République Tunisienne
Présidence du Gouvernement

Ecole Nationale d'Administration
24, Avenue du Dr Calmette Mutuelle-ville 1082 Tunis
Tél. (+216) 848 300 Fax (+216) 794 188
www.ena.nat.tn

STATISTIQUE ET CALCUL DE PROBABILITE (COURS)

Par
Hassen MZALI
Professeur en méthodes quantitatives

Septembre 2013

Introduction

Généralités

Les statistiques

Le mot « *statistiques* », au pluriel, désigne l'ensemble des données chiffrées qui regroupent toutes les observations faites sur des faits relatifs à un même phénomène qui concerne un groupe d'individus ou d'objets. Ces données sont essentiellement tirées des recensements de la population, des déclarations du registre d'état civil ou d'enquêtes appropriées et sont groupées sous forme de tableaux, de graphiques et d'indicateurs statistiques. On trouve des statistiques qui concernent la démographie, l'emploi, la santé, l'industrie, les transports, le commerce intérieur, le commerce extérieur, les indices de prix, la finance ..etc.

La statistique

Le mot « *statistique* », au singulier, désigne la discipline scientifique constituée par l'ensemble des procédés, des techniques ou des méthodes d'analyse visant, d'une part, à fournir, au moyen d'un nombre limité de caractéristiques, une Description simple et la plus complète possible d'une population envisagée sous l'angle d'un caractère donné. D'autre part, la statistique permet d'interpréter les caractéristiques ainsi déterminées afin de tirer des conclusions concernant la population étudiée et de prendre des décisions.

La statistique, en tant que méthode d'analyse des données quantitatives et qualitatives , comporte deux niveaux :

- *La statistique descriptive*, qui consiste en la collecte et la présentation de données, ainsi que leur première analyse. Le but est de représenter d'une manière compréhensible et utilisable l'information fournie par les données.
- *La statistique inférentielle ou statistique mathématique* , qui cherche à trouver les caractéristiques de la population mère à partir des observations faites sur un échantillon. Elle prend la suite de la statistique descriptive et fait appel au calcul des probabilités.

Une opération statistique se déroule en général en 4 étapes :

1. La collecte des données. Cette étape peut se réaliser de deux manières :

   - par recensement, auquel cas l'information porte sur la totalité des individus qui forment la population en question. (exemple : le recensement de la population et de l'habitat effectué par l'Institut National de la Statistique, une fois tous les dix ans).

   - par sondage, auquel cas on se limite à une partie de la population qu'on appelle échantillon. Un échantillon est un sous ensemble de la population totale. Il doit être représentatif, c'est-à-dire doit être choisi de telle sorte qu'il ait la même structure et les mêmes propriétés fondamentales de l'ensemble dont il est issu (population mère).

2. La représentation et l'organisation statistique des données. Cette étape peut se réaliser, soit à l'aide par de graphiques soit à l'aide d'indicateurs statistiques.
3. La modélisation. On distingue deux types de modèles : les modèles explicatifs et les modèles prévisionnels
4. L'interprétation des résultats.

Concepts statistiques de base

• *Population statistique :*

C'est l'ensemble de référence, c'est-à-dire l'ensemble des unités observées, qui constitue l'objet de l'étude de la statistique que l'on cherche à connaître. (La population des étudiants de l'université de Tunis, la population des salariés d'une entreprise industrielle). Une population ne signifie pas exclusivement un ensemble de personnes physiques, mais peut concerner des personnes morales ou des objets (entreprises, exploitations agricoles, universités, ampoules, voitures). Une population doit être bien définie. Sa définition est importante car elle conditionne l'homogénéité des unités observées et aussi la fiabilité des résultats.

• *Individu*

Un individu ou unité statistique est tout élément de la population ou de l'échantillon. La totalité des individus correspond à la population.

• *Caractère et modalité*

Pour chaque individu extrait d'une population ou d'un échantillon, on relève la valeur d'une ou plusieurs de ses caractéristiques. Le caractère ou variable statistique est un aspect particulier de l'individu que l'on désire étudier.

On distingue deux types de caractères : caractère qualitatif et caractère quantitatif.

Caractère qualitatif ou variable qualitative

Un caractère est dit **qualitatif** lorsqu'il est lié à une observation qui n'est pas mesurable.

(Exemple : lors de l'étude de la population estudiantine on peut s'intéresser à quelques unes de ses caractéristique telles que : La section du baccalauréat ; Le milieu de résidence (urbain, rural) ; Le sexe (masculin, féminin) ; La région de résidence (nord, centre, sud) ; L'état matrimonial (marié, veuf, divorcé, célibataire)….

Les **modalités** d'un caractère sont simplement les différentes rubriques d'une nomenclature définie a priori et associées à un caractère qualitatif. Une modalité est donc une des réponses possibles à un caractère. Par exemple le caractère milieu de résidence comporte deux modalités : milieu rural et milieu urbain. Ces modalités doivent former

une partition, c'est à dire doivent être exhaustives et disjointes : pour chaque individu on doit pouvoir lui associer une modalité et une seule. La **Nomenclature** désigne l'ensemble des modalités d'un caractère précédées d'un numéro.

- **Présentation du tableau statistique associé à un caractère qualitatif**

Modalités (numérotées) M_i	Effectifs n_i	Fréquences f_i
M_1 (1)	n_1	f_1
M_2 (2)	n_2	f_2
.	.	.
	.	.
M_r (r)	n_r	f_r
Ensemble	N	1

L'effectif total ***N*** est le nombre total d'individus observés

$$N = n_1 + n_2 + \cdots + n_r = \sum_{i=1}^{r} n_i$$

L'effectif n_i d'une modalité, appelé aussi fréquence absolue, est le nombre de fois où la modalité numéro i a été observée.

La fréquence relative f_i d'une modalité est le rapport de l'effectif n_i à l'effectif total N

$$f_i = \frac{n_i}{N} = \frac{n_i}{\sum_{i=1}^{r} n_i}$$

$$\sum_{i=1}^{r} f_i = 1$$

Remarque : les fréquences relatives peuvent être exprimées en pourcentage.

Lorsque les modalités ne permettent pas l'exhaustivité, c'est à dire lorsqu'il y a des individus qu'on ne peut classer dans le tableau, on peut rajouter une modalité, en bas du tableau, qu'on appelle « autres « ou « non réponses »

Caractère quantitatif ou variable quantitative

Lorsque les observations relatives à un caractère sont mesurables, le caractère est dit **quantitatif** (taille, âge, poids, moyenne du baccalauréat, superficie du logement,….). A chaque modalité correspond

un nombre différent.

Exemple : lors de l'étude de la population estudiantine, on peut s'intéresser à quelques-unes de ses caractéristique telles que :

- Le nombre d'enfants par ménage, Le nombre d'années d'études. nombre de voitures par ménage,…)
- L'âge, le poids, la taille, Le revenu des parents, la facture de l'électricité et du gaz, les dépenses en loyer, …)

On distingue deux types de caractères quantitatifs :

- Les caractères quantitatifs **discrets,** auxquels cas, les valeurs possibles de la variable sont des nombres isolés (en général des nombres entiers comme par exemple le nombre d'enfants d'un ménage, le nombre de voyages effectués à l'étranger …
- Les caractères quantitatifs **continus**, auxquels cas, les valeurs possibles de la variable sont a priori en nombre infini dans un intervalle de valeurs (comme par exemple la taille, l'âge, moyenne du baccalauréat ….)

Remarque : certains caractères discrets sont de préférence traités en tant que caractères continus. Exemple : le nombre d'ouvriers dans chaque entreprise, nombre de places de cinémas associées à chaque salle, ..)

- **Présentation du tableau statistique associé à un caractère quantitatif discret**

Valeur observées x_i	Effectifs n_i	Fréquences f_i	Fréquences cumulées $F_i \uparrow$
x_1	n_1	f_1	$F_1=0$
x_2	n_2	f_2	$F_2=f_1$
.	.	.	$F_3=f_1+f_2$
.	.	.	.
	.	.	.
x_p	n_p	f_p	F_p
Ensemble	N	1	

Fréquences cumulées croissantes F_i: le cumul des fréquences associées aux valeurs du caractère inférieures strictement à la valeur x_i

$$F_i = \sum_{j=1}^{i-1} f_j \quad \text{pour } i = 2, 3, \ldots p. \quad \text{et} \quad F_1 = 0$$

- **Présentation du tableau statistique associé à un caractère quantitatif continu**

Classes numérotée $[b_{i-1}- b_i[$	Centres c_i	Effectifs n_i	Fréquences f_i	Fréquence cumulée $F_i \uparrow$
$[b_0 - b_1[$	c_1	n_1	f_1	$F_1= f_1$
$[b_1 - b_2[$	c_2	n_2	f_2	$F_2= f_1+ f_2$
$[b_2 - b_3[$	.	.	.	$F_3= f_1+ f_2+ f_3$
.	.	.	.	.
	.	.	.	.
$[b_{p-1}- b_p[$	c_p	n_p	f_p	F_p
Ensemble		N	1	

Remarque : Par convention, les classes sont fermées à gauche et ouvertes à droite. Une classe est dite bornée si : $b_{i-1} \neq -\infty, \quad b_i \neq +\infty$

- le centre d'une classe bornée est : $\boxed{c_i = \frac{b_{i-1} + b_i}{2}, \quad c_i \approx x_i}$
- l'amplitude d'une classe bornée est : $\boxed{a_i = b_i - b_{i-1}}$

L'opérateur somme $\sum$

L'opérateur $\sum$ (lettre grecque *sigma* majuscule) permet d'écrire de manière compactée la somme d'une variable indicée entre deux bornes.

On peut par exemple écrire : $x_1 + x_2 + x_3 + x_4 + x_5 = \sum_{i=1}^{5} x_i$

D'une manière générale : $x_1 + x_2 + \cdots + x_n = \sum_{i=1}^{n} x_i$

Cette formule se lit de bas en haut : somme de i égal 1 à i égal n de x indice i

Remarque : On peut établir les résultats suivants :

1) $\sum_{i=1}^{i=n} x_i + \sum_{i=1}^{i=n} y_i = \sum_{i=1}^{i=n} (x_i + y_i)$

2) $\sum_{i=1}^{i=n} a = na$

3) $\sum_{i=1}^{i=n} (x_i + a) = \sum_{i=1}^{i=n} x_i + na$

4) $\sum_{i=1}^{i=n} a x_i = a \sum_{i=1}^{i=n} x_i$

5) $\sum_{i=1}^{i=n} x_i y_i \neq \sum_{i=1}^{i=n} x_i \sum_{i=1}^{i=n} y_i$

6) $\sum_{i=1}^{i=n} \left(\frac{x_i}{y_i} \right) \neq \frac{\sum_{i=1}^{i=n} x_i}{\sum_{i=1}^{i=n} y_i}$

7) $\left(\sum_{i=1}^{i=n} x_i \right)^2 \neq \sum_{i=1}^{i=n} x_i^2$

8) $\sum_{i=1}^{i=n} (x_i + y_i)^2 = \sum_{i=1}^{i=n} x_i^2 + \sum_{i=1}^{i=n} y_i^2 + 2 \sum_{i=1}^{i=n} x_i y_i$

9) $\sum_{i=1}^{n} \sum_{j=1}^{j=m} x_i y_j = \sum_{j=1}^{j=m} \sum_{i=1}^{i=n} x_i y_j = \sum_{i=1}^{i=n} x_i \sum_{j=1}^{j=m} y_j = \sum_{j=1}^{j=m} y_j \sum_{i=1}^{i=n} x_i$

Chapitre 1 :
Séries statistiques à un seul caractère

Chapitre I : Séries Statistiques à un seul caractère

Série statistique simple

Variable discrète

On appelle série statistique d'une variable discrète tout ensemble de couples $\{(x_i, n_i)\}$, $i = 1,....p$ ou encore $\{(x_i, f_i)\}$, $i = 1,....p$, où les x_i désignent les valeurs possibles prises par la variable et les n_i les effectifs correspondants.

Variable continue

On appelle série statistique d'une variable continue tout ensemble de couples $\{([b_i, b_{i+1}[\ ,\ n_i)\}$, $i = 1,....p$ ou encore $\{([b_i, b_{i+1}[,\ f_i)\}$, $i = 1,....p$.

Principales représentations graphiques

Les tableaux statistiques donnent un premier résumé statistique des résultats d'une enquête. Cependant, dans le cas où la variable présente plusieurs modalités ou dans le cas ou nous avons à comparer deux ou plusieurs distributions, il est préférable de représenter les résultats à l'aide de graphiques.

Cas d'une variable qualitative

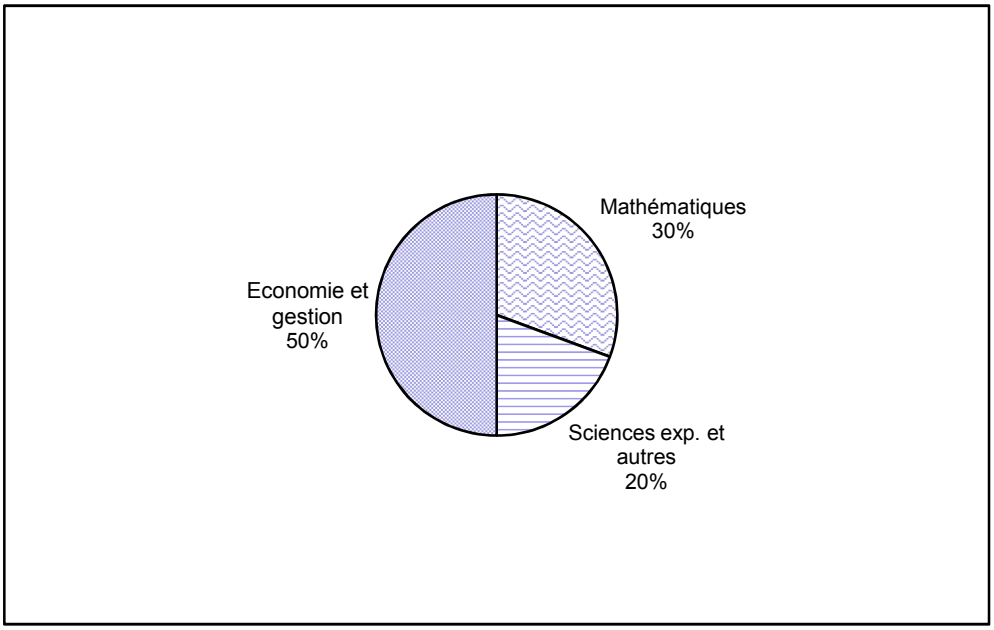
Il y a plusieurs manières de représenter graphiquement une variable qualitative. Le choix du type de la représentation graphique dépend des différentes modalités du caractère. On distingue essentiellement le diagramme circulaire, appelé aussi diagramme à secteurs et le diagramme en tuyaux d'orgues, appelé aussi diagramme à bandes ou encore diagramme en barres.

Exemple : D'après une enquête menée à l'Eole Supérieure de Commerce de Tunis, la répartition de 50 étudiants selon la section du baccalauréat est reportée dans le tableau suivant :

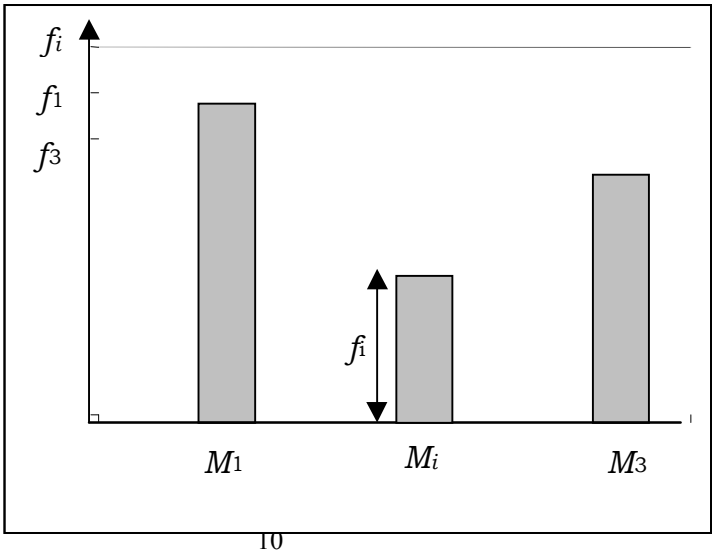
Section du baccalauréat M_i	Effectifs n_i	Fréquences f_i	Angles α_i
Economie et gestion	25	0,5	180°
Mathématiques	15	0,3	108°
Sciences exp. et autres	10	0,2	72°
Ensemble	50	1	360°

- Le principe de la représentation du diagramme à secteurs est le suivant : effectif total représenté par un disque, modalité représentée par un secteur circulaire dont la surface est proportionnelle à la fréquence, angle de chaque secteur est égale à :

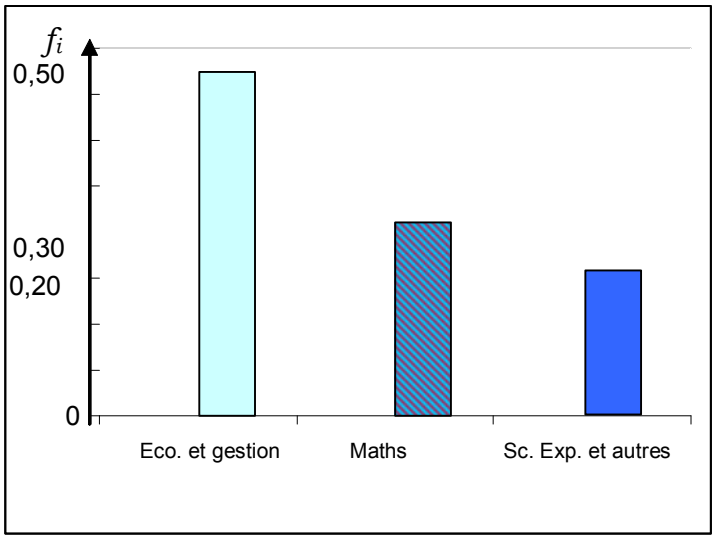
$$\alpha_i° = 360° \times f_i$$



Le principe de la représentation du diagramme à bandes ou en tuyaux d'orgues est le suivant : Association à chacune des modalités M_i du caractère, qui sont placées sur un axe horizontal, une bande verticale ayant une hauteur proportionnelle à la fréquence f_i (ou à l'effectif n_i). Les bases des bandes doivent être égales et équidistantes.



Concernant l'exemple de la distribution de la population active agricole en Tunisie, la représentation par un diagramme en tuyaux d'orgue est la suivante :

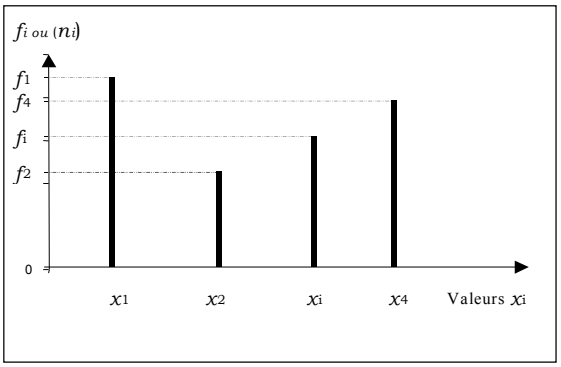


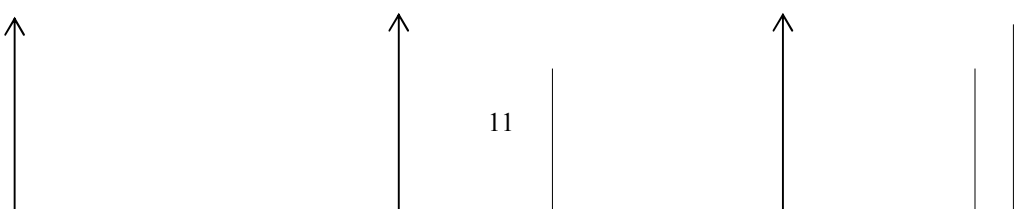
Cas d'une variable quantitative

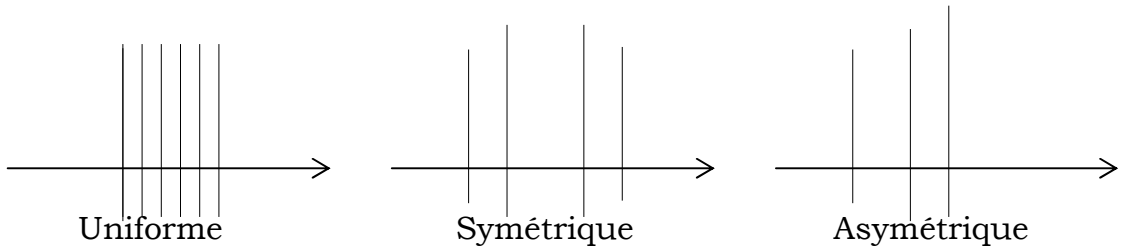
Série statistique discrète

La représentation utilisée s'appelle diagramme en bâton.

Il s'agit de la figure obtenu sur un repère cartésien en associant à chaque point de coordonnés $(x_i, 0)$ un segment vertical dont la longueur est proportionnelle à la fréquence f_i (ou à l'effectif n_i).







L'intérêt de cette représentation est double. D'une part, elle permet de donner une idée générale sur la forme de la distribution. D'autre part, elle permet de repérer les valeurs aberrantes.

Exemple : La distribution du même échantillon d'étudiants selon le nombre de personnes par ménage est résumée dans le tableau suivant :

Valeurs x_i	Effectifs n_i	Fréquences f_i	Fréquences cumulées $F_i \uparrow$
1	20	0,40	0
2	15	0,30	0,40
3	10	0,20	0,70
4	5	0,10	0,90
Total	50	1,00	

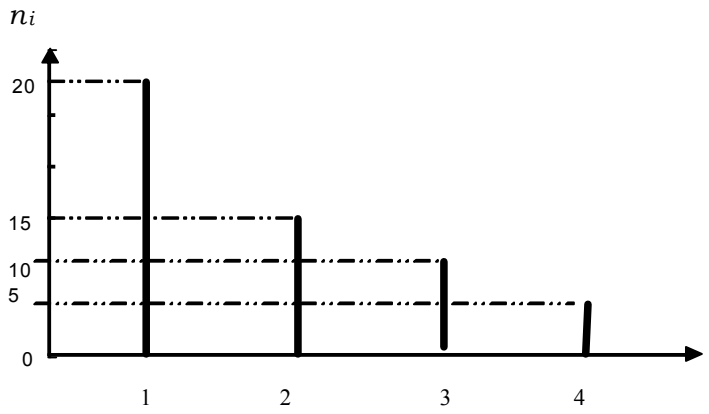


Diagramme en bâtons des effectifs

Série statistique continue

Graphiquement, on représente une série statistique continue par un histogramme. Il s'agit d'une figure obtenue sur un repère cartésien en représentant pour chaque classe $[b_{i-1} - b_i[$ un rectangle de surface S_i proportionnelle à l'effectif n_i ou à la fréquence f_i. Les rectangles de l'histogrammes sont contigus.

Principe de construction de l'histogramme

$S_i = base \times hauteur = a_i \times h_i = n_i \times a^*$ d'où, $h_i = \frac{n_i}{a_i} \times a^* = d_i \times a^*$

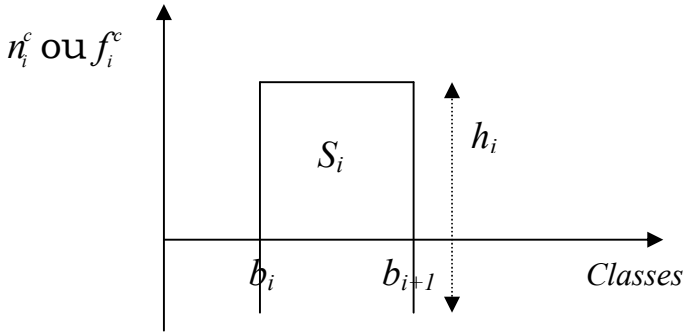
Le a* est appelée amplitude de référence. Elle est choisie arbitrairement de manière à faciliter la représentation graphique (valeurs sur l'axe des ordonnées).

La hauteur h_i est dans ce cas appelée effectif corrigé qu'on note par n_i^c.

La densité d_i d'une classe est : $d_i = \frac{n_i}{a_i}$. Il s'agit du nombre d'individus par unité d'amplitude.

Remarque : on peut utiliser les fréquences corrigées à la place des effectifs corrigés.

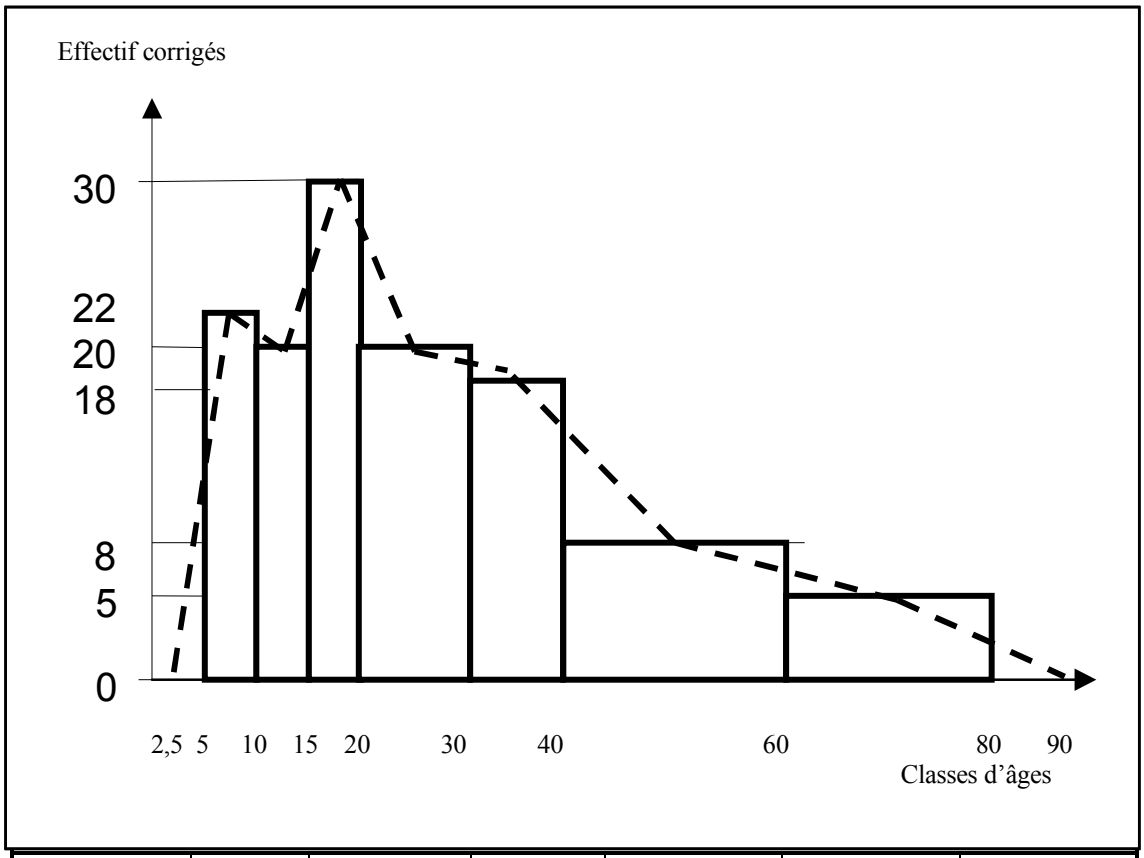
- Dans le cas de classes d'**amplitudes égales,** il n'est pas nécessaire de calculer les fréquences corrigées ou les effectifs corrigés. On peut utiliser directement les effectifs ou les fréquences comme hauteurs des rectangles. En revanche, dans le cas de classes d'**amplitudes inégales,** les hauteurs des rectangles doivent être proportionnelles à la **densité**, afin d'avoir une **surface** proportionnelle à l'effectif.



Exemple : La répartition de 100 individus par classes d'âges est donnée par le tableau suivant :

Classes d'âges	**Effectifs** n_i	**Amplitudes** a_i	**Densité** d_i	**Effectifs corrigés** n_i^c	**Fréquences** f_i	**Fréquences corrigées** f_i^c
[5 , 10[	11	5	2,2	22	0,11	0,22
[10 , 15[	10	5	2	20	0,10	0,20
[15 , 20[	15	5	3	30	0,15	0,30
[20 , 30[	20	10	2	20	0,20	0,20
[30 , 40[	18	10	1,8	18	0,18	0,18

[40 , 60[	16	20	0,8	8	0,16	0,8



[60 , 80[	10	20	0,5	5	0,10	0,5
Total	100				1	

Remarque : Dans certains cas, la borne inférieure de la première classe et la borne supérieure de la dernière classe ne sont pas données. Par convention, on retient comme amplitude de la première classe celle de la deuxième classe et comme amplitude de la dernière classe celle de l'avant dernière classe.

Polygone des fréquences

Il s'agit d'une ligne brisée reliant les milieux des sommets des rectangles de l'histogramme. La fermeture se fait par deux points sur l'axe des abscisses situés respectivement à un demi-intervalle de la borne inférieure de la première classe et de la borne supérieure de la dernière classe. Dans notre exemple, le polygone des effectifs est présenté par la ligne en pointillée gras.

Fréquences cumulés croissantes, fonction de répartition et diagrammes

cumulatif

On appelle fonction de répartition d'une variable statistique quantitative toute application définie par :

$$F : \Re \to [0 \ , \ 1]$$
$$x_i \mapsto F(x_i) = prop\ (X < x_i)$$

F (x_i) est égale à la proportion des individus ayant une valeur du caractère strictement inférieur à x_i.

Cas d'une variable statistique discrète

On donne, dans le tableau suivant, la distribution du même échantillon d'étudiants selon le nombre de personnes par ménage.

x_i	**Effectifs** n_i	**Fréquences** f_i	**Fréquence cumulée** $F_i \uparrow$
1	20	0,40	0
2	15	0,30	0,40
3	10	0,20	0,70
4	5	0,10	0,90
Ensemble	50	1,00	

$$F(1) = prop(x < 1) = 0$$
$$F(1{,}5) = prop(x < 1{,}5) = prop(x = 1) = 0{,}40$$
$$F(2) = prop(x < 2) = prop(x = 1) = 0{,}40$$
$$F(2{,}5) = prop(x < 2{,}5) = prop(x = 1) + prop(x = 2) = 0{,}70$$
$$F(3) = prop(x < 3) = prop(x = 1) + prop(x = 2) = 0{,}70$$
$$F(4) = prop(x < 4) = prop(x = 1) + prop(x = 2) + prop(x = 3) = 0{,}90$$

Ainsi, la formulation de la fonction de répartition de cette distribution statistique est :

$$F = \begin{cases} 0 & \text{si } x \le 1 \\ 0{,}40 & \text{si } 1 < x \le 2 \\ 0{,}70 & \text{si } 2 < x \le 3 \\ 0{,}90 & \text{si } 3 < x \le 4 \\ 1 & \text{si } x > 4 \end{cases}$$

La représentation graphique de la fonction de répartition, appelée

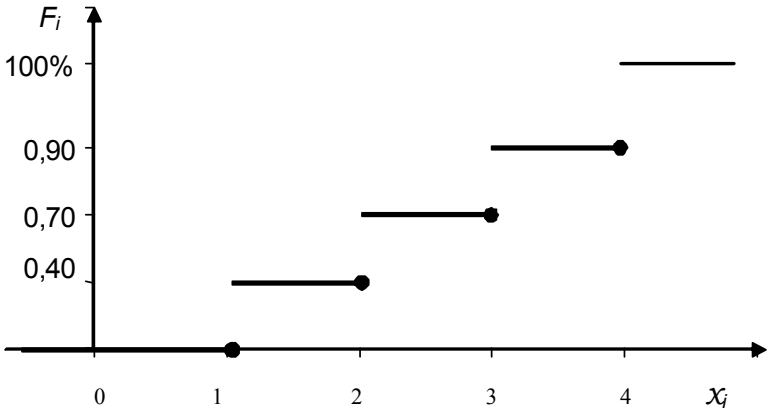


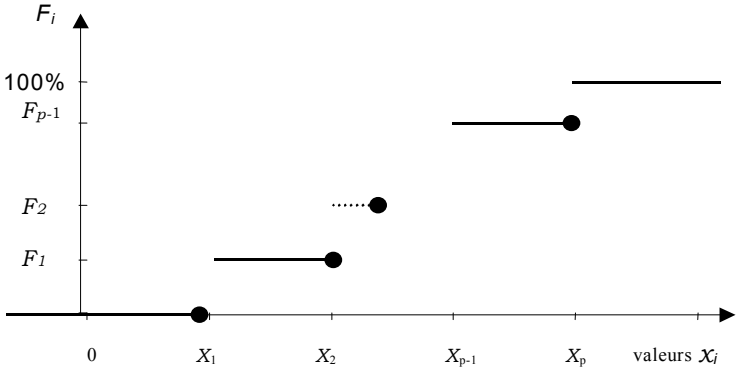
diagramme cumulatif ou diagramme intégral, est :

Ce diagramme permet de visualiser l'évolution des fréquences cumulées liées aux valeurs de la variable. Le caractère étant discret, la courbe des fréquences cumulées croissante est la représentation graphique d'une fonction en escalier.

D'une manière générale, La fonction de répartition est constante par intervalle. Sa formulation est la suivante :

$$F = \begin{cases} 0 & x \le x_1 \\ f_1 & x_1 < x \le x_2 \\ f_1 + f_2 & x_2 < x \le x_3 \\ : & \\ f_1 + f_2 + \ldots\ldots + f_{\text{p-1}} & x_{p-1} < x \le x_p \\ 1 & x > x_p \end{cases}$$

La représentation graphique de la fonction de répartition, appelée diagramme intégral, est :



L'intérêt de la représentation graphique est qu'elle permet de retrouver pour toute valeur de x_i donnée, la proportion des individus ayant une valeur de la variable strictement inférieure à x_i.

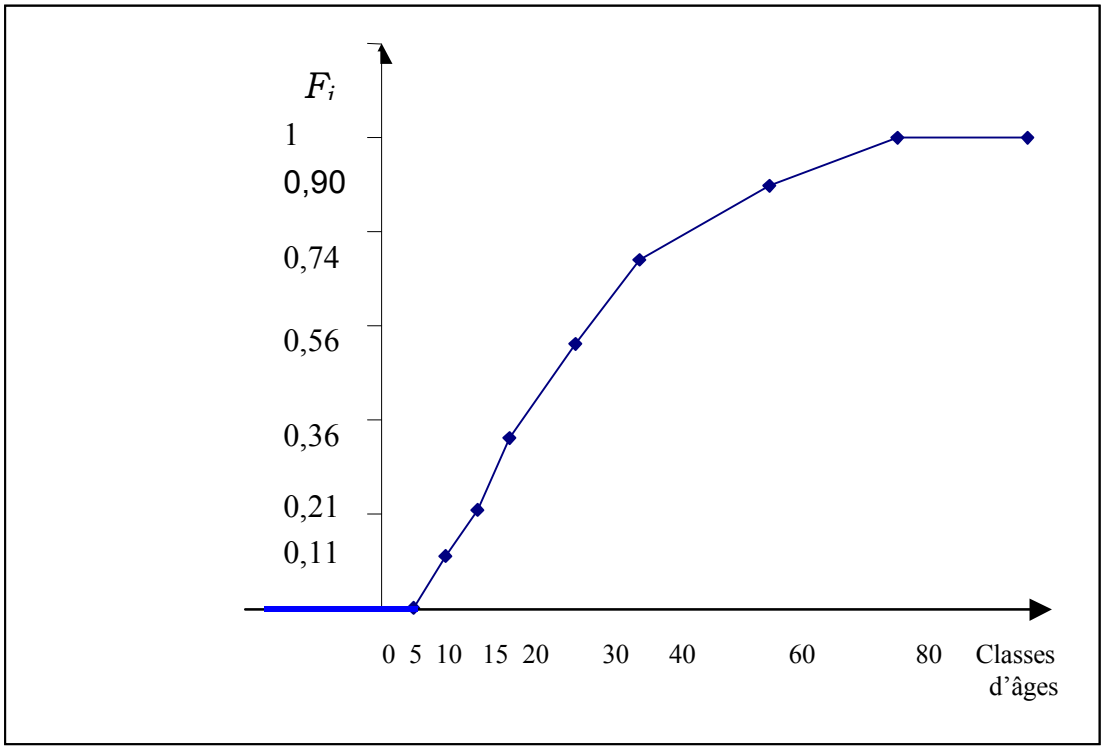
Cas d'une variable statistique continue

Classes d'âges	**Effectifs** n_i	**Effectifs cumulés croissants** $n_i \uparrow$	**Fréquences relatives** f_i	**Fréquences cumulées croissantes** $F_i \uparrow$ $F(b_i)$
[5 , 10[	11	11	0,11	0,11
[10 , 15[	10	21	0,10	0,21
[15 , 20[	15	36	0,15	0,36
[20 , 30[	20	56	0,20	0,56
[30 , 40[	18	74	0,18	0,74
[40 , 60[	16	90	0,16	0,9
[60 , 80[	10	100	0,10	1
total	100		1	

La lecture des fréquences cumulées croissantes se fait par rapport à la borne supérieure de chaque classe.

La représentation graphique de la fonction de répartition appelée courbe cumulative est la suivante :

D'une manière générale, la courbe cumulative, dans le cas d'une variable continue, est une ligne brisée obtenue en joignant différents points de coordonnés (b_i , F_i) où *bi* désigne la borne supérieure de la classe *i,* et F_i la fréquence cumulée croissante correspondante.



Remarque : on peut aussi représenter graphiquement la courbe des fréquences cumulées décroissantes, lesquelles sont définies par la proportion des individus ayant une valeur du caractère supérieure ou égale à la borne inférieure de la classe *i*.

Chapitre 2 :
Les paramètres de position d'une série statistique

Chapitre II : Les paramètres de position d'une série statistique

La représentation graphique d'une série statistique nous donne une idée assez générale sur la distribution. Pour confirmer certaines impressions sur la série et pour en donner plus de précision, nous serons amenés à trouver une ou plusieurs valeurs centrales de la variable, capables de résumer la série en caractérisant l'ordre de grandeur des observations. De telles valeurs centrales sont appelées paramètres de tendance centrale ou caractéristiques de position. Un indicateur de position doit être défini de manière rigoureuse et objective, doit tenir compte de l'ensemble des observations de la série et doit être exprimé dans la **même unité** que la variable.

Le mode

On appelle mode ou valeur dominante d'une série statistique la valeur observée de la variable ayant le plus grand effectif (ou la fréquence la plus élevée). On note généralement le mode M_O.

Remarques :

Le mode est exprimé dans la même unité que la variable.

- Si toutes les modalités ont la même fréquence alors la distribution statistique ne possède pas de mode. On parle alors de distribution uniforme.
- Lorsqu'une série possède un seul mode, on dit que la distribution est unimodale. En revanche, lorsqu'elle en possède deux ou plusieurs elle est respectivement qualifiée de bimodale et multimodale.

Le calcul du mode dépend de la nature de la variable, discrète ou continue.

Cas d'une variable discrète

Exemple 1 :

On considère les notes obtenues en statistique par un groupe de 20 étudiants : 7, 13, 5, 15, 12, 9, 7, 8, 14, 16, 13, 6, 13, 10, 13, 12, 10, 7, 12, 13.

Le mode de cette série correspond à la note la plus fréquente, soit $M_0 = 13$, valeur qui apparaît cinq fois. L'interprétation en est que la note la plus fréquente est 13.

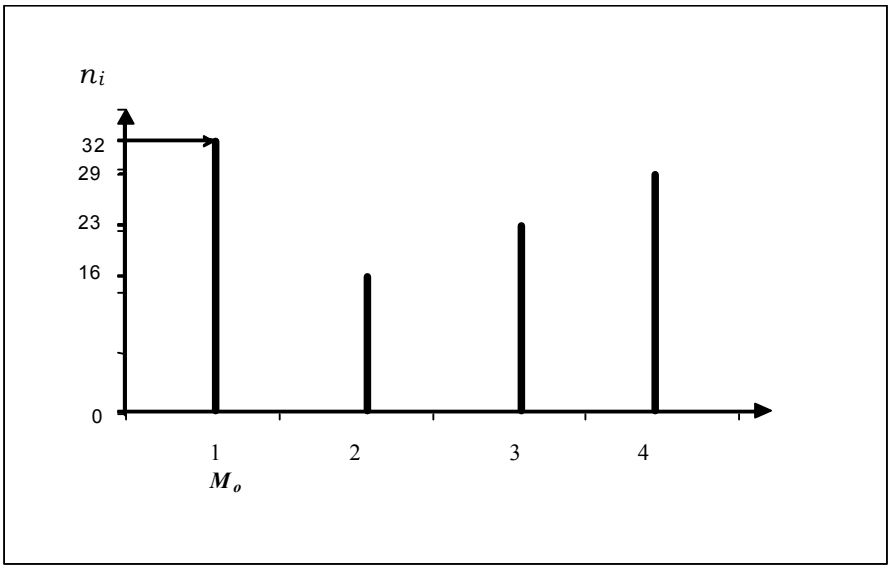
Exemple 2 :

On considère une distribution statistique d'une population de 100 ménages selon le nombre d'enfants :

Valeurs x_i	Effectifs n_i	Fréquences f_i
1	32	0,32
2	16	0,16
3	23	0,23
4	29	0,29
Total	100	1,00

Le mode de cette série est : M_o=1. Il signifie que la plupart des ménages ont **un seul** enfant.

Graphiquement, le mode correspond à l'abscisse du bâton le plus élevé.



Cas d'une variable continue

Dans le cas d'une variable continue groupée en classes, on parle plutôt de classe modale. La classe modale est la base du rectangle ayant la hauteur la plus élevée.

Cependant, on distingue deux cas selon que les amplitudes des classes sont égales ou inégales.

Cas d'amplitudes identiques

Dans ce cas, la classe modale est la classe d'effectif n_i le plus élevé, soit $[b_{i-1} - b_i[$. L'effectif de la classe qui précède la classe modale est n_{i-1} et celui de la classe qui suit la classe modale est n_{i+1}. La détermination du mode à partir de la classe modale se fait de la façon générale suivante :

$$M_0 = b_{i-1} + a_i\left(\frac{m_1}{m_1 + m_2}\right) = \frac{b_{i-1} \times m_2 + b_i \times m_1}{m_1 + m_2}$$

avec :

b_{i-1} : borne inférieure de la classe modale

b_i : borne supérieure de la classe modale

a_i : amplitude de la classe modale

$m_1 = n_i - n_{i-1}$

$m_2 = n_i - n_{i+1}$

Exemple : Soit la distribution de la population de 20 ménages selon le revenu des deux parents :

Classe de Revenu en DT	Amplitudes	Effectifs n_i	Fréquences f_i
[200-300[	100	40 } m_1 {	0,20
[300-400[	100	***60***	***0,30***
[400-500[	100	30 } m_2 {	0,15
[500-600[	100	50	0,25
[600-700[	100	20	0,10
Total		**200**	**1**

La classe modale est la classe ayant la fréquence la plus élevée. C'est la classe $[300-400[$ dans notre exemple. Dans ce cas, le mode est calculé par :

$$M_0 = 300 + 100\left(\frac{60-40}{(60-40)+(60-30)}\right) = 340 \text{ DT.}$$

On interprète en disant que le salaire le plus fréquent est de 340 Dinars.

Remarque : *On peut aussi utiliser les fréquences relatives au lieu des effectifs.*

Dans ce cas, on aura :

$$M_0 = 300 + 100\left(\frac{0{,}30-0{,}20}{(0{,}30-0{,}20)+(0{,}30-0{,}15)}\right) = 340 \text{ DT.}$$

Cas d'amplitudes inégales

Dans le cas où les amplitudes sont différentes, la classe modale est la classe de densité (ou de fréquence corrigée) la plus élevée, ou encore d'effectif corrigé le plus élevé.

Le mode est donné par : $M_0 = b_{i-1} + a_i\left(\frac{m_1}{m_1+m_2}\right) = \frac{b_{i-1}\times m_2 + b_i \times m_1}{m_1+m_2}$

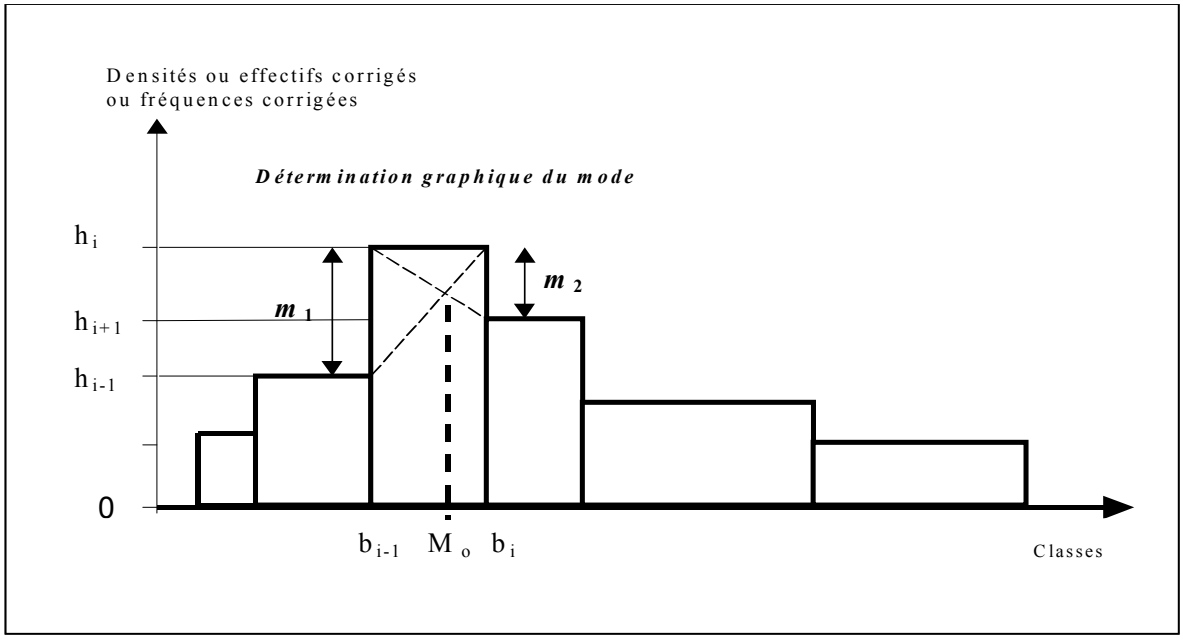
b_{i-1} : borne inférieure de la classe modale

b_i : borne supérieure de la classe modale

a_i : amplitude de la classe modale

$m_1 = h_i - h_{i-1}$, $(m_1 = n_i^c - n_{i-1}^c)$;
$m_2 = h_i - h_{i+1}$, $(m_2 = n_i^c - n_{i+1}^c)$
où h_i, h_{i-1}et h_{i+1} sont les effectifs corrigés

Classes numérotée $[b_{i-1}–b_i[$	**Amplitudes** a_i	**Effectifs** n_i^c	**Fréquences corrigées** f_i^c ou h_i
$[b_0–b_1[$	a_1	n_1	f_1^c ou h_1
$[b_1–b_2[$	a_2	n_2	f_2^c ou h_2
	.	n_{i-1}^c } m_1 {	f_{i-1}^c ou h_{i-1}
$[b_{i-1}–b_i[$	a_i	n_i^c	f_i^c ou h_i
	.	n_{i+1}^c } m_2 {	f_{i+1}^c ou h_{i+1}
$[b_{p-1}–b_p[$	a_p	n_p^c	f_p^c ou h_p
Ensemble		N	1



Exemple : Soit la répartition de 100 personnes selon leur âge :

Classes d'âges	Effectifs n_i	Amplitudes a_i	Densités d_i	Effectifs corrigés n_i^c ou h_i
[5 , 10[	11	5	2,2	22
[10 , 15[	10	5	2	**20**
[15 , 20[	15	**5**	3	**30**
[20 , 30[	20	10	2	**20**
[30 , 40[	18	10	1,8	18
[40 , 60[	16	20	0,8	8
[60 , 80[	10	20	0,5	5
total	100			

La plus grande hauteur appartient à la classe $[15-20[$. Donc : $M_0 \in [15-20[$

et $M_0 = 15 + 5\left(\frac{30-20}{(30-20)+(30-20)}\right) = 17{,}5 = \frac{15\times 10 + 20\times 10}{10+10}$

On interprète en disant que l'âge observé le plus fréquemment est d'environ 17ans et 6mois.

Remarque : On peut aussi utiliser les fréquences corrigées à la place des effectifs corrigés. Dans ce cas on aura

$$M_0 = 15 + 5\left(\frac{0{,}3-0{,}2}{(0{,}3-0{,}2)+(0{,}3-0{,}2)}\right) = 17{,}5 = \frac{15\times 0{,}10 + 20\times 0{,}10}{0{,}1+0{,}1}$$

La médiane

Soit une série statistique ordonnée par valeurs croissantes ou décroissante. La médiane, notée généralement Me, est la valeur de la variable qui partage la population en deux groupes d'effectifs égaux. En d'autres termes, la médiane est la valeur de la variable située au « **milieu** » d'une série ordonnée telle que la moitié des individus prenne une valeur qui lui soit inférieure, l'autre moitié prenant par conséquent une valeur qui lui soit supérieure.

Comme pour le mode, le calcul de la médiane dépend de la nature de la variable, discrète ou continue.

Cas d'une variable discrète

La détermination de la médiane d'une série statistique nécessite d'abord de ranger par ordre croissant (ou décroissant) les valeurs observées.

- Si la série comporte un nombre impair de valeurs, soit N valeurs, la médiane sera la valeur de rang $(\frac{N+1}{2})$.
- Si la série comporte un nombre pair de valeurs, on parle d'intervalle médian. Ce dernier est défini par :

$$]\text{la } (\tfrac{N}{2})\text{ième valeur , la } (\tfrac{N}{2}+1)\text{ième valeur}].$$

Toute valeur appartenant à cet intervalle fait fonction de médiane.

Remarque : certains proposent de choisir comme médiane le centre de l'intervalle médian. La médiane, dans ce cas, n'est pas forcément une valeur observée.

Exemple 1 *:*

On considère la répartition de 9 ménages selon le nombre d'enfants par ménage.

nombre d'enfants par ménage	0	0	1	1	**2**	3	3	3	4
Rang (ordre croissant)	1	2	3	4	**5ième**	6	7	8	9
	4 observations				***Mé***	4 observations			

La médiane, dans ce cas, correspond à la cinquième valeur : $Mé = 2$ enfants par ménage. On dit qu'il y a autant de ménage qui ont moins de 2 enfants que de ménage qui ont plus de 2 enfants.

Exemple 2 *:*

On considère la répartition de 10 ménages selon le nombre d'enfants par ménage.

nombre d'enfants par ménage	0	0	1	1	**2**	**3**	3	3	4	4
Rang (ordre croissant)	1	2	3	4	**5ième**	**6ième**	7	8	9	10
	4 observations				***Intervalle médian***		4 observations			

Dans ce cas on parle plutôt d'intervalle médian]2 , 3], correspondant à la]cinquième valeur , sixième valeur].

Remarque : certains retiennent comme valeur médiane le centre de l'intervalle médian, soit $Mé = \frac{2+3}{2} = 2{,}5$ enfants. Cette valeur ne correspond pas à une valeur réellement observée.

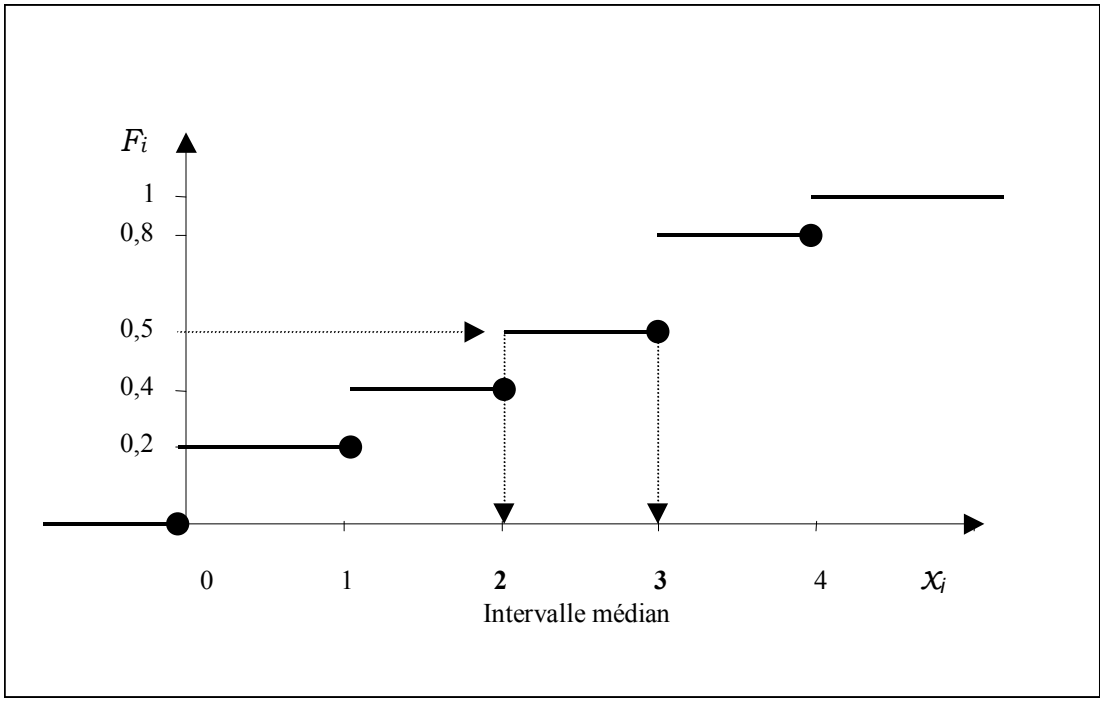
Exemple 3 *:*

nombre d'enfants par ménage	0	0	1	1	**2**	**2**	3	3	3	4
Rang (ordre croissant)	1	2	3	4	**5ième**	**6ième**	7	8	9	10
	4 observations				***Intervalle médian***		4 observations			

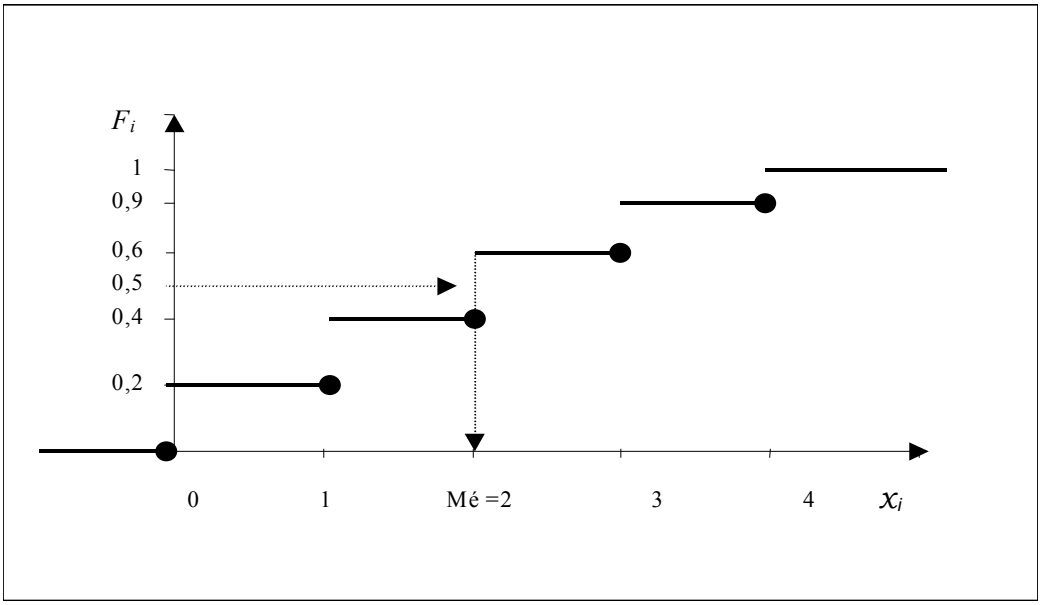
Dans le cas de cette distribution statistique, l'intervalle médian est :]2 , 2]. La valeur médiane est donc égale à 2 .

Les représentation groupée des données des exemple 2 et 3 nous donnent les deux tableaux suivants :

Exemple 2			
x_i	**Effectifs** n_i	**Fréquences** f_i	**Fréquences cumulées** $F_i \uparrow$
0	2	0,2	0
1	2	0,2	0,2
2	1	0,1	0,4
3	3	0,3	**0,5**
4	2	0,2	0,8
Ensemble	10	**1**	



Exemple 3			
x_i	**Effectifs** n_i	**Fréquences** f_i	**Fréquences cumulées** $F_i \uparrow$
0	2	0,2	0
1	2	0,2	0,2
2	2	0,2	0,4
3	3	0,3	0,6
4	1	0,1	0,9
Ensemble	10	**1**	



Cas d'une variable continue

Il n'y a aucune différence de calcul pour la médiane selon que les classes sont d'amplitudes constantes ou variables.

Le calcul de la médiane dans le cas de variable continue passe, d'abord, par la détermination de la classe médiane. Ensuite, par interpolation linéaire, on peut calculer la valeur précise de la médiane à l'intérieur de la classe médiane.

Soit $[b_{i-1}–b_i[$ la classe médiane, a_i l'amplitude de la classe médiane, N_i l'effectif cumulé croissant de la classe médiane, N_{i-1} l'effectif cumulé croissant de la classe avant la classe médiane et N l'effectif total.

L'expression de la médiane est donnée par :

$$Mé = b_{i-1} + a_i \left(\frac{\frac{N}{2} - N_{i-1}}{N_i - N_{i-1}} \right)$$

La même démarche pourrait être utilisée en remplaçant les fréquences absolues par les fréquences relatives :

$$Mé = b_{i-1} + a_i \left(\frac{0{,}5 - F_{i-1}}{F_i - F_{i-1}} \right),$$

où F_i désigne la fréquence cumulée croissante de la classe médiane, F_{i-1} la fréquence cumulée croissante de la classe qui précède la classe médiane.

Exemple : En reprenant notre exemple sur la répartition des 100 individus selon leur âge :

Classes d'âges	Effectifs n_i	Effectifs cumulée croissants $n_i \uparrow$	Fréquences f_i	Fréquences cumulées croissantes $F_i \uparrow$
[5 , 10[	11	11	0,11	0,11
[10 , **15[**	10	**21**	0,10	**0,21**
[15 **, 20[** → **Mé** → **[20 , 30 [**	→ 20	**36** $\frac{N}{2}=50$ **56**	→	
[30 , 40[	18	74	0,18	0,74
[40 , 60[	16	90	0,16	0,9
[60 , 80[	10	100	0,10	1
Total	100		1	

Le calcul, par interpolation linéaire, de la médiane donne :

20 ------ 0,36

Mé ------ 0,50

30 ---- ---- 0,56

$$\frac{Mé-20}{30-20}=\frac{0{,}50-0{,}36}{0{,}56-0{,}36}$$

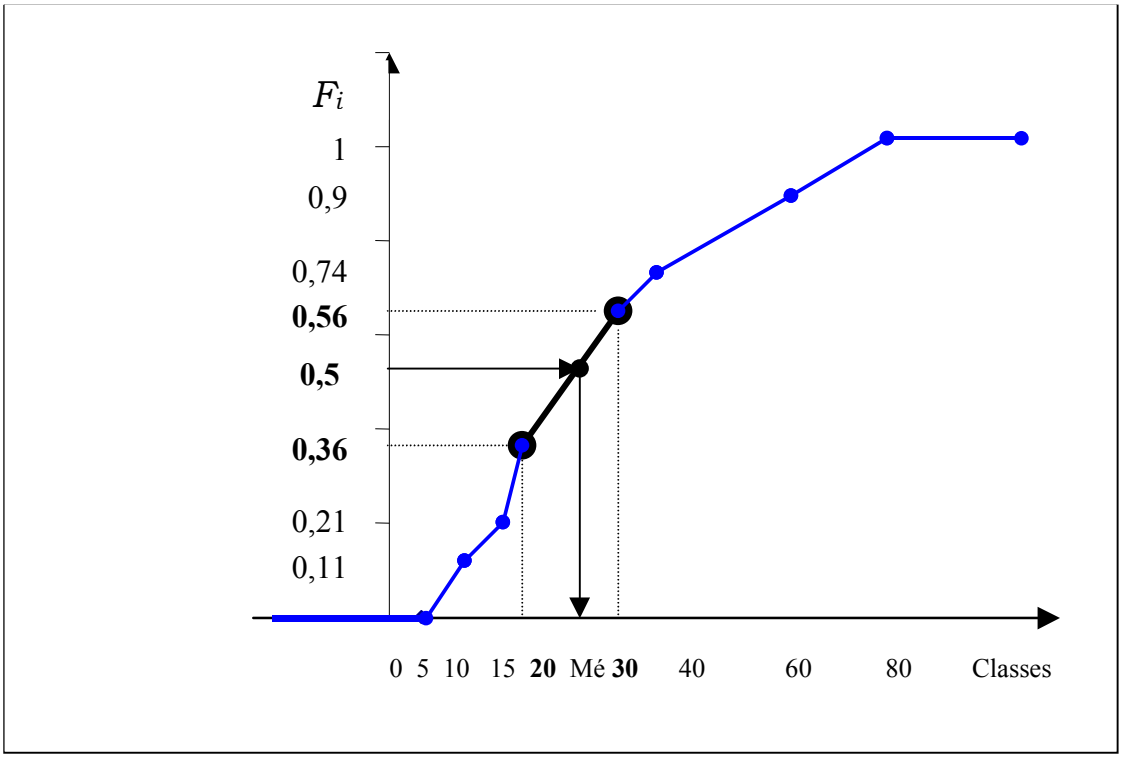
Ou encore, en utilisant les effectifs cumulés croissants :

20 -------- 36

Mé ------- 50

30 --------- 56

$$\frac{Mé-20}{30-20}=\frac{50-36}{56-36}$$



Dans notre exemple : $\frac{N}{2} = 50$. La classe médiane est la classe à laquelle appartient la valeur médiane, c'est à dire la classe $[20-30[$, d'où :

$$Mé = 20 + 10\left(\frac{50-36}{56-36}\right) = 27\text{ans}$$

C'est à dire que 50% des individus sont âgés de moins de 27 ans.

La moyenne arithmétique

La moyenne arithmétique, dite simplement moyenne est notée $\bar{x}$, est la caractéristique de tendance centrale la plus usuelle.

Cas de données non groupées

<u>En entend par données non groupées, celles qui ne sont pas présentées dans un tableau statistique</u>.

Soit une série statistique de N observations : $x_1, x_2, x_3, \cdots, x_n$. La moyenne arithmétique (appelée simple) de ces observations est donnée par :

$$\boxed{\bar{x} = \frac{x_1 + x_2 + x_3 + \cdots + x_n}{N} = \frac{\sum_{i=1}^{i=N} x_i}{N}}$$

Exemple :

On observe les notes en statistique d'un groupe d'étudiants :

14, 16, 12, 9, 11, 16, 7, 9, 7, 9. La moyenne simple de ces notes est :

$$\bar{x} = \frac{14+16+12+9+11+16+7+9+7+9}{10} = 11$$

Cas de données groupées

Dans le cas d'une variable discrète :

$$\boxed{\bar{x} = \frac{n_1x_1 + n_2x_2 + n_3x_3 + \cdots + n_px_p}{N} = \frac{\sum_{i=1}^{i=p} n_ix_i}{N}}$$

où x_i, et n_i, $i = 1, 2, \cdots, p$ représentent respectivement la valeur du caractère et l'effectif correspondant, et p est le nombre de valeurs prises par la variable.

Dans le cas d'une variable continue, où les données sont groupées en classes, on applique la même formule, en remplaçant les valeurs x_i par

les centres de classes. Dans ce cas on a : $\boxed{\bar{x} = \frac{\sum_{i=1}^{i=p} n_i c_i}{N} = \frac{\sum_{i=1}^{i=p} n_i x_i}{N}}$

Exemple : soit la distribution par classe d'âges suivante :

Classes d'âges	**Effectifs** n_i	f_i	**Centre de classe** c_i (*noté aussi*) x_i	$n_i x_i$
[5 , 10[	11	0,11	7,5	82,5
[10 , 15[	10	0,10	12,5	125
[15 , 20[	15	0,15	17,5	262,5
[20 , 30[	20	0,20	25	500
[30 , 40[	18	0,18	35	630
[40 , 60[	16	0,16	50	800
[60 , 80[	10	0,10	70	700
total	100	1		3100

L'âge moyen est donné par : $\bar{x} = \frac{3100}{100} = 31\,ans$

Remarques :

La somme des écarts à la moyenne arithmétique est nulle :

- pour des données non groupées $\sum_{i=i}^{N}(x_i - \bar{x}) = 0$. En effet :

$$\sum_{i=i}^{N}(x_i - \bar{x}) = \sum_{i=1}^{i=N} x_i - \sum_{i=1}^{i=N} \bar{x} = \sum_{i=1}^{i=N} x_i - N.\bar{x} = N.\bar{x} - N.\bar{x} = 0$$

- pour des données groupées en classes, On a $\sum_{i=i}^{N} n_i(c_i - \bar{x}) = 0$.

En effet : $\sum_{i=1}^{i=N} n_i c_i - N.\bar{x} = N.\bar{x} - N.\bar{x} = 0$

La moyenne arithmétique $\bar{x}$ d'une population d'effectif N composée de k sous-populations d'effectifs N_k et de moyenne $\bar{x}_k$ est égale à :

$$\boxed{\bar{x} = \frac{N_1\bar{x}_1 + ...N_p\bar{x}_p}{N} = \frac{\sum_{i=1}^{i=p} N_i\bar{x}_i}{N}}$$

La moyenne arithmétique est le critère le plus fréquemment utilisé pour définir une valeur moyenne d'observations d'une variable additive comme par exemple : la taille, le poids, l'âge, ...etc. Il y a d'autres variables dont le calcul de la moyenne se traite autrement, comme, par exemple, le taux de chômage ou d'inflation, le taux de change, la vitesse sur différents parcours, ...etc.

Autres moyennes

La moyenne géométrique

La moyenne géométrique d'une variable, notée généralement G, est égale à la racine N-ième du produit des N valeurs observées de cette variable. Elle est utilisée souvent dans le calcul des taux de croissance moyens et de certains indices statistiques synthétiques.

Cas de données non groupées

La moyenne géométrique simple est donnée par :

$$\boxed{G = \sqrt[N]{x_1 . x_2 . x_3 . \cdots . x_n}}$$

Cas de données groupées

La moyenne géométrique pondérée est définie par :

$$\boxed{G = \sqrt[N]{x_1^{n_1} . x_2^{n_2} . x_3^{n_3} . \cdots . x_p^{n_p}} = \left(x_1^{n_1} . x_2^{n_2} . x_3^{n_3} . \cdots . x_p^{n_p} \right)^{\frac{1}{N}}}$$

La moyenne géométrique peut être exprimée en fonction des fréquences relatives de la manière suivante :

$$\boxed{G = x_1^{\frac{n_1}{N}} . x_2^{\frac{n_2}{N}} . x_3^{\frac{n_3}{N}} . \cdots . x_p^{\frac{n_p}{N}} = x_1^{f_1} . x_2^{f_2} . x_3^{f_3} . \cdots . x_p^{f_p}}$$

Remarques :

Le logarithme de la moyenne géométrique est égale à la moyenne arithmétique des logarithmes des x_i.

En pratique, le calcul de la moyenne géométrique passe par le logarithme. Ainsi, dans le cas des données non groupées, on a :

$$\boxed{ln\, G = \frac{1}{N} \sum_{i=1}^{N} ln x_i}$$

et dans le cas des données groupées par classes, on a :

$$\boxed{ln\, G = \frac{1}{N} \sum_{i=1}^{p} n_i\, ln x_i}$$

La moyenne géométrique est utilisée quand les valeurs de la variable sont liées de façon multiplicative les unes aux autres.

La moyenne géométrique d'un produit de deux variables est égale au produit de leurs moyennes géométriques.

La moyenne géométrique d'un rapport de deux variables ($\neq 0$) est égale au rapport de leurs moyennes géométriques.

Exemple :

L'étude des bénéfices d'une entreprise sur 5 ans montre que les

bénéfices ont augmenté de 6% pendant les deux premières années, de 10% pendant les deux années suivantes et de 8% pendant la dernière année. Quel est l'augmentation moyenne sur 5 ans ?

En utilisant la moyenne arithmétique des taux observés, et en désignant par $\bar{x}$, le taux moyen ainsi défini, on obtient :

$$\bar{x} = \frac{2 \times 6\% + 2 \times 10\% + 1 \times 8\%}{5} = 8\%$$

Mais ce résultat est un résultat **erroné**. En effet :

Soit F_0 le bénéfice de l'entreprise au début de la période d'étude.

- A la fin de la première année le bénéfice augmente de 6%. Il est égal à $F_1 = F_0 \times 1{,}06$

- A la fin de la deuxième année : $F_2 = F_1 \times 1{,}06 = F_0 \times (1{,}06)^2$

- A la fin de la troisième année : $F_3 = F_2 \times 1{,}10 = F_0 \times (1{,}06)^2 \times 1{,}1$

- A la fin de la quatrième année : $F_4 = F_3 \times 1{,}10 = F_0 \times (1{,}06)^2 \times (1{,}1)^2$

- A la fin de la 5ième année : $F_5 = F_4 \times 1{,}08 = F_0 \times (1{,}06)^2 \times (1{,}1)^2 \times 1{,}08$

Le taux de croissance annuel moyen, c, doit satisfaire la relation :

$$F_5 = F_0 (1+c)^5$$

On peut alors écrire :

$$F_5 = F_0 \times (1{,}06)^2 \times (1{,}1)^2 \times 1{,}08 = F_0 (1+c)^5$$

$$\Rightarrow (1{,}06)^2 \times (1{,}1)^2 \times 1{,}08 = (1+c)^5$$

$$\Leftrightarrow \left((1{,}06)^2 \times (1{,}1)^2 \times 1{,}08\right)^{\frac{1}{5}} = 1 + c$$

(C'est l'écriture de la moyenne géométrique des augmentations)

$$\Leftrightarrow \frac{2}{5}\ln(1{,}06) + \frac{2}{5}\ln(1{,}1) + \frac{1}{5}\ln(1{,}08) = \ln(1+c)$$

$$\Leftrightarrow \frac{2\ln(1{,}06) + 2\ln(1{,}1) + \ln(1{,}08)}{5} = \ln(1+c)$$

$$\Leftrightarrow 0{,}07682 = \ln(1+c)$$

$$\Leftrightarrow e^{0{,}07682} = 1 + c$$

$$\Rightarrow c = 0{,}0798$$

$\ln(1+c)$ apparaît ainsi comme la moyenne arithmétique des logarithmes des taux de croissance. $(1+c)$ est donc la moyenne géométrique des différents taux de croissance du bénéfice.

On peut dire que l'augmentation annuelle moyenne est de $7{,}98\%$

La moyenne harmonique

La moyenne harmonique, notée *H*, est égale à l'inverse de la moyenne arithmétique des inverses des valeurs :

Cas de données non groupées

$$H = \frac{N}{\sum_{i=1}^{i=N}(\frac{1}{x_i})} = \frac{1}{\frac{1}{N}\sum_{i=1}^{i=N}\frac{1}{x_i}}$$

Cas de données groupées

$$H = \frac{1}{\frac{1}{N}\sum_{i=1}^{i=p}\left(\frac{n_i}{x_i}\right)} = \frac{1}{\sum_{i=1}^{i=p}\left(\frac{f_i}{x_i}\right)}$$

Remarques :

L'inverse de la moyenne harmonique est égale à la moyenne arithmétique des inverses des x_i.

La moyenne harmonique est généralement employée lorsque la variable observée est égale au rapport de deux variables exprimées dans deux unités différentes, par exemple le prix d'un bien exprimé en unités monétaires par unité de bien, la vitesse exprimée en unités de distance par unité de temps.

Exemple :

Un étudiant a consacré la même somme de 36 D pendant trois ans à l'achat de livres aux prix respectifs de 4 D, 6 D et 9 D le livre.

Dans ce cas le prix d'achat moyen d'un livre **n'est pas** la moyenne arithmétique des prix : $\bar{x} = \frac{4+6+9}{3} = 6{,}33 \text{ D}$

En effet, l'étudiant a dépensé durant les trois ans $3 \times 36 = 108\text{D}$. Il a acheté :

$\frac{36}{4} = 9$ livres pendant la première année,

$\frac{36}{6} = 6$ durant la deuxième année

et $\frac{36}{9} = 4$ au cours de la troisième année.

Il a donc acheté : $36\left(\frac{1}{4}+\frac{1}{6}+\frac{1}{9}\right)=9+6+4=19$ livres

et le prix moyen d'un livre est donc :

$$H=\frac{3\times 36}{19}=\frac{3\times 36}{36\left(\frac{1}{4}+\frac{1}{6}+\frac{1}{9}\right)}=\frac{3}{\left(\frac{1}{4}+\frac{1}{6}+\frac{1}{9}\right)}=5{,}68\text{D}.$$

H est donc la moyenne harmonique des différents prix 4, 6 et 9.

La moyenne quadratique

La moyenne quadratique d'une variable statistique, notée Q, est égale à la racine carrée de la moyenne arithmétique des carrées des valeurs de la variable.

Cas de données non groupées

$$\boxed{Q=\sqrt{\frac{1}{N}\sum_{i=1}^{i=N}x_i^2}}$$

Cas de données groupées

$$\boxed{Q=\sqrt{\frac{1}{N}\sum_{i=1}^{i=p}n_i x_i^2}=\sqrt{\sum_{i=1}^{i=p}f_i x_i^2}}$$

Exemple :

Quelle est la mesure du « côté moyen » de trois plaques métalliques carrées dont les côtés mesurent 3 cm, 6 cm et 9 cm.

$$\bar{x}=\frac{3+6+9}{3}=6\text{ cm}$$

Le calcul de la moyenne arithmétique des côtés est faux. En effet, les superficies des plaques sont : 9 cm², 36 cm² et 81 cm².

La superficie moyenne est de :

$$\bar{x}=\frac{9+36+81}{3}=42\text{ cm}$$

Ainsi, le côté moyen mesure : $\text{c}=\sqrt{42}$. Il s'agit de la moyenne quadratique des côtés :

$$\text{c}=\sqrt{\frac{1}{3}(3^2+6^2+9^2)}=\sqrt{42}\text{ cm}.$$

Remarques :

La moyenne quadratique est souvent utilisée dans le calcul de la variance

(voir le section suivante).

Les moyennes quadratique et arithmétique tiennent compte davantage des valeurs les plus élevées de la série statistique. En revanche, Les moyennes géométrique et harmonique réduisent l'influence des observations les plus élevées.

Les relations existantes entre les différentes moyennes est :

$$\boxed{x_{\min} \leq H \leq G \leq \bar{x} \leq Q \leq x_{\max}}$$

Exemple :

Calculer les moyennes arithmétique, géométrique, harmonique et quadratique de la série suivante :2, 5, 11, 18.

$$\bar{x} = \frac{2+5+11+18}{4} = 9 \quad ; \quad G = \sqrt[4]{2 \times 5 \times 11 \times 18} = (2 \times 5 \times 11 \times 18)^{\frac{1}{4}} = 6{,}67$$

$$H = \frac{4}{\frac{1}{2}+\frac{1}{5}+\frac{1}{11}+\frac{1}{18}} = 4{,}72 \quad ; \quad Q = \sqrt{\frac{1}{4}(2^2+5^2+11^2+18^2)} = 10{,}88$$

On peut vérifier la relation établie entre les différentes moyennes :

$$2 < 4{,}72 < 6{,}67 < 9 < 10{,}88 < 18$$

Chapitre 3 :
Les paramètres de dispersion et de forme

Chapitre III : Les caractéristiques de dispersion et de forme

Très souvent les indicateurs de tendance centrale (mode, médiane et moyenne) s'avèrent insuffisants pour permettre de résumer à eux seuls et de comparer deux ou plusieurs séries statistiques. Prenons, à titre d'exemple, les deux séries de notes en statistique obtenues par deux groupes d'étudiants :

Groupe I	1	3	4	10	10	16	17	19
Groupe II	8	9	10	10	10	10	11	12

Nous pouvons constater que les deux séries ont un même mode (*Mo*=10), une même médiane (*Mé*=10) et une même moyenne $(\bar{x} = 10)$. Cependant, leur distribution se fait d'une manière nettement différente. En effet, pour le groupe II, les notes ne s'écartent pas trop des valeurs centrales ($Mé = \bar{x} = 10$). Ce qui n'est pas le cas pour le groupe I. D'où la nécessité de calculer d'autres indicateurs capables de rendre compte des écarts entre les différentes valeurs observées et la valeur centrale. Ces indicateurs, qui nous informent sur la variabilité des valeurs observées, sont appelés indicateurs de dispersion.

L'étendue

On appelle étendue d'une série statistique, la différence entre la plus élevée et la plus faible des valeurs observées, soit :

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

L'étendue est un indicateur de dispersion. Il est simple et facile à calculer. Toutefois, il est très sensible aux valeurs extrêmes « aberrantes ».

Les écarts interquantiles

Il s'agit des écarts entre les premiers et les derniers principaux quantiles.

Les quantiles

Comme pour la médiane où l'on s'est intéressé à la valeur de la variable qui partage la population en deux parties d'égal effectif, on s'intéresse ici aux valeurs qui partagent la population en quatre, en dix ou en cent parties de même effectif. Ces valeurs sont appelées respectivement quartiles, déciles et centiles.

D'une manière générale, on appelle quantile d'ordre α, La valeur de la variable x_α telle que $\alpha\%$ des valeurs observées lui sont inférieures.

On peut alors écrire : $F(x_\alpha) = \alpha\%$, où F désigne la fonction de répartition de la variable. La détermination des différents quantiles se fait de la même manière que la médiane (par interpolation linéaire).

Les principaux quantiles sont les quartiles, les déciles et les centiles

Les quartiles

Les quartiles, en nombre de trois notés Q_1, Q_2 et Q_3, sont les valeurs d'une variable, rangées par ordre croissant ou décroissant, qui partagent la population étudiée en quatre parties de même effectif. L'expression des trois quartiles peut être dérivée de la même manière que la médiane.

Soit $[b_i\ b_{i+1}[$ la classe d'amplitude a_i à laquelle appartient Q_1, N_i l'effectif cumulé croissant de cette classe, N_{i-1} l'effectif cumulé croissant de la classe précédant la classe $[b_i\ b_{i+1}[$ et N l'effectif total.

L'expression du premier quartile est donnée par :

$$Q_1 = b_i + a_i\left(\frac{\frac{N}{4} - N_{i-1}}{N_i - N_{i-1}}\right) = b_i + a_i\left(\frac{0{,}25 - F_{i-1}}{F_i - F_{i-1}}\right)$$

Q_1 (premier quartile) : valeur de la variable telle que 25% des observations lui soient inférieures

Si Q_2 appartient à $[b_i\ b_{i+1}[$ alors :

$$Q_2 = b_i + a_i\left(\frac{\frac{N}{2} - N_{i-1}}{N_i - N_{i-1}}\right) = b_i + a_i\left(\frac{0{,}5 - F_{i-1}}{F_i - F_{i-1}}\right) = Mé$$

Q_2 (deuxième quartile) : valeur de la variable telle que 50% des observations lui soient inférieures

Si Q_3 appartient à $[b_i\ b_{i+1}[$ alors :

$$Q_3 = b_i + a_i\left(\frac{\frac{3}{4}N - N_{i-1}}{N_i - N_{i-1}}\right) = b_i + a_i\left(\frac{0{,}75 - F_{i-1}}{F_i - F_{i-1}}\right)$$

Q_3 (troisième quartile) : valeur de la variable telle que 75% des observations lui soient inférieures

Les déciles

Les déciles, en nombre de neuf, notés D_1, D_2,$\cdots$et D_9 sont les valeurs de la variable qui partagent la population en dix sous-populations de même effectif.

De la même manière, on peut définir les expressions des déciles :

$$D_1 = b_i + a_i\left(\frac{\frac{1}{10}N - N_{i-1}}{N_i - N_{i-1}}\right) = b_i + a_i\left(\frac{0{,}10 - F_{i-1}}{F_i - F_{i-1}}\right)$$

$$D_2 = b_i + a_i\left(\frac{\frac{2}{10}N - N_{i-1}}{N_i - N_{i-1}}\right) = b_i + a_i\left(\frac{0{,}20 - F_{i-1}}{F_i - F_{i-1}}\right)$$

$$D_9 = b_i + a_i\left(\frac{\frac{9}{10}N - N_{i-1}}{N_i - N_{i-1}}\right) = b_i + a_i\left(\frac{0{,}90 - F_{i-1}}{F_i - F_{i-1}}\right)$$

D_1 (premier décile) : valeur de la variable telle que 10% des observations lui soient inférieures.

D_2 (deuxième décile) : valeur de la variable telle que 20% des observations lui soient inférieures.

D_9 (neuvième décile) : valeur de la variable telle que 90% des observations lui soient inférieures.

Remarque : $Q_2 = Mé = D_5$

Les centiles

Les centiles, en nombre de 99, notés $C_1, C_2, \cdots,$ et C_{99} et appelés aussi percentiles, sont les valeurs de la variable qui partagent la population en cent sous-populations d'égal effectifs.

On peut définir les centiles de la manière suivante :

$$C_1 = b_i + a_i\left(\frac{\frac{1}{100}N - N_{i-1}}{N_i - N_{i-1}}\right) = b_i + a_i\left(\frac{0{,}01 - F_{i-1}}{F_i - F_{i-1}}\right)$$

$$C_2 = b_i + a_i\left(\frac{\frac{2}{100}N - N_{i-1}}{N_i - N_{i-1}}\right) = b_i + a_i\left(\frac{0{,}02 - F_{i-1}}{F_i - F_{i-1}}\right)$$

$$D_{99} = b_i + a_i\left(\frac{\frac{99}{100}N - N_{i-1}}{N_i - N_{i-1}}\right) = b_i + a_i\left(\frac{0{,}99 - F_{i-1}}{F_i - F_{i-1}}\right)$$

C_1 (premier centile) : valeur de la variable telle que 1% des observations lui soient inférieures.

C_2 (deuxième centile) : valeur de la variable telle que 2% des observations lui soient inférieures.

C_{99} (99ième centile) : valeur de la variable telle que 99% des observations lui soient inférieures.

Remarques :

$Q_2 = Me = D_5 = C_{50}$ et $Q_3 = C_{75}$

Les quartiles, les déciles et les centiles permettent de calculer les différents intervalles interquantiles. La longueur de ces intervalles correspond aux écarts interquantiles qui sont des indicateurs de dispersion. Plus la longueur de l'intervalle est grande, plus la dispersion est forte.

On distingue :

L'intervalle interquartile, qui contient 50% des observations, est :

$$[Q_1 \ , \ Q_3].$$

L'écart interquartile est égal à : $e_Q = Q_3 - Q_1$.

L'intervalle interdécile, qui contient 80% des observations, est :

$$[D_1 \ , \ D_9].$$

L'écart interdécile est égal à : $e_D = D_9 - D_1$.

L'intervalle intercentile, qui contient 98% des observations, est :

$$[C_1 \ , \ C_{99}].$$

L'écart intercentile est égal à : $e_C = C_{99} - C_1$.

Exemple : Soit la répartition de 100 individus par classe d'âges :

Classes d'âges	**Effectifs** n_i	**Effectifs cumulées** $n_i \uparrow$	**Fréquences cumulées croissantes** $F_i \uparrow$
[5 , 10[	11	11	0,11
[10 , 15[	10	21	0,21
[15 , 20[	15	36	0,36
[20 , 30[	20	56	0,56
[30 , 40[	18	74	0,74
[40 , 60[	16	90	0,9
[60 , 80[	10	100	1
total	100		

Calculons les quantiles et les intervalles interquartiles.

$$Q_1 \in [15-20[\Rightarrow Q_1 = 15 + 5\left(\frac{0{,}25-0{,}21}{0{,}36-0{,}21}\right) = 16{,}33 \text{ ans}$$

Ce qui signifie que 25% des individus sont âgés de moins de 16 ans et 4 mois.

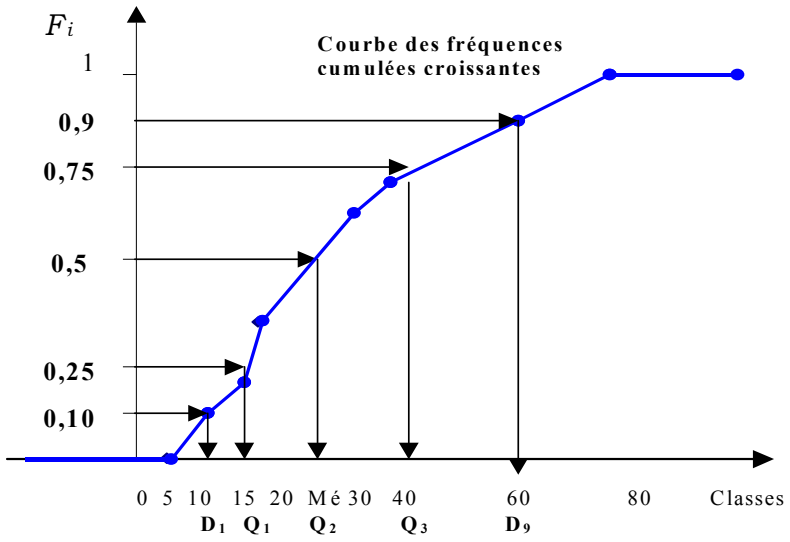
$$Q_3 \in [40-60[\Rightarrow Q_3 = 40 + 20\left(\frac{0{,}75-0{,}74}{0{,}90-0{,}74}\right) = 41{,}25 \text{ ans}$$

Ce qui signifie que 75% des individus sont âgés de moins de 41 ans et 3 mois.

$$D_1 \in [5-10[\Rightarrow D_1 = 5 + 5\left(\frac{0{,}10-0}{0{,}11-0}\right) = 9{,}5 \text{ ans}$$

Ce qui signifie que 10% des individus sont âgés de moins de 9 ans et 6 mois.

En ce qui concerne le neuvième décile, on peut lire sa valeur directement sur le tableau. Il s'agit de la borne supérieure de la classe ayant une fréquence cumulée croissante égale à 0,90 (puisque la valeur



0,9 figure dans la colonne des F_i du tableau). Donc $D_9 = 60$.

Ce qui signifie que 90% des individus sont âgés de moins de 60 ans.

Mesure de la dispersion autour de la moyenne

Exemple 1 :

Considérons les notes suivantes en statistique d'un groupe de 6 étudiants :

2, 17, 7,18, 3, 13.

La moyenne des notes est : $\bar{x} = 10$. Faut-il conclure alors que ce groupe est homogène ? En d'autres termes, faut-il conclure que les étudiants ont le même niveau ? La réponse est non, car 50% des étudiants seulement ont la moyenne.

Pour mesurer cette dispersion autour de la moyenne on peut calculer les différentes distances (écarts) entre la moyenne et les notes observées. On obtient :

$$x_1 - \bar{x} = 2 - 10 = -8,\ \ x_2 - \bar{x} = 17 - 10 = 7,\ \ x_3 - \bar{x} = 7 - 10 = -3$$
$$x_4 - \bar{x} = 18 - 10 = 8,\ \ x_5 - \bar{x} = 3 - 10 = -7,\ \ x_6 - \bar{x} = 13 - 10 = 3$$

Calculons maintenant la moyenne des six distances :

$$\frac{(x_1 - \bar{x}) + (x_2 - \bar{x}) + (x_3 - \bar{x}) + (x_4 - \bar{x}) + (x_5 - \bar{x}) + (x_5 - \bar{x})}{6}$$

$$\frac{\sum_{i=1}^{i=6} x_i - 6\bar{x}}{6} = \frac{1}{6}\sum_{i=1}^{i=6}(x_i - \bar{x}) = \frac{1}{6}(-8 + 7 - 3 + 8 - 7 + 3) = 0$$

Remarque : On a toujours : $\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x}) = 0$.

Ceci traduit le fait que certains étudiants ont des notes supérieures à la moyenne et d'autres ont des notes qui lui sont inférieures (certaines différences sont positives et d'autres sont négatives).

Une première solution consiste à prendre les valeurs absolues de ces écarts et de calculer leur moyenne.

Ecart absolu moyen par rapport à la moyenne

L'écart absolu moyen par rapport à la moyenne, noté $\bar{e}_{\bar{x}}$, d'une série statistique est égal à la moyenne arithmétique de la valeur absolu des écarts entre les valeurs observées et leur moyenne.

Cas de données non groupées

$$\boxed{\bar{e}_{\bar{x}} = \frac{1}{N}\sum_{i=1}^{i=N}|x_i - \bar{x}|}$$

Cas de données groupées

$$\boxed{\bar{e}_{\bar{x}} = \frac{1}{N}\sum_{i=1}^{i=p} n_i|x_i - \bar{x}| = \sum_{i=1}^{i=p} f_i|x_i - \bar{x}|}$$

Cet indicateur de dispersion tient compte de tous les écarts entre chaque valeur observée et la moyenne. Ces écarts sont exprimés dans la

même unité que la variable. Le calcul de l'écart absolue moyen n'est pas commode pour le calcul algébrique (expression de la valeur absolue).

Variance et écart type

Une solution alternative consiste à considérer la moyenne des carrés des différences (dans ce cas toutes les valeurs négatives deviennent positives).

$$\frac{1}{6}\sum_{i=1}^{i=6}(x_i - \bar{x})^2 = \frac{1}{6}\left((-8)^2 + 7^2 + (-3)^2 + 8^2 + (-7)^2 + 3^2\right) = 40{,}66$$

On peut calculer maintenant la racine carrée de la moyenne des carrés des différences pour retrouver la moyenne des écarts par rapport à la moyenne.

$$\sqrt{\frac{1}{6}\sum_{i=1}^{i=6}(x_i - \bar{x})^2} = \sqrt{40{,}66} = 6{,}37$$

Donc, certains étudiants (les bons) auront approximativement la note moyenne (10) plus 6,37, les autres (les mauvais) auront la note moyenne (10) moins 6,37.

On appelle **variance** d'une variable la moyenne des carrés des écarts des valeurs de cette variable à sa moyenne :

Cas de données non groupées

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

Cas de données groupées

$$\boxed{V(x) = \frac{1}{N}\sum_{i=1}^{N} n_i (x_i - \bar{x})^2 = \sum_{i=1}^{N} f_i (x_i - \bar{x})^2 \text{ où } f_i = \frac{n_i}{N}}$$

Remarques :

La variance peut être écrite sous une autre forme dite « formule développée » :

Cas de données non groupées

$$\boxed{V(x) = \left(\frac{1}{N}\sum_{i=1}^{N} x_i^2\right) - \bar{x}^2}$$

Cas de données groupées

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

Cette formule développée de la variance est plus aisée à retenir et plus rapide à calculer.

La variance d'une série statistique correspond à la plus petite des moyennes des carrés des écarts par rapport à une constante k :

$$\boxed{V(x) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2 \leq \frac{1}{N}\sum_{i=1}^{N}(x_i - k)^2, \quad \forall k}.$$

La variance d'une variable y définie par : $y = ax + b$, est :

$$V(y) = a^2 V(x).$$

La variance est exprimée dans le carré de l'unité de la variable. Par exemple, la variance de la variable âge est exprimée en « années au carré (année^2)». C'est la raison pour laquelle on ne doit pas interpréter la variance, mais plutôt sa racine carrée.

On appelle **écart type** que l'on le note par σ_x, La racine carrée de la variance. Il est utilisé comme un indicateur de la dispersion de la série statistique :

$$\boxed{\sigma_x = \sqrt{V(x)}}$$

L'écart type est exprimé dans la même unité de mesure que la variable. Plus l'écart type est grand, plus la dispersion des observations autour de la moyenne de la variable est forte.

Exemple:

Considérons les notes suivantes en statistique d'un groupe de 4 étudiants :

8, 12, 9, 11

Calculer l'écart type des notes et comparer le résultat obtenu avec le résultat de l'exemple 1.

$$\bar{x} = \frac{1}{4}\sum_{i=1}^{4} x_i = \frac{40}{4} = 10$$

$$\begin{aligned} V(x) &= \frac{1}{4}\sum_{i=1}^{4}(x_i - \bar{x})^2 \\ &= \frac{1}{4}\left((8-10)^2 + (12-10)^2 + (9-10)^2 + (11-10)^2\right) \\ &= 2{,}5. \end{aligned}$$

$$\sigma_x = \sqrt{V(x)} = 1{,}58$$

La dispersion des notes dans l'exemple 1 est deux fois plus importante que celle de l'exemple 2. Le second groupe d'étudiant est un groupe plus homogène que le groupe 1.

Exemple 3 :

Soit la répartition de 100 salariés selon leur salaire mensuel :

Salaire en (DT)	Effectifs n_i	Centres de classe x_i	x_i^2	$x_i n_i$	$n_i x_i^2$
[200 - 300[	15	250	62500	3750	937500
[300 - 400[	20	350	122500	7000	2450000
[400 - 600[	35	500	250000	17500	8750000
[600- 700[	15	650	422500	9750	6337500
[700 - 900[	10	800	640000	8000	6400000
[900 - 1100[	5	1000	1000000	5000	5000000
Total	**100**	**-------**		**51000**	**29875000**

Calculer la variance et l'écart type des salaires.

$$\bar{x} = \frac{1}{100}\sum_{i=1}^{6} n_i x_i = \frac{51000}{100} = 510 \text{ Dinars}.$$

$$V(x) = \frac{1}{N}\sum_{i=1}^{k} n_i x_i^2 - \bar{x}^2 = \frac{1}{100} \times 29875000 - 510^2 = 38650 \text{ (Dinars)}^2$$

$$\sigma_x = \sqrt{V(x)} = \sqrt{38650} = 196{,}59 \text{ Dinars}.$$

Variance intra-population et variance inter-populations

On considère une population P de taille N composée de deux sous-populations : P_1 et P_2. L'effectif et la moyenne de chaque sous-population sont :

$$N_1 \text{ , } \bar{x}_1 \text{ pour } P_1$$
$$N_2 \text{ , } \bar{x}_2 \text{ pour } P_2$$

Ou $N = N_1 + N_2$

- *Calculons la moyenne arithmétique de la population P.*

On sait que :

$$\bar{x}_1 = \frac{1}{N_1}\sum_{i=1}^{i=N_1} x_i \Rightarrow N_1 \bar{x}_1 = \sum_{i=1}^{i=N_1} x_i$$

et

$$\bar{x}_2 = \frac{1}{N_2}\sum_{i=1}^{i=N_2} x_i \Rightarrow N_2\bar{x}_2 = \sum_{i=1}^{i=N_2} x_i$$

La moyenne de la population *P* est donnée par :

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{i=N} x_i = \frac{1}{N}\left(\sum_{i=1}^{i=N_1} x_i + \sum_{i=1}^{N_2} x_i\right) = \frac{1}{N}\left(N_1\bar{x}_1 + N_2\bar{x}_2\right)$$

- *Calculons la variance de la population P.*

Soit c une constante, on peut écrire :

$$\frac{1}{N}\sum_{i=1}^{N}(x_i - c)^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - c - \bar{x} + \bar{x})^2 = \frac{1}{N}\sum_{i=1}^{N}\left[(x_i - \bar{x}) + (\bar{x} - c)\right]^2$$

$$= \frac{1}{N}\sum_{i=1}^{i=N}\left[(x_i - \bar{x})^2 + (\bar{x} - c)^2 + 2(x_i - \bar{x})(\bar{x} - c)\right]$$

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

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

Comme $\sum_{i=1}^{i=N}(x_i - \bar{x}) = 0$, on obtient alors :

$$\frac{1}{N}\sum_{i=1}^{N}(x_i - c)^2 = V(x) + (\bar{x} - c)^2$$

$$\boxed{V(x) = \frac{1}{N}\sum_{i=1}^{N}(x_i - c)^2 - (\bar{x} - c)^2, \qquad \forall c}$$

Revenons maintenant au calcul de la variance de la population *P.*

$$V(x) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2 = \frac{1}{N}\sum_{i=1}^{N_1}(x_i - \bar{x})^2 + \frac{1}{N}\sum_{i=1}^{N_2}(x_i - \bar{x})^2$$

Par définition la variance de P_1 est donné par : $V_1(x) = \frac{1}{N_1}\sum_{i=1}^{N_1}(x_i - \bar{x}_1)^2$

et celle de P_2 par : $V_2(x) = \frac{1}{N_2}\sum_{i=1}^{N_2}(x_i - \bar{x}_2)^2$

$$\Rightarrow \begin{cases} N_1V_1(x) = \sum_{i=1}^{N_1}(x_i - \bar{x}_1)^2 \\ N_2V_2(x) = \sum_{i=1}^{N_2}(x_i - \bar{x}_2)^2 \end{cases}$$

En utilisant le résultat précédent et en prenant $c = \bar{x}$ (où $\bar{x}$ est la moyenne de la population *P*), les deux variances $V_1(x)$ et $V_2(x)$ peuvent être exprimées sous la forme suivante :

$$V_1(x)=\frac{1}{N_1}\sum_{i=1}^{N_1}(x_i-\bar{x})^2-(\bar{x}_1-\bar{x})^2 \Rightarrow \frac{1}{N_1}\sum_{i=1}^{N_1}(x_i-\bar{x})^2=V_1(x)+(\bar{x}_1-\bar{x})^2$$

$$\Rightarrow \sum_{i=1}^{N_1}(x_i-\bar{x})^2 = N_1V_1(x)+N_1(\bar{x}_1-\bar{x})^2$$

$$V_2(x)=\frac{1}{N_2}\sum_{i=1}^{N_2}(x_i-\bar{x})^2-(\bar{x}_2-\bar{x})^2 \Rightarrow \frac{1}{N_2}\sum_{i=1}^{N_2}(x_i-\bar{x})^2=V_2(x)+(\bar{x}_2-\bar{x})^2$$

$$\Rightarrow \sum_{i=1}^{N_2}(x_i-\bar{x})^2 = N_2V_2(x)+N_2(\bar{x}_2-\bar{x})^2$$

$$V(x)=\frac{1}{N}\left[N_1V_1(x)+N_1(\bar{x}_1-\bar{x})^2+N_2V_2(x)+N_2(\bar{x}_2-\bar{x})^2\right]$$

$$\boxed{V(x)=\underbrace{\frac{1}{N}\left[N_1V_1(x)+N_2V_2(x)\right]}_{\text{Moyenne des variances}}+\underbrace{\frac{1}{N}\left[N_1(\bar{x}_1-\bar{x})^2+N_2(\bar{x}_2-\bar{x})^2\right]}_{\text{Variance des moyennes}}}$$

La moyenne des variances est notée par : $\overline{V(x)}$.

La variance des moyennes est notée par : $V(\bar{x})$.

La variance totale est décomposée en deux parties :

$$\boxed{V(x)=\overline{V(x)}+V(\bar{x})}$$

- La première composante, $\overline{V(x)}$, nous renseigne sur la dispersion au sein de chaque population. On l'appelle variance intra-population.
- La deuxième composante, $V(\bar{x})$, nous indique la dispersion de la moyenne de chaque sous population par rapport à la moyenne de la population totale. On l'appelle variance inter-populations.

Exemple :

La distribution des salaires dans une entreprise *E*, composée de deux établissements, est la suivante :

Etablissement 1		**Etablissement 2**	
Salaires en 10^2 Dinars	Effectifs n_i	Salaires en 10^2 Dinars	Effectifs n_i
[4 – 8[	40	[8 – 12[	60
[8 – 12[	30	[12 – 20[	50
[12 - 28[	20	[20 - 40[	30

1) Calculer la moyenne des salaires pour l'entreprise E :

2) Calculer la variance totale des salaires dans l'entreprise E.

3) Décomposer la variance totale en variance intra-établissements et inter-établissements. Commenter.

Solution :

Etablissement 1				
Classes	n_i	x_i	$n_i\, x_i$	$n_i\, x_i^2$
[4 – 8[	40	6	240	1440
[8 – 12[	30	10	300	3000
[12 - 28[	20	20	400	8000
Total	**90**		**940**	**12440**

Etablissement 2				
Classes	n_i	x_i	$n_i\, x_i$	$n_i\, x_i^2$
[8 – 12[	60	10	600	6000
[12 – 20[	50	16	800	12800
[20 - 40[	30	30	900	27000
Total	**140**		**2300**	**45800**

1) *Calcul de la moyenne des salaires pour l'entreprise E :*

$$\overline{X} = \frac{1}{N}\sum_{i=1}^{i=6} n_i x_i = \frac{1}{230} \times (940 + 2300) = 14{,}08 \times 10^2 \text{DT}.$$

2) *Calcul de la variance totale des salaires de l'entreprise E :*

$$V(x) = \frac{1}{N}\sum_{i=1}^{i=6} n_i x_i^2 - \bar{x}^2 = \frac{1}{230}(12440 + 45800) \times 10^4 - 14{,}08^2 \times 10^4$$

$$V(x) = 54{,}97 \times 10^6$$

3) *Décomposition de la variance totale des salaires de l'entreprise E :*

La variance totale est donnée par :

$$V(x) = \overline{V(x)} + V(\bar{x})$$

La moyenne des variances est :

$$\overline{V(x)} = \frac{1}{N}\left[N_1 V_1(x) + N_2 V_2(x)\right]$$

- La variance des salaires de l'établissement 1 est :

$$V_1(x) = \frac{1}{90} \times 12440 \times 10^4 - 10{,}444^2 \times 10^4 = 29{,}13 \times 10^4$$

- La variance des salaires de l'établissement 2 est :

$$V_2(x) = \frac{1}{140} \times 45800 \times 10^4 - 16{,}42^2 \times 10^4 = 57{,}26 \times 10^4$$

$$\overline{V(x)} = \frac{1}{230}\left[90 \times 29{,}13 + 140 \times 57{,}26) \times 10^4\right] = 46{,}25 \times 10^2$$

La variance des moyenne est :

$$V(\bar{x}) = \frac{1}{N}\left[N_1(\bar{x}_1 - \bar{x})^2 + N_2(\bar{x}_2 - \bar{x})^2\right] = \frac{1}{N}\sum_{j=1}^{j=2} N_j \bar{x}_j^2 - \bar{x}^2$$

$$V(\bar{x}) = \frac{1}{230}\left[90 \times 10{,}44^2 + 140 \times 16{,}42^2\right] \times 10^4 - 14{,}08^2 \times 10^4 = 8{,}51 \times 10^4$$

$$V(x) = \overline{V(x)} + V(\bar{x}) = (46{,}25 + 8{,}51) \times 10^4 = 54{,}76 \times 10^4$$

La variance des salaires est donc imputable pour une grande partie à la variance intra-établissement des salaires.

Remarque :

Plus généralement, La variance totale $V(x)$ d'une population *P,* de taille *N* composée de *k* sous-populations, $P_1, P_2, \ldots, P_k$, de tailles respectives $N_1, N_2, \ldots, N_k$, de moyennes respectives $\bar{x}_1, \bar{x}_2, \ldots\ldots, \bar{x}_k$, et de variances respectives $V_1(x), V_2(x), \ldots\ldots, V_k(x)$, est donnée par :

$$V(x) = \underbrace{\frac{1}{N}\sum_{k=1}^{K} N_k V_k(x)}_{\text{Variance intra-population}} + \underbrace{\frac{1}{N}\sum_{k=1}^{K} N_k (\bar{x}_k - \bar{x})^2}_{\text{Variance inter-populations}}$$

Avec $N = N_1 + N_2 + \cdots + N_k$, et $\bar{x} = \frac{1}{N}\sum_{k=1}^{K} N_k \bar{x}_k$

Le coefficient de variation

Aussi bien l'écart-type que les indicateurs de tendance centrale (mode, médiane et moyenne) sont exprimés dans la même unité de mesure de la variable. D'autre part, l'écart-type dépend de l'ordre de grandeur des observations de la variable. Ainsi, pour comparer la dispersion de deux ou plusieurs distributions exprimées dans des unités différentes, il est indispensable d'utiliser un indicateur de dispersion indépendant de l'unité de mesure et de l'ordre de grandeur des valeurs observées. Pour ce faire, on utilise **Le coefficient de variation**, qu'on note par : *CV*(*x*), et que l'on définit par :

$$\boxed{CV(x) = \frac{\sigma_x}{\bar{x}}}$$

Remarque : Le coefficient de variation est un nombre pur sans unité. C'est un indicateur de dispersion relatif.

Exemple :

Reprenons l'exemple de la distribution des salaires dans une entreprise *E*, composée de deux établissements :

Etablissement 1		**Etablissement 2**	
Salaires en10^2 Dinars	Effectifs n_i	Salaires en10^2 Dinars	Effectifs n_i
[4 – 8[	40	[8 – 12[	60
[8 – 12[	30	[12 – 20[	50
[12 - 28[	20	[20 - 40[	30

On peut résumer les caractéristiques principales de ces deux établissements dans le tableau suivant

	Etablissement 1	**Etablissement 2**
Moyenne	10,44. 10^2	16,42. 10^2
Variance	29,13. 10^4	57,26. 10^4
Ecart-type	5,39. 10^2	7,56. 10^2
Coefficient de variation	O,51	0,46

- La comparaison directe des écarts-types indique une dispersion des salaires plus forte dans l'établissement 2 que dans l'établissement 1 (7,56. 10^2 contre 5,39. 10^2).
- La comparaison des dispersions à partir du coefficient de variation, indique au contraire une dispersion plus forte (O,51) pour l'établissement 1, que pour l'établissement 2 (0,46).
- En conclusion, on peut dire que les salaires sont plus dispersés dans l'établissement 1 que dans l'établissement 2.

Mesure de la dispersion autour de la médiane

L'écart absolu moyen par rapport à la médiane, noté $\overline{e}_{Mé}$ d'une série statistique est égal à la moyenne arithmétique de la valeur absolue des écarts entre les valeurs observées et leur médiane.

Cas de données non groupées

$$\overline{e}_{Mé} = \frac{1}{N}\sum_{i=1}^{i=N}\left|x_i - Mé\right|$$

Cas de données groupées

$$\overline{e}_{Mé} = \frac{1}{N}\sum_{i=1}^{i=p} n_i\left|x_i - Mé\right| = \sum_{i=1}^{i=p} f_i\left|x_i - Mé\right|$$

Cet indicateur de dispersion tient compte de tous les écarts entre chaque valeur observée et la médiane. Ces écarts sont exprimés dans la

même unité que la variable.

Remarque :

Pour toute série statistique on a :

$$\boxed{\overline{e}_{Mé} \leq \overline{e}_{\bar{x}} \leq \sigma_x}$$

Moments d'une série statistique

Moments non centrés

Cas de données non groupées

Le moment non centré d'ordre *r*, qu'on note $m_r(x)$, d'une série statistique est :

$$\boxed{m_r(x) = \frac{1}{N}\sum_{i=1}^{i=N} x_i^r}$$

ii) Cas de données groupées

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

Remarque :

Le moment non centré d'ordre 1 est : $m_1 = \bar{x}$

Le moment non centré d'ordre 2 est : $m_2 = \overline{(x^2)}$

Moments centrés

Cas de données non groupées

Le moment centré d'ordre *r* , qu'on note $\mu_r(x)$, d'une série statistique est :

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

Cas de données groupées

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

Remarque :

Le moment centré d'ordre 1 est : $\mu_1 = 0$

Le moment centré d'ordre 2 est : $\mu_2 = V(x)$

A partir de la formule développée de la variance, on a :

$$\mu_2 = V(x) = m_2 - (m_1)^2$$

En général, les moments centrés d'ordre pair donnent une indication sur la dispersion des observations autour de la moyenne. Les moments centrés d'ordre impair donnent une indication sur le degré de symétrie de la distribution.

Exemple : La répartition de 100 individus par classes d'âges est donnée par le tableau suivant :

Classes d'âges	n_i	f_i	**Centres** x_i	$n_i x_i$	$n_i x_i^2$	$n_i \lvert x_i - \bar{x} \rvert$	$n_i \lvert x_i - Mé \rvert$
[5 , 10[	11	0,11	7,5	82,5	618,75	258,5	214,5
[10 , 15[	10	0,10	12,5	125	1562,5	185	145
[15 , 20[	15	0,15	17,5	262,5	4593,75	202,5	142,5
[20 , 30[	20	0,20	25	500	12500	120	40
[30 , 40[	18	0,18	35	630	22050	72	144
[40 , 60[	16	0,16	50	800	40000	304	368
[60 , 80[	10	0,10	70	700	49000	390	430
total	100	1		3100	130325	1532	1484

$\bar{x} = \dfrac{3100}{100} = 31$ ans, $Mé = 27$ ans, $\bar{e}_{\bar{x}} = 15{,}32$ ans, $\bar{e}_{Mé} = 14{,}84$ ans

$V(x) = 1303{,}25 - (31)^2 = 342{,}25$ (années)^2

$\sigma_x = \sqrt{342{,}25} = 18{,}5$ ans

On remarque bien que : $\boxed{\bar{e}_{Mé} \leq \bar{e}_{\bar{x}} \leq \sigma_x}$

Indicateurs de forme

Les polygones des fréquences nous livrent une représentation approximative de la distribution réelle des fréquences. Pour avoir une idée satisfaisante et plus précise sur la forme de la distribution, il est recommandé de calculer des indicateurs de forme. On distingue les indicateurs d'asymétrie et les indicateurs d'aplatissement. Ces indicateurs sont sans unité de mesure. Ils sont indépendants d'un changement d'échelle et/ou d'origine.

Asymétrie

Une distribution est dite symétrique si les observations se répartissent dans la même proportion de part et d'autre des trois valeurs centrales (mode, médiane et moyenne).

Les mesures d'asymétrie permettent de quantifier le degré de déviation de la forme de distribution par rapport à une distribution symétrique.

i) Le coefficient d'asymétrie de Fisher, qu'on note par γ_1 :

$$\gamma_1 = \frac{\mu_3}{\sigma^3} = \frac{\text{moment centré d'ordre 3}}{(\text{écart - type})^3}$$

La distribution est dite symétrique dans le cas où $\gamma_1 = 0$.

La distribution est dite étalée à gauche dans le cas où $\gamma_1 < 0$.

La distribution est dite étalée à droite dans le cas où $\gamma_1 > 0$.

ii) Le coefficient d'asymétrie de Yule, basé sur les quartiles, qu'on note par C_Y :

$$C_Y = \frac{(Q_3 - Q_2) - (Q_2 - Q_1)}{(Q_3 - Q_1)}$$

La distribution est dite symétrique dans le cas où $C_Y = 0$.

La distribution est dite étalée à gauche dans le cas où $C_Y < 0$.

La distribution est dite étalée à droite dans le cas où $C_Y > 0$.

iii) Le coefficient d'asymétrie de Pearson, basé sur la moyenne, le mode et l'écart-type, qu'on note par C_P :

$$C_P = \frac{\bar{x} - Mo}{\sigma_x}$$

La distribution est dite symétrique dans le cas où $C_P = 0$.

La distribution est dite étalée à gauche dans le cas où $C_P < 0$.

La distribution est dite étalée à droite dans le cas où $C_P > 0$.

Aplatissement

Une distribution est d'autant plus « plate » que la dispersion des observations autour des valeurs centrales est forte.

i) Le coefficient d'aplatissement de Pearson, qu'on note par β :

$$\beta = \frac{\mu_4}{\sigma^4} = \frac{\text{moment centré d'ordre 4}}{(\text{écart - type})^4}$$

La distribution est dite normale dans le cas où $\beta = 3$.

La distribution est dite hyponormale (plus aplatie que la normale) dans le cas où $\beta < 3$.

La distribution est dite hypernormale (moins aplatie que la normale) dans le cas où $\beta > 3$.

ii) Le coefficient d'aplatissement de Fisher, qu'on note par γ_2 :

$$\gamma_2 = \frac{\mu_4}{\sigma^4} - 3 = \beta - 3$$

La distribution est dite normale dans le cas où $\gamma_2 = 0$.

La distribution est dite hyponormale (plus aplatie que la normale) dans le cas où $\gamma_2 < 0$.

La distribution est dite hypernormale (moins aplatie que la normale) dans le cas où $\gamma_2 > 0$.

Chapitre 4 :
Concentration d'une série statistique

Chapitre IV : Concentration d'une série statistique

L'étude de concentration a pour objet de mesurer et de mettre en exergue d'éventuelles inégalités de répartition d'une valeur globale totale. Cette étude n'est pas centrée sur l'individu, elle est plutôt globale. L'analyse porte davantage sur la répartition de la masse totale. Elle permet de compléter l'analyse de la dispersion relative d'une distribution. Les domaines d'applications sont nombreux : concentration des salaires, des revenus, des superficies agricoles, …etc.

Le concept de concentration a été élaboré dans les années 1910-1914 par le statisticien italien Corrado Gini (1884-1965).

L'étude de la concentration porte sur toute série positive La notion de concentration ne s'applique qu'à des variables quantitatives continues à valeurs positives cumulables, celles où le cumul a un sens.

La question fondamentale, à laquelle on doit répondre est, par exemple :

La masse salariale totale est-elle répartie d'une manière égalitaire ? Dans le cas où elle s'est faite d'une manière inégalitaire, on observe un faible nombre d'individus détenir une grande partie de cette masse, la partie restante étant détenue par un grand nombre d'individus.

Valeurs globales et valeurs globales relatives

Soit X une variable statistique continue. On considère la série statistique correspondante.

On appelle **valeur globale** associée au couple (x_i, n_i), le produit défini par :

$$VG_i = n_i x_i$$

On appelle **valeur globale totale**, qu'on note VGT :

$$VGT = \sum_{i=1}^{p} n_i x_i$$

On appelle **valeur globale relative** associée au couple (x_i, n_i), le rapport, qu'on note q_i, défini par :

$$q_i = \frac{x_i n_i}{\sum_{i=1}^{p} x_i n_i}$$

On appelle **valeur globale relative cumulée croissante** associée la valeur x_i, centre de la classe $[b_{i-1}, b_i[$, qu'on note Q_i, :

$$Q_i = \sum_{j=1}^{j=i} q_j$$

Exemple :

Superficie en ha	*Centres* x_i	*Effectifs* n_i	Valeurs globales $n_i x_i$	Valeurs globales relatives $q_i = \frac{n_i x_i}{\sum n_i x_i}$	Valeurs globales relatives cumulées croissantes Q_i
[1 – 5[	3	11	33	0,008	0,008
[5 – 10[	7,5	12	90	0,022	0,030
[10 – 20[	15	15	225	0,056	0,086
[20 – **50**[	35	26	910	0,229	**0,315**
[50 – 100[	75	36	2700	0,685	1
Total		100	**3958**	1	

On peut interpréter la cinquième ligne en disant que les exploitation qui ont moins de 50 ha se partagent 31,5% de la superficie totale qui est égale à 3958 ha.

Médiale

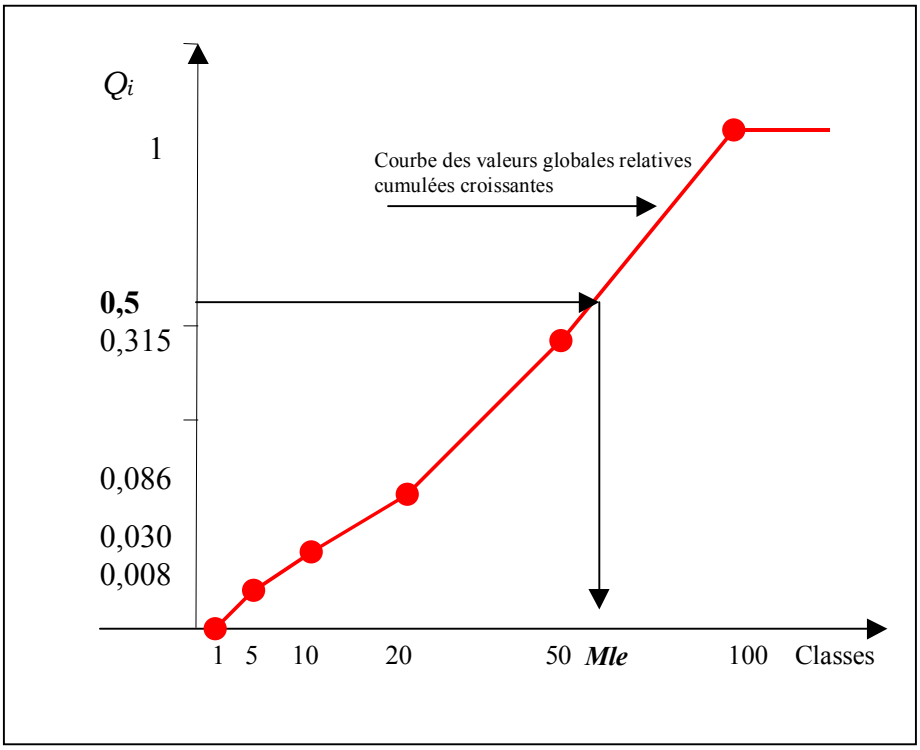
Définition

On appelle médiale d'une série statistique, qu'on note par Mle, la valeur de la variable telle que :

$$Q(Mle) = 0{,}5 = 50\%$$

Détermination graphique

La médiale est déterminée graphiquement comme étant l'abscisse du point d'ordonnée 0,5 de la courbe des valeurs globales relatives cumulées croissantes. Cette courbe est définie en tant qu'une ligne brisée obtenue sur un repère cartésien, en joignant les points de coordonnées $(b_i\ ,Q_i)$, où b_i désigne la borne supérieure de la classe $[b_{i-1}\ , b_i[$ et Q_i la valeur globale cumulée croissante.



Calcul de la médiale

La médiale se détermine, par interpolation linéaire, de la même manière que la médiane. Seulement, les calculs ne se font plus sur les fréquences cumulées croissantes de la série statistique, mais sur les valeurs globales relatives cumulées croissantes.

Le calcul de la médiale passe d'abord par la détermination de la classe médiale. Dans un deuxième temps, on détermine la valeur précise de la médiale par interpolation linéaire.

Soit $[b_{i-1}–b_i[$ la classe médiale, a_i l'amplitude de la classe médiane, Q_i la valeur globale relative cumulée croissante de la classe médiale, Q_{i-1} la valeur globale relative cumulée croissante de la classe qui précède la classe médiale.

L'expression de la médiale est alors donnée par :

$$Mle = b_{i-1} + a_i\left(\frac{0{,}5 - Q_{i-1}}{Q_i - Q_{i-1}}\right)$$

Dans notre exemple, la classe médiale à laquelle appartient la médiale, est la classe $[50-100[$ d'où :

Superficie en ha	*Amplitudes* a_i	Valeurs globales $n_i x_i$	Valeurs globales relatives $q_i = \frac{n_i x_i}{\sum n_i x_i}$	Valeurs globales relatives cumulées croissantes Q_i
[1 – 5[	4	33	0,008	0,008
[5 – 10[	5	90	0,022	0,030
[10 – 20[	10	225	0,056	0,086
[20 , 50[→ *Mle* [50 , 100 [	30 **50**	910 2700	0,229 0,685	**0,315** $\frac{1}{2} = 0,5$ **1**
Total		**3958**	1	

Le calcul de la médiale par interpolation linéaire donne :

$$\frac{Mle - 50}{100 - 50} = \left(\frac{0,50 - 0,315}{1 - 0,315} \right)$$

$$Mle = 50 + 50\left(\frac{0,50 - 0,315}{1 - 0,315} \right) = 63,5\text{ha}$$

On interprète en disant que les exploitations qui ont individuellement moins de 63,5 ha totalisent 50% de la superficie totale.

Ecart médiale- médiane

On appelle écart médiale-médiane d'une série statistique, qu'on note par ΔM , le nombre défini par :

$$\boxed{\Delta M = Mle - Mé}$$

Cet écart nous fournit un premier renseignement sur la concentration d'une distribution statistique. Son interprétation se fait par rapport à l'étendue de la série. En d'autres termes, on calcule :

$$\boxed{\frac{\Delta M}{\text{Intervalle de variation}}}$$

- Si ΔM est grand par rapport à l'intervalle de variation, alors la concentration est forte
- Si ΔM est petit par rapport à l'intervalle de variation, alors la concentration est faible
- S'il y a absence de concentration ou situation d'équipartition parfaite, alors ΔM est nul.

Dans notre exemple, la médiane est égale à :

$$Mé = 20 + 30\left(\frac{0{,}50 - 0{,}38}{0{,}64 - 0{,}38}\right) = 33{,}84\text{ha}$$

Donc, l'écart médiale-médiane est :

$$\boxed{\Delta M = 63{,}50 - 33{,}84 = 29{,}66}$$

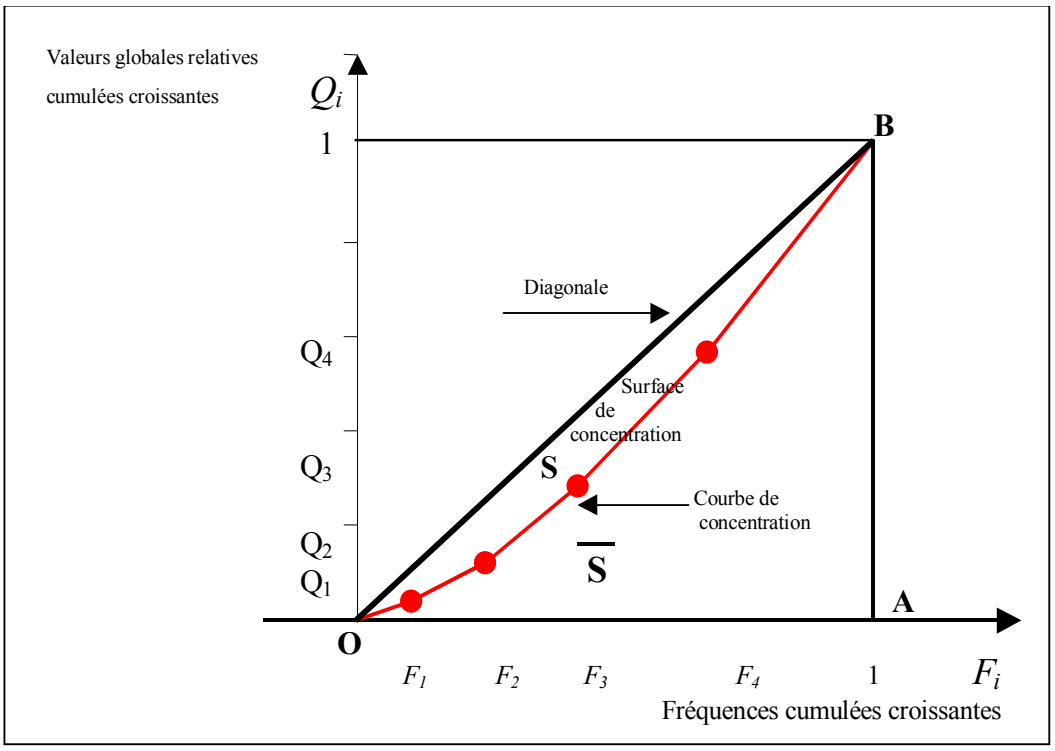
Par conséquent, l'écart médiale-médiane relatif est : $\boxed{\frac{29{,}66}{100 - 1} = 0{,}29}$

On peut dire que la concentration est relativement moyenne.

Courbe de concentration

Définition

On appelle courbe de concentration (ou courbe de Lorenz), Le polygone obtenu en joignent, les points de coordonnées (F_i, Q_i), dans un repère orthonormé, où les F_i sont portés sur l'axe des abscisses et les Q_i sur l'axes des ordonnées. Cette représentation se fait dans un carré de côte égal à l'unité.



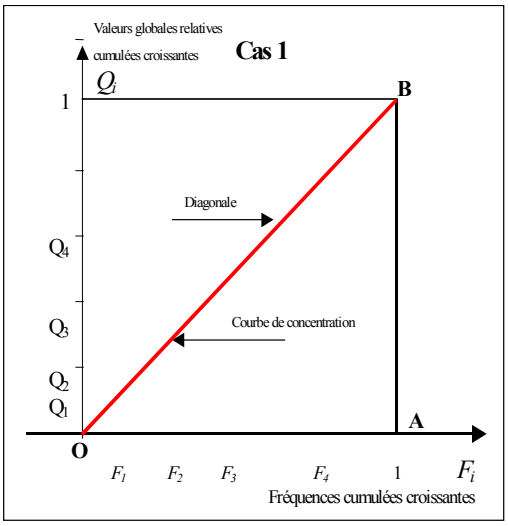
Interprétation

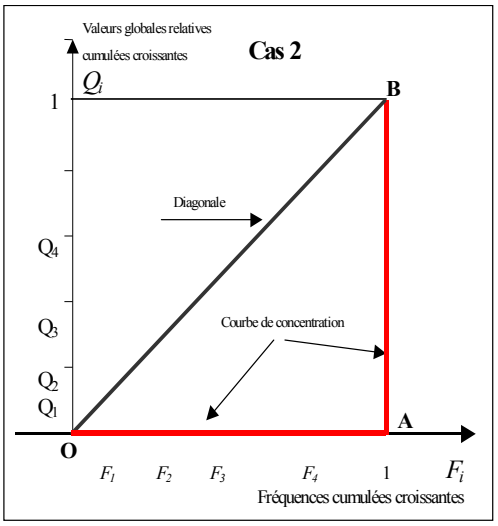
Plus la courbe de concentration se rapproche de la diagonale, plus la répartition est égalitaire, et plus la courbe s'éloigne de la diagonale, plus la distribution est concentrée, c'est-à-dire inégalement répartie.

Cas extrêmes

Cas 1. La courbe de concentration se confond avec la diagonale. C'est le cas d'une équirépartition parfaite. On dit aussi que la concentration est nulle.

Cas 2. La courbe de concentration se confond avec les côtés OA et AB du triangle OAB. C'est le cas, hypothétique, où un seul individu possède toute la richesse. On dit aussi que la série est totalement concentrée.





Indice de concentration de Gini

Surface de concentration

On appelle surface de concentration, qu'on note par S, la surface comprise entre la diagonale principale OB et la courbe de concentration. Plus la courbe s'éloigne de la diagonale et plus la surface de concentration est grande.

Remarque :

La courbe de concentration se situe toujours en dessous de la diagonale car on a, pour toute valeur de x : $F(x) \geq Q(x)$

Définition de l'indice de Gini

On appelle indice de Gini (ou indice de concentration), le rapport entre l'aire de la surface de concentration et l'aire du triangle OAB. On le note par I_G. :

$$I_G = \frac{\text{Aire de la surface de concentration}}{\text{Aire du triangle OAB}} = \frac{S}{\frac{1}{2}} = 2S.$$

Remarque :

L'indice de Gini est compris entre [0 , 1]

Dans le **Cas 1,** où la courbe de concentration se confond avec la diagonale, l'indice de Gini est égal à **zéro**.

Dans le **Cas 2**, où la courbe de concentration se confond avec les côtés OA et AB du triangle OAB, l'indice est égal à **un**.

Plus l'indice de Gini tend vers 1, plus la concentration est forte.

Plus l'indice de Gini tend vers 0, plus la concentration est faible

Calcul de l'indice de Gini

Pour le calcul de l'indice de Gini, on retient la méthode des trapèzes. Celle-ci consiste à calculer l'aire de la surface complémentaire à S par rapport à l'aire du triangle OAB. Pour ce faire, il suffit de créer une nouvelle colonne $f_i(Q_i + Q_{i-1})$.

L'indice est alors égal à :

$$\boxed{I_G = 1 - \sum_{i=1}^{p} f_i(Q_i + Q_{i-1})}$$

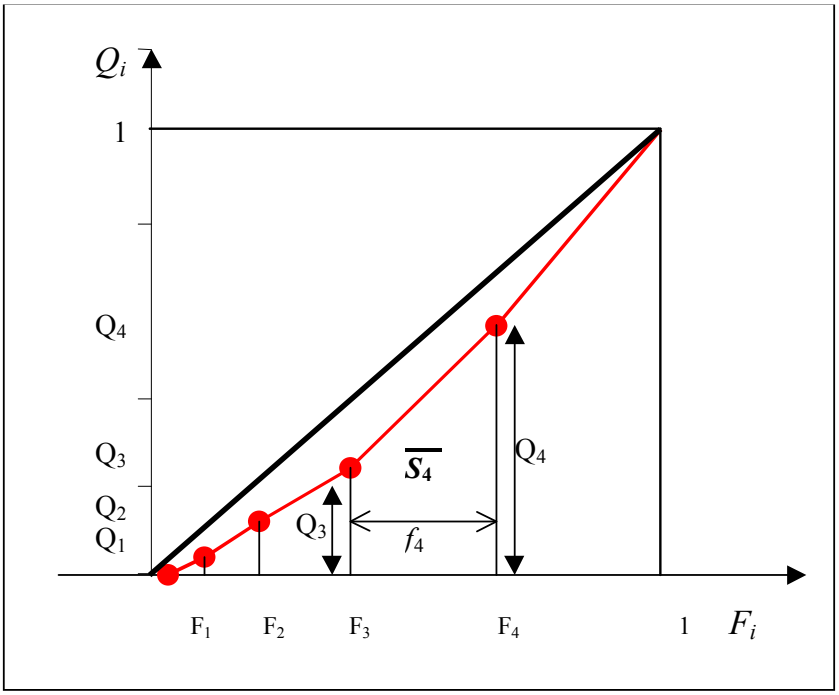
La surface de chaque trapèze est :

$$\overline{S}_i = \frac{(\text{grande base} + \text{petite base}) \times \text{hauteur}}{2}$$

$$= \frac{(Q_{i-1} + Q_i) \times (F_i - F_{i-1})}{2} = \frac{(Q_{i-1} + Q_i) \times (f_i)}{2}$$

Par exemple :

$$\overline{S}_4 = \frac{(Q_3 + Q_4) \times (F_4 - F_3)}{2} = \frac{(f_4) \times (Q_3 + Q_4)}{2}$$



Superficie en ha	x_i	n_i	f_i	F_i	$n_i x_i$ (VG_i)	$q_i = \frac{n_i x_i}{\sum n_i x_i}$	Q_i	$f_i(Q_i + Q_{i-1})$
[1 – 5[	3	11	0,11	0,11	33	0,008	0,008	0,00088
[5 – 10[	7,5	12	0,12	0,23	90	0,022	0,030	0,00456
[10 – 20[	15	15	0,15	0,38	225	0,056	0,086	0,0174
[20 – 50[	35	26	0,26	0,64	910	0,229	0,315	0,10426
[50 – 100[	75	36	0,36	1	2700	0,685	1	0,4734
Total		100	1		3958	1		0,6005

En reprenant notre exemple concernant la répartition des exploitations agricoles, l'indice de Gini est égal à :

$$I_G = 1 - 0{,}6005 = 0{,}3995 \approx 0{,}4 .$$

Cette valeur indique que la concentration est relativement moyenne.

Chapitre 5 :
Les indices statistiques

Chapitre V : Les indices statistiques

L'analyse économique et sociale fait appel à la comparaison et l'étude de l'évolution de grandeurs simples, telles que la production de blé, le prix de pétrole, ….etc. La comparaison peut se faire dans le temps ou dans l'espace, moyennant le calcul d'un rapport entre deux valeurs de cette grandeur, prises par conséquent en deux périodes différentes ou dans deux lieux différents.

Il est aussi important de pouvoir suivre l'évolution de grandeurs complexes telles que la production agricole, les exportations d'un pays, …etc. Ces comparaisons se font au moyen d'indices synthétiques.

Ainsi, on distingue deux types d'indices : *L'indice élémentaire* et *l'indice synthétique.*

Les Indices élémentaires

L'indice élémentaire permet de calculer l'évolution d'une grandeur simple (comme, par exemple, le prix ou la production d'un bien donné), soit dans le temps, auquel cas on appelle cet indice un indice élémentaire temporel, soit entre deux lieux géographiques différents, auquel cas on parle d'indice élémentaire spatial.

Définition

On peut définir l'indice élémentaire temporel ou spatial comme un nombre pur (sans dimension) résultant du rapport de deux valeurs prises par la même grandeur, soit à deux dates différentes, soit sur deux espaces différents.

Soit x_1 la valeur de la grandeur G à la date $t = 1$ et x_0 la valeur de la variable à la date $t = 0$.

L'indice élémentaire de la grandeur G est donné par :

$$\boxed{I_{1/0} = \frac{x_1}{x_0} \times 100}$$

La date $t = 1$ est appelée date courante ou période courante, ou encore situation courante, dans le cas d'un indice spatial. La date $t = 0$ est dite date de référence, ou période de base, ou encore situation de base, dans le cas d'un indice spatial.

Exemple 1:

Le prix d'un billet d'avion Tunis - Toulouse est passé de 310 D en 1985 à 400 D en 1998.

L'indice de prix dans ce cas est donné par :

$$I_{98/85} = \frac{P_{98}}{P_{85}} \times 100 = \frac{400}{310} \times 100 = 107{,}75$$

On dit que le prix d'un billet d'avion a augmenté de (107,75-100), soit 7,75% entre 1995 et 1998.

Exemple 2:

Le loyer d'un studio à Tunis est de 240 D, alors qu'à Bizerte il est de 120 D.

Dans ce cas l'indice de loyer entre Bizerte et Tunis est de :

$$I_{\text{Bizerte/Tunis}} = \frac{P_{\text{Bizerte}}}{P_{\text{Tunis}}} \times 100 = \frac{120}{240} \times 100 = 50\%$$

Donc le loyer d'un studio à Bizerte est le moitié de celui à Tunis.

Les propriétés de l'indice élémentaire :

La circularité ou transitivité

Cette propriété est intéressante dans le cas d'un changement de l'année de base.

Si une grandeur économique prend les valeurs x_0, x_1 et x_2 respectivement aux dates $t = 0, 1 \text{ et } 2$, l'indice élémentaire satisfait :

$$\boxed{I_{2/0} = \frac{1}{100} \times I_{2/1} \times I_{1/0}}$$

Ainsi on a :

$$\boxed{\underbrace{I_{2/1}}_{\text{base 1}} = \frac{I_{2/0}}{\underbrace{I_{1/0}}_{\text{base 0}}} \times 100}$$

Démonstration :

$$I_{2/0} = \frac{x_2}{x_0} \times 100 = \frac{x_2}{x_0} \times 100 \times \frac{100 \times x_1}{100 \times x_1} = \frac{1}{100} \times \frac{x_2}{x_1} \times 100 \frac{x_1}{x_0} \times 100$$

$$I_{2/0} = \frac{1}{100} \times I_{2/1} \times I_{1/0}$$

D'une manière générale :

$$\boxed{I_{t/0} = 100 \times \left(\frac{I_{t/t-1}}{100} \times \frac{I_{t-1/t-2}}{100} \times \frac{I_{t-2/t-3}}{100} \times \cdots \times \frac{I_{1/0}}{100} \right)}$$

Exemple :

Le prix d'un bien *Z* pour trois dates est donné dans le tableau suivant :

Date	Prix
1985	150
1990	210
1995	230

Calculons les différents indices élémentaires:

$$I_{90/85} = \frac{210}{150} \times 100 = 140\%, \quad I_{95/85} = \frac{230}{150} \times 100 = 153{,}3$$

$$I_{95/90} = \frac{230}{210} \times 100 = 109{,}5$$

On vérifie que :

$$I_{95/85} = \frac{1}{100} \times I_{95/90} \times I_{90/85} = \frac{1}{100} \times 109{,}5 \times 140 = 153{,}3\%$$

Ainsi, pour comparer deux variables entre deux dates, il suffit de faire le rapport de leur indice.

La réversibilité :

Cette propriété est prenante dans le cas du calcul d'indice spatial car le choix de l'espace de référence est arbitraire.

La propriété de la réversibilité peut être présentée sous la forme suivante :

$$\boxed{I_{1/0} \times I_{0/1} = 10^4} \text{ ou encore } \boxed{I_{0/1} = \frac{10^4}{I_{1/0}}}$$

Démonstration :

$$100 \times 100 = \frac{x_1}{x_0} \times \frac{x_0}{x_1} \times 100 \times 100 \Rightarrow 10^4 = I_{1/0} \times I_{0/1}$$

Exemple :

En reprenant les données de l'exemple précédent, on peut vérifier que :.

$$I_{90/85} = \frac{210}{150} \times 100 = 140\%, \quad I_{85/90} = \frac{150}{210} \times 100 = 71{,}428\%$$

$$I_{90/85} \times I_{85/90} = 140 \times 71{,}428 \approx 10^4$$

Autres propriétés de l'indice élémentaire :

Si $a = bc$ alors l'indice élémentaire de a est donné par :

$$\boxed{I_{1/0}(a) = I_{1/0}(b).\, I_{1/0}(c).\frac{1}{100}}$$

Exemple :

Supposons que $I_{1/0}(p) = 110\%$ et $I_{1/0}(q) = 120\%$.

La recette étant égale au produit du prix par la quantité, $R = pq$, l'indice élémentaire de la recette est :

$$I_{1/0}(R) = I_{1/0}(p) \times I_{1/0}(q) \times \frac{1}{100} = 110 \times 120 \times \frac{1}{100} = 132\%$$

Entre la date 0 et la date 1 la recette a augmenté de 32%

Si $a = \frac{b}{c}$, alors l'indice élémentaire de a est donné par :

$$\boxed{I_{1/0}(a) = \frac{I_{1/0}(b)}{I_{1/0}(c)} \times 100}$$

Les indices synthétiques

Soit G une grandeur complexe composée de plusieurs autres grandeurs simples :

$$G = \left\{g^1, g^2, \cdots, g^k\right\}$$

Pour chaque grandeur simple $g^i, i = 1, 2, \cdots, k$ on peut calculer un indice élémentaire simple :

$$I_{t/0}(g^i) = \frac{g_t^i}{g_0^i} \times 100, \quad i = 1, 2, \cdots, k$$

On peut résumer cette série d'indices élémentaires par un indice synthétique noté $I_{t/0}(G)$.

En économie on s'intéresse souvent aux variations des prix, des quantités et de la valeur globale (prix fois quantités). Ainsi, on peut calculer trois indices synthétiques, à savoir l'indice des prix, l'indice des quantités et l'indice de valeur globale.

Soient p_0^i, q_0^i respectivement le prix et la quantité du bien i à la date 0, et p_t^i, q_t^i respectivement le prix et la quantité du même bien à la date t. Considérons un panier composé de k biens.

Les valeurs globales de ce panier évaluées à la date 0 et à la date 1 sont données respectivement par :

$$V_0 = \sum_{i=1}^{i=k} p_0^i q_0^i \text{ et } V_t = \sum_{i=1}^{i=k} p_t^i q_t^i$$

L'indice de la valeur globale est donné par :

$$I_{t/0}(V) = I_{t/0}(p.q) = \frac{V_t}{V_0} \times 100 = \frac{\sum_{i=1}^{i=k} p_t^i q_t^i}{\sum_{i=1}^{i=k} p_0^i q_0^i} \times 100$$

Par exemple, $I_{t/0}(V) = 130\%$ signifie que la valeur du panier a augmenté de 30% entre la date 0 et la date *t*. A ce niveau, une question importante se pose : quelle est l'origine de cette augmentation ? résulte-elle de l'augmentation des prix, des quantité ou des deux ?

En effet, dans ce cas, plusieurs cas de figures peuvent se présenter :

les prix augmentent et les quantités restent constantes.

les quantités augmentent et les prix restent constants.

les prix augmentent et les quantités baissent, mais la hausse des prix l'emporte sur la baisse des quantités.

les quantités augmentent et les prix baissent, mais la hausse des quantités l'emporte sur la baisse des prix.

les quantités et les prix augmentent simultanément.

Afin de cerner avec précision les origines de la variation, on fixe les quantités et on calcule un indice de prix, ensuite on fixe les prix et on calcule un indice de quantités

Généralement, on distingue deux types d'indices selon que l'on fixe les quantités ou les prix à la date de base 0 ou à la date courante *t*. Dans le premier cas, lorsque l'on fixe les prix ou les quantités à la date de base 0 : on calcule les indices synthétiques de Laspeyres. Dans le deuxième cas, lorsque l'on fixe les prix ou les quantités à la date courante *t*, on calcule les indices synthétiques de Paasche.

Indices synthétiques de Laspeyres

Indice de prix de Laspeyres:

Cet indice indique l'évolution de la valeur d'un panier de biens à composition constante. Les quantités fixes sont évaluées à la date de base 0 :

$$L_{t/0}^{P} = \frac{\sum_{i=1}^{i=k} p_t^i q_0^i}{\sum_{i=1}^{i=k} p_0^i q_0^i} \times 100$$

Indice de quantité de Laspeyres:

Cet indice indique l'évolution de la valeur d'un panier de biens à prix constants. Les prix constants sont évalués à la date de base 0 :

$$L^{q}_{t/0} = \frac{\sum_{i=1}^{i=k} p^{i}_{0} q^{i}_{t}}{\sum_{i=1}^{i=k} p^{i}_{0} q^{i}_{0}} \times 100$$

Indices synthétiques de Paasche

Indice de prix de Paasche :

Cet indice indique l'évolution de la valeur d'un panier de biens à composition constante. Les quantités fixes sont évaluées à la date courante *t*.

$$P^{P}_{t/0} = \frac{\sum_{i=1}^{i=k} p^{i}_{t} q^{i}_{t}}{\sum_{i=1}^{i=k} p^{i}_{0} q^{i}_{t}} \times 100$$

Indice de quantité de Paasche :

Cet indice indique l'évolution de la valeur d'un panier de biens à prix constants. Les prix constants sont évalués à la date courante *t* :

$$P^{q}_{t/0} = \frac{\sum_{i=1}^{i=k} p^{i}_{t} q^{i}_{t}}{\sum_{i=1}^{i=k} p^{i}_{t} q^{i}_{0}} \times 100$$

Les coefficients budgétaires

On appelle coefficient budgétaire associé au bien *i* , la part de la dépense consacrée à ce bien. Ainsi, les coefficients budgétaires d'un bien *i,* respectivement à la date 0 et à la date t sont :

$$W^{i}_{0} = \frac{p^{i}_{0} q^{i}_{0}}{\sum_{i=1}^{i=k} p^{i}_{0} q^{i}_{0}} \quad \text{et} \quad W^{i}_{t} = \frac{p^{i}_{t} q^{i}_{t}}{\sum_{i=1}^{i=k} p^{i}_{t} q^{i}_{t}}$$

Les coefficients budgétaires ont les propriétés suivantes :

$$0 \le W_{i} \le 1$$

$$\sum_{i=1}^{k} W_{i} = 1$$

Indices synthétiques de Laspeyres et moyenne arithmétique

L'indice synthétique de Laspeyres peut être défini comme étant la moyenne arithmétique des indices élémentaires pondérés par les coefficients budgétaires de la date de base.

Démonstration : Nous allons la faire pour le cas de l'indice de prix de Laspeyres. Le cas de l'indice de quantité de Laspeyres se fait d'une manière similaire.

$$L_{t/0}^{P} = \frac{\sum_{i=1}^{i=k} p_t^i q_0^i}{\sum_{i=1}^{i=k} p_0^i q_0^i} \times 100$$

$$L_{t/0}^{P} = \sum_{i=1}^{i=k} \frac{p_t^i q_0^i}{\sum_{i=1}^{i=k} p_0^i q_0^i} \times 100$$

En multipliant et en divisant par p_0^i, on obtient :

$$L_{t/0}^{P} = \sum_{i=1}^{i=k} \frac{q_0^i p_t^i}{\sum_{i=1}^{i=k} p_0^i q_0^i} \times \frac{p_0^i}{p_0^i} \times 100$$

$$= \sum_{i=1}^{i=k} \underbrace{\frac{q_0^i p_0^i}{\sum_{i=1}^{i=k} p_0^i q_0^i}}_{W_0^i} \times \underbrace{\frac{p_t^i}{p_0^i} \times 100}_{I_{t/O}^i(p)}$$

$$\boxed{L_{t/0}^{P} = \sum_{i=1}^{i=k} W_0^i \times I_{t/O}^i(p)}$$

Indices synthétiques de Paasche et moyenne harmonique

L'indice synthétique de Paasche peut être défini comme étant la moyenne harmonique des indices élémentaires pondérés par les coefficients budgétaires de la date courante.

Démonstration : Nous allons le faire pour le cas de l'indice de prix de Paasche. Le cas de l'indice de quantité de Paasche se fait d'une manière similaire.

$$P_{t/0}^{P} = \frac{\sum_{i=1}^{i=k} p_t^i q_t^i}{\sum_{i=1}^{i=k} p_0^i q_t^i} \times 100$$

On calcul l'inverse de l'indice de Paasche

$$\frac{1}{P^P_{t/0}} = \frac{\sum_{i=1}^{i=k} p^i_0 q^i_t}{\sum_{i=1}^{i=k} p^i_t q^i_t} \times \frac{1}{100}$$

$$\frac{1}{P^P_{t/0}} = \sum_{i=1}^{i=k} \frac{p^i_0 q^i_t}{\sum_{i=1}^{i=k} p^i_t q^i_t} \times \frac{1}{100}$$

En multipliant et en divisant par p^i_t, on obtient :

$$\frac{1}{P^P_{t/0}} = \sum_{i=1}^{i=k} \frac{p^i_0 q^i_t}{\sum_{i=1}^{i=k} p^i_t q^i_t} \frac{p^i_t}{p^i_t} \times \frac{1}{100}$$

$$= \sum_{i=1}^{i=k} \underbrace{\frac{p^i_t q^i_t}{\sum_{i=1}^{i=k} p^i_t q^i_t}}_{W^i_t} \left(\underbrace{\frac{p^i_0 \times 1}{p^i_t \quad 100}}_{\frac{1}{I^i_{t/0}(p)}} \right)$$

$$\boxed{\frac{1}{P^P_{t/0}} = \sum_{i=1}^{i=k} W^i_t \times \frac{1}{I^i_{t/0}(p)}}$$

Limites et extension des indices de laspeyres et de Paasche

Limites

Les deux indices de Laspeyers et de Paasche ne sont pas **réversibles**

$$\boxed{L_{t/0} \times L_{0/t} \neq 10^4}$$

et

$$\boxed{P_{t/0} \times P_{0/t} \neq 10^4}$$

Ces propriétés sont valables pour les deux indices, prix et quantité.

Les indices de Laspeyers et de Paasche ne vérifient pas la propriété de **circularité** :

$$\boxed{L_{t/0} \neq \frac{1}{100} \times L_{t/t'} \times L_{t'/0}}$$

et

$$\boxed{P_{t/0} \neq \frac{1}{100} \times P_{t/t'} \times P_{t'/0}}$$

D'une manière générale l'indice de Paasche est toujours inférieur ou égal à l'indice de Laspeyers.

L'indice de Laspeyers surestime l'évolution des prix.

L'indice de Paasche sous-estime l'évolution des prix.

Indices de Fisher

On peut définir un troisième indice, dit indice de Fisher, comme la moyenne géométrique des deux indices de Paasche et de Laspeyers.

L'indice de prix de Fisher est donc :

$$\boxed{F_{t/0}^{P} = \left[P_{t/0}^{P} \times L_{t/0}^{P}\right]^{\frac{1}{2}} = \sqrt{P_{t/0}^{P} \times L_{t/0}^{P}}}$$

L'indice de quantités de Fisher est donc :

$$\boxed{F_{t/0}^{q} = \left[P_{t/0}^{q} \times L_{t/0}^{q}\right]^{\frac{1}{2}} = \sqrt{P_{t/0}^{q} \times L_{t/0}^{q}}}$$

Propriétés de l'indice de Fisher

L'indice de Fisher est compris entre ceux de Laspeyers et Paasche

$$\boxed{P \leq F \leq L}$$

L'indice de Fisher est réversible :

$$\boxed{F_{t/0} \times F_{0/t} = 10^4}$$

L'indice de Fisher n'est pas transitif :

$$\boxed{F_{t/0} \neq \frac{1}{100} \times F_{t/t'} \times F_{t'/0}}$$

Remarques :

L'indice de la valeur globale ou de la recette totale peut être exprimé en fonction des trois indices : Laspeyers, Paasche et Fisher

$$\boxed{I(V) = I(pq) = \frac{L^{P} \times P^{q}}{100} = \frac{L^{q} \times P^{P}}{100} = \frac{F^{P} \times F^{q}}{100}}$$

Exemple 1

On dispose des données suivantes sur les prix et les quantités de deux biens en 1995 et 1998 :

	Bien 1		Bien 2	
	Prix	Quantité	Prix	Quantité
1995	10	5	25	10
1998	15	6	32	14

1) Calculer les indices de prix et de quantité de Laspeyers, de Paasche et de Fisher.

2) Calculer l'indice de la valeur globale et vérifier que :

$$I(V) = I(pq) = \frac{L^p \times P^q}{100} = \frac{L^q \times P^p}{100} = \frac{F^p \times F^q}{100}$$

1) Le calcul des indices

$$L^p_{98/95} = \frac{\sum_{i=1}^{i=2} p^i_{98} q^i_{95}}{\sum_{i=1}^{i=2} p^i_{95} q^i_{95}} \times 100 = \frac{15 \times 5 + 32 \times 10}{10 \times 5 + 25 \times 10} \times 100 = 131{,}67$$

$$L^q_{98/95} = \frac{\sum_{i=1}^{i=2} p^i_{95} q^i_{98}}{\sum_{i=1}^{i=2} p^i_{95} q^i_{95}} \times 100 = \frac{10 \times 6 + 25 \times 14}{10 \times 5 + 25 \times 10} \times 100 = 136{,}67$$

$$P^p_{98/95} = \frac{\sum_{i=1}^{i=2} p^i_{98} q^i_{98}}{\sum_{i=1}^{i=2} p^i_{95} q^i_{98}} \times 100 = \frac{15 \times 6 + 32 \times 14}{10 \times 6 + 25 \times 14} \times 100 = 131{,}22$$

$$P^q_{98/95} = \frac{\sum_{i=1}^{i=2} p^i_{98} q^i_{98}}{\sum_{i=1}^{i=2} p^i_{98} q^i_{95}} \times 100 = \frac{15 \times 6 + 32 \times 14}{15 \times 5 + 32 \times 10} \times 100 = 136{,}20$$

$$F^p_{98/95} = \sqrt{P^p_{98/95} \times L^p_{98/95}} = 131{,}44$$

$$F^q_{98/95} = \sqrt{P^q_{98/95} \times L^q_{98/95}} = 136{,}43$$

On remarque que $P \le F \le L$

2) L'indice de la valeur

$$I_{98/95}(V) = \frac{\sum_{i=1}^{i=2} p^i_{98} q^i_{98}}{\sum_{i=1}^{i=2} p^i_{95} q^i_{95}} \times 100 = \frac{15 \times 6 + 32 \times 14}{10 \times 5 + 25 \times 10} \times 100 = 179{,}34$$

$$I_{98/95}(V) = I_{98/95}(pq) = \frac{L^p \times P^q}{100} = \frac{L^q \times P^p}{100} = \frac{F^p \times F^q}{100}$$

$$I_{98/95}(V) = I_{98/95}(pq) = \frac{L^p \times P^q}{100} = \frac{L^q \times P^p}{100} = \frac{F^p \times F^q}{100}$$

$$= 131{,}67 \times 136{,}20 \times \frac{1}{100} = 136{,}67 \times 131{,}22 \times \frac{1}{100}$$

$$= 179{,}34$$

Exemple 2

Bien	1995 : (0)		1998 : (t)		Indices élementaires	Indices élementaires	$p_{95}q_{95}$	$p_{98}q_{98}$	W_{95}^i	W_{98}^i	$p_{98}q_{95}$	$p_{95}q_{98}$
	p_{95}	q_{95}	p_{98}	q_{98}	$I_{98/85}(p)$	$I_{98/85}(q)$						
A	12	6	15	7	125	116	72	105	0,33	0,37	90	84
B	5	13	8	11	160	84	65	88	0,30	0,31	104	55
C	8	10	10	9	125	90	80	90	0,37	0,32	100	72
Total							**217**	**283**	1	1	**294**	**211**

1) Le calcul des indices synthétiques

$$L_{98/95}^p = \frac{\sum_{i=1}^{i=3} p_{98}^i q_{95}^i}{\sum_{i=1}^{i=3} p_{95}^i q_{95}^i} \times 100 = \frac{294}{217} \times 100 = 135$$

$$L_{98/95}^q = \frac{\sum_{i=1}^{i=3} p_{95}^i q_{98}^i}{\sum_{i=1}^{i=3} p_{95}^i q_{95}^i} \times 100 = \frac{211}{217} \times 100 = 97$$

$$P_{98/95}^p = \frac{\sum_{i=1}^{i=3} p_{98}^i q_{98}^i}{\sum_{i=1}^{i=3} p_{95}^i q_{98}^i} \times 100 = \frac{283}{211} \times 100 = 134$$

$$P_{98/95}^q = \frac{\sum_{i=1}^{i=3} p_{98}^i q_{98}^i}{\sum_{i=1}^{i=3} p_{98}^i q_{95}^i} \times 100 = \frac{283}{294} \times 100 = 96$$

$$F_{98/95}^p = \sqrt{P_{98/95}^p \times L_{98/95}^p} = 134{,}4$$

$$F_{98/95}^q = \sqrt{P_{98/95}^q \times L_{98/95}^q} = 96{,}49$$

L'indice de la valeur est :

$$I_{98/95}(V) = \frac{\sum_{i=1}^{i=3} p_{98}^{i} q_{98}^{i}}{\sum_{i=1}^{i=3} p_{95}^{i} q_{95}^{i}} \times 100 = \frac{283}{217} \times 100 = 130{,}41$$

On peut vérifier que :

$$L_{98/95}^{P} = \sum_{i=1}^{i=3} W_{95}^{i} \times I_{98/95}^{i}(p)$$

$$= (0{,}33 \times 125) + (0{,}30 \times 160) + (0{,}37 \times 125) = 130$$

$$L_{98/95}^{q} = \sum_{i=1}^{i=3} W_{95}^{i} \times I_{98/95}^{i}(q)$$

$$= (0{,}33 \times 116) + (0{,}30 \times 84) + (0{,}37 \times 90) = 97$$

et que :

$$\frac{1}{P_{98/95}^{P}} = \sum_{i=1}^{i=k} W_{98}^{i} \times \frac{1}{I_{98/95}^{i}(p)}$$

$$= (0{,}37 \times \frac{1}{125}) + (0{,}31 \times \frac{1}{160}) + (0{,}32 \times \frac{1}{125}) = \frac{1}{134}$$

$$\frac{1}{P_{98/95}^{q}} = \sum_{i=1}^{i=k} W_{98}^{i} \times \frac{1}{I_{98/95}^{i}(q)}$$

$$= (0{,}37 \times \frac{1}{116}) + (0{,}31 \times \frac{1}{84}) + (0{,}32 \times \frac{1}{90}) = \frac{1}{96}$$

Chapitre 6 :
Introduction à l'analyse des distributions à deux variables

Chapitre VI : Introduction à l'analyse des distributions à deux variables

On considère une population de N individus mesurés simultanément par les deux caractères X et Y, de modalités $x_1,\ldots x_i,\ldots x_L$ pour la variable X et $y_1,\ldots y_j,\ldots y_K$ pour la variable Y. On note par n_{ij} le nombre d'individus appartenant à la fois à une classe de rang i (*pour la variable X*) et à une classe de rang j (*pour la variable Y*).

Présentation d'un tableau à double entrée

Exemple

On considère le tableau suivant, relatif à une population de 100 ménages, où X désigne le nombre d'enfants du ménage et Y est le nombre de pièces du logement.

Y_j / X_i	**3** ($j=1$)	**4** ($j=2$)	**5** ($j=3$)	**Total**
2 ($i=1$)	15	10	5	**30**
3 ($i=2$)	$\rightarrow$30	5	10	**45**
4 ($i=3$)	10	5	0	**15**
5 ($i=4$)	10	0	0	**10**
Total	**65**	**20**	**15**	**100**

Remarques :

La valeur **30** indique que, parmi les 100 ménages observés, il y a 30 ménages qui ont 3 enfants et qui habitent dans des logements de 3 pièces.

La valeur **65** indique que, parmi les 100 ménages observés, il y a 65 ménages habitent dans des logements de 3 pièces.

La valeur **45** indique que, parmi les 100 ménages observés, il y a 45 ménages qui ont 3 enfants.

Tableau de contingence

Y_j / X_i	y_1	y_2	.....	y_j	.....	y_K	**Total**
x_1	n_{11}	n_{12}		n_{1j}		n_{1K}	$n_{1.}$
x_2	n_{21}	n_{22}		n_{2j}		n_{2K}	$n_{2.}$
⋮							
x_i	n_{i1}	n_{i2}		n_{ij}		n_{iK}	$n_{i.}$
⋮							
x_L	n_{L1}	n_{L2}		n_{Lj}		n_{LK}	$n_{L.}$
Total	$n_{.1}$	$n_{.2}$		$n_{.j}$		$n_{.K}$	N

Les effectifs situés à l'intérieur du tableau sont notés par n_{ij}, où n_{ij} désigne le nombre de fois où la modalité x_i de la variable *X* et la modalité y_j de la variable *Y* ont été observées simultanément.

L'effectif $n_{i.}$, appelé effectif marginal de *X*, est le nombre total d'observations de la modalité x_i de la variable *X* quelque soit la modalité de la variable :

$$\boxed{n_{i.} = \sum_{j=1}^{j=K} n_{ij}}$$

L'effectif $n_{.j}$, appelé effectif marginal de *Y*, représente le nombre total d'observations de la modalité y_j de la variable *Y* quelque soit la modalité de la variable X :

$$\boxed{n_{.j} = \sum_{i=1}^{i=L} n_{ij}}$$

L'effectif total de la distribution conjointe, noté *N*, peut être obtenu à partir de l'effectif marginal de X ou bien à partir de l'effectif marginal de *Y* :

$$\boxed{N = \sum_{i=1}^{i=L} n_{i.} = \sum_{j=1}^{j=K} n_{.j} = \sum_{i=1}^{i=L} \sum_{j=1}^{j=K} n_{ij}}$$

Remarque : la distribution conjointe des variables *X* et *Y* peut être définie à partir des fréquences relatives :

$$f_{ij} = \frac{n_{ij}}{N} \qquad \text{avec } f_{i.} = \frac{n_{i.}}{N} = \sum_{j=1}^{j=K} f_{ij} \;\; ; \;\; f_{.j} = \frac{n_{.j}}{N} = \sum_{i=1}^{i=L} f_{ij} \quad \text{et} \quad \sum_{i=1}^{i=L}\sum_{j=1}^{j=K} f_{ij} = 1$$

Distributions marginales

Définition

A partir de la distribution conjointe des variables *X* et *Y,* on peut déduire la distribution marginale de chacune des deux variables. Ceci nous permet d'analyser séparément la distribution de chacune des deux variables.

On appelle distribution marginale de la variable *X, la donnée des L couples* $(x_i, n_{i.})$.

On appelle distribution marginale de la variable *Y, la donnée des K couples* $(y_j, n_{.j})$.

Ces deux distributions peuvent se présenter sous forme de tableaux statistiques.

Distribution marginale de *X*

X_i	**Effectif marginal**
x_1	$n_{1.}$
x_2	$n_{2.}$
⋮	
x_i	$n_{i.}$
⋮	
x_L	$n_{L.}$
Total	N

Distribution marginale de *Y*

Y_j	**Effectif marginal**
y_1	$n_{.1}$
y_2	$n_{.2}$
⋮	
y_j	$n_{.j}$
⋮	
y_K	$n_{.K}$
Total	N

Remarque : la distribution marginale de chacune des variables *X* et *Y* peut être définie à partir des fréquences relatives :

$$f_{i.} = \frac{n_{i.}}{N} \qquad \text{et} \qquad f_{.j} = \frac{n_{.j}}{N}$$

Exemple

En reprenant l'exemple de la distribution des 100 ménages selon le nombre d'enfants du ménage et le nombre de pièces du logement, la distribution marginale selon chacun des deux caractères peut se

présenter de la manière suivante :

Distribution marginale de X

X_i	**Effectif marginal**
2	**30**
3	**45**
4	**15**
5	**10**
Total	**100**

Distribution marginale de Y

Y_j	**Effectif marginal**
3	**65**
4	**20**
5	**15**
Total	**100**

Distributions conditionnelles

Définition

On appelle distribution conditionnelle de Y pour $X = x_i$, la distribution des individus correspondant à une modalité x_i de la variable X suivant les modalités de la variable Y.

On appelle distribution conditionnelle de X pour $Y = y_j$, la distribution des individus correspondant à une modalité y_j la variable Y suivant les modalités de la variable X.

Ces deux distributions peuvent se présenter sous forme de tableaux statistiques.

Distribution conditionnelle de X sachant $Y = Y_j$	
$X/_{Y=Y_j}$	$n_{i/j}$
x_1	n_{1j}
⋮	
x_i	n_{ij}
x_L	n_{Lj}
Total	$n_{.j}$

Distribution conditionnelle de Y sachant $X = X_i$	
$Y/_{X=x_i}$	$n_{j/i}$
y_1	n_{i1}
⋮	
y_j	n_{ij}
y_K	n_{iK}
Total	$n_{i.}$

Remarque : la distribution *conditionnelle* de chacune des variables X et Y peut être définie à partir des fréquences relatives .

Dans le cas de la distribution conditionnelle de X pour $Y = y_j$, on a :

$$f_{i/j} = \frac{n_{ij}}{n_{.j}} = \frac{\frac{n_{ij}}{N}}{\frac{n_{.j}}{N}} = \frac{f_{ij}}{f_{.j}} \quad ; \quad \sum_{i=1}^{i=L} f_{i/j} = 1$$

Dans le cas de la distribution conditionnelle de Y pour $X = x_i$, on a :

$$f_{j/i} = \frac{n_{ij}}{n_{i.}} = \frac{\frac{n_{ij}}{N}}{\frac{n_{i.}}{N}} = \frac{f_{ij}}{f_{i.}} \quad ; \quad \sum_{j=1}^{j=K} f_{j/i} = 1$$

Exemple

En reprenant l'exemple de la distribution des 100 ménages selon le nombre d'enfants du ménage et le nombre de pièces du logement, la distribution conditionnelle de X sachant $Y = 4$ et la distribution conditionnelle de Y sachant $X = 3$ se présentent ainsi :

Distribution conditionnelle de X sachant $Y = 4$	
2	10
3	5
4	5
5	0
Total	**20**

Distribution conditionnelle de Y sachant $X = 3$	
3	30
4	5
5	10
Total	**45**

Dépendance et indépendance entre les variables *X* et *Y*

Les variables X et Y sont dites statistiquement indépendantes lorsque la distribution de la variables X ne dépend pas de la variable Y ou vice versa. Dans ce cas, la connaissance de la variable Y ne donne aucune information sur la variables X, auquel cas, toutes les distributions conditionnelles de la variables X sont identiques à la distribution marginale de la variables X.

L'indépendance se traduit en termes de fréquences relatives par :

$$f_{ij} = f_{i.} \times f_{.j} \qquad \forall \ i, j$$

Chapitre 7 :
Corrélation et Ajustement linéaire

Chapitre VII : Corrélation et Ajustement linéaire

Dans le cadre de ce chapitre, on s'intéresse à l'étude d'une éventuelle relation entre deux variables statistiques. En d'autres termes, nous allons voir, d'abord, comment déterminer le sens de la liaison entre ces deux variables, ensuite, comment mesurer l'intensité ou le degré de la liaison entre elles, et enfin fournir une expression mathématique de la liaison entre ces deux variables.

La covariance entre *X* et *Y*

Définition

La covariance est égale à la moyenne des écarts des couples (x_i, y_i) de *X* et *Y* par rapport au point $(\bar{x}, \bar{y})$.

$$\boxed{Cov(x,y) = \frac{1}{N}\sum_{i=1}^{i=N}(x_i - \bar{x})(y_i - \bar{y})}$$

La covariance indique le sens de la relation entre les variables *X* et *Y*. Ainsi, On peut distinguer les cas suivants :

Si $Cov(x,y) > 0$, alors on peut dire que la relation entre les deux variables est positive. Dans ce cas, ces deux variables varient dans le même sens.

Si $Cov(x,y) < 0$, alors on peut dire que la relation entre les deux variables est négative. Dans ce cas, ces deux variables varient en sens inverse.

Si $Cov(x,y) = 0$, alors on peut dire qu'il n'y a pas de relation entre les deux variables. Dans ce cas, les variations de l'une n'entraînent pas la variation de l'autre.

Propriétés

i) $\boxed{Cov(ax+b, cy+d) = ac.Cov(x,y)}$

Démonstration

$$Cov(ax+b, cy+d) = \frac{1}{N}\sum_{i=1}^{i=N}\left[(ax_i + b) - (a\bar{x} + b)\right]\left[(cy_i + d) - (c\bar{y} + d)\right]$$

$$= \frac{1}{N}\sum_{i=1}^{i=N}(ax_i - a\bar{x})(cy_i - c\bar{y}) = \frac{1}{N}\sum_{i=1}^{i=N}\left[a(x_i - \bar{x})\right].\left[c(y_i - \bar{y})\right]$$

$$= \frac{1}{N}\sum_{i=1}^{i=N}(a \times c)\left[(x_i - \bar{x})(y_i - \bar{y})\right] = (a \times c) \times Cov(x,y)$$

ii) $\boxed{Cov(y,x) = Cov(x,y)}$

Démonstration

$$Cov(x,y) = \frac{1}{N}\sum_{i=1}^{i=N}(x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{N}\sum_{i=1}^{i=N}(y_i - \bar{y})(x_i - \bar{x}) = Cov(y,x)$$

iii) $\boxed{Cov(x,x) = V(x)}$

Démonstration

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

iv) $\boxed{Cov(x,y) = \left[\frac{1}{N}\sum_{i=1}^{i=N}x_i y_i\right] - [\bar{x}.\bar{y}]}$

Démonstration

$$Cov(x,y) = \frac{1}{N}\sum_{i=1}^{i=N}(x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{N}\sum_{i=1}^{i=N}(x_i y_i - \bar{y}x_i - \bar{x}y_i + \bar{x}\bar{y})$$

$$= \frac{1}{N}\left[\sum_{i=1}^{i=N}(x_i y_i) - \sum_{i=1}^{i=N}(\bar{y}x_i) - \sum_{i=1}^{i=N}(\bar{x}y_i) + \sum_{i=1}^{i=N}(\bar{x}.\bar{y})\right]$$

$$= \frac{1}{N}\left[\sum_{i=1}^{i=N}(x_i y_i) - \bar{y}\underbrace{\sum_{i=1}^{i=N}(x_i)}_{N.\bar{x}} - \bar{x}\underbrace{\sum_{i=1}^{i=N}(y_i)}_{N.\bar{y}} + \underbrace{\sum_{i=1}^{i=N}(\bar{x}.\bar{y})}_{N.\bar{y}.\bar{x}}\right]$$

$$= \frac{1}{N}\left[\sum_{i=1}^{i=N}(x_i y_i) - N.\bar{x}.\bar{y} - N.\bar{x}.\bar{y} + N.\bar{y}.\bar{x}\right]$$

$$= \frac{1}{N}\left[\sum_{i=1}^{i=N}(x_i y_i) - N.\bar{y}.\bar{x}\right] = \left[\frac{1}{N}\sum_{i=1}^{i=N}(x_i y_i)\right] - (\bar{x}.\bar{y})$$

Le coefficient de corrélation linéaire entre *X* et *Y*

Définition

Le coefficient de corrélation linéaire est un nombre sans dimension qui permet de mesurer le degré ou l'intensité de la liaison **linéaire** entre deux variables statistiques. Ainsi, la formule du coefficient de corrélation linéaire entre *X* et *Y* est :

$$\boxed{r_{x,y} = \frac{Cov(x,y)}{\sigma_x \sigma_y} = \frac{Cov(x,y)}{\sqrt{V(x)}\sqrt{V(y)}}}$$

La covariance indique le sens de la relation entre les variables *X* et *Y*.

Ainsi, On peut distinguer les cas suivants :

Si $r_{x,y} > 0$, les deux variables varient dans le même sens.

Si $r_{x,y} < 0$, les deux variables varient en sens inverse.

Si $r_{x,y} = 0$, les deux variables sont <u>linéairement</u> indépendantes.

Propriétés

i) $\boxed{r_{ax+b,cy+d} = (\text{signe de } a) \times (\text{signe de } c) \times r_{x,y}}$

Démonstration

$$r_{ax+b,cy+d} = \frac{Cov(ax+b, cy+d)}{\sqrt{V(ax+b)}\sqrt{V(cy+d)}} = \frac{(a \times c).Cov(x,y)}{|a|\sqrt{V(x)} \times |c|\sqrt{V(y)}}$$

$$= \frac{(a \times c).Cov(x,y)}{|a| \times |c|\sqrt{V(x)}\sqrt{V(y)}} = \frac{(a \times c)}{|a| \times |c|} \frac{Cov(x,y)}{\sqrt{V(x)}\sqrt{V(y)}}$$

$$= (\text{signe de } a) \times (\text{signe de } c) \times r_{x,y}$$

$\boxed{r_{y,x} = r_{x,y}}$

Démonstration

$$r_{x,y} = \frac{Cov(x,y)}{\sigma_x \sigma_y} = \frac{Cov(y,x)}{\sigma_y \sigma_x} = r_{y,x}$$

$\boxed{r_{x,x} = 1}$

Démonstration

$$r_{x,x} = \frac{Cov(x,x)}{\sigma_x \sigma_x} = \frac{V(x)}{\sigma_x \sigma_x} = 1$$

$\boxed{-1 \le r \le +1}$

Interprétation de la valeur de $r_{x,y}$

Si $r_{x,y} = 1$: on dit qu'il y a une parfaite corrélation <u>linéaire</u> positive entre les deux variables.

Si $r_{x,y} = -1$: on dit qu'il y a une parfaite corrélation <u>linéaire</u> négative entre les deux variables.

Si $r_{x,y} = 0$, on dit qu'il y a absence de corrélation <u>linéaire</u> entre les deux variables.

On dit qu'il y a une forte corrélation linéaire entre les deux variables (ou forte dépendance linéaire) si r est proche de ± 1. En revanche, si r est proche de zéro, on dit qu'il y a une faible corrélation linéaire entre les

deux variables.

Ajustement linéaire d'un nuage de points

On considère deux variables statistiques quantitatives x *et* y et on s'intéresse à une relation éventuelle entre elles.

La représentation du nuage de points peut nous renseigner sur l'allure de la distribution à deux caractères. La forme de la relation entre les deux variables peut être mise en évidence graphiquement par les courbes de régression.

Généralement, on exprime y en fonction de x, on parle alors de la droite de régression de y sur x (ou de y en x). Dans ce cas, on cherche à expliquer la variable y par la variable x. De ce fait, y est dite variable expliquée ou variable endogène et x est appelée variable explicative ou variable exogène.

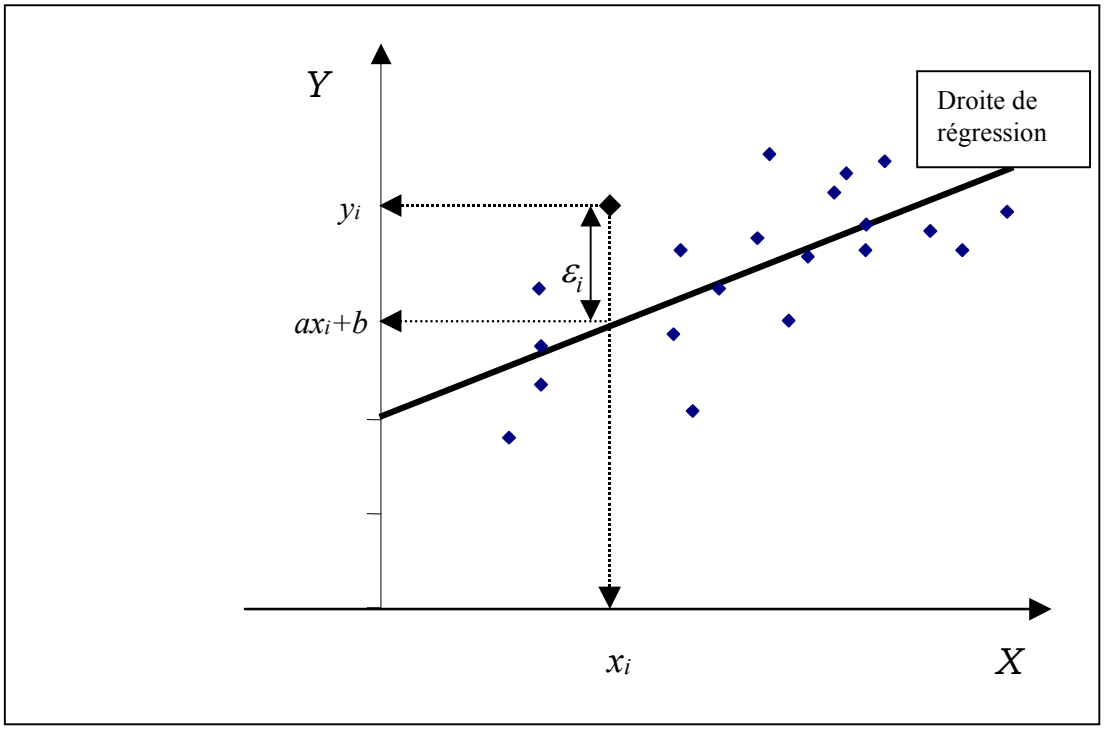
La droite de régression de y sur x

On considère N observations sur les deux variables x et y. Ces observations peuvent être représentées par un nuage de points. D'une manière générale, l'ajustement d'un nuage de point par une fonction mathématique, revient à estimer les valeurs des coefficients de cette fonction de telle sorte que sa courbe représentative se rapproche au mieux du nuage de points.

Lorsqu'il s'agit d'une liaison linéaire entre les deux variables, on parle alors d'ajustement linéaire. L'ajustement linéaire consiste à estimer les coefficients de la droite de régression du type $y = ax + b$, c'est à dire à trouver la valeur de a et celle de b.

Cette droite est supposée refléter l'évolution moyenne de la variable y (variable expliquée) en fonction de la variable explicative x.

La méthode d'ajustement que nous allons exposer est appelée « *méthode des **M**oindres **C**arrés **O**rdinaires* » ou simplement « *MCO* ».



Critère des moindres carrés

Considérons N couples d'observations (x_i, y_i) tels que :

$$y_i = (ax_i + b) + \varepsilon_i$$

où ε_i représente le résidu du couple (x_i, y_i). On peut alors écrire :

$$\varepsilon_i = y_i - (ax_i + b)$$

La méthode *MCO* consiste à ajuster le nuage de points par une droite de manière à minimiser la somme des carrés des distances entre les points du nuage et cette droite. Ceci revient à minimiser la somme des carrés des résidus.

Remarque : On minimise la somme des carrés des résidus et non la somme des résidus car : $\sum_{i=1}^{i=N} \varepsilon_i = 0$

Détermination des deux paramètres a et b par *la méthode MCO.*

$$\varepsilon_i = y_i - (ax_i - b) \Rightarrow \varepsilon_i^2 = (y_i - ax_i - b)^2$$

La somme des carrés des résidus pour $i = 1,2,\cdots,N$ est donnée par :

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

Les deux conditions de premier ordre de la minimisation de cette fonction *f* par rapport à *a* et à *b* sont :

$$\frac{\partial\left(\sum_{i=1}^{i=N}\varepsilon_i^2\right)}{\partial a}=0 \text{ et } \frac{\partial\left(\sum_{i=1}^{i=N}\varepsilon_i^2\right)}{\partial b}=0$$

$$\frac{\partial\sum_{i=1}^{i=N}e_i^2}{\partial a}=2\sum_{i=1}^{i=N}(y_i-ax_i-b)(-x_i)=0\Rightarrow\sum_{i=1}^{i=N}(y_i-ax_i-b)(x_i)=0 \qquad (1)$$

$$\frac{\partial\sum_{i=1}^{i=N}e_i^2}{\partial b}=2\sum_{i=1}^{i=N}(y_i-ax_i-b)(-1)=0\Rightarrow\sum_{i=1}^{i=N}(y_i-ax_i-b)=0 \qquad (2)$$

Le développement de ces deux équations nous donne :

$$(1)\Rightarrow\sum_{i=1}^{i=N}(y_ix_i-ax_i^2-bx_i)=\sum_{i=1}^{i=N}y_ix_i-a\sum_{i=1}^{i=N}x_i^2-b\sum_{i=1}^{i=N}x_i=0 \qquad (3)$$

$$(2)\Rightarrow\sum_{i=1}^{i=N}(y_i-ax_i-b)=\sum_{i=1}^{i=N}y_i-a\sum_{i=1}^{i=N}x_i-Nb=0 \qquad (4)$$

En divisant les deux membres de l'équation (4) par N, on obtient :

$$\frac{\sum_{i=1}^{i=N}y_i}{N}-a\frac{\sum_{i=1}^{i=N}x_i}{N}-\frac{Nb}{N}=0$$

Sachant que :

$$\bar{x}=\frac{1}{N}\sum_{i=1}^{i=N}x_i \text{ et que } \bar{y}=\frac{1}{N}\sum_{i=1}^{i=N}y_i$$

L'équation (4) devient :

$$\bar{y}-a\,\bar{x}-b=0 \qquad (5)$$

En remplaçant, dans l'équation (3), b par : $\bar{y}-a\,\bar{x}$ (d'après l'équation (5)), on a :

$$\sum_{i=1}^{i=N}y_ix_i-a\sum_{i=1}^{i=N}x_i^2-(\bar{y}-a\bar{x})\sum_{i=1}^{i=N}x_i=0$$

$$\Leftrightarrow\sum_{i=1}^{i=N}y_ix_i-a\sum_{i=1}^{i=N}x_i^2-\bar{y}\underbrace{\sum_{i=1}^{i=N}x_i}_{N\bar{x}}+a\bar{x}\underbrace{\sum_{i=1}^{i=N}x_i}_{N\bar{x}}=0$$

$$\Leftrightarrow\sum_{i=1}^{i=N}y_ix_i-a\sum_{i=1}^{i=N}x_i^2-N\bar{x}.\bar{y}+aN\left[\bar{x}\right]^2=0$$

$$\Leftrightarrow\sum_{i=1}^{i=N}y_ix_i-N\bar{x}.\bar{y}=a\left(\sum_{i=1}^{i=N}x_i^2-N\left[\bar{x}\right]^2\right)$$

Ainsi, on obtient la valeur estimée de la pente de la droite de régression :

$$\boxed{\hat{a} = \frac{\sum_{i=1}^{i=N} y_i x_i - N\bar{x}\bar{y}}{\sum_{i=1}^{i=N} x_i^2 - N\bar{x}^2}}$$

et par là la valeur estimée de b :

$$\boxed{\hat{b} = \bar{y} - \hat{a}\bar{x}}$$

Remarque :

On peut aussi calculer la valeur estimée de la pente de la droite de régression en utilisant l'une de ces deux expressions

$$\boxed{\hat{a} = \frac{\sum_{i=1}^{i=N} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{i=N} (x_i - \bar{x})^2}} \text{ ou } \boxed{\hat{a} = \frac{Cov(x,y)}{V(x)}}$$

Enfin, l'équation de la droite de régression est donnée par :

$$\boxed{y_i = \hat{a}x_i + \hat{b}}$$

Remarque :

La droite de régression passe par le point moyen de coordonnées $(\bar{x}, \bar{y})$. En effet, Comme, $\hat{b} = \bar{y} - \hat{a}\bar{x}$, on a alors $\bar{y} = \hat{a}\bar{x} + \hat{b}$.

L'étude de la droite de régression de y sur x permet de prévoir y en fonction x :

$$\boxed{\hat{y} = \hat{a}x + \hat{b}}$$

La droite de régression de x sur y

On peut exprimer x en fonction de y. Dans ce cas, on appelle x une variable endogène ou expliquée et y une variable exogène ou explicative, et on parle de la droite de régression de x sur y :

$$x = a'y + b'.$$

En utilisant la méthode des moindres carrés ordinaires, on retrouve la valeur de a' et de b' exprimées par :

$$\boxed{\hat{a}' = \frac{\sum_{i=1}^{i=N} y_i x_i - N\bar{x}\bar{y}}{\sum_{i=1}^{i=N} y_i^2 - N\bar{y}^2}} \text{ et } \boxed{\hat{b}' = \bar{x} - \hat{a}'\bar{y}}$$

On peut montrer aussi que :

$$\hat{a}' = \frac{\sum_{i=1}^{i=N}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{i=N}(y_i - \bar{y})^2} = \frac{\frac{1}{N}\sum_{i=1}^{i=N}(x_i - \bar{x})(y_i - \bar{y})}{\frac{1}{N}\sum_{i=1}^{i=N}(y_i - \bar{y})^2} = \frac{Cov(x,y)}{V(y)}$$

Remarque :

L'étude de la droite de régression de x sur y permet de prévoir x en fonction de y :

$$\boxed{\hat{x} = \hat{a}'\, y + \hat{b}'}$$

Décomposition de la variance totale

$$\sum_{i=1}^{i=N}(y_i - \bar{y})^2 = \sum_{i=1}^{i=N}\left[(y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})\right]^2$$

$$= \sum_{i=1}^{i=N}\left[y_i - \hat{y}_i\right]^2 + \sum_{i=1}^{i=N}\left[\hat{y}_i - \bar{y}\right]^2 + 2\sum_{i=1}^{i=N}\underbrace{(y_i - \hat{y}_i)}_{(1)} \times \underbrace{(\hat{y}_i - \bar{y})}_{(2)}$$

$$\underbrace{(\hat{y}_i - \bar{y})}_{(2)} = (\hat{a}.x_i + \hat{b}) - (\hat{a}.\bar{x} + \hat{b}) = \hat{a}(x_i - \bar{x})$$

$$\underbrace{(y_i - \hat{y}_i)}_{(1)} = (y_i - \bar{y}) - (\hat{y}_i - \bar{y}) = (y_i - \bar{y}) - \hat{a}(x_i - \bar{x})$$

$$(1)\times(2) = \left[\hat{a}(x_i - \bar{x})\right] \times \left[(y_i - \bar{y}) - \hat{a}(x_i - \bar{x})\right]$$

$$\sum_{i=1}^{i=N}(1)\times(2) = \hat{a}\left[\sum_{i=1}^{i=N}\left[(x_i - \bar{x})(y_i - \bar{y})\right] - \hat{a}\sum_{i=1}^{i=N}(x_i - \bar{x})^2\right]$$

or $\hat{a} = \dfrac{\sum_{i=1}^{i=N}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{i=N}(x_i - \bar{x})^2} \Rightarrow \hat{a}\sum_{i=1}^{i=N}(x_i - \bar{x})^2 = \sum_{i=1}^{i=N}(x_i - \bar{x})(y_i - \bar{y})$

d'où $\sum_{i=1}^{i=N}(1)\times(2) = \hat{a}\left[\sum_{i=1}^{i=N}\left[(x_i - \bar{x})(y_i - \bar{y})\right] - \sum_{i=1}^{i=N}\left[(x_i - \bar{x})(y_i - \bar{y})\right]\right] = 0$

$$\boxed{\underbrace{\sum_{i=1}^{i=N}(y_i - \bar{y})^2}_{\text{Somme des Carrés Totale}} = \underbrace{\sum_{i=1}^{i=N}\left[y_i - \hat{y}_i\right]^2}_{\text{Somme des Carrés des Résidus}} + \underbrace{\sum_{i=1}^{i=N}\left[\hat{y}_i - \bar{y}\right]^2}_{\text{Somme des Carrés Expliquée}}}$$

$$\boxed{\text{SCT} = \text{SCR} + \text{SCE}}$$

En divisant par les deux membres par N on obtient l'équation d'analyse de la variance.

$$\underbrace{\frac{1}{N}\sum_{i=1}^{i=N}(y_i-\bar{y})^2}_{\text{VARIANCE TOTALE}}=\underbrace{\frac{1}{N}\sum_{i=1}^{i=N}\left[y_i-\hat{y}_i\right]^2}_{\text{VARIANCE RESIDUELLE}}+\underbrace{\frac{1}{N}\sum_{i=1}^{i=N}\left[\hat{y}_i-\bar{y}\right]^2}_{\text{VARIANCE EXPLIQUEE}}$$

Coefficient de détermination

L'équation d'analyse de la variance nous permet d'avoir une idée sur la qualité d'ajustement. Afin de mesurer la qualité de cet ajustement, on défini le coefficient de détermination, noté R^2, par la part de la variance expliquée dans la variance totale :

$$R^2=\frac{\text{VARIANCE EXPLIQUEE}}{\text{VARIANCE TOTALE}}=\frac{\text{SCE}}{\text{SCT}}=1-\frac{\text{SCR}}{\text{SCT}}$$

Remarque :

on peut retenir le coefficient de détermination comme étant le carré du coefficient de corrélation linéaire entre x et y.

$$R^2=(r_{x,y})^2=\left(\frac{Cov(x,y)}{\sigma_x\sigma_y}\right)^2$$

Le coefficient de détermination est aussi égal au produit des pentes des deux droites de régression, de y sur x et de x sur y.

$$R^2=a\times a'$$

En effet,

$$R^2=\left[\frac{Cov(x,y)}{\sigma_x\sigma_y}\right]^2=\frac{Cov(x,y)\times Cov(y,x)}{V(x)\times V(y)}=\frac{Cov(x,y)}{V(x)}\times\frac{Cov(y,x)}{V(y)}=a\times a'$$

Remarque :

$$r_{x,y}=(\text{signe de } Cov(x,y))\times\sqrt{\hat{a}\times\hat{a}'}$$

Interprétation de la valeur de R²

Si $R^2=1$: on dit qu'il y a dépendance totale ou liaison fonctionnelle entre les deux variables. Les deux droites de régression, de y sur x et de x sur y, sont alors confondues.

Si $R^2=0$, on dit qu'il y a indépendance totale ou liaison nulle entre les deux variables. Les deux droites de régression sont alors perpendiculaires .

Si $0<R^2<1$: on dit qu'il y a liaison relative entre les deux variables.

On dit que la qualité d'ajustement est bonne si R^2 est proche de 1. En

revanche, si R^2 est proche de zéro, on dit que la qualité de l'ajustement est mauvaise.

Exemple :

Cas 1	
x	y
2	8
5	12
9	18
11	24

Cas 2	
x	y
5	90
8	12
10	1
2	4
9	45

Pour les deux cas, on détermine les deux droites de régression, en utilisant les formules de a, a', b et b'.

Cas 1 :

1) $y_i = ax_i + b \Rightarrow \hat{y}_i = 1{,}71x_i + 3{,}93$

2) $x_i = a'y_i + b' \Rightarrow \hat{x}_i = 0{,}56y_i - 2{,}05$

Dans ce cas : $\hat{a} \times \hat{a}' = 1{,}71 \times 0{,}56 = 0{,}9576$

Cas 2 :

1) $y_i = ax_i + b \Rightarrow \hat{y}_i = -1{,}5x_i + 40{,}66$

2) $x_i = a'y_i + b' \Rightarrow \hat{x}_i = -0{,}011y_i + 7{,}14$

Dans ce cas : $\hat{a} \times \hat{a}' = -1{,}5 \times (-0{,}011) = 0{,}0165$

Dans le cas 1, le produit $\hat{a} \times \hat{a}' = 0{,}9576$ est proche de 1 alors que dans cas 2, le même produit $\hat{a} \times \hat{a}' = 0{,}0165$ est proche de zéro. L'examen des données (cas 1) montre que x et y varient dans le même sens et que la variation de x conditionne celle de y. Par contre, l'examen des données (cas 2) indique que la variation de y est indépendante de celle de x. Ainsi, on remarque que lorsque les deux variables sont liées entre elles, le produit $\hat{a} \times \hat{a}'$ est proche de 1. Ce même produit sera proche de zéro dans le cas contraire.

Ajustement non linéaire

L'ajustement linéaire suppose que la forme de la fonction reliant y et x est linéaire du type : $y = ax + b$. Cependant, dans d'autres cas, la

relation entre y et x semble être plutôt non linéaire.

Exemple 1

La fonction permettant de représenter le nuage de points est une fonction hyperbolique du type :

$$y = \frac{b}{x^a},\ b > 0$$

Comment peut-on estimer b et a ?

Nous sommes en présence d'une relation non linéaire entre y et x. Afin d'utiliser la méthode des MCO, il faut d'abord retrouver, moyennant une transformation, dans ce cas logarithmique, une forme linéaire :

On a :

$$y = \frac{b}{x^a} = bx^{-a} \Rightarrow \log y = \log bx^{-a} = \log b - a \log x$$

Supposons que : $\log b = \beta$ et $-a = \alpha$, le modèle linéaire est alors de la forme : $\boxed{\log y = \alpha \log x + \beta}$

En utilisant la méthode des MCO, on peut retrouver l'expression α et de β :

$$\boxed{\hat{\alpha} = \frac{Cov(\log x, \log y)}{V(\log x)}} \text{ et } \boxed{\hat{\beta} = \overline{\log y} - \hat{\alpha}\overline{\log x}}$$

On peut maintenant retrouver la valeur de b et la valeur de a :

$$\log b = \beta \Rightarrow \hat{b} = e^{\hat{\beta}}$$

$$-a = \alpha \Rightarrow \hat{a} = -\hat{\alpha}$$

Exemple 2

La fonction permettant de représenter le nuage de points est une fonction parabolique du type :

$$y = ax^2 + b$$

Comment peut-on estimer b et a ?

Nous sommes en présence d'une relation non linéaire entre y et x. Afin d'utiliser la méthode des MCO, il suffit de poser $z = x^2$. On obtient ainsi une forme linéaire entre y et z :

$$\boxed{y = az + b}$$

En utilisant la méthode des MCO, on peut retrouver l'expression a et de b:

$$\boxed{\hat{a} = \frac{Cov(z, y)}{V(z)}} \text{ et } \boxed{\hat{b} = \overline{y} - \hat{a}\overline{z}}$$

Remarque : Le choix entre l'ajustement linéaire et l'ajustement non linéaire peut être basé sur la forme générale du nuage de points. En effet, si cette forme est linéaire on applique directement la méthode des MCO. Dans le cas inverse (forme de nuage non linéaire), on doit au préalable passer par une transformation appropriée afin d'obtenir une relation linéaire qu'on peut estimer par les MCO.

- Exemple illustratif du calcul des coefficients de la régression de Y sur X ainsi que la décomposition de la variance totale.

X : note obtenue en test d'intelligence.

Y : note obtenue en statistique.

Pour calculer la valeur de $\hat{a}$ et $\hat{b}$, on effectue les calculs suivants :

i	x_i	y_i	x_i^2	y_i^2	$x_i y_i$
1	2	6	4	36	12
2	1	5	1	25	5
3	7	15	49	225	105
4	5	11	25	121	55
5	3	9	9	81	27
Total	**18**	**46**	**88**	**488**	**204**

On a :

- $\bar{x} = \frac{1}{N}\sum_{i=1}^{i=N} x_i = \frac{1}{5} \times 18 = 3{,}6 \quad \overline{x^2} = \frac{1}{N}\sum_{i=1}^{i=N} x_i^2 = \frac{1}{5} \times 88 = 17{,}6$
- $V(x) = \overline{x^2} - (\bar{x})^2 = 17{,}6 - 12{,}96 = 4{,}64$
- $\bar{y} = \frac{1}{N}\sum_{i=1}^{i=N} y_i = \frac{1}{5} \times 46 = 9{,}2 \quad \overline{y^2} = \frac{1}{N}\sum_{i=1}^{i=N} y_i^2 = \frac{1}{5} \times 488 = 97{,}6$
- $V(y) = \overline{y^2} - (\bar{y})^2 = 97{,}6 - 84{,}64 = 12{,}96$
- $\overline{xy} = \frac{1}{N}\sum_{i=1}^{i=N} x_i y_i = \frac{1}{5} \times 204 = 40{,}8$
- $Cov(x, y) = \overline{xy} - \vec{x}.\bar{y} = 40{,}8 - (3{,}6 \times 9{,}2) = 7{,}68$
- $\hat{a} = \frac{Cov(x, y)}{V(x)} = \frac{7{,}68}{4{,}64} = 1{,}65$
- $\hat{b} = \bar{y} - \hat{a}.\bar{x} = 9{,}2 - 1{,}65 \times 3{,}6 = 3{,}26$

Donc, la droite de régression est :

$$y_i = 1{,}65x_i + 3{,}26$$

Signifie que pour celui qui a eu zéro en test d'intelligence, sa note en statistique est en moyenne égale à 3,26. Un point supplémentaire obtenu en test d'intelligence entraîne une augmentation de 1,65 point de la note en statistique.

Cette droite de régression nous permet d'avoir une estimation de la note en statistique d'un individu ayant obtenu 5 en test d'intelligence. En effet, sa note en statistique est estimée à : $1{,}65 \times 5 + 3{,}26 = 11{,}51$

- $r_{x,y} = \dfrac{Cov(x,y)}{\sigma_x \sigma_y} = \dfrac{7{,}68}{7{,}75} = 0{,}99$

On dit qu'il y a parfaite corrélation linéaire positive entre les deux notes obtenues.

- $R^2 = (r_{x,y})^2 = 0{,}98$
- Décomposition de la variance totale

i	$\hat{y}_i = 1{,}65x_i + 3{,}2$	$y_i - \hat{y}_i$	$(y_i - \hat{y}_i)^2$	$\hat{y}_i - \bar{y}$	$(\hat{y}_i - \bar{y})^2$
1	6,56	-0,56	0,3136	-2,64	6,9696
2	4,91	0,09	0,0081	-4,29	18,4041
3	14,81	0,19	0,036	5,61	31,4721
4	11,51	-0,51	0,2601	2,31	5,3361
5	8,21	0,79	0,6241	-0,99	0,9801
Total			**1,2419**		**63,162**

La variance expliquée est égale à : $\dfrac{1}{N}\sum_{i=1}^{i=N}[\hat{y}_i - \bar{y}]^2 = \dfrac{63{,}162}{5} = 12{,}64$

La variance résiduelle est égale à : $\dfrac{1}{N}\sum_{i=1}^{i=N}[y_i - \hat{y}_i]^2 = \dfrac{1{,}2419}{5} = 0{,}25$

On peut remarquer que La variance totale est égale à la somme de ces deux variances : $\underbrace{VT}_{12{,}96} = \underbrace{VE + VR}_{12{,}89}$

Bibliographie

Bavaud, F. (1998) *Modèles et données: Une introduction à la Statistique uni-, bi- et trivariée*. L'Harmattan, Paris.

Bernard GRAIS (2000), « Techniques statistiques », Tome 1 : Statistique descriptive, Tome 2 : Méthodes statistiques, Editions Dunod, collection Economie.

Bernard PY (1990), « Exercices corrigés de statistique descriptive », 3ème édition Economica.

Calot, G (1975)., « Cours de statistique descriptive », Dunod, Paris,

Droesbeke, J.- J. (1997), « *Éléments de* Statistique », Ellipses, 3ème édition

Goldfarb, B., Pardoux, C. (2000) *Introduction à la méthode statistique*, 3ème édition. Dunod.

J.L. BOURSIN, "Comprendre les statistiques descriptives", A. Colin.

Lévy, M.-L. (1979), *Comprendre les statistiques*, Points Économie

M. Lethielleux (1998) , « Statistique descriptive », Editions Dunod, collection Express.

Reuchlin, M(1991). « Précis de statistique », Paris: PUF, Le Psychologue, (5e éd.).

Rouanet , H., Leroux, B. & Bert, M.-C (1987). « «Statistique en sciences humaines: procédures naturelles », Paris: Dunod,.

Vincent Giard, « Statistique descriptive pour les gestionnaires », Editions Economica.

Wonnacott, T.H., Wonnacott, R.J. (1991) *Statistique*, 4ème édition. Economica.