STATISTIQUE DESCRIPTIVE

S. DERRADJI

Table des Matières

Chapitre III

Distribution à deux caractères

Introduction

Ce document est conçu dans le cadre des cours à distance. C'est une introduction à la statique descriptive. Son contenu correspond au programme du module intitulé: Les Statistiques en Biomécanique que j'ai dispensé aux étudiants master II biomécanique. Il se distingue par l'autosuffisance, l'auto apprentissage et l'autoévaluation.

Il est constitué de trois chapitres:

Chapitre I : définitions de base.

Chapitre II : caractéristiques de position et de dispersion.

Chapitre III : distribution à deux caractères.

CHAPITRE I

DEFINITIONS DE BASE

1.1 Définitions

Définition 1.1.1

a) Une population est un ensemble d'objets, de choses soumis à une étude statistique.
b) Les éléments de cet ensemble sont appelés individus.

Exemple 1.1.1 L'ensemble des Algériens, les salariés d'une entreprise, les étudiants d'une université constituent des populations.

Définition 1.1.2 Un échantillon est un sous-ensemble d'une population.

Exemple 1.1.2 Au lieu de prendre toute la population des étudiants d'une université, on prend n étudiants. L'ensemble de ces n étudiants est un échantillon.

Définition 1.1.3 Une variable statistique (ou caractère) est un aspect particulier des individus auxquels on s'intéresse.

Exemple 1.1.3 L'âge des étudiants, le nombre d'enfants d'un salarié peuvent être choisis comme variables statistiques (caractères).

L'ensemble des observations élémentaires d'une variable statistique forme l'ensemble des modalités de ce caractère.

Exemple 1.1.4 Pour la population des étudiants, si la variable statistique est l'âge alors les modalités sont l'ensemble des âges des étudiants. Si elle est le sexe, il y a deux modalités : masculin et féminin.

Mathématiquement, une variable statistique peut donc être définie comme une application entre la population et l'ensemble des modalités. On la note X Elle associe à chaque individu x_i une modalité w_i.

Définition 1.1.4 Une variable statistique est dite qualitative si l'ensemble des modalités n'est pas un ensemble de nombres.

Exemple 1.1.5

L'ensemble des modalités da la variable statistique sexe est$\{masculin, féminin\}$.

L'ensemble des modalités da la variable statistique récolte de blé est $\{très\ bonne, bonne, médiocre, très\ mauvaise\}$.

Définition 1.1.5 Une variable statistique est dite quantitative si l'ensemble des modalités est un ensemble de nombres.

Exemple 1.1.6 l'âge, le salaire, la production de blé correspondent à des variables statistiques quantitatives.

Définition 1.1.6 Une variable statistique quantitative est dite discrète si ses modalités sont des nombres isolés les uns des autres. Il s'agit souvent de nombres entiers.

Exemple 1.1.7 les variables statistiques âge, nombre d'enfants sont des variables statistiques quantitatives.

Définition 1.1.7

a) Une variable statistique quantitative est dite continue si ses modalités peuvent prendre toutes les valeurs d'un intervalle réel.
b) Ces valeurs sont regroupées dans des intervalles de valeurs numériques appelées classes. Elles sont notées$[e_i; e_{i+1}[$.
c) La moyenne des extrémités d'une classe qu'on note $\bar{x}_i$ est appelée centre de la classe et elle est donnée par :

$$\bar{x}_i = \frac{e_i + e_{i+1}}{2}$$

Définition 1.1.8

a) Le nombre d'individus présentant la modalité m_i (variable qualitative) ou x_i (variable quantitative discrète) ou une modalité incluse dans $[e_i; e_{i+1}[$(variable quantitative continue) est appelé effectif et il est noté n_i. S'il y a k modalités d'une variable statistique, les effectifs correspondant à chaque modalité sont notés: $n_1, n_2, \ldots, n_k$.
b) La somme des effectifs est appelée effectif total. On le note n. Il est donc égal au nombre d'individus d'une population :

$$n = \sum_{i=1}^{k} n_i$$

Définition 1.1.9

a) La fréquence associée à une modalité, ou à un ensemble de modalités regroupées en classes est la proportion d'individus présentant cette modalité ou cet ensemble de modalités, par rapport à l'ensemble des individus (effectif total).
b) La fréquence associée à la $i^{ème}$ modalité, la $i^{ème}$ classe est notée f_i . Elle est donnée par :

$$f_i = \frac{n_i}{n} = \frac{effectif}{effectif\ total}$$

Définition 1.1.10 L'ensemble des couples(x_i, n_i) ou (x_i, f_i) est appelé distribution statistique de la variable statistique.

Exemple 1.1.8 Soit la série de valeurs représentant les notes de 20 étudiants : 14, 16, 12, 12, 8, 10, 12, 8, 16, 10,16, 10,12, 8, 12, 8, 10, 14,10, 10.

Mettre ses données sous la forme d'un tableau dont les colonnes: 1, 2 et 3 représentent respectivement les modalités, les effectifs et les fréquences.

Solution

x_i	n_i	f_i
8	4	0.2
10	6	0.3
12	5	0.25
14	2	0.1
16	3	0.15
total	20	1

Les couples $(8,4)$, $(10,6)$, $(12,5)$, $(14,2)$, $(16,3)$ ou les $(8,0.2)$, $(10,0.3)$, $(12,0.25)$, $(14,0.1)$, $(16,0.15)$ représentent la distribution statistique de la variable statistique moyenne.

1.2 Lecture des données et représentations graphiques des distributions à une variable

Les données peuvent être présentées dans un tableau et/ou un graphique.

1.2.1 Lecture d'un tableau de données

Considérons le tableau suivant:

Catégorie	Nbre de logements (en milliers)	Part des logements (%)
Résidences principales	27161	84.2
Résidences secondaires	3235	10.0
Résidences vacantes	1864	5.8
Total	32260	100%

Ce tableau résume les informations suivantes:

1) La population est l'ensemble de logements en France en 2007.
2) Chaque logement est un individu de cette population.
3) Il est caractérisé par la catégorie à laquelle il appartient : Résidence principale, Résidence secondaire, Résidence vacante. La variable statistique étudiée est donc la catégorie de logements.
4) Elle a trois modalités: Résidence principale, Résidence secondaire, Résidence vacante.
5) L'effectif des Résidences principales est $n_1 = 27161000$.

6) L'effectif des Résidences secondaires est $n_2 = 3235000$.
7) L'effectif des Résidences vacantes est $n_2 = 1864000$.
8) L'effectif total des résidences est $n = 32260000$.

1.2.2 Représentations graphiques des distributions à une variable

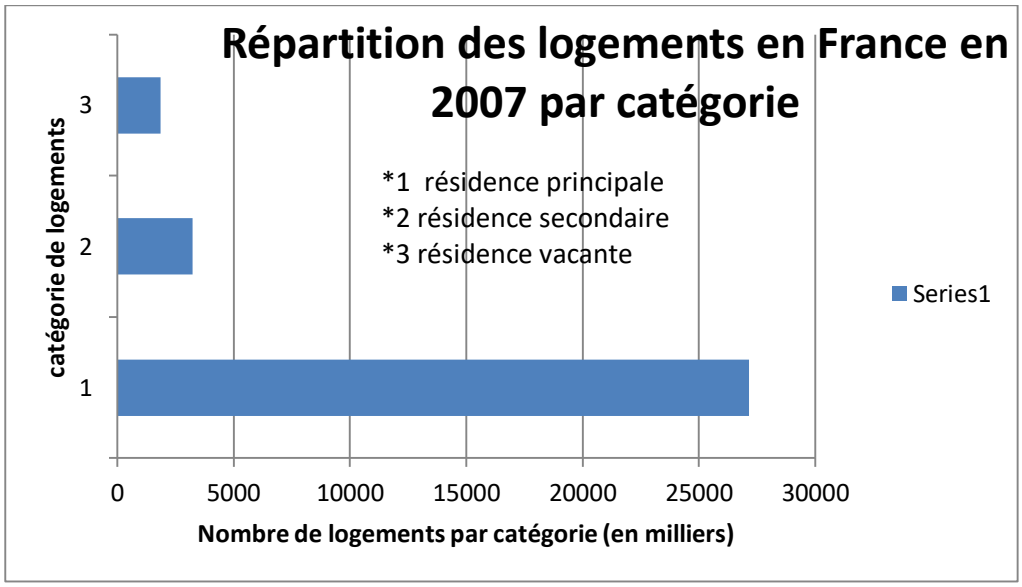
1.2.2.1 Variable statistique qualitative

1) Diagrammes à bandes

Dans un diagramme à bandes, On associe à chaque modalité une bande verticale ou horizontale dont la hauteur représente l'effectif ou la fréquence de cette modalité.

Exemple 1.2.1

Le tableau ci-dessus peut être représenté par le diagramme à bandes suivant **:**



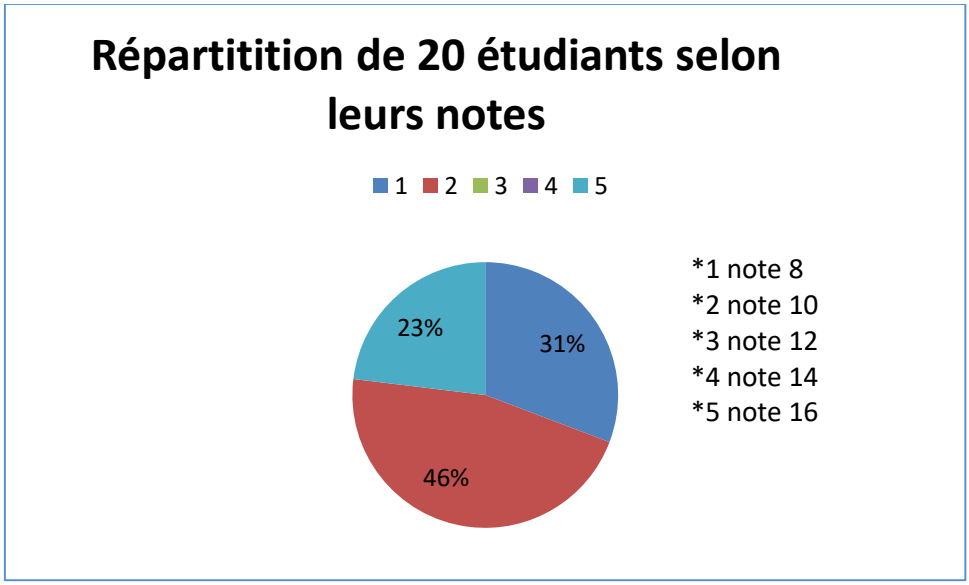
2) Diagramme à secteurs circulaires

Définition 1.2.1

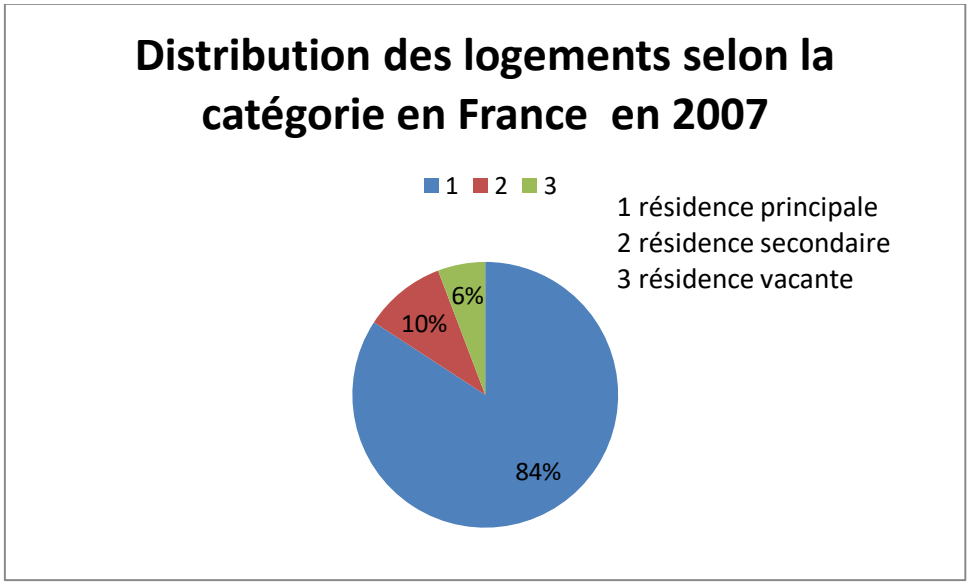
a) Un diagramme à secteurs circulaires est un graphique qui divise un disque en secteurs angulaires.
b) Le nombre de secteurs est égal au nombre de modalités.
c) Les angles au centre de ces secteurs sont proportionnels aux effectifs (ou aux fréquences) de chaque modalité.
d) L'angle au centre α_i, en degré, associé à la modalité x_i d'effectif n_i est donné par $\alpha_i = \frac{n_i}{n} 360 = f_i 360$

Exemple 1.2.2

a) Le tableau de l'exemple 1.1.9 peut être représenté par le diagramme à secteurs



b) Le tableau des logements en France en 2007 dessus peut être représenté par le diagramme à secteurs



1.2.2.2 Variable statistique quantitative discrète

a) Lecture d'un tableau de données

Soit le tableau suivant donnant les notes obtenues par 40 étudiants ayant participé à l'épreuve de statistiques.

x_i(note)	n_i	f_i
1	2	0.05
2	2	0.05
3	5	0.125
4	6	0.15
5	9	0.225
6	5	0.125
7	4	0.1
8	4	0.1
9	2	0.05
10	1	0.025
Total	n =40	1

Ce tableau résume les informations suivantes:

.1) La population est l'ensemble des étudiants ayant participé à l'épreuve de statistiques.

1) La population est l'ensemble des étudiants ayant participé à l'épreuve de statistiques.
2) Chaque étudiant est un individu de cette population
3) Il est caractérisé par la note ayant obtenu.
4) Ce caractère a 10 modalités :1, 2,…,10
5) L'effectif total des étudiants est $n = 40$.
6) La fréquence f_1 des étudiants ayant obtenu 1 est 5%, leur effectif est $n_1 = 2$.

   La fréquence f_2 des étudiants ayant obtenu 2 est 5%, leur effectif est $n_2 = 2$.

   La fréquence f_3 des étudiants ayant obtenu 3 est 12.5%, leur effectif est $n_5 = 5$.

   ⋮ ⋮ ⋮ ⋮

   ⋮ ⋮ ⋮ ⋮

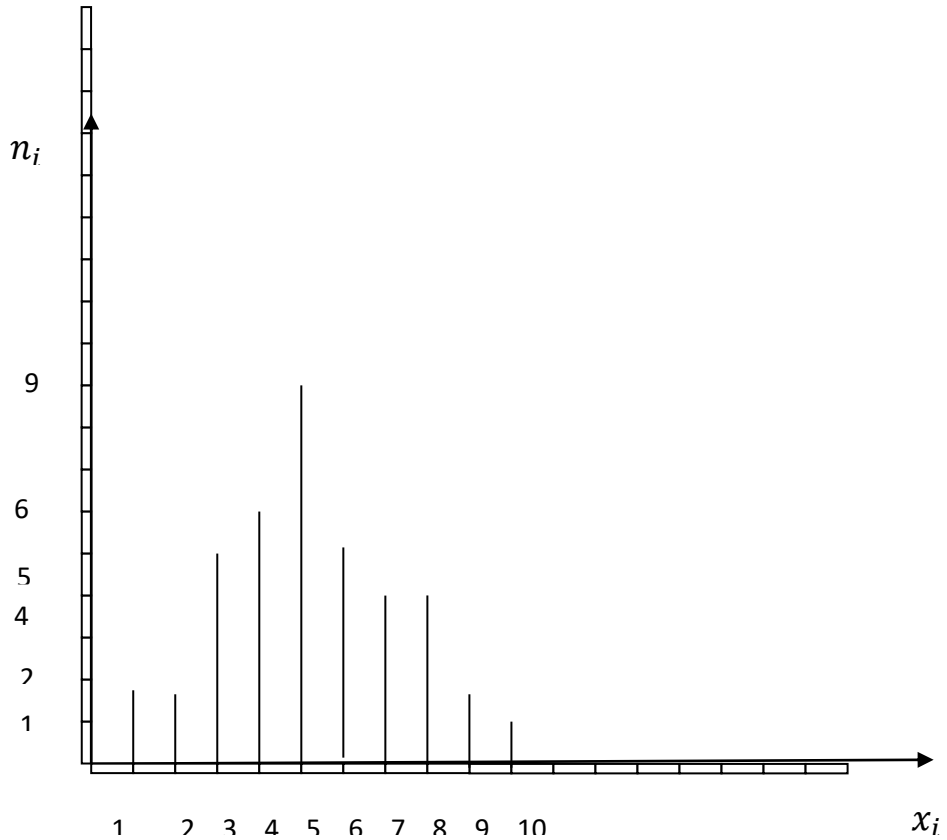
   ⋮ ⋮ ⋮ ⋮

   La fréquence f_{10} des étudiants ayant obtenu 10 est 2.5%, leur effectif est $n_{10} = 1$.

b) Représentation graphique

La représentation graphique d'une distribution d'une variable quantitative discrète est un diagramme en bâtons.

Les valeurs de la variable sont portées en abscisses et en ordonnées figurent les effectifs ou les fréquences.



1.2.2.3 Variable statistique quantitative continue

Définition 1.2.2 Un histogramme est un digramme formé d'un ensemble de rectangles contigües dont la base est déterminée par les extrémités de la classe et dont la surface doit être proportionnelle à l'effectif (ou à la fréquence) de celle-ci.

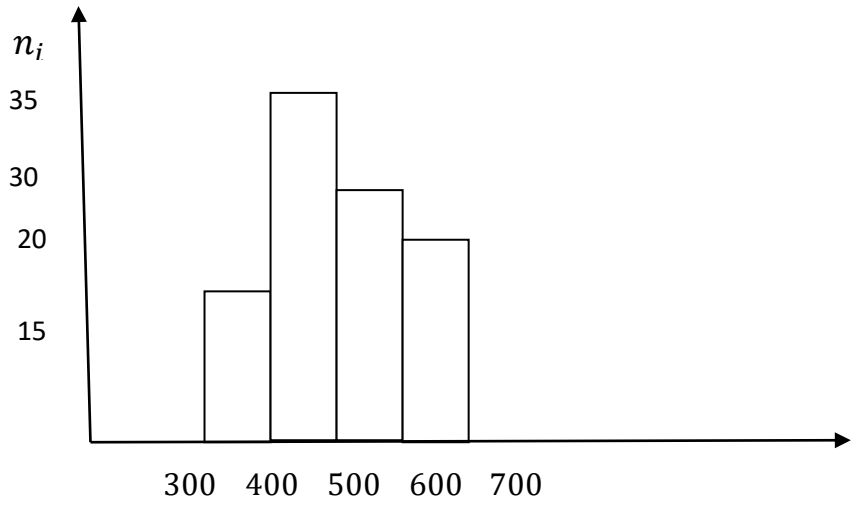
Si les classes sont toutes de même amplitude, il suffit pour réaliser l'histogramme de porter en ordonnée les effectifs ou les fréquences. En revanche, lorsqu'elles ne le sont pas, pour le réaliser sans risque d'erreur, on porte en abscisse les extrémités des classes et en ordonnée les effectifs par unité d'amplitude $\frac{n_i}{a_i}$ appelées densités que l'on note d_i, ou les fréquences par unité d'amplitude $\frac{f_i}{a_i}$, nommées densités de fréquence et notées d_i'.Ainsi la surface de chaque rectangle est $\frac{n_i}{a_i} \times a_i = n_i$ ou $\frac{f_i}{a_i} \times a_i = f_i$.

Exemple 1.2.3 cas où les amplitudes sont égales

Soit le tableau suivant donnant la durée de vie de 50 ampoules

Durée de vie D'une ampoule (hr.)	n_i	$f_i\%$
$[300;400[$	15	15
$[400;500[$	35	35
$[500;600[$	30	30
$[600;700[$	20	20

Où n_i les effectfs et $f_i\%$ les fréquences en pourcentage.

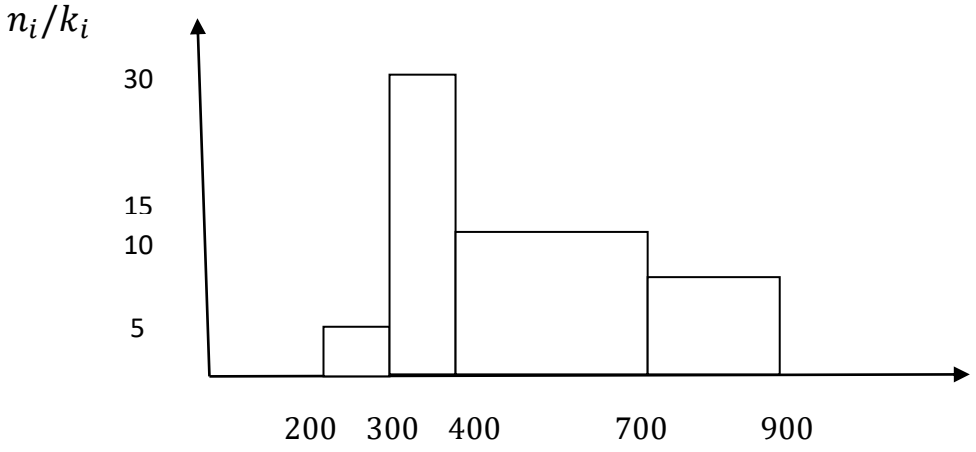


Exemple 1.2.4 cas où les amplitudes sont différentes

Soit le tableau suivant donnant la durée de vie de 50 ampoules

Durée de vie D'une ampoule (hr.)	n_i	$f_i\%$	Ampl a_i	$k_i = \frac{a_i}{a}$	$\frac{n_i}{k_i}$
$[200;300[$	5	5	100	1	5
$[300;400[$	30	30	100	1	30
$[400;700[$	45	45	300	3	15
$[700;900[$	20	20	200	2	10

Où n_i les effectfs, $f_i\%$ les fréquences en pourcentage, a_i les amplitudes des classes et a est la plus petite amplitude,



1.3 Courbes Cumulatives

1.3.1 variable quantitative discrète

Définition 1.3.1 L'effectif cumulé croissant associé à la modalité x_i, noté N_i est égal au nombre d'individus dont la valeur de la variable statistique est inférieur ou égal à x_i.

Définition 1.3.2 La fréquence cumulée croissante associée à la modalité x_i, noté F_i est égale à la proportion d'individus dont la valeur de la variable statistique est inférieur ou égal à x_i.

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

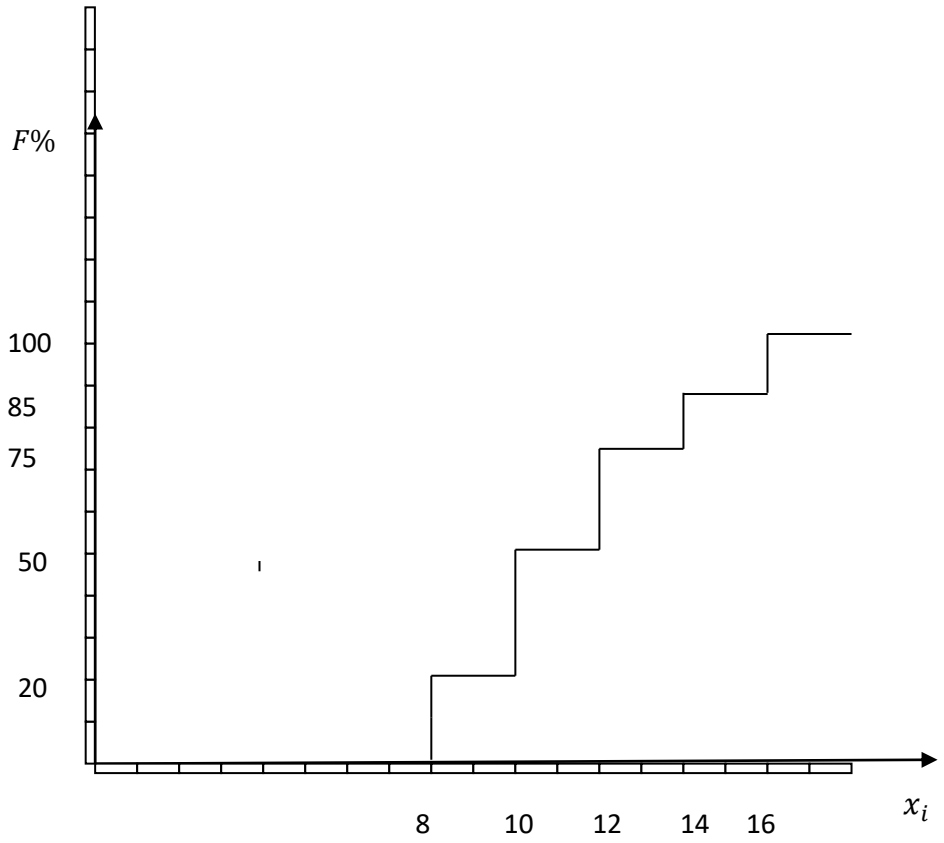
Définition 1.3.3 La courbe cumulative des effectifs est la courbe obtenue en traçant pour chaque point (x_i, N_i) dans un système d'axes orthogonaux un segment parallèle à l'axe des abscisses et un autre parallèle à l'axe des ordonnées.

Définition 1.3.4 La courbe cumulative des fréquences est la courbe obtenue en traçant pour chaque point (x_i, F_i) dans un système d'axes orthogonaux un segment parallèle à l'axe des abscisses et un autre parallèle à l'axe des ordonnées.

Exemple 1.3.1 Soit le tableau donnant les notes de 20 étudiants ayant participé à une épreuve de statistiques

Déterminons les effectifs et les fréquences cumulés.

x_i(note)	n_i	f_i%	F_i%	N_i
8	4	20	20	4
10	6	30	50	10
12	5	25	75	15
14	2	10	85	17
16	3	15	100	20
Total	n =20	100%		



1.3.2 variable quantitative continue

Définition 1.3.5 La courbe cumulative des fréquences est la ligne polygonale qui joint, dans un système d'axes orthogonaux, des points (e_i, F_i).

Exemple 1.3.2

Soit le tableau suivant donnant la durée de vie de 50 ampoules

Déterminons les effectifs cumulés et les fréquences cumulées de l'exemple 1.2.3.

Durée de vie D'une ampoule (hr.)	n_i	$f_i\%$	F_i	N_i
$[300;400[$	15	15	15	15
$[400;500[$	35	35	50	50
$[500;600[$	30	30	80	80
$[600;700[$	20	20	100	100

Où n_i les effectfs et $f_i\%$ les fréquences en pourcentage, F_i les fréquences cumulatives et N_i les effectifs cumulatifs.

Si $x \in [e_i; e_{i+1}[$ alors $F(x) = F(e_i) + \frac{F(e_{i+1})-F(e_i)}{e_{i+1}-e_i}(x - e_i)$

Exemple 1.3.3

Soit le tableau suivant donnant la répartition de 66 étudiants selon leurs revenus.

Classe(D.A)	Centre de la classe $\bar{x_i}$	n_i	$f(x)\%$	$F(x)\%$	N_i
$[0; 2000[$	1000	7	10.6	10.6	7
$[2000; 4000[$	3000	19	28.8	39.4	26
$[4000; 6000[$	5000	14	21.2	60.6	40
$[6000; 8000[$	7000	9	13.7	74.3	49
$[8000; 10000[$	9000	6	9.1	83.4	55
$[10000; 12000[$	11000	6	9.1	92.5	61
$[12000; 14000[$	13000	2	3.0	95.5	63
$[14000; 16000[$	15000	3	4.5	100	66
Total		66	100		

1) Déterminer $F(6000)$.

D'après ce tableau, $F(6000) = 60.6\%$.

2) Déterminer $F(7500)$.

$7500 \in [6000; 8000[$ alors d'après la formule ci-dessus on a :

$$F(7500) = F(6000) + \frac{F(8000) - F(6000)}{8000 - 6000}(7500 - 6000)$$

$$F(7500) = 60.6 + \frac{74.3 - 60.6}{8000 - 6000}(7500 - 6006)$$

$$F(7500) = 60.6 + \frac{13.7}{2000}1500 = 70.88\%$$

Remarque. De nombreux calculs pratiques s'effectuent à l'aide de $F(x)$. Par exemple, si on s'intéresse au pourcentage d'étudiants dont le revenu est compris entre deux limites x_1 et x_2 ou supérieur à x_3.

Alors dans le premier cas le pourcentage d'étudiants est $F(x_2) - F(x_1)$ et dans le deucième cas est $1 - F(x_3)$

CHAPITRE II

CARACTERISTIQUES DE POSITION

ET DE DISPERSION

2.1 Introduction

Les caractéristiques de position sur l'ordre de grandeur de la variable statistique Elles se situent entre la plus petite et la grande valeurs des observations.

Les principales caractéristiques de position sont : la moyenne, la médiane, le mode et les quantiles.

2.2 Moyenne

1er cas : Observations non groupées.

Soit $x_1, x_2, x_3, \dots, x_n$ n observations, alors leur moyenne est donnée par :

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \tag{2.1}$$

Exemple 2.2.1

Soit les données suivantes : 46, 54, 42, 46, 32, alors leur moyenne est

$$\bar{x} = \frac{46 + 54 + 42 + 46 + 32}{5} = \frac{220}{5} = 44$$

2ème cas : Observations groupées.

Soit r groupes de données :

x_i(note)	n_i	f_i%
x_1	n_1	f_1
x_2	n_2	f_2
⋮	⋮	⋮
⋮	⋮	⋮
⋮	⋮	⋮
x_e	n_r	f_r

Alors

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

Ou

$$\overline{x} = \sum_{i=1}^{r} f_i x_i \qquad (2.2)$$

Exemple 2.2.2

Soit les données suivantes 46, 54, 42, 46, 32. Regroupons les dans le tableau suivant :

x_i(note)	n_i	f_i%
32	1	20
42	1	20
46	2	40
64	1	20

Alors

$$\bar{x} = \frac{\sum_{i=1}^{r} n_i x_i}{n} = \frac{32+42+46+64}{5} = \frac{220}{5} = 44$$

$$\overline{x} = \sum_{i=1}^{4} f_i x_i = 0.2 * 32 + 0.2 * 42 + 0.4 * 46 + 0.2 * 64 = 44$$

Remarque. Dans le cas où la variable statistique est continue, on remplace dans (2.2) x_i par $\overline{x_\iota}$ le centre de la classe i.

Exemple 2.2.3

Le tableau suivant donne la répartition de 81 salariés selon leurs salaires journaliers.

Salaire $x_i(\ D.A)$	$\overline{x_\iota}$	n_i	f_i
$[400;450[$	425	15	0.185
$[450;500[$	475	20	0.246
$[500;550[$	525	25	0.308
$[550;600[$	575	10	0.123
$[600;6500[$	625	11	0.135
Total		81	1

La moyenne des salaires est :

$$\bar{x} = \frac{\sum_{i=1}^{r} n_i x_i}{n} = \frac{15 * 425 + 20 * 475 + 25 * 525 + 10 * 575 + 11 * 625}{81} = \frac{41625}{81} = 513.89$$

2.3 Médiane

La médiane est une autre caractéristique de position pour une variable statistique.

Définition 2.3.1 La médiane est la valeur de la variable statistique telle qu'il y ait autant s'observations au dessous d'elle qu'au dessus. C'est-à-dire, la valeur correspondant à 50% des observations.

2.3.1 Calcul de la médiane dans le cas discret

Pour calculer la moyenne on doit

1) Classer (ordonner) les observations selon l'ordre croissant ou décroissant.
2) Si le nombre d'observations n est impair, alors la médiane est donnée par :

$$m_e = x_p, \quad \text{où } p = \frac{n+1}{2} \tag{2.3}$$

Sinon elle est donnée par :

$$m_e = \frac{x_p + x_{p+1}}{2}, \text{ où } p = \frac{n}{2} \tag{2.4}$$

Exemple 2.3.1

Calculons la médiane des données suivantes :

3, 4, 4, 5, 7, 10, 6, 8, 6

1) Ordonnons ces nombre selon l'ordre croissant :

3, 4, 4, 5, 6, 6, 7, 8, 10

2) $n = 9$ est impair donc

$$m_e = x_p, \text{ où } p = \frac{n+1}{2} = \frac{9+1}{2} = 5$$

$$m_e = x_5 = 6$$

Exemple 2.3.2

Calculons la médiane des données suivantes :

9, 3, 4, 7, 8, 7, 5, 2

1) Ordonnons ces nombre selon l'ordre croissant :

2, 3, 4, 5, 7, 7, 8, 9

2) $n = 8$ est pair donc

$$m_e = \frac{x_p + x_{p+1}}{2}, \text{ où } p = \frac{n}{2} = \frac{8}{2} = 4$$

$$m_e = \frac{x_4 + x_5}{2} = \frac{5+7}{2} = 6$$

2.3.2 Calcul de la moyenne dans le cas continue

Puisque la médiane est la valeur de la variable statistique correspondant à 50% de données, il suffit donc de déterminer $m_e$telle que $F(m_e) = 0.5$ où $m_e = F^{-1}(0.5)$.

Dans ce cas, m_e est donnée par :

$$m_e = x_j + \frac{\left(F(m_e) - F(x_j)\right)}{\left(F(x_{j+1}) - F(x_j)\right)}\left(x_{j+1} - x_j\right) \qquad (2.5)$$

x_j et x_{j+1} désignent les limites inférieur et supérieur de la classe dans laquelle se trouve la médiane.

Exemple 2.3.3

Le tableau suivant donne la répartition de 81 salariés selon leurs salaires journaliers.

Salaire $x_i(D.A)$	$\overline{x_i}$	n_i	$f_i\%$	$F_i\%$
$[400;450[$	425	15	18.51	18.51
$[450;500[$	475	20	24.69	43.2
$[500;550[$	525	25	30.86	74.06
$[550;600[$	575	10	12.34	86.4
$[600;6500[$	625	11	13.58	100
Total		81	100%	

D'après ce tableau, la médiane $m_e \in [500;550[$. Donc de (6) on a :

$$m_e = x_j + \frac{\left(F(m_e) - F(x_j)\right)}{\left(F(x_{j+1}) - F(x_j)\right)}\left(x_{j+1} - x_j\right) = 500 + \frac{(0.5 - 0.432)}{(0.741 - 0.432)}(550 - 500)$$

$$m_e = 500 + \frac{0.068}{0.309} \times 50 = 511$$

Remarque. Lorsque un ensemble de données contient des valeurs extrêmes, la médiane est souvent une mesure préférable de la tendance centrale.

2.4 Mode

Définition 2.4.1 Le mode correspond à la valeur de l'observation qui a la plus grande fréquence

Exemple 2.4.1

Soit les données suivantes ; 46, 54, 42, 46, 32

46 apparait deux fois donc le mode est 46.

Dans le cas continue, on parle de classe modale au lieu de mode. Elle le détermine de la manière suivante :

1) Si les classes sont toutes de même amplitude, la classe modale est la classe d'effectif (fréquence) le plus élevé.
2) Si les classes sont d'amplitudes différentes, on calcule pour chaque le nombre $d_i = \frac{n_i}{a_i}$ appelé densité d'effectif. Alors la classe de densité d'effectif la plus élevée est la classe modale.

Exemple 2.4.1

La classe modale de l'exemple 2.3.3 est $[500; 550[$ car la fréquence la plus élevée est30.86%.

Remarque. IL est possible que plusieurs valeurs apparaissent avec la même fréquence et que cette fréquence est la plus élevée. Dans ce cas, plus d'un mode existe.

Si les données ont exactement deux modes, on dit que les données sont bimodales. Si elles sont plus de deux modes, on dit qu'elles sont multimodales.

Exemple 2.4.2

Soit les données suivantes: 3, 4 , 4, 5, 6, 6, 7, 8,10.

La valeur 4 apparait deux fois.

La valeur 6 apparait deux fois.

Les autres valeurs apparaissent chacune une fois

Donc les valeurs 4 et 6 sont deux modes de ces données. Dans ce cas, on dit que ces donnes sont bimodales.

Exemple 2.4.3

. Soit les données suivantes: 3, 4 , 4, 5, 5, 6, 6, 7, 8.

Les valeurs 4, 5 et 6 apparaissent chacune deux fois et les autres apparaissent chacune une fois. Donc les valeurs 4, 5 et 6 sont trois modes de ces données.

Exemple 2.4.4

Soit les données représentant les notes de 20 étudiants :

14, 16, 12, 12, 8, 10, 12, 8, 16, 10

16, 10, 12, 8, 12, 8, 10, 14, 10, 10

x_i(note)	n_i	f_i%
8	4	20
10	6	30
12	5	25
14	2	10
16	3	15

La valeur 10 est le mode de ces données car sa fréquence (30) est la plus élevée.

2.5 Quantiles

Définition 2.5.1 Le quantile d'ordre α%, noté, q_α , est la valeur de la variable statistique telle que α% des valeurs observées sont inférieures ou égales àq_α.

Le quantile est déterminé soit graphiquement, soit analytiquement.

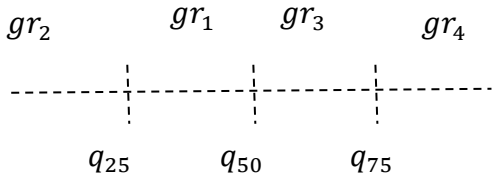
a) Graphiquement, le quantile d'ordre α%, q_α , est l'abscisse d'un point de la courbe cumulative des fréquences dont l'ordonnées est α% . C'est-à-dire, q_α est solution de l'équation $F(q_\alpha) = \alpha$.
b) Analytiquement, le quantile d'ordre α%, q_α , est donné par

$$q_\alpha = x_j + (x_{j+1} - x_j)\frac{(F(q_\alpha) - F(x_j))}{(F(x_{j+1}) - F(x_j))}$$

2.5.1 Principaux Quantiles

Les quantiles les plus utilisés sont les quartiles, les déciles et les centiles.

1) Les quartiles sont les trois valeurs de la variable statistique, notées q_{25}, q_{50} et q_{75} qui partagent en quatre groupes d'effectifs égaux. 25% (resp. 50% et 75%) des individus de la population ont une valeur de la variable inférieure ou égale à q_{25}(resp. q_{50} et q_{75}).



2) Les déciles sont les neufs valeurs de la variable statistique, notées $q_{10}, q_{20,} \ldots, q_{90}$ qui partagent les observations en dix groupes d'effectifs égaux :10% des observations sont inférieures ou égales à q_{10}, 20% sont inférieures ou égales à$q_{20,} \ldots$, 90% sont inférieures ou égales àq_{90}
3) Les centiles sont les 99 valeurs de la variable statistique, notées $q_1, q_{2,} \ldots, q_{99}$ qui partagent les observations en cent groupes d'effectifs égaux :1% des observations sont inférieures ou égales à q_1, 2% sont inférieures ou égales à $q_{2,} \ldots$, 99% sont inférieures ou égales à$q_{99}\%$

Remarque.

Le quantile d'ordre $\alpha\%$ peut aussi être défini comme la valeur de la variable qui partage la population en deux sous populations, telles que dans la première il y'a $n \times \alpha\%$ individus et dans la seconde $n \times (100 - \alpha)\%$ individus.

L'effectif cumulé croisant associé au quantile d'ordre $\alpha\%$ est donc égal à $n \times \alpha\%$, soit $N(q_\alpha) = n \times \alpha\%$.

2.5.2 Détermination des quantiles

1) **Cas d'une variable statistique discrète**

Si les données sont classées (ordonnées) par ordre croissant ou décroissant alors le quantile d'ordre $\alpha\%$, q_α, est la valeur de la variable statistique telle que $\alpha\%$ des valeurs sont inférieures ou égales à q_α.

Exemple 2.5.1

Soit les données :

3310, 3355, 3450, 3480, 3480, 3490, 3520, 3540, 3550, 3650, 3730, 3925.

Calculons le 85ème centile.

Rappelons que le $p^{\text{ème}}$ centile est la valeur telle que $p\%$ des observations sont inférieures ou égales à cette valeur.

Pour calculer le $p^{\text{ème}}$ centile, on procède de la manière suivante :

a) Classer les données selon l'ordre croissant (décroissant)
b) Calculer

$$i = n \times \frac{p}{100}$$

Où p est le centile considéré n est le nombre d'observations.

c) Si n n'est pas un entier, alors le $p^{ème}$ centile est la valeur dont la position correspond au plus petit entier supérieur à i, sinon il est la valeur moyenne des observations$i$ et $i+1$

Dans notre cas, $i = n \times \frac{p}{100} = 12 \times \frac{85}{100} = 10.2$.

Puisqu'il n'est pas un entier, la position du 85ème centile est 11. Donc le 85ème centile est 3730.

Calculons le 50ème centile qui correspond à la médiane.

$$i = n \times \frac{p}{100} = 12 \times \frac{50}{100} = 6$$

Puisqu'il est entier, le 50ème centile est la moyenne des observations 6 et 7. C'est-à-dire,

$$q_{50} = \frac{3490+3520}{2} = 3505$$

Calculons q_{25} et q_{75} (le 1er quartile et le 3ème quartile)

Pour q_{25}

On a :

$i = n \times \frac{p}{100} = 12 \times \frac{25}{100} = 3$.Puisque i est un entier, alors le 1er quartile est la moyenne des 3ème et 4ème observations :

$$q_{25} = \frac{3450+3480}{2} = 3465$$

Pour q_{75}

On a :

$i = n \times \frac{p}{100} = 12 \times \frac{75}{100} = 9$.Puisque i est un entier, alors le 3ème quartile est la moyenne des 9ème et 10ème observations :

$$q_{25} = \frac{3550+3650}{2} = 3600$$

Cas d'une variable statistique continue

a) Puisque q_α est la valeur de la variable statistique qui correspond à $\alpha\%$ des données qui lui sont inférieures ou égales, il suffit de déterminer q_α tel que $F(q_\alpha) = \alpha$.

$$q_\alpha = x_j + (x_{j+1} - x_j)\frac{(F(q_\alpha) - F(x_j))}{(F(x_{j+1}) - F(x_j))}$$

x_j et x_{j+1}désignent les limites inférieure et supérieure de la classe dans laquelle se trouve q_α.

Exemple 2.5.2

Soit le tableau suivant donnant la répartition de 66 étudiants selon leurs revenus.

Classe(D.A)	Centre de la classe $\overline{x_i}$	n_i	$f(x)\%$	$F(x)\%$	N_i
$[0;2000[$	1000	7	10.6	10.6	7
$[2000;4000[$	3000	19	28.8	39.4	26
$[4000;6000[$	5000	14	21.2	60.6	40
$[6000;8000[$	7000	9	13.7	74.3	49
$[8000;10000[$	9000	6	9.1	83.4	55
$[10000;12000[$	11000	6	9.1	92.5	61
$[12000;14000[$	13000	2	3.0	95.5	63
$[14000;16000[$	15000	3	4.5	100	66
Total		66	100		

Calculons q_{75}

L'inspection de ce tableau montre que q_{75}est légèrement supérieur à 8000D.A puisque 74.3% de étudiants on un revenu inférieur 8000D.A. Donc $q_{75} \in [8000;10000[$. Par conséquent $x_j = 8000$ et $x_{j+1} = 10000$, $F(x_j) = 74.3\%$, $F(x_{j+1}) = 83.4\%$ et $F(q_{75}) = 75\%$.

$q_{75} = x_j + (x_{j+1} - x_j)\frac{(F(q_{75})-F(x_j))}{(F(x_{j+1})-F(x_j))}$=8000+$(10000 - 8000)\frac{(75-74.3)}{(83.4-74.3)} = 8140$

Donc 75% des étudiants ont un revenu inférieur à 8140.

Calcul du quantile d'ordre 30% (ou 3ème décile).

D'après ce tableau $q_{30} \in [2000;4000[$. Par conséquent $x_j = 2000$ et $x_{j+1} = 4000$, $F(x_j) = 10.6\%$, $F(x_{j+1}) = 39.4\%$ et $F(q_{30}) = 30\%$

$q_{30} = x_j + (x_{j+1} - x_j)\frac{(F(q_{30})-F(x_j))}{(F(x_{j+1})-F(x_j))}$=2000+$(4000 - 2000)\frac{(30-10.6)}{(39.4-10.6))}$

$$q_{30} = 3347$$

Donc 30% des étudiants ont un revenu inférieur à 3347.

2.6 Caractéristiques de dispersion

2.6.1 Etendue

Définition 2.6.1 L'étendue représente la différence entre les valeurs extrêmes.

$$e = x_n - x_1$$

Exemple 2.6.1

Soit les données: -10, -9, -8, 8, 9, 10 alors l'étendue est

$$e = 10 - (-10) = 20$$

2.6.2 Intervalle interquartile

Définition 2.6.2 L'intervalle interquartile, noté *I*, est la différence entre les deux quartiles q_{25} et q_{75}

$$I = q_{75} - q_{25}$$

Exemple 2.6.2

L'intervalle interquartile des données de l'exemple 2.6.1 est

$$I = q_{75} - q_{25} = 8140 - 3000 = 5140$$

2.6.3 Variance et Ecart type

Définition 2.6.3

a) La variance des données non groupées: $x_1, x_2, \dots, x_n$ est le nombre s^2 donné par

$$s^2 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n}$$

On peut aussi l'écrire:

$$s^2 = \sum_{i=1}^{n} \frac{{x_i}^2}{n} - \bar{x}^2$$

b) La variance des données groupées: $x_1, x_2, \dots, x_r$ est le nombre s^2 donné par

$$s^2 = \sum_{i=1}^{r} f_i \, (x_i - \bar{x})^2$$

Où $f_i = \frac{n_i}{n}$ est la fréquence et r est le nombre de groupes.

On peut aussi l'écrire :

$$s^2 = \sum_{i=1}^{n} f_i \, {x_i}^2 - \bar{x}^2$$

c) L'écart type des données dans les deux cas est le nombre s.

Exemple 2.6.3

Soit les données non groupées suivantes : 46, 54, 42, 46, 32, alors leur moyenne est

$$\overline{x} = \frac{46 + 54 + 42 + 46 + 32}{5} = \frac{220}{5} = 44$$

Et leur variance est

$$s^2 = \sum_{i=1}^{5} \frac{(x_i - \bar{x})^2}{5} = \frac{(46-44)^2 + (54-44)^2 + (42-44)^2 + (46-44)^2 + (32-44)^2}{5}$$

$$s^2 = \frac{4 + 100 + 4 + 4 + 144}{5} = 51.2$$

l'écart type est $s = 7.16$

Exemple 2.6.4

Regroupons les données de l'exemple 2.6.3 dans le tableau suivant :

x_i(note)	n_i	f_i%
32	1	20
42	1	20
46	2	40
54	1	20

Alors

$$\bar{x} = \sum_{i=1}^{4} f_i x_i = 0.2 * 32 + 0.2 * 42 + 0.4 * 46 + 0.2 * 64 = 44$$

$s^2 = \sum_{i=1}^{n} f_i \, {x_i}^2 - \bar{x}^2 = 0.2 * 32^2 + 0.2 * 42^2 + 0.4 * 46^2 + 0.2 * 54^2 - 44^2 = 51.2$

l'écart type est $s = 7.16$