Université Mohammed V
Faculté des Sciences de Rabat
Département de Mathématiques

2022–2023
Filière SMI (S3)
Statistique Descriptive et Probabilités

Statistique descriptive univariée et bivariée

1 Statistique descriptive univariée

1.1 Vocabulaire Statistique

La ***Statistique*** est la science du traitement de l'information et de la prise de décision. Elle englobe un ensemble de méthodes et théories appliquées à l'analyse des données associées à une simulation ou à un phénomène dont le comportement ne peut être décrit avec certitude mais plutôt être analysé dans un contexte d'incertitude. Le but de l'utilisation de ces méthodes est d'arriver à des conclusions pratiques pour éventuellement proposer des recommandations et des mesures correctives s'il y a lieu.

Dans un autre autre sens on appelle ***une statistique*** une donnée ou une information tirée d'une population (ou d'un échantillon), c'est la collection des données numériques (chiffres), relatif à un phénomène, à une activité etc : gestion financière (états, banques, assurances, entreprises...), démographie, contrôles de qualité, études de marché, sciences expérimentales (biologie, psychologie...).

La ***Statistique Descriptive*** (univariée, bivariée, multivariée) a pour objet de proposer une description simple, clairement présentée et aussi complète

que possible d'un ensemble des données (ou informations) que l'on possède sur un sujet.

Ci-après quelques définitions de base pour développer le vocabulaire statistique.

Définition 1.1.

1. ***La science statistique :*** *Méthode scientifique du traitement des données. La statistique s'applique dans la plupart des disciplines : agronomie, biologie, démographie, économie, sociologie, linguistique, psychologie, ...*

2. ***Statistique Descriptive et Inférentielle : La Statistique Descriptive*** *à pour objectif de traiter les données, et d'en dégager certaines conclusions.* ***La Statistique Inférentielle*** *est la statistique inductive a pour objectif de tirer des conclusions et des décisions sur une population à partir d'un échantillon (sous-population) tiré de cette population.*

3. ***Population :*** *La population est l'ensemble des éléments sur lesquels porte une étude statistique.*

   Exemple : Dans une usine fabriquant des produits d'éclairage, on a mesuré la durée de vie de certaines type lampes. L'ensemble des lampes fabriquées au cours de cette étude constitue la population.

4. ***Individu :*** *L'individu (ou aussi appelé unité statistique) est l'un des éléments de la population, qui est soumis à une étude statistique.*

   Exemple : Dans l'exemple précédent de l'usine fabriquant des produits d'éclairage. Chaque lampe est un individu de la population.

5. ***Échantillon :*** *L'échantillon est un sous ensemble tiré aléatoirement d'une population (la taille de l'échantillon est raisonnable par rapport à la taille de la population, lorsque cette dernière est impossible de la tirer entièrement, on a recours à l'échantillonnage).*

6. ***Caractère et Modalité :*** *Le caractère (ou* ***variable statistique****) représente l'objectif de l'étude statistique, c'est la caractéristique étudiée sur tous les individus de la population (on le note par des lettres majuscules X, Y, ...). Le résultat pris par chaque individu est appelé* ***modalité****.*

   Exemple : Dans l'exemple précédent de l'usine fabriquant des produits d'éclairage. L'objectif est de mesurer la durée de vie de certaines type lampes. Donc la variable statistique est X : "durée de vie des lampes". les résultats peuvent êtres : $5h$, $0h$, $2h$, ... *ces derniers sont les modalités de la variables.*

Les modalités nous permettent de distinguer les types de la variable.

Définition 1.2 (Variable qualitative et Variable quantitative)**.**

1. *Un caractère (ou une variable) est dit* ***quantitatif*** *si ses modalités sont mesurables, sinon le caractère est dit* ***qualitatif****.*
2. ***Un caractère quantitatif*** *peut être*

   (a) discret : *Lorsque le caractère statistique prend un nombre fini de valeurs : entre deux valeurs successives de modalités il n'existe pas de valeur pour une autre modalité (nombre d'enfants, nombre de pièces, ...),*

   (b) continu : *Lorsque le caractère statistique peut prendre toutes les valeurs d'un intervalle de l'ensemble des nombres réels : entre deux valeurs successives de modalités il existe toujours une autre valeur pour une modalité (durée de vie, taille, temps d'appel, ...).*
3. ***Un caractère qualitatif*** *peut être*

   (a) ordinal : *lorsqu'on peut établir un ordre pour les modalités (Appréciation de la qualité d'un produit consommé : excellente, bonne, moyenne, mauvaise),*

(b) nominal : *lorsqu'on ne peut pas établir un tel ordre (La couleur : jaune, bleue, verte, rouge, ...).*

1.2 Tableau Statistique

— On appelle série statistique la suite des valeurs prises par une variable (un caractère) X sur les unités d'observation, ces unités forment les modalités de la variable.

— Le nombre d'unités d'observation est noté n, c'est la taille totale de la population.

— Les valeurs de la variable X sont notées $x_1, x_2, ..., x_n$.

$$\mathcal{S}_n = \{x_1, x_2, ..., x_n\}$$

— Cette série statistique est **non-groupée** et d'où vient la définition du ***tableau statistique***.

— Le tableau statistique permet de regrouper la série en **modalité/effectif** : on compte le nombre d'observations associé à chaque modalité puis on les dresse dans un tableau (tableau statistique, ou distribution statistique)

Exemple 1.1. *Soit la variable X représentant "l'état civil de* 20 *employés dans une entreprise".*

La série statistique des valeurs prises par X est la suivante :

$$M-M-D-C-C-M-C-C-C-M-C-M-V-M-V-D-C-C-C-M,$$

où, C : célibataire, M : marié(e), V : veuf(ve), D : divorcé(e).
Le tableau statistique associé à cette série statistique est le suivant :

Modalité (x_i)	*Effectif* (n_i)
C	9
M	7
V	2
D	2
Σ	20

Exemple 1.2. *On a relevé une population de* 50 *ménages et la variable X représentant "le nombre de personnes par ménage". Les valeurs de la variable sont*

1−1−1−1−1−2−2−2−2−2−2−2−2−2−3−3−3−3−3−3−3−3−3−3−3

3−3−3−3−4−4−4−4−4−4−4−4−4−5−5−5−5−5−6−6−6−8−8.

Le tableau statistique associé à cette série statistique est le suivant :

Modalité (x_i)	*Effectif* (n_i)
1	5
2	9
3	15
4	10
5	6
6	3
8	2
Σ	50

Remarque 1.1. *En plus des effectifs, le tableau statistique contient d'autres informations telles que : les fréquences, les effectifs cumulés et les fréquences cumulées.*

1.3 Effectif, fréquence, fréquence cumulée et effectif cumulé

– Cas discret :

Soit une série statistique à n observations et à k modalités discrètes $x_1, x_2, ..., x_k$. Le tableau statistique (complet) associé à cette série statistique est le suivant :

Modalité (x_i)	Effectif (n_i)	Fréquence (f_i)	Fréquence cumulée (F_i)	Effectif cumulé (N_i)
x_1	n_1	f_1	F_1	N_1
x_2	n_2	f_2	F_2	N_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_i	n_i	f_i	F_i	N_i
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_k	n_k	f_k	$F_k = 1$	$N_k = n$
Σ	n	1	$/\ /$	$/\ /$

avec, pour $i = 1, 2, ..., n$:

- n_i est l'effectif associé à la modalité x_i.
- $f_i = \frac{n_i}{n}$ est la fréquence associée à la modalité x_i.

- $F_i = \sum_{j=1}^{i} f_j$ est la fréquence cumulée associée à la modalité x_i.
- $N_i = \sum_{j=1}^{i} n_j$ est l'effectif cumulé associé à la modalité x_i.

De plus, on a $\sum_{i=1}^{k} n_i = n$, $\sum_{i=1}^{k} f_i = 1$ et $F_i = F_{i-1} + f_i$, ...

– Cas continu :

Une variable quantitative continue peut prendre une infinité de valeurs possibles. Soit une série statistique à n observations : $x_1, x_2, ..., x_n$, on regroupe les données sous forme de k classes.

Le tableau statistique (complet) associé à cette série statistique est le suivant :

Modalité (x_i)	Effectif (n_i)	Fréquence (f_i)	Fréquence cumulée (F_i)	Effectif cumulé (N_i)
$[v_1, v_2[$	n_1	f_1	F_1	N_1
$[v_2, v_3[$	n_2	f_2	F_2	N_2
$\vdots$ $\vdots$	$\vdots$ $\vdots$	$\vdots$ $\vdots$	$\vdots$ $\vdots$	$\vdots$ $\vdots$
$[v_i, v_{i+1}[$	n_i	f_i	F_i	N_i
$\vdots$ $\vdots$	$\vdots$ $\vdots$	$\vdots$ $\vdots$	$\vdots$ $\vdots$	$\vdots$ $\vdots$
$[v_k, v_{k+1}[$	n_k	f_k	$F_k = 1$	$N_k = n$
Σ	n	1	/ /	/ /

Si $[v_i, v_{i+1}[$ désigne la classe i, alors, pour $i = 1, 2, ..., k$:

- n_i, f_i, F_i et N_i se calculent de la même manière que dans le cas discret et on a les mêmes propriétés,
- $c_i = \frac{v_i + v_{i+1}}{2}$ est le centre de la classe $[v_i, v_{i+1}[$,
- $a_i = v_{i+1} - v_i$ est l'amplitude de la classe $[v_i, v_{i+1}[$,
- Il arrive que l'amplitude des classes extrêmes soit indéterminée,
- **Détermination du nombre de classes :**
  Le nombre de classes ne devrait, en généraln être ni inférieur à 5 ni supérieur à 20. De préférence, il varie entre 5 et 12 classes.
  En pratique on peut utiliser une formule pour déterminer le nombre de classes : Il s'agit de la formule de **Sturges** (la plus utilisée) ou la formule de **Yule** :
  Soient n la taille de la population et k le nombre de classes à utiliser, alors :

$$\text{Formule de Sturges : } k = 1 + \frac{10}{3}\log_{10}(n).$$

$$\text{Formule de Yule : } k = 2.5\sqrt[4]{n}.$$

On arrondit le nombre de classe k à l'entier le plus proche.

On calcule l'amplitude des classes : $A = \frac{e}{k}$, avec $e = x_{\max} - x_{\min}$ est l'**étendue de la série**.

A partir de la plus petite valeur observée, on obtient les bornes de classes en additionnant successivement par A et on retrouve l'intervalle de chaque classe (qui a la même amplitude de tous les intervalles).

Exemple 1.3. *Les données suivantes sont les durées de vie en heures de* 30 *lampes miniatures.*

419 451 412 412 375 397 429 407 454 375 393 357 456 355 364 414 413 425 467 345 432 392 329 422 426 439 381 451 413 421

on suit, en général, les étapes suivantes :

1. *On ordonne :*

   329 345 355 357 364 375 375 381 392 393 397 407 412 412 413 413 414 419 421 422 425 426 429 432 439 451 451 454 456 467

2. *La* ***formule de Sturges*** *donne le nombre de classes :*
   $k = 1 + 3,33 \log_{10}(30) \simeq 6 \Rightarrow$ *6 classes.*

3. *On calcule l'étendue :* $e = x_{\max} - x_{\min} = 476 - 329 = 138.$

4. *On calcule l'amplitude des classes :* $A = \dfrac{e}{k} = \dfrac{138}{6} = 23.$

5. *On obtient le tableau des classes ci-dessous :*

Classe	*Effectif*	*Fréquence*	*Fréquence cumulée*
[329,352[	*2*	*0,07*	*0,07*
[352,375[	*3*	*0,1*	*0,17*
[375,398[	*6*	*0,2*	*0,37*
[398,421[	*7*	*0,23*	*0,6*
[421,444[	*7*	*0,23*	*0,83*
[444,467]	*5*	*0,17*	*1*
Total	*30*	*1*	*/ /*

— $n_3 = 6$ *: Effectif de la* $3^{ème}$ *classe,*

— $f_3 = \dfrac{n_3}{n} = \dfrac{6}{30} = 0,2$ *: Fréquence de la* $3^{ème}$ *classe,*

— $F_3 = F_2 + f_3 = 0,17 + 0,2 = 0,37$ *: Fréquence cumulée de la* $3^{ème}$ *classe : représente la proportion des lampes ayant moins de 398.*

– Cas qualitatif :

Lorsque la variable est qualitative, on ne calcule pas les fréquences cumulées et les effectifs cumulés, ils n'ont pas de sens en statistique (...). les modalités $x_i, i = 1, ..., k$, deviennent des qualités et le tableau statistique (complet) a la forme suivante :

Modalité (x_i)	Effectif (n_i)	Fréquence (f_i)
x_1	n_1	f_1
x_2	n_2	f_2
$\vdots$ $\vdots$	$\vdots$ $\vdots$	$\vdots$ $\vdots$
x_i	n_i	f_i
$\vdots$ $\vdots$	$\vdots$ $\vdots$	$\vdots$ $\vdots$
x_k	n_k	f_k
$\sum$	n	1

avec, $f_i = \frac{n_i}{n}$, pour $i = 1, 2, ..., n$, est la fréquence associée à la modalité x_i.

Exemple 1.4. (1) *Reprenons l'exemple 1.1 sur la variable X représentant "l'état civil de 20 employés dans une entreprise".*
Le tableau statistique (complet) associé à cette série statistique est le suivant :

Modalité (x_i)	*Effectif* (n_i)	*Fréquence* (f_i)
C	9	0.45
M	7	0.35
V	2	0.10
D	2	0.10
Σ	20	1

(2) *Reprenons l'exemple 1.2 de* 50 *ménages où la variable* X *représentant "le nombre de personnes par ménage".*
Le tableau statistique (complet) associé à cette série statistique est le suivant :

Modalité (x_i)	*Effectif* (n_i)	*Effectif cumulé* (N_i)	*Fréquence* (f_i)	*Fréquence cumulée* (F_i)
1	5	5	0.10	0.10
2	9	14	0.18	0.28
3	15	29	0.30	0.58
4	10	39	0.20	0.78
5	6	45	0.12	0.90
6	3	48	0.06	0.96
8	2	50	0.04	1.00
Σ	50	/ /	1	/ /

Exemple 1.5 (**Autre Exemple**). *La répartition de* 40 *familles d'un certain quartier de la ville selon le nombre d'enfants par famille est donnée par le tableau suivant :*

Nb d'enfants par famille	*Effectif*	*Fréquence*	*Fréquence cumulée*
0	*4*	*0,1*	*0,1*
1	*8*	*0,2*	*0,3*
2	*8*	*0,2*	*0,5*
3	*7*	*0,175*	*0,675*
4	*6*	*0,15*	*0,825*
5	*4*	*0,1*	*0,925*
6	*3*	*0,075*	*1*
Total	*40*	*1*	*/ /*

— $n_4 = 7$ *(resp.* $f_4 = \dfrac{n_4}{n} = \dfrac{7}{40} = 0,175$*) : Effectif (resp. Fréquence) de la* $4^{ème}$ *observation,*

— $n = \Sigma n_i = n_1 + ... + n_7 = 40$ *: Effectif total,*

— $F_1 = f_1 = 0,1$ *: Fréquence cumulée de la* $1^{ère}$ *observation,*

— $F_2 = f1 + f_2 = 0,1 + 0,2 = 0,3$ *: Fréquence de la* $2^{ème}$ *observation,*

— $F_4 = f_1 + f_2 + f_3 + f_4 = F_3 + f_4 = 0,5 + 0,175 = 0,675$ *: Fréquence cumulée de la* $4^{ème}$ *observation,*

— $F_7 = f_1 + ... + f_7 = F_6 + f_7 = 0,925 + 0,075$***=1*** *: Fréquence cumulée de la dernière observation.*

1.4 Graphiques

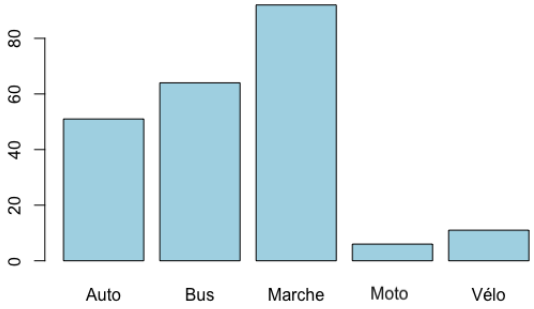
Les représentations graphiques ont l'avantage de renseigner immédiatement sur l'allure générale de la distribution. Elles facilitent l'interprétation des données recueillies.

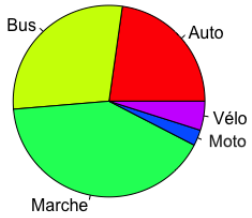
1.4.1 Variable qualitative

Pour les variables qualitatives, on trace l***e diagramme en barres (tuyaux d'orgue) des effectifs (ou des fréquence), ou le diagramme en secteur (circulaire).***

Exemple 1.6. *Moyens de transport des étudiants pour se rendre à l'université.*

Modalité	*Effectif*	*Fréquence*
Auto	*51*	*0,23*
Bus	*64*	*0,29*
Marche	*92*	*0,4*
Moto	*6*	*0,03*
Vélo	*11*	*0,05*
Total	*224*	*1*





Tuyaux d'orgue *Diagramme circulaire*

$$\underline{\textit{Bus :}}\ \textit{Angle} = 0,29 \times 360 = 102,85.$$

En général : pour une modalité ayant une fréquence f_i l'angle associée est calculée par la formule suivante :

$$\textit{Angle}_i = f_i \times 360$$

.

1.4.2 Variable quantitative discrète

Dans le cas où la variable est quantitative discrète, on trace le ***diagramme en bâtons des effectifs (ou des fréquences).***

Exemple 1.7. *On reprend l'exemple 1.5 de la répartition du nombre d'enfants de* 40 *famille (page 17).*

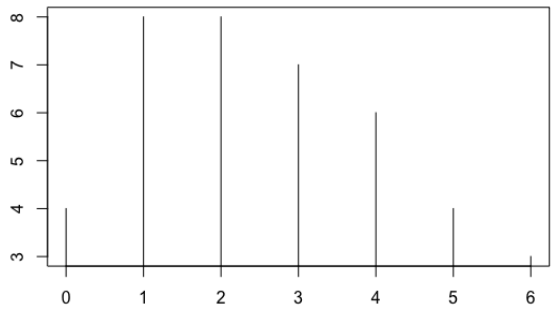


FIGURE 1 – Diagramme en bâtons des effectifs du nombre d'enfants de 40 famille

Définition 1.3 (Polygone des effectifs (ou des fréquences))**.**

le polygone des effectifs (ou des fréquences) est obtenu en joignant les sommets de chaque bâton par des segments de droites. la seule utilité est de présenter l'allure générale de la distribution des fréquences (ou des effectifs).

1.4.3 Variable quantitative continue

Dans le cas où la variable est quantitative continue, on trace l'histogramme (des fréquences ou des effectifs), en regroupant les données sous forme de classes. Les amplitudes de ces classes peuvent être égales ou non.

Si les classes sont définies et ont la même amplitude A, on trace des rectangles dont la base est l'amplitude A de la classe la hauteur est égale

à la fréquence f_i (ou l'éffectif n_i) associée à chaque classe.

Si les classes ne sont pas définies, on utilise la **formule de Sturges** qui permettra de donner le nombre de classe selon la taille de la population.

Exemple 1.8. *Reprenons l'exemple 1.3 (page 13) :*

l'histogramme des fréquences associée cette distribution est donné dans la figure suivante :

Classe	*Effectif*	*Fréquence*
[329,352[	*2*	*0,07*
[352,375[	*3*	*0,1*
[375,398[	*6*	*0,2*
[398,421[	*7*	*0,23*
[421,444[	*7*	*0,23*
[444,467]	*5*	*0,17*
Total	*30*	*1*

Définition 1.4 (Polygone des effectifs (ou des fréquences))**.**

Le polygone des effectifs (ou des fréquences) est obtenu en joignant les milieux des sommets de chaque rectangle de l'histogramme par des

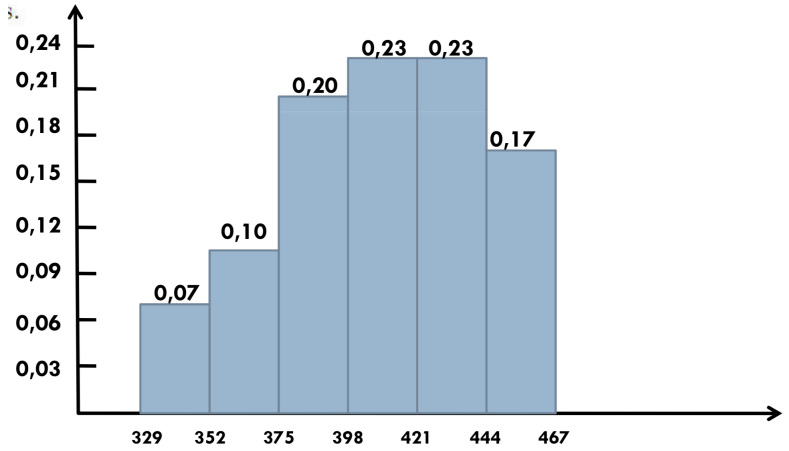


FIGURE 2 – Histogramme des fréquences

segments de droites. la seule utilité est de présenter l'allure générale de la distribution des fréquences (ou des effectifs), c'est le graphe commun entre les deux caractères continu et discret.

Exemple 1.9. *Reprenons l'exemple 1.3 (page 13). On trace le polygone des fréquences en reliant les centres des classes (figure 3).*

Remarque 1.2. — *Dans le cas où les classes ont des amplitudes différentes, pour chaque classe on trace un rectangle dont la base est l'amplitude a_i de la classe mais dont la hauteur est égale à la fréquence corrigée*

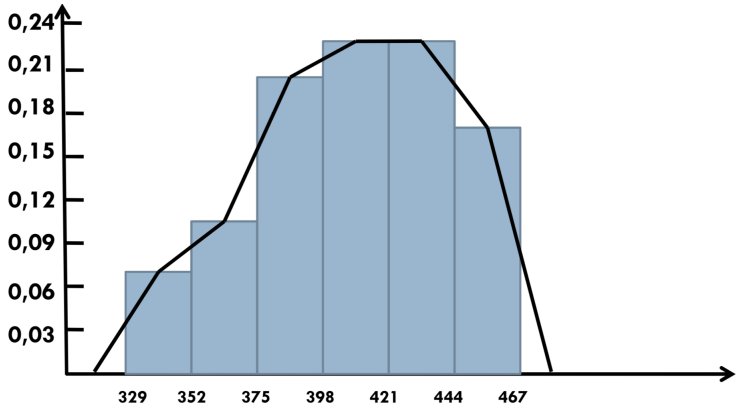


FIGURE 3 – Polygone des fréquences

f_i^c obtenue ainsi :

$$f_i^c = \frac{f_i}{a_i} \times a_0$$

où f_i est la fréquence de la classe $[x_i, x_{i+1}[$ et a_0 l'amplitude de base choisie (c'est généralement la plus petite, ou on prend $a_0 = 1$).

— *On pourra aussi travailler avec les effectifs corrigés n_i^c de la même manière :*

$$n_i^c = \frac{n_i}{a_i} \times a_0.$$

— ***La correction des effectifs (ou des fréquences) sert seulement à tracer l'histogramme des effectifs (ou des fréquences) et***

à la définition de la classe modale.

1.4.4 Courbe cumulative croissante (Fonction de répartition)

– Cas discret :

La représentation de la fonction cumulative croissante (appelée aussi fonction de répartition) est réalisée au moyen des fréquences cumulées. Cette fonction est définie de $\mathbb{R}$ dans $[0, 1]$ et vaut, pour $i = 1, 2, ..., k$ (où k est le nombre de modalités discrètes) :

$$F(x) = \begin{cases} 0 & \text{si } x < x_1 \\ F_i & \text{si } x_i \leq x < x_{i+1} \\ 1 & \text{si } x \geq x_k \end{cases}$$

Exemple 1.10. *Reprenons l'exemple 1.2 de* 50 *ménages où la variable* X *représentant "le nombre de personnes par ménage" (page 8). la fonction de répartition (fonction cumulative) est représentée comme suit :*

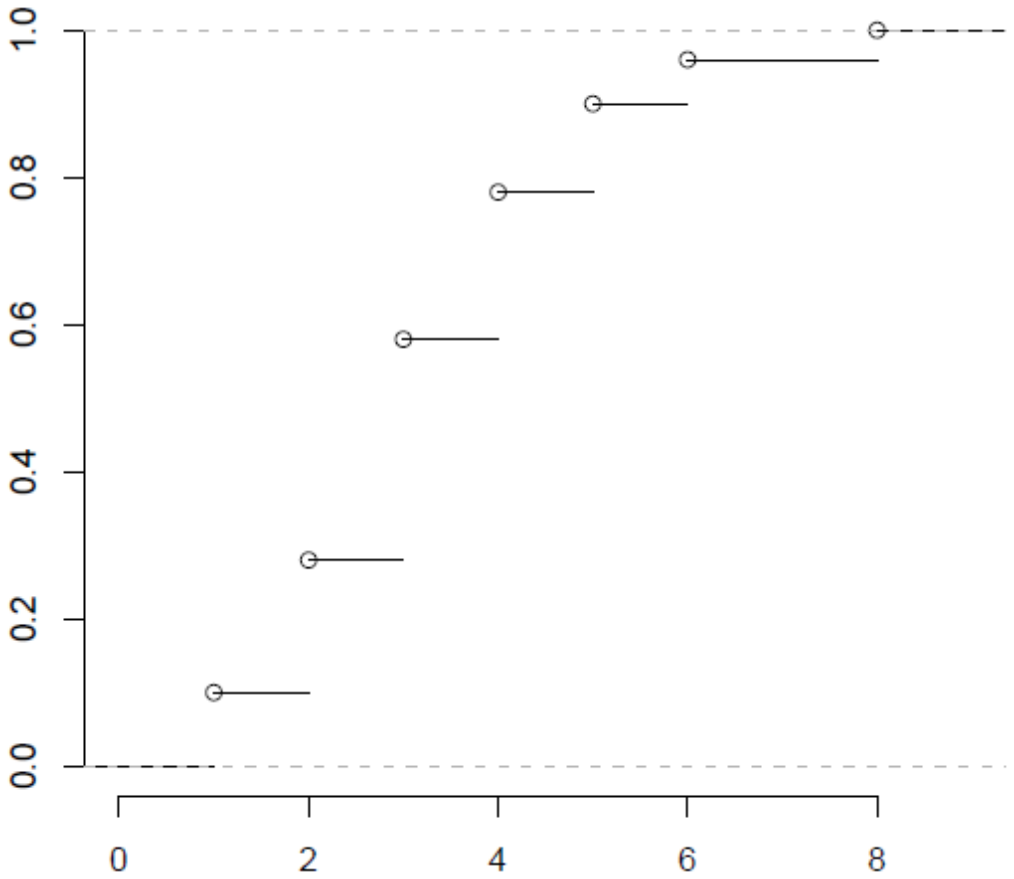


FIGURE 4 – Courbe cumulative croissante de la variable discrète X

– Cas continu :

La courbe cumulative des fréquences de d'une distribution statistique (définie par des classes $[x_i, x_{i+1}[$ et des fréquences cumulées F_i) s'obtient en liant les points $A_i(x_{i+1}, F_i)$ par des segments (pour $i = 1, 2, ..., k$, avec k est le nombre de classe, $x_{k+1} = x_{\max}$ et $x_1 = x_{\min}$). Il s'agit d'une fonction continue définie de $\mathbb{R}$ dans $[0, 1]$ dont sa limite en $-\infty$ vaut 0 et en $+\infty$ vaut 1.

Exemple : La courbe cumulative de l'exemple de la page 13 est la suivante :

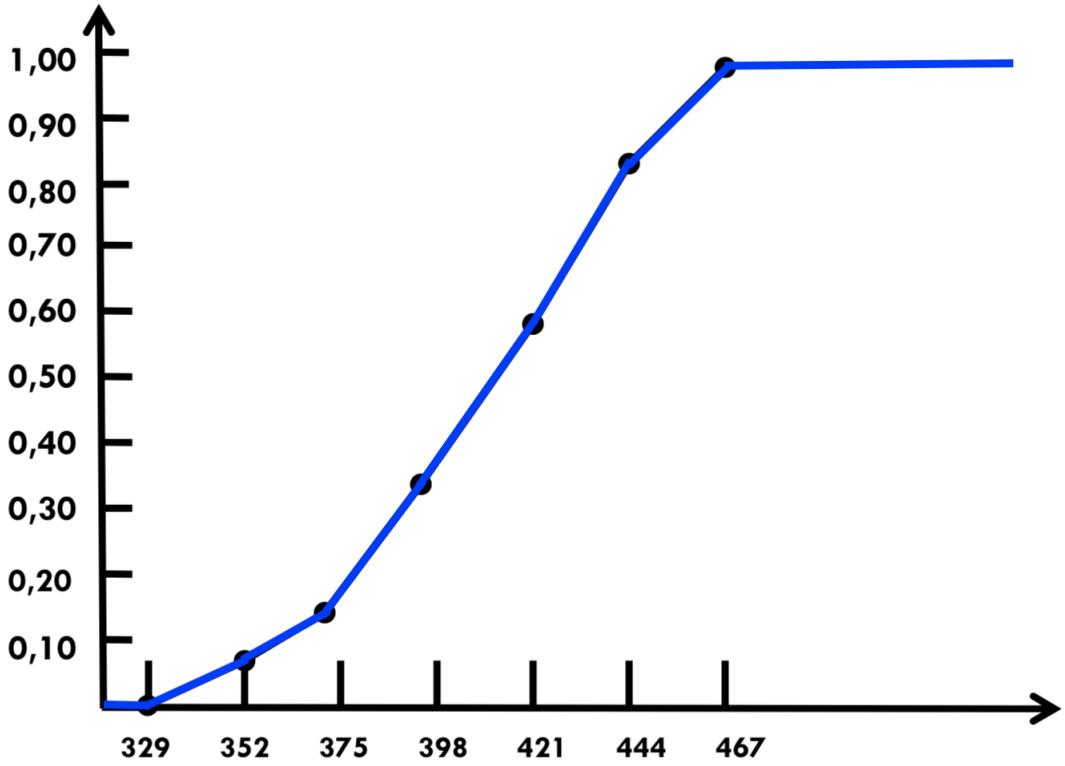


FIGURE 5 – Courbe cumulative croissante de la variable continue X

Les tableaux et les graphes permettent d'obtenir une première image de la distribution des données. Pour améliorer ou éclaircir plus cette image, on introduit de nouveaux indicateurs statistiques qui caractériseront la distribution : On distingue des mesure de tendance centrale, des mesures de dispersion et des mesures de forme (on peut aussi chercher des aspects particuliers : valeurs extrêmes, groupe de valeurs, ...). Ces mesures ne sont calculées que dans le cas d'un caractère **quantitatif (non groupé ou groupé : discret, continu)**.

1.5 Mesures de tendance centrale

1.5.1 Mode

Variable quantitative discrète

Définition 1.5. *Le **mode**, noté m_o est la valeur de la variable ayant le plus grand effectif (ou la plus grande fréquence). Si la série admet deux modes on dit que la distribution et bimodale et note les deux modes.*

Exemple 1.11. *On reprend l'exemple de la page 17 du nombre d'enfants par famille.*

Nb d'enfants par famille	*Effectif*	*Fréquence*	*Fréquence cummulée*
0	*4*	*0,1*	*0,1*
1	*8*	*0,2*	*0,3*
2	*8*	*0,2*	*0,5*
3	*7*	*0,175*	*0,675*
4	*6*	*0,15*	*0,825*
5	*4*	*0,1*	*0,925*
6	*3*	*0,075*	*1*
Total	*40*	*1*	*/ /*

On dispose de deux modes $m_o = 1$ ou $m_o = 2$. Il s'agit d'une série statistique ***bimodale****.*

Variable quantitative continue (classe modale)

Définition 1.6. — *La classe modale est la classe de la variable ayant le plus grand effectif (ou la plus grande fréquence).*
— *On peut considérer le mode comme la valeur milieu de la classe modale.*
— *Si les classes ont des amplitudes inégales, alors la classe modale est la classe associée au plus grand effectif corrigé ou la plus grande fréquence corrigée.*

Exemple 1.12. *On reprend l'exemple 1.3 des lampes.*

Classe	*Effectif*	*Fréquence*	*Fréquence cummulée*
[329,352[	*2*	*0,07*	*0,1*
[352,375[	*3*	*0,1*	*0,17*
[375,398[	*6*	*0,2*	*0,37*
[398,421[	*7*	*0,23*	*0,6*
[421,444[	*7*	*0,23*	*0,83*
[444,467]	*5*	*0,17*	*1*
Total	*30*	*1*	

Ici aussi, on dispose de deux classes modales $[398, 421[$ *et* $[421, 444[$*, directement puisque les classes ont la même amplitude.*

1.5.2 Moyenne

La moyenne constitue l'un des paramètres fondamentaux de tendance centrale mais non suffisant pour caractériser une distribution. Complémentaire du mode. La moyenne constitue la mesure la plus calculée et la plus utilisée lors de la description de séries statistiques. Il existe plusieurs types de moyennes, chacun adapté à des situations précises :

Moyenne arithmétique

La moyenne arithmétique (souvent appelée moyenne), notée $\bar{x}$, d'une variable dans une série statistique est définie par :

— ***Cas discret :***

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{k} n_i x_i = \frac{n_1 x_1 + ... + n_k x_k}{n} = \sum_{i=1}^{k} f_i x_i,$$

où $x_1, .., x_k$ sont les différentes valeurs de la variable.

— ***Cas continu :***

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{k} n_i c_i = \frac{n_1 c_1 + ... + n_k c_k}{n_1 + ... + n_k} = \sum_{i=1}^{k} f_i c_i,$$

où $c_i = \dfrac{v_i + v_{i+1}}{2}$ est le centre de la classe $[v_i, v_{i+1}[$.

— ***Cas où les données ne sont pas groupées :***

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

où n est la taille de la série statistique.

Exemple 1.13. *On reprend l'exemple 1.5 du nombre d'enfants par famille (page 17).*

Nombre d'enfants par famille	*Effectif*	*Fréquence*	*Fréquence cumulée*
0	*4*	*0,1*	*0,1*
1	*8*	*0,2*	*0,3*
2	*8*	*0,2*	*0,5*
3	*7*	*0,175*	*0,675*
4	*6*	*0,15*	*0,825*
5	*4*	*0,1*	*0,925*
6	*3*	*0,075*	*1*
Total	*40*	*1*	

La moyenne est : $\bar{x} = \dfrac{4 \times 0 + 8 \times 1 + ... + 3 \times 6}{40} = \dfrac{107}{40} = 2,675.$

$\Rightarrow$ *Environ, les familles ont en moyenne 3 enfants.*

Exemple 1.14. *On reprend l'exemple 1.3 des lampes (page 13).*

Classe	*Effectif*	*Fréquence*	*Fréquence cumulée*
[329,352[	*2*	*0,07*	*0,1*
[352,375[	*3*	*0,1*	*0,17*
[375,398[	*6*	*0,2*	*0,37*
[398,421[	*7*	*0,23*	*0,6*
[421,444[	*7*	*0,23*	*0,83*
[444,467]	*5*	*0,17*	*1*
Total	*30*	*1*	

Les centres de classes sont : $c_1 = \dfrac{329+352}{2} = 340,5, ..., c_6 = \dfrac{444+467}{2} = 455,5.$

Et la moyenne est : $\bar{x} = \dfrac{2 \times 340,5 + ...5 \times 455,5}{30} = \dfrac{12262}{30} = 408,7333.$

$\Rightarrow$ *Environ, les lampes durent an moyenne* $409h.$

Moyenne géométrique

La moyenne géométrique est appliquée à des mesures de grandeurs dont la croissance est géométrique ou exponentielle.

La moyenne géométrique, notée $\bar{x}_G$, d'une variable dans une série statistique est définie par :

— ***Cas discret :***

$$\bar{x}_G = \sqrt[n]{\prod_{i=1}^{k} x_i^{n_i}} = \prod_{i=1}^{k} x_i^{f_i},$$

où $x_1, .., x_k$ sont les différentes valeurs de la variable.

— ***Cas continu :***

$$\bar{x}_G = \sqrt[n]{\prod_{i=1}^{k} c_i^{n_i}} = \prod_{i=1}^{k} c_i^{f_i},$$

où $c_i = \dfrac{v_i + v_{i+1}}{2}$ est le centre de la classe $[v_i, v_{i+1}[$.

— ***Cas où les données ne sont pas groupées :***

$$\bar{x}_G = \sqrt[n]{\prod_{i=1}^{n} x_i},$$

où n est la taille de la série statistique.

Remarque 1.3. *On peut écrire la moyenne géométrique comme l'exponentielle de la moyenne arithmétique des logarithmes des valeurs observées, on a par exemple pour les données non groupées :*

$$\bar{x}_G = \exp \ln \bar{x}_G = \exp \ln \sqrt[n]{\prod_{i=1}^{n} x_i} = \exp \frac{1}{n} \sum_{i=1}^{n} \ln x_i$$

Exemple 1.15. *Si les taux d'intérêt pour* 4 *ans successif sont respectivement de* 5, 10, 15, *et* 10%. *Quelle est le montant retrouvé après* 4 *si on place* $1000DH$

- *Après 1 an on aura :* $1000 \times 1.05 = 1050DH$.
- *Après 2 ans on aura :* $1000 \times 1.05 \times 1.1 = 1155DH$.
- *Après 3 ans on aura :* $1000 \times 1.05 \times 1.1 \times 1.15 = 1328.25DH$.
- *Après 4 ans on aura :* $1000 \times 1.05 \times 1.1 \times 1.15 \times 1.1 = 1461.075DH$.

Si on calcule la moyenne arithmétique des taux on obtient

$$\bar{x} = \frac{1.05 + 1.10 + 1.15 + 1.10}{4} = 1.10.$$

Si on calcule la moyenne géométrique des taux, on obtient

$$\bar{x}_G = (1.05 \times 1.10 \times 1.15 \times 1.10)^{1/4} = 1.099431377.$$

Le bon taux moyen est bien $\bar{x}_G$ *et non* $\bar{x}$*, car si on applique* 4 *fois le taux moyen* $\bar{x}_G$ *aux* $1000DH$*, on obtient*

$$1000DH \times \bar{x}_G^4 = 1000 \times 1.0994313774 = 1461.075DH.$$

Moyenne harmonique

La moyenne harmonique est utilisée lorsqu'on veut déterminer un rapport moyen dans des domaines où il existe des liens de proportionnalité inverse. Par exemple, pour une distance donnée, le temps de trajet est d'autant plus court que la vitesse est élevée.

La moyenne harmonique, notée $\bar{x}_H$, d'une variable dans une série statistique est définie par :

— ***Cas discret :***

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

où $x_1, .., x_k$ sont les différentes valeurs de la variable.

— ***Cas continu :***

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

où $c_i = \dfrac{v_i + v_{i+1}}{2}$ est le centre de la classe $[v_i, v_{i+1}[$.

— ***Cas où les données ne sont pas groupées :***

$$\bar{x}_H = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}},$$

où n est la taille de la série statistique.

Exemple 1.16. *Un cycliste parcourt* 4 *étapes de* $100km$. *Les vitesses respectives pour ces étapes sont de* $10km/h$, $30km/h$, $40km/h$ *et* $20km/h$. *Quelle était sa vitesse moyenne ?*

Un raisonnement simple nous dit qu'il a parcouru la première étape en $10h$, *la deuxième en* $3h20$ *la troisième en* $2h30$ *et la quatrième en* $5h$. *Il a donc parcouru le total des* $400km$ *en* $10 + 3h20 + 2h30 + 5h = 20h50 = 20.8333h$, *sa vitesse moyenne est donc*

$$Moyenne = \frac{400}{20.8333} = 19.2km/h.$$

Si on calcule la moyenne arithmétique des vitesses, on obtient

$$\bar{x} = \frac{10 + 30 + 40 + 20}{4} = 25km/h.$$

Si on calcule la moyenne harmonique des vitesses, on obtient

$$\bar{x}_H = \frac{4}{\frac{1}{10} + \frac{1}{30} + \frac{1}{40} + \frac{1}{20}} = 19.2km/h.$$

La moyenne harmonique est donc la manière appropriée de calculer la vitesse moyenne.

Remarque 1.4. *Il est possible de montrer que la moyenne harmonique est toujours inférieure ou égale à la moyenne géométrique qui est toujours inférieure ou égale à la moyenne arithmétique*

$$\bar{x}_H \leq \bar{x}_G \leq \bar{x}$$

1.5.3 Médiane

Variable quantitative discrète

La médiane, notée M_e, est la valeur de la variable qui partage la série en deux parties égales. Pour déterminer la médiane M_e, on utilise les valeurs

ordonnées définies comme suit :

$$x_{\min} = x_{(1)} \leq x_{(2)} \leq ... \leq x_{(n)} = x_{\max}$$

où $x_{(i)}$ est la $i^{ème}$ valeur dans la série ordonnée.

On distingue alors les deux cas suivants :

— Si n est impair alors la médiane est $M_e = x_{\left(\frac{n+1}{2}\right)}$,

— Si n est pair alors la médiane est $M_e = \dfrac{x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}}{2}$.

Exemple 1.17. — *Série* : $5 - 2 - 6 - 8 - 1 - 9 - 3$.

Ordonner $\Rightarrow 1 - 2 - 3 - 5 - 6 - 8 - 9$.

$n = 7$ est impair. Donc la médiane est $M_e = x_{(4)} = 5$

— *Série* : $3 - 2 - 7 - 1 - 8 - 5 - 9 - 2$.

Ordonner $\Rightarrow 1 - 2 - 2 - 3 - 5 - 7 - 8 - 9$.

$n = 8$ est impair. Donc la médiane est $M_e = \dfrac{x_{(4)} + x_{(5)}}{2} = \dfrac{3+5}{2} = 4$

Variable quantitative continue

La médiane, notée M_e, est la valeur de la variable telle que $F(M_e) = 0,5$ (ou $M_e = F^{-1}(0,5)$) où F est la fréquence cumulée. Pour retrouver la valeur de M_e on utilise la méthode d'**interpolation linéaire**

Exemple 1.18. *Dans un atelier mécanique, on a fabriqué des tiges sur un tour automatique, les diamètres de ces tiges sont données dans le tableau suivant :*

Classe	*Effectif*	*Fréquence*	*Fréquence cumulée*
[36,5 ;37,5[	*3*	*0,05*	*0,05*
[37,5 ;38,5[	*7*	*0,12*	*0,17*
*[38,5 ;**39,5**[*	*17*	*0,28*	***0,45***
*[39,5 ;**40,5**[*	*18*	*0,3*	***0,75***
[40,5 ;41,5[	*9*	*0,15*	*0,90*
[41,5 ;42,5[	*4*	*0,07*	*0,97*
[42,5 ;43,5]	*2*	*0,03*	*1*
Total	*60*	*1*	

On connait la valeur de la fréquence cumulée égale à $0,5$ *et on cherche la valeur* M_e *de la variable telle que* $F(M_e) = 0,5$.

Puisque $0,5 \in [0,45; 0,75]$ *alors* $M_e \in [39,5; 40,5]$.

On a alors :

$$\frac{M_e - 39,5}{40,5 - 39,5} = \frac{0,5 - 0,45}{0,75 - 0,45}$$

Donc la médiane est

$$M_e = 39,5 + 1 \times \frac{0,05}{0,3} = 39,6667$$

Remarque 1.5. *En général, si* $M_e \in [x_i, x_{i+1}[$ *alors*

$$M_e = x_i + (x_{i+1} - x_i) \times \frac{0,5 - F(x_i)}{F(x_{i+1}) - F(x_i)}$$

Remarque 1.6. *La médiane peut aussi être déterminée graphiquement à travers la courbe cumulée croissante (fonction de répartition), c'est l'abscisse du point d'ordonnée* 0.5.

Dans le cas de distribution uni-modale, la médiane est fréquemment comprise entre la moyenne arithmétique et le mode, et plus près de la moyenne que du mode. Si la distribution est symétrique, ces trois caractéristiques de tendance centrale sont confondues (figure 6).

1.5.4 Quantiles

La notion de quantile d'ordre p (où $0 < p < 1$) généralise la médiane.

Formellement un quantile est donné par l'inverse de la fonction de répartition :

$$x_p = F^{-1}(p).$$

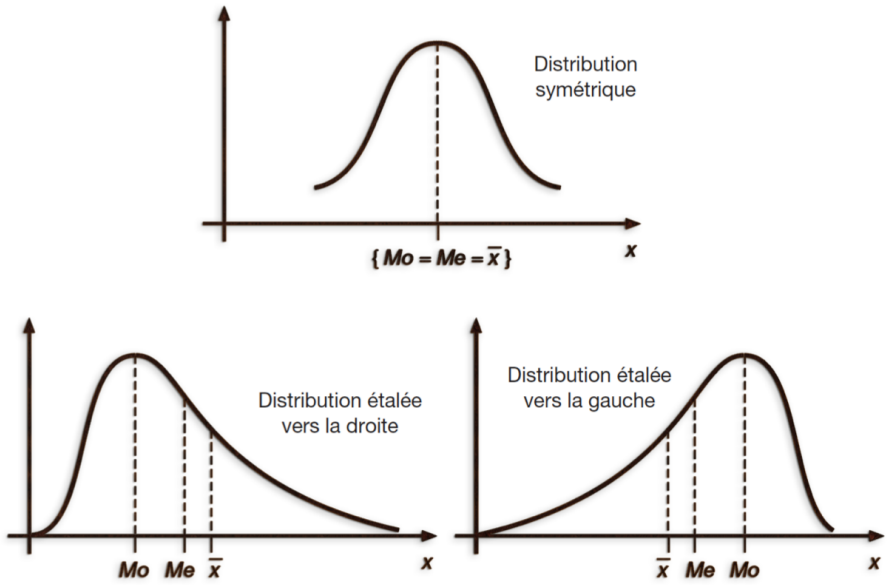


FIGURE 6 – positions possibles pour la moyenne, la médiane et le mode

Dans le cas discret / données non groupées, le quantile x_p d'ordre p se calcule de la manière suivante (n est la taille de la population) :

— Si np est un nombre entier, alors

$$x_p = \frac{x_{(np)} + x_{(np+1)}}{2}.$$

— Si np n'est pas un nombre entier, alors

$$x_p = x_{(\lceil np \rceil)},$$

où $\lceil np \rceil$ représente le plus petit nombre entier supérieur ou égal à np.

Dans le cas continu, on procède de la même manière que dans le calcul de la médiane, par interpolation linéaire, pour Q_1 on cherche l'intervalle qui contient une fréquence cumulée supérieure ou égale à 0.25 et pour Q_3 on cherche l'intervalle qui contient une fréquence cumulée supérieure ou égale à 0.75.

Remarque 1.7. — *La médiane est le quantile d'ordre* $p = 1/2$*, qui donne le même résultat précédent.*

— *On utilise souvent :*

$x_{1/4} = Q_1$ *le premier quartile,*

$x_{3/4} = Q_3$ *le troisième quartile,*

$x_{1/10} = D_1$ *le premier décile,*

$x_{9/10} = D_9$ *le neuvième décile.*

Exemple 1.19. *Soit la série statistique ordonnée de taille* $n = 10$ *suivante :*

$$12, 13, 15, 16, 18, 19, 22, 24, 25, 27$$

- Le premier quartile : Comme $np = 0.25 \times 10 = 2.5$ *n'est pas un entier, on a :*

$$x_{1/4} = Q_1 = x_{(\lceil 2.5 \rceil)} = x_{(3)} = 15.$$

- *La médiane : Comme* $np = 0.5 \times 10 = 5$ *est un entier, on a*

$$x_{1/2} = Me = \frac{x_{(5)} + x_{(6)}}{2} = (18 + 19)/2 = 18.5.$$

- *Le troisième quartile : Comme* $np = 0.75 \times 10 = 7.5$ *n'est pas un entier, on a :*

$$x_{3/4} = Q_3 = x_{(\lceil 7.5 \rceil)} = x_{(8)} = 24.$$

1.6 Mesures de dispersion

1.6.1 Étendue

Définition 1.7. *L'étendue, notée e, est la différence entre les valeurs maximale et minimale de la variable.*

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

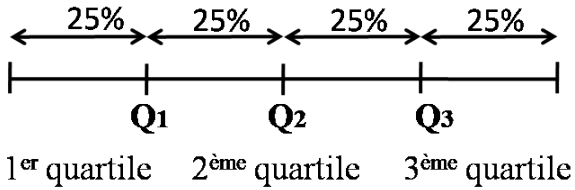
Exemple 1.20. *On considère la série suivante :*

$$12 - 15,5 - 17 - 9 - 19 - 5 - 8,5.$$

Les valeurs maximale et minimale sont respectivement : $x_{\max} = 19$ *et* $x_{\min} = 5$, *alors l'étendue est* $e = 19 - 5 = 14$.

1.6.2 Écart interquartile

Les trois valeurs Q_1, Q_2 et Q_3 avec $Q_1 \leq Q_2 \leq Q_3$ partagent la série en 4 parties égales.



— $[Q_1, Q_3]$ est ***l'intervalle interquartile***, il contient 50% des observations.
— $EIQ = Q_3 - Q_1$ est ***l'écart interquartile***.

Exemple 1.21. *On reprend l'exemple des tiges, dont le tableau statistique est le suivant :*

Classe	*Effectif*	*Fréquence*	*Fréquence cumulée*
[36,5 ;37,5[	*3*	*0,05*	*0,05*
[37,5 ;38,5[	*7*	*0,12*	*0,17*
[38,5 ;39,5[	*17*	*0,28*	***0,45***
*[39,5 ;**40,5[***	*18*	*0,3*	***0,75***
[40,5 ;41,5[	*9*	*0,15*	*0,90*
[41,5 ;42,5[	*4*	*0,07*	*0,97*
[42,5 ;43,5]	*2*	*0,03*	*1*
Total	*60*	*1*	

— $0,25 \in [0,17; 0,45] \Rightarrow Q_1 \in [38,5; 39,5]$ *et*

$$Q_1 = 38,5 + \frac{(39,5-38,5)(0,25-0,17)}{0,45-0,17} = 38,7857.$$

— $Q_2 = M_e = 39,6667$ *calculée précédemment.*

— $Q_3 = 40,5$ *se lit directement de la table ci-dessus.*

— $EIQ = Q_3 - Q_1 = 40,5 - 38.7857 = 1.7143.$

1.6.3 Variance et écart-type

Pour mesurer la dispersion d'une série, on peut s'intéresser à la moyenne des carrées des distances des valeurs à la moyenne. il s'agit de la variance, notée $V(x)$ (ou aussi s^2), qui est toujours strictement positive et d'unité le carré de l'unité de la distribution. Elle se calcule de la manière suivante :

– ***Cas discret :***

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

où k est le nombre de modalités discrètes et $\bar{x} = \frac{1}{n}\sum_{i=1}^{k} n_i x_i$ est la moyenne de la distribution.

– ***Cas continu :***

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

où k est le nombre de classes ; $c_i = (v_i + v_{i+1})/2$ et le centre de la classe $[v_i, v_{i+1}[$ et et $\bar{x} = \frac{1}{n}\sum_{i=1}^{k} n_i c_i$ est la moyenne de la distribution.

– ***Cas où les données ne sont pas groupées :***

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

où $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ est la moyenne de la distribution.

Remarque 1.8 (Formule de Konig)**.** *Par le théorème de Konig on peut simplifier le calcul de la variance de la manière suivante (c'est la moyenne des carrés moins le carré de la moyenne) :*

– ***Cas discret :***

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

– ***Cas continu :***

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

– ***Cas où les données ne sont pas groupées :***

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

Notons que l'unité de la variance est le carré de l'unité de la distribution.

Pour revenir à l'unité de la distribution, on introduit, l'**écart-type** qui est la racine carrée de la variance, noté σ (ou encore s). Sa formule est :

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

Exemple 1.22. *On reprend l'exemple du nombre d'enfants des* 40 *femmes (page 17) :*

On rappelle que $\overline{x} = 2,675$.

La variance est alors

$$V(X) = \frac{4 \times (0 - 2,675)^2 + 8 \times (1 - 2,675)^2 + ... + 3 \times (6 - 2,675)^2}{40}$$
$$= \frac{4 \times 0^2 + 8 \times 1^2 + ... + 3 \times 6^2}{40} - 2,675^2$$

c'est à dire

$$V(X) = 3.019375$$

Et l'écart-type est :

$$s \simeq 1.7376$$

Exemple 1.23. *On reprend l'exemple des lampes (page 13)*
On rappelle que $\overline{x} = 408,7333$.
La variance est alors

$$V(X) = \frac{2 \times (340,5 - 408,7333)^2 + 3 \times (363,5 - 408,7333)^2 ... + 5 \times (455,5 - 408,7333)^2}{30}$$
$$= \frac{2 \times 340,5^2 + 3 \times 363,5^2 ... + 5 \times 455,5^2}{30} - 408,7333^2$$

c'est à dire

$$V(X) \simeq 1110.3395$$

Et l'écart-type est :

$$s \simeq 33.3217$$

Il existe d'autres indicateur de dispersion, on cite par exemple :

Remarque 1.9 (L'écart moyen absolu et L'écart médian absolu)**.**

— *L'écart moyen absolu, noté e_{moy}, est la somme des valeurs absolues des écarts à la* ***moyenne*** *divisée par le nombre d'observations :*

$$e_{moy} = \frac{1}{n}\sum_{i=1}^{n} |x_i - \bar{x}|,$$

— *L'écart médian absolu, noté e_{med}, est la somme des valeurs absolues des écarts à la* ***médiane*** *divisée par le nombre d'observations :*

$$e_{med} = \frac{1}{n}\sum_{i=1}^{n} |x_i - M_e|,$$

1.6.4 Moment et Moment centré

Définition 1.8 (Moment et Moment centré)**.**

— *On appelle moment d'ordre $r \in \mathbb{N}$, noté m_r, le paramètre*

$$m_r = \frac{1}{n}\sum_{i=1}^{n} x_i^r.$$

— *On appelle moment centré d'ordre* $r \in \mathbb{N}$, *noté* μ_r, *le paramètre*

$$\mu_r = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^r.$$

Les moments généralisent la plupart des paramètres. On a en particulier : $m_1 = \bar{x}$, $\mu_2 = V(X)$, ...

Nous allons voir plus loin que les moments d'ordres supérieurs ($r = 3, 4$) sont utilisés pour mesurer la symétrie et l'aplatissement d'une distribution.

Les formule données concerne les séries non groupées, pour retrouver les autres formules, il suffit de procéder de la même manière que dans le calcul de la variance par exemple.

1.6.5 Coefficient de variation

Définition 1.9. *On appelle* ***coefficient de variation*** *d'une variable* X *le nombre, noté* CV, *définit par :*

$$CV = \frac{s}{\bar{x}}$$

— *Si* $CV < 0.15$ *la série statistique est dite* ***très homogène***,
— *Si* $0.15 \leq CV < 0.85$ *la série statistique est dite* ***homogène***,
— *Si* $CV \geq 0.85$ *la série statistique est dite* ***non homogène***,

Remarque 1.10. *Le coefficient de variation permet d'apprécier la représentativité de la moyenne par rapport à l'ensemble des données. Une série est fortement dispersée quand sont coefficient de variation est supérieur à* 0.85. *Elle est faiblement dispersée quand le coefficient de variation est proche de* 0.

Le coefficient de variation permet aussi de comparer les dispersions de plusieurs séries qui ne sont pas exprimées dans les mêmes unités ou des séries ayant des moyennes différentes.

Exemple 1.24. *Un candidat à un examen a obtenu les notes suivantes :*

$$x_1 = 15\ ;\ x_2 = 10\ ;\ x_3 = 6\ ;\ x_4 = 9\ ;\ x_5 = 11\ ;\ x_6 = 5\ ;\ x_7 = 12\ ;\ x_8 = 7\ ;$$
$$x_9 = 16\ ;\ x_{10} = 8$$

La moyenne est $\bar{x} = \frac{15+10+...+16+8}{10} = 9,9$.

La variance est $s^2 = V(X) = \frac{15^2+10^2+...+16^2+8^2}{10} - 9,9^2 = 12,09$, *et l'écart type est* $s = 3,4771$.

Donc le coefficient de variation est

$$CV = \frac{3,4771}{9,9} \simeq 0,35$$

Comme $0,15 \leq 0,35 < 0,85$, *la série est* ***homogène***.

1.7 Mesures de forme

1.7.1 Coefficient d'asymétrie

Définition 1.10 (Coefficient d'asymétrie de Fisher)**.**
Le ***coefficient d'asymétrie*** *de Fisher est défini par :*

$$\gamma_3 = \frac{\mu_3}{s^3}$$

— Si $\gamma_3 > 0$ la série présente une **asymétrie à droite**,
— Si $\gamma_3 = 0$ la série est **symétrique**,
— Si $\gamma_3 < 0$ la série présente une **asymétrie à gauche**.

Définition 1.11 (Coefficient d'asymétrie de Pearson)**.**
Le ***coefficient d'asymétrie*** *de Pearson est défini par :*

$$A_P = \frac{3(\bar{x} - M_e)}{s}$$

— Si $A_P > 0$ la série présente une **asymétrie à droite**,
— Si $A_P = 0$ la série est **symétrique**,
— Si $A_P < 0$ la série présente une **asymétrie à gauche**.

En fait, pour ce coefficient, comme déjà mentionné dans la figure 6, il suffit de comparer la moyenne et la médiane.

Définition 1.12 (Coefficient d'asymétrie de Yule)**.**

Le ***coefficient d'asymétrie*** *de Yule est basé sur les positions des* 3 *quartiles* (Q_1, M_e, Q_3)*, et est normalisé par la distance interquartile :*

$$A_Y = \frac{Q_3 + Q_1 - 2M_e}{Q_3 - Q_1}.$$

— Si $A_Y > 0$ la série présente une **asymétrie à droite**,
— Si $A_Y = 0$ la série est **symétrique**,
— Si $A_Y < 0$ la série présente une **asymétrie à gauche**.

Exemple 1.25. *On reprend l'exemple des notes.*

$$x_1 = 15\ ;\ x_2 = 10\ ;\ x_3 = 6\ ;\ x_4 = 9\ ;\ x_5 = 11\ ;\ x_6 = 5\ ;\ x_7 = 12\ ;\ x_8 = 7\ ;$$
$$x_9 = 16\ ;\ x_{10} = 8.$$

On a $\bar{x} = 9.9$*,* $M_e = 9.5$ *donc* $A_P = \dfrac{3(9.9 - 9.5)}{s} > 0$*. Donc la série est étalée vers la droite.*

1.7.2 Coefficient d'aplatissement

Définition 1.13 (Coefficient d'aplatissement de Fisher)**.** *Le coefficient d'aplatisseme de Fisher, noté* γ_4 *est défini par la relation :*

$$\gamma_4 = \frac{\mu_4}{s^4} - 3,$$

— Si $\gamma_4 > 0$ la série présente une **léptokurtique** (aigue),
— Si $\gamma_4 \simeq 0$ la série est **mésokurtique** (normale),
— Si $\gamma_4 < 0$ la série présente une **platykurtique** (aplatie).

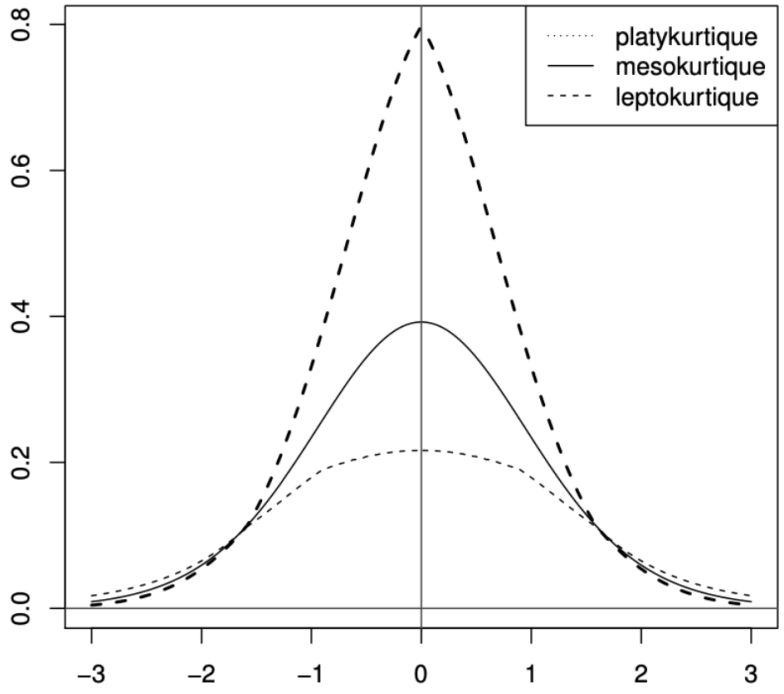


FIGURE 7 – Comparaison des aplatissements

1.8 Diagramme en boite (Box plot)

Définition 1.14. *La **boîte à moustaches** (ou **diagramme en boîte**, ou encore **box plot** en anglais) est un diagramme résumant les indicateurs d'une série statistique : médiane, quartiles Q_1 et Q_3. Elle est généralement utilisée pour comparer plusieurs séries de même unité.*

Ce diagramme est composé de :

— un rectangle qui s'étend du premier au troisième quartile. Le rectangle est divisé par une ligne correspondant à la médiane,

— ce rectangle est complété par deux demis-segments (moustaches) de limites :

$$a = \max(Q_1 - 1.5 \times EIQ; x_{\min}) \qquad \text{et} \qquad b = \min(Q_3 + 1.5 \times EIQ; x_{\max}),$$

— on identifie ensuite la plus petite et la plus grande observation comprise entre ces bornes. Ces observations sont appelées "**valeurs adjacentes**",

— on trace les segments de droites reliant ces observations au rectangle,

— les valeurs qui ne sont pas comprises entre les valeurs adjacentes, sont représentées par des points et sont appelées "**valeurs extrêmes**".

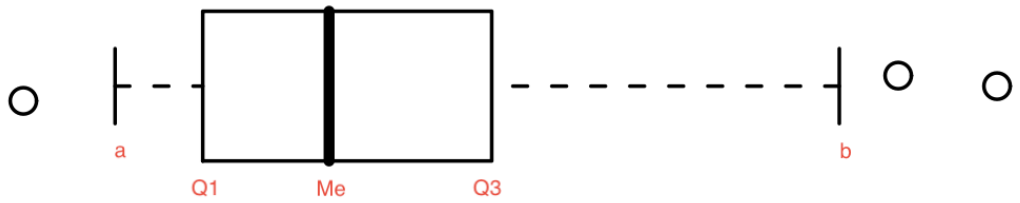


FIGURE 8 – Exemple de diagramme en boite

1.9 Changement d'origine et d'unité

Définition 1.15. *On appelle changement d'**origine** l'opération consistant à ajouter (ou soustraire) la même quantité $a \in R$ à toutes les observations x_i d'une variable X de taille n et on note la nouvelle variable Y ayant les observations y_i telle que :*

$$y_i = a + x_i, \qquad i = 1, ..., n$$

Définition 1.16. *On appelle changement d'**unité** l'opération consistant à multiplier (ou diviser) par la même quantité $b \in R$ toutes les observations x_i d'une variable X de taille n et on note la nouvelle variable Y ayant les observations y_i telle que :*

$$y_i = bx_i, \qquad i = 1, ..., n$$

Définition 1.17. *On appelle changement d'**origine** et d'**unité** l'opération consistant à multiplier (ou diviser) par la même quantité $b \in R$ toutes les observations x_i d'une variable X de taille n puis rajouter (ou soustraire) la même quantité $b \in R$ et on note la nouvelle variable Y ayant les observations y_i telle que :*

$$y_i = a + bx_i, \qquad i = 1, ..., n$$

Propriétés 1.1. *Si on effectue un changement d'origine et d'unité sur une variable X, alors*

— *sa moyenne est affectée du même changement d'origine et d'unité :*

$$\bar{y} = a + b\bar{x};$$

— *sa variance est affectée par le carré du changement d'unité et pas par le changement d'origine :*

$$V(Y) = b^2 V(X).$$

Remarque 1.11. — *Les* ***paramètres de position*** *sont tous affectés par un changement d'origine et d'unité.*

— *Les* ***paramètres de dispersion*** *sont tous affectés par un changement d'unité mais pas par un changement d'origine.*

— *Les* ***paramètres de forme*** *et d'****aplatissement*** *ne sont affectés ni par un changement d'unité ni par un changement d'origine.*

Définition 1.18 (Variable centrée réduite)**.** — *Une variable est dite centrée si sa moyenne est nulle.*

— *Une variable est dite réduite si sa variance est égale à* 1*.*

— *Une variable est dite centrée et réduite si sa moyenne est nulle et sa variance est égale à* 1*.*

Propriétés 1.2. *Soit X une variable ayant la moyenne $\bar{x}$ et la variance $V(X)$, alors la variable Z telle que :*

$$Z = \frac{X - \bar{x}}{\sigma_X}.$$

est ***centrée et réduite****.*

1.10 Exercice corrigé

Exercice

On a relevé la taille (en cm) de 50 étudiantes de la filière **SMI**, les résultats sont regroupés dans le tableaux suivant

Classe	$[151.5, 155.5[$	$[155.5, 159.5[$	$[159, 5; 163, 5[$	$[163, 5; 167, 5[$	$[167, 5; 171, 5[$
Effectif	10	12	11	7	10

1. Caractériser la distribution *(la population et sa taille, l'individu, la variable et son type).*
2. Dresser le tableau statistique complet (calculer les fréquences, les fréquences cumulées et les effectifs cumulés)
3. Tracer le diagramme correspondant.
4. Quelle est la classe modale ?
5. Définir et représenter la courbe cumulative croissante.
6. Calculer la moyenne et la variance.
7. Calculer le coefficient de variation. Interpréter le résultat.

8. Calculer la médiane ainsi que le premier et le troisième quantile.

9. Quelle est la fréquence des étudiantes ayant au moins $165cm$?

Corrigé

(1) Caractériser la distribution *(la population et sa taille, l'individu, la variable et son type).*
Population étudiée : Les étudiantes de la filière SMI ; Taille : 50 ;
L'individu : une étudiante de la filière SMI ;
Variable : "taille en cm des étudiantes" ; Type : Quantitative continue.

(2) Le tableau statistique est le suivant :

Classe	n_i	f_i	F_i	N_i
$[151.5, 155.5[$	10	0.20	0.20	10
$[155.5, 159.5[$	12	0.24	0.44	22
$[159, 5; 163, 5[$	11	0.22	0.66	33
$[163, 5; 167, 5[$	7	0.14	0.80	40
$[167, 5; 171, 5[$	10	0.20	1.00	50
Σ	50	1.00	//	//

(3) Le diagramme correspondant : Puisque la variable est quantitative continue, on trace l'histogramme des effectif ou des fréquence. Et puisque les classes

sont d'amplitudes égales alors on trace directement l'histogramme.

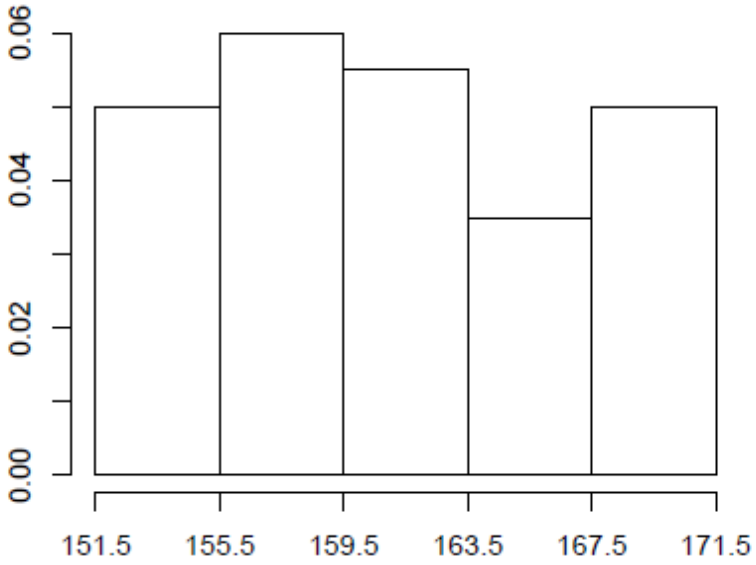


FIGURE 9 – Histogramme des fréquences

(4) Puisque les classes sont d'amplitudes égales alors on retrouve directement la classe qui contient l'effectif (ou la fréquence) le plus élevé(e) : il s'agit de la classe des taille entre 155.5 et 159.5 centimètre.

(5) La courbe cumulative croissante (fonction de répartition) est définie par les points $A_i(x_{i+1}, F_i)$ donnés dans le tableau statistique.

(6) La moyenne : $\bar{x} = \frac{1}{n}\sum_{i=1}^{k} n_i c_i = \sum_{i=1}^{k} f_i c_i$, avec $c_i = \frac{x_i + x_{i+1}}{2}$ est le centre

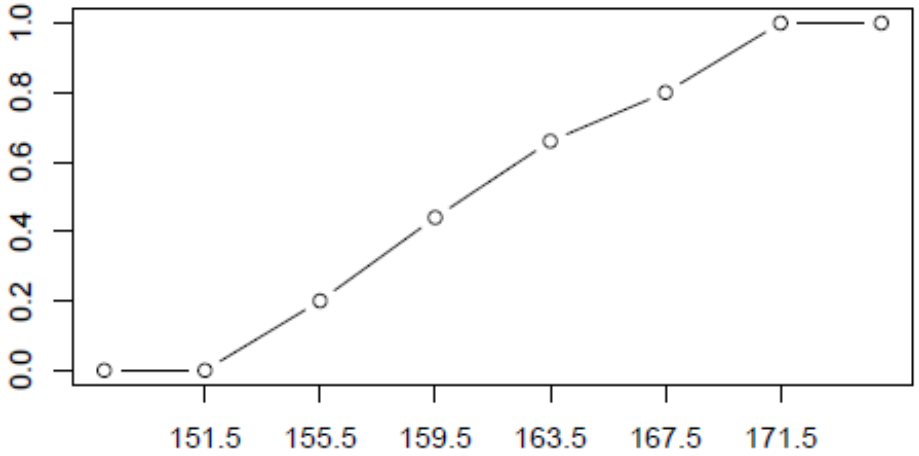


FIGURE 10 – Fonction cumulative (fonction de répartition)

de la classe $[x_i, x_{i+1}[$.

$$\bar{x} = \frac{10 \times 153.5 + 12 \times 157.5 + 11 \times 161.5 + 7 \times 165.5 + 10 \times 169.5}{50}$$
$$= 0.20 \times 153.5 + 0.24 \times 157.5 + 0.22 \times 161.5 + 0.14 \times 165.5 + 0.20 \times 169.5$$
$$= 161.1cm.$$

(6) La variance : $S^2 = \frac{1}{n}\sum_{i=1}^{k} n_i c_i^2 - (\bar{x})^2 = \sum_{i=1}^{k} f_i c_i^2 - (\bar{x})^2$.

$$\frac{1}{n}\sum_{i=1}^{k} n_i c_i^2 = \frac{10 \times 153.5^2 + 12 \times 157.5^2 + 11 \times 161.5^2 + 7 \times 165.5^2 + 10 \times 169.5^2}{50}$$

$$= 25984.73 cm^2$$

$$\sum_{i=1}^{k} f_i c_i^2 = 0.20 \times 153.5^2 + 0.24 \times 157.5^2 + 0.22 \times 161.5^2 + 0.14 \times 165.5^2 + 0.20 \times 169.5^2.$$

$$S^2 = 25984.73 - 161.1^2 = 31.52 cm^2.$$

(7) Le coefficient de variation

$$CV = \frac{S}{\bar{x}} \times 100 = \frac{\sqrt{31.52}}{161.1} \times 100 = 03.49\%.$$

Interprétation : la série est très homogène.

Classe	F_i	N_i
$[151.5, 155.5[$	0.20	10
$[155.5, 159.5[$	0.44	22
$[159, 5; 163, 5[$	0.66	33
$[163, 5; 167, 5[$	0.80	40
$[167, 5; 171, 5[$	1.00	50

(8) La médiane : $M_e \in]159, 5; 163, 5[$:

$$M_e = 159.5 + \frac{0.50 - 0.44}{0.66 - 0.44} \times (163.5 - 159.5) \simeq 160.59cm$$

Le premier quartile : $Q_1 \in]155, 5; 159, 5[$:

$$Q_1 = 155.5 + \frac{0.25 - 0.20}{0.44 - 0.20} \times (159.5 - 155.5) \simeq 156.33cm$$

Le troisième quartile : $Q_3 \in]163, 5; 167, 5[$:

$$Q_3 = 163.5 + \frac{0.75 - 0.66}{0.80 - 0.66} \times (167.5 - 163.5) \simeq 166.07cm$$

$$\Rightarrow \qquad EIQ = Q_3 - Q_1 \simeq 9.74cm$$

(9) Quelle est la fréquence des étudiantes ayant au moins $165cm$?

Par interpolation, on cherche d'abord la fréquence f des étudiantes ayant moins de $165cm$: puisque $165 \in]163,5; 167,5[$, alors par interpolation linéaire on a :

$$\frac{f - 0.66}{0.80 - 0.66} = \frac{165 - 163.5}{167.5 - 163.5}$$

qui donne $f = 0.66 + \frac{165-163.5}{167.5-163.5} \times (0.80 - 0.66) = 0.7125$

Donc la proportion (fréquence) des étudiantes ayant au moins $165cm$ est égale à $1 - 0.7125 = 0.2875$.

2 Statistique descriptive bivariée

2.1 Introduction

La statistique descriptive bivariée permet de décrire simultanément deux variables et, par le fait même, de donner une information sur la relation possible entre les deux variables. Si les valeurs de la première variable sont affectées par celles de la seconde variable, on dira que les deux variables sont liées ; à l'inverse, s'il n y a pas de lien entre les deux variables on dira qu'elles sont indépendantes.

L'étude statistique peut se porter sur n'importe quel type de ces variables, on peut avoir :

— deux variables qualitatives ;

— une variable quantitative et l'autre qualitative ;

— deux variable quantitative.

Dans ce chapitre, on va traiter le cas des deux variables quantitatives (continues ou discrètes). Ces deux variables seront représentées par X et Y .

Formellement, on considère une série statistique double non groupée

$$\{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$$

de n observations mesurées par les deux variables X et Y simultanément.

Ces données peuvent être groupées dans un tableau (comme dans le cas d'une variable, par modalité/effectif). Ce tableau est appelé tableau croisé (distribution conjointe) :

	y_1	$\dots$	y_j	$\dots$	y_J
x_1	n_{11}	$\dots$	n_{1j}	$\dots$	n_{1J}
$\vdots$		$\vdots$		$\vdots$	
x_i	n_{i1}	$\dots$	n_{ij}	$\dots$	n_{iJ}
$\vdots$		$\vdots$		$\vdots$	
x_I	n_{I1}	$\dots$	n_{Ij}	$\dots$	n_{IJ}

2.2 Nuage de points

On représente dans le plan une distribution statistique à deux variables quantitatives par un ensemble de points A_i ($i \in \{1, ..., n\}$). Les coordonnées du point A_i sont (x_i, y_i). Chaque point représente alors un individu de la population.

On considère l'exemple suivant :

X	2	6	7	5	4	1	3
Y	5	9	12	9	8	1	5

Chaque observation est un point dans le plan. On peut donc représenter ces données sous forme d'un **nuage de points** comme montré dans la figure ci-dessous.

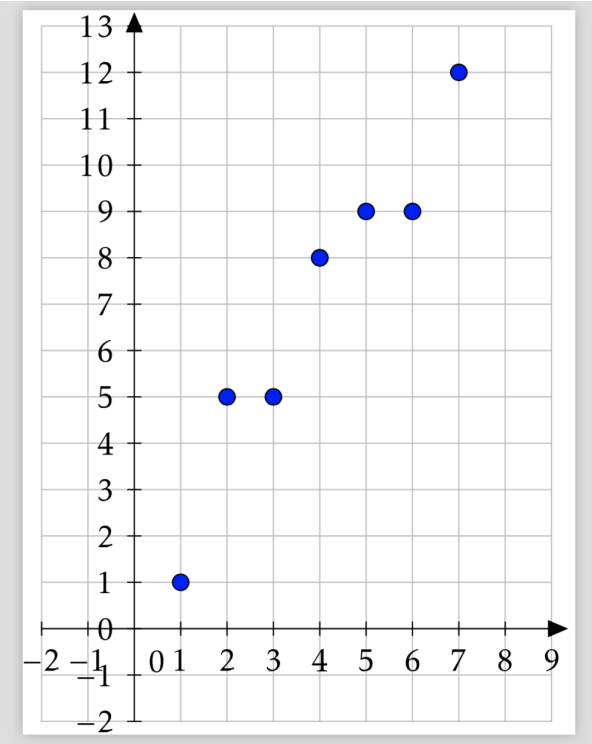


FIGURE 11 – Nuage de points du couple (X, Y)

2.3 Ajustement linéaire

2.3.1 Covariance & corrélation

On considère une série statistique double (non groupée) mesurée par deux variables X et Y ayant n observations : $\mathcal{S}_X = \{x_1, ..., x_n\}$ et $\mathcal{S}_Y = \{y_1, ..., y_n\}$

La liaison entre les deux variables X et Y est mesurée à travers plusieurs

indicateurs, on cite en particulier :

Définition 2.1. *La* ***Covariance*** *entre deux variables X et Y, notée $cov(X,Y)$, est donnée par la formule suivante :*

$$cov(X,Y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \left(\frac{1}{n}\sum_{i=1}^{n} x_i y_i\right) - \bar{x}\bar{y}.$$

On remarque que la covariance dépend des unités de mesures dans lesquelles sont exprimées les variables. Pour enlever l'effet de ces unité on réduit par les écarts-type des deux variable, d'où la définition du coefficient de corrélation :

Définition 2.2. *Le* ***coefficient de corrélation linéaire*** *entre X et Y, noté $cor(X,Y)$ (ou encore r ou ρ), est donné par :*

$$cor(X,Y) = \frac{cov(X,Y)}{\sqrt{V(X)}\sqrt{V(Y)}}$$

Le **coefficient de corrélation linéaire** entre X et Y est un indicateur sans unité permet de mesurer l'**intensité de la liaison** entre les deux variables X et Y et il est toujours compris entre -1 et 1.

2.3.2 Propriétés

On a :

— toujours $-1 \leq cor(X, Y) \leq 1$,

— si $cor(X, Y)$ est proche de 1 alors les variables X et Y sont **positivement corrélées** : si X croit alors Y croit (et vis versa) linéairement,

— si $cor(X, Y)$ est proche de -1 alors les variables X et Y sont **négativement corrélés** : si X croit alors Y décroît (et vice versa) linéairement,

— pratiquement, si $|cor(X, Y)| \geq 0,8$ alors les deux variables X et Y sont **fortement corrélées**,

— si $cor(X, Y)$ est proche de 0 alors les variables X et Y sont **non corrélées** : si X croit (ou décroît) ce n'influence pas sur le comportement de Y.

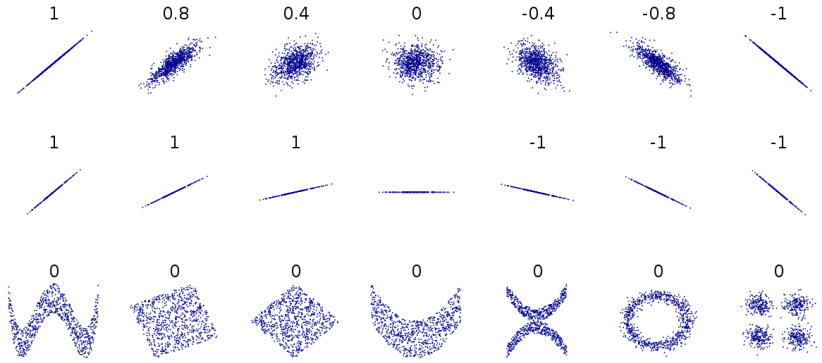


FIGURE 12 – Différentes situations de corrélation entre deux variables

Comme mentionné, le coefficient de corrélation peut être négatif. On définie **le coefficient de détermination** qui est égal au carré du coefficient de corrélation (noté R^2) :

$$R^2 = (cor(X, Y))^2$$

et qui mesure l'adéquation entre le modèle et les données observées ou encore à quel point l'équation de régression est adaptée pour décrire la distribution des points.

Exemple 2.1. *Considérons la série double précédente.*
Calculons $cov(X, Y)$ *et* $cor(X, Y)$.
D'abord, les moyennes de X *et* Y *sont respectivement :*

$$\bar{x} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{2+6+7+5+4+1+3}{7} = \frac{28}{7} = \boldsymbol{4}$$

$$\bar{y} = \frac{\sum_{i=1}^{n} Y_i}{n} = \frac{5+9+12+9+8+1+5}{7} = \frac{49}{7} = \boldsymbol{7}$$

Donc la covariance entre X *et* Y *est :*

$$cov\,(X, Y) = \frac{-2 \times (-2) + 2 \times 2 + 3 \times 5 + 1 \times 2 + 0 \times 1 - 3 \times (-6) - 1 \times (-2)}{7}$$

$$= \frac{4+4+15+2+0+18+2}{7} = \frac{45}{7} \simeq \boldsymbol{6.4286}$$

De plus,

$$V(X) = \frac{(-2)^2+2^2+3^2+1^2+0^2+(-3)^2+(-1)^2}{7}$$
$$= \frac{4+4+9+1+0+9+1}{7} = \frac{28}{7} = \boldsymbol{4}$$

$$V(Y) = \frac{(-2)^2+2^2+5^2+2^2+1^2+(-6)^2+(-2)^2}{7}$$
$$= \frac{4+4+25+4+1+36+4}{7} = \frac{78}{7} \simeq \boldsymbol{11.1429}$$

Donc le coefficient de corrélation entre X et Y est :

$$cor\,(X,Y) = \frac{cov\,(X,Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}} = \frac{6.4286}{\sqrt{4}\times\sqrt{11.1429}} = \boldsymbol{0.9629}$$

On a $cor(X,Y) \simeq 0.9629 \geq 0,8$ donc les variables X et Y sont ***très corrélées positivement.***

2.4 Droite de régression

Lorsque X et Y sont fortement corrélées alors Y est liée linéairement à X. C'est à dire, on peut écrire une équation linéaire entre les deux variables de la forme

$$Y = aX + b.$$

Cette droite est appelée **droite de régression**, elle est la droite qui ajuste au mieux un nuage de points au sens des *moindres carrés*. La variable X s'appelle la variable **explicative** et la variable Y s'appelle la variable **dépendante**.

Comment trouver les meilleurs valeurs (estimations) $\hat{a}$ et $\hat{b}$ de a et b qui ajustent au mieux le nuage ? on doit minimiser les résidus (erreurs) entre les valeurs réelle y_i et les valeurs ajustées (notée y_i^*) .

Il s'agit du principe des moindres carrés qui consiste à chercher la droite qui minimise la somme des carrés des résidus $e_i = y_i - \hat{a}x - \hat{b}$ pour tout $i = 1, ..., n$.

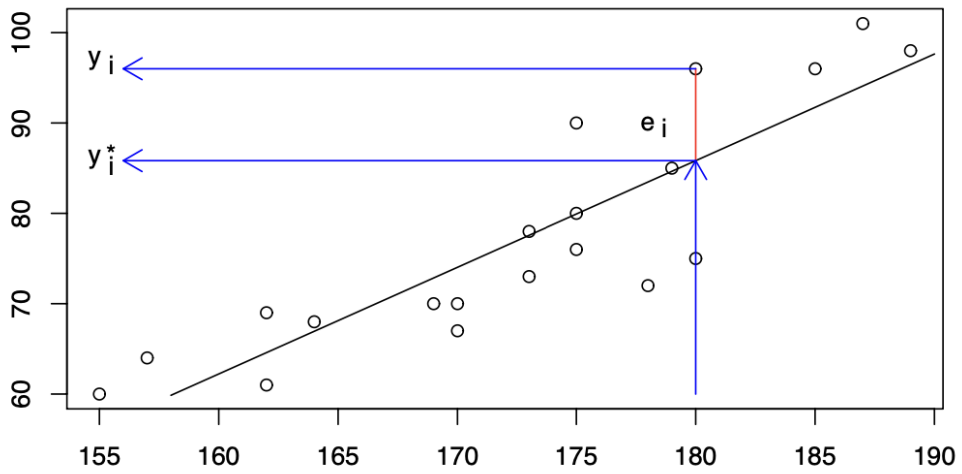


FIGURE 13 – Exemple de nuage de points et résidu de l'équation de régression

On définit la fonction suivante f à deux variables a et b tel que :

$$f(a,b) = \sum_{i=1}^{n} (y_i - ax_i - b)^2.$$

Alors les meilleures valeurs $\hat{a}$ et $\hat{b}$ s'obtiennent en utilisant une méthode très connue appelée **méthode des moindres carrées ordinaires**. Cette méthode consiste à minimiser la fonction f.

En utilisant des techniques de dérivation on trouve :

Théorème 2.1. *Les coefficients a et b qui minimisent le critère des*

moindres carrés ordinaire sont donnés par :

$$\hat{a} = \frac{cov(X,Y)}{Var(X)} \qquad et \qquad \hat{b} = \overline{y} - \hat{a}\overline{x}.$$

Exercice. *Démontrer ce théorème.*

Remarque 2.1. — $\hat{Y} = \hat{a}X + \hat{b}$ *s'appelle l'équation de la* ***droite de régression*** *de* Y *en* X.

— $\hat{a}$ *est la pente de la droite.*

— $\hat{b}$ *est l'ordonnée à l'origine.*

— $\hat{a} = \dfrac{cov(X,Y)}{V(X)} = \dfrac{cov(X,Y)\sqrt{V(Y)}}{V(X)\sqrt{Var(Y)}} = cor(X,Y)\dfrac{\sigma_Y}{\sigma_X}$
où σ_X *(respectivement* σ_Y*) est l'écart- type de* X *(respectivement de* Y*).*

— *La droite de régression obtenue passe toujours par le point moyen* $(\bar{x}, \bar{y})$.

— *La droite de régression de* Y *en* X *n'est pas la même que la droite de régression de* X *en* Y.

Exemple 2.2. *Reprenons l'exemple précédent. La pente est :*

$$\hat{a} = \frac{cov(X,Y)}{V(X)} = \frac{6.4286}{4} = 1.60715$$

et l'ordonnée à l'origine est :

$$\hat{b} = \overline{y} - \hat{a}\overline{x} = 7 - 1,60715 \times 4 = 0.5714$$

La droite de régression a pour équation :

$$\hat{Y} = 1.60715X + 0.5714.$$

Cette droite est représentée dans la figure suivante :

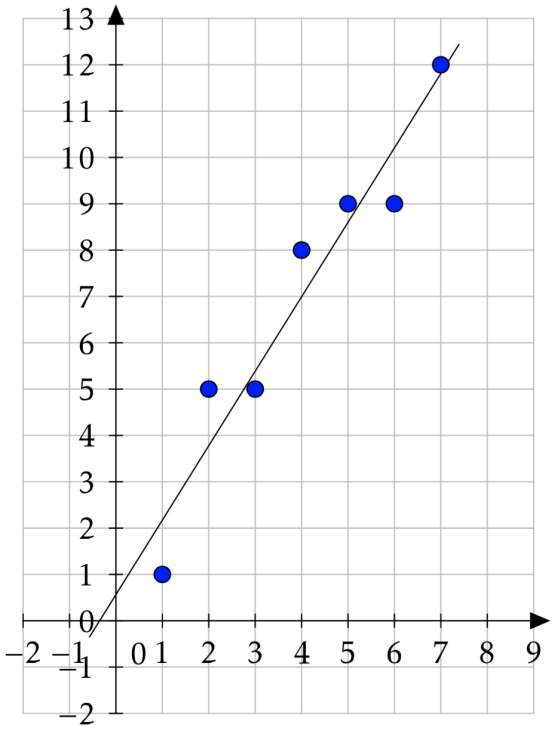


FIGURE 14 – Ajustement linéaire de Y en X

2.5 Prédiction

La droite de régression permet de donner une approximation future.

Si une nouvelle valeur de variable X est disponible alors on peut calculer, par l'équation de la droite de régression linéaire, la valeur **prédite** pour la variable Y correspondante.

Exemple 2.3. *Supposons que, pour la série double de données précédente, on ait la nouvelle valeur* $x_8 = 4.5$. *Alors la valeur prédite de* x_8 *par la droite de régression est égale à :*

$$\hat{y}_8 = \hat{a}x_8 + \hat{b} = 1.60715 \times 4.5 + 0.5714 \simeq \boldsymbol{7.8075}.$$

2.6 Distribution conjointe (tableau croisé)

De façon générale, quand on étudie simultanément les deux variables quantitatives X et Y, ayant respectivement les modalités $x_1, x_2, ..., x_I$ et $y_1, y_2, ..., y_J$, alors le tableau de la distribution conjointe (ou tableau croisé) des deux variables est présenté comme suit ($n = \sum_{i=1}^{I} n_{i.} = \sum_{j=1}^{J} n_{.j} = \sum_{i=1}^{I} \sum_{j=1}^{J} n_{ij}$) :

	y_1	...	y_j	...	y_J	total
x_1	n_{11}	...	n_{1j}	...	n_{1J}	$n_{1.}$
⋮		⋮		⋮		⋮
x_i	n_{i1}	...	n_{ij}	...	n_{iJ}	$n_{i.}$
⋮		⋮		⋮		⋮
x_I	n_{I1}	...	n_{Ij}	...	n_{IJ}	$n_{I.}$
total	$n_{.1}$	...	$n_{.j}$	...	$n_{.J}$	$n_{..} = n$

Les modalités x_i et y_j peuvent être des valeurs discrètes (dans le cas d'une variable quantitative discrète) ou intervalles (dans le cas d'une variable quantitative continue).

Les $n_{i.}$ et $n_{.j}$ sont appelées les **effectifs marginaux**. Dans ce tableau,

— $n_{i.}$ représente le nombre de fois que la modalité x_i apparait,

— $n_{.j}$ représente le nombre de fois que la modalité y_j apparait,
— n_{ij} (resp. $f_{ij} = \frac{n_{ij}}{n}$) représente le nombre de fois (resp. la fréquence) que les modalités x_i et y_j apparaissent ensemble.

On a : $\sum_{i=1}^{I} n_{ij} = n_{.j}$, pour $j = 1, ..., J$, $\sum_{j=1}^{J} n_{ij} = n_{i.}$, pour $i = 1, ..., I$,

2.6.1 Distribution marginale

Le tableau croisé compte deux distributions marginales : la distribution marginale de X et la distribution marginale de Y.

Distribution marginale de X : elle est composée des modalités de la variable X et les effectifs marginaux correspondants quelles que soit la valeur de la modalité de Y.

X	x_1	x_2	...	x_I	total
$n_{i.}$	$n_{1.}$	$n_{2.}$	...	$n_{I.}$	n

Distribution marginale de Y : elle est composée des modalités de la variable Y et les effectifs marginaux correspondants quelles que soit la valeur de la modalité de X.

Y	y_1	y_2	...	y_J	total
$n_{.j}$	$n_{.1}$	$n_{.2}$	...	$n_{.J}$	n

Remarque 2.2. *On aura des distributions à une variable ; ce qui permettra d'appliquer toutes les propriétés vues dans la statistique descriptive univariée : moyenne, variance, écart-type, coefficient de variation, médiane, quantiles, mode, ... ; toutes ces propriétés seront appelées des indicateurs marginaux. Exemple : moyenne marginale de X, écart-type marginal de Y, ...*

2.6.2 Covariance et corrélation

— La **Covariance** entre X et Y est donnée par :

$$cov(X,Y) = \frac{1}{n}\sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij}(x_i - \bar{x})(y_j - \bar{y}) = \left(\frac{1}{n}\sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij}x_i y_j\right) - \bar{x}\bar{y}.$$

— Le coefficient de corrélation entre X et Y est donné par :

$$cor(X,Y) = \frac{cov(X,Y)}{\sqrt{V(X)}\sqrt{V(Y)}}.$$

— Pour tout $a, b \in \mathbb{R}$, on a : $V(aX+bY) = a^2V(X)+b^2V(Y)+2ab\,cov(X,Y)$.
— Les variables X et Y sont indépendantes si et seulement si

$$f_{ij} = f_{i.} \times f_{.j} \Longleftrightarrow n_{ij} = \frac{n_{i.} \times n_{.j}}{n}, \qquad \text{pour } i = 1, ..., I \text{ et } j = 1, ..., J.$$

— Les variables X et Y sont indépendantes si et seulement si les lignes (resp. colonnes) du tableau croisé associé sont proportionnelles entre elles.
— Si $cov(X,Y) = 0$ alors les variables X et Y sont indépendantes.

Exemple 2.4. *Soit le tableau suivant associé à deux variables X et Y.*

X \ Y	-2	0	2	*total*
0	2	4	12	18
1	4	8	24	36
total	6	12	36	54

Les variables X et Y sont indépendantes.

2.7 Exercice corrigé

Exercice

Soit le tableau suivant donnant la distribution du couple (X, Y).

X \ Y	0	1
$[0.5, 1.5[$	21	8
$[1.5, 2.5[$	23	15
$[2.5, 3.5[$	10	23

1. Quelles sont les distributions marginales de X et de Y ?
2. Calculer les moyennes et les variances marginales de X et de Y.
3. Calculer le coefficient de variation marginale de Y. Interpréter.
4. Les variables X et Y sont elles indépendantes ?
5. Calculer la moyenne et la variance de la variable $Z = 0.165X + 0.13Y$.

Corrigé

1. La distribution marginale de X est donnée dans le tableau suivant :

X	effectif
$[0.5, 1.5[$	29
$[1.5, 2.5[$	38
$[2.5, 3.5[$	33
Σ	100

La distribution marginale de Y est donnée dans le tableau suivant :

Y	effectif
0	54
1	46
Σ	100

2. On trouve :

$$\overline{x} = \frac{1}{100}\sum_{i=1}^{3} n_{i.}c_i = \frac{29 \times 1 + 38 \times 2 + 33 \times 3}{100} = 2.04,$$

$$\overline{y} = \frac{1}{100}\sum_{j=1}^{2} n_{.j}y_j = \frac{54 \times 0 + 46 \times 1}{100} = 0.46,$$

$$V(X) = s_x^2 = \left(\frac{1}{100}\sum_{i=1}^{3} n_{i.}c_i^2\right) - (\overline{x})^2 = 4.78 - 2.04^2 = 0.6184,$$

$$V(Y) = s_y^2 = \left(\frac{1}{100}\sum_{j=1}^{2} n_{.j}y_j^2\right) - (\overline{y})^2 = 0.2484.$$

3. $CV_Y = \frac{s_y}{\overline{y}} = \frac{\sqrt{0.2484}}{0.46} = 1.083473 \simeq 108\%$. la distribution de Y est hétérogène.

4. Rappelons que les variables X et Y sont indépendantes si et seulement si

$$n_{ij} = \frac{n_{i.} \times n_{.j}}{n}, \forall i = 1, 2, 3 \quad et \quad j = 1, 2.$$

X \ Y	0	1	Σ
$[0.5, 1.5[$	21	8	29
$[1.5, 2.5[$	23	15	38
$[2.5, 3.5[$	10	23	33
Σ	54	46	100

Or, on a (contre exemple)

$$n_{21} = 23 \neq \frac{n_{2.} \times n_{.1}}{n} = \frac{38 \times 54}{100} = 20.52,$$

donc les variables X et Y sont liées.

5. $V(Z) = V(0.165X + 0.13Y) = 0.165^2 V(X) + 0.13^2 V(Y) + 2 \times 0.165 \times 0.13\ cov(X, Y)$,

   avec, la covariance entre X et Y :

$$s_{xy} = cov(X, Y) = \left(\frac{1}{100} \sum_{i=1}^{3} \sum_{j=1}^{2} n_{ij} c_i y_j \right) - \overline{x} \times \overline{y} = 0.1316$$