

Chapitre 1

Introduction à la statistique descriptive

La statistique est une science qui consiste à collecter les données, les traiter , les analyser, les interpréter et les présenter en vue d'étudier un phénomène donné.

Les statistiques sont les informations et les données collectées auprès des établissements.

La statistique descriptive est l'ensemble des méthodes permettant de décrire, de résumer , de présenter les données.

1) Vocabulaires de base

Définitions 1.1.

1. **Population** est l'ensemble des éléments ayant un caractère commun à étudier. On la note Ω .
2. **Individu ou Unité statistique** est l'élément de la population. On le note ω .
3. **Taille de la population** est le nombre des éléments de la population. On le note N .
4. **Caractère ou variable statistique** est une caractéristique ou propriété qui décrit un individu d'une population. Elle est observée ou mesurée sur les individus d'une population. On le note X .
5. **Modalité** est la valeur d'une variable statistique. On la note x .
6. **L'effectif** noté n_i d'une modalité x_i est le nombre d'individus ayant cette modalité, ou le nombre de fois que la modalité x_i est répétée.
7. **L'effectif cumulé croissant** noté N_i est le nombre d'individus présentant au plus la modalité x_i .
8. **L'effectif cumulé décroissant** noté N_i^d est le nombre d'individus présentant au moins la modalité x_k .
9. **La fréquence** f_i de la modalité x_i est le nombre d'individus possédant ce caractère divisé par la taille de la population. On écrit : $f_i = \frac{n_i}{N}$.

10. **Le Pourcentage ou fréquence en pourcentage** d'un caractère s'exprime comme suit : $p_i = f_i \times 100$.
11. **Fréquence cumulée croissante** F_i est la proportion d'individus ayant des modalités du caractère étudié inférieures ou égales à x_i .
12. **Fréquence cumulée décroissante** F_i^d est la proportion d'individus ayant des modalités du caractère étudié supérieures ou égales à x_i .
13. **Distribution d'un caractère** X est l'ensemble des couples : $\{(x_1, n_1), (x_2, n_2), \dots, (x_k, n_k)\}$.
On peut écrire aussi la distribution sous la forme suivante : $\{(x_1, f_1), (x_2, f_2), \dots, (x_k, f_k)\}$.

Exemple :

Au cours d'un semestre, on a observé 20 élèves d'une école dont les notes sont données par les tableaux suivants :

élève 1	élève 2	élève 3	élève 4	élève 5	élève 6	élève 7	élève 8	élève 9	élève 10
12	15	12	9	14	16	12	10	6	16

élève 11	élève 12	élève 13	élève 14	élève 15	élève 16	élève 17	élève 18	élève 19	élève 20
15	18	14	12	13	10	9	12	16	11

1. La population ici c'est "les élèves". On peut écrire :

$$\Omega = \{\text{élève 1, élève 2, ..., élève 20}\}.$$

2. L'unité statistique c'est un élève. On peut écrire $\omega_1 = \text{élève 1}$, $\omega_2 = \text{élève 2}$,...
3. La taille de la population c'est $N = 20$, le nombre total des élèves.
4. La variable statistique c'est $X = \text{"la note de l'élève"}$.
On peut écrire $X(\omega_1)$ pour dire la note de l'élève 1, en voyant le tableau on peut voir que $X(\omega_1) = 12$ et $X(\omega_{13}) = 14$...
5. Les modalités sont les notes obtenues : 12 c'est une modalité, 15 c'est une modalité, 25 ce n'est pas une modalité parce que les notes sont comprises entre 0 et 20. Les modalités dans notre exemples sont $\{12, 15, 9, 14, 16, 10, 6, 18, 13, 11\}$.
6. L'effectif de la modalité 12 est 5, c'est le nombre des élèves qui ont la note 12. L'effectif de 14 c'est 2,...
7. L'effectif cumulé croissant de la modalité 12, c'est le nombre des élèves qui ont eu une note au plus égale à 12 ; c'est à dire, 12 ou moins que 12, il existe 11 élèves : $N(12) = 11$.
8. L'effectif cumulé décroissant de la modalité 12, c'est le nombre des élèves qui ont eu une note au moins égale à 12 ; c'est à dire, 12 ou plus que 12, il existe 14 élèves : $N^d(12) = 14$.

9. La fréquence de la modalité 12 c'est $f(12) = \frac{5}{20} = 0.25$.
10. Le pourcentage de la modalité 12 c'est $p(12) = 0.25 \times 100 = 25\%$.
11. La fréquence cumulé croissante de la modalité 12 c'est $F(12) = \frac{11}{20} = 0.55$.
12. La fréquence cumulé décroissante de la modalité 12 c'est $F^d(12) = \frac{14}{20} = 0.70$.
13. La distribution des notes est comme suit :

$$\{(12, 5), (15, 2), (9, 2), (14, 2), (16, 3), (10, 2), (6, 1), (18, 1), (13, 1), (11, 1)\}.$$

Remarque : La somme de tous les effectifs est égale à la taille de la population.

La somme de tous les fréquences est égale à 1.

Définitions 1.2. Il existe deux types des variables statistiques :

1. **Variable qualitative :** Les modalités sont des qualités ; ne sont pas des nombres. Il y a deux types des variables qualitatives :
 - (a) **Variables qualitatives nominales :** quand on ne peut pas faire un ordre croissant/décroissant.
 - (b) **Variables qualitatives ordinales :** quand on peut faire un ordre croissant/décroissant.
2. **Variable quantitative :** Les modalités sont des quantités ; ce sont des nombres. Il y a deux types des variables quantitatives :
 - (a) **Variables quantitatives discrètes :** quand les modalités admettent un nombre fini des possibilités.
 - (b) **Variables quantitatives continues :** quand les modalités admettent des possibilités infinies ; classés par des intervalles.

Exemples :

1. Dans une ville, on veut compter le nombre des voitures pour chaque couleur : les voitures noires, les voitures bleues, les voitures blanches...
la variable ici c'est la couleur des voitures ; les modalités possibles ce sont : noire, bleue, blanche,... ce ne sont pas des nombres, donc c'est une variable qualitative. De plus, on ne peut pas dire que noire $>$ bleue $>$ blanche... donc c'est une variable qualitative nominale.
2. Dans un amphi, on veut compter le nombre des étudiants pour chaque mention : mention passable, assez bien, bien ...
la variable ici c'est la mention ; les modalités possibles sont : passable, assez bien, bien,... ce ne sont pas des nombres, donc c'est une variable qualitative. De plus, on sait que passable $<$ assez bien $<$ bien $<$... donc c'est une variable qualitative ordinale.

3. Dans un village, on veut compter le nombre des familles pour chaque nombre d'enfants : les familles ayant 0 enfant, les familles ayant 1 enfant, les familles ayant 2 enfants...

la variable ici c'est le nombre des enfants ; les modalités possibles ce sont : 0,1,2,... ce sont des nombres, donc c'est une variable quantitative. De plus, le nombre des enfants est limité (n'est pas infini), donc c'est une variable quantitative discrète.

4. Dans un quartier, on veut compter le nombre des personnes pour chaque taille : les personnes ayant 120cm, les personnes ayant 172,26 cm, les personnes ayant 180,54cm...

la variable ici c'est la taille des personnes ; les modalités possibles ce sont : 100cm, 100,01 cm, 140,57cm, 180,54cm, ... ce sont des nombres, donc c'est une variable quantitative. De plus, les tailles possibles sont infinies et on doit les classer par des intervalles, donc c'est une variable quantitative continues.

Comment déterminer les classes d'une variable continues ?

Soit X une variable statistique continue. Supposons qu'on veut les classer en k classes. Alors, on suit les démarches suivantes :

1. On prend x_{min} la plus petite modalité et x_{max} la plus grande modalité.
2. On calcule $e = \frac{x_{max} - x_{min}}{k}$.
3. Les classes sont :

$$[x_{min}, x_{min} + e[, [x_{min} + e, x_{min} + 2e[, [x_{min} + 2e, x_{min} + 3e[, \dots, [x_{min} + (k-1)e, x_{max}].$$

Exemple :

Supposons que dans une variable X , on a les modalités suivantes :

$$1.21, 1.16, 1.18, 1.26, 1.28, 1.20, 1.17, 1.36, 1.28, 1.27$$

On veut les classer en 4 classes ; c'est à dire, $k = 4$.

La plus petite modalité est $x_{min} = 1.16$ et la plus grande est $x_{max} = 1.36$.

On calcule $e = \frac{x_{max} - x_{min}}{k} = \frac{1.36 - 1.16}{4} = \frac{0.20}{4} = 0.05$.

Donc, les classes obtenues sont :

$$[1.16, 1.21[, [1.21, 1.26[, [1.26, 1.31[, [1.31, 1.36[.$$

2) Représentation des données

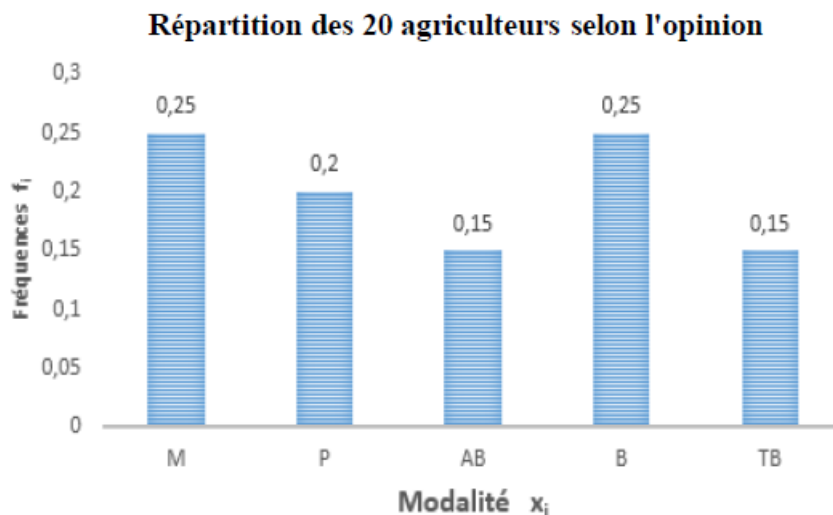
2.1) La représentation des données qualitatives

1. **Le tableau statistique** : c'est un tableau dans lequel on met les modalités (si la variable est ordinale, il faut les mettre ordonnées), l'effectif pour chaque modalité, la fréquence pour chaque modalité, si la variable est ordinale on ajoute l'effectif et la fréquence cumulé croissant/ décroissant.

Exemple :

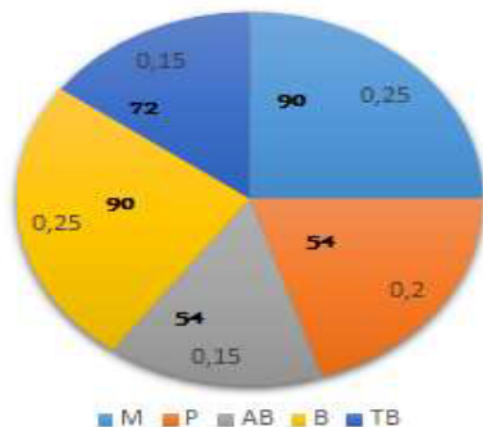
Modalité	effectif	fréquence	effectif cumulé croissant	fréq. cum. croissant
M	5	0.25	5	0.25
P	4	0.2	9	0.45
AB	3	0.15	12	0.6
B	5	0.25	17	0.85
TB	3	0.15	20	1

2. **Diagramme en barre** : ce diagramme est formé de rectangles d'hauteurs proportionnelles aux effectifs (ou fréquences) des modalités associées. Ces rectangles ont les mêmes bases et sont constants.



3. **Diagramme circulaire** : ce diagramme est formé de disque dont les secteurs sont proportionnels aux effectifs n_i (ou fréquences f_i) des modalités associées. En terme d'angle, on associe à chaque modalité x_i l'angle θ_i du secteur i tel que : $\theta_i = \frac{360 \times n_i}{N} = 360 \times f_i$.

Répartition des 20 agriculteurs selon l'opinion



2.2) La représentation des données quantitatives discrètes

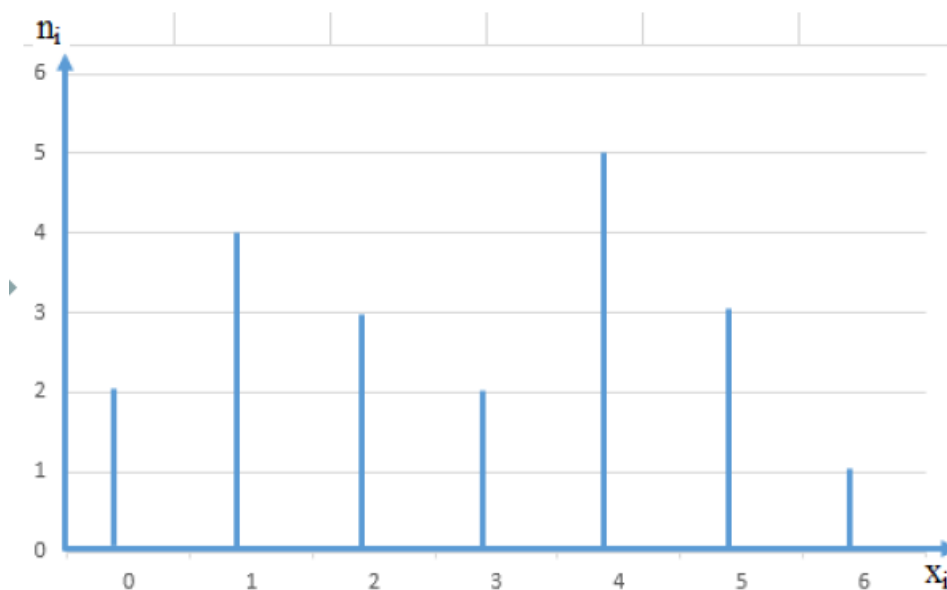
1. **Le tableau statistique** : c'est un tableau dans lequel on met les modalités ordonnées, l'effectif pour chaque modalité, la fréquence pour chaque modalité, l'effectif et la fréquence cumulés croissants/ décroissants.

Exemple :

<i>Modalité</i>	<i>effectif</i>	<i>fréquence</i>	<i>effectif cumulé croissant</i>	<i>fréq. cum. croissant</i>
0	2	0.1	2	0.1
1	4	0.2	6	0.3
2	3	0.15	9	0.45
3	2	0.1	11	0.55
4	5	0.25	16	0.8
5	3	0.15	19	0.95
6	1	0.05	20	1

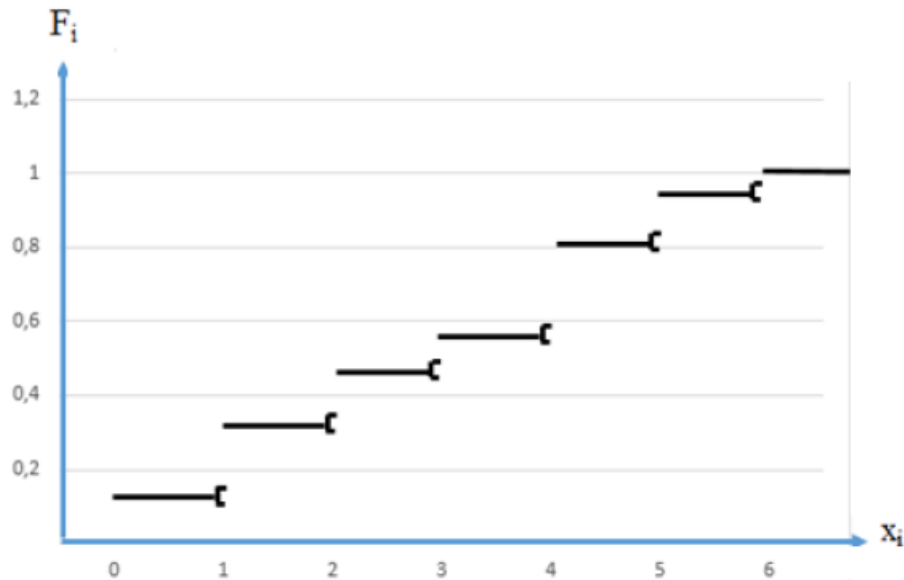
2. **Diagramme en bâtons** : ce diagramme est formé de segments dont chaque modalité x_i est associée à un segment de hauteur h_i proportionnel à l'effectif n_i ou à la fréquence f_i .

Exemple :



3. **Diagramme cumulatif** : ce diagramme est formé des paliers qui sont horizontaux et les fréquences cumulées croissantes $F(x)$ qui sont constantes sur chaque intervalle $[x_{i-1}, x_i[$.

Exemple :



2.2) La représentation des données quantitatives continues

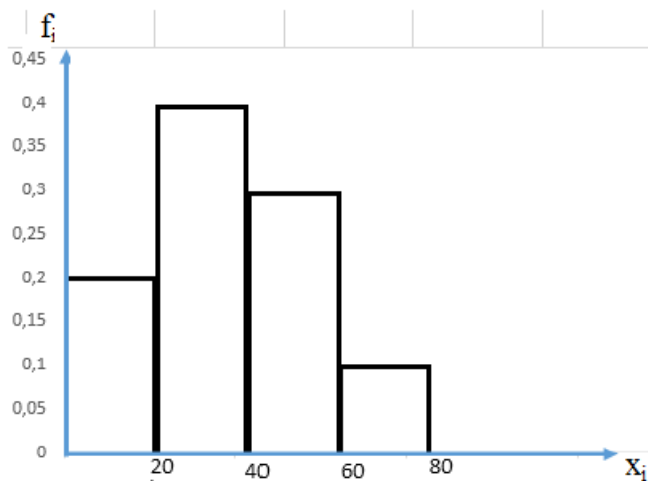
1. **Le tableau statistique** : c'est un tableau dans lequel on met les modalités ordonnées, l'effectif pour chaque modalité, la fréquence pour chaque modalité, l'effectif et la fréquence cumulés croissants/ décroissants.

Exemple :

<i>Modalité</i>	<i>effectif</i>	<i>fréquence</i>	<i>effectif cumulé croissant</i>	<i>fréq. cum. croissant</i>
$[0, 20[$	4	0.2	4	0.2
$[20, 40[$	8	0.4	12	0.6
$[40, 60[$	6	0.3	18	0.9
$[60, 80]$	2	0.1	20	1

2. **Histogramme** : c'est un diagramme formé de rectangles composés chacun d'une base égale à l'amplitude d'une classe et d'aire proportionnelle à l'effectif (ou fréquence).

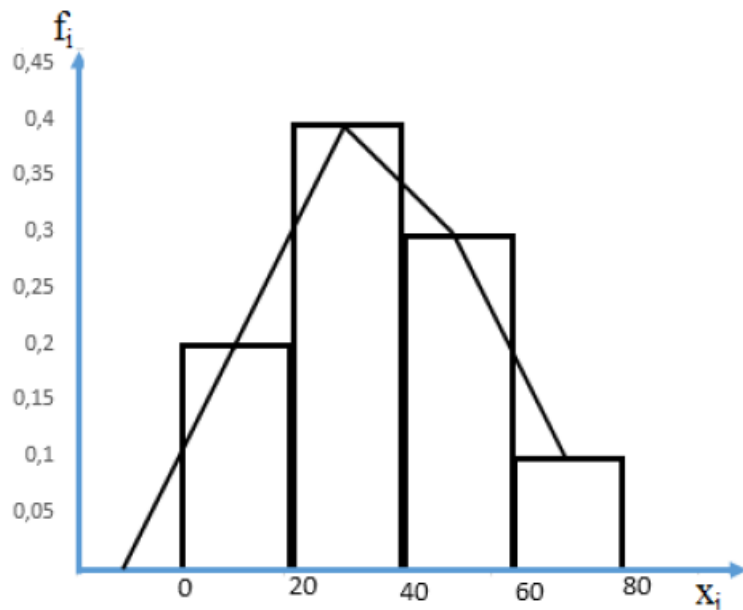
Exemple :



3. **Polygone de fréquence** ce diagramme est formé en joignant, par des segments

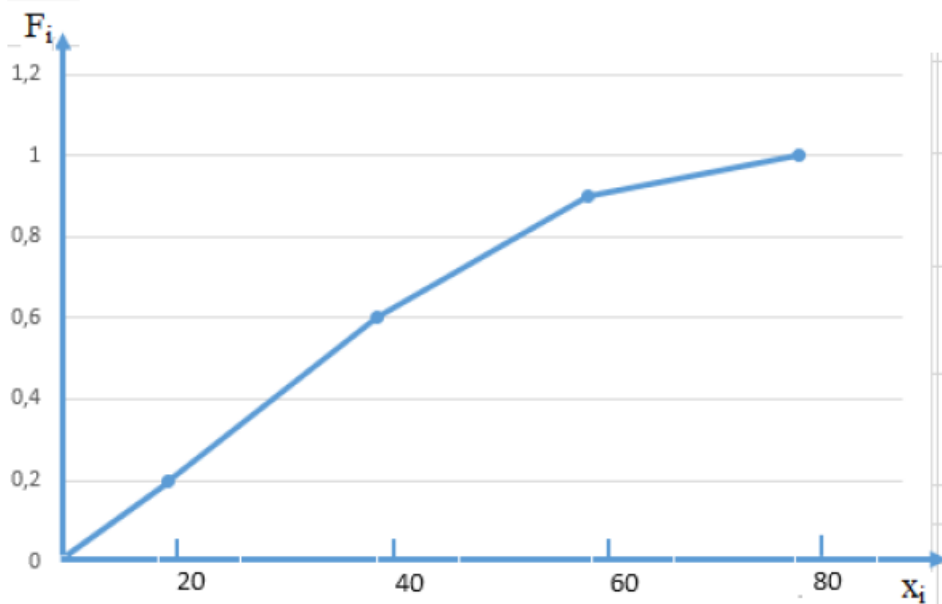
de droites, les milieux des sommets de chaque rectangle de l'histogramme. Ainsi, il doit commencer par 0 et se terminer par 0.

Exemple :



4. **Diagramme cumulatif :** ce diagramme est formé en joignant les points de coordonnées (x_i, F_i) pour chaque classe $[x_{i-1}, x_i[$ où F_i est la fréquence cumulée croissante de cette classe.

Exemple :



3) Caractéristiques de tendance centrale

3.1) Les moyennes

1. **La moyenne arithmétique :** c'est la somme de toutes les modalités divisée par la taille de la population. Notée $\bar{x}$.

On peut aussi la calculer par la formule : $\bar{x} = \frac{\text{somme de } n_i x_i}{\text{taille de popul.}}$

Exemple :

Considérons la série statistique $\{5, 8, 5, 7, 9, 5, 4, 1, 7, 4\}$. La taille de la population c'est 10. Donc,

$$\bar{x} = \frac{5 + 8 + 5 + 7 + 9 + 5 + 4 + 1 + 7 + 4}{10} = \frac{55}{10} = 5.5$$

Remarquer que l'effectif de 1 est 1, l'effectif de 5 est 3, l'effectif de 8 est 1, l'effectif de 7 est 2, l'effectif de 9 est 1 et l'effectif de 4 est 2. Donc on peut retrouver $\bar{x}$ par la formule :

$$\bar{x} = \frac{1 \times 1 + 3 \times 5 + 1 \times 8 + 2 \times 7 + 1 \times 9 + 2 \times 4}{10} = \frac{1 + 15 + 8 + 14 + 9 + 8}{10} = \frac{55}{10} = 5.5$$

Remarque : Dans le cas d'une variable continue avec des intervalles $[x_i, x_{i+1}[$, pour calculer la moyenne on considère le centre $c_i = \frac{x_{i+1} + x_i}{2}$ comme modalité.

2. **La moyenne géométrique :** La moyenne géométrique est utilisée dans le calcul du taux d'accroissement moyen et de certains indices statistiques. Notée G .

Pour la calculer on calcule d'abord $\log(G) = \frac{\text{la somme de } n_i \cdot \log(x_i)}{\text{la taille de popul.}}$. Ensuite, $G = e^{\log(G)}$.

Remarque : La moyenne arithmétique est très sensible aux valeurs extrêmes de la série alors que la moyenne géométrique est peu sensible à ces dernières.

Exemple :

Considérons la série statistique $\{5, 8, 5, 7, 9, 5, 4, 1, 7, 4\}$. La taille de la population c'est 10.

$$\begin{aligned} \log(G) &= \frac{3 \times \log(5) + 1 \times \log(1) + 1 \times \log(8) + 2 \times \log(7) + 1 \times \log(9) + 2 \times \log(4)}{10} \\ &\approx \frac{3 \times 1.61 + 1 \times 0 + 1 \times 2.08 + 2 \times 1.95 + 1 \times 2.20 + 2 \times 1.39}{10} \\ &= \frac{15.79}{10} \approx 1.58 \end{aligned}$$

Donc, $G \approx e^{1.58} = 4.85$.

3. **La moyenne harmonique :** est utilisée dans le calcul des études du pouvoir d'achat, des durées moyennes, des moyennes de rapports et de pourcentages. La moyenne harmonique, notée H , d'une variable statistique X de distribution est donnée par :

$$H = \frac{\text{la taille de population}}{\text{la somme de } n_i \times \frac{1}{x_i}}.$$

Remarque : La moyenne harmonique est plus sensible aux plus petites valeurs de la série qu'aux plus grandes.

Exemple :

Considérons la série statistique $\{5, 8, 5, 7, 9, 5, 4, 1, 7, 4\}$. La taille de la population

c'est 10.

$$\begin{aligned}
 H &= \frac{10}{3 \times \frac{1}{5} + 1 \times \frac{1}{1} + 1 \times \frac{1}{8} + 2 \times \frac{1}{7} + 1 \times \frac{1}{9} + 2 \times \frac{1}{4}} \\
 &\approx \frac{10}{3 \times 0.2 + 1 \times 1 + 1 \times 0.125 + 2 \times 0.14 + 1 \times 0.11 + 2 \times 0.25} \\
 &= \frac{10}{2.615} \approx 3.82
 \end{aligned}$$

4. **La moyenne quadratique** : est utilisée dans le calcul de certains paramètres de dispersion. La moyenne quadratique, notée Q , d'une variable statistique X est donnée par :

$$Q = \sqrt{\frac{\text{somme de } n_i \times x_i^2}{\text{taille de popul.}}}$$

Exemple :

Considérons la série statistique $\{5, 8, 5, 7, 9, 5, 4, 1, 7, 4\}$. La taille de la population c'est 10.

$$\begin{aligned}
 Q &= \sqrt{\frac{3 \times 5^2 + 1 \times 1^2 + 1 \times 8^2 + 2 \times 7^2 + 1 \times 9^2 + 2 \times 4^2}{10}} \\
 &= \sqrt{\frac{3 \times 25 + 1 \times 1 + 1 \times 64 + 2 \times 49 + 1 \times 81 + 2 \times 16}{10}} \\
 &= \sqrt{\frac{351}{10}} = \sqrt{35.1} \approx 5.92
 \end{aligned}$$

Proposition : Pour toute série statistique, on a toujours :

$$H < G < \bar{x} < Q$$

3.2) La médiane

Définition 3.2.1

La médiane M d'une distribution statistique ordonnée est la valeur qui la partage en deux parties égales. C'est la valeur du caractère qui dépasse pour la première fois 0,5 en fréquence cumulée croissante.

Exemple :

Dans le tableau suivant on présente les modalités avec leurs fréquences cumulées croissantes :

Modalité	x_1	x_2	x_3	x_4	x_5	x_6
Fréquence cumulée croissante	0,2	0,35	0,45	0,6	0,85	1

Alors la médiane c'est $M = x_4$, car c'est la première modalité qui dépasse 0,5 en fréquence cumulée croissante.

Proposition :

Si la variable est quantitative continue et toutes les classes ont la même amplitude, alors la première classe $]x_i, x_{i+1}[$ qui dépasse 0,5 en fréquence cumulée croissante s'appelle la classe médiane.

Dans ce cas, la médiane se calcule comme suit :

$$M = x_i + \frac{0,5 - F_{i-1}}{f_i} \times (x_{i+1} - x_i),$$

avec f_i c'est la fréquence de la classe $]x_i, x_{i+1}[$, F_{i-1} c'est la fréquence cumulée croissante de la classe précédente.

Exemple

Considérons la distribution suivante :

La classe	$[1, 3[$	$[3, 5[$	$[5, 7[$	$[7, 9[$	$[9, 11]$
La fréquence	0, 15	0, 25	0, 35	0, 15	0, 1
La fréquence cumulée croissante	0, 15	0, 4	0, 75	0, 9	1

La classe médiane c'est la première classe qui dépasse 0,5 en fréquence cumulée croissante, donc c'est $[5, 7[$.

La fréquence de la classe médiane c'est 0,35, la fréquence cumulée croissante de la classe précédente c'est 0,4. Donc,

$$M = 5 + \frac{0,5 - 0,4}{0,35} \times (7 - 5) = 5 + \frac{0,1}{0,35} \times 2 \approx 5,29.$$

3.3) Le mode

Définition 3.3.1

Le mode M_o d'une distribution statistique est sa valeur la plus fréquente. C'est la valeur du caractère qui correspond à l'effectif le plus grand ou à la fréquence la plus importante.

Le mode permet ainsi de connaître la valeur la plus probable du caractère.

Exemple :

Dans le tableau suivant on présente le nombre des femmes qui préfèrent chaque couleur des habits :

La couleur	Blanc	Noir	Vert	Bleu	Rouge	Jaune
Le nombre des femmes	15	31	9	12	40	14

Alors le mode c'est $M_o = \text{Rouge}$, car c'est la modalité qui a le plus grand effectif = 40.

Remarque :

On peut avoir plusieurs modes si le plus grand effectif se répète plusieurs fois.

Proposition :

Si la variable est quantitative continue et toutes les classes ont la même amplitude, alors la classe $]x_i, x_{i+1}[$ qui a le plus grand effectif s'appelle la classe modale.

Dans ce cas, le mode se calcule comme suit :

$$M_o = x_i + \frac{n_i - n_{i-1}}{2n_i - n_{i-1} - n_{i+1}} \times (x_{i+1} - x_i),$$

avec n_i c'est l'effectif de la classe $]x_i, x_{i+1}[$, n_{i-1} c'est l'effectif de la classe précédente, et n_{i+1} c'est l'effectif de la classe suivante.

Exemple

Considérons la distribution suivante :

La classe	[1, 3[	[3, 5[	[5, 7[	[7, 9[	[9, 11]
L'effectif	4	5	9	3	6

La classe modale c'est la classe qui a le plus grand effectif, donc c'est $[5, 7[$.

L'effectif de la classe modale c'est 9, l'effectif de la classe précédente c'est 5, et l'effectif de la classe suivante c'est 3. Donc,

$$M_o = 5 + \frac{9 - 5}{2 \times 9 - 5 - 3} \times (7 - 5) = 5 + \frac{4}{10} \times 2 = 5,8.$$

4) Caractéristiques de dispersion : la variance et l'écart-type

La variance et l'écart-type mesurent la dispersion des valeurs autour de la moyenne arithmétique. Plus les valeurs sont regroupées autour de la moyenne, plus la variance est petite. Plus elles sont dispersées, plus la variance est grande.

Définition 4.1

Pour une série statistique X de distribution $\{(x_1, n_1), (x_2, n_2), \dots, (x_k, n_k)\}$, la variance est

$$V(X) = \frac{1}{N} \times (n_1(x_1 - \bar{x})^2 + n_2(x_2 - \bar{x})^2 + \dots + n_k(x_k - \bar{x})^2),$$

avec N est la taille de population, n_i est l'effectif de la modalité x_i , et $\bar{x}$ est la moyenne arithmétique.

Remarque

On peut calculer la variance autrement avec la formule :

$$V(X) = \left(\frac{1}{N} \times (n_1 \times x_1^2 + n_2 \times x_2^2 + \dots + n_k \times x_k^2) \right) - \bar{x}^2.$$

Remarque

Dans le cas d'une variable continue, on remplace la modalité $]x_i, x_{i+1}[$ par son centre $c_i = \frac{x_i + x_{i+1}}{2}$.

Exemple

Dans l'exemple qu'on a fait sur la moyenne arithmétique, on a la série statistique $\{5, 8, 5, 7, 9, 5, 4, 1, 7, 4\}$. On avait la moyenne arithmétique : $\bar{x} = 5,5$. On présente la tableau statistique :

La modalité	1	4	5	7	8	9
L'effectif	1	2	3	2	1	1

La taille de population est $N = 10$. Donc, la variance est :

$$\begin{aligned}
 V &= \frac{1}{10} \times (1 \times (1 - 5,5)^2 + 2 \times (4 - 5,5)^2 + 3 \times (5 - 5,5)^2 + 2 \times (7 - 5,5)^2 + (8 - 5,5)^2 + (9 - 5,5)^2) \\
 &= \frac{20,25 + 4,5 + 0,75 + 4,5 + 6,25 + 12,25}{10}, \\
 &= \frac{48,5}{10}, \\
 &= 4,85
 \end{aligned}$$

Définition 4.2

L'écart-type σ se calcule par la formule :

$$\sigma(X) = \sqrt{V(X)}.$$

Exemple :

Dans l'exemple précédent, on avait la variance $V = 4,85$. Alors, l'écart-type est

$$\sigma = \sqrt{4,85} \approx 2,2.$$