CHAPITRE 1

Éléments de base

Dans ce chapitre, nous commencerons par définir la statistique et rappellerons ensuite quelques notions mathématiques nécessaires pour comprendre le contenu du livre. Parmi celles-ci, nous aborderons les différents types de variables, la sommation, comment arrondir, et les fréquences. Suivra une explication des échelles de mesure qui sont à la base d'une bonne utilisation de la statistique. Nous terminerons avec une brève description des notions de population et d'échantillon.

Définition de la statistique

Il existe un grand nombre de définitions de la statistique mais nous n'en présenterons qu'une seule car elle est suffisamment complète :

> « Ensemble des méthodes qui ont pour objet la collecte, le traitement et l'interprétation des données d'observation relatives à un groupe de personnes ou d'objets ». (*Grand Dictionnaire de la psychologie.* Larousse, 1991).

Notons que dans l'ouvrage présent, nous ne traiterons pas de la collecte des données (méthodes d'échantillonnage, expériences, etc.) mais plutôt de leur traitement (ou analyse) ainsi que de leur interprétation. Notons aussi que le mot « statistique » utilisé au

singulier représente la science (la matière de cet ouvrage donc) alors que le pluriel, « les statistiques », désigne les données obtenues ou calculées (voir la dernière section de ce chapitre). Il faut donc faire attention au terme que l'on utilise. On se sert de *la* statistique pour obtenir *des* statistiques (ex. le salaire médian, l'âge moyen, etc.). Ce qui suit concerne avant tout *la* statistique.

L'opinion du grand public face à cette science est parfois empreinte de méfiance et de scepticisme. On entend des phrases du type, « On peut tout faire dire à la statistique » ou « La statistique, ce n'est que des mensonges » ! D'ailleurs, n'est-ce pas le Premier ministre victorien, Disraeli, qui a dit : « Il y a trois types de mensonges : les mensonges, les gros mensonges, et la statistique » ? Un des objets de ce livre sera d'expliquer comment « faire de la statistique », processus empreint de rigueur et régi par des règles et des conventions. Avec les connaissances présentées ici, nous souhaitons donner au lecteur les outils nécessaires pour faire de la statistique de base et comprendre les statistiques obtenues. Cela lui permettra, s'il le faut, de détecter les approches erronées et les interprétations douteuses émises par certains. Nous faisons nôtre le commentaire de Robert Abelson[1], qui souligne qu'au lieu de rejeter sans arrière-pensée tout énoncé qui s'appuie sur des chiffres, une approche plus raisonnable consiste à connaître assez de statistique pour pouvoir distinguer les conclusions honnêtes et utiles des déclarations insensées et malhonnêtes.

1. Robert Abelson, *Statistics as Principled Argument*, Hillsdale, New Jersey, Lawrence Erlbaum Associates, 1995

Notions mathématiques essentielles

■ Variables

Une variable est une caractéristique ou propriété représentée par un nom (ex. « Poids », « Couleur ») ou par un symbole (*X*, *Y*, etc.). Il existe deux grandes catégories de variables : les variables catégorielles (ou qualitatives) et les variables quantitatives (ou de mesure). On appelle modalités les différentes valeurs possibles d'une variable. Dans les variables catégorielles, les modalités sont des catégories. Par exemple, la variable « Sexe » possède deux modalités : « Masculin » et « Féminin » ; la variable « Couleur » plusieurs modalités (« Rouge », « Noir », « Blanc », etc.). Les données sont soit l'effectif soit la fréquence de chaque catégorie. Dans les variables quantitatives, les modalités sont des valeurs numériques mesurables. Le poids, par exemple, peut être une variable dans une étude, et être accompagnée de plusieurs valeurs (ex. 63, 72, 84 kg). Notons que le mot « facteur » est parfois utilisé à la place de variable, notamment lorsque l'on cherche à expliquer les valeurs d'une variable en fonction d'une autre que l'on appelle facteur.

Une variable quantitative est dite « discrète » si l'étendue des valeurs possibles est dénombrable (ex. le nombre de personnes dans une famille, le nombre de mots dans une phrase, etc.). Elle est dite « continue » lorsque les valeurs possibles ne sont pas parfaitement dénombrables dans le sens où la mesure peut toujours être plus précise. Par exemple, la taille, le poids, l'âge, etc.

Dans le monde de l'expérimentation on distingue les variables indépendantes, dépendantes et de contrôle. Une variable indépendante (ou variable explicative) correspond à ce qui est manipulé ou choisi par le chercheur (ex. le type de médicament) ; elle figure sur l'abscisse (l'axe horizontal) d'un graphe (voir la figure 1.1). Une variable dépendante (ou variable à expliquer) correspond à la mesure obtenue sur laquelle l'étude porte (ex. le temps de guérison) ; elle figure sur l'ordonnée (l'axe vertical) du graphe. Enfin, une variable de contrôle concerne l'aspect qui doit être contrôlé pour ne pas interférer dans l'étude (ex. l'âge du patient). Afin de réussir une recherche, il faut souvent contrôler plusieurs variables qui pourraient venir « biaiser » le résultat.

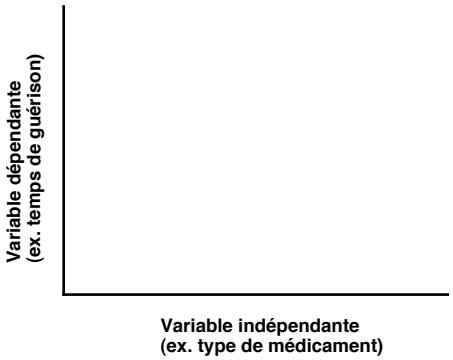

Figure 1.1 – Axes d'un graphe montrant l'emplacement des variables indépendante et dépendante

■ Sommation

Prenons deux variables, *X* et *Y*, avec leurs valeurs respectives représentées par *x* et *y* :

Tableau 1.1 – Les valeurs de deux variables

x	*y*
1	2
2	3
5	4
4	5

S'il faut additionner toutes les valeurs de *X*, nous utiliserons le symbole $\sum$ (sigma majuscule) accompagné d'un petit *x* pour désigner les valeurs, ce qui donne $\sum x$. Donc dans ce cas, $\sum x = 12$. Qu'en est-il de la somme des valeurs de *Y*? $\sum y = 14$. Le symbole sigma sera présent dans de nombreuses formules du livre, comme dans les cas suivants : $\sum x^2$, $\sum y^2$, $\sum xy$. Pour obtenir ces sommes, il faut agrandir le tableau présenté ci-dessus et créer des colonnes pour chaque opération demandée. Pour $\sum x^2$ et $\sum y^2$, il faut élever les *x* et les *y* au carré et ensuite prendre la somme de chaque colonne, et pour $\sum xy$, il faut multiplier les valeurs de *X* et de *Y* et ensuite calculer la somme, comme on le voit dans le tableau 1.2.

Tableau 1.2 – Calculs supplémentaires avec les valeurs de *X* et *Y*

x	y	x^2	y^2	xy
1	2	1	4	2
2	3	4	9	6
5	4	25	16	20
4	5	16	25	20
$\sum x = 12$	$\sum y = 14$	$\sum x^2 = 46$	$\sum y^2 = 54$	$\sum xy = 48$

D'autres opérations qui pourront être requises sont tout aussi simples. Par exemple, si l'on trouve $(\sum x)^2$, il faut d'abord prendre la somme des x (= 12) et ensuite l'élever au carré (cela donne 144). De même pour $(\sum y)^2$ qui correspond à $14^2 = 196$. Attention ! $\sum x^2$ et $(\sum x)^2$ ne donnent pas le même résultat, comme nous venons de le voir ($\sum x^2 = 46$ et $(\sum x)^2 = 144$). Il faut donc être sur le qui-vive ! Enfin, si l'on voit $(\sum x)(\sum y)$, il faut tout simplement prendre la somme des x (à savoir 12) et la somme des y (à savoir 14), et ensuite les multiplier (cela donne 168).

■ Arrondir

Nous pensons tous savoir arrondir des nombres mais en fait les règles peuvent parfois être complexes. Pour cet ouvrage d'introduction, nous proposons une approche simple. Prenons l'exemple d'un nombre avec trois décimales que l'on souhaite arrondir :

- si le troisième chiffre est supérieur ou égal à 5, alors on augmente le deuxième chiffre après la virgule d'une unité. Par exemple, 8,546 est arrondi à 8,55 ;
- si le troisième chiffre est moins grand que 5, alors on l'enlève tout simplement ; donc 8,332 est arrondi à 8,33.

■ Fréquence

Le mot « fréquence » possède deux sens qu'il ne faut pas confondre :

- La fréquence absolue, que l'on nomme habituellement « effectif » : elle correspond au nombre d'objets, d'éléments ou d'individus dans une catégorie. Elle se note généralement par le symbole « n ».

– La fréquence relative, qui est le rapport entre l'effectif d'une catégorie et l'effectif total ; elle est souvent exprimée en pourcentage, qui représente ce rapport multiplié par 100. Son symbole est « *f* » ou « % » (pour le pourcentage). Prenons un exemple : on trouve 13 mots de trois syllabes dans un texte de 50 mots. La fréquence absolue du nombre de mots de trois syllabes (ou l'effectif) est donc 13 ; la fréquence relative est 13/50 = 0,26, et le pourcentage est 26 %.

Un mot de prudence ici. Très souvent, les deux fréquences apportent des informations différentes et l'une ne doit pas occulter l'autre, au moins dans la discussion de résultats. Par exemple, nous savons qu'il y a eu 4,26 % d'accidentés de la route en 2007 avec séquelles majeures dans un pays de l'Union européenne. Dire que cela est assez peu ne tient pas compte du fait qu'il s'agit tout de même de 4 400 personnes dont la vie est malheureusement affectée de manière permanente ! Il n'est donc pas surprenant que certains s'en prennent à la statistique lorsqu'ils constatent l'utilisation d'un seul type de fréquence, et la mise de côté de l'autre, à des fins oratoires dans un débat.

■ Les échelles de mesure

En statistique nous « mesurons », dans le sens où nous faisons correspondre des nombres aux individus, objets ou événements qui nous intéressent, selon certaines règles. Pour mesurer, nous utilisons des « échelles de mesure » qui ont des propriétés propres. Le choix de l'échelle est le nôtre mais une fois qu'il est fait, les outils statistiques qui sont utilisables avec l'échelle sont fixés. Il est donc important que le choix soit judicieux car les calculs statistiques en dépendent.

Prenons un exemple pour illustrer le fait que pour une même variable, on peut parfois utiliser une échelle différente. Nous souhaitons classer Marie, Joanne et Susanne selon leur taille. Dans un premier temps, nous pourrions décider d'ordonner les trois personnes selon leur taille et d'attribuer un rang à chacune d'elle. Par exemple, Joanne, qui est la plus grande, recevrait le nombre 1 ; Marie, qui se situe entre les deux, aurait le nombre 2, et Susanne, la moins grande, serait représentée par le nombre 3. Nous venons d'utiliser une échelle ordinale dans laquelle les nombres (dits ordinaux) servent à marquer l'ordre ou le rang.

Avec l'utilisation d'une telle échelle, nous ne savons pas quelle est la distance entre ces personnes, c'est-à-dire, la différence en centimètres entre elles. Pour le savoir, il nous faut quantifier la taille des trois personnes et utiliser des nombres dits cardinaux qui désignent une quantité. Nous constatons alors que Joanne mesure 1 m 75, Marie est toute proche avec 1 m 74, et Susanne mesure 1 m 60. Grâce à une telle échelle, dite de rapports, nous connaissons maintenant la distance exacte qui sépare les trois personnes et constatons que Joanne est à peine plus grande que Marie alors que Susanne est bien moins grande que les deux autres. Cette dernière échelle nous donne donc davantage d'informations (elle est plus puissante) que la première utilisée. En somme, les nombres, bien qu'en apparence égaux, ne le sont pas forcément. Ils peuvent provenir d'échelles différentes et ainsi avoir des propriétés qui sont différentes les unes des autres. Nous venons de voir qu'un nombre peut marquer un rang (nombre ordinal) ou désigner une quantité (nombre cardinal). Lorsque l'on traite de données en statistique, il est donc important de toujours savoir le type de nombres, et donc l'échelle, dont il s'agit. Nous allons maintenant passer en revue quatre échelles de mesure, de la moins puissante à la plus puissante, et allons donner les caractéristiques de chacune.

■ Échelle nominale

Avec cette échelle, on nomme des catégories en utilisant des nombres nominaux. Voici quelques exemples de nombres qui proviennent de cette échelle : les numéros de téléphone et de sécurité sociale ; les numéros sur les tenues de sport ; les codes postaux, etc. Les opérations possibles avec cette échelle sont limitées : on peut indiquer si deux catégories (ou nombres) sont identiques ou pas ($444 \neq 22$) ; on peut compter le nombre de valeurs dans chaque catégorie (ex. le nombre d'enfants, le nombre d'adultes, etc.) ; on peut également regrouper les catégories.

■ Échelle ordinale

Avec cette échelle, on peut marquer le rang (l'ordre), et ce grâce à des nombres ordinaux. Voici quelques exemples : la position d'arrivée dans une course, l'ordre de naissance dans une famille et la dureté des minéraux (échelle de Mohs). Les opérations possibles

avec ces nombres sont d'abord celles de l'échelle nominale ; en plus, on peut indiquer si un nombre est plus grand, plus petit, ou égal à un autre nombre (ce qui n'était pas le cas de l'échelle précédente). Par contre, l'échelle ne permet pas d'indiquer l'importance de la différence. Par exemple, quand Michel dit qu'il est arrivé deuxième dans une course, cela ne nous dit pas à quelle distance se trouvaient le premier et le troisième. D'ailleurs, il est parfois plus « avantageux » de dire qu'on était second (échelle ordinale) que d'annoncer que le premier est arrivé trois minutes plus tôt et le troisième deux secondes après soi (échelle de rapports) !

■ Échelle d'intervalles

Cette échelle marque la quantité, l'ordre de grandeur, à l'aide de nombres cardinaux. Voici quelques exemples : la température en degrés Celsius ou Fahrenheit, les dates du calendrier, les valeurs d'un test de QI, etc. Notons que cette échelle ne contient pas de zéro absolu. En effet, 0 degré Celsius ne correspond pas à l'absence de chaleur et l'an 0 ne représente pas le début de l'histoire. Les opérations possibles avec les nombres de cette échelle sont celles de l'échelle nominale et de l'échelle ordinale. En plus, on peut additionner ou soustraire des nombres ou une constante (chose impossible avec les deux précédentes échelles) mais on ne peut pas faire de multiplication ou de division car on changerait alors le rapport entre les nombres (ceci à cause de l'absence d'un zéro absolu).

■ Échelle de rapports

Comme l'échelle précédente, l'échelle de rapports marque la quantité, l'ordre de grandeur, à l'aide de nombres cardinaux. Cette échelle permet toutes les opérations de l'échelle d'intervalles, mais, en plus, on peut faire des multiplications et des divisions sans que cela ne change le rapport entre les nombres (ceci à cause de la présence du zéro absolu, à savoir l'absence de l'élément mesuré). Parmi des exemples de cette échelle on trouve la taille, le poids, la distance, l'âge, etc. Quant à la température, l'échelle Kelvin est une échelle de rapports car 0 Kelvin représente le point triple de l'eau (coexistence des trois états : liquide, solide et gazeux).