Chapitre 1

Statistiques descriptives

1.1 Introduction

La statistique descriptive est un ensemble de méthodes permettant de décrire et d'analyser des phénomènes susceptibles d'être dénombrés et classés. Elle a pour but de décrire et non d'expliquer.

1.1.1 Concepts de base

Les observations constituent la source des informations statistiques. Avant de débuter l'étude, il faut définir l'ensemble étudié et les critères de la description chiffrée.

1. Les ensembles étudiés sont appelés **population**.
2. Les éléments de la population sont appelés **individus ou unités statistiques**.
3. Un sous ensemble de la population est **un échantillon** et sa taille correspond à son cardinal.
4. Les critères étudiés constituent des caractères ; et un caractère permet de déterminer une partition de la population.

Exemple 1.1. Nous résumons les différents concepts dans cet exemple :

Population : l'ensemble des tous les employés d'une usine.

Individu : chaque employé de l'usine.

Caractère : le salaire, l'état matrimonial, le nombre d'enfants,... etc.

Les modalités du caractère : marié, célibataire, divorcé et veuf sont les les modalités de l'état matrimonial, par exemple.

1.1.2 Types de caractères

On distingue deux types de caractères : qualitatif et quantitatif.

Caractère qualitatif

Définition 1.1. Un caractère est dit qualitatif lorsque ses modalités ne sont pas mesurables.

Exemple 1.2. Les couleurs du pelage : l'ensemble des modalités est

$$\{\text{noir}, \text{marron}, \text{blanc}, \ldots\}.$$

Caractère quantitatif

Définition 1.2. Un caractère est dit quantitatif lorsque ses modalités sont des nombres. On lui donne souvent le nom de **variable statistique**.

Une variable statistique peut être :

Discrète : Si elle prend des valeurs isolées.

Exemple 1.3. Le nombre d'enfants d'une famille.

Continue : Lorsqu'elle peut prendre n'importe quelle valeur dans son domaine de variation.

Remarque 1.1. Dans le cas continu, le nombre de ces valeurs est toujours très grand. Dans ce cas, on regroupe toutes ces valeurs en classes.
En général, toutes les grandeurs liées à l'espace (longueur, surface, volume, ...), au temps (age), à la masse (poids, teneur, ...) ou à des combinaisons (vitesse, débit, ...) sont des variables statistiques continues.

1.2 Tableaux statistiques et représentations graphiques

Soit une population composée de n individus, sur laquelle on a étudié un caractère possédant k valeurs possibles. Ces valeurs x_1, x_2, ..., x_k sont des modalités (cas qualitatif) ou des nombres (cas quantitatif).

Soient :

$$\begin{array}{ll} n_1 & \text{le nombre d'individus ayant pris la valeur } \quad x_1 \\ n_2 & \text{le nombre d'individus ayant pris la valeur } \quad x_2 \\ \vdots & \\ n_k & \text{le nombre d'individus ayant pris la valeur } \quad x_k \end{array}$$

n_i est appelé **fréquence** ou **effectif** de la valeur x_i et n est l'effectif total.

On appelle **fréquence relative** ou **effectif relatif** de la valeur x_i la quantité :

$$f_i = \frac{n_i}{n} \text{ (ou en \% } f_i = \frac{n_i}{n} \times 100\%).$$

C'est la proportion d'individus ayant pris la valeur x_i.

Remarque 1.2.

$$\begin{aligned} \sum_{i=1}^{k} n_i &= n_1 + n_2 + \ldots + n_k = n. \\ \sum_{i=1}^{k} f_i &= f_1 + f_2 + \ldots + f_k = \frac{n_1}{n} + \frac{n_2}{n} + \ldots + \frac{n_k}{n} = 1. \end{aligned}$$

Modalités x_i	Effectifs n_i	Fréquences relatives $f_i = \frac{n_i}{n}$
x_1	n_1	$f_1 = \frac{n_1}{n}$
x_2	n_2	$f_2 = \frac{n_2}{n}$
$\vdots$	$\vdots$	$\vdots$
x_k	n_k	$f_k = \frac{n_k}{n}$
Total	$\sum_{i=1}^{k} n_i = n$	$\sum_{i=1}^{k} f_i = 1$

Tableau des effectifs et des fréquences relatives

1.2.1 Tableau statistique relatif à un caractère qualitatif et sa représentation graphique

Exemple 1.4. On veut étudier les lois de Mendel sur le caractère couleur de la fleur de Balsamine. Pour cela on étudiera le croisement des plantes hétérozygotes. On obtient quatre couleurs : pourpre, rose, blanc-lavande et blanche.

Population : les plantes de Balsamine.

Individu : une plante.

Caractère étudié : couleur de la fleur.

Modalités x_i	Effectif n_i	Fréquences relatives f_i	f_i en %
Pourpre	1790	0.5778	57.78 %
Rose	547	0.1766	17.66 %
Blanc-Lavande	548	0.1769	17.69 %
Blanche	213	0.0688	6.88%
Total	3098	1	100 %

Représentation graphique

L'information résumée dans un tableau statistique se traduit par un graphique pour en réaliser une synthèse visuelle.

a) Représentation par tuyaux d'orgue (diagramme en colonnes)

Dans ce cas, le graphe s'obtient en construisant autant de colonnes que des modalités du caractère qualitatif. Ces colonnes sont des rectangles de bases constantes et de hauteurs proportionnelles aux fréquences relatives.

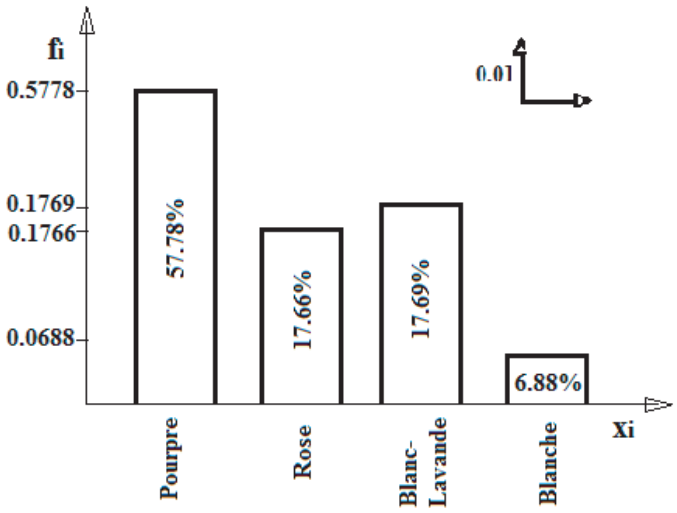


FIGURE 1.1 – Représentation en tuyaux d'orgue des fréquences relatives.

b) Représentation par le diagramme circulaire (camembert)

$$\theta_i = 360° \times f_i = 360° \times \frac{n_i}{n}.$$

Les angles correspondant de l'exemple sont :

$$\begin{aligned}
\theta_1 &= 360^\circ \times 0.5778 = 208.01^\circ \longrightarrow \text{Pourpre;} \\
\theta_2 &= 360^\circ \times 0.1766 = 63.58^\circ \longrightarrow \text{Rose;} \\
\theta_3 &= 360^\circ \times 0.1769 = 63.68^\circ \longrightarrow \text{Blanc-Lavande;} \\
\theta_4 &= 360^\circ \times 0.0688 = 24.77^\circ \longrightarrow \text{Blanche.}
\end{aligned}$$

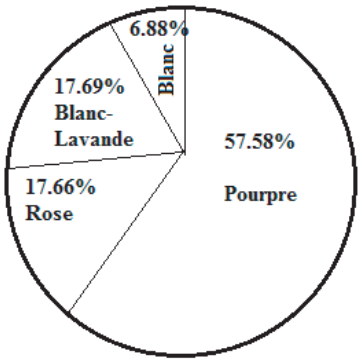


FIGURE 1.2 – Diagramme en camembert des fréquences relatives.

1.2.2 Tableaux statistiques relatifs à un caractère quantitatif et représentations graphiques

1) Cas d'une variable statistique discrète :

Exemple 1.5. Lors d'un contrôle d'une chaîne de médicaments, on s'intéresse au nombre de comprimés défectueux dans un lot. L'étude de 200 lots a donné les résultats suivants :
75 lots ont 0 comprimés défectueux ;
53 lots ont 1 comprimé défectueux ;
39 lots ont 2 comprimés défectueux ;
23 lots ont 3 comprimés défectueux ;
9 lots ont 4 comprimés défectueux ;
1 lot a 5 comprimés défectueux.

Population : l'ensemble des lots des médicaments.

Individu : un lot.

Caractère étudié : nombre de comprimés défectueux.

Modalités : 0, 1, 2, 3, 4 et 5.

Les fréquences relatives obtenues sont données dans le tableau suivant :

Modalités (Nbre de comprimés défectueux)	Nbre de lots n_i	Fréq. rel. $f_i = \frac{n_i}{n}$	Fréq. rel. en %
0	75	0.375	37.5 %
1	53	0.265	26.5 %
2	39	0.195	19.5 %
3	23	0.115	11.5 %
4	9	0.045	4.5 %
5	1	0.005	0.5 %
Total	200	1	100%

Représentation graphique :

On utilise le diagramme en batons pour représenter les effectifs n_i et les fréquences relatives f_i. Dans le cas du graphe des fréquences relatives, en joignant les sommets des batons on obtient le polygone des fréquences relatives.

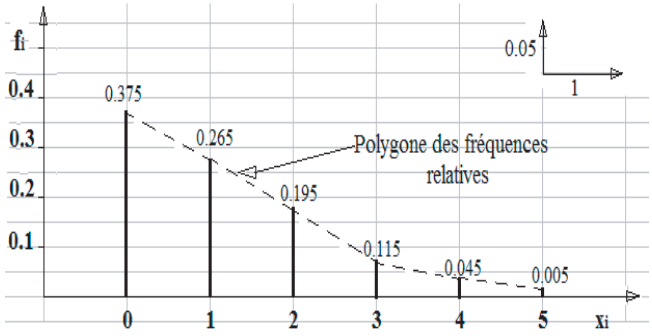


FIGURE 1.3 – Diagramme en escalier des fréquences relatives.

2) Cas d'une variable statistique continue :

Lorsque la variable statistique est continue les données sont regroupées en classes

$$[e_0, e_1[,\ [e_1, e_2[,\ [e_2, e_3[,\ ...,\ [e_{k-1}, e_k[.$$

Les modalités x_i représentent les centres c_i des classes $[e_{i-1}, e_i[$, avec :

- $c_i = \frac{e_{i-1}+e_i}{2}$, $i \in \{1, 2, ..., k\}$;
- e_{i-1} est appelé l'extrémité inférieure de la classe $[e_{i-1}, e_i[$;
- e_i est appelé l'extrémité supérieure de la classe $[e_{i-1}, e_i[$;
- $a_i = e_i - e_{i-1}$ est l'amplitude de la classe $[e_{i-1}, e_i[$;
- $e_k - e_0$ est appelé l'étendu de la variable statistique.

Exemple 1.6. Une étude faite sur la taille d'un groupe d'étudiants (en mètre) a donné les résultats suivants :

Classes	Centre c_i	Effectif n_i	Fréq. rel. f_i	(f_i en %)
$[1.5; 1.6[$	1.55	8	0.08	(8%)
$[1.6; 1.7[$	1.65	33	0.33	(30%)
$[1.7; 1.8[$	1.75	31	0.31	(31%)
$[1.8; 1.9[$	1.85	22	0.22	(22%
$[1.9; 2[$	1.95	6	0.06	(6%
Total		100	1	(100%)

Population : les étudiants du groupe.

Individu : un étudiant.

Caractère étudié : la taille d'un étudiant.

Représentation graphique :

Dans le cas d'une variable statistique continue on utilise l'histogramme.

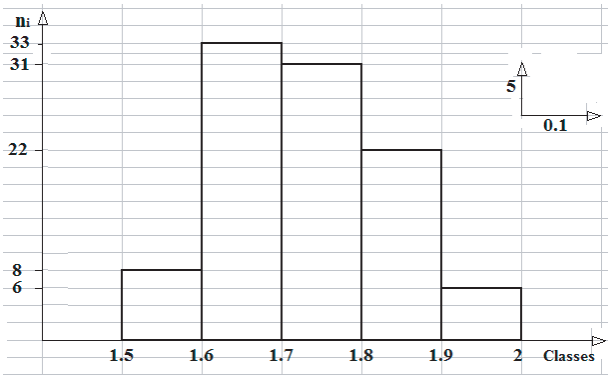


FIGURE 1.4 – L'histogramme des effectifs de l'exemple 1.6.

L'histogramme dans le cas des amplitudes inégales :

Dans ce cas les classes ont des amplitudes différentes. Du coup, il faut effectuer des corrections pour tenir compte des différences d'amplitude. Il convient de diviser les fréquences par leurs amplitudes correspondantes et on obtient ainsi, **l'amplitude corrigée** (h_i).

Exemple 1.7. Supposons que l'on regroupe les données de l'exemple précédent en classe d'amplitudes inégales.

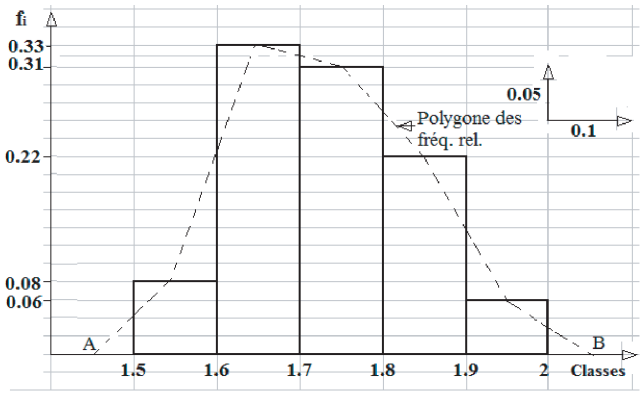


FIGURE 1.5 – L'histogramme des fréquences relatives de l'exemple 1.6.

Classes	Amplitude de la classe a_i	Effectif n_i	Fréq. rel. f_i f_i	Amplitude corrigée $h_i = \frac{f_i}{a_i}$
$[1.5; 1.7[$	0.2	41	0.41	2.05
$[1.7; 1.8[$	0.1	31	0.31	3.1
$[1.8; 2[$	0.2	28	0.28	1.4

Représentation graphique :

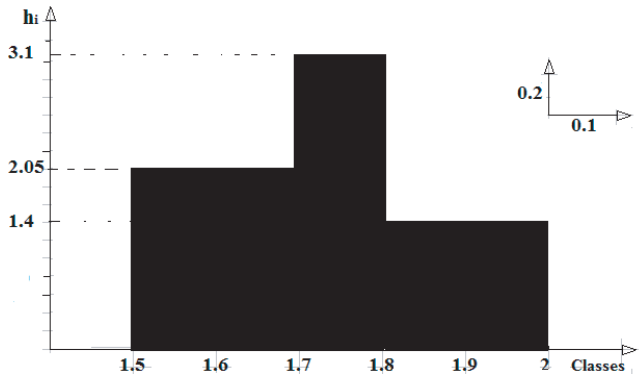


FIGURE 1.6 – Histogramme avec amplitudes inégales.

1.3 Fréquences relatives cumulées et effectifs cumulés

La fréquence relative cumulée F_i est la somme des fréquences relatives correspondantes aux valeurs de la variable statistique inférieure à x_{i+1}.

$$\begin{aligned} F_1 &= f_1; \\ F_2 &= f_1 + f_2; \\ F_3 &= f_1 + f_2 + f_3; \\ &\vdots \\ F_i &= f_1 + f_2 + \ldots + f_j + \ldots + f_i. \end{aligned}$$

La fréquence relative cumulée F_i indique la proportion des individus pour lesquels la variable statistique est inférieure à x_{i+1}. De la même façon on définit les effectifs cumulés

$$N_i = \sum_{j=1}^{i} n_j.$$

D'une manière équivalente, la fréquence relative cumulée est donnée par :

$$F_i = \frac{N_i}{n}.$$

1.3.1 Variable statistique discrète

Exemple 1.8. On reprend l'exemple du cas discret (l'exemple 1.5 des comprimés défectueux).

moins de x_i	N_i	F_i
moins de 0	0	0
moins de 1	$0 + 75 = 75$	$0 + f_1 = 0.375$
moins de 2	$75 + 53 = 128$	$f_1 + f_2 = 0.64$
moins de 3	167	0.835
moins de 4	190	0.95
moins de 5	199	0.995
moins de x_k $(x_k > 5)$	$199 + 1 = 200$	1

Remarque 1.3.

$$\sum_{i=1}^{k} f_i = 1;$$

$$F_i = 0 \quad \text{si} \quad x_i < x_1;$$

$$F_i = 1 \quad \text{si} \quad x_i > x_k,$$

où x_1 est la plus petite valeur observée et x_k est la plus grande valeur observée.

Soit X une variable statistique discrète et $x_1, x_2, ..., x_k$ les valeurs rangées dans l'ordre croissant. La fonction de répartition d'une v.s. discrète est définie de $\mathbb{R}$ dans $[0,1]$ et est donnée par :

$$F(x) = \begin{cases} 0, & \text{si } x < x_1\,; \\ f_1, & \text{si } x_1 \leq x < x_2\,; \\ f_1 + f_2, & \text{si } x_2 \leq x < x_3\,; \\ \vdots & \\ f_1 + f_2 + \ldots + f_i, & \text{si } x_i \leq x < x_{i+1}\,; \\ \vdots & \\ 1, & \text{si } x \geq x_k. \end{cases}$$

Exemple 1.9. Écrivons la fonction de répartition de la variable statistique X de l'exemple des comprimés défectueux (l'exemple 1.5) :

$$F(x) = \begin{cases} 0, & \text{si } x < 0\,; \\ 0.375, & \text{si } 0 \leq x < 1\,; \\ 0.64, & \text{si } 1 \leq x < 2\,; \\ 0.835, & \text{si } 2 \leq x < 3\,; \\ 0.95, & \text{si } 3 \leq x < 4\,; \\ 0.995, & \text{si } 4 \leq x < 5\,; \\ 1, & \text{si } x \geq 5. \end{cases}$$

La courbe cumulative est la représentation graphique des fréquences relatives cumulées. Dans le cas discret, la courbe cumulative est une courbe en escalier (voir figure 1.7), dont les paliers horizontaux ont pour coordonnées (x_i, F_i).

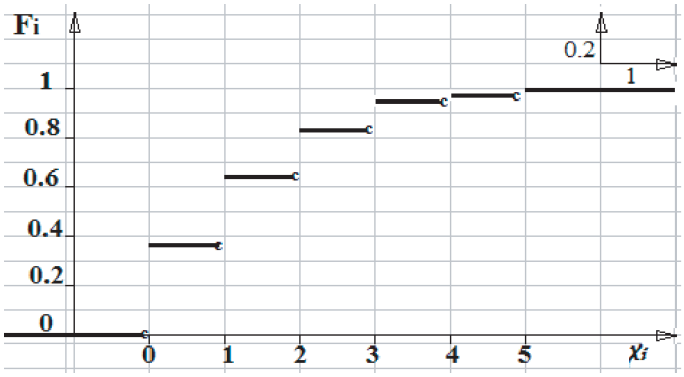


FIGURE 1.7 – Courbe des fréquences relatives cumulées (courbe cumulative).

1.3.2 Variable statistique continue

Exemple 1.10. On reprend l'exemple précédent sur la taille des étudiants. Les effectifs et les fréquences relatives cumulées sont données dans le tableau suivant :

moins de x_i	Effectifs cumulés N_i	Fréq. relatives $F_i = \frac{N_i}{n}$	cumulées (en %)
moins de 1.5	0	0	(0 %)
moins de 1.6	8	0.08	(8 %)
moins de 1.7	41	0.41	(41 %)
moins de 1.8	72	0.72	(72 %)
moins de 1.9	94	0.94	(94 %)
moins de x $(x \geq 2)$	100	1	(100 %)

La courbe cumulative est donnée dans la figure suivante :

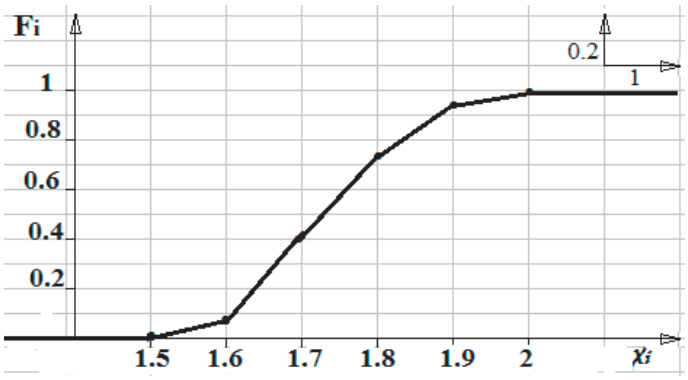


FIGURE 1.8 – Courbe des fréquences relatives cumulées du cas continu.

1.4 Paramètres d'une variable statistique

Lorsqu'on observe une représentation graphique d'une série statistique, on peut en tirer deux observations :

1. Paramètres de tendance centrale ou de position : valeurs situées au centre de la distribution statistique.
2. Paramètres de dispersion : fluctuations des observations autour de la valeur centrale, mesurées par des écarts à celles-ci.

1.4.1 Moyenne arithmétique

La moyenne arithmétique est la somme de toutes les valeurs observées divisée par le nombre total des observations.

(a) Cas d'une variable statistique discrète (données non groupées) :

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.

La moyenne arithmétique notée $\overline{x}$ de cette série statistique, est définie par :

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

Remarque 1.4.

$$\begin{aligned} \overline{x} = \frac{1}{n} \sum_{i=1}^{k} n_i x_i &= \frac{n_1 x_1 + n_2 x_2 + \ldots + n_k x_k}{n} \\ &= \frac{n_1}{n} x_1 + \frac{n_2}{n} x_2 + \ldots + \frac{n_k}{n} x_k \\ &= f_1 x_1 + f_2 x_2 + \ldots + f_k x_k \\ &= \sum_{i=1}^{k} f_i x_i, \end{aligned}$$

où f_i est la fréquence relative.

Exemple 1.11. La moyenne arithmétique de 200 lots de comprimés de l'exemple 1.5 est :

$$\begin{aligned} \overline{x} &= \frac{1}{n} \sum_{i=1}^{k} n_i x_i \\ &= \frac{75.0 + 53.1 + 39.2 + 23.3 + 9.4 + 1.5}{200} = \frac{241}{200} = 1.205. \end{aligned}$$

(b) Cas d'une variable statistique continue (données groupées) :

Dans le cas d'une variable statistique continue, les observations sont groupées dans des classes ; et nous avons la même formule que le cas discret, sauf qu'on remplace les x_i par les centres des classes x_i :

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

Exemple 1.12. La moyenne arithmétique de la taille des étudiants de l'exemple 1.6 est :

$$\begin{aligned}\overline{x} &= \frac{1}{n}\sum_{i=1}^{k} n_i c_i \\ &= \frac{(8\times 1.55)+(33\times 1.65)+(31\times 1.75)+(22\times 1.85)+(6\times 1.95)}{100} \\ &= \frac{173.5}{100} = 1.735.\end{aligned}$$

(c) Propriétés de la moyenne arithmétique :

La moyenne de la somme des écarts d'un ensemble d'observations à leurs moyenne arithmétique est nulle $\frac{1}{n}\sum_{i=1}^{k} n_i(x_i-\overline{x})=0$.

En effet,

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{k} n_i(x_i-\overline{x}) &= \frac{1}{n}\sum_{i=1}^{k}(n_i x_i - n_i\overline{x}) = \frac{1}{n}\sum_{i=1}^{k} n_i x_i - \frac{1}{n}\sum_{i=1}^{k} n_i\overline{x} \\ &= \frac{1}{n}\sum_{i=1}^{k} n_i x_i - \overline{x}\frac{1}{n}\sum_{i=1}^{k} n_i = \frac{1}{n}\sum_{i=1}^{k} n_i x_i - \overline{x}\frac{n}{n} \\ &= \frac{1}{n}\sum_{i=1}^{k} n_i x_i - \overline{x} = \overline{x}-\overline{x} = 0.\end{aligned}$$

1.4.2 La médiane

La médiane est la valeur pour laquelle, on a le même nombre d'individus à gauche et à droite dans un échantillon. Elle correspond au milieu de la distribution.

(a) Cas d'une variable statistique discrète :

Pour déterminer la médiane d'un échantillon ou d'une population :

1. on classe les individus par ordre croissant ;
2. on prend celui du milieu.

Remarque 1.5. la médiane Me est la valeur qui se trouve au centre de la série statistique :

• Si la valeur de l'effectif total n est impaire, i.e. $n=2p+1$, alors la médiane Me est la valeur qui se trouve à l'ordre $p+1$ (ou $\frac{n+1}{2}$) :

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

• Si la valeur de l'effectif total n est paire, i.e. $n = 2p$, alors la médiane Me est la moyenne des valeurs qui se trouve à l'ordre p et $p+1$ (ou $\frac{n}{2}$ et $\frac{n}{2}+1$) :

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

Exemple 1.13. Soit un échantillon de 11 personnes dont le poids en kg est :

$$45, 68, 89, 74, 55, 62, 56, 74, 49, 52, 63.$$

Les poids classés par ordre croissant sont :

$$\underbrace{45, 68, 49, 52, 55, 56}_{5}, \underbrace{62}_{x_6=Me}, \underbrace{63, 68, 74, 74, 89}_{5}.$$

Si le nombre d'individus est pair, $n = 12$, on prend la moyenne entre les deux valeurs centrales.

$$\underbrace{45, 68, 49, 52, 55, 55, 56}_{6}, \underbrace{62, 63, 68, 74, 74, 89}_{6}.$$

$$45, 68, 49, 52, 55, 55, \underbrace{56, 62}, 63, 68, 74, 74, 89.$$

La médiane $Me = \frac{x_{\frac{n}{2}} + x_{\frac{n}{2}+1}}{2} = \frac{x_{\frac{12}{2}} + x_{\frac{12}{2}+1}}{2} = \frac{x_6 + x_7}{2} = \frac{56+62}{2} = 59$ kg.

(b) Cas d'une variable statistique continue :

Dans ce cas la médiane est donnée par la formule suivante :

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

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

Exemple 1.14. Prenons le tableau des fréquences relatives cumulées de l'exemple 1.6. Pour déterminer la classe médiane, il suffit de tirer une ligne horizontale partant du point 0.5 (50%) de l'axe des fréquences relatives cumulées dans la courbe cumulative, arriver à l'ogive on descend une ligne verticale jusqu'à l'axe des x, et la classe où se situe le point d'intersection est la classe médiane. La classe médiane correspond, aussi, à la classe où les effectifs cumulées atteint ou dépasse pour la première fois le 50%.
Dans notre exemple, la classe médiane est la classe $[1.7; 1.8[$.

- $L_i = 1.7$;
- $n = 100$;
- $\sum_{i=1}^{<Me} n_i = 33 + 8$;
- $n_{Me} = 31$;
- $a = 0.1$;

$$Me = L_i + (\frac{\frac{n}{2} - \sum_{i=1}^{<Me} n_i}{n_{Me}}).a = 1.7 + (\frac{\frac{100}{2} - (33+8)}{31})0.1 = 1.7290.$$

1.4.3 Le mode

(a) Cas d'une variable statistique discrète :

Le mode Mo d'un ensemble d'observations est l'observation que l'on rencontre le plus fréquemment et il correspond à la modalité x_i ayant le plus grand effectif.

Remarque 1.6. Une distribution observée peut avoir plusieurs modes. Lorsqu'une distribution observée possède un seul mode, on parle de distribution unimodale. Lorsqu'une distribution observée possède deux modes, on parle de distribution bimodale.

Exemple 1.15.
1. Dans l'exemple des comprimés défectueux (exemple 1.5), le mode est 0.
2. Dans l'exemple des observations suivantes : 2, 2, 3, 5, 6, 6, 7, 8, 9, 7, 10 ; y a trois modes : 2, 6 et 7.

(b) Cas d'une variable statistique continue :

Dans le cas des données groupées en classes, le mode se calcule par la formule :

$$Mo = L_i + (\frac{\Delta_1}{\Delta_1 + \Delta_2}).a$$

- L_i : borne inférieure de la classe modale (classe correspondant au plus grand effectif) ;
- Δ_1 : excédent de l'effectif de la classe modale par rapport à l'effectif de la classe précédente ;
- Δ_2 : excédent de l'effectif de la classe modale par rapport à l'effectif de la classe suivante ;
- a : amplitude de la classe modale.

Exemple 1.16. Prenons les données de l'exemple du cas quantitatif continu (exemple 1.6) et calculons le mode.

- La classe modale est $[1.6; 1.7[$;
- $L_i = 1.6$;
- $\Delta_1 = 33 - 8$;
- $\Delta_2 = 33 - 31$;
- $a = 0.1$.

Le mode est

$$Mo = L_i + (\frac{\Delta_1}{\Delta_1 + \Delta_2}).a = 1.6 + (\frac{33 - 8}{(33 - 8) + (33 - 31)}) \times 0.1 = 1.6926 \text{ mètres.}$$

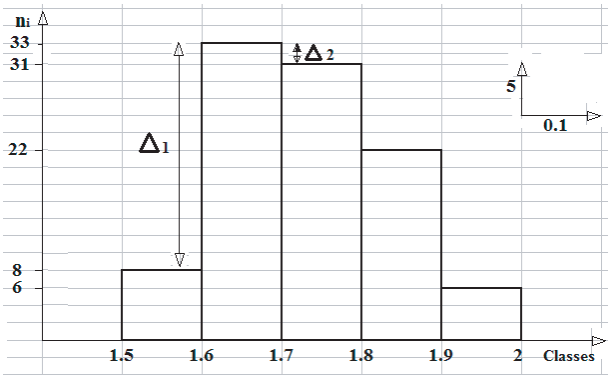


FIGURE 1.9 – L'histogramme des effectifs avec une illustration de Δ_1 et Δ_2.

1.5 Paramètres de dispersion

Les deux ensembles d'observations suivants :

$$X = \{6, 6, 7, 7, \underbrace{8}_{\overline{x}=Me}, 9, 9, 10, 10\} \text{ et } Y = \{1, 2, 4, 6, \underbrace{8}_{\overline{y}=Me}, 10, 12, 14, 15\}$$

ont la même moyenne et la même médiane $\overline{x} = \overline{y} = Me = 8$, mais ils sont différents. Le premier ensemble est moins dispersé que le deuxième.

1.5.1 Étendu

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

Exemple 1.17. L'étendu de X est $e = 10 - 6 = 4$ et l'étendu de Y est $e = 15 - 1 = 14$.

1.5.2 Variance et écart type

- **Variance**

  La variance notée $V(X)$ est la moyenne des carrés des écarts des observations à la moyenne. Elle est définie par :

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

  Propriétés

  1) $V(X) \geq 0$;

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

  3) Lorsqu'on compare les observations de deux variables statistiques X et Y, celle qui possède l'écart-type le plus élevé est la plus dispersée.

- **Écart type**

  L'écart-type noté σ_X est la racine carrée de $V(X)$:

$$\sigma_X = \sqrt{V(X)} = \sqrt{\sum_{1}^{k} f_i(x_i - \overline{x})^2}.$$

Exemple 1.18. Cas quantitatif discret

La variance des 200 lots de médicaments (exemple 1.5)) est :

$$\begin{aligned} V(X) &= \frac{1}{n}\sum_{1}^{k} n_i x_i^2 - \overline{x}^2 \\ &= \frac{(75\times 0^2)+(53\times 1^2)+(39\times 2^2)+(23\times 3^2)+(9\times 4^2)+(1\times 5^2)}{200} - (1.205)^2 \\ &= 1.473 \end{aligned}$$

et l'écart-type $\sigma_X = \sqrt{V(X)} = \sqrt{1.473} = 1.214$.

Exemple 1.19. Cas quantitatif continu

La variance de l'exemple 1.6 est :

$$\begin{aligned} V(X) &= \frac{1}{n}\sum_{1}^{k} n_i c_i^2 - \overline{x}^2 \\ &= \frac{(8\times 1.55^2)+(33\times 1.65^2)+(31\times 1.75^2)+(22\times 1.85^2)+(6\times 1.95^2)}{100} - (1.735)^2 \\ &= 0.010875. \end{aligned}$$

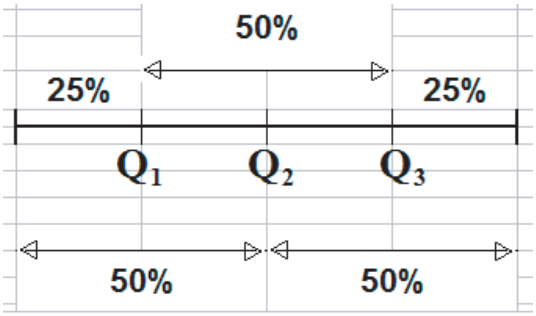
L'écart-type $\sigma_X = \sqrt{V(X)} = \sqrt{0.010875} = 0.1043$.

1.5.3 Les quartiles

On utilise couramment les quartiles Q_1, Q_2 et Q_3.
Q_1 est le quartile d'ordre $\frac{1}{4}$, représente 25% de l'échantillon ;
Q_2 est le quartile d'ordre $\frac{1}{2}$, représente 50% de l'échantillon ;
Q_3 est le quartile d'ordre $\frac{3}{4}$, représente 75% de l'échantillon.

Intervalle interquartile

(Q_1, Q_3) contient 50% de la population laissant à droite 25% et à gauche 25%. Cet intervalle est donnée par : $Q_1 - Q_3$.
Pour déterminer l'intervalle interquartile, il faut déterminer d'abord Q_1 et Q_3.



Le premier quartile Q_1

(a) Cas discret

Q_1 est la valeur x_i dont le rang (la position) est le plus petit entier qui suit $\frac{n}{4}$.

Exemple 1.20. Dans l'exemple des observations suivantes : 2, 3, 4, 5, 6, 6, 7, 7, 8, 9, 10 ; on a :
$n = 11$ et $\frac{n}{4} = \frac{11}{4} = 2.25$. Le plus petit entier qui suit $\frac{n}{4} = 2.25$ est 3, alors Q_1 est la troisième valeur. D'où $Q_1 = x_3 = 4$.

Exemple 1.21. Reprenons l'exemple des comprimés défectueux (exemple 1.5), où on a $\frac{n}{4} = \frac{200}{4} = 50$ et la valeur pour laquelle les effectifs cumulés atteignent ou dépassent pour la première fois 50 est "moins 1", c'est à dire la valeur de 0, alors $Q_1 = 0$.

(b) Cas continu

Dans ce cas le premier quartile est donné par la formule suivante :

$$Q_1 = L_i + \left(\frac{\frac{n}{4} - \sum_{i=1}^{<Q_1} n_i}{n_{Q_1}}\right).a$$

- L_i : borne inférieure de la classe de Q_1 ;
- n : effectif total ;
- $\sum_{i=1}^{<Q_1} n_i$: somme des effectifs correspondant à toutes les classes inférieures à la classe de Q_1 ;
- n_{Q_1} : effectif de la classe de Q_1 ;
- a : amplitude de la classe de Q_1.

Exemple 1.22. Prenons le tableau de l'exemple 1.6. Pour déterminer la classe de Q_1, il suffit de tirer une ligne horizontale partant du point 0.25 (25%) de l'axe des fréquences relatives cumulées dans la courbe cumulative, arriver à l'ogive on descend une ligne verticale jusqu'à l'axe des x et la classe où se situe le point d'intersection est la classe de Q_1.
La classe de Q_1 correspond, à la classe où les effectifs cumulées atteignent ou dépassent pour la première fois 25% de l'effectif total.
Dans notre exemple, la classe de Q_1 est la classe $[1.6; 1.7[$.

- $L_i = 1.6$;
- $n = 100$;
- $\sum_{i=1}^{<Q_1} n_i = 8$;
- $n_{Q_1} = 33$;
- $a = 0.1$;

$$Q_1 = L_i + (\frac{\frac{n}{4} - \sum_{i=1}^{<Q_1} n_i}{n_{Q_1}}).a = 1.6 + (\frac{\frac{100}{4} - 8}{33})0.1 = 1.65515.$$

Le troisième quartile Q_3

(a) Cas discret

Q_3 est la valeur x_i dont le rang (la position) est le plus petit entier qui suit $\frac{3n}{4}$.

Exemple 1.23. Dans le même exemple des observations : 2, 3, 4, 5, 6, 6, 7, 7, 8, 9, 10 ; on a :
$n = 11$ et $\frac{3n}{4} = \frac{3\times 11}{4} = 8.25$. Le plus petit entier qui suit $\frac{3n}{4} = 8.25$ est 9, alors Q_3 est la $9^{\text{ème}}$ valeur. D'où $Q_3 = x_9 = 8$.

Exemple 1.24. Dans l'exemple des comprimés défectueux (exemple 1.5), on a $\frac{3n}{4} = \frac{3\times 200}{4} = 150$ et la valeur où les effectifs cumulés atteignent ou dépassent pour la première fois 150 est "moins 3", c'est à dire la valeur de 2, alors $Q_3 = 2$.

(b) Cas continu

Dans le cas continu, le troisième quartile est donné par la formule suivante :

$$Q_3 = L_i + (\frac{\frac{3n}{4} - \sum_{i=1}^{<Q_3} n_i}{n_{Q_3}}).a$$

- L_i : borne inférieure de la classe de Q_3 ;
- n : effectif total ;
- $\sum_{i=1}^{<Q_3} n_i$: somme des effectifs correspondant à toutes les classes inférieures à la classe de Q_3 ;
- n_{Q_3} : effectif de la classe de Q_3 ;
- a : amplitude de la classe de Q_3.

Exemple 1.25. Pour déterminer la classe de Q_3 de l'exemple 1.6, également, il suffit de tirer une ligne horizontale partant du point 0.75 (75%) de l'axe des fréquences relatives cumulées dans la courbe cumulative, arriver à l'ogive on descend une ligne verticale jusqu'à l'axe des x et la classe où se situe le point d'intersection est la classe de Q_3.
La classe de Q_3 correspond, à la classe où les effectifs cumulées atteignent ou dépassent pour la première fois 75% de l'effectif total.
Dans notre exemple, la classe de Q_3 est la classe $[1.8; 1.9[$.

- $L_i = 1.8$;
- $n = 100$;
- $\sum_{i=1}^{<Q_3} n_i = 8 + 33 + 31$;
- $n_{Q_3} = 22$;
- $a = 0.1$;

$$Q_3 = L_i + (\frac{\frac{3n}{4} - \sum_{i=1}^{<Q_3} n_i}{n_{Q_3}}).a = 1.8 + (\frac{\frac{3\times100}{4} - (8 + 33 + 31)}{22})0.1 = 1.8136.$$