

Chapter 1

Statistique descriptive

1.1 Introduction

La statistique vise à décrire, à résumer et à interpréter des phénomènes dont le caractère essentiel est la variabilité. Elle peut être décomposée en deux phases au moins : le recueil ou la collecte des données statistiques, et leur analyse ou leur interprétation.

a. Le recueil des données : peut être réalisé soit par la simple observation des phénomènes, soit par l'expérimentation.

Exemple : le rôle de quelques substances (N, P, K) dans la production de biomasse chez les végétaux.

Lorsque les données sont très nombreuses, ou particulièrement difficiles à obtenir, il sera nécessaire pour la mise en œuvre rationnelle du recueil de définir des méthodes d'échantillonnage.

Exemple : la numération des mammifères d'une aire protégée : inventaire et recensement.

b. L'analyse et l'interprétation des données : l'analyse statistique se subdivise en deux étapes :

- *La statistique déductive ou descriptive :* elle a pour but de résumer et de présenter les données observées sous la forme la plus accessible (simplification et réduction des données, à la fois visuelle et conceptuelle).
- *L'analyse inductive ou inférence statistique :* est l'ensemble des méthodes permettant de formuler en termes probabilistes un jugement sur une population, à partir des résultats observés sur un échantillon extrait au hasard de cette population. Les méthodes statistiques les plus classiques sont celles de l'estimation (estimation par domaine de confiance) et celles de l'épreuve d'hypothèse.

1.2 Vocabulaire

1.2.1 Population :

L'ensemble sur lequel porte notre étude statistique. Cet ensemble est noté Ω .

Exemple : On considère l'ensemble des étudiants de la spécialité biologie cellulaire.
 $\Rightarrow \Omega$: les étudiants de la spécialité biologie cellulaire.

1.2.2 Unités statistiques (individus) :

Les éléments d'une populations, unités d'observation ou de mesure pour lesquelles des données sont recueillies ou dérivées.

Exemple : On a compté le nombre de loges capsulaires sur 1905 coquelicots.

$\Rightarrow \Omega$: L'ensemble des individus constituant l'espèce "coquelicots".

Les unités statistiques : c'est l'ensemble des 1905 valeurs observées de la variable "nombre de loges capsulaires" : 16, 3, 17, 2, ..., 13, 9, 11.

1.2.3 Échantillon :

On appelle échantillon un sous-ensemble de la population Ω . Un échantillon est dit représentatif si chaque élément de la population a une probabilité connue et non nulle d'en faire partie.

Exemple 1 : Pour Ω = "Ensemble des étudiants de maîtrise de biologie cellulaire".

- un échantillon peut être obtenu en prenant les 30 premiers étudiants par ordre alphabétique. Cet échantillon est non représentatif car tous les autres étudiants ont alors une probabilité nulle d'en faire partie.
- On se propose de tirer au sort 30 étudiants (sur les 90). Cet échantillon 1 est représentatif car tous les étudiants ont une probabilité non nulle (égale à 90) d'en faire partie.

Exemple 2 : on a pesé 100 poissons provenant d'un lac.

- Ω : L'ensemble des poissons du lac.
- L'échantillon : c'est l'ensemble des 100 poissons.
- Les unités statistiques : c'est l'ensemble des 100 valeurs observées de la variable "poids du poisson" 180, 200, 150, 180, ..., 320, 195.

1.2.4 Variable (caractère) :

L'ensemble des unités statistiques de la population sont caractérisées par une ou plusieurs variables (caractères).

Exemple : Pour Ω = "Ensemble des étudiants de maîtrise de biologie cellulaire", on peut considérer les variables suivantes : nom, prénom, genre, note contrôle continu, nombre d'absence.

1.2.5 Modalités :

Les modalités d'une variable statistique sont les différentes valeurs que peut prendre celle-ci. On distingue quatre types de variables selon leurs valeurs possibles :

- **Les variables qualitatives nominales** sont celles dont les valeurs sont des attributs sans ordre naturel (ex. type de sol, couleur de cheveux).
- **Les variables qualitatives ordinales** sont celles dont les valeurs sont des attributs avec un ordre naturel (ex. activité faible, moyenne, forte).

- **Les variables quantitatives discrètes** sont celles dont les valeurs sont le résultat d'un dénombrement (ex. nombre d'enfants dans une famille, nombre de bactéries dans une colonie).
- **Les variables quantitatives continues** sont celles dont les valeurs sont des mesures (ex. taille, poids, distance)

Remarque :

1. Une variable qualitative nominale est dite dichotomique si elle ne possède que deux modalités (ex. genre : féminin ou masculin, présence : présent ou absent). Pour cette variable on peut faire un codage (ex : présent : 0 "signifie absent" ou 1 "signifie" présent)
2. Le type d'une variable dépend de sa nature, mais aussi de la manière dont elle a été mesurée et/ou codée et/ou du choix de la personne qui mène l'étude.

1.2.6 Distributions statistiques

On appelle distribution statistique empirique d'un n-échantillon la donnée d'un regroupement $(c_1, n_1), (c_2, n_2), \dots, (c_k, n_k)$ construit de la manière suivante :

1. Pour les variables discrètes, c_j est la j^{ime} valeur possible de la variable et n_j le nombre de fois où cette valeur a été observée (c'est l'effectif). Les valeurs c_j sont également appelées modalités de la variable aléatoire.
2. Pour les variables continues, on construit une partition de l'ensemble des valeurs en k intervalles : $c_1 = [a_0, a_1[, c_2 = [a_1, a_2[, \dots, c_k = [a_{k-1}, a_k[$, et on compte le nombre de valeurs observées dans chaque intervalle. Les valeurs c_j sont appelées classes.

Fréquence empirique : pour une distribution statistique empirique donnée, on appelle fréquence empirique de la modalité ou de la classe c_j le nombre

$$f_j = \frac{n_j}{n}$$

où $n = \sum_{j=1}^k n_j$

Remarques : Dans le cas d'une variable continue, la distribution statistique empirique dépend du choix des classes. Il n'existe pas de choix optimal du nombre de classes k . Il est conseillé de manière à obtenir une distribution régulière qui représente au mieux la distribution des données, de choisir des classes de même amplitude et d'utiliser pour leur nombre la règle de Sturges :

$$k = 1 + \frac{10 \ln n}{3 \ln 10}$$

1.3 Paramètres de position et de dispersion

1.3.1 Paramètres de position d'un échantillon

Ils sont utilisés pour résumer la position centrale d'une distribution à l'aide d'un indice.

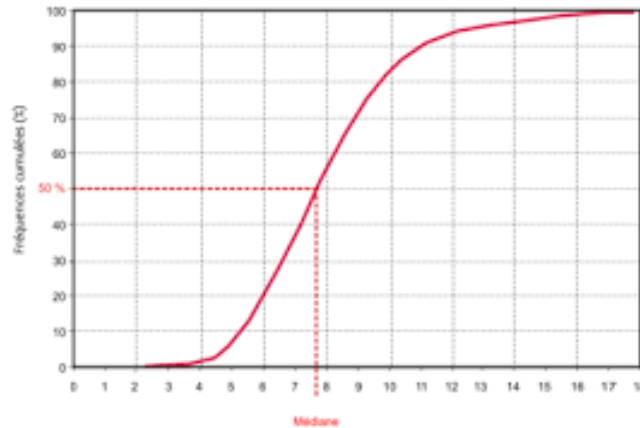


FIGURE 1.1: Détermination de la valeur médiane graphiquement

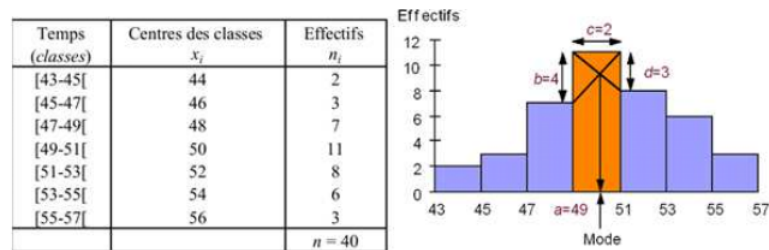


FIGURE 1.2: Classe modale et détermination du mode graphiquement

a. La moyenne empirique : est le paramètre de position noté $\bar{x}$ et défini par :

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

pour un échantillon, et par :

$$\bar{x} = \frac{1}{k} \sum_{j=1}^n n_i \times c_j$$

b. La médiane empirique $me(x)$: est l'indice de position dont la valeur est telle que 50% des observations lui sont supérieures. Si l'on classe les observations par ordre croissant des x_i alors :

- Si n est impair $me(x) = x_{(\frac{n+1}{2})}$
- Si n est pair $me(x) = \frac{x_{\frac{n}{2}} + x_{\frac{n}{2}+1}}{2}$

c. Le mode ou classe modale : est le paramètre de position $mo(x)$ dont la valeur est celle de la modalité ou la classe la plus fréquente.

Remarques : Moyennes et médianes ne s'appliquent qu'aux variables quantitatives ou ordonnées (médiane). Pour une variable qualitative il faut attribuer une valeur numérique (un codage) à chacune de ses modalités. La valeur obtenue pour la moyenne ou la médiane dépend alors du codage choisi. Lorsque ce choix est arbitraire (ex : catégories socio-professionnelles, code postal), la valeur de moyenne ou de médiane n'a aucune interprétation.

1.3.2 Paramètres de dispersion d'un échantillon

Ils ont utilisés pour résumer la dispersion des observations d'une variable aléatoire.

a. La variance : est le paramètre de dispersion, noté $s^2(x)$ qui vaut:

$$s^2(x) = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{1}{n-1} \sum_{j=1}^k c_j (c_i - \bar{x})^2 = \sum_{j=1}^k \frac{n_j}{n-1} (c_i - \bar{x})^2.$$

On appelle écart-type empirique $s(x)$.

Remarque : l'unité dans laquelle est exprimée la variance est celle de la variable au carré, c'est pourquoi on lui préfère l'écart-type qui a la même unité que la variable concernée.

b. L'étendue : est l'indice $e(x) = \max(x_i) - \min(x_i)$

c. Le coefficient de variation : est l'indice $cv(x) = \frac{s(x)}{\bar{x}} \times 100$

Remarque :

Le coefficient de variation permet d'exprimer la variabilité en terme relatif alors que la variance (ou l'écart-type) l'exprime en terme absolue. Il ne dépend pas de l'unité choisie et permet donc de comparer les variabilités de deux variables qui ne sont pas exprimées dans la même unité.

Exemple : On a mesuré la taille et le poids de 10 poissons. Les deux n-échantillons sont :

$x = 23, 20, 17, 15, 30, 25, 24, 27, 22, 19$ pour la taille en centimètres.

$y = 250, 220, 150, 180, 350, 250, 200, 240, 200, 100$ pour le poids exprimé en gramme.

On a $\bar{x} = 22.2$ cm, $s(x) = 4.59$ cm, $cv(x) = 20.6\%$ $\bar{y} = 214$ cm, $s(x) = 67.03$ cm, $cv(x) = 31.2\%$.

1.4 Représentations graphiques

On considère la distribution empirique d'un échantillon : $(c_1, n_1), (c_2, n_2) \dots (c_k, n_k)$, et sa distribution des fréquences empiriques $\frac{n_1}{n}, \frac{n_2}{n} \dots \frac{n_k}{n}$

Cas discret : on appelle graphique des fréquences le graphe où on reporte sur l'axe des abscisses les modalités c_j et où on trace au dessus de chacune d'elles un segment vertical de hauteur proportionnelle à la fréquence empirique.

Cas continu : on appelle histogramme des fréquences le graphe où on reporte sur l'axe des abscisse les classes et au dessus de chacune d'elles un rectangle dont l'aire est égale ou proportionnelle à la fréquence associée. $c_j =]a_{j-1}, a_j]$

Remarque 1 :

- L'unité de mesure d'un histogramme est une aire. Lorsque toutes les classes ont même largeur ; alors seulement l'aire est proportionnelle à la hauteur et il est possible de définir une unité sur l'axe des ordonnées.

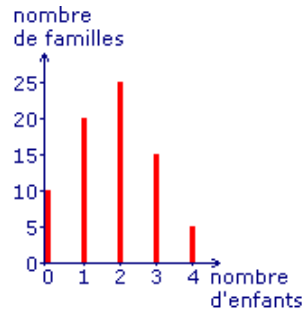


FIGURE 1.3: Représentation graphique d'une variable discrète par un diagramme en bâtons

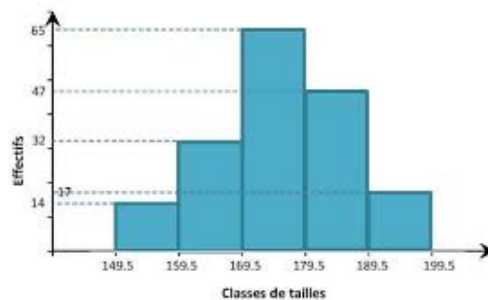


FIGURE 1.4: Représentation graphique d'une variable continue par l'histogramme

- Lien entre fréquence empirique et fréquence théorique (ou probabilité). Pour un n -échantillon, on mesure la probabilité de l'intervalle $[a, b]$ par la surface du rectangle : $P_X[a, b] \simeq h \times \delta_x$.

Si on peut observer un plus grand nombre d'individus, on peut exprimer cette probabilité sous la forme : $P_X[a, b] \simeq h_1\delta_{x_1} + h_2\delta_{x_2} + \dots = \sum_i h_i\delta_{x_i} = \delta F = F_{i+1} - F_i$.

Quand on fait tendre δx vers 0, l'équation $\delta F_i = h_i\delta_{x_i}$ devient $dF = h dx$ et $\frac{dF}{dx} = f(x)$ est la densité de probabilité de X .

1.5 Exercices

Exercice 1 :

Parmi ces assertions, préciser celles qui sont vraies, celles qui sont fausses.

1. On appelle variable, une caractéristique que l'on étudie.
2. La tâche de la statistique descriptive est de recueillir des données.
3. La tâche de la statistique descriptive est de présenter les données sous forme de tableaux, de graphiques et d'indicateurs statistiques.
4. En Statistique, on classe les variables selon différents types.
5. Les valeurs des variables sont aussi appelées modalités.
6. Pour une variable qualitative, chaque individu statistique ne peut avoir qu'une seule modalité.