

Maurice LETHIELLEUX

Ancien maître de conférence
à l'université Paris II - Panthéon - Assas

Céline CHEVALIER

Maître de conférence
à l'université Paris II - Panthéon - Assas

Exercices de statistique et probabilités

3^e édition

DUNOD

Tout le catalogue sur
www.dunod.com



Le pictogramme qui figure ci-contre mérite une explication. Son objet est d'alerter le lecteur sur la menace que représente pour l'avenir de l'écrit, particulièrement dans le domaine de l'édition technique et universitaire, le développement massif du photocopillage.

Le Code de la propriété intellectuelle du 1^{er} juillet 1992 interdit en effet expressément la photocopie à usage collectif sans autorisation des ayants droit. Or, cette pratique s'est généralisée dans les établissements

d'enseignement supérieur, provoquant une baisse brutale des achats de livres et de revues, au point que la possibilité même pour

les auteurs de créer des œuvres nouvelles et de les faire éditer correctement est aujourd'hui menacée.

Nous rappelons donc que toute reproduction, partielle ou totale, de la présente publication est interdite sans autorisation de l'auteur, de son éditeur ou du

Centre français d'exploitation du droit de copie (CFC, 20, rue des Grands-Augustins, 75006 Paris).



© Dunod, 2017

11 rue Paul Bert, 92240 Malakoff

ISBN 978-2-10-076047-3

Le Code de la propriété intellectuelle n'autorisant, aux termes de l'article L. 122-5, 2^o et 3^o a), d'une part, que les « copies ou reproductions strictement réservées à l'usage privé du copiste et non destinées à une utilisation collective » et, d'autre part, que les analyses et les courtes citations dans un but d'exemple et d'illustration, « toute représentation ou reproduction intégrale ou partielle faite sans le consentement de l'auteur ou de ses ayants droit ou ayants cause est illicite » (art. L. 122-4).

Cette représentation ou reproduction, par quelque procédé que ce soit, constituerait donc une contrefaçon sanctionnée par les articles L. 335-2 et suivants du Code de la propriété intellectuelle.

Sommaire

Fiche 1	Généralités et représentations graphiques des séries à un caractère	1
Fiche 2	Caractéristiques de tendance centrale et de dispersion. Concentration	9
Fiche 3	Indices de prix – Indices en volume – Indices en valeur	26
Fiche 4	Séries statistiques à deux variables : ajustement par les moindres carrés	34
Fiche 5	Les séries chronologiques	47
Fiche 6	Principes du calcul des probabilités	57
Fiche 7	Loi de probabilité d'une variable aléatoire discrète	67
Fiche 8	Variables aléatoires continues et lois de probabilités continues usuelles	83
Fiche 9	Convergences et approximation par la loi de Poisson ou la loi normale	99
Fiche 10	Échantillons et simulations	112
Fiche 11	Estimation ponctuelle	121
Fiche 12	Estimation par intervalles de confiance	136
Tables	Normale, Student, Chi-deux...	148
Annexe	Niveau de difficulté des exercices	156

Généralités et représentations graphiques des séries à un caractère

I Rappels de cours

- **Population** : en statistique descriptive, c'est un ensemble d'**individus**. Chaque individu est décrit selon une ou plusieurs caractéristiques désignées par **variable** ou **caractère**.
- **Unité statistique** : c'est une autre façon de désigner un individu.
- **Modalités** : ce sont les différentes caractéristiques d'une variable. Chaque individu présente une et une seule modalité à la fois (exhaustivité et disjonctivité).
- **Variable quantitative** : les modalités sont mesurables ou repérables. Lorsque ces modalités sont des nombres isolés, cette variable est quantitative **discrète** ; sinon cette variable est quantitative **continue**.
- **Variable qualitative** : les différentes modalités ne sont pas mesurables ou repérables.
- **Variable qualitative ordinale** : on peut établir une hiérarchie entre les modalités.
- **Sondage** : l'information est recueillie sur une partie de la population qui constitue un **échantillon**.
- **Série statistique** : suite de données (ou de variables) recueillies concernant des individus.

Sur une population ou un échantillon de n individus, chaque individu présente l'une des p modalités de la variable. Ces modalités sont notées $M_1, M_2, \dots, M_i, \dots, M_p$.
 $n_1, n_2, \dots, n_i, \dots, n_p$ sont les effectifs ou fréquences absolues des différentes modalités.

$f_i = \frac{n_i}{n}, f_1, f_2, \dots, f_i, \dots, f_p$, sont les fréquences relatives des diverses modalités.

Les fréquences peuvent être exprimées en pourcentage.

La distribution statistique d'une variable selon ses modalités est présentée dans un tableau.

Modalités	Effectifs n_i	Fréquences f_i	Fréquences en %
M_1	n_1	f_1	$f_1 \times 100$
M_2	n_2	f_2	$f_2 \times 100$
...
M_i	n_i	f_i	$f_i \times 100$
...
M_p	n_p	f_p	$f_p \times 100$
Total	n	1	100

Pour une variable quantitative continue, les données sont regroupées en classes.

- **L'amplitude**, ou longueur de la classe, est la différence entre l'extrémité et l'origine de la classe.
- **Fonction de répartition** (variable quantitative) :
 $F(x)$ est la fréquence relative (ou les effectifs) des individus dont la valeur de la variable est inférieure ou égale à x .
 $G(x) = 1 - F(x)$ est la fréquence relative (ou les effectifs) des individus dont la valeur de la variable est supérieure à x .
- **La courbe des fréquences cumulées croissantes** est le graphe de la fonction F .
- **La courbe des fréquences cumulées décroissantes** est le graphe de la fonction G .
- **Diagramme en bâtons** : c'est la représentation graphique de la distribution d'une variable quantitative discrète.
- **Histogramme** : c'est la représentation graphique sous forme de rectangles de la distribution d'une variable quantitative continue après regroupement des données en classes.

II Exercices

1. Représentations graphiques d'une variable qualitative

Le tableau suivant donne la répartition des 500 salariés d'une entreprise selon le mode de transport utilisé pour se rendre du domicile au lieu de travail.

Si un salarié utilise plusieurs modes de transport, celui retenu dans la classification est celui de la distance parcourue la plus longue.

Mode de transport	Symbole	Effectifs	Fréquences	Fréquences en %
Voiture	V	60	0,12	12
RER	R	120	0,24	24
Métro	M	160	0,32	32
Autobus	A	80	0,16	16
Bicyclette	B	80	0,16	16
Total		500	1	100

1. Les modalités d'une variable sont disjointes et exhaustives, expliquez ce que cela signifie.
2. Indiquer les difficultés à réaliser une classification pertinente pour les modalités de la variable utilisée dans cet exercice.
3. Indiquer comment on obtient les 4^e et 5^e colonnes à partir de la 3^e colonne.
4. Indiquer le principe essentiel pour faire un diagramme ou une représentation graphique d'une distribution statistique d'une variable qualitative.
Représenter les données du tableau à l'aide d'un diagramme circulaire.
5. Indiquer d'autres modes de représentations graphiques pour des variables qualitatives.

S o l u t i o n

1. Disjointes signifie que les modalités ne se recouvrent pas afin qu'un même individu ne puisse pas être classé dans plusieurs modalités.

Exhaustives signifie que chaque individu peut être classé selon les modalités existantes.

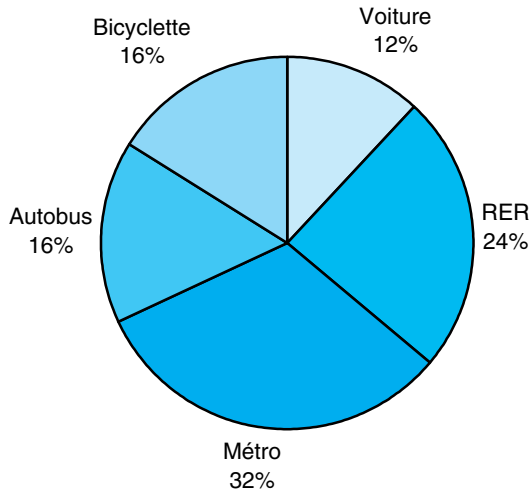
En résumé, chaque individu est classé selon une et une seule modalité de la variable, ce qui explique que le total des individus répertoriés dans les diverses modalités fasse 500.

2. Cet exercice montre qu'il est difficile avec les données précédentes de trouver une classification pertinente. En effet, les individus qui vont à pied à leur travail ou en deux roues motorisées ne sont pas pris en compte dans cette classification. De plus, la classification qui s'appuie sur la distance parcourue la plus longue par les salariés utilisant plusieurs modes de transport, n'est pas forcément la plus pertinente. Ceci n'est qu'un exercice, mais avant de recueillir des données, il faut penser à la façon de les traiter.

3. $f_1 = \frac{n_1}{n} = \frac{60}{500} = 0,12$ $f_2 = \frac{120}{500} = 0,24 \dots$

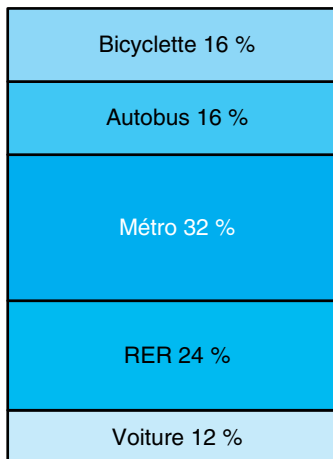
4. Le principe de base d'un diagramme représentant des données qualitatives est que les différentes aires du diagramme sont proportionnelles aux effectifs ou fréquences. Ainsi la modalité *V* (voiture) doit représenter 12 % de l'aire totale, *R* 24 %, etc. Le diagramme le plus usuel est le diagramme circulaire, souvent désigné par « camembert ».

Les aires des secteurs étant proportionnelles aux angles qu'ils forment, les angles des secteurs représentant les différentes modalités sont aussi proportionnels aux effectifs.



5. Un autre type de graphique est le diagramme en barres ou en bandeaux ; chaque bandeau a une hauteur proportionnelle à l'effectif de la modalité qu'il représente. Les outils informatiques permettent de réaliser une grande variété de représentations graphiques.

Diagramme à barres



2. Représentation graphique d'une variable quantitative discrète

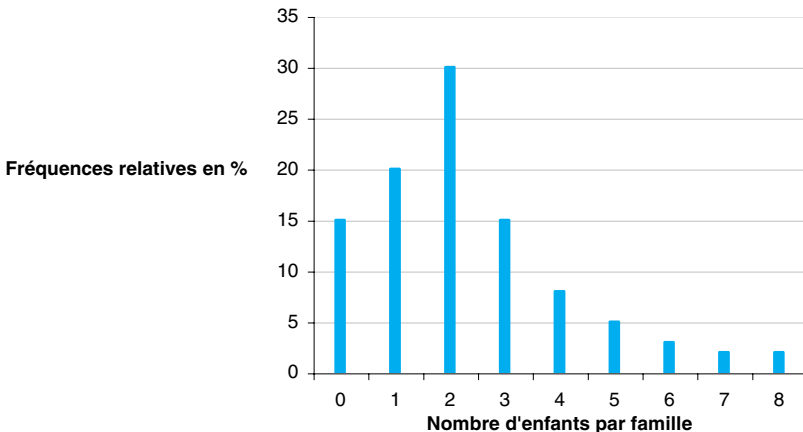
Le tableau suivant donne la distribution de 200 familles selon le nombre d'enfants.

Nombre d'enfants	Effectifs	Fréquences relatives	Fréquences relatives en %	Fréquences cumulées croissantes en %
0	30	0,15	15	15
1	40	0,20	20	35
2	60	0,30	30	65
3	30	0,15	15	80
4	16	0,08	8	88
5	10	0,05	5	93
6	6	0,03	3	96
7	4	0,02	2	98
8	4	0,02	2	100
Total	200			

1. Faire le diagramme en bâtons de cette distribution.
2. Comment obtenir la dernière colonne du tableau à partir de la précédente ?
3. Indiquer les propriétés de la fonction de répartition F .
4. Déterminer la fonction de répartition de cette distribution.
5. Tracer la courbe des fréquences cumulées croissantes, c'est-à-dire le graphe de F .

S o l u t i o n

1. On porte en abscisse les différentes modalités de la variable et en ordonnée les effectifs ou les fréquences relatives. Comme l'énoncé ne le précise pas, on choisit dans ce cas les fréquences relatives en pourcentage, c'est souvent la façon la plus lisible de présenter des données.



2. $35 = 20 + 15$; $65 = 15 + 20 + 30 = 35 + 30$. Les lignes de la dernière colonne s'obtiennent par sommation des lignes de la colonne précédente du haut vers le bas en s'arrêtant à la ligne correspondant à un nombre donné d'enfants. Ainsi 35 % des familles ont 0 ou 1 enfant ; 65 % des familles ont, au plus, deux enfants.

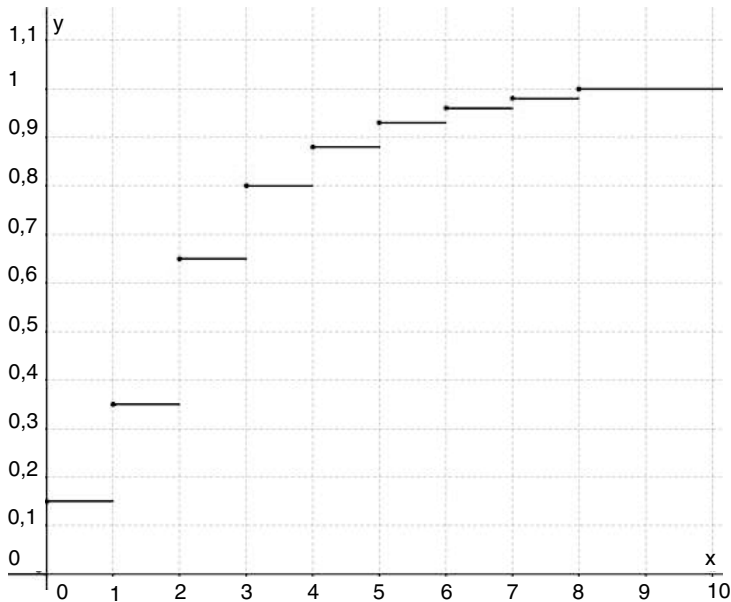
3. $F(x)$ est la fréquence des individus dont la variable est inférieure ou égale à x . Il en résulte : $0 \leq F(x) \leq 1$.

F est croissante au sens large : $b > a$ implique $F(b) \geq F(a)$.

4. La fonction F pour une variable discrète est constante par morceaux, c'est une fonction en escalier.

Si : $0 \leq x < 1$ $F(x) = 0,15$
 $1 \leq x < 2$ $F(x) = 0,35$
 $2 \leq x < 3$ $F(x) = 0,65$
 $3 \leq x < 4$ $F(x) = 0,80$
 $4 \leq x < 5$ $F(x) = 0,88$
 $5 \leq x < 6$ $F(x) = 0,93$
 $6 \leq x < 7$ $F(x) = 0,96$
 $7 \leq x < 8$ $F(x) = 0,98$
 $x \geq 8$ $F(x) = 1$

5. Courbe des fréquences cumulées ascendantes.



3. Représentation graphique d'une variable continue

Une enquête a été réalisée auprès de 500 salariés d'une entreprise pour étudier la distribution des salaires nets mensuels en euros.

Salaires mensuels (milliers d'euros)	Effectifs n_i	Fréquences relatives f_i	Effectifs cumulés croissants	Fréquences cumulées croissantes	Fréquences cumulées croissantes en %	Fréquences cumulées décroissantes en %
[1,2 à 1,6[100	0,20				
[1,6 à 2,0[150	0,30				
[2,0 à 2,8[100	0,20				
[2,8 à 3,6[80	0,16				
[3,6 à 4,4[50	0,10				
[4,4 à 6,0[20	0,04				
Total	500					

1. Compléter le tableau précédent.
2. Indiquer comment on construit un histogramme et tracer l'historgramme de cette distribution.
3. Tracer la courbe des fréquences cumulées croissantes (en %) et la courbe des fréquences cumulées décroissantes (en %).

S o l u t i o n

1. Tableau complété

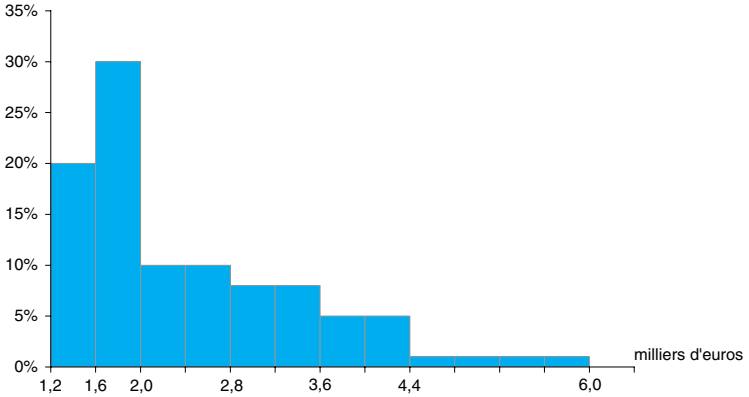
Salaires mensuels (milliers d'euros)	Effectifs n_i	Fréquences relatives f_i	Effectifs cumulés croissants	Fréquences cumulées croissantes	Fréquences cumulées croissantes en %	Fréquences cumulées décroissantes en %
[1,2 à 1,6[100	0,20	100	0,20	20	100
[1,6 à 2,0[150	0,30	250	0,50	50	80
[2,0 à 2,8[100	0,20	350	0,70	70	50
[2,8 à 3,6[80	0,16	430	0,86	86	30
[3,6 à 4,4[50	0,10	480	0,96	96	14
[4,4 à 6,0[20	0,04	500	1,00	100	4
Total	500	1				

2. Construction de l'historgramme

En ordonnée, on porte les fréquences par classe d'amplitude 400 euros ce qui correspond aux deux premières classes. La classe suivante qui va de 2 000 à 2 800 euros a une amplitude de 800 euros et une fréquence de 20 %. Ceci revient à répartir 10 % des effectifs dans une classe fictive d'amplitude 400 euros qui s'étend de 2 000 à

2 400 euros et 10 % dans une classe qui s'étend de 2 400 euros à 2 800 euros. On porte donc pour la classe qui s'étend de 2 000 à 2 800 euros une hauteur de 10 %. On raisonne de cette façon pour terminer l'histogramme.

Histogramme des fréquences

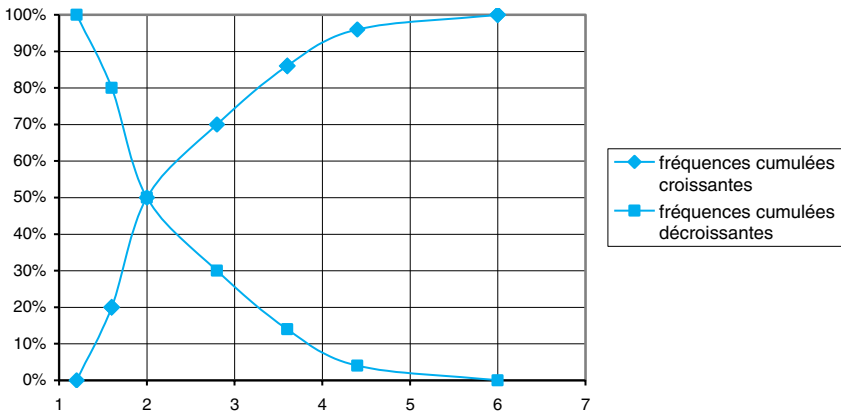


3. Courbes des fréquences cumulées

Pour repérer les points qui figurent sur la courbe des fréquences cumulées croissantes, on prend en abscisse l'extrémité des classes.

Pour repérer les points qui figurent sur la courbe des fréquences cumulées décroissantes, on prend en abscisse l'origine des classes.

Courbes des fréquences cumulées



Caractéristiques de tendance centrale et de dispersion. Concentration

I Rappel de cours

- Une **série statistique** x_1, x_2, \dots, x_n est une suite d'observations d'une variable X . Dans le cas où les observations x_i sont observées avec les effectifs n_i ou avec les fréquences f_i on présente les données sous forme de tableau.

Variable	x_1		x_i		x_p	Total
Effectif	n_1		n_i		n_p	n
Fréquence	f_1		f_i		f_p	1 ou 100 %

Caractéristiques de valeur centrale et de position

- **La moyenne arithmétique :**

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{i=n} x_i \quad \text{ou} \quad \bar{x} = \frac{1}{n} \sum_{i=1}^{i=p} n_i x_i \quad \text{ou} \quad \bar{x} = \sum_{i=1}^{i=p} f_i x_i \quad \text{ou} \quad \bar{x} = \frac{1}{100} \sum_{i=1}^{i=p} f_i x_i$$

Lorsque chaque modalité x_i de la variable X est observée une seule fois, c'est la première formule qui s'applique sinon c'est la deuxième. La troisième correspond à des fréquences relatives exprimées entre 0 et 1 ; la dernière à des fréquences relatives exprimées en pourcentage.

- **La médiane** se détermine de telle façon qu'il y ait autant d'observations supérieures à la médiane que d'observations inférieures à la médiane.
- **Le mode** correspond à la valeur de la variable observée avec la plus grande fréquence ou le plus grand effectif. Sa détermination est un peu plus délicate pour une

variable continue (on définit plutôt une **classe modale**). Il existe des distributions à plusieurs modes.

- **La moyenne géométrique : m_g**

$$m_g = (x_1 x_2 \dots x_n)^{\frac{1}{n}} = \prod_{i=1}^{i=n} x_i^{\frac{1}{n}} \text{ ou } m_g = (x_1^{n_1} x_2^{n_2} \dots x_p^{n_p})^{\frac{1}{n}} = \left(\prod_{i=1}^{i=p} x_i^{n_i} \right)^{\frac{1}{n}}$$

- **La moyenne harmonique : m_h**

$$\frac{1}{m_h} = \frac{1}{n} \left(\frac{1}{x_1} + \frac{1}{x_2} + \dots + \frac{1}{x_n} \right) = \frac{1}{n} \sum_{i=1}^{i=n} \frac{1}{x_i}$$

$$\text{ou } \frac{1}{m_h} = \frac{1}{n} \left(\frac{n_1}{x_1} + \frac{n_2}{x_2} + \dots + \frac{n_p}{x_p} \right)$$

L'inverse de la moyenne harmonique est la moyenne arithmétique des quantités $\frac{1}{x_i}$.

- **Les quartiles Q_1, Q_2, Q_3** correspondent à des effectifs de 25 %, 50 %, 75 % sur des observations qui sont ordonnées par ordre croissant. Ainsi, 25 % des observations sont inférieures au premier quartile Q_1 , 50 % inférieures au deuxième quartile (qui est égal à la médiane), 75 % inférieures au troisième quartile Q_3 .
- **Les déciles** correspondent à des effectifs de 10 %, 20 %, ..., 90 %. Il existe donc 9 déciles notés D_1, D_2, \dots, D_9 . Ainsi, 10 % des observations sont inférieures à D_1 .
- **Les centiles ou percentiles** correspondent à 1 %, 2 %, ..., 99 %.

Caractéristiques de dispersion

- **L'étendue** est la différence entre la plus grande et la plus petite des observations.
- **L'écart entre les quartiles $Q_3 - Q_1$** , ou entre les déciles $D_9 - D_1$.
- **La variance V** .
Simple

$$V = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{1}{n} \sum_{i=1}^n x_i^2 - \bar{x}^2 = \text{moyenne des } x^2 - (\text{moyenne des } x)^2$$

Pondérée

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

- **L'écart type σ** est la racine carrée de la variance V .

La concentration

- La notion de **concentration** permet de mesurer les inégalités dans la répartition d'unités statistiques (individus) distribuées selon la taille. La concentration est

forte si un petit nombre d'individus se partagent une part importante de la masse de la variable (exemple masse des salaires).

- La **médiale** : les individus dont la variable (caractère) est inférieure à la médiale se partagent la moitié de la masse totale.
- La **concentration** s'observe visuellement sur la courbe de **Lorentz** et se mesure par l'indice de **Gini** ou par l'écart entre la médiale et la médiane. (Voir exercice.)

Moments centrés et non centrés, asymétrie et aplatissement

- **Moments non centrés** d'ordre r : $m_r = \frac{1}{n} \sum_i n_i x_i^r$
- **Moments centrés** d'ordre r : $\mu_r = \frac{1}{n} \sum_i n_i (x_i - \bar{x})^2$
- **L'asymétrie** est mesurée par le coefficient γ_1 de Fisher : $\gamma_1 = \frac{\mu_3}{\sigma^3}$
- **L'aplatissement** est mesuré par le coefficient γ_2 de Fisher : $\gamma_2 = \frac{\mu_4}{\mu_2^2} - 3$

II Exercices

1. Calculs élémentaires sur des séries statistiques

Les notes obtenues à un test dans deux groupes de 7 personnes sont les suivantes :

- groupe 1 : 16 ; 08 ; 10 ; 12 ; 14 ; 11 ; 13.
- groupe 2 : 08 ; 20 ; 12 ; 06 ; 16 ; 18 ; 04.

1. Calculer les moyennes m_1 et m_2 des notes dans les groupes 1 et 2 en donnant les formules algébriques.

Calculer les variances V_1 et V_2 puis les écarts types σ_1 et σ_2 en utilisant successivement les deux formules différentes pour le calcul.

Commenter les résultats obtenus et indiquer s'ils sont conformes à ce que l'on pourrait attendre.

2. Calculer les médianes et les étendues des deux séries.

3. Démontrer le résultat suivant :

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

4. Démontrer le résultat suivant :

$$E_a = \sum_{i=1}^n (x_i - a)^2 = \sum_{i=1}^n (x_i - \bar{x})^2 + \sum_{i=1}^n (\bar{x} - a)^2$$

En déduire que E_a est minimale pour $a = \bar{x}$.

S o l u t i o n

$$\begin{aligned} 1. m_1 &= \frac{1}{n} \left(\sum_{i=1}^{i=n} x_i \right) = \frac{1}{7} \left(\sum_{i=1}^{i=7} x_i \right) = \frac{1}{7} (x_1 + x_2 + \dots + x_7) \\ &= \frac{1}{7} (16 + 8 + 10 + 12 + 14 + 11 + 13) = 12 \end{aligned}$$

$$\begin{aligned} m_2 &= \frac{1}{n} \left(\sum_{i=1}^{i=n} y_i \right) = \frac{1}{7} \left(\sum_{i=1}^{i=7} y_i \right) = \frac{1}{7} (y_1 + y_2 + \dots + y_7) \\ &= \frac{1}{7} (8 + 20 + 12 + 6 + 16 + 18 + 14) = 12 \end{aligned}$$

La $i^{\text{ème}}$ observation de la série 1 est notée x_i (on pourrait aussi la noter x_{1i}).

La $j^{\text{ème}}$ observation de la série 2 est notée y_j (on pourrait aussi la noter x_{2j}).

Calcul de la variance de la première série en utilisant la première formule.

$$\begin{aligned} V_1 &= \frac{1}{n} \left(\sum_{i=1}^n (x_i - m_1)^2 \right) \\ &= \frac{1}{7} [(16 - 12)^2 + (8 - 12)^2 + (10 - 12)^2 + \dots + (13 - 12)^2] = 6 \end{aligned}$$

Calcul de la variance de la première série en utilisant la seconde formule.

$$\begin{aligned} V_1 &= \frac{1}{n} \sum_{i=1}^n x_i^2 - m_1^2 = \frac{1}{7} (16^2 + 8^2 + \dots + 11^2 + 13^2) - 12^2 \\ &= \frac{1\ 050}{7} - 144 = 150 - 144 = 6 \end{aligned}$$

Calcul de la variance de la deuxième série.

$$\begin{aligned} V_2 &= \frac{1}{n} \left(\sum_{i=1}^n (y_i - m_2)^2 \right) \\ &= \frac{1}{7} [(8 - 12)^2 + (20 - 12)^2 + \dots + (4 - 12)^2] = 33,14 \\ \sigma_1 &= \sqrt{V_1} = \sqrt{6} = 2,45 \quad \sigma_2 = \sqrt{V_2} = \sqrt{33,14} = 5,76 \end{aligned}$$