Table des matières

Chapitre 1

Statistiques descriptives

Contenu

La Statistique descriptive est l'ensemble de méthodes scientifiques permettant de collecter, décrire et analyser des données observées.

1.1 Vocabulaire statistique

❶ **Population** : est l'ensemble des individus ou d'objets de même nature sur lequel porte l'étude.

❷ **Individus** : Individus ou unités statistiques sont les éléments de la population.

❸ **Echantillon** : est un sous ensemble de la population.

❹ **Variable statistique** : le caractère est la propriété que l'on se propose d'observer dans la population ou l'échantillon. Un caractère qui fait le sujet d'une étude porte aussi le nom de variable statistique X.

❺ **Modalité statistique** : On appelle une modalité (ou catégorie) les différentes situations (niveaux) possibles d'une variable statistique.
On distingue deux types de variables statistiques

Variables quantitatives

Sont les variables qu'on peut mesurer, elles sont caractérisées par des valeurs numérique. Variables dont les modalités sont des nombres.
Une variable statistique quantitative peut être :
Continue : lorsqu'elle peut prendre des nombres issus d'un intervalle de nombres réels (résultats de mesures).
Discrète : si elle prend des valeurs isolées.
Temporelle : Ce sont des variables quantitatives particulières qui utilisent les unités de mesure du temps. Il existe deux types, le type date (date de naissance : 26/04/1994) et le type horaire (heures d'étude : 6h).

Example 1.1.

variable	*modalités possibles*	*type de variable*
la taille	*1.70m, 1.60m, 1.65m, 1.75m*	*quantitative continue*
le nombre des étudiants	*30, 50, 60, 80*	*quantitative discrète*

Variables qualitatives

Ce sont des variables qui ne sont pas mesurables (n'ont pas de valeurs numériques). Variables dont les modalités sont des mots.

Les variables statistiques qualitatives peuvent être :

Ordinales : ce sont des variables dont les modalités s'ordonnent selon leur sens.

Nominales : ce sont des variables dont les modalités ne peuvent être ordonnées selon leur sens.

Example 1.2.

variable	*modalités possibles*	*type de variable*
couleur des yeux	*noir, bleu, vert, marron*	*qualitative nominale*
degré de satisfaction face à son niveau de vie	*très satisfait, satisfait, insatisfait*	*qualitative ordinale*

❻ **Série statistique** : La forme la plus simple de présentation des données statistiques relatives à un seul caractère ou variable, consiste à une simple énumération des valeurs prises par le caractère.

❼ **Effectif total** : On appelle effectif total n le nombre total d'individus dans la population.

❽ **Effectif** : l'effectif ou fréquence absolue noté n_i est le nombre des éléments statistiques relatifs à une modalité donnée.

❾ **Effectif cumulé croissant** : On appelle effectif cumulé croissant noté $n_i^c \uparrow$ le nombre d'individus qui correspondent au même caractère (modalité) et aux caractères précédents.

❿ **Effectif cumulé déroissant** : On appelle effectif cumulé décroissant noté $n_i^c \downarrow$ le nombre d'individus qui correspondent au même caractère (modalité) et aux caractères suivants.

① **Fréquence** : on appelle fréquence ou fréquence relative noté f_i, le rapport entre l'effectif d'une valeur et l'effectif total $\frac{n_i}{n}$.

② **Fréquence cumulée croissante** : on appelle fréquence cumulée croissante noté $f_i^c \uparrow$, le rapport entre l'effectif cumulé croissant d'une valeur et l'effectif total $\frac{n_i^c \uparrow}{n}$.

③ **Fréquence cumulée décroissante** : on appelle fréquence cumulée décroissante noté $f_i^c \downarrow$, le rapport entre l'effectif cumulé déroissant d'une valeur et

l'effectif total $\frac{n_i^c\downarrow}{n}$.

Example 1.3. *les notes de 9 étudiants d'un groupe*

Note	n_i	$n_i^c \uparrow$	$n_i^c \downarrow$	*Fréquence* f_i	$f_i^c \uparrow$	$f_i^c \downarrow$
5	*2*	*2*	*9*	*2/9*	*2/9*	*1*
6	*1*	*3*	*7*	*1/9*	*1/3*	*7/9*
8	*3*	*6*	*6*	*1/3*	*2/3*	*6/9*
12	*2*	*8*	*3*	*2/9*	*8/9*	*3/9*
16	*1*	*9*	*1*	*1/9*	*1*	*1/9*
Total	$n = 9$			$\sum_{i=1}^{5} f_i = 1$		

④ **Classe (Intervalle)** : On appelle classe un groupement de valeurs d'une variable selon des intervalles qui peuvent être égaux ou inégaux. On l'utilise surtout lorsque la variable étudiée est quantitative continue.

Pour chaque classe on peut définir :

- Une limite inférieure

- Une limite supérieure

- Intervalle de classe (amplitude)= limite (sup)- limite (inf)

- Centre de classe $c_i = \frac{limite\,(sup) + limite\,(inf)}{2}$.

Example 1.4. *: Le taux de glucose sanguin (glycémie) chez 14 sujets en g/l*

classe	*Centre de classe* c_i	n_i	$n_i^c \uparrow$	$n_i^c \downarrow$	*Fréquence* f_i	$f_i^c \uparrow$	$f_i^c \downarrow$
[0,85 ; 0,91[	*0,88*	*3*	*3*	*14*	*3/14*	*3/14*	*1*
[0,91 ; 0,97[	*0,94*	*5*	*8*	*11*	*5/14*	*4/7*	*11/14*
[0,97 ; 1,03[	*1*	*3*	*11*	*6*	*3/14*	*11/14*	*6/14*
[1,03 ; 1,09[	*1,06*	*2*	*13*	*3*	*1/7*	*13/14*	*3/14*
[1,09 ; 1,15[	*1,12*	*1*	*14*	*1*	*1/14*	*1*	*1/14*
Total		*n=14*			$\sum_{i=1}^{5} f_i = 1$		

1.2 Description des données

Selon le type de la variable étudiée. Il existe deux formes de présentation pour décrire une série de données statistiques sont : les tableaux et les représentations graphiques.

1.2.1 Tableaux

Le tableau est utilisable quelle que soit la nature des données, il sert à présenter les données d'une façon exacte et complète .

1.2.2 Graphiques

L'objectif des graphiques est de faire ressortir une vision systhématique du phénomène étudié en illustrant une tendance générale et en donnant une image globale des résultats.

Histogramme

Les histogrammes sont des surfaces qui permettent la représentation d'une variable quantitative continue. L'aire de chaque surface est égale à l'effectif cor-

respondant à une classe.

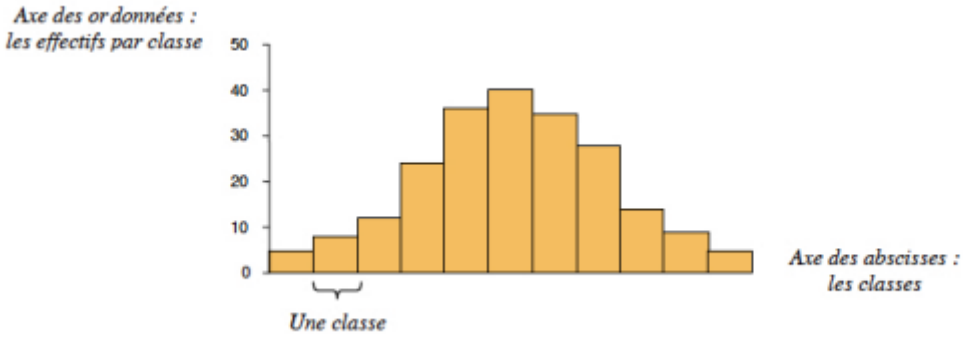


Diagramme en bâtons

Un diagramme en bâtons est une représentation graphique de données statistiques à l'aide de segments.

Example 1.5.

Diamètre	*12*	*13*	*14*	*15*	*16*	*17*	*18*
Effectif	*2*	*5*	*3*	*4*	*6*	*5*	*3*

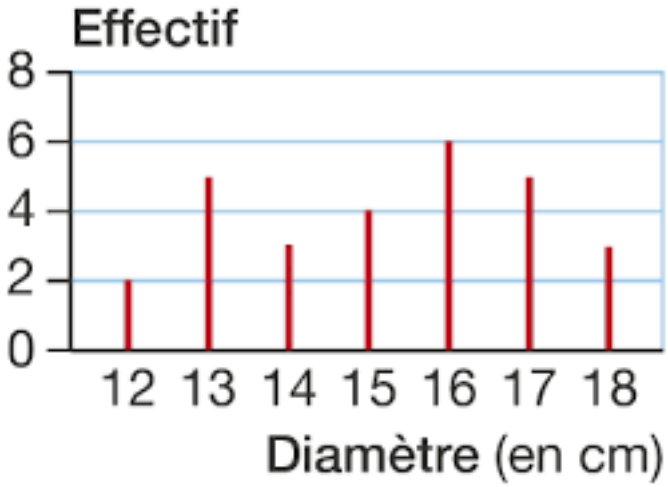


Diagramme en barres

Un diagramme en barres est une représentation graphique réservée surtout pour la distribution d'une variable qualitative à l'aide de rectangles de même largeur.

Example 1.6.

Situation familiale	*Célibataire*	*Divorcé(e)*	*Marié(e)*	*veuf(ve)*
Effectif	*9*	*2*	*7*	*2*

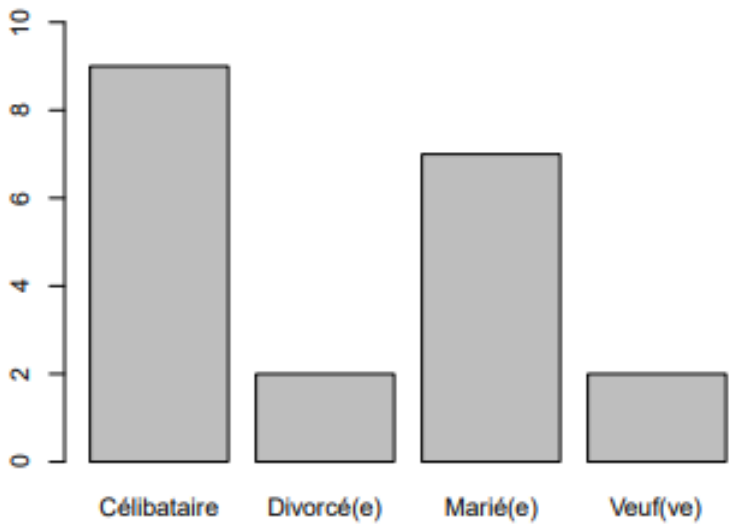


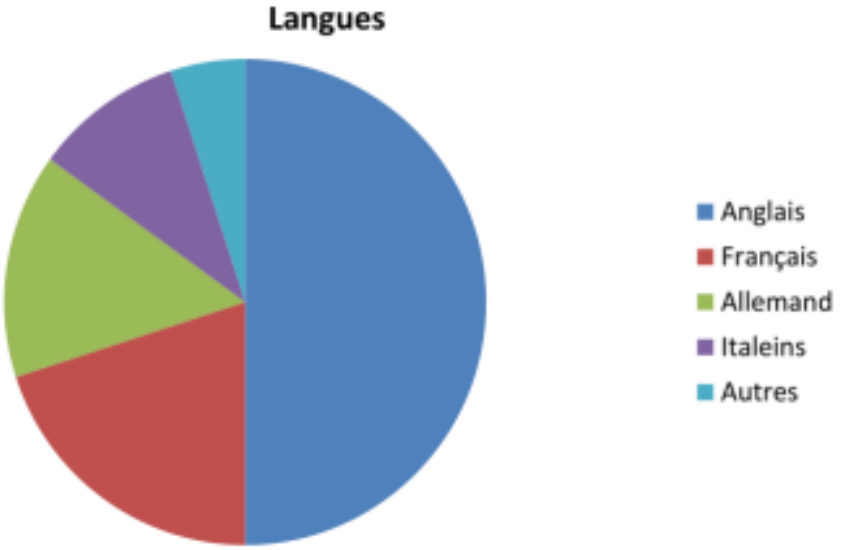
Diagramme circulaire (Camembert)

On dessine sur un disque des sections correspondants au modalités du caractère dont les angles sont proportionnels au pourcentages.

$$\alpha_i = 360^0 * f_i = 360^0 * \frac{n_i}{n}$$

Example 1.7.

Langue	*Nombre d'étudiants*	f_i	α_i
Anglais	*500*	*0.5*	180^0
Français	*200*	*0.2*	72^0
Allemand	*150*	*0.15*	54^0
Italien	*100*	*0.1*	36^0
Autre	*50*	*0.05*	18^0



1.3 Paramètres de position

Paramètres de tendance centrale ou de position : valeurs situées au centre de la distribution statistique qui sont la moyenne, le mode et la médiane.

1.3.1 Moyenne arithmétique

Cas d'une variable statistique discrète

Soient X une variable statistique discrète et $x_1, x_2, ..., x_k$ ses valeurs pour lesquelles correspondent les effectifs $n_1, n_2, ..., n_k$, avec $n = \sum_{i=1}^{k} n_i$ l'effectif total.
On appelle moyenne de X la quantité

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{k} n_i x_i = \sum_{i=1}^{k} f_i x_i.$$

Example 1.8.

x_i	0	1	2	3	4
n_i	2	3	1	1	1

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{5} n_i x_i = \frac{1}{8}\Big(0\times 2 + 1\times 3 + 2\times 1 + 3\times 1 + 4\times 1\Big) = \frac{12}{8} = 1.5.$$

Cas d'une variable statistique continue

Les observations sont groupées dans des classes, alors

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{k} n_i c_i = \sum_{i=1}^{k} f_i c_i.$$

Example 1.9.

classe	c_i	n_i
[1,2[	1.5	3
[2,3[	2.5	1
[3,4[	3.5	2

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{3} n_i c_i = \frac{1}{6}\Big(3\times 1.5 + 1\times 2.5 + 2\times 3.5\Big) = \frac{14}{6} = 2.33.$$

1.3.2 Mode

Cas d'une variable statistique discrète

Le mode *Mo* est la valeur x_i ayant le plus grand effectif.

Example 1.10.

x_i	2	3	5	6	7	8	9	10
n_i	2	1	1	2	2	1	1	1

On a trois modes : $Mo = 2, 6, 7$

Cas d'une variable statistique continue

Dans ce cas le mode se calcule par la formule

$$Mo = L_i + \left(\frac{d_1}{d_1 + d_2}\right) a$$

• L_i : la borne inférieure de la classe modale (classe correspondant au plus grand effectif)

• d_1 = l'effectif de la classe modale- l'effectif de la classe précédente ($n_i - n_{i-1}$).

• d_2 = l'effectif de la classe modale- l'effectif de la classe suivante ($n_i - n_{i+1}$).

• a : l'amplitude de la classe modale.

Example 1.11.

classe	n_i
[1,60-1,65[	3
[1,65-1,70[	8
[1,70-1,75[	2

• *La classe modale est :* $[1,65 - 1,70[$.

• $L_i = 1,65$.

- $d_1 = 8 - 3 = 5.$
- $d_2 = 8 - 2 = 6.$
- $a = 1,70 - 1,65 = 0.05 \quad donc \quad Mo = 1,65 + \left(\frac{5}{5+6}\right) 0.05 = 1,67$

1.3.3 Médiane

Cas d'une variable statistique discrète

La médiane *Me* est la valeur qui se trouve au centre d'une série de nombres rangés par ordre croissant.

- Si n est paire, alors

$$Me = \frac{x_{\frac{n}{2}} + x_{\frac{n}{2}+1}}{2}$$

- Si n est impaire, alors

$$Me = x_{\frac{n+1}{2}}$$

Example 1.12. *Le nombre d'enfants de 6 familles est le suivant*

$$7,\ 3,\ 1,\ 1,\ 5,\ 2$$

On ordonnee d'abord les valeurs :

$$\underbrace{1,\ 1,\ 2}_{3},\ \underbrace{3,\ 5,\ 7}_{3}$$

On a $n = 6$ *paire donc* $Me = \frac{x_3 + x_4}{2} = \frac{2+3}{2} = 2.5.$

Example 1.13. *Le nombre d'enfants de 7 familles est le suivant*

$$3,\ 2,\ 1,\ 0,\ 0,\ 1,\ 2$$

On ordonnee d'abord les valeurs :

$$\underbrace{0,\ 0,\ 1,}_{3}\ \underbrace{1}_{Me=x_4=1}\ ,\ \underbrace{2,\ 2,\ 3}_{3}$$

On a $n = 7$ impaire donc $Me = x_4 = 1$.

Cas d'une variable statistique continue

Dans ce cas la médiane est donnée par

$$Me = L_i + \left(\frac{\frac{n}{2} - \sum_{i=1}^{<Me} n_i}{n_{Me}}\right) a$$

- L_i : la borne inférieure de la classe médiane (classe qui divise l'effectif en deux)
- $\sum_{i=1}^{<Me} n_i$ = la somme des effectifs correspondant à toutes les classes inférieures à la classe médiane.
- n_{Me} = l'effectif de la classe médiane.
- a : l'amplitude de la classe médiane.

Example 1.14. *D'après l'exemple (1.4), on obtient*

- *La classe médiane est :* $[0.91 - 0.97[$.
- $L_i = 0.91$.
- $n = 14$.
- $\sum_{i=1}^{<Me} n_i = 3$
- n_{Me}= *5* .
- $a = 0.97 - 0.91 = 0.06$

donc $Me = 0.91 + \left(\frac{7-3}{5}\right) 0.06 = 0.958$

1.3.4 Les quartiles

Cas d'une variable statistique discrète

Les quartiles sont les trois valeurs qui partagent la distribution en quatre parties égales. On les appelle respectivement :

• **Le premier quartile Q_1** représente 25% de l'échantillon c'est à dire Q_1 est la valeur x_i dont la position est le plus petit entier qui suit $\frac{n}{4}$.

• **Le deuxième quartile Q_2** représente 50% de l'échantillon

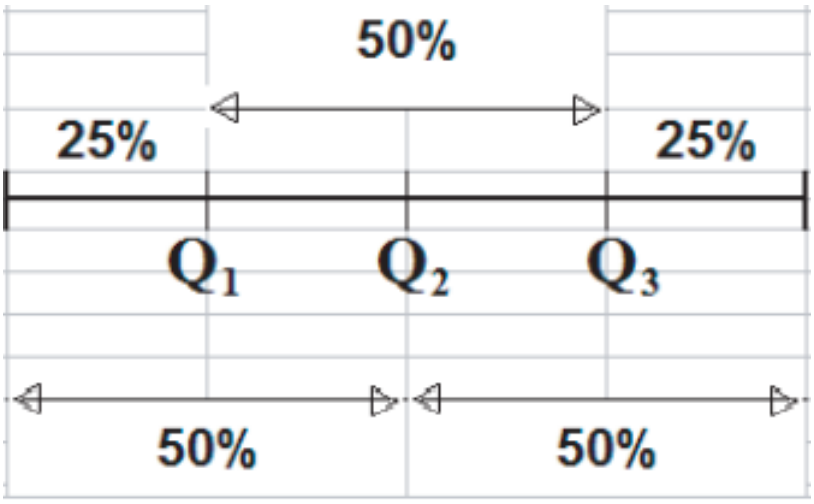
• **Le troisième quartile Q_3** représente 75% de l'échantillon c'est à dire Q_3 est la valeur x_i dont la position est le plus petit entier qui suit $\frac{3n}{4}$.

L'écart interquartile

L'écart interquartile est la différence entre le dernier et le premier quartile :

$$I_Q = Q_3 - Q_1$$

.



Example 1.15. *Dans l'exemple des observations suivantes*

x_i	1	3	5	7	9
n_i	1	2	1	2	2
n_i^c	1	3	4	6	8

• *On a* $n = 8$ *et* $\frac{n}{4} = 2$ *donc* Q_1 *est la deuxième valeur* $Q_1 = x_2 = 3$.

• *On a* $n = 8$ *et* $\frac{3n}{4} = 6$ *donc* Q_3 *est la sixième valeur* $Q_3 = x_6 = 7$.

1.4 Paramètres de dispersion

Les paramètres de dispersion sont les paramètres qui résument la dispersion des valeurs autour de la valeur centrale

1.4.1 Etendue

On appelle étendue e, la différence entre la plus grande valeur et la plus petite valeur observée.

$$e = x_{max} - x_{min}$$

Example 1.16. *Les notes de 10 étudiants sont les suivantes*

$$2,\ 3,\ 10,\ 10,\ 11,\ 12,\ 15,\ 18,\ 19,\ 20$$

donc

$$e = x_{max} - x_{min} = 20 - 2 = 18$$

1.4.2 Variance

On appelle une variance la moyenne arithmétique des carrés des écarts entre les valeurs d'une variable et la moyenne arithmétique.

$$\begin{aligned} V(X) &= \frac{1}{n}\sum_{i=1}^{k} n_i(x_i-\bar{x})^2 = \frac{1}{n}\sum_{i=1}^{k} n_i x_i^2 - \bar{x}^2 \\ &= \sum_{i=1}^{k} f_i(x_i-\bar{x})^2. = \sum_{i=1}^{k} f_i x_i^2 - \bar{x}^2. \end{aligned}$$

1.4.3 Ecart-type

On appelle écart-type noté σ_X (ou écart quadratique moyen) la racine carrée de la variance.

$$\sigma_X = \sqrt{V(X)}$$

1.4.4 Le coefficient de variation

Le coefficient de variation noté CV se définit par

$$CV = \frac{\sigma_X}{\bar{x}}$$

Example 1.17.

x_i	0	1	2	3	4
n_i	2	3	1	1	1

$\bar{x} = 1.5$

$$\begin{aligned} V(X) &= \frac{1}{n}\sum_{i=1}^{k} n_i x_i^2 - \bar{x}^2 \\ &= \frac{1}{8}\sum_{i=1}^{5} n_i x_i^2 - (1.5)^2 \\ &= \frac{1}{8}\left(2\times 0^2 + 3\times 1^2 + 1\times 2^2 + 1\times 3^2 + 1\times 4^2\right) - 2.25 \\ &= \frac{32}{8} - 2.25 \\ &= 1.75 \end{aligned}$$

l'écart-type

$$\sigma_X = \sqrt{V(X)} = \sqrt{1.75} = 1.3$$

et le coefficient de variation

$$CV = \frac{\sigma_X}{\bar{x}} = \frac{1.3}{1.5} = 0.87$$

1.5 Paramètres de forme

1.5.1 Asymétrie

Il existe plusieurs coefficients d'asymétrie, les principaux sont les suivants :

• Coefficient d'asymétrie de Pearson :

$$A_P = \frac{\bar{x} - Mo}{\sigma_X}.$$

• Coefficient d'asymétrie de Yule :

$$A_Y = \frac{Q_1 + Q_3 - 2Q_2}{Q_3 - Q_1}.$$

Remarque

• Un coefficient positif indique que la distribution est plus étalée à droite.

• Un coefficient négatif indique que la distribution est plus étalée à gauche.

• Un coefficient nul indique que la distribution est symétrique.

1.5.2 Aplatissement

L'aplatissement est mesuré par :

• Le coefficient d' aplatissement de Pearson :

$$AP_P = \frac{m_4}{\sigma_X^4}$$

où m_4 est le moment centré d'ordre 4 définie par

$$m_4 = \frac{1}{n}\sum_{i=1}^{k} n_i(x_i - \bar{x})^4$$

• Le coefficient d' aplatissement de Fisher :

$$AP_F = \frac{m_4}{\sigma_X^4} - 3$$

Remarque

• Si $AP_F = 0$ alors la distibution est dite "normale" ou "mésokurtique".

• Si $AP_F < 0$ alors la distibution est dite plus aplatie que la "normale" ou "platy-kurtique".

• Si $AP_F > 0$ alors la distibution est dite moins aplatie que la "normale" ou "lepto-kurtique".