Mathématique Statistique

Dr. HAFIRASSOU ZINEB

Centre Universitaire Abdelhafid Boussouf - Mila

Institut des Sciences et de la Technologie

Département de SNV

Email : z.hafirassou@centre-univ-mila.dz

5.0 Février 2023

Table des matières

Chapitre 02 : Statistiques descriptives

1. Vocabulaire statistique

Rappel

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).

Exemple

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.

Exemple

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 décroissant** : 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 décroissant d'une valeur et l'effectif total $\frac{n_i^c \downarrow}{n}$.

Exemple les notes de 9 étudiants d'un groupe

Note	n_i	$n_i^c \uparrow$	$n_i^c \downarrow$	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	2/3
12	2	8	3	2/9	8/9	1/3
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

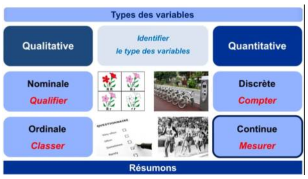
- Une limite supérieure

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

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

Exemple Le taux de glucose sanguin (glycémie) chez 14 sujets en g/l

classe	c_i	n_i	$n_i^c \uparrow$	$n_i^c \downarrow$	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	3/7
[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			$\sum_{i=1}^{5} f_i = 1$		



Résumé

2. Description des données

2.1. Tableaux

Rappel

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

2.2. Graphiques

Rappel

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 ***Définition***

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

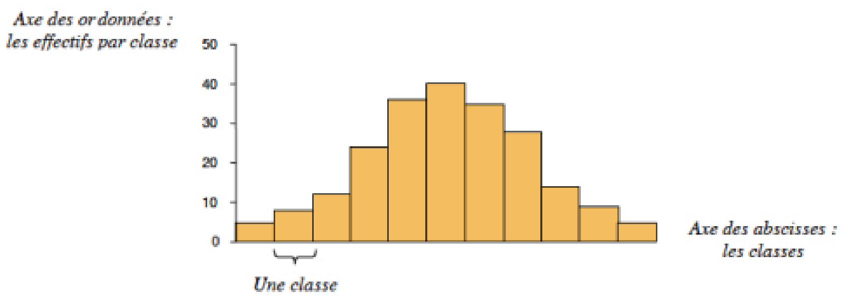


Diagramme en bâtons

Définition

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

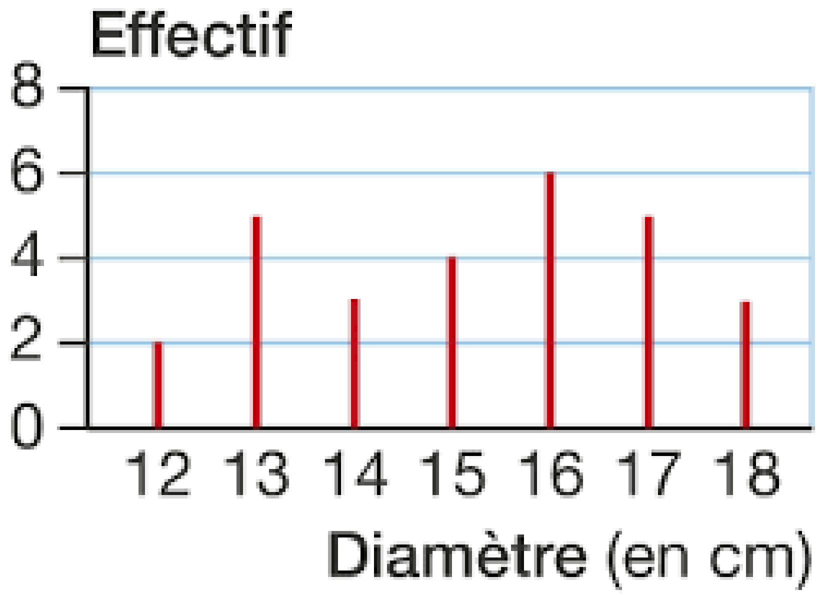


Diagramme en barres

Définition

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.

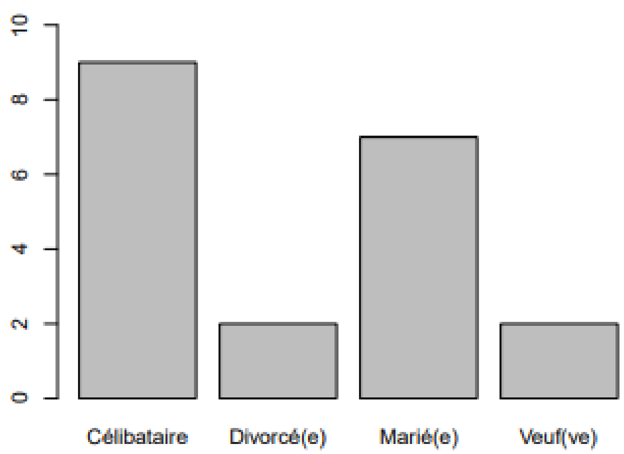


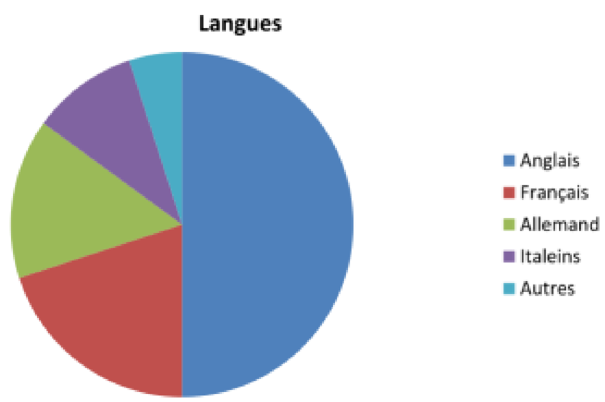
Diagramme circulaire (Camembert) *Définition*

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

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

Exemple

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



3. Paramètres de position

3.1. Moyenne arithmétique

Cas d'une variable statistique discrète *Définition*

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

On appelle moyenne de X la quantité

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

Exemple

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

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{5} n_i x_i$$

$$= \frac{1}{8}(0 \times 2 + 1 \times 3 + 2 \times 1 + 3 \times 1 + 4 \times 1) = \frac{12}{8} = 1.5.$$

Cas d'une variable statistique continue ***Définition***

Les observations sont groupées dans des classes, alors

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

Exemple

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

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{3} n_i c_i$$

$$= \frac{1}{6}(3 \times 1.5 + 1 \times 2.5 + 2 \times 3.5) = \frac{14}{6} = 2.33.$$

3.2. Mode

Cas d'une variable statistique discrète ***Définition***

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

Exemple

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 — ***Définition***

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.

Exemple

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

- La classe modale est : $[1,65 - 1,70[$.
- $L_i = 1.65$.
- $d_1 = 8 - 3 = 5$
- $d_2 = 8 - 2 = 6$
- $a = 1.70 - 1.65 = 0.05$ donc $Mo = 1.65 + \left(\frac{5}{5+6}\right)0.05 = 1.67$

3.3. Médiane

Cas d'une variable statistique discrète — ***Définition***

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

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

Exemple

Le nombre d'enfants de 6 familles est le suivant : 7, 3, 1, 1, 5, 2

On ordonne d'abord les valeurs

On a $n = 6$ paire donc

$$Me = \frac{x_3 + x_4}{2} = \frac{2+3}{2} = 2.5$$

Exemple

Le nombre d'enfants de 7 familles est le suivant : 3, 2, 1, 0, 0, 1, 2

On ordonne d'abord les valeurs

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

Cas d'une variable statistique continue ***Définition***

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.

3.4. Les quartiles

Cas d'une variable statistique discrète ***Définition***

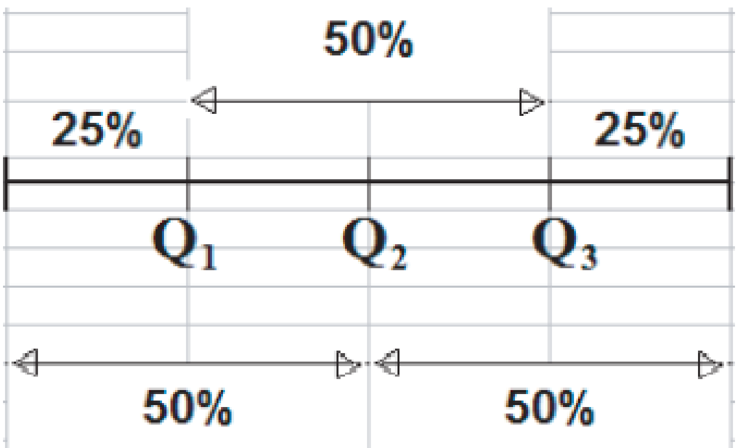
Les quartiles sont les trois valeurs qui partagent la distribution en quatre parties égales. On les appelle respectivement :

- **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 :

$I_Q = Q_3 - Q_1$.



Exemple

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 $Q_3 = x_6 = 7$.

Pour voir la vidio cliquer ici[1]

4. Paramètres de dispersion

4.1. Etendue

Définition

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

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

Exemple

Les notes de 10 étudiants sont les suivantes : 2, 3, 10, 10, 11, 12, 15, 18, 19, 20

donc

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

4.2. Variance

Définition

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] https://www.youtube.com/watch?v=bEf1sPqNBkk&t=119s

4.3. Ecart-type

Définition

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

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

4.4. Le coefficient de variation

Définition

Le coefficient de variation noté CV se définit par $CV = \frac{\sigma_X}{\bar{x}}$

Exemple

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

$\bar{x} = 1.5$

La variance

$V(X) = \frac{1}{n} \sum_{i=1}^{k} n_i x_i^2 - \bar{x}^2$

$= \frac{1}{8} \sum_{i=1}^{5} n_i x_i^2 - (1.5)^2$

$= \frac{1}{8}(2 \times 0^2 + 3 \times 1^2 + 1 \times 2^2 + 1 \times 3^2 + 1 \times 4^2) - 2.25$

$= \frac{2}{8} - 2.25 = 1.75$

L'écart-type

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

et le coefficient de variation

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

5. Paramètres de forme

5.1. Asymétrie

Définition

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.

5.2. Aplatissement

Définition

L'aplatissement est mesuré par :

- Le coefficient d' aplatissement de Pearson :

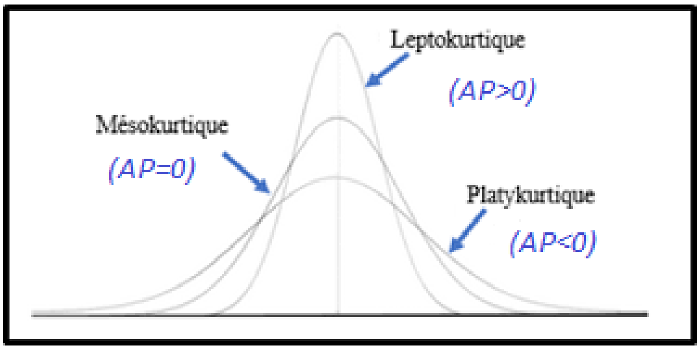
$$AP_P = \frac{m_4}{\sigma_X^4}$$

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 "platykurtique".
- Si $AP_F > 0$ alors la distibution est dite moins aplatie que la "normale" ou "leptokurtique".

6. Test d'acquisitions

6.1. Exercice

[solution n°1 p. 16]

La population est l'ensemble des individus ou d'objets de même nature sur lequel porte l'étude.

- ○ Vrai
- ○ faux

6.2. Exercice

Préciser la nature de la variable statistique:

❶ Lieu de résidence

❷ Sexe

❸ Le groupe sanguin

❹ Nombre de globule blanc

❺ La taille

❻ Âge

❼ Nombre de langues parlées

❽ Le niveau d'obésité

❾ Couleur des yeux

6.3. Exercice

[solution n°2 p. 16]

Les paramètres de dispersion sont :

- ☐ Etendue
- ☐ Le mode
- ☐ La variance
- ☐ La moyenne

6.4. Exercice

[solution n°3 p. 16]

La médiane de la série statistique suivante : 1,2,8,12,18 est :

6.5. Exercice

Le staff médical d'une grande entreprise fait des statistiques sur la pratique du sport par mois de ses employés. Les observations sur 88 employés sont les suivantes

Note	n_i	$n_i^c \uparrow$	$n_i^c \downarrow$	Fréquence f_i	$f_i^c \uparrow$	$f_i^c \downarrow$
8	7					
12	20					
16	23					
20	19					
24	14					
28	5					
Total	n = 88					

❶ Déterminer la population, le caractère étudié et donner sa nature.

❷ Compléter le tableau.

❸ Représenter graphiquement la série statistique.

❹ Calculer le mode, la moyenne et la médiane.

❺ Déterminer les quartiles et l'écart interquartiles.

❻ Calculer l'étendue, la variance, l'écart-type et le coefficient de variation.

❼ Calculer le coefficient d'asymétrie de Pearson, et donner la conclusion nécessaire.

Solutions des exercices

Solution n°1

[exercice p. 14]

La population est l'ensemble des individus ou d'objets de même nature sur lequel porte l'étude.

- (x) Vrai
- ( ) faux

Solution n°2

[exercice p. 14]

Les paramètres de dispersion sont :

- [x] Etendue
- [ ] Le mode
- [x] La variance
- [ ] La moyenne

Solution n°3

[exercice p. 14]

La médiane de la série statistique suivante : 1,2,8,12,18 est :

8