimation du résultat que produirait un recensement réel. Grâce aux statistiques, nous pouvons toutefois mesurer, dans un certain sens, le degré d'approximation d'une telle estimation, pourvu que l'échantillon soit aléatoire, donc construit selon un mécanisme aléatoire.

Quelques techniques d'échantillonnage

Il existe plusieurs moyens de créer un échantillon aléatoire. Voici une liste des méthodes les plus répandues :

- **L'échantillon aléatoire simple**. Les unités échantillon sont sélectionnées selon un mécanisme aléatoire qui garantit que la probabilité de faire partie de l'échantillon est la même pour toutes les unités de la population. Si N désigne le nombre d'unités de population et n le nombre d'unités échantillon, la sélection d'un échantillon aléatoire simple est comparable au tirage de n balles d'une urne contenant N balles portant les chiffres 1 à N qui permettent d'identifier les unités échantillon (la sélection doit se faire sans remise, c'est-à-dire sans qu'aucune balle choisie ne soit remise dans l'urne avant le tirage suivant).

- **L'échantillon systématique**. Nous supposons que le nombre d'éléments unités N est un multiple de la taille d'échantillon n et que les éléments se trouvent sur une liste. Nous définissons alors l'intervalle d'échantillonnage $p = N/n$ et nous sélectionnons un chiffre au hasard inférieur ou égal à p. Si r désigne le chiffre tiré au hasard, l'échantillon systématique est défini comme l'ensemble d'unités identifiées par les chiffres $r, r + p, r + 2p, \ldots, r + (n - 1)p$. Si N n'est pas un multiple de n, nous supposons alors que le nombre entier le plus proche de N/n est l'intervalle d'échantillonnage. Dans certaines conditions, un échantillon sélectionné par échantillonnage systématique ressemble fortement à l'échantillon aléatoire simple.
- **L'échantillon aléatoire stratifié**. Si nous disposions d'informations supplémentaires sur la population, nous pourrions alors diviser la population selon ces informations en plusieurs groupes (sous-populations) appelés **strates**, chacune d'entre elles contenant des unités qui partagent une caractéristique. Nous prélevons ensuite dans chaque strate un échantillon aléatoire simple d'une taille adéquate.

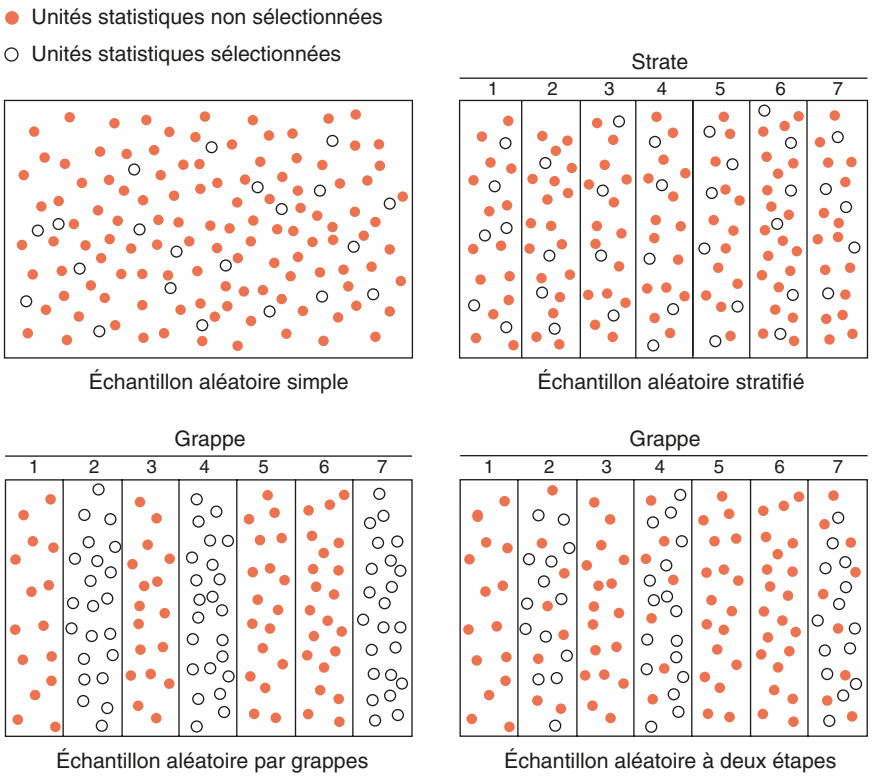
Figure 1.3 Exemple typique d'une enquête statistique

Crédit photo : Ambient Ideas/Shutterstock.

- **L'échantillon aléatoire par grappes ou par *clusters***. Supposons que la population soit divisée en un certain nombre de sous-ensembles, appelés grappes. Dès lors, il convient de prélever un échantillon aléatoire simple de ces grappes, puis d'y inclure toutes les unités qui appartiennent aux grappes sélectionnées.
- **L'échantillon aléatoire à deux étapes**. Nous pouvons le considérer comme un échantillon par grappes dans lequel l'enquête se limite à un échantillon aléatoire simple tiré de chaque grappe sélectionnée, plutôt que d'étudier toutes les unités contenues dans la grappe sélectionnée. Nous parlons alors de deux étapes d'échantillonnage : les grappes sont prélevées lors de la première étape, et les unités statistiques à l'étude sont prélevées lors de la seconde étape.

La figure 1.4 permet de visualiser des échantillons de type aléatoire simple, stratifié, par grappes (ou *cluster*) et à deux étapes.

Figure 1.4 Les techniques d'échantillonnage aléatoire les plus utilisées



Exemple 1.1

Prélevez un échantillon aléatoire simple de taille 10 de la population de l'exercice 1.6.

Solution

La population est composée de 92 unités numérotées de 1 à 92. Nous devons sélectionner au hasard 10 nombres entiers entre 1 et 92. À cet effet, nous pouvons utiliser l'un des nombreux logiciels (Excel, par exemple) dont sont dotés les ordinateurs. Supposons que les nombres suivants aient été tirés :

28, 62, 24, 49, 59, 72, 5, 16, 83, 11

L'échantillon aléatoire simple est alors composé des unités de population ainsi numérotées.

Exemple 1.2

Vingt étudiants $N = 20$ sont inscrits à un cours de master. Nous souhaitons à présent choisir un échantillon de la taille $n = 5$.

Solution
Nous commençons par dresser la liste des étudiants (par ordre alphabétique par exemple). Étant donné que l'intervalle d'échantillonnage est de $p = 20/5 = 4$, nous sélectionnons au hasard un nombre entier entre 1 et 4. Si nous avons sélectionné $r = 3$, l'échantillon systématique est composé des unités de population numérotées 3, 7, 11, 15, 19.

Exemple 1.3

Prélevez de la population de l'exercice 1.5 un échantillon aléatoire stratifié de taille 15 et stratifiez-le par sexes.

Solution
Lorsque l'on stratifie une population par sexes, on classe les unités dans deux strates, hommes et femmes. Ensuite, on prélève un échantillon aléatoire simple de chaque strate. Cela présuppose la création de deux listes séparées et la numérotation progressive des unités de chaque strate. Si l'on suppose que l'intégralité de la taille de l'échantillon est divisée en deux parties proportionnelles aux tailles des strates, la taille de l'échantillon aléatoire stratifié obtenu par le prélèvement d'un échantillon aléatoire simple est de 11 pour la strate des hommes et de 4 pour l'échantillon aléatoire simple pour la strate des femmes.

Exemple 1.4

L'enquête Emploi effectuée par l'Istat est un exemple d'échantillonnage aléatoire en deux étapes. Le but de cette enquête est de recueillir des informations sur le marché de l'emploi, telles que le nombre de personnes actives occupées, celles en recherche d'emploi, la distribution par secteurs d'activité, le statut (salarié ou indépendant), les qualifications, etc. La population est composée de personnes âgées de 15 ans et plus. Lors de la première étape, des communes sont tirées au hasard en vue de constituer un échantillon aléatoire. Lors de la deuxième étape, on prélève au hasard des ménages de cet échantillon aléatoire des communes.

Lorsque l'échantillon est construit suivant une des techniques aléatoires susmentionnées, des méthodes d'inférence statistique permettent d'extrapoler à l'ensemble de la population les données obtenues sur cet échantillon, à la suite d'un traitement adapté. Les enquêtes-échantillons réalisées par les instituts de statistiques nationaux, ou tout autre institut public, s'appuient généralement sur des échantillons aléatoires. De nombreuses autres enquêtes, menées pour la plupart par des organismes privés, préfèrent l'échantillonnage non probabiliste pour diverses raisons, notamment pour limiter les

coûts. Voici une liste brève et non exhaustive de quelques techniques d'échantillonnage non aléatoire :

- **L'échantillon par la méthode des quotas**. La population est divisée en strates suivant quelques caractéristiques structurelles de la population. Ensuite sont définis des quotas, à savoir le nombre de personnes à sonder pour chaque strate. Puis l'enquêteur choisit les personnes à interviewer par strate.
- **L'échantillon boule de neige**. On sélectionne un premier groupe de personnes à interviewer à partir de la population cible. Ces personnes désignent ensuite d'autres sujets qui appartiennent à la même population et qui présentent les mêmes caractéristiques. Cette approche est communément utilisée pour des populations rares ou faute de liste.
- **L'échantillon basé sur le jugement**. Cette méthode permet d'identifier des personnes expertes dans le domaine étudié (témoins privilégiés), en raison de leur formation ou de leur métier. On a recours à cette approche lorsque le sujet étudié est complexe ou délicat. Un exemple d'échantillonnage au jugé est la méthode de Delphes, qui consiste à collecter des informations d'un groupe d'experts que l'on consulte à propos de certains sujets ou situations.

Nous rappelons que toute enquête statistique s'appuie sur une **conception de l'enquête** qui réunit tous les aspects pertinents en vue de la réalisation de l'étude : notamment, définir la population cible, identifier les variables à observer, définir les outils de collection, estimer les coûts et la manière de les couvrir, choisir le plan d'échantillonnage et, en cas d'enquête-échantillon, identifier la méthode d'analyse des données.

Le questionnaire

Un **questionnaire** est une suite coordonnée de questions que l'on soumet à un ensemble d'unités statistiques (un échantillon ou une population). Construire un questionnaire est une tâche délicate de longue haleine, sachant que la qualité du questionnaire contribue grandement à la qualité des résultats obtenus à l'issue de l'enquête. Toutes les questions peuvent être classées dans les catégories suivantes :

- **Les questions ouvertes**. On ne propose pas de réponses prédéfinies au répondant.
- **Les questions fermées**. Le répondant choisit une réponse dans une liste de réponses prédéfinies.
- **Les questions mixtes**. Les questions sont assorties d'une liste de réponses, mais le répondant peut donner une réponse autre que celles proposées.
- **Les questions filtres**. Elles permettent de passer d'un ensemble de questions directement à un autre, afin d'éviter au répondant de s'attarder sur des questions sans pertinence pour lui.
- **Les questions structurées**. Il s'agit de questions assorties de réponses à des options prédéfinies et permettant aux répondants de choisir parmi une variété de combinaisons de réponses possibles, souvent présentées sous forme tabulaire.

La formulation des questions fermées mérite une attention particulière. Le nombre de réponses possibles et l'ordre de présentation peuvent influer sur le résultat. Nous associons souvent ces réponses à des grilles d'évaluation. L'**échelle de Likert** en est un bon exemple, très populaire en raison de sa simplicité. Les répondants expriment leur degré d'accord ou de désaccord concernant une suite d'affirmations, par exemple « tout à fait

d'accord », « d'accord », « ni en désaccord ni d'accord », « pas d'accord » et « pas du tout d'accord ». Un score est attribué à chaque modalité de réponse, par exemple (1, 2, 3, 4, 5) ou (2, 1, 0, –1, –2). La somme obtenue par chaque répondant pour l'ensemble des questions montre où il ou elle se situe par rapport au phénomène étudié.

Compte tenu de l'incidence du questionnaire sur la qualité des résultats du sondage, on procède généralement à quelques opérations au préalable, qui visent à tester le questionnaire. Le terme « prétest » est utilisé lorsqu'une version provisoire du questionnaire est remise à un petit groupe de personnes afin de repérer des problèmes que les répondants pourraient rencontrer lorsqu'ils répondent aux questions. Une autre possibilité consiste à remettre deux versions du questionnaire qui se différencient par un seul élément – par exemple, la formulation des questions – de deux échantillons similaires. Le but de cette démarche est de déterminer si les deux formulations produisent des résultats différents. Sinon, il est également possible de faire une enquête pilote, réalisée auprès d'un petit échantillon.

Le lecteur peut se reporter au questionnaire utilisé par Statistics Canada *pour l'enquête sur les ménages de 2011. L'annexe de ce chapitre propose un extrait de ce questionnaire.*

Les modes d'administration du questionnaire

Le questionnaire est administré lors d'une interview qui peut se dérouler en face à face, par téléphone, par courrier électronique ou par Internet. La méthode en face à face a été pendant longtemps le mode le plus répandu. Le répondant et l'enquêteur sont en contact direct : l'enquêteur se rend au domicile ou sur le lieu de travail du répondant, lui lit les questions préalablement imprimées, tout en présentant éventuellement du contenu visuel et en notant les réponses directement sur le questionnaire. De nos jours, les enquêteurs se servent d'ordinateurs portables pour administrer le questionnaire sous forme électronique. Ce cas de figure de l'enquête en face à face porte le nom de CAPI (*computer aided personal interview*). Dans l'**interview en face à face**, l'enquêteur utilise des questions supplémentaires, absentes du questionnaire, pour permettre aux répondants de mieux comprendre le sens des questions. Des interviews en face à face où le répondant remplit lui-même le questionnaire sont également très répandues.

Lors de l'**interview par téléphone**, l'enquêteur appelle la personne à interviewer, pose les questions et consigne les réponses sur le questionnaire imprimé. La méthode CATI (*computer assisted telephone interview*) est la plus répandue de nos jours : l'interview se déroule au moyen d'un questionnaire électronique et les informations ainsi collectées peuvent être directement encodées et enregistrées *via* un outil informatique. De cette façon, l'enquêteur peut rectifier des réponses incohérentes ou erronées, vu que les réponses sont traitées et vérifiées de sorte à déceler toute incohérence au moment de l'enregistrement.

Le **questionnaire par voie postale** consiste à envoyer au domicile ou sur le lieu de travail des personnes à interviewer une enveloppe contenant une lettre explicative, le questionnaire et une enveloppe de retour, habituellement affranchie et préadressée. L'approche est souvent mixte : l'envoi du questionnaire par la poste, puis une relance faite par téléphone (l'enquêteur poursuit parfois le remplissage du questionnaire ou administre le questionnaire entier à ce moment-là). Des expériences où des enquêtes comprenaient un questionnaire autoadministré ont été faites dernièrement : les personnes remplissent le questionnaire sur leur propre ordinateur avec la méthode CASI (*computer assisted self interview*).

L'**interview sur Internet** peut se faire de deux manières. Le questionnaire est envoyé par courrier électronique après la création d'une liste de diffusion. Les répondants

retournent ensuite le questionnaire par courrier électronique. Il peut également être diffusé sur un site web. Un courrier électronique est alors envoyé aux personnes à interviewer, accompagné du lien vers le site et des identifiants nécessaires pour se connecter et remplir le questionnaire.

Le tableau 1.1 présente une comparaison schématique des avantages et des inconvénients des différents modes d'administration d'un questionnaire.

Tableau 1.1 Avantages et inconvénients des différents modes d'administration d'un questionnaire

	Interview en face à face	Interview par téléphone	Interview par voie postale	Interview sur Internet
Avantages	• Faible taux de refus • Qualité élevée des informations recueillies, entre autres en raison des stimuli visuels et multimédias (avec la méthode CAPI) et de la possibilité d'utiliser des questions ouvertes et complexes • Possibilité de vérifier la composition de l'échantillon et l'identité de la personne interviewée	• Moins coûteuse en argent et en temps comparée au face-à-face • Facilité de joindre les personnes ciblées • Contrôle progressif de la composition de l'échantillon • Traitement en direct des informations recueillies (avec la méthode CATI)	• Peu coûteuse	*Courrier électronique :* • Peu coûteuse en temps et en argent *Site web :* • Peu coûteuse en temps et en argent • Présentation de stimuli multimédias, liens vers d'autres sites web • Données disponibles immédiatement pour analyse
Inconvénients	• Coûteuse en argent et en temps • Difficulté de surveiller les activités des enquêteurs, avec une incidence possible sur les répondants • Faible probabilité d'obtenir des informations sur des sujets délicats	• Impossibilité de fournir des stimuli visuels • Besoin de limiter le temps d'administration du questionnaire (questionnaire court) • Distorsion éventuelle en raison du contact avec l'enquêteur • Impossibilité de contacter les personnes sans téléphone	• Taux de réponse généralement faible • Problèmes de qualité liés à l'auto-administration • Impossibilité de vérifier l'identité de la personne interviewée	*Courrier électronique et site web :* • Faible taux de réponse • Problèmes de représentativité de l'échantillon *Site web :* • Faible taux de réponse • Problèmes de représentativité de l'échantillon

Figure 1.5 Cadre expérimental exemplaire : photo du centre de recherche de Rothamsted au Royaume-Uni, célèbre pour être un des plus anciens centres de recherche agricole au monde

Crédit photo : Centre de recherche de Rothamsted.

1.8.2 La méthode expérimentale

Le terme **expérience** s'emploie lorsque des personnes, des animaux ou des objets sont soumis à un traitement afin d'observer leur réponse à celui-ci. Par traitement, nous entendons une condition expérimentale bien spécifique pendant laquelle les unités statistiques sont observées. La condition expérimentale est déterminée par le niveau d'une ou de plusieurs variables, appelées **variables explicatives ou indépendantes**.

Exemple 1.5

Un ingénieur en développement produit souhaite étudier la résistance à la traction d'une nouvelle fibre synthétique destinée à la production de chemises. On sait grâce à des expériences antérieures que le pourcentage de coton dans la composition du tissu agit sur la résistance à la traction. On sait également que ce pourcentage doit se situer entre 10 et 40 % du poids total du tissu si l'on veut conférer certaines propriétés au produit fini, par exemple le rendre infroissable. L'ingénieur décide ainsi de définir plusieurs niveaux pour la variable indépendante, à savoir 15, 20, 25, 30 et 35, et prévoit 10 observations pour chaque niveau. (Exemple tiré de Montgomery, *Design and Analysis of Experiments*, 2005, p. 69.)

Analyse

Il s'agit ici d'une expérience avec une seule variable indépendante : le pourcentage de coton. Les traitements sont les différents niveaux de la variable indépendante,

donc les 5 % de coton. Le résultat est la résistance à la traction. Lors de l'expérience, il pourrait s'avérer judicieux de randomiser l'ordre des tests afin de neutraliser son effet éventuel sur le résultat (ici, l'effet produit par la température croissante de la machine).

Exemple 1.6

Pour concevoir une pile destinée à une utilisation dans des conditions où la température varie fortement, un ingénieur souhaite étudier l'effet de deux variables sur la durée de vie de la pile : le matériel utilisé pour la plaque et la température. Il décide de tester trois matériaux différents, A, B et C, et d'observer la durée à trois températures différentes, 15, 70 et 125 degrés. Il décide en outre de mener quatre essais pour chaque combinaison des deux variables. (Exemple tiré de Montgomery, *op. cit.*, 2005, p. 199.)

Analyse
Cet exemple possède deux variables indépendantes : le matériel et la température. Il y a 9 traitements, en raison des 9 combinaisons possibles avec les deux variables indépendantes. La durée de vie de la pile constitue la variable dépendante.

Exemple 1.7

Afin d'étudier l'effet d'une campagne publicitaire, une agence a conduit une expérience sur un échantillon de 200 étudiants. La durée du spot, diffusé pendant un programme télévisé de 50 minutes, était de 40 secondes dans la version courte et de 80 secondes dans la version longue. Le message a été répété deux et quatre fois. À la fin du programme, les sujets ont été questionnés sur le contenu du spot publicitaire et ont dû donner leur avis concernant le produit.

Analyse
Cet exemple contient deux variables indépendantes : la durée et la fréquence des spots publicitaires. Le nombre de traitements est de quatre, car c'est le nombre de combinaisons possibles entre deux durées et deux fréquences. Il existe au moins deux variables dépendantes : le niveau d'attention et le degré d'intérêt pour le produit suscité par la publicité. Un problème survient lorsqu'on assigne les étudiants aux traitements. Répartir de manière aléatoire 50 étudiants à chaque traitement pourrait limiter les effets des « **variables confondantes** » – il s'agit de l'ensemble des circonstances et des éléments susceptibles d'influer sur le résultat, en dehors de la variable indépendante – telles que l'âge, le sexe, le genre, la catégorie sociale, etc.

Exemple 1.8

Lors d'une étude clinique, un groupe de patients appelé **groupe expérimental**, ou traité, se voit administrer un nouveau médicament que l'on souhaite étudier,

alors que le deuxième groupe, nommé **groupe témoin** ou de contrôle, reçoit le traitement standard, c'est-à-dire celui utilisé habituellement pour la maladie en question.

Analyse
La variable indépendante est le nouveau médicament. Il existe deux traitements : le nouveau et l'ancien médicament. La variable dépendante est le résultat de la thérapie : guérison ou non-guérison. Il convient de répartir de manière aléatoire les patients entre les traitements, afin de neutraliser l'effet qu'ont les variables confondantes sur le résultat. Outre l'esquisse que nous présentons ici, nous insistons sur le fait que les essais portant sur les médicaments sont très complexes : ils sont soumis à des régulations très strictes de sorte à garantir l'innocuité des traitements et doivent suivre des procédures clairement définies dans le **protocole expérimental**.

1.8.3 Les études observationnelles ou de terrain

Dans les **études observationnelles** ou **de terrain**, il n'existe ni population finie, ni unités statistiques que le chercheur décide de répartir entre différents traitements. En réalité, les unités se répartissent elles-mêmes d'une certaine manière.

Exemple 1.9

Un exemple classique d'étude observationnelle porte sur les risques du tabagisme sur la santé. Imaginons qu'un groupe de chercheurs observe 500 sujets, dont 150 fumeurs et 350 non-fumeurs âgés entre 30 et 40 ans, qui ont accepté d'être suivis pendant 20 ans pour dépister des maladies communément liées au tabagisme, telles que des problèmes cardiaques, le cancer du poumon, etc.

Analyse
La variable explicative ou indépendante est le fait d'appartenir ou non au groupe des fumeurs : de cette manière, il existe deux traitements auxquels les sujets se sont assignés eux-mêmes. La variable dépendante est la survenue d'une des maladies liées au tabagisme. Les fumeurs jouent le rôle du groupe expérimental, les non-fumeurs constituent pour ainsi dire le groupe de contrôle. Le but de cette étude est de déterminer si la fréquence des maladies observées est significativement supérieure[6] parmi les fumeurs que parmi les non-fumeurs. Dans ce type d'étude, une attention particulière est accordée au contrôle des variables de nuisance qui pourraient accroître le nombre d'occurrences des maladies : il est important que les deux groupes, le groupe expérimental et le groupe de contrôle, soient aussi semblables que possible en ce qui concerne le sexe, le genre, l'âge et toute autre caractéristique qui pourrait avoir un effet sur la variable dépendante.

6 L'expression « significativement supérieur » s'utilise ici dans le sens de « considérablement plus grand ». Nous donnerons sa signification technique précise dans les chapitres dédiés à l'inférence statistique.

Exemple 1.10

Pour démontrer que les études observationnelles peuvent aboutir à des conclusions biaisées, Freedman *et al.* (2007) examinent l'exemple de la *pellagre*. Cette maladie fut observée pour la première fois en 1700 par le médecin espagnol Gaspar Casal parmi les pauvres vivant aux Asturies.

Au début des années 1800, la maladie se répandit à travers l'Europe et provoqua la mort de milliers de personnes, notamment en France, en Autriche et en Roumanie. À cette époque, les épidémiologistes considéraient la pellagre comme une maladie infectieuse résultant de mauvaises conditions d'hygiène et transmise par les simulies, une espèce de mouche hématophage. Cette conclusion s'appuya sur l'observation que la diffusion géographique était la même pour la maladie et l'insecte, et que ce dernier était particulièrement actif au printemps, la saison qui enregistrait le plus de cas de la maladie.

Analyse

Établir un lien entre la pellagre et les conditions environnementales susmentionnées (mauvaise hygiène et présence des simulies) laissa penser que ces dernières étaient à l'origine de la maladie et de sa propagation. Au début des années 1900, l'épidémiologiste américain Joseph Goldberg démontra que la pellagre était en réalité due à la malnutrition (certains aliments, tels que le maïs, contiennent peu de niacine) et qu'il ne s'agissait pas d'une maladie infectieuse. Cet exemple montre l'importance d'une interprétation prudente et méticuleuse lorsque l'on croit observer un lien de cause à effet entre les variables explicatives et les variables dépendantes.

1.9 La matrice de données

Après avoir vu comment « produire » des données à l'aide d'enquêtes, d'expériences ou d'études de terrain, nous allons à présent aborder la façon d'organiser les données en vue de les traiter ou de les diffuser à autrui pour étude ou analyse.

La meilleure façon d'organiser les données statistiques est de le faire sous forme de matrice : la **matrice de données**. Chaque ligne correspond à une unité et chaque colonne à une variable. Ainsi, la matrice est composée d'autant de lignes que d'unités et d'autant de colonnes que de variables observées.

Exemple 1.11

L'exercice 1.6 donne un exemple de matrice de données. Les données portent sur sept variables quantitatives observées chez 92 personnes. À titre illustratif, voici les dix premières lignes de la matrice (voir tableau 1.2).

Analyse

Chaque ligne du tableau nous renseigne sur les valeurs prises par les variables observées sur une seule unité. Les colonnes nous fournissent les valeurs que prend une variable donnée sur toutes les unités statistiques de l'ensemble étudié. Par exemple, la première ligne nous montre que l'unité 1 a 22 ans, pèse 79 kg, et ainsi de suite. La quatrième colonne nous indique des informations sur la taille de chaque unité : la première unité mesure 184 cm, la deuxième mesure 168 cm, et ainsi de suite.

Tableau 1.2 **Extrait de la matrice de données décrite dans l'exercice 1.6**

Unité	Âge (en années)	Poids (en kg)	Taille (en cm)	Tour de cou (en cm)	Tour de poitrine (en cm)	Tour de taille (en cm)	Cuisse (en cm)	Avant-bras (en cm)	Poignet (en cm)
1	22	79	184	38,5	94	83	59	28,9	18,2
2	22	70	168	34,0	96	88	60	25,2	16,6
3	23	70	172	36,2	93	85	59	27,4	17,1
4	23	90	187	42,1	100	89	63	30,0	19,2
5	23	73	184	35,5	92	77	56	27,2	18,2
6	23	85	197	38,0	97	85	59	29,7	18,3
7	24	84	181	34,4	97	100	63	27,7	17,7
8	24	95	190	39,0	105	94	66	30,6	18,8
9	24	71	180	35,7	93	82	56	28,3	17,3
10	24	95	185	39,2	102	99	71	30,3	18,7

Il est parfois impossible d'obtenir toutes les données prévues par l'enquête statistique. Certaines unités peuvent refuser de coopérer (non-réponse totale) ou communiquent des informations incomplètes (non-réponse partielle). Dans le premier cas, quelques lignes de la matrice de données sont manquantes ; dans le deuxième cas, les lignes qui correspondent à des unités avec non-réponse partielle contiennent des cellules vides. Pour atteindre l'objectif de l'étude, il est préférable que la matrice soit complète. Mais comment remplir les cellules vides ? Il existe des techniques statistiques pour remédier à ce problème tout en garantissant des résultats statistiques valables. Nous allons revenir à la matrice de données, le problème de non-réponses et d'autres aspects ayant trait à la qualité des données dans le chapitre suivant, après avoir abordé les concepts nécessaires à une analyse plus approfondie de ces sujets.

1.10 La statistique descriptive et l'inférence statistique

La statistique se divise généralement en deux branches : la **statistique descriptive** et l'**inférence statistique**. La statistique descriptive réunit des outils conçus pour organiser, présenter et résumer des données. Faire des tableaux, dessiner des graphes, calculer des moyennes ou toute autre mesure permettant de faire un résumé sont des pratiques censées améliorer notre compréhension du phénomène à l'étude, grâce à une représentation simplifiée des données.

L'inférence statistique, en revanche, traite des méthodes qui nous permettent de tirer une conclusion générale de l'observation des résultats d'un échantillon, comme lors d'enquêtes-échantillons, où un échantillon aléatoire est analysé afin de recueillir des informations sur la population dont est issu l'échantillon. L'inférence statistique se divise en deux branches : les **tests d'hypothèses** et l'**estimation des paramètres**. Dans les deux domaines, le but est d'acquérir des connaissances sur des grandeurs numériques (par exemple, la moyenne de la population) qui décrivent un aspect de la population étudiée.

Ces quantités sont appelées des paramètres. Avec la technique des tests d'hypothèses, les données de l'échantillon et la théorie des probabilités sont exploitées afin de décider si une hypothèse sur un paramètre est rejetée ou non, comme l'hypothèse selon laquelle la moyenne de la population est inférieure ou égale à une valeur donnée. Quant à la technique de l'estimation des paramètres, elle consiste à utiliser les données de l'échantillon et la théorie des probabilités de sorte à imputer une valeur ou un ensemble de valeurs à un paramètre de la population.

Dans les enquêtes s'appuyant sur des données réelles, il n'existe aucune démarcation stricte entre les deux domaines : le chercheur peut recourir à des outils de la statistique descriptive (tableaux de fréquences, représentations graphiques, etc.) pour décrire les données disponibles, par exemple une étape préalable à la formulation d'une hypothèse ou à l'estimation d'un paramètre auquel il s'intéresse.

Dans cet ouvrage, nous aborderons la statistique descriptive dans les chapitres 2 à 11, les probabilités dans les chapitres 12 à 16 et l'inférence statistique dans les chapitres restants. Nous informons le lecteur que la statistique descriptive sera présentée avec des notations adaptées à une population finie de N unités. Il est évidemment possible d'appliquer aux données d'échantillon, *mutatis mutandis*, les formules et les procédures que nous traitons ici.

1.11 Des calculs statistiques élémentaires : les différences relatives et les ratios

Le moment est venu d'introduire quelques notions de calcul utiles dans la comparaison d'agrégats statistiques. Soit a et b deux grandeurs exprimées dans la même unité de mesure. La **différence relative** entre les deux (si nous prenons a comme valeur de référence par exemple) est définie comme étant le rapport :

$$\frac{b-a}{a}$$

Ce rapport est une grandeur sans dimension (un nombre pur). La multiplication par 100 de la différence relative nous donne alors la **différence en pourcentage**, ou **taux de variation** :

$$\text{Différence en pourcentage} = \frac{b-a}{a} \times 100$$

Exemple 1.12

Le 31 décembre 2015, les populations résidentes en Italie et en Espagne s'élevaient respectivement à 60,6 millions et à 46,5 millions. La différence en pourcentage entre ces deux grandeurs, prenant la population de l'Espagne comme valeur de référence, est la suivante :

$$\frac{60,6 - 46,5}{46,5} \times 100 = 30,3$$

La population italienne est donc 30,3 % plus importante que la population espagnole.

La comparaison de deux grandeurs, *a* et *b*, peut se faire par un rapport, soit *a/b*, soit *b/a*. Le rapport *a/b* démontre combien de fois *a* contient *b* : s'il est de 1,5, cela signifie que *a* est égal à 1,5 fois *b* ; s'il est de 0,8, alors *a* est 0,8 fois *b* et ainsi de suite. La multiplication par 100 de ces rapports nous donne des **rapports en pourcentage**.

Si un ensemble de grandeurs est exprimé dans la même unité de mesure, $a_1, a_2, \ldots, a_k$, il est possible de comparer les grandeurs individuelles au total global $A = a_1 + a_2 + \ldots + a_k$ avec les rapports en pourcentage suivants :

$$\frac{a_1}{A} \times 100,\ \frac{a_2}{A} \times 100,\ \ldots,\ \frac{a_k}{A} \times 100$$

La somme est évidemment 100.

Les rapports susmentionnés – parfois appelés **rapports partie au tout** – expriment les grandeurs individuelles $a_1, a_2, \ldots, a_k$ en pourcentage du total *A* et permettent de facilement comprendre leur importance relative comparée au total.

Exemple 1.13

Le tableau 1.3 donne un exemple de rapports en pourcentage sur la fréquence d'activités artistiques dans l'UE-28.

Tableau 1.3 Fréquence d'activités artistiques dans l'UE-28

Fréquence	Pourcentage
Chaque jour	6
Au moins une fois par semaine (pas chaque jour)	12
Au moins une fois par mois (pas chaque semaine)	12
Au moins une fois par an (pas chaque mois)	5
Aucune activité durant les 12 derniers mois	65
Total	100

Source : Eurostat, 2015.

Chacun des pourcentages ci-dessus est exprimé par le rapport (multiplié par 100) entre le nombre de personnes âgées de 16 ans et plus, qui font partie d'une catégorie donnée, et le total de la population de 16 ans et plus.

En comparant les grandeurs $a_1, a_2, \ldots, a_k$ les unes aux autres, on définit les rapports appelés **rapports partie à partie**. On a fréquemment recours à ces rapports dans l'analyse des phénomènes démographiques. Le **rapport de masculinité** en est un exemple. Il est exprimé par le rapport entre le nombre d'hommes et de femmes dans une population, multiplié par 100.

Exemple 1.14

Le 31 décembre 2015, la population résidente de l'Union européenne comptait 249,4 millions d'hommes et 260,9 millions de femmes (*source* : Eurostat). Le rapport de masculinité était :

$$\frac{249,4}{260,9} \times 100 = 95,6$$

Il y avait donc 95,6 hommes pour 100 femmes.

D'autres rapports de ce type peuvent se trouver parmi les indicateurs de la structure par âges d'une population. Nous pouvons citer ici l'**indice de vieillesse**, qui est le rapport (multiplié par 100) entre la population âgée de 65 ans et plus et la population qui a moins de 15 ans, et le **taux de dépendance des personnes âgées**, qui s'obtient en divisant la population de plus de 65 ans par la population âgée de 15 à 64 ans, puis en multipliant ce résultat par 100.

Nous devons également mentionner ces rapports, appelés **taux**, où le numérateur est le nombre d'occurrences d'un événement durant une période et où le dénominateur est le nombre de personnes exposées à cet événement durant cette période. Le **taux de natalité** en est un exemple : c'est le rapport entre le nombre de naissances vivantes durant une année donnée et la population en milieu d'année[7], multiplié par 1 000. Ce taux reflète le nombre de naissances pour 1 000 habitants sur une année.

Exemple 1.15

Le tableau de l'exercice 1.8 fait état des naissances vivantes et de la population en milieu d'année de quelques pays membres de l'UE en 2018. Grâce à ces données, on peut calculer les taux de natalité. Par exemple, le taux de natalité de l'Autriche est de 9,66 (sur 1 000 habitants).

Pour conclure, n'oublions pas les **rapports de densité**, où le numérateur est le niveau d'une variable donnée et où le dénominateur est le niveau (longueur, surface, période, etc.) d'une deuxième variable à laquelle la première est liée. Voici quelques exemples : densité de population (nombre d'habitants par kilomètre carré), nombre de véhicules par kilomètre de route, revenu par habitant, etc.

7 *Remarque* : le facteur de multiplication n'est pas toujours 1 000. Quand la proportion entre les quantités que l'on compare est inférieure à un millième, on utilise 10 000 comme multiplicateur, ou même une puissance de 10 plus grande encore. Si la comparaison s'effectue entre un petit nombre (c'est-à-dire des événements rares tels que des crimes ou des maladies orphelines) et un grand nombre tel que la population d'un pays, le multiplicateur utilisé est 100 000. Le but est de rendre le taux plus lisible, avec au moins un chiffre avant la virgule décimale.

Résumé

Ce chapitre donne une appréciation d'ensemble de la statistique et présente un bref aperçu du développement de cette discipline du XVI^e^ au XX^e^ siècle, tout en soulignant son rôle dans la recherche scientifique et dans la vie de tous les jours. Les termes techniques les plus significatifs (population, échantillon, unité statistique, variable) sont ensuite présentés et expliqués. La suite consiste en une analyse des différents types de variables (nominales, ordinales, discrètes, continues) et en l'étude de la façon de les mesurer.

La production des données occupe une place centrale, qui établit une distinction entre enquêtes, études expérimentales et études observationnelles. Dans la section couvrant les enquêtes statistiques, nous illustrons les méthodes fondamentales visant à construire un échantillon aléatoire, tout en abordant les difficultés liées à la formulation du questionnaire et à ses modes d'administration. La définition générale d'une expérience est donnée à travers plusieurs exemples tirés d'applications concrètes. Nous avons traité le sujet des études observationnelles ou de terrain d'une manière similaire.

La dernière partie du chapitre se consacre à la définition de quelques outils statistiques élémentaires (rapports, taux, indices) dont l'usage permet de comparer des agrégats statistiques de façon pertinente.