Statistiques Descriptives

L1 Économie

Seynabou Gueye, Mehdi Guelmamen, Emilien Macault

Faculté de Droit, Économie et Administration
Université de Lorraine

Chapitre 4 : Les caractéristiques de dispersion

Sommaire

Introduction

Définition

Les indicateurs de dispersion complètent les mesures de tendances centrales. La dispersion est la tendance qu'ont les valeurs de la distribution d'un caractère à s'écarter ou à se disperser, de part et d'autre d'une valeur centrale de référence (la médiane ou la moyenne). Elles permettent de comparer des séries statistiques de même nature.

Introduction

Remarque

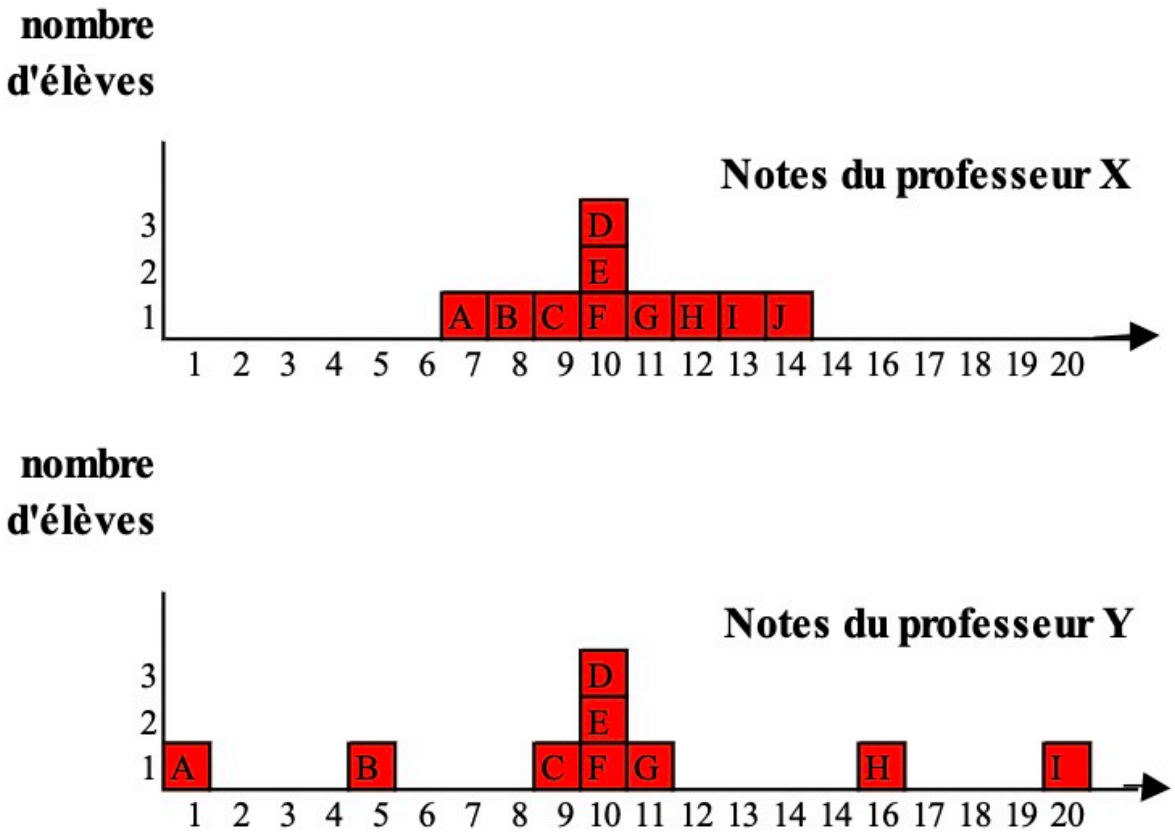
On distingue la dispersion **absolue** (mesurée dans l'unité de mesure du caractère) et la dispersion **relative** (mesurée par un nombre sans dimension). Les indicateurs nous aident à déterminer si les données sont trop éloignées de la valeur centrale afin de voir si cette dernière est suffisante pour représenter la population à l'étude.

Introduction

- Exemple: au baccalauréat, les copies sont analysées après correction. Les moyennes sont examinées et il arrive qu'elles divergent fortement d'un enseignant à l'autre, dans une même discipline.

Introduction

- Exemple: au baccalauréat, les copies sont analysées après correction. Les moyennes sont examinées et il arrive qu'elles divergent fortement d'un enseignant à l'autre, dans une même discipline.



Sommaire

Sommaire

Dispersion par rapport à la médiane

Les caractéristiques de dispersion par rapport à la médiane indiquent dans quelle mesure un échantillon s'écarte de la médiane. On appelle **intervalle de variation** l'intervalle I_{var} tel que :

$$I_{var} = [X_{min}; X_{max}]$$

avec X_{min} la plus petite valeur et X_{max} la plus grande valeur. L'intervalle de variation contient 100% des modalités de la série. De I_{var} nous pouvons tirer l'**étendue**:

$$E = X_{max} - X_{min}$$

Remarque

L'étendue sert de supplément à d'autres mesures, mais elle est rarement utilisée comme seule mesure de dispersion étant donné qu'elle est sensible aux valeurs extrêmes.

Sommaire

Dispersion par rapport à la médiane

Définition

L'**intervalle interquartile** est l'étendue de la distribution sur laquelle se trouvent concentrée la moitié des éléments dont les valeurs de X les plus proches de la médiane. Il contient 50% des modalités de la série. On exclut alors de la distribution les 25% des valeurs les plus basses et les 25% des valeurs les plus élevées de X.

Dispersion par rapport à la médiane

Définition

L'**intervalle interquartile** est l'étendue de la distribution sur laquelle se trouvent concentrée la moitié des éléments dont les valeurs de X les plus proches de la médiane. Il contient 50% des modalités de la série. On exclut alors de la distribution les 25% des valeurs les plus basses et les 25% des valeurs les plus élevées de X.

Cet intervalle s'écrit :

$$I_Q = [Q_1; Q_3]$$

Dispersion par rapport à la médiane

Définition

L'**intervalle interquartile** est l'étendue de la distribution sur laquelle se trouvent concentrée la moitié des éléments dont les valeurs de X les plus proches de la médiane. Il contient 50% des modalités de la série. On exclut alors de la distribution les 25% des valeurs les plus basses et les 25% des valeurs les plus élevées de X.

Cet intervalle s'écrit :

$$I_Q = [Q_1; Q_3]$$

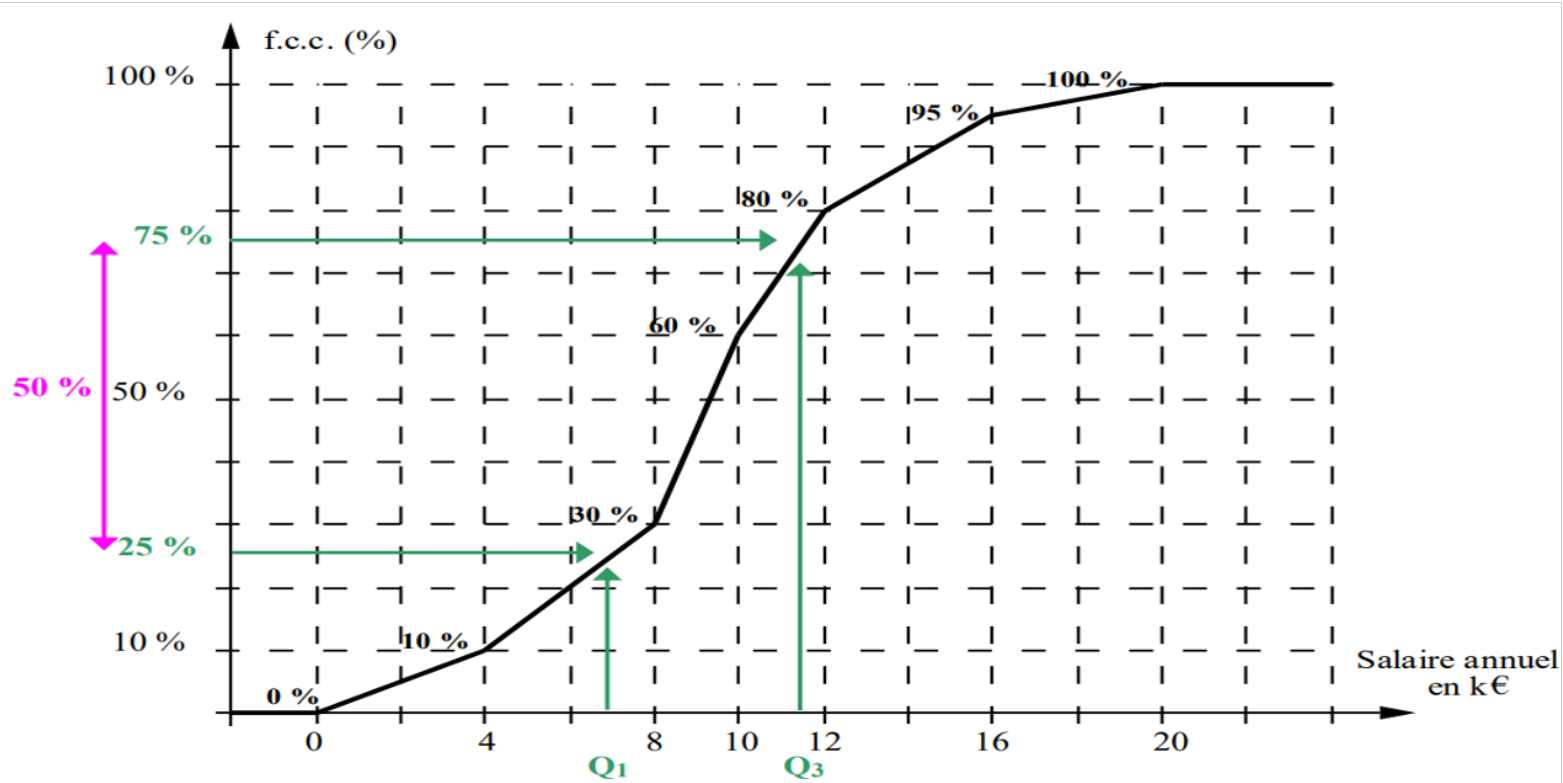
D'où l'on tire l'**écart interquartile**:

$$e_Q = Q_3 - Q_1$$

Dispersion par rapport à la médiane

Exemple : détermination graphique d'un intervalle et d'un écart interquartile.

- La courbe cumulative croissante suivante est associée à une distribution de salaires annuels exprimés en k€ dans une population de salariés. On lit sur le graphique que $[Q_1; Q_3] = [7; 11,5]$. Soit $e_Q = 4,5$k€.



Dispersion par rapport à la médiane

Définition

L'**intervalle interdécile** est l'étendue de la distribution sur laquelle se trouvent concentrée 80% des éléments dont les valeurs de X sont les plus proches de la médiane. Intervalle interdécile contient 80% des modalités de la série. On exclut alors de la distribution les 10% des valeurs les plus faibles et les 10% des valeurs les plus fortes de X

Dispersion par rapport à la médiane

Définition

L'**intervalle interdécile** est l'étendue de la distribution sur laquelle se trouvent concentrée 80% des éléments dont les valeurs de X sont les plus proches de la médiane. Intervalle interdécile contient 80% des modalités de la série. On exclut alors de la distribution les 10% des valeurs les plus faibles et les 10% des valeurs les plus fortes de X

Cet intervalle s'écrit :

$$I_D = [D_1; D_9]$$

Dispersion par rapport à la médiane

Définition

L'**intervalle interdécile** est l'étendue de la distribution sur laquelle se trouvent concentrée 80% des éléments dont les valeurs de X sont les plus proches de la médiane. Intervalle interdécile contient 80% des modalités de la série. On exclut alors de la distribution les 10% des valeurs les plus faibles et les 10% des valeurs les plus fortes de X

Cet intervalle s'écrit :

$$I_D = [D_1; D_9]$$

D'où l'on obtient l'**écart interdécile**:

$$e_D = D_9 - D_1$$

Dispersion par rapport à la médiane

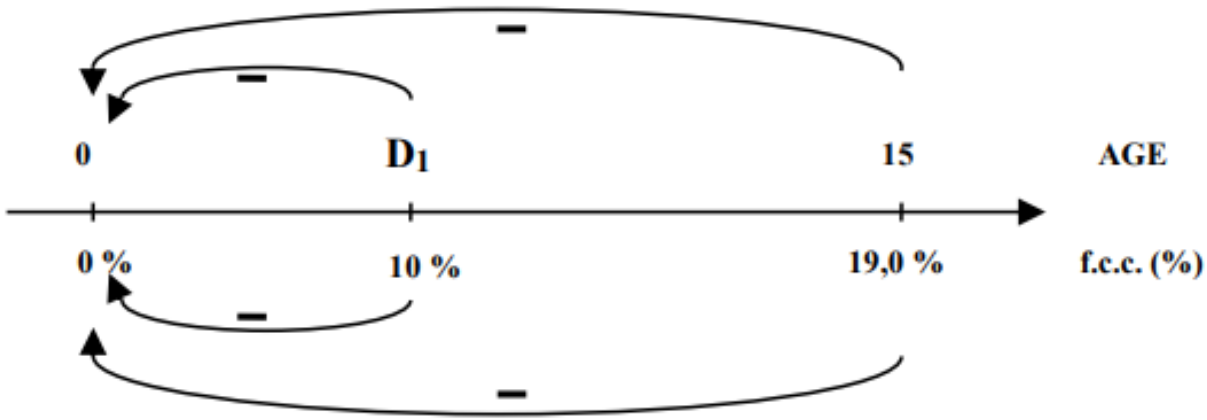
Exemple : détermination d'un intervalle interdécile par interpolation linéaire.

- Soit la répartition de la population française par tranche d'âge, au recensement de 1990 (Source : Annuaire statistique de la France, INSEE, 1993) :

TRANCHE D'AGE	Fréquence f_i (%)	f.c.c. $f_i \uparrow$ (%)	
Classe contenant D_1 : [0 ; 15 [	19,0	19,0	10 %
[15 ; 20 [	7,5	26,5	
[20 ; 25 [	7,5	34,0	
[25 ; 30 [	7,6	41,6	
[30 ; 40 [	15,1	56,7	
[40 ; 50 [	13.0	69,7	
[50 ; 60 [	10,4	80,1	
Classe contenant D_9 : [60;100 [	19,9	100,0	90 %
	100 %		

Dispersion par rapport à la médiane

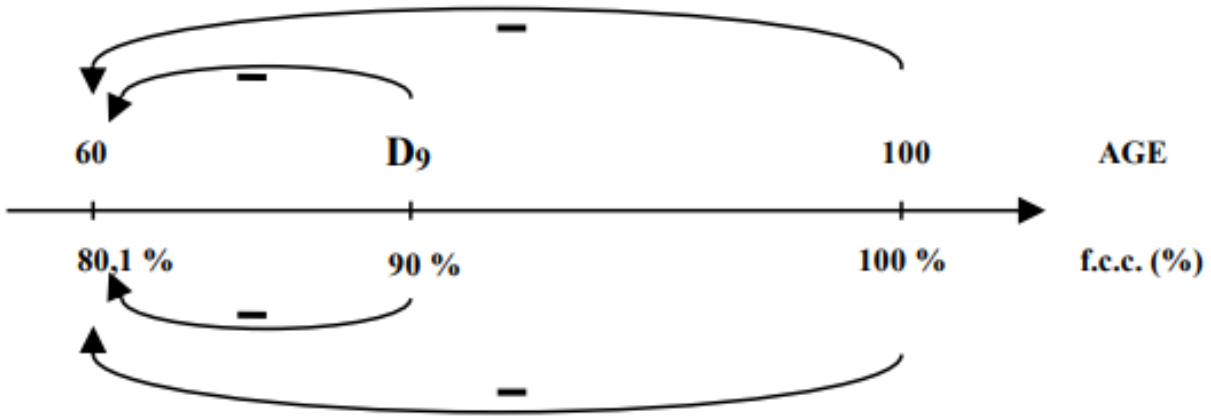
- Au niveau du tableau, la classe contenant le 1er décile D_1 est la classe $[0; 15[$. Elle permet de dépasser 10% des observations.
- Déterminons D_1 et D_9 par interpolation linéaire:



$$\frac{D_1 - 0}{15 - 0} = \frac{10 - 0}{19 - 0} \Rightarrow D_1 = 15 \times \frac{10}{19} \Rightarrow D_1 = 7,9 \text{ ans}$$

Dispersion par rapport à la médiane

- Au niveau du tableau, la classe contenant le 9e décile D_9 est la classe $[60; 100[$. Elle permet de dépasser 90% des observations.



$$\frac{D_9 - 60}{90 - 80,1} = \frac{100 - 60}{100 - 80,1} \Rightarrow \frac{D_9 - 60}{9,9} = \frac{40}{19,9} \Rightarrow D_9 = 9,9 \times 2 + 60 = 79,9$$

Dispersion par rapport à la médiane

- Interprétation
- Soit l'écart interdécile : $e_D = 79,9 - 7,9 = 72$ ans;

Interprétation de l'intervalle interdéciles

En 1990, 80% des français étaient âgés entre 7,9 ans et 79,9 ans. De plus, 10 % avaient moins de 7,9 ans ou plus 79,9 ans.

Interprétation de l'écart interdéciles

Le 10% ème individu le plus âgé est plus âgé de 72 ans du 10% ème individu le plus jeune.

- Calculons le rapport $\frac{D_9}{D_1} = \frac{79,9}{7,9} = 10,1$. Les 10% des plus âgés sont 10 fois plus vieux que les 10% des plus jeunes. Ce rapport est souvent utilisé avec les revenus (T.Piketty notamment).

Dispersion par rapport à la médiane

Définition

L'**intervalle intercentile** est l'étendue de la distribution sur laquelle se trouve concentrée 98% des éléments dont les valeurs de X sont les plus proches de la médiane. L'intervalle intercentile contient 98% des modalités de la série. On exclut alors de la distribution les 1% des valeur les plus faibles et les 1% des valeurs les plus fortes de X.

Cet intervalle s'écrit :

Dispersion par rapport à la médiane

Définition

L'**intervalle intercentile** est l'étendue de la distribution sur laquelle se trouve concentrée 98% des éléments dont les valeurs de X sont les plus proches de la médiane. L'intervalle intercentile contient 98% des modalités de la série. On exclut alors de la distribution les 1% des valeur les plus faibles et les 1% des valeurs les plus fortes de X.

Cet intervalle s'écrit :

$$I_C = [C_1; C_{99}]$$

D'où l'on tire l'**écart intercentile**:

Dispersion par rapport à la médiane

Définition

L'**intervalle intercentile** est l'étendue de la distribution sur laquelle se trouve concentrée 98% des éléments dont les valeurs de X sont les plus proches de la médiane. L'intervalle intercentile contient 98% des modalités de la série. On exclut alors de la distribution les 1% des valeur les plus faibles et les 1% des valeurs les plus fortes de X.

Cet intervalle s'écrit :

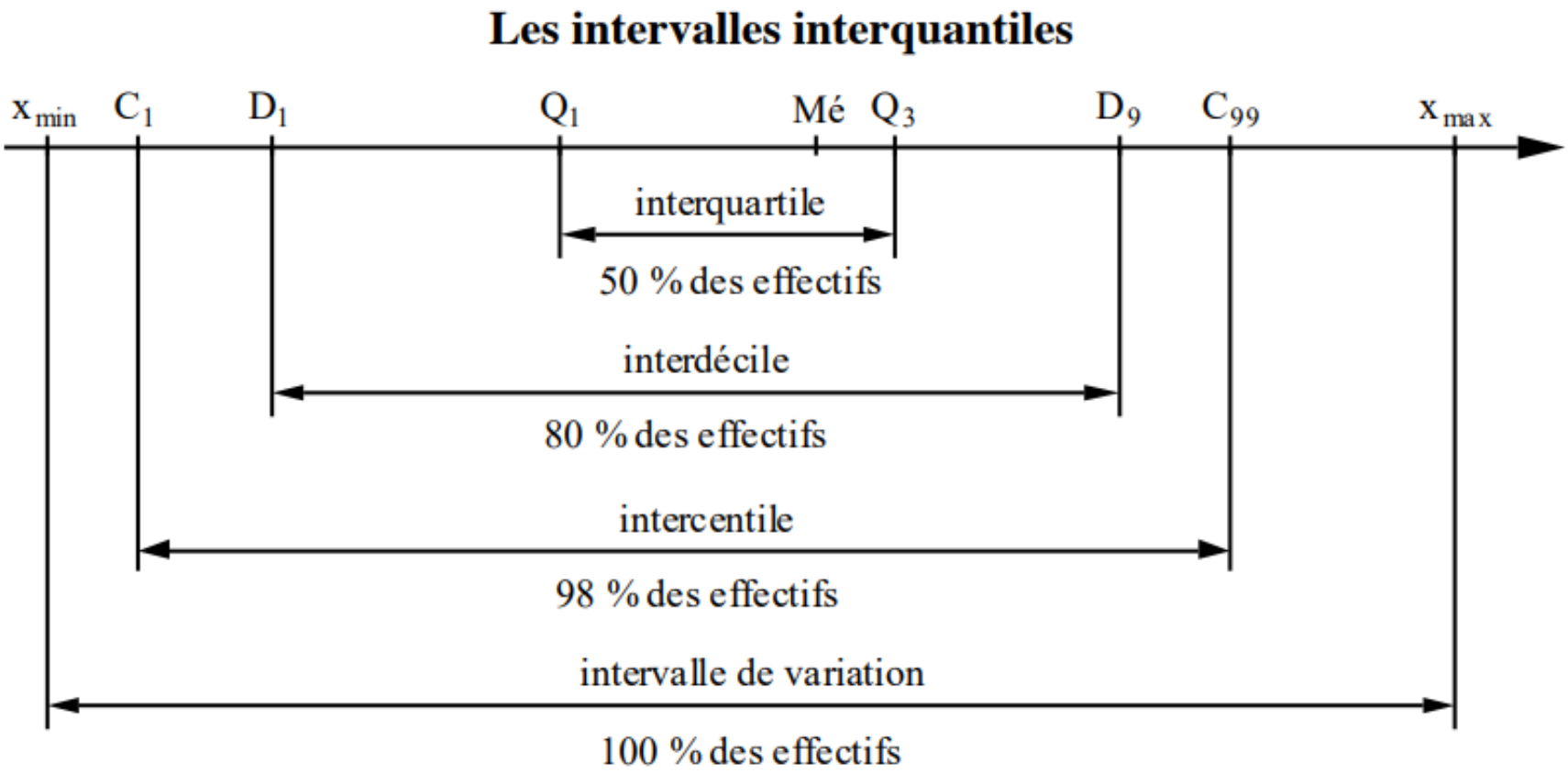
$$I_C = [C_1; C_{99}]$$

D'où l'on tire l'**écart intercentile**:

$$e_C = C_{99} - C_1$$

Dispersion autour de la médiane

Illustration:



Sommaire

Sommaire

Dispersion autour de la moyenne

Les **moments** sont les outils les plus importants pour mesurer la forme d'une distribution. On appelle **moment d'ordre** r ($r \in \mathbb{N}$) d'un caractère X le nombre:

Dispersion autour de la moyenne

Les **moments** sont les outils les plus importants pour mesurer la forme d'une distribution. On appelle **moment d'ordre** r $(r \in \mathbb{N})$ d'un caractère X le nombre:

$$m_r(X) = \frac{\sum_{i=1}^{p} n_i x_i^r}{\sum_{i=1}^{p} n_i} = \sum_{i=1}^{p} f_i x_i^r$$

Dispersion autour de la moyenne

Les **moments** sont les outils les plus importants pour mesurer la forme d'une distribution. On appelle **moment d'ordre** r ($r \in \mathbb{N}$) d'un caractère X le nombre:

$$m_r(X) = \frac{\sum_{i=1}^{p} n_i x_i^r}{\sum_{i=1}^{p} n_i} = \sum_{i=1}^{p} f_i x_i^r$$

Cas particuliers

- $m_0 = 1$
- $m_1 = \overline{x}$
- $m_2 = Q^2$ (moyenne quadratique au carré)

Dispersion autour de la moyenne

On appelle **moment centré d'ordre** r $(r \in \mathbb{N})$ d'un caractère X le nombre:

Dispersion autour de la moyenne

On appelle **moment centré d'ordre** r $(r \in \mathbb{N})$ d'un caractère X le nombre:

$$\mu_r(X) = \frac{\sum_{i=1}^{p} n_i (x_i - \overline{x})^r}{\sum_{i=1}^{p} n_i} = \sum_{i=1}^{p} f_i (x_i - \overline{x})^r$$

Remarque

D'une manière générale, les moments centrés d'ordre pair renseignent sur la dispersion des observations autour de la moyenne $\overline{x}$, et les moments centrés d'ordre impair sur la dissymétrie de la distribution (que nous aborderons plus tard).

Dispersion autour de la moyenne

Définition

On appelle **variance** d'une variable X son moment centré d'ordre 2. C'est la moyenne des carrés des écarts à la moyenne. Elle mesure globalement la variation d'un caractère de part et d'autre de la moyenne arithmétique.

Elle est donnée par:

Dispersion autour de la moyenne

Définition

On appelle **variance** d'une variable X son moment centré d'ordre 2. C'est la moyenne des carrés des écarts à la moyenne. Elle mesure globalement la variation d'un caractère de part et d'autre de la moyenne arithmétique.

Elle est donnée par:

$$\mathbf{Var(X)} = \frac{\sum_{i=1}^{n} n_i(x_i - \overline{x})^2}{\sum_{i=1}^{n} n_i} = \frac{1}{N}\sum_{i=1}^{n} n_i(x_i - \overline{x})^2$$

Dispersion autour de la moyenne

Définition

On appelle **variance** d'une variable X son moment centré d'ordre 2. C'est la moyenne des carrés des écarts à la moyenne. Elle mesure globalement la variation d'un caractère de part et d'autre de la moyenne arithmétique.

Elle est donnée par:

$$\mathbf{Var(X)} = \frac{\sum_{i=1}^{n} n_i(x_i - \overline{x})^2}{\sum_{i=1}^{n} n_i} = \frac{1}{N}\sum_{i=1}^{n} n_i(x_i - \overline{x})^2$$

Remarque

La variance est parfois dénotée σ_X^2 ou simplement σ^2 lorsque le contexte est clair.

Dispersion autour de la moyenne

Démonstration: on sait que $var(X) = \frac{1}{N}\sum_{i=1}^{n} n_i(x_i - \overline{x})^2$. On peut développer cette expression

Dispersion autour de la moyenne

Démonstration: on sait que $var(X) = \frac{1}{N}\sum_{i=1}^{n} n_i(x_i - \overline{x})^2$. On peut développer cette expression :

$$\begin{aligned}
var(X) = \quad &= \frac{1}{N}\sum_{i=1}^{k} n_i \left(x_i^2 - 2\overline{x}x_i + \overline{x}^2\right) \\
&= \frac{1}{N}\sum_{i=1}^{k} n_i x_i^2 - 2\overline{x}\frac{1}{N}\sum_{i=1}^{k} n_i x_i + \overline{x}^2\frac{1}{N}\sum_{i=1}^{k} n_i \\
&= \frac{1}{N}\sum_{i=1}^{k} n_i x_i^2 - 2\overline{x}\overline{x} + \overline{x}^2 \\
&= \frac{1}{N}\sum_{i=1}^{k} n_i x_i^2 - 2\overline{x}^2 + \overline{x}^2 \\
&= \frac{1}{N}\sum_{i=1}^{k} n_i x_i^2 - \overline{x}^2.
\end{aligned}$$

Dispersion autour de la moyenne

Le théorème de König-Huygens nous permet de déterminer la variance d'une variable statistique à partir des moments d'ordres 1 et 2, sans utiliser les écarts.

Théorème

$$\mathbf{Var}(X) = m_2(X) - m_1^2(X) = \frac{\sum_{i=1}^{n} n_i x_i^2}{\sum_{i=1}^{n} n_i} - \left[\frac{\sum_{i=1}^{n} n_i x_i}{\sum_{i=1}^{n} n_i}\right]^2$$

Avec m_1 la moyenne arithmétique $\overline{x}$ et m_2 la moyenne quadratique au carré Q^2.

Dispersion autour de la moyenne

- Soit $a \in \mathbb{R}$ une constante :

Propriété 1

$\mathbf{Var}(a) = 0$

Propriété 2

$\mathbf{Var}(aX) = a^2\mathbf{Var}(X)$

- Remarque: la variance est **toujours positive ou nulle** car elle est une somme de carré.

Dispersion autour de la moyenne

- Soit X_i et Y_i des séries statistiques définies par $Y_i = aX_i + b$ où a et b sont deux réels (constants) même niveau.

Propriété 3

$$\mathbf{Var}(Y) = \mathbf{Var}(aX + b) = a^2 Var(X)$$

Dispersion autour de la moyenne

Exemples:

- $\mathbf{Var}(2X + 1) = 2^2\mathbf{Var}(X) = 4\mathbf{Var}(X)$
- $\mathbf{Var}(-X + 2) = -1^2\mathbf{Var}(X) = \mathbf{Var}(X)$
- $\mathbf{Var}(4X - 3) = 4^2\mathbf{Var}(X) = 16\mathbf{Var}(X)$
- $\mathbf{Var}(153\ 678\ 987\ 432) = 0$

Sommaire

Dispersion autour de la moyenne

Définition

L'écart-type d'une série est la racine carrée la variance. Il est noté $\sigma(X)$ et mesure la dispersion absolue, i.e. la dispersion absolue autour de la moyenne. Il s'exprime dans la même unité que les valeurs observées. Plus l'écart type est grand, plus la dispersion des observations autour de la moyenne est importante.

Dispersion autour de la moyenne

Définition

L'écart-type d'une série est la racine carrée la variance. Il est noté $\sigma(X)$ et mesure la dispersion absolue, i.e. la dispersion absolue autour de la moyenne. Il s'exprime dans la même unité que les valeurs observées. Plus l'écart type est grand, plus la dispersion des observations autour de la moyenne est importante.

Il est donné par :

$$\sigma(X) = \sqrt{\mathbf{Var}(X)}$$

Dispersion autour de la moyenne

Définition

L'écart-type d'une série est la racine carrée la variance. Il est noté $\sigma(X)$ et mesure la dispersion absolue, i.e. la dispersion absolue autour de la moyenne. Il s'exprime dans la même unité que les valeurs observées. Plus l'écart type est grand, plus la dispersion des observations autour de la moyenne est importante.

Il est donné par :

$$\sigma(X) = \sqrt{\mathbf{Var}(X)}$$

Remarque

L'écart-type $\sigma(X)$ s'interprète également comme la moyenne quadratique Q des écarts à la moyenne arithmétique $\overline{x}$. Variance et écart-type sont des indicateurs très utilisés en finance de marché, puisqu'ils représentent le **risque** associé à un actif.

Dispersion autour de la moyenne

- Soit c une constante.

Propriété 1

$\sigma(c) = 0$

- Soit a et b deux sont deux réels (constants) même niveau.

Propriété 2

$\sigma(aX + b) = |a|\sigma(X)$

Remarque

La seconde propriété rend l'écart-type plus naturel à manipuler que la variance.

Sommaire

Dispersion autour de la moyenne

Définition

Le **coefficient de variation** CV d'une variable statistique est le ratio entre l'écart-type et la moyenne exprimé sous la forme d'un pourcentage. Il s'agit d'une mesure de dispersion **relative**, sans dimension et indépendant de l'unité de la série.

Dispersion autour de la moyenne

Définition

Le **coefficient de variation** CV d'une variable statistique est le ratio entre l'écart-type et la moyenne exprimé sous la forme d'un pourcentage. Il s'agit d'une mesure de dispersion **relative**, sans dimension et indépendant de l'unité de la série.

$$CV = \frac{\sigma}{\overline{x}} \times 100$$

C'est un indicateur de l'homogénéité de la population. On considère qu'un coefficient de variation inférieur à 15% indique que la population est homogène, tandis qu'un coefficient supérieur à 15% indique que les valeurs sont relativement dispersées. Plus il est proche de 0, plus les données sont homogènes.

Dispersion autour de la moyenne

Application:

Notes	Effectifs
$[0, 5[$	4
$[5, 10[$	17
$[10, 15[$	26
$[15, 20[$	3
Total	50

Table: Distribution des notes obtenues à un devoir de Statistiques pour 50 étudiants

- Calculer la moyenne arithmétique $\overline{x}$.
- Calculer l'écart-type σ.
- Calculer le coefficient de variation de cette distribution. Interpréter.

Dispersion autour de la moyenne

On dresse le tableau suivant:

Notes	Centre de classe x_i	n_i	$n_i x_i$	$n_i x_i^2$
$[0,5[$	2,5	4	10	25
$[5,10[$	7,5	17	127,5	956,25
$[10,15[$	12,5	26	325	4062,5
$[15,20[$	17,5	3	52,5	918,75
Total	/	50	515	5962,5

- Calcul de la moyenne : $\overline{x} = \frac{1}{N}\sum_{i=1}^{4} n_i x_i = \frac{515}{50} = 10,3$. La note moyenne est de 10,3.
- Calcul de l'écart-type : $\sigma(x) = \sqrt{\mathbf{Var}(x)} = \left[\frac{1}{N}\sum_i n_i x_i^2 - \overline{x}^2\right]^{0,5} = \left[\frac{1}{50} \times 5962,5 - (10,3^2)\right]^{0,5} = 3,63$. La dispersion moyenne autour de 10,3 est égale à 3,63.

Dispersion autour de la moyenne

Calcul du coefficient de variation:

$$CV = \frac{\sigma(x)}{\overline{x}} = \frac{3,63}{10,3} = 0,352$$

Le coefficient de variation CV est de 35,2%$>$ 15%. La distribution des notes est hétérogène, elle est donc relativement dispersée.

Sommaire

Application

Sur une population, on a dosé le taux de cholestérol, exprimé en cg/L (x_i) et on a obtenu les résultats suivants :

Classes	$[84, 96[$	$[96, 108[$	$[108, 120[$	$[120, 132[$	$[132, 144[$	$[144, 156[$	$[156, 168[$
Effectifs n_i	2	8	16	19	2	2	1

1. Quel est le caractère étudié?
2. Déterminer le mode.
3. Déterminer la médiane.
4. Déterminer les quartiles Q_1 et Q_3.
5. Déterminer la moyenne arithmétique et l'écart-type.
6. Déterminer le coefficient de variation. Interpréter.

Correction

1. Caractère étudié : le taux de cholestérol. Tableau statistique :

Classes	n_i	Centres	f_i	N_i	$f_i c_i$	$f_i c_i^2$
$[84, 96[$	02	90	0,040	2	3,600	324
$[96, 108[$	8	102	0,160	10	16,320	1664,640
$[108, 120[$	16	114	0,320	26	36,480	4158,720
$[120, 132[$	19	126	0,380	45	47,788	6032,880
$[132, 144[$	2	138	0,040	47	5,520	761,760
$[144, 156[$	2	150	0,040	49	6,000	900
$[156, 168[$	1	162	0,020	50	3,240	524,880
Total	50		1		119,040	14366,880

2. Classe modale : $Mo \in [120, 132[$. D'où l'on obtient le mode
$Mo = 120 + 12\left[\frac{(19-16)}{(19-16)+(19-2)}\right] = 121,8$

Correction

3. $\frac{N}{2} = 25$. $Mo \in [108, 120[$. Par interpolation linéaire on trouve:

$$M = 108 + (120 - 108)\left[\frac{25 - 10}{26 - 10}\right] = 119,25$$

4. $Q_1 = X_{N/4} = 12,5 \Rightarrow Q_1 \in [108, 120[$, alors

$$Q_1 = 108 + (120 - 108)\left[\frac{12,5 - 10}{26 - 10}\right] = 109,88$$

$$Q_3 = X_{[3N/4]} = 37.5 \implies Q_3 \in [120,\ 132[\,,\ donc$$

$$Q_1 = 120 + (132\ - 120) \times \left[\frac{37.5 - 26}{45 - 26}\right] = 127.26.$$

5. Moyenne arithmétique et variance : $\overline{x} = \sum_{i=1}^{7} f_i c_i = 119,04$

$$var(x) = \sum_{i=1}^{7} f_i c_i^2 - (\overline{x})^2 = 14366.880 - (119.040)^2 = 196.358.$$

donc

$$\sigma(x) = \sqrt{var(x)} = \sqrt{196.358} = 14.012.$$

6. $CV = \frac{\sigma}{\overline{x}} = 11,7\% < 15\%$. L'échantillon est bien homogène.