1

STATISTIQUE DESCRIPTIVE

Table des matières

L'activité statistique comprend deux aspects complémentaires : la *statistique descriptive*, partie indispensable de tout projet d'analyse de données; et *l'inférence statistique*, une étape qui suit souvent — mais pas toujours — l'analyse descriptive. Un sondage, une enquête, une recherche scientifique donnent lieu à des données si massives qu'on ne peut les déchiffrer qu'après les avoir réduites à des dimensions compréhensibles. Supposons, par exemple, que l'administrateur d'une petite municipalité rassemble des données recueillies auprès de 5 000 ménages de la ville — des données sur le nombre de personnes qui y vivent, ou le nombre d'enfants, ou la consommation d'eau. Le résultat est un indigestible tableau de 5 000 lignes et autant de colonnes que de questions posées. Les techniques qui permettent de donner un sens à ces données — les tableaux, les graphiques, ainsi que les « mesures statistiques » telles les moyennes et les ratios — font partie de la statistique descriptive.

Si la ville ne comprend que ces 5 000 ménages, alors l'étude est un *recensement*, et les 5 000 ménages constituent ce qu'on appelle une *population* : c'est la totalité des objets d'intérêt.

Dans ce cas les analyses descriptives suffisent. Mais un recensement est très couteux. À l'échelle d'un pays ou d'une compagnie multinationale, c'est un projet colossal. À l'échelle d'une PME aussi, toute proportion gardée.

Il est donc généralement nécessaire de se limiter à une partie de la population, un *échantillon*. Les données qu'on en tire seront quand même soumises à une analyse descriptive : qu'il s'agisse d'un échantillon ou d'une population, il faut que les données soient résumées. Mais l'étude ne peut s'arrêter là, car son objet, c'est la population, pas l'échantillon. L'échantillon n'est qu'une image de la population, une image qu'on souhaite fidèle, qui *peut* ressembler à la population, mais qui n'est jamais parfaite. Il révèle — à peu près — certaines des caractéristiques de la population, mais il peut aussi « révéler » des choses qui ne sont pas vraies. Il peut présenter des familles plus nombreuses dans le secteur Sud que dans le secteur Nord, sans que cela soit le cas dans la population. Dans quelle mesure peut-on se fier aux observations faites sur un échantillon et déduire qu'elles sont également vraies (ou à peu près vraies) de la population ? C'est à ces questions que doivent répondre les techniques d'inférence statistique.

Des questions que nous n'aborderons formellement qu'au chapitre 6, lequel sera précédé de trois chapitres dans lesquels nous développerons les éléments de la *théorie des probabilités*, base théorique de l'inférence statistique. Dans le présent chapitre, cependant, et dans le suivant, nous nous concentrons sur la statistique descriptive et donc nous n'y ferons pas la distinction — très importante, par ailleurs — entre *population* et *échantillon*.

1.1 INTRODUCTION : VARIABLES ET DISTRIBUTIONS

Pour commencer, un peu de vocabulaire. Le tableau A.01 en annexe (dont une partie est reproduite ci-dessous) présente une série de données concernant des professeurs d'université. L'ensemble des professeurs constitue la **population**. Les membres d'une population sont appelés des **unités statistiques**, ou simplement des **unités**. Chaque ligne du tableau représente une unité, identifiée par un numéro (qui peut aussi être un nom) dans la première colonne. Chaque colonne représente une **variable**, et les nombres ou lettres qui y figurent sont les **valeurs**, ou **modalités** de la variable. La variable « Sexe », par exemple, a pour modalités les lettres *F* et *M* ; les valeurs de la variable « Date d'entrée » sont des entiers de 1980 à 2012. Les valeurs de la variable « Département » sont les noms des différents départements de l'université.

Extrait du tableau A.01 - Quelques données sur un groupe de professeurs

Identité	Sexe	Date d'entrée	Département	Salaire à l'entrée	Salaire en 2012	Expérience
1	F	1995	Management	16 598	109 268	22
2	M	1984	Management	9 386	134 244	27
3	F	2008	Management	34 446	81 170	22
4	M	1990	Sc. économiques	15 962	159 532	30
5	M	1999	Marketing	23 413	153 600	20
6	M	1995	Sc. comptables	19 838	140 175	28
7	M	2007	Management	34 541	107 395	21
8	F	2001	Finance	30 797	126 751	48
9	M	2004	Sc. comptables	20 726	109 893	22
10	M	1990	Finance	13 038	126 751	28
11	M	2007	Sc. comptables	30 005	91 786	21
…	…	…	…	…	…	…

On distingue deux catégories de variables : les variables *quantitatives*, comme les salaires et le nombre de mois d'expérience, sont celles dont les modalités sont des quantités ; et les variables *qualitatives*, dont les modalités sont des caractéristiques non mesurables, comme le département et le sexe.

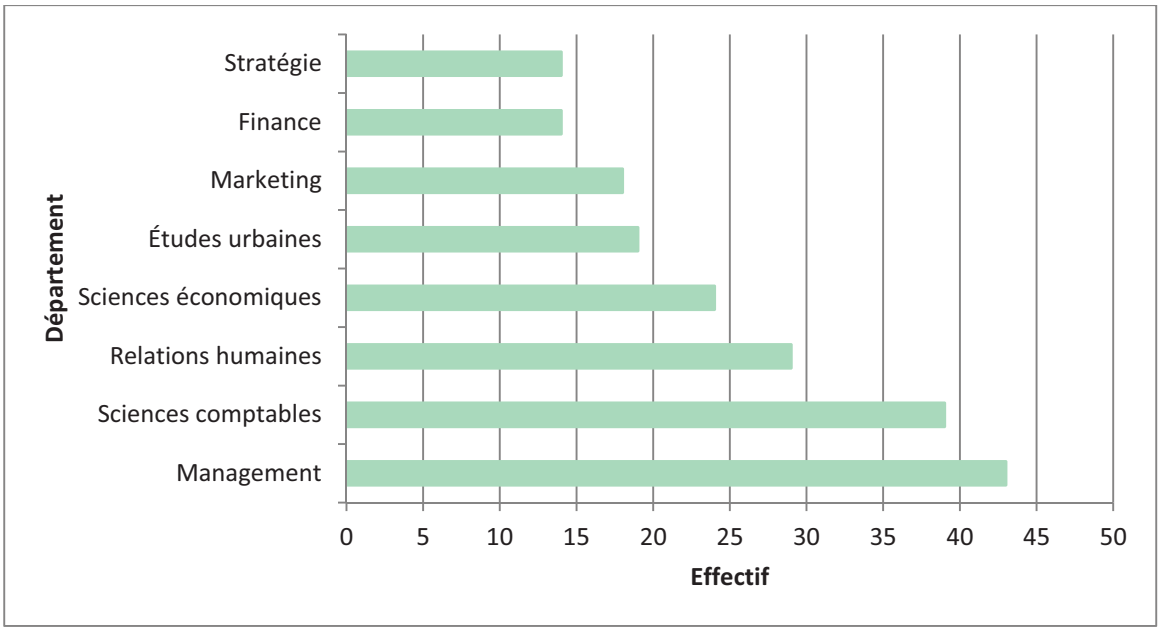
On peut résumer les caractéristiques d'une variable en énumérant ses modalités et en indiquant la fréquence de chacune dans la population. Le tableau 1.1 résume les caractéristiques de la variable « Département ». L'effectif correspondant à une valeur donnée *x*, c'est le nombre d'unités pour lesquelles la variable prend la valeur *x*. *L'effectif total n* est le nombre d'unités dans la population (ou l'échantillon). La *fréquence* d'une valeur est l'effectif de cette valeur divisé par *n*. La somme des fréquences est nécessairement égale à 1. La fréquence est parfois multipliée par 100, de façon à représenter un pourcentage. Cette correspondance entre les valeurs d'une variable et les effectifs ou fréquences correspondantes est appelé *distribution*.

Tableau 1.1 Distribution de la variable « Département » - Données du tableau A.01

Valeurs	Effectif	Fréquence
Études urbaines	19	0,095
Finance	14	0,070
Management	43	0,215
Marketing	18	0,090
Relations humaines	29	0,145
Sciences comptables	39	0,195
Sciences économiques	24	0,120
Stratégie	14	0,070
	200	1

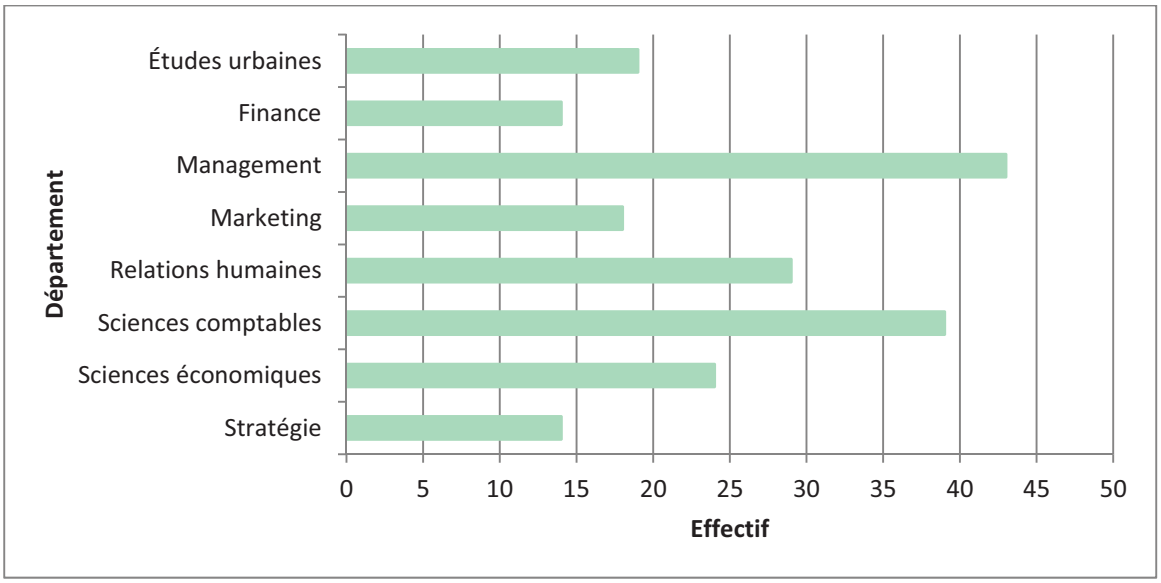
La figure 1.1 présente un graphique de cette distribution, appelé **diagramme à barres**. Les barres sont placées en ordre croissant de haut en bas, mais d'autres arrangements sont possibles.

Figure 1.1 Présentation graphique de la distribution de la variable « Département »
Données du tableau 1.1



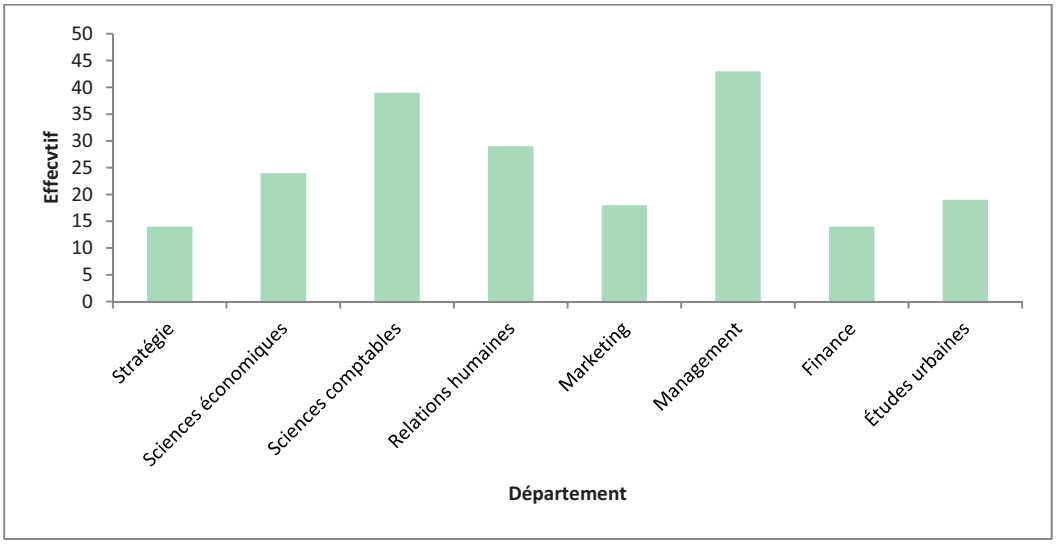
Parfois les barres sont présentées dans l'ordre alphabétique des noms de la variable « département » (Figure 1.2).

Figure 1.2 Données du tableau 1.1 - Ordre alphabétique



Les barres auraient aussi pu être verticales, bien que lorsqu'il s'agit d'une variable qualitative comme le « département », la présentation horizontale est privilégiée, normalement pour des raisons essentiellement de commodité : les « valeurs » d'une variable qualitative étant des mots, parfois des phrases, elles s'inscrivent mieux dans une marge à gauche que le long d'un axe horizontal (Figure 1.3).

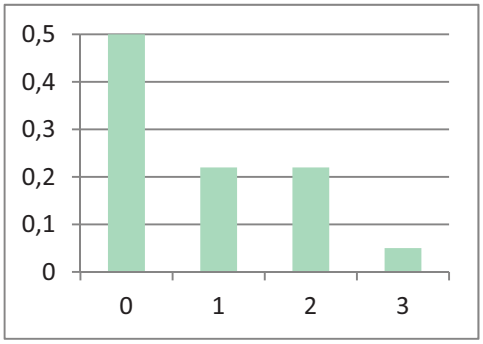
Figure 1.3 Données du tableau 1.1 - Présentation en barres verticales



Pour des données quantitatives discrètes, une distribution peut être présentée graphiquement par un **diagramme à bâtons**, comme à la figure 1.4.

Figure 1.4 Distribution du nombre d'enfants dans les familles de trois enfants ou moins

Nombre d'enfants	Fréquence
0	0,4830
1	0,2130
2	0,2161
3	0,0879*
Total	1,00



Source : Institut de la statistique du Québec, *Le Québec chiffres en mains*, Édition 2010
* Fréquence estimée

Lorsque les valeurs d'une variable sont nombreuses, un graphique qui présente l'effectif de chacune des valeurs est souvent trop chargé pour être lisible, comme on le voit dans la figure 1.5 a) l'allure générale de la distribution est embrouillée par les détails. La figure 1.5 b) présente la distribution déterminée au tableau 1.2, dans lequel les valeurs sont groupées en classes de 5. L'effectif d'une classe est le nombre d'unités dont la valeur est incluse dans la classe.

Tableau 1.2 Distribution de l'âge de la population du Canada en 2011

Âge	Fréquence	Âge	Fréquence	Âge	Fréquence	Âge	Fréquence
]0 ; 5]	0,0561	]25 ; 30]	0,0648	]50 ; 55]	0,0794	]75 ; 80]	0,0276
]5 ; 10]	0,0541	]30 ; 35]	0,0646	]55 ; 60]	0,0699	]80 ; 85]	0,0210
]10 ; 15]	0,0574	]35 ; 40]	0,0650	]60 ; 65]	0,0613	]85 ; 90]	0,0128
]15 ; 20]	0,0651	]40 ; 45]	0,0695	]65 ; 70]	0,0455	]90 ; 95]	0,0051
]20 ; 25]	0,0654	]45 ; 50]	0,0799	]70 ; 75]	0,0344	]95 ; 100]	0,0013