Statistique descriptive

Remita Nawel
Université Badji Mokhtar, Faculté des Sciences, TC SNV 1

2016-2017

Il n'existe pas de définition précise de la statistique mais, pour nous, nous allons utiliser la définition suivante :
On appelle statistique l'ensemble des méthodes ou des techniques permettant d'analyser ou de traiter des ensembles d'observations que nous appellerons données.
Les méthodes utilisées relèvent des mathématiques et font largement appel à l'outil informatique pour leur mise en œuvre.
Remarque : Il ne faut pas confondre la statistique qui est la science qui vient d'être définie et une statistique qui est un ensemble de données chiffrées sur un sujet précis.

Nous distinguerons trois phases importantes dans l'évolution de la statistique :

- De l'antiquité et jusqu'à la fin du 19ème siècle, la statistique est restée principalement un ensemble de techniques de dénombrement.
- De la fin du 19ème siècle aux années 1960 s'est construit la statistique mathématique surtout grâce à l'école anglaise (K. Pearson, W. Gosset (Student), R. Fisher, J. Neyman, ...).
- Depuis les années 1960, et avec le développement des outils informatiques et graphiques, la statistique a connu un développement considérable.

- **Population :** ensemble concerné par une étude statistique (noté Ω).
- **Individu** : (unité statistique) : tout élément de la population (noté $\omega \in \Omega$).
- **Echantillon** : sous ensemble de la population sur lequel sont réalisées les observations.
- **Enquête** : Opération consistant à observer (mesurer, questionner, ...) l'ensemble des individus d'un échantillon.
- **Recensement** : enquête dans laquelle l'échantillon observé est la population toute entière (enquête exhaustive).
- **Sondage** : enquête dans laquelle l'échantillon observé est un sous ensemble strict de la population (enquête non exhaustive).

- **Caractère** : c'est une caractéristique définie sur la population et observée sur l'échantillon. Le caractère peut être :
  - Qualitatif nominal (sexe, profession, situation familiale, ...)
  - Qualitatif ordinal (grade militaire, grade dans l'enseignement supérieur, ...)
  - Quantitatif discret (nombre d'enfants, nombre de chambre dans un appartement, ...) appelé aussi **variable statistique discrète.**
  - Quantitatif continu (taille, âge, vitesse, poids, taux, ...) appelé aussi **variable statistique continue.**
- **Modalités** : les différentes valeurs prises par chaque caractère.

Pour mener à bien une étude statistique il est souhaitable de suivre la démarche suivante :
- Tout d'abord on commence par collecter les observations brutes qu'on regroupe dans un tableau appelé tableau de données.
- Dans beaucoup de situations les observations peuvent être différentes et en nombre important, et la tableau de données n'est pas facilement exploitable, pour cela on synthétise les observations en les regroupant dans le tableau statistique.
- On complète l'étude en faisant des représentations graphiques et des calculs numériques afin de donner une bonne interprétaion des données observées.

Schématiquement la démarche statistique est résumée comme suit :

$\boxed{\textbf{Tableaux de données}}$

Synthétiser

$\boxed{\textbf{Tableaux statistiques}}$

Visualiser

$\boxed{\textbf{Représentations graphiques}}$

Résumer

$\boxed{\textbf{Résumés numériques}}$

Exemple 1. L'étude de la situation familliale de 30 employés d'une entreprise est résumée dans le tableau suivant :

01	**02**	**03**	**04**	**05**	**06**	**07**	**08**	**09**	**10**
C	M	C	C	D	V	M	M	C	D
11	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
C	M	C	C	M	D	C	C	M	M
21	**22**	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**
C	C	M	C	V	C	C	M	C	C

Exemple 2. On étudie le nombre d'enfants dans une famille dans une cité habitée par 100 familles. On a obtenu les résultats suivants :

01	**02**	**03**	**04**	**...**	**98**	**99**	**100**
2	0	4	2	...	6	3	1

Exemple 3. On étudie le poids en Kg de 150 nouveaux né dans une maternité. On a obtenu les résultats suivants :

01	**02**	**03**	**04**	**...**	**148**	**149**	**150**
2.356	3.102	2.212	4.125	...	3.256	3.894	2.352

Soit X le caractère étudié,
n la taille de l'échantillon,
k le nombre de modalités du caractère X ,
$x_1, \cdots, x_k$ les modalités d'effectifs respectifs $n_1, \cdots, n_k$.
Fréquence de la modalité x_i est $f_i = \frac{n_i}{n}$.
Pourcentage de la modalité x_i est $p_i = \frac{n_i}{n} \times 100$.
Effectif cumulé croissant à la modalité x_i est $n_i^c \nearrow = n_1 + n_2 + \cdots + n_i$.
Effectif cumulé décroissant à la modalité x_i est
$n_i^c \searrow = n_k + n_{k-1} + \cdots + n_i$.
La série statistique est formée par l'ensemble des données $(x_i)_{i \leq k}$.
La présentation d'une série statistique se fait par le biai d'un tableau, cette présentation diffère suivant la nature du caractère étudié.

Exemple 1. On reprend l'exemple sur la situation familliale de 30 employés.

X : Situation familliale	Effectif n_i	Fréquence f_i	Pourcentage p_i
C	16	0.5333	53.33
M	9	0.3000	30
D	3	0.1000	10
V	2	0.0667	6.67
Total	30	1	100

Exemple 2. On reprend l'exemple sur la distribution des familles suivant le nombre d'enfants.

X : Nombre d'enfants	Effectif n_i	Fréquence f_i	% p_i	$n_i^c \nearrow$	$n_i^c \searrow$
0	11	0.11	11	11	100
1	16	0.16	16	27	89
2	21	0.21	21	48	73
3	25	0.25	25	73	52
4	17	0.17	17	90	27
5	8	0.08	8	98	10
6 et plus	2	0.02	2	100	2
Total	100	1	100	—	—

Exemple 3. On reprend l'exemple sur la distribution des nouveaux né suivant leur poids.
On ordonne les données par ordre croissant et on calcule la longueur de la série statistique :

$$\textbf{étendue} = e = x_n - x_1$$

x_1 est la plus petite observation
x_n est la plus grande observation
n étant le nombre d'observations.

Pour construire le tableau statistique on doit regrouper les données dans des intervalles qu'on appelle **classes**, pour cela on doit d'abord déterminer les nombre de classes défini par

$$n_{classes} = \begin{cases} \sqrt{n} \text{ si } n < 50 \\ 1 + \frac{10}{3} \log_{10} n \text{ si } n \geq 50 \end{cases}$$

Par la suite on détermine la longueur des classes qu'on appelle **amplitude** telle que

$$a = \frac{e}{n_{classes}}.$$

Pour notre exemple on a $x_1 = 2.212$ et $x_n = 4.593$ et $n \geq 50$, alors

$$e = 4.593 - 2.212 = 2.381$$

$$n_{classes} = 1 + \frac{10}{3}\log_{10} 150 = 8.2536 \approx 8$$

$$a = \frac{e}{n_{classes}} = \frac{2.381}{8} \approx 0.2976 \approx 0.3$$

Tableau statistique

Variable continue (Caractère quantitatif continu)

X : Nombre d'enfants	Centre x_i	Effectif n_i	Fréquence f_i	% p_i	$n_i^c \nearrow$	$n_i^c \searrow$
$[2.2; 2.5[$	2.35	5	0.0333	3.33	5	150
$[2.5; 2.8[$	2.65	11	0.0733	7.33	16	145
$[2.8; 3.1[$	2.95	21	0.1400	14	37	134
$[3.1; 3.4[$	3.25	39	0.2600	26	76	113
$[3.4; 3.7[$	3.55	35	0.2333	23.33	111	74
$[3.7; 4.0[$	3.85	20	0.1333	13.33	131	39
$[4.0; 4.3[$	4.15	13	0.0867	8.67	144	19
$[4.3; 4.6[$	4.45	6	0.0400	4	150	6
Total	—	150	1	100	—	—

Exemple 1. On reprend l'exemple sur la situation familliale de 30 employés

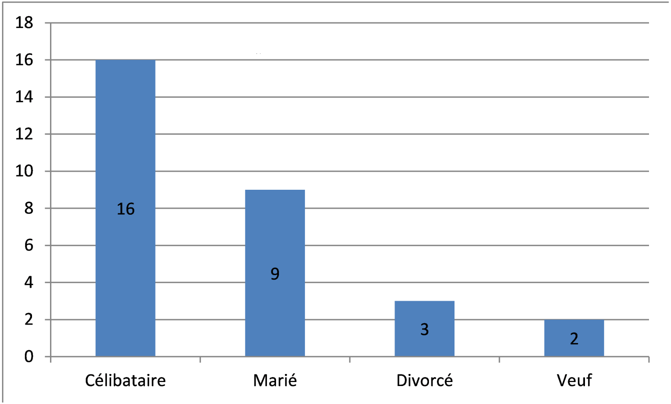


Figure: Diagramme à bandes

Diagramme circulaire

Secteur pour "Célibataire"

$$\begin{array}{c} \theta_C \longrightarrow 16 \\ 360 \longrightarrow 30 \end{array} \Longrightarrow \theta_C = \frac{16 \times 360}{30} = 192^\circ$$

Secteur pour "Marié"

$$\begin{array}{c} \theta_M \longrightarrow 9 \\ 360 \longrightarrow 30 \end{array} \Longrightarrow \theta_M = \frac{9 \times 360}{30} = 108^\circ$$

Diagramme circulaire

Secteur pour "Divorcé"

$$\begin{array}{c} \theta_D \longrightarrow 3 \\ 360 \longrightarrow 30 \end{array} \implies \theta_D = \frac{3 \times 360}{30} = 36^\circ$$

Secteur pour "Veuf"

$$\begin{array}{c} \theta_V \longrightarrow 2 \\ 360 \longrightarrow 30 \end{array} \implies \theta_V = \frac{2 \times 360}{30} = 24^\circ$$

Représentation graphique

Caractère qualitatif (distribution des effectifs)

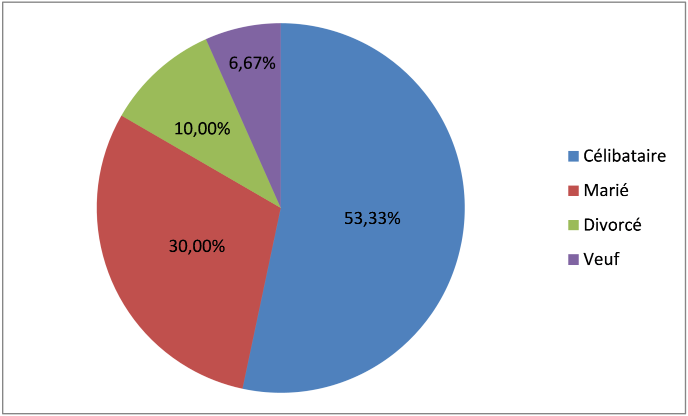


Figure: Diagramme circulaire

Exemple 2. On reprend l'exemple sur la distribution des familles suivant le nombre d'enfants.

X : Nombre d'enfants	Effectif
0	11
1	16
2	21
3	25
4	17
5	8
6 et plus	2
Total	100

Représentation graphique

Variable discrète (Caractère quantitatif discret) (distribution des effectifs)

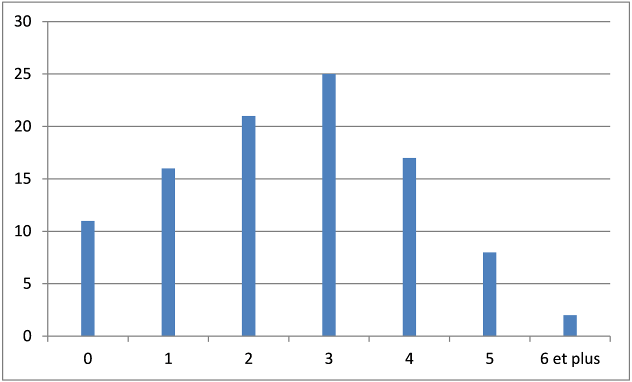


Figure: Diagramme en bâtons

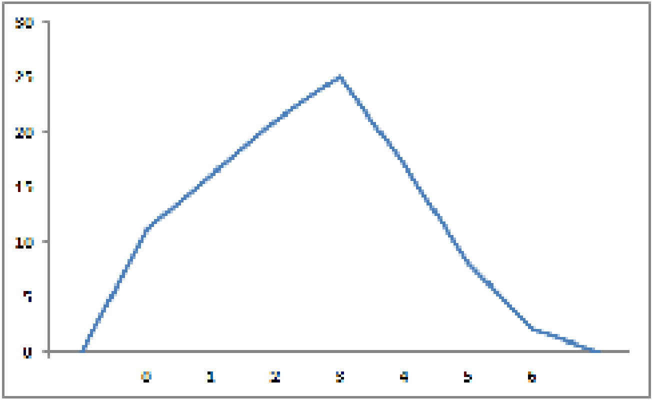


Figure: Polygône des effectifs

Exemple 3. On reprend l'exemple sur la distribution des nouveaux né suivant leur poids.

Histogramme (cas de classes de même amplitude)

X : Nombre d'enfants	Centre	Effectif
$[2.2; 2.5[$	2.35	5
$[2.5; 2.8[$	2.65	11
$[2.8; 3.1[$	2.95	21
$[3.1; 3.4[$	3.25	39
$[3.4; 3.7[$	3.55	35
$[3.7; 4.0[$	3.85	20
$[4.0; 4.3[$	4.15	13
$[4.3; 4.6[$	4.45	6
Total	—	150

Histogramme (cas de classes de même amplitude)

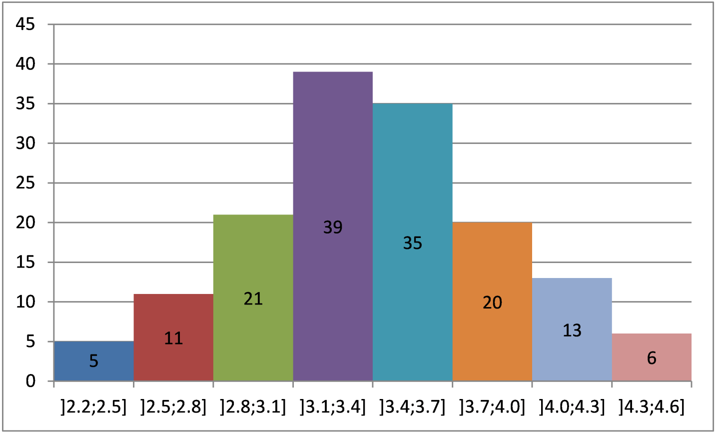


Figure: Histogramme

Représentation graphique

Variable continue (Caractère quantitatif continu) (distribution des effectifs)

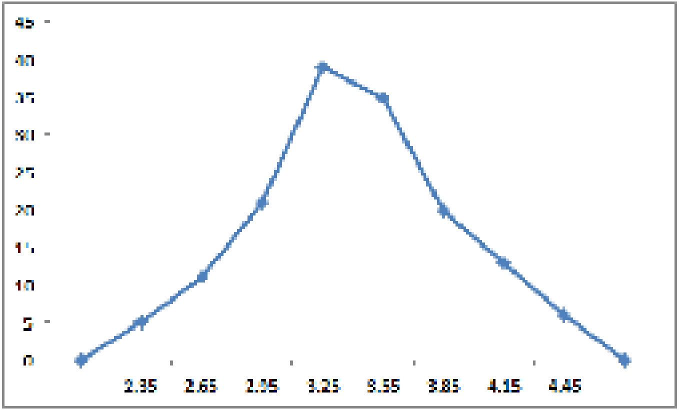


Figure: Polygône des effectifs

Exemple 2. On reprend l'exemple sur la distribution des familles suivant le nombre d'enfants.

X : Nombre d'enfants	Effectif cumulé croissant
0	11
1	27
2	48
3	73
4	90
5	98
6 et plus	100
Total	—

Représentation graphique

Variable discrète (distribution des effectifs cumulés croissants)

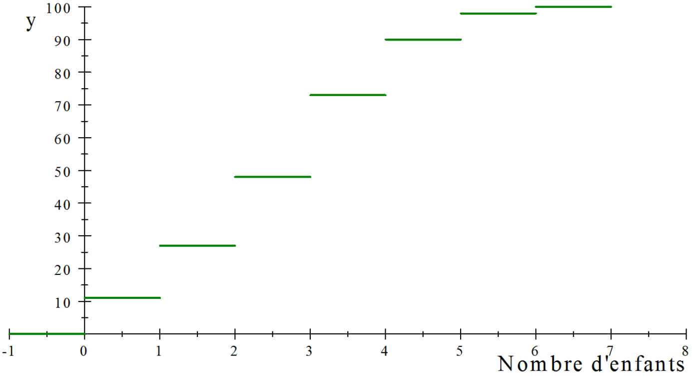


Figure: Courbe des effectifs cumulés croissants

Exemple 3. On reprend l'exemple sur la distribution des nouveaux né suivant leur poids.

Courbe des effectifs cumulés croissants

X : Nombre d'enfants	Centre	Effectif cumulé croissant
$[2.2; 2.5[$	2.35	5
$[2.5; 2.8[$	2.65	16
$[2.8; 3.1[$	2.95	37
$[3.1; 3.4[$	3.25	76
$[3.4; 3.7[$	3.55	111
$[3.7; 4.0[$	3.85	131
$[4.0; 4.3[$	4.15	144
$[4.3; 4.6[$	4.45	150
Total	—	—

Courbe des effectifs cumulés croissants

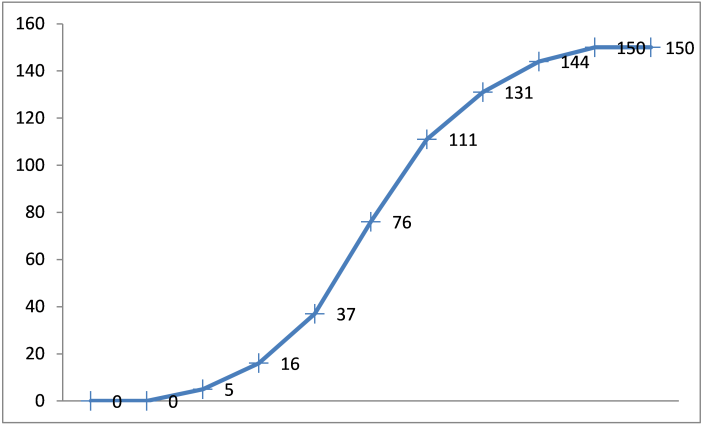


Figure: Courbe des effectifs cumulés croissants

L'ensemble des données d'une série statistique est d'un maniement lourd et difficile. Il s'impose donc de définir un ensemble de valeurs caractéristiques (ou paramètres) qui permettent une représentation condensée de l'information contenue dans la série statistique. On distingue deux catégories de valeurs typiques :

1. Les paramètres du 1^{er} ordre ou paramètres de position : **moyenne arithmétique**, **mode** et **médiane**.
2. Les paramètres du $2^{ème}$ ordre ou paramètres de dispertion : **écart-type**, **coefficient de variation** et **écart interquartile**.

Le mode, noté Mo, d'un caractère (qualitatif ou quantitatif) est la modalité la plus observée, c'est à dire celle qui a le plus grand effectif ou la plus grande fréquence. Le mode peut ne pas exister et s'il existe, ne pas être unique.

- L'ensemble 2, 2, 5, 7, 9, 9, 9, 10, 10, 11, 12, 18 a pour mode $Mo = 9$
- L'ensemble 3, 5, 8, 10, 12, 15, 16 n'a pas de mode.
- L'ensemble 2, 2, 2, 5, 7, 9, 9, 9, 10, 10, 11, 12, 18 a deux modes $Mo_1 = 2$ et $Mo_2 = 9$, on dit que c'est une série bimodale.
- Pour l'exemple 1 le mode est $Mo =$ Célibataire.
- Pour l'exemple 2 le mode est $Mo = 3$.
- Pour l'exemple 3 la classe modale est $[3,1;3,4[$ et le mode sera le centre de la classe c'est à dire $Mo = 3,250$.

La moyenne arithmétique d'une série statistique $x_1, x_2, \cdots, x_k$, d'un caractère quantitatif X, et d'effectifs respectifs $n_1, n_2, \cdots, n_k$ est donné par le nombre réel $\overline{X}$ défini par

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

où x_i est la modalité i de la variable X dans le cas d'une variable discrète et c'est le centre de la classe i dans le cas d'une variable continue et $n = \sum_{i=1}^{k} n_i$.

Exemple : On reprend l'exemple 2

X : Nombre d'enfants	Effectif	$n_i x_i$
0	11	0
1	16	16
2	21	42
3	25	75
4	17	68
5	8	40
6 et plus	2	12
Total	100	253

d'où $\overline{X} = \frac{253}{100} = 2,53$.

Soit une série statistique d'une variable X ayant pour modalités (ordonnées par ordre croissant) $x_1 < x_2 < \cdots < x_n$. On appelle médiane de X, le nombre Me, s'il existe, qui partage la série statistique en deux parties d'égal effectif.

Considérons les notes de 19 étudiants obtenues à l'examen de statistiques :

$$9, 15, 4, 8, 16, 10, 11, 10, 5, 12, 9, 10, 7, 12, 15, 11, 6, 13, 12.$$

On ordonne la série par ordre croissant, on a alors

$$4, 5, 6, 7, 8, 9, 9, 10, 10, 11, 11, 11, 12, 12, 12, 13, 15, 15, 16.$$

La moitié de la série est 9 alors

$$\underbrace{4, 5, 6, 7, 8, 9, 9, 10, 10}_{\text{9 valeurs}}, \boxed{11}, \underbrace{11, 11, 12, 12, 12, 13, 15, 15, 16}_{\text{9 valeurs}}.$$

la médiane est le nombre qui partage la série en deux parties de même efectif, la valeur qui réalise ceci est 11 donc $Me = 11$.

On ajoute maintenant la note 7 à la série précédente, on a alors une série de 20 notes :

$$4, 5, 6, 7, 8, 9, 9, 10, 10, 10, 11, 11, 11, 12, 12, 12, 13, 15, 15, 16.$$

La moitié de la série est 10 alors

$$\underbrace{4, 5, 6, 7, 7, 8, 9, 9, 10, 10}_{\text{10 valeurs}}, \underbrace{11, 11, 12, 12, 12, 13, 15, 15, 16}_{\text{10 valeurs}}.$$

la médiane est le nombre qui partage la série en deux parties de même efectif, elle se trouve entre le dernier 10 et le premier 11, dans ce cas on prendra la valeur moyenne de ces deux notes, alors $Me = \frac{10+11}{2} = 10,5$.

D'une manière générale soit X est une variable discrète prenant les valeurs ordonnées $x_1, x_2, \cdots, x_p, x_{p+1,\ldots}, x_n$, alors si

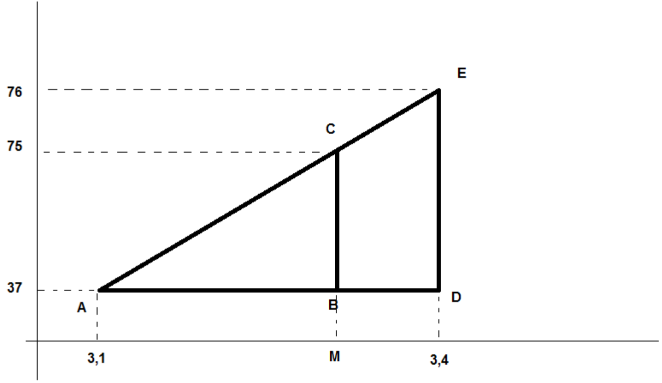
$$
\begin{aligned}
n &= 2p \Longrightarrow Me = \frac{x_p + x_{p+1}}{2} \\
n &= 2p+1 \Longrightarrow Me = x_{p+1}.
\end{aligned}
$$

On considère l'exemple des nouveaux né

X : Poids en Kg	Centre	n_i	$n_i^c \nearrow$	$n_i x_i$
$[2.2; 2.5[$	2,35	5	5	11,75
$[2.5; 2.8[$	2,65	11	16	29,15
$[2.8; 3.1[$	2,95	21	37	61,95
$[3.1; 3.4[$	3,25	39	76	126,75
$[3.4; 3.7[$	3,55	35	111	124,25
$[3.7; 4.0[$	3,85	20	131	77
$[4.0; 4.3[$	4,15	13	144	53,95
$[4.3; 4.6[$	4,45	6	150	26,7
Total		150		511,5

On détermine d'abord la classe médiane qui correspond à la moitié des effectifs $\left(\frac{n}{2} = 75\right)$ d'où $Me \in [3,1; 3,4[$.

Pour déterminer *Me* on considère la courbe des effectifs cumulés croissants mais on ne s'intéressera qu'à la portion de la classe $[3,1; 3,4[$



D'après le théorème de Thales on a

$$\frac{AB}{AD} = \frac{BC}{DE} \Rightarrow \frac{M \quad 3,1}{3,4 \quad 3,1} = \frac{75 \quad 37}{76 \quad 37}$$

$$\Rightarrow \quad M = \frac{38}{39}0,3 + 3,1$$

$$\Rightarrow \quad M \quad 3,392\,\text{Kg}$$

On a aussi pour cet exemple $\overline{X} = \frac{511,5}{150} \quad 3,41\,\text{Kg}$

L'écart type d'une série statistique $(x_1, x_2, \cdots, x_k)$, d'un caractère X, et d'effectif respectif respectif $(n_1, n_2, \cdots, n_k)$ est donné par le nombre réel σ_X défini par

$$\sigma_X = \sqrt{\frac{1}{n}\sum_{i=1}^{n} n_i \left(x_i - \overline{X}\right)^2}$$

où $n = \sum_{i=1}^{n} n_i$.
x_i est la modalité i de la variable X dans le cas d'une variable discrète et c'est le centre de la classe i dans le cas d'une variable continue.
Remarque 1. On peut aussi décrire σ_X sous la forme suivante

$$\sigma_X = \sqrt{\frac{1}{n}\sum_{i=1}^{n} n_i x_i^2 - \overline{X}^2}.$$

Remarque 2. On appelle variance de la variable X, noté $Var\,(X)$, le carré de l'écart type. $Var\,(X) = \sigma_X^2$.

Exemple : On reprend l'exemple de la distribution des nouveaux né suivant leur poids.

X : Poids en Kg	x_i	n_i	$n_i^c \nearrow$	$n_i x_i$	$n_i x_i^2$
$[2.2; 2.5[$	$2,35$	5	5	$11,75$	$27,6125$
$[2.5; 2.8[$	$2,65$	11	16	$29,15$	$77,2475$
$[2.8; 3.1[$	$2,95$	21	37	$61,95$	$182,7525$
$[3.1; 3.4[$	$3,25$	39	76	$126,75$	$411,9375$
$[3.4; 3.7[$	$3,55$	35	111	$124,25$	$441,0875$
$[3.7; 4.0[$	$3,85$	20	131	77	$296,45$
$[4.0; 4.3[$	$4,15$	13	144	$53,95$	$223,8925$
$[4.3; 4.6[$	$4,45$	6	150	$26,7$	$118,815$
Total		150		$511,5$	$1779,795$

On a

$$\begin{aligned}
\overline{X} &= \frac{1}{n}\sum_{i=1}^{8} n_i x_i = \frac{511,5}{150} \approx 3,41 Kg \\
\sigma_X^2 &= \frac{1}{n}\sum_{i=1}^{8} n_i x_i^2 - \overline{X}^2 = \frac{1779,795}{150} - 3,41^2 \\
&\Longrightarrow \sigma_X^2 \approx 0,2372 \Longrightarrow \sigma_X \approx 0,4870 Kg.
\end{aligned}$$

Le coefficient de variation est un paramètre de dispersion relative exprimé en pourcentage et défini par

$$CV_X = 100\frac{\sigma_X}{\overline{X}}.$$

Pour l'exemple précédent on obtient

$$\begin{aligned} CV_X &= 100\frac{0,4870}{3,41} \\ &\Longrightarrow 14,28\%. \end{aligned}$$

Les quartiles d'une série statistique sont les valeurs qui partagent la série en quatre parties de même effectif. Alors il existe trois quartiles, le premier quartile Q_1, le deuxième quartile Q_2 et le troisième quartile Q_3. Le deuxième quartile Q_2 étant la médiane Me.
L'écart interquartile est la différence entre le troisième et le premier quartile et on le note

$$IQR = Q_3 - Q_1.$$

Cas d'une variable discrète
Exemple : Considérons les notes de 21 étudiants obtenues à l'examen de biostatistique

$$4, 5, 6, 7, 7, 8, 9, 9, 10, 10, 10, 11, 11, 11, 12, 12, 13, 13, 15, 15, 16.$$

On a $n = 21 = 2 \times 10 + 1 \implies p = 10$ d'où
$Q_2 = Me = x_{p+1} = x_{11} = 10.$

$$\underbrace{4, 5, 6, 7, 7, 8, 9, 9, 10, 10}, \boxed{10}, \underbrace{11, 11, 11, 12, 12, 13, 13, 15, 15, 16}.$$

et $p = 10 = 2 \times 5 \implies p = 5$ d'où $Q_1 = \frac{x_5 + x_6}{2} = \frac{7+8}{2} = 7,5.$

$$\underbrace{4, 5, 6, 7, 7}, \boxed{8{,}5}, \underbrace{8, 9, 9, 10, 10}, \boxed{10}, \underbrace{11, 11, 11, 12, 12}, \boxed{12{,}5}, \underbrace{13, 13, 15, 15, 16}.$$

Cas d'une variable continue

Exemple : On considère l'exemple des nouveaux né.

X : Poids en Kg	x_i	n_i	$n_i^c \nearrow$
$[2.2; 2.5[$	2,35	5	5
$[2.5; 2.8[$	2,65	11	16
$[2.8; 3.1[$	2,95	21	37
$[3.1; 3.4[$	3,25	39	76
$[3.4; 3.7[$	3,55	35	111
$[3.7; 4.0[$	3,85	20	131
$[4.0; 4.3[$	4,15	13	144
$[4.3; 4.6[$	4,45	6	150
Total		150	

On a $\frac{n}{4} = 37,5$ alors $Q_1 \in [3,1;3,4[$ d'où

$$Q_1 = \frac{37,5 - 37}{76 - 37} \times (3,4 - 3,1) + 3,1 = 3,104Kg$$

et $\frac{3n}{4} = 112,5$ alors $Q_3 \in [3,7;4,0[$ d'où

$$Q_3 = \frac{112,5 - 111}{131 - 111} \times (3,7 - 4,0) + 3,7 = 3,723Kg$$

d'où

$$\begin{aligned} IQR &= Q_3 - Q_1 = 3,723 - 3,104 \\ &\Longrightarrow IQR = 0,619Kg. \end{aligned}$$