الجمهورية الجزائرية الديمقراطية الشعبية

République Algérienne Démocratique et Populaire

وزارة التعليم العالي والبحث العلمي

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université Ferhat Abbas / Sétif 1
Faculté des Sciences Économiques, Commerciales et de Gestion
Département d'enseignement fondamental

جامعة فرحات عباس سطيف1

كلية العلوم الاقتصادية، التسيير والتجارة

قسم التعليم الاساسي

Ferhat ABBAS University of Setif

STATISTIQUES 1

Polycopié pédagogique

Présenté par :

BOURIOUNE Tahar

Maitre de Conférences B

Année universitaire : 2023-2024

Table des matières

Liste des tableaux

Liste des figures

INTRODUCTION GENERALE

L'objet de la statistique est la collecte des données, leur traitement puis leur analyse/interprétation.

Trois branches s'en distinguent : la théorie des sondages *(*collecte), la statistique descriptive (traitement) et la statistique mathématique (analyse/interprétation).

La méthodologie de la statistique se résume au recueil des données (choix et taille de la population, choix du caractère, …), leur traitement (classement, présentation, synthèse, …) puis leur analyse & interprétation *(*tests d'hypothèse, *modélisation*, …).

Selon le nombre des variables étudiées, on distingue la statistique univariée, bivariée ou multivariée.

CH1. STATISTIQUE UNIVARIEE

Introduction : vocabulaire statistique

Un ensemble d'objets ou de personnes d'une étude statistique est appelé **population**. Un élément de cette population est appelé **individu**. Les individus d'une population possèdent des caractéristiques communes appelées **caractères.**

L'étude statistique porte sur un caractère. Les **modalités** d'un caractère sont les différents états qu'il peut prendre. Un **caractère** est dit **qualitatif** si ses modalités ne sont pas mesurables (exemple : sexe) : il se subdivise en **ordinal** (modalités hiérarchisables, exemple : bon, passable, médiocre) et en **nominal** (modalités non hiérarchisables, exemple : sexe).

Un **caractère** est **quantitatif** si ses modalités sont mesurables ; les mesures sont alors les valeurs d'une **variable statistique VS** (exemple : un âge, une taille...).

La **VS** est dite **discrète** si elle ne prend que des valeurs isolées (exemple : entières). La **VS** est **continue** si elle peut prendre toutes les valeurs d'un intervalle (exemple : 0 - 4).

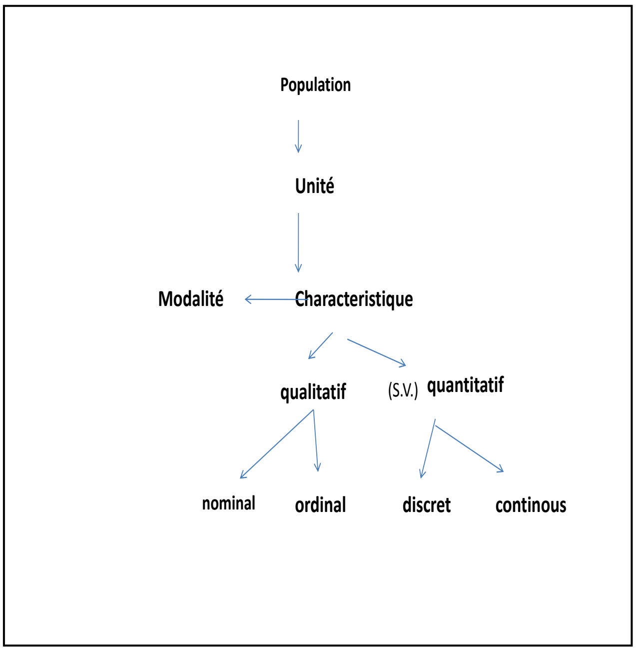
L'**effectif N** d'une population est le nombre total d'individus de cette population. L'effectif **n** est l'effectif de l'échantillon.

L'**effectif (n_i)** de la modalité d'un caractère est le nombre d'individus possédant cette modalité.

La **fréquence (f_i)** de la modalité d'un caractère est le nombre d'individus possédant cette modalité (n_i) divisé par l'effectif total de la population (N) ou l'effectif total de l'échantillon (n).

Population, individu, caractère, modalité, type de caractère (qualitatif, quantitatif), variable statistique vs (vs discrète, vs continue), tableau statistique, représentation graphique, ...constituent l'essentiel du vocabulaire de la statistique descriptive univariée (graphe 1)

Figure 1 vocabulaire statistique



1.1 TABLEAU STATISTIQUE & REPRESENTATION GRAPHIQUE

1.1.1 Tableau statistique (TS)

Le TS diffère selon le type de caractère :

.Caractère Qualitatif

Tableau 1 TF d'un caractère qualitatif

Caractère$_i$	n_i	f_i	$f_i^{\%}$
modalité$_1$	n_1	f_1	$f_1^{\%}$
modalité$_2$	n_2	f_2	$f_2^{\%}$
…			
modalité$_n$	n_n	f_n	$f_n^{\%}$

.VS discrète

Tableau 2 TF d'une vs discrète

x_i	n_i	f_i	$f_i^{\%}$	$n_i\uparrow$	$n_i\downarrow$	$f_i^{\%}\uparrow$	$f_i^{\%}\downarrow$
x_1	n_1	f_1	$f_1^{\%}$	$n_1\uparrow$	$n_1\downarrow$	$f_1^{\%}\uparrow$	$f_1^{\%}\downarrow$
…							
x_n	n_n	f_n	$f_n^{\%}$	$n_n\uparrow$	$n_n\downarrow$	$f_n^{\%}\uparrow$	$f_n^{\%}\downarrow$

.VS continue

Tableau 3 TF d'une vs continue

cl_i	n_i	c_i	f_i	$f_i^{\%}$	$n_i\uparrow$	$n_i\downarrow$	$f_i^{\%}\uparrow$	$f_i^{\%}\downarrow$
$[e_0, e_1[$	n_0	c_o	f_0	$f_0^{\%}$	$n_0\uparrow$	$n_0\downarrow$	$f_0^{\%}\uparrow$	$f_0^{\%}\downarrow$
…….								
$[e_i, e_{i+1}[$	n_i	c_i	f_i	$f_i^{\%}$	$n_i\uparrow$	$n_i\downarrow$	$f_i^{\%}\uparrow$	$f_i^{\%}\downarrow$

1.1.2 Représentation graphique (RG)

La représentation graphique des données relatives à un caractère unique repose sur la **proportionnalité des hauteurs ou des aires** des graphiques **aux effectifs** ou aux fréquences des différentes modalités du caractère.

1.1.2.1 Caractère qualitatif

Pour un caractère qualitatif, on utilise principalement deux types de représentation graphique : la représentation en **tuyaux d'orgue (barres ou colonnes)** et la représentation en **secteurs**. Lorsque le caractère étudié est la répartition géographique d'une population, la représentation graphique est un **cartogramme**.

a) Tuyaux d'orgue

Nous portons en abscisses les modalités, de façon arbitraire. Nous portons en ordonnées des colonnes dont *la **hauteur** est proportionnelle aux effectifs, ou aux fréquences, de chaque modalité (* $h_i / n_i = k i = \overline{1...m}$ *)*

. EXEMPLE

Le tableau suivant nous donne le taux d'analphabétisme (A) pour les garçons et les filles de plus de 10 ans, de 1966 à 2002 en Algérie selon l'ONS.

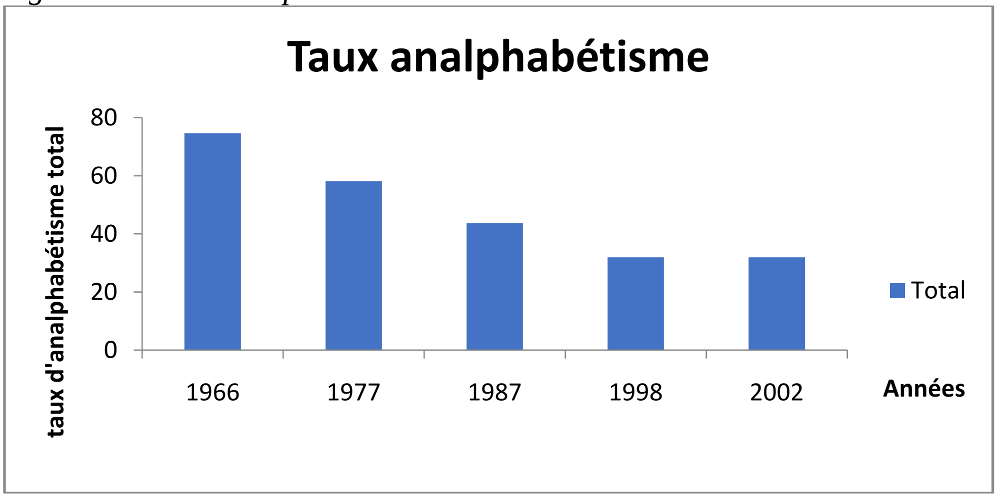
Le caractère étudié, le taux d'analphabétisme, est un caractère qualitatif. La représentation appropriée est le graphe en tuyaux d'orgue ou en secteurs.

Tableau 4 EX.1 caractère qualitatif

Ans	Garçons	Filles	Total
1966	62.30	85.40	74.60
1977	48.20	74.30	58.10
1987	30.75	56.66	43.62
1998	23.65	40.27	31.90
2002	18.20	35.00	26.50

Source : ONS

Figure 2 Taux d'analphabétisme total



b) Secteurs

Les diagrammes circulaires, ou semi-circulaires, consistent à partager un disque ou un demi disque, en secteurs, correspondant aux modalités observées et dont *la **surface** est proportionnelle à l'effectif ou à la fréquence de la modalité :*

$(s_i / n_i = k \quad i = \overline{1...m})$

Ces diagrammes conviennent aux données politiques ou socio-économiques.

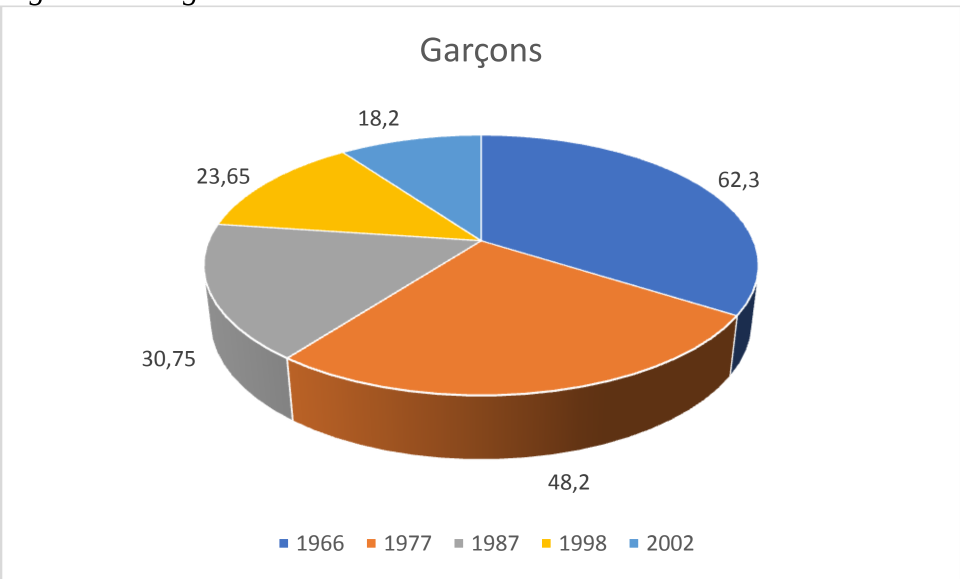
. EXEMPLE

Reprenons la distribution précédente et donnons sa représentation graphique en secteurs pour les garçons.

Tableau 5 EX.2 caractère qualitatif

Ans	Garçons	Filles	Total
1966	62.30	85.40	74.60
1977	48.20	74.30	58.10
1987	30.75	56.66	43.62
1998	23.65	40.27	31.90
2002	18.20	35.00	26.50

Figure 3 Diagramme circulaire



e) Cartogramme

Un cartogramme est une carte géographique dont les secteurs géographiques sont coloriés avec une couleur différente suivant l'effectif ou suivant la fréquence du caractère étudié.

Figure 4 Cartogramme

Source : Youtube

1.1.2.2 Caractère quantitatif

La variable statistique (VS) est la mesure du caractère. Celle-ci peut être discrète (Vd) ou continue (Vc).

Il existe deux types de représentation graphique d'une distribution statistique à caractère quantitatif :

— Le diagramme différentiel correspond à une représentation des effectifs ou des fréquences.

— Le diagramme intégral correspond à une représentation des effectifs cumulés ou des fréquences cumulées.

. VS discrète

— Diagramme différentiel : **diagramme en bâtons** des effectifs ou des fréquences.

La différence avec le cas qualitatif consiste en ce que les abscisses sont ici les valeurs ordonnées de la variable statistique.

Exemple

Soit la distribution suivante, calculer f_i% , Fi et sa représentation graphique?

Tableau 6 EX.3 vs discrète

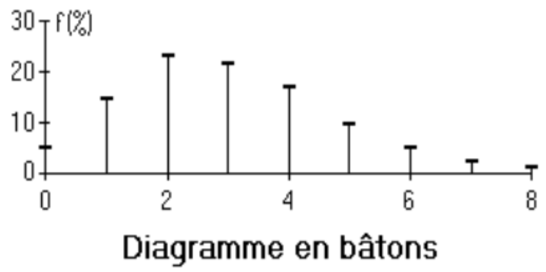
Nombre d'appels téléphoniques par minute	Nombre de minutes
0	93
1	261
2	416
3	393
4	308
5	174
6	93
7	42
8 et plus	20
TOTAL	1 800

. Solution

Tableau 7 Solution EX.3

Nombre d'appels téléphoniques par minute	Nombre de minutes (effectifs)	Fréquence (%)	Fréquence cumulée (%)
0	93	5,2	5,2
1	261	14,5	19,7
2	416	23,1	42,8
3	393	21,8	64,6
4	308	17,1	81,7
5	174	9,7	91,4
6	93	5,2	96,6
7	42	2,3	98,9
8 et plus	20	1,1	100,0
TOTAL	1 800	100,0	

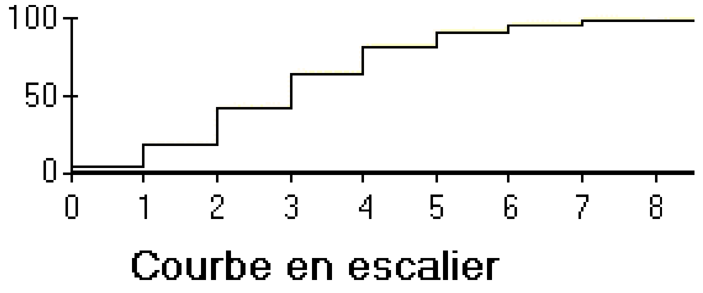
Figure 5 Diagramme en bâtons



---Diagramme intégral : **courbe en escaliers** des effectifs cumulés ou des fréquences cumulées.

La représentation graphique **intégrale** correcte est la ***courbe en escalier***: les fréquences des diverses valeurs de la variable statistique correspondent aux hauteurs des marches de la courbe en escalier.

Figure 6 Courbe en escalier



$f^{\%\uparrow}_{i} = F(x) = \sum_{j=1}^{i} f_j^{\%} \quad x_i \leq X < X_{i+1}$

. VS continue

Les observations sont regroupées en classes.

Chaque classe possède une certaine amplitude (a_i), qui est la longueur de l'intervalle définissant la classe. *Le rapport entre l'effectif d'une classe et son amplitude s'appelle la* ***densité d'effectif (d_i)***.

Le rapport entre la fréquence d'une classe et son amplitude s'appelle la ***densité de fréquence (d_{ir})***.

$di = n_i / a_i \;;\; d_{ir} = f_i / a_i$

— Diagramme différentiel **:** **histogramme des densités**.

Nous portons en abscisse les classes représentant les modalités et en ordonnées des rectangles dont la longueur (hauteur) est proportionnelle à la densité d'effectif ou à la densité de fréquence.

L'**aire** d'un rectangle de cet histogramme est alors proportionnelle à l'effectif ou à la fréquence de la classe.

. EXEMPLE

Soit la distribution suivante, tracer sa représentation graphique ?

Tableau 8 EX.4 vs continue

CA (millions DA)	ni
0 - 0.25	13712
0.25 - 0.50	10674
0.50 - 1.00	11221
1.00 - 2.50	15496
2.50 - 5.00	10043
5.00 - 10.00	3347
10.00-20.00	3147
Total	67640

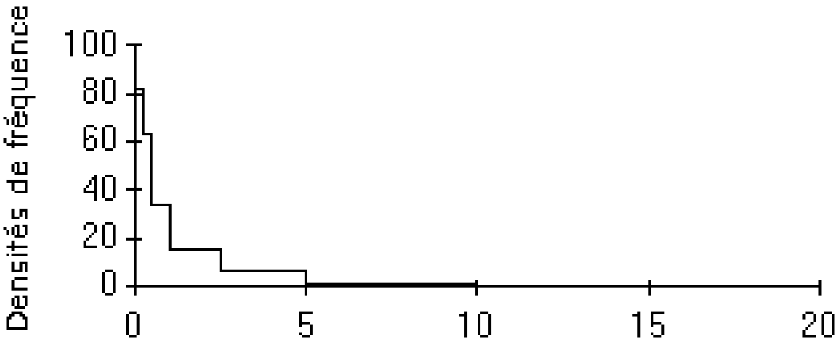
Solution

Tableau 9 Solution1 EX.4

CA (millions DA)	ni	ai	fi	fi%	di
0 - 0.25	13712	0,25	0,203	20,3	81,09
0.25 - 0.50	10674	0,25	0,158	15,8	63,12
0.50 - 1.00	11221	0,5	0,166	16,6	33,18
1.00 - 2.50	15496	1,5	0,229	22,9	15,27
2.50 - 5.00	10043	2,5	0,148	14,8	5,94
5.00 - 10.00	3347	5	0,049	4,9	0,99
10.00-20.00	3147	50	0,047	4,7	0,09
Total	67640		1,000	100,000	

. Solution

Figure 7 Solution Histogramme



— Diagramme intégral : **courbe cumulative** des effectifs ou des fréquences.

La courbe cumulative des fréquences doit représenter la fonction de répartition de la variable statistique. La représentation graphique intégrale correcte est la courbe cumulative des fréquences. Pour que chaque point expérimental représente la fonction de répartition, il faut prendre pour abscisses les ***<u>bornes supérieures des classes</u>*** et, pour ordonnées, les fréquences cumulées correspondantes.

. EXEMPLE

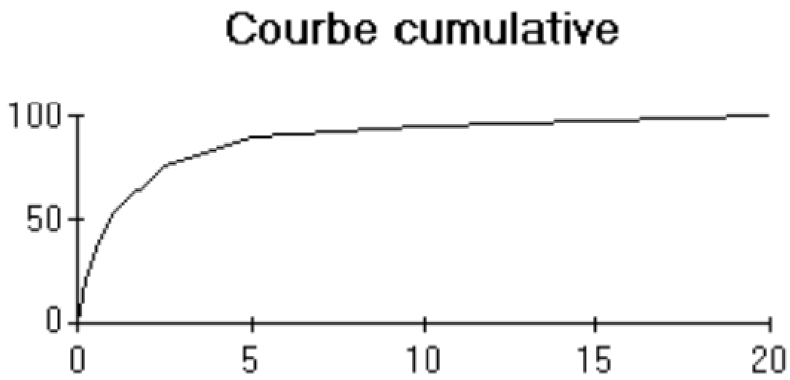
Soit la distribution précédente, tracer la représentation graphique des effectifs cumulées Fi ?

. Solution

Tableau 10 Solution2 EX.4

CA (millions DA)	ni	ai	fi	fi%	di	Fi
0 - 0.25	13712	0,25	0,203	20,3	81,09	20,3
0.25 - 0.50	10674	0,25	0,158	15,8	63,12	36,1
0.50 - 1.00	11221	0,5	0,166	16,6	33,18	52,7
1.00 - 2.50	15496	1,5	0,229	22,9	15,27	75,6
2.50 - 5.00	10043	2,5	0,148	14,8	5,94	90,4
5.00 - 10.00	3347	5	0,049	4,9	0,99	95,4
10.00-20.00	3147	50	0,047	4,7	0,09	100,0
Total	67640		1,000	100		

Figure 8 Courbe cumulative



1.2 PARAMETRES CARACTERISTIQUES

Le but de l'étude statistique est aussi de résumer les données par des paramètres ou synthétiseurs.

Il existe 3 types de paramètres :

— *paramètres de position (ou de tendance centrale)*

— *paramètres de dispersion*

— *paramètres de forme (asymétrie, aplatissement, concentration)*

1.2.1 PARAMETRES DE POSITION

Les paramètres de position (mode, médiane, moyenne) permettent de savoir autour de quelles valeurs se situent les valeurs d'une variable statistique.

1.2.1.1 Mode (M_o)

Le mode, noté *Mo*, est la modalité qui admet **la plus grande fréquence corrigée**: $M_o \Leftrightarrow f'(Mo) = Max\,(f'i)\ ;\ i \in [\,1, p\,]$

Il est parfaitement défini pour une variable qualitative ou une variable quantitative discrète.

. Détermination algébrique

. VS discrète

i. Série numérique

Soit la série X=2, 4, 1, 3, 6, 7, 6, 8, 9, 7

-on ordonne la série : X=1, 2, 3, 4, 6, 6, 7,7, 8, 9

-on calcule le mode : Mo=modalité qui a le plus grand effectif (ni) :

$M_{1o}=6; M_{2o}=7$

ii. Série classée

Le mode Mo d'un tableau des effectifs est le caractère qui a le plus grand effectif (ni). $M_o \Leftrightarrow n\,(M_o) = Max\,(n_i)\ ;\ i \in [\,1, p\,]$

. EXEMPLE : Soit la variable Xi suivante, calculer Mo ?

Tableau 11 EX.5 Mode

Xi	ni
5	2
10	3
20	10
8	2

$M_o = 20$

. VS continue

Pour une variable quantitative continue nous parlerons de **classe modale** : c'est la classe dont la *densité de fréquence* (ou l'effectif corrigé) est maximum :

$cl_o \Leftrightarrow \mathrm{dir(Clo)} = (d_{ir})_{max};\ i \in [\,1, p\,]$

Une VS peut présenter plusieurs modes locaux : on dit alors qu'elle est **plurimodale**.

Si les classes ont même amplitude, la densité est remplacée par l'effectif ou la fréquence et nous retrouvons la définition précédente.

. EXEMPLE : Soit la Cl_i suivante : calculer Mo ?

Tableau 12 EX.6 Mode

cl_i	fi%
[0-10[	10
[10-20[	20
[20-40[	30
[40-45[	40

. *Méthode de la densité d_{ir}*

-on calcule d'abord la classe modale cl_o

Tableau 13 Solution1 EX.6

cl_i	fi%	ai	dir
[0-10[	10	10	1
[10-20[	20	10	2
[20-40[	30	20	1.5
[40-45[	40	5	8

Clo↔ dir(Clo)=dir_{max}=[40-45[

-on calcule ensuite le mode par interpolation linéaire:

M_o =40+(45-40) (8-1.5)/(8-1.5)+(8-0)=42

. *Méthode du coefficient*

-on calcule d'abord la classe modale Clo : c'est la classe ayant la plus grande fréquence corrigé.

Notons : $k_i = a_i / a$ et $f'_i = f_i / k_i$

Tableau 14 Solution2 EX.6

cl_i	fi%	ai	k_i=ai/a	f '%
[0-10[	10	10	2	5
[10-20[	20	10	2	10
[20-40[	30	20	4	7.5
[40-45[	40	5	1	40

Clo $\leftrightarrow$f'$_i$%(Clo)=f'%$_{max}$=[40-45[

-on calcule ensuite le mode par interpolation linéaire :

Mo =40+(45-40) (40-7.5) / (40-7.5) +(40-0) =42

. Détermination graphique

. VS discrète

i. Série classée

Soit la série Xi suivante : représenter Mo ?

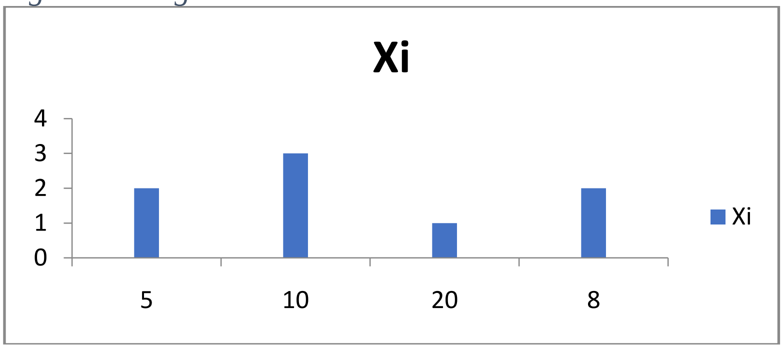
Tableau 15 EX.7 Mode

Xi	ni
5	2
10	3
20	1
8	2

-On représente d'abord le graphe de Xi (graphe en bâtons)

-On marque ensuite le bâton de plus grande hauteur comme étant Mo.

Figure 9 Diagramme en bâtons

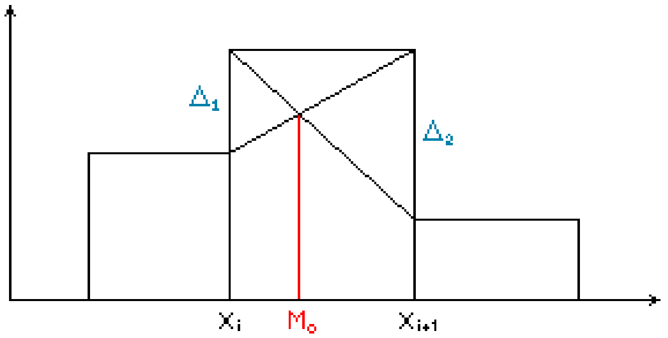


Mo=10

. VS continue

Nous définissons le **mode**, pour une variable quantitative continue, en tenant compte des densités de fréquence (ou des effectifs corrigés) des 2 classes adjacentes à la classe modale par la méthode suivante.

Figure 10 Graphe Mode : cas continu



La classe modale [xi, xi+ 1 [étant déterminée, le mode *Mo* vérifie :

$$\frac{M_0 - x_i}{x_{i+1} - x_i} = \frac{\Delta_1}{\Delta_1 + \Delta_2} \Rightarrow Mo = xi + (x_{i+1} - x_i)\frac{\Delta_1}{\Delta_1 + \Delta_2} \quad / \ \Delta_1 = n'_0 - n'_{0-1} \quad \Delta_2 = n'_0 - n'_{0+1}$$

Remarque :

Lorsque les classes adjacentes à la classe modale ont des densités de fréquences égales, le mode coïncide avec le centre de la classe modale.

Le mode dépend beaucoup de la répartition en classes.

Une variable statistique peut présenter plusieurs modes locaux : on dit alors qu'elle est **plurimodale**.

Cette situation est intéressante : elle met en évidence l'existence de plusieurs sous-populations, donc l'hétérogénéité de la population étudiée.

. EXEMPLE

Soit le tableau de la distribution précédente, calculer Mo graphiquement ?

Tableau 16 EX.8 Mode

cl_i	fi%	ai	k_i	f'
[0-10[	10	10	2	5
[10-20[	20	10	2	10
[20-40[	30	20	4	7.5
[40-45[	40	5	1	40

. Solution

1. On trace l'histogramme

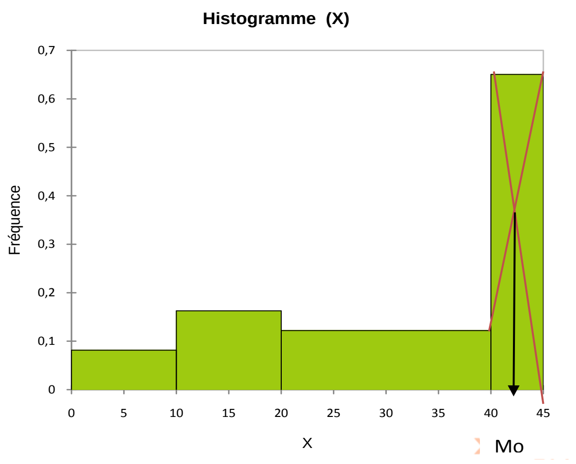
2. On détermine la classe modale : Clo $\leftrightarrow f'_i\%(Clo)=f'\%_{max}$=[40-45[

3. On calcule ensuite le mode Mo conformément à la règle des « ciseaux » et au théorème de THALES.

$$Mo = xi + (x_{i+1} - x_i)\frac{\Delta_1}{\Delta_1 + \Delta_2}$$

Mo =40+(45-40) (40-7.5) / (40-7.5) +(40-0) =42

Figure 11 Histogramme de la distribution



Mo

1.2.1.2 Médiane (Me)

La médiane *Me* est la modalité telle que l'effectif des observations dont les modalités sont inférieures à *Me* est égal à l'effectif des observations dont les modalités sont supérieures à *Me*.

Cette définition n'a de sens que si les modalités sont toutes ordonnées :

$M_e \Leftrightarrow F(M_e)= 0,5$

. Détermination algébrique

· VS discrète

i. série numérique

Soit la série X=2, 4, 1, 3, 6, 7, 6, 8, 9, 7

-on ordonne la série : X=1, 2, 3, 4, 6, 6, 7,7, 8, 9

- on examine la parité de la suite :

.si la suite possède un nombre impair de termes n=(2k+1);k∈ N: $X_e = \frac{X_{n+1}}{2}$

.si la suite possède un nombre pair de termes n=(2k);k∈ N:$X_e=\frac{X_{n/2}+X_{n/2}+1}{2}$

Dans notre exemple n=10 donc $X_e=\frac{X_{n/2}+X_{n/2+1}}{2}=\frac{X_5+X_6}{2}=6$

ii. série classée (tableau de fréquence)

Si le nombre d'observations N est pair, Me $=(X_1+X_2)/2$: $X_1 \leftrightarrow n^{\uparrow}(X_1) \geq n/2$

$X_2 \leftrightarrow n^{\uparrow}(X_2) \geq n/_2 +1$

. EXEMPLE

Soit le tableau des fréquences (TF) suivant, calculer Me ?

Tableau 17 EX.9 Médiane

Xi	ni	ni↑
5	2	2
10	3	5
20	1	6
8	2	8

8

N =8 *X1↔n(X1)≥n/2=10* *X2↔n(X2)≥n/2 +1= 10*

Donc *Me =(X1+X2)/2 =10*

·VS continue

-On calcule d'abord l'intervalle médian $\left]a\ b\right[$: Cle ↔ F(Cle)≥50

-On calcule ensuite par interpolation linéaire Xe= a+(b-a) $\frac{(50-n_{e-1}^{\uparrow}}{n_e}$

. EXEMPLE

Reprenons la distribution du tableau 12, calculer Me ?

Tableau 18 EX.10 Médiane

cl_i	fi%
[0-10[	10
[10-20[	20
[20-40[	30
[40-45[	40

. Solution

Tableau 19 Solution EX.10

cl_i	fi%	Fi
[0-10[	10	10
[10-20[	20	30
[20-40[	30	60
[40-45[	40	100

Cle ↔ F(Cle) ≥50 =[20-40[

Me=20+20 (50-30)/30=33.33

L'intervalle [20-40[est l'intervalle médian.

Dans l'intervalle médian, la médiane est calculée par interpolation linéaire : Xe==a+(b-a). $\frac{(50-n^{\uparrow}_{e-1}}{n_e}$ =33.33

x_e=a+(b-a). $\frac{rg(n/2)}{n_e}$ / rg(n/2)= n/2 - n $^{\uparrow}$ a

Remarque

La médiane ne dépend que de l'ordre des modalités, elle n'est donc pas influencée par les observations aberrantes.

La médiane partage l'histogramme des fréquences en 2 parties d'aires égales.

. *Détermination graphique*

. VS discrète

i. Série classée

. On trace le graphe des fréquences (effectifs) cumulées ascendants (courbe en

escalier)

. On détermine Me ↔ n(Me)≥N/2 ou F(Me)≥50

. EXEMPLE

Soit le TF suivant d'une VS discrète, calculer Me graphiquement ?

Tableau 20 EX.11 Médiane

Xi	fi%
1	10
2	20
3	30
4	10
6	20
8	10

. Solution

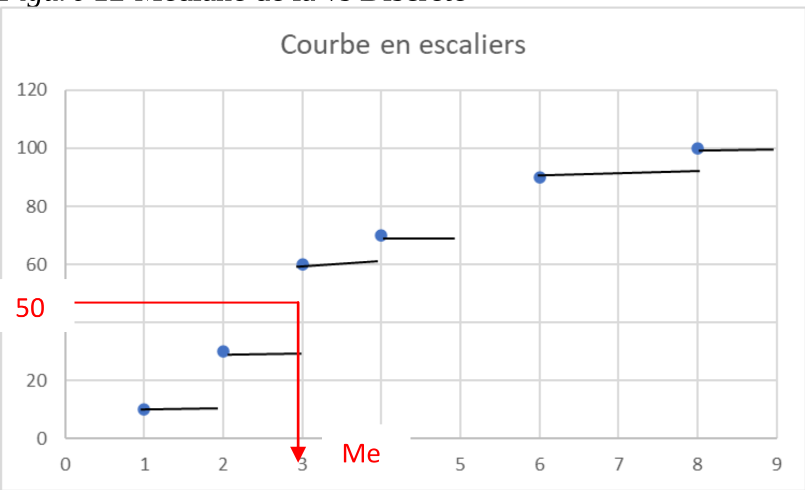
. On calcule d'abord Fi

Tableau 21 Solution EX.11

Xi	fi%	Fi
1	10	10
2	20	30
3	30	60
4	10	70
6	20	90
8	10	100

. On détermine ensuite Me=F(Me) ≥50 =3 , on le fixe sur le graphe de Fi.

Figure 12 Médiane de la vs Discrète



VS continue

. On trace le graphe des fréquences (effectifs) cumulées ascendants

. On fixe la classe médiane : Cle↔ F(Cle)≥50

. On calcule Me par la méthode de « l'égalité des pentes »

. EXEMPLE

Soit la distribution suivante, calculer Me ?

Tableau 22 EX.12 Médiane

cl_i	fi%
[0-10[	10
[10-20[	20
[20-40[	30
[40-45[	40

Solution

Tableau 23 Solution1 EX.12

cl_i	fi%	Fi
[0-10[	10	10
[10-20[	20	30
[20-40[	30	60
[40-45[	40	100

. Cle ↔ F(Cle)≥50 =[20-40[

Par la méthode de l'égalité des pentes, nous obtenons :

. (F(b)-F(a)))/(b—a)=(50-F(a))/(Me-20)

. (Me-20)/(b—a)=(50-F(a))/(F(b)-F(a))

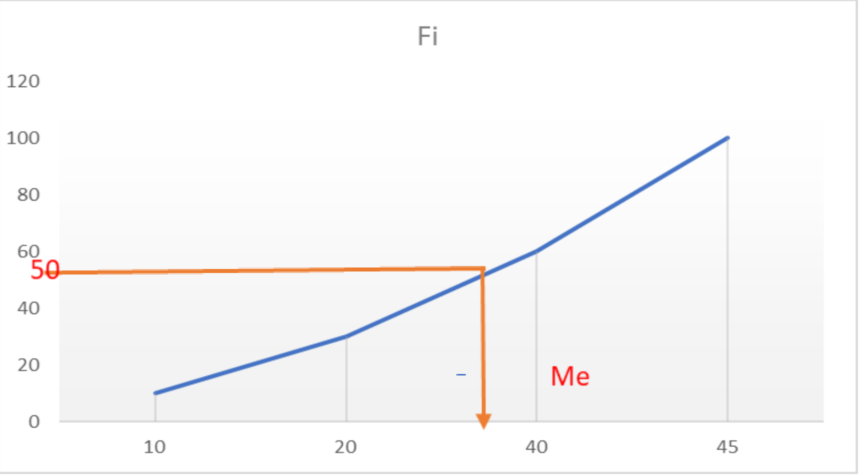
Par conséquent on déduit Me :

Me=a+(b-a)(50-F(a))/(F(b)-F(a))=20+20.(50-30)/30=33.33

Tableau 24 Solution2 EX.12

cl_i	ei	ei+1	fi%	Fi	Ci
[0-10[	0	10	10	10	5
[10-20[	10	20	20	30	15
[20-40[	20	40	30	60	30
[40-45[	40	45	40	100	42,5

Figure 13 Médiane de la vs continue



1.2.1.3 Moyenne ($\overline{X}$)

Tableau 25 Moyenne Arithmétique simple et pondérée

	Vs discrète	**Vs continue**
Moyenne arith. simple	$\frac{1}{n}\sum_i x_i$	$\frac{1}{n}\sum_i c_i$
Moyenne arith. pondérée	$\frac{1}{n}\sum_i n_i x_i$	$\frac{1}{n}\sum_i n_i c_i$

Propriétés de la moyenne

$$\overline{x+y} = \bar{x} + \bar{y}$$

$$\overline{\lambda x} = \lambda \bar{x}$$

$$\overline{x - \bar{x}} = 0 \quad \overline{\left(x - \bar{x}\right)}^2 < \overline{\left(x - x_\alpha\right)}^2 \qquad x_\alpha \neq \bar{x}$$

. Détermination algébrique

. VS discrète

i. Série numérique

On calcule ici la moyenne par la méthode de la moyenne simple.

. EXEMPLE

Soit la série X_i suivante, calculer la moyenne ?

$$Xi = 1,2,3,4,5,6,7,8,9,10$$

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{10} Xi = \frac{(1+2+\ldots\ldots+10)}{10} = 5.5$$

ii. Série classée

On calcule ici la moyenne arithmétique par la méthode de « la moyenne pondérée ».

. EXEMPLE

Reprenons le tableau 15 précédent, calculer la moyenne ?

Tableau 26 EX.13 Moyenne

X_i	n_i
1	1
2	8
3	12
4	14
5	6
6	4
7	3
8	2
	50

. Solution

Tableau 27 Solution EX.13

X_i	n_i	f_i	f_iX_i
1	1	0.02	0.02
2	8	0.16	0.32
3	12	0.24	0.72
4	14	0.28	1.12
5	6	0.12	0.6
6	4	0.08	0.48
7	3	0.06	0.42
8	2	0.04	0.32
	50	1	4

$$\overline{X} = f_i \ X_i = 4$$

. Changement de variable :

$x=a\,x'+b \Rightarrow x'=(x-b)/a$; $\overline{x} = a\,\overline{x'} + b$ $v(x)= a^2\, v(x')$; $b=x_0\,(c_0)$

[centre de classe médiane]

VSdiscrète

Tableau 28 Changement de variable d'une vs discrète

x_i	n_i	$x'i=(x_i-x_0)/a$	$n_ix'_i$	$n_ix'^2_i$

VS continue

Tableau 29 Changement de variable d'une vs continue

cl_i	n_i	c_i	$c'_i=(c_i-c_0)/a$	$n_i c'_i$	$n_i c'^2_i$
$[e_0,e_1[$					
$[e_1,e_2[$					
$[e_i,e_{i+1}[$					

. Moyenne conditionnée :. φ **-moyenne**

Soit $X = \{(xi, ni)\}$, $i \in$ [1,p], une v.s. qtitive discrète à valeurs dans R+

Soit φ: : R+* $\rightarrow$R une application monotone (injection croissante ou décroissante) continue.

Alors φ $(X) = \{(\varphi(xi), ni)\}$, $i \in$ [1,p], est une v. s. quantitative discrète à valeurs dans R.

On peut calculer sa moyenne $\overline{\varphi(X)} = \frac{1}{n}\sum_i n_i\varphi(x_i)$.

Comme φ est une injection continue, il existe un unique $\overline{\varphi}_\Psi \in \mathbb{R}_+^*$ tel que $\overline{\varphi(x)} = \varphi(\overline{x_\Psi})$

$\overline{x}_\Psi$ = est appelé la φ **-moyenne** de X.

. EXEMPLES de φ **-moyennes.**

1. $\varphi(X) = X^r$ $\overline{\varphi(X)} = \overline{X^r} = \frac{1}{n}\sum_i n_i X^r = \overline{X}^r \Rightarrow \overline{X}_r = (\frac{1}{n}\sum_i n_i X^r)^{\frac{1}{r}}$

r=1 $\overline{X}a = \frac{1}{n}\sum_i n_i X^r$

r=2 $\overline{X}q = \sqrt{\frac{1}{n}\sum_i n_i X^2}$

r=-1 $\overline{X}h = \frac{n}{\sum_i \frac{n_i}{x_i}}$

2. $\varphi(X) = \ln X$; $\overline{\varphi(X)} = \overline{\ln X} = \frac{1}{n}\sum_i n_i \ln X_i = \ln(\overline{X}) \Rightarrow \overline{X_g} = e(\frac{1}{n}\sum_i n_i \ln X_i) = [\prod_i x_i^{ni}]^{\frac{1}{N}}$

Tableau 30 Les moyennes de grandeurs simples et pondérées

	simples	pondérées
$\overline{x}$	$\frac{1}{n}\sum_i x_i$	$\frac{1}{n}\sum_i n_i x_i$
$\overline{x_h}$	$\frac{n}{\sum_i \frac{1}{x_i}}$	$\frac{n}{\sum_i \frac{n_i}{x_i}}$
$\overline{X_g}$	$(\prod_i x_i]^{\frac{1}{N}}$	$(\prod_i x_i^{ni}]^{\frac{1}{N}}$
$\overline{X}q =$	$\sqrt{\frac{1}{n}\sum_i X^2}$	$\sqrt{\frac{1}{n}\sum_i n_i X^2}$

$\overline{X}_h < \overline{X_g} < \overline{X} < \overline{X}_q$

Il y a égalité si, et seulement si, toutes les valeurs de *X* sont égales.

$\overline{X_g}$ est bien adaptée à l'étude des phénomènes de croissance.

$\overline{x_h}$ est utilisée pour les calculs d'indices économiques.

$\overline{X}q$ est utilisée en statistique pour le calcul de l'écart type.

. EXEMPLE

i. Série numérique

Soit la série X_i *suivante, calculer* $\overline{X}_h$, $\overline{X}g$, $\overline{X}_a$ et $\overline{X}_q$ puis confirmer la relation théorique qui existe entre elles?

X_i =12, 4, 3, 8, 7, 2

. Solution

Tableau 31 Solution moyenne composée

Xi	1/Xi	Xi²	Xi	log Xi
12	0.083	144	12	1.079
4	0.250	16	4	0.602
3	0.333	9	3	0.477
8	0.125	64	8	0.903
7	0.143	49	7	0.845
2	0.500	4	2	0.301
36	1.435	286	36	4.208

$$\overline{X}h = \frac{n}{\sum_i \frac{1}{X_i}} = \frac{6}{1.435} = 4.18$$

$$\overline{X}g = \left[\prod_i X_i\right]^{\frac{1}{n}} = \sqrt[6]{16128} = 5.03$$

ou

$$\overline{X}g = 10^{\left(\frac{1}{n}\sum \log X_i\right)} = 10^{\left(\frac{4.207}{6}\right)} = 5.03$$

$$\overline{X}a = \frac{1}{n}\sum_i Xi = \frac{36}{6} = 6$$

$$\overline{X}q = \sqrt{\frac{1}{n}\sum_i X_i^2} = \sqrt{\frac{286}{6}} = 6.90$$

La relation théorique $\overline{X}h \leq \overline{X}g \leq \overline{X}a \leq \overline{X}q$ est confirmée.

ii. Série classée

Soit la distribution suivante, calculer les moyennes $\overline{X}hp$, $\overline{X}gp$, $\overline{X}ap$ et $\overline{X}qp$

puis confirmer la relation théorique qui existe entre elles ?

Tableau 32 EX.14 Moyenne composée

Xi	ni
2	2
8	3
17	1
12	3
3	2

. Solution

Tableau 33 Solution EX.14

Xi	ni	ni Xi	ni Xi²	ni/Xi	log Xi	n logXi
2	2	4	8	1	0.30	0.60
8	3	24	192	0.38	0.90	2.71
17	1	17	289	0.06	1.23	1.23
12	3	36	432	0.25	1.08	3.24
3	2	6	18	0.67	0.48	0.95
	11	87	939	2.35		8.73

$$\overline{X}hp = \frac{\sum_i ni}{\sum_i \frac{ni}{Ci}} = \frac{11}{2.35} = 4.68$$

$$\overline{X}gp = \left[\prod_i X_i^{ni}\right]^{\frac{1}{\sum_i ni}} = (541458432)^{1/11} = 6.22$$

ou

$$\overline{X}gp = 10^{\left(\sum_i \frac{ni \log Ci}{\sum_i ni}\right)} = 10^{\left(\frac{8.73}{11}\right)} = 10^{(0.794)} = 6.22$$

$$\overline{X}ap = \frac{\sum_i ni\ Ci}{\sum_i ni} = \frac{87}{11} = 7.90$$

$$\overline{X}qp = \left(\frac{\sum_i ni\ C_i^2}{\sum_i ni}\right)^{1/2} = \left(\frac{939}{11}\right)^{1/2} = \sqrt{85.36} = 9.24$$

La relation théorique $\overline{X}hp \leq \overline{X}gp \leq \overline{X}ap \leq \overline{X}qp$ est confirmée.

. VS Continue

Soit la distribution de 20 exploitations agricoles selon leur superficie (en hectares), *calculer* les moyennes pondérées $\overline{X}hp$, $\overline{X}gp$, $\overline{X}ap$ et $\overline{X}qp$ puis confirmer la relation théorique qui existe entre elles?

Tableau 34 EX.15 Moyenne composée

Cli	ni
0-5	**3**
5-10	12
10-20	4
20-50	1
	20

. Solution

Tableau 35 Solution EX.15

Cli	ni	Ci	ni / Ci	ni C²i	ni Ci	ni log Ci
0-5	3	2.5	1.20	18.75	7.50	1.19
5-10	12	7.5	1.60	675.00	90.00	10.50
10-20	4	15	0.27	900.00	60.00	4.70
20-50	1	35	0.03	1225.00	35.00	1.54
	20		3.10	2818.75	192.50	17.94

$$\overline{\mathrm{X}}\mathrm{hp} = \frac{\sum_i ni}{\sum_i \frac{ni}{Ci}} = \frac{20}{3.10} = 6.45$$

$$\overline{\mathrm{X}}\mathrm{g}p = \left[\prod_i X_i^{ni}\right]^{\frac{1}{\sum_i ni}} = \left(8.76977E+17\right)^{1/20} = 7.89$$

ou

$$\overline{\mathrm{X}}\mathrm{gp} = 10^{\left(\sum_i \frac{ni \text{ log Ci}}{\sum_i ni}\right)} = 10^{\left(\frac{17.94}{20}\right)} = 10^{(0.897)} = 7.89$$

$$\overline{X}_{ap} = \frac{\sum_i ni\ Ci}{\sum_i ni} = \frac{192.5}{20} = 9.625$$

$$\overline{X}_{qp} = \left(\frac{\sum_i ni\ C_i^2}{\sum_i ni} \right)^{1/2} = \left(\frac{2818.75}{20} \right)^{1/2} = \sqrt{140.9375} = 11.87$$

La relation théorique $\overline{X}_{hp} \leq \overline{X}_{gp} \leq \overline{X}_{ap} \leq \overline{X}_{qp}$ est confirmée.

1.2.2 PARAMETRES DE DISPERSION

1.2.2.1 Etendue

L'étendue de X : $E = x_{max} - x_{min}$

Ce paramètre est souvent utilisé dans les contrôles de fabrication, pour lesquels on donne, a priori, des marges de construction. Son intérêt est limité par le fait qu'il dépend uniquement des valeurs extrêmes, qui peuvent être des valeurs aberrantes.

. EXEMPLE

. VS discrète

. Série numérique

Soit la série numérique Xi = 1, 3, 2, 3, 3, 2, 4 . Calculer l’étendue E ?

Réponse

E=4-1= 3

. Série classée

Soit le TF ci-dessous, calculer l'étendue E ?

Tableau 36 EX.16 Etendue

Xi	ni
1	1
2	2
3	3
4	1

. Réponse

E= 4-3=2

. VS continue

Soit la Cli ci-dessous, calculer l'étendue E ?

Tableau 37 EX.17 Etendue

Cli	ni
0-5	3
5-10	12
10-20	4
20-50	1
	20

. Réponse

Tableau 38 Solution EX.17

Cli	ni	Ci
0-5	3	2.5
5-10	12	7.5
10-20	4	15
20-50	1	35
	20	

E= $C_{MAX} - C_{MIN}$ =35 – 7.5 = 27.5

1.2.2.2 Quartiles

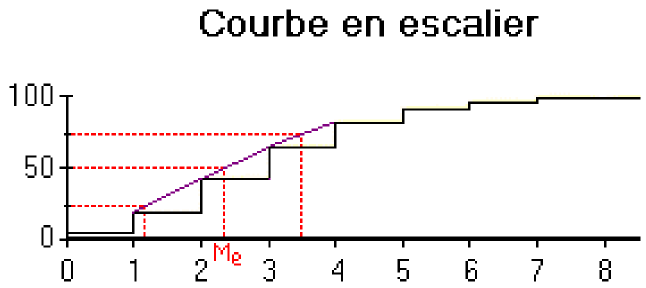
. Détermination algébrique

. VS discrète

Pour une variable statistique réelle discrète *X*, la courbe des fréquences cumulées est une **courbe en escalier**. S'il existe une valeur de *x* pour laquelle la fréquence cumulée est 0,25 (resp. 0,50, 0,75), le quartile correspondant est cette valeur de *X*.

Sinon, les quartiles seront déterminés par **interpolation linéaire** entre deux valeurs.

Figure 14 Détermination graphique des quartiles. cas discret



i. Série numérique

. EXEMPLE

Soit la série numérique Xi=2, 4, 1, 3, 6, 7, 6, 8, 9,7. Calculer Q_1, Q_3 ?

. Réponse

-On ordonne la série : 1,2,3,4,5,6,6,7,7,8,9

$$- Q_1 \Leftrightarrow X_{k \geq \frac{N}{4}} \quad /K \in N \quad \Leftrightarrow X_{k \geq \frac{10}{4}} = X_3 = 3$$

$$- Q_3 \Leftrightarrow X_{k \geq \frac{3N}{4}} \quad /K \in N \quad \Leftrightarrow X_{k \geq \frac{30}{4}} = X_8 = 7$$

ii. Série classée

Soit le TF suivant, calculer Q_1, Q_3 ?

Tableau 39 EX.18 Quartiles

X	n_i
1	1
2	8
3	12
4	14
5	6
6	4
7	3
8	2
	50

. Réponse

Tableau 40 Solution EX.18

X	n_i	$ni\uparrow$
1	1	1
2	8	9
3	12	21
4	14	35
5	6	41
6	4	45
7	3	48
8	2	50
	50	

$$Q_1 \Leftrightarrow n^{\uparrow}\left(Q_1\right) \geq \frac{N}{4} = 21$$

$$Q_1 \quad = 3$$

$$Q_3 \Leftrightarrow n^{\uparrow}\left(Q_3\right) \geq \frac{3N}{4} = 41$$

$$Q_3 \quad = 5$$

. VS continue

On appelle **quartiles** les nombres réels Q_1, Q_2, Q_3, pour lesquels les fréquences cumulées de X sont respectivement 0,25, 0,50, 0,75. Les quartiles partagent l'étendue en quatre intervalles qui ont le même effectif. Le deuxième quartile, Q2, est égal à la médiane.

Soit la distribution de la vs continue, calculer Q_1, Q_3 ?

Tableau 41 EX.19 Quartiles

cl_i	fi%
[0-10[	10
[10-20[	40
[20-30[	40
[30-40[	10

. Solution

Tableau 42 Solution EX.19

cl_i	fi%	Fi
[0-10[	10	10
[10-20[	40	50
[20-30[	40	90
[30-40[	10	100

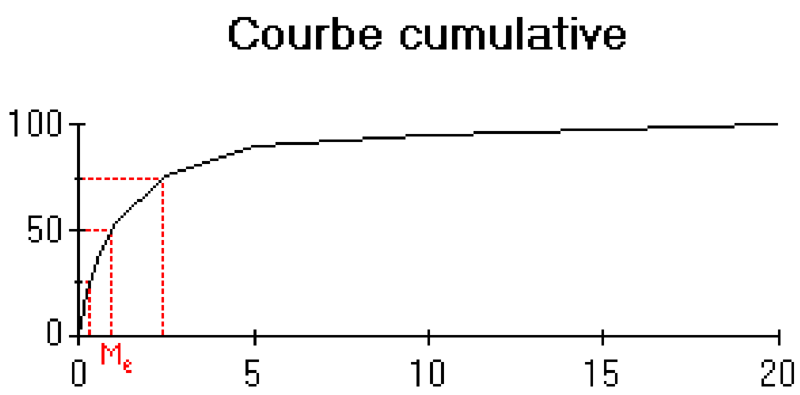
– On calcule Cl (Q1) d'abord: Cl(Q1) $\Leftrightarrow F(Cl(Q1) \geq 25 \;\Rightarrow$ Cl(Q1)=$[10, 20[$

–*On* calcule ensuite Q1: Q1=a+(b-a). $\dfrac{\text{25-F(a)}}{f\%(Q1)} = 10 + 10.\dfrac{25-10}{40} = 13.75$

– On calcule Cl (Q_3) d'abord: Cl(Q3) $\Leftrightarrow F(Cl(Q3) \geq 75 \;\Rightarrow$ Cl(Q1)=$[20,30[$

–*On* calcule ensuite Q_3: Q_3=a+(b-a). $\dfrac{\text{75-F(a)}}{f\%(Q3)} = 20 + 10.\dfrac{75-50}{40} = 26.25$

Figure 15 détermination graphique des quartiles: cas continu



1.2.2.3 Intervalle interquartile

$I_Q = Q3 - Q1$ est l'intervalle [$Q1$, $Q3$]. Il contient 50 % des valeurs de X.

. Détermination algébrique

.VS discrète

. EXEMPLE

Calculer I_Q sur la base du TF précédent (Tableau 39):

$$Q_3 \Leftrightarrow n^{\uparrow}(Q_3) \geq \frac{3N}{4} = 41$$

$$Q_3 \quad = 5$$

$$Q_1 \Leftrightarrow n^{\uparrow}(Q_1) \geq \frac{N}{4} = 21$$

$$Q_1 \quad = 3$$

On en déduit qu'IQ = 5 –3 = 2.

VS continue

Calculer IQ sur la base de la distribution de la VS continue précédente (Tableau 41):

– On calcule Cl (Q_3) d'abord: Cl(Q3) $\Leftrightarrow F(Cl(Q3) \geq 75 \;\Rightarrow$ Cl(Q1)=$[20,30[$

–*On* calcule ensuite Q_3: Q_3=a+(b-a). $\dfrac{75\text{-F(a)}}{f\%(Q3)} = 20 + 10.\dfrac{75-50}{40} = 26.25$

– On calcule Cl (Q1) d'abord: Cl(Q1) $\Leftrightarrow F(Cl(Q1) \geq 25 \;\Rightarrow$ Cl(Q1)=$[10,20[$

–*On* calcule ensuite Q1: Q1=a+(b-a). $\dfrac{25\text{-F(a)}}{f\%(Q1)} = 10 + 10.\dfrac{25-10}{40} = 13.75$

On en déduit que $I_Q = Q_3 - Q_1$ = 26.25 -13.75 = 12.50

1.2.2.4 Déciles et percentiles

Les 9 déciles sont les nombres réels qui partagent l'étendue en dix intervalles de même effectif. Utilisation : en matière de salaires, le rapport interdécile est un paramètre de dispersion fréquemment utilisé.

Les 99 percentiles sont les nombres réels qui partagent l'étendue en cent intervalles de même effectif.

. Détermination algébrique

. VS discrète

. Série numérique

Soit Xi= 12, 8, 7, 7, 5, 4, 4, 3, 2, 1, 0. Calculer D_3, C_{12} ?

. Solution

. On ordonne d'abord la série : Xi=0, 1, 2, 3, 4, 4, 5, 7, 7, 8, 12.

. On calcule D_3 :

$$D_3 = X_{k \geq \frac{3N}{10} = 4} \quad k \in \mathrm{N} \quad \Rightarrow \quad D_3 = X_4 = 3$$

$$C_{12} = X_{k \geq \frac{12N}{100} = 2} \quad k \in \mathrm{N} \quad \Rightarrow C_{12} = X_2 = 1$$

. Série classée

Reprenons la distribution précédente de vs discrète et calculons D_2 et C_{67} ?

Tableau 43 EX.20 Déciles

X	n_i
1	1
2	8
3	12
4	14
5	6
6	4
7	3
8	2
	50

. Réponse

Tableau 44 Solution EX.20

X	n_i	$n_i \uparrow$
1	1	1
2	8	9
3	12	21
4	14	35
5	6	41
6	4	45
7	3	48
8	2	50
	50	

$$D_2 \Leftrightarrow n^{\uparrow}(D_2) \geq \frac{2N}{10} = 21$$

$$D_2 = 3$$

$$C_{67} \Leftrightarrow n^{\uparrow}(C_{67}) \geq \frac{67N}{100} = 35$$

$$C_{67} = 4$$

. VS continue

Soit la distribution de la vs continue ci-après, calculer D_6 et C_{80} ?

Tableau 45 EX.21 Déciles

cl_i	fi%
[0-10[	10
[10-20[	30
[20-40[	40
[40-50[	20

. Réponse

Tableau 46 Solution EX.21

cl_i	fi%	Fi
[0-10[	10	10
[10-20[	30	40
[20-40[	40	80
[40-50[	20	100

D_6= ?

$-On\ calcule\ d'abord\ Cl\left(D_6\right) : Cl\left(D_6\right) \Leftrightarrow F\left[Cl\left(D6\right)\right] \geq 60 \Rightarrow Cl(D6) = [20\ 40[$

$-On$ calcule ensuite D_6: $D_6 = 20 + 20.\dfrac{60-40}{40} = 30$

C80 ?

$-On\ calcule\ d'abord\ Cl\left(C80\right) : Cl\left(C80\right) \Leftrightarrow F\left[Cl\left(C80\right)\right] \geq 80 \Rightarrow Cl(C80) = [20\ 40[$

$-On$ calcule ensuite C_{80}: $C_{80} = 20 + 20.\dfrac{80-40}{40} = 40$

. Détermination graphique

. VS discrète

Reprenons la distribution précédente de la vs discrète et calculons D_4 graphiquement ?

Tableau 47 EX.22 Déciles

Xi	fi%
1	10
2	20
3	30
4	10
6	20
8	10

. Solution

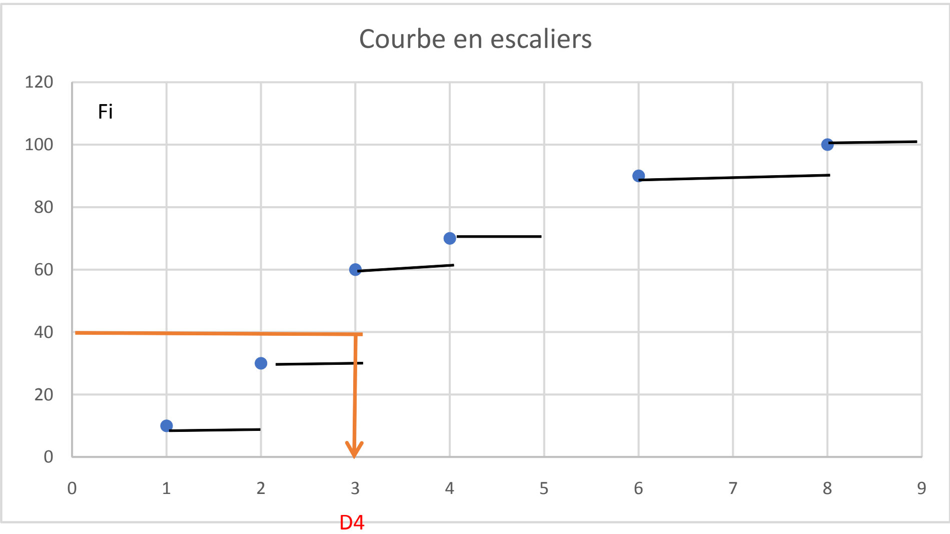
. On calcule d'abord F_i

Tableau 48 Solution EX.22

Xi	fi%	Fi
1	10	10
2	20	30
3	30	60
4	10	70
6	20	90
8	10	100

. On détermine $D_4 \leftrightarrow F(D4) \geq 40 = 3$ puis on le fixe sur le graphe de Fi

Figure 16 Détermination graphique de D_4 : cas discret



.VS continue

Reprenons la distribution de la vs continue précédente présentée ci-dessous et calculons C_{80} graphiquement?

Tableau 49 EX.23 Centiles

cl_i	fi%
[0-10[	10
[10-20[	20
[20-40[	30
[40-45[	40

. Solution

Tableau 50 Solution EX.23

cl_i	fi%	Fi
[0-10[	10	10
[10-20[	20	30
[20-40[	30	60
[40-45[	40	100

. On calcule d'abord Cl(C_{80}) : Cl(C_{80}) ↔ F(C_{80}) ≥80 =[40-45[

. Ensuite, on calcule Cl(C_{80}) : Par la méthode de l'égalité des pentes, nous obtenons :

. $(F(b)-F(a)))/(b—a)=(80-F(a))/(C_{80}-40)$

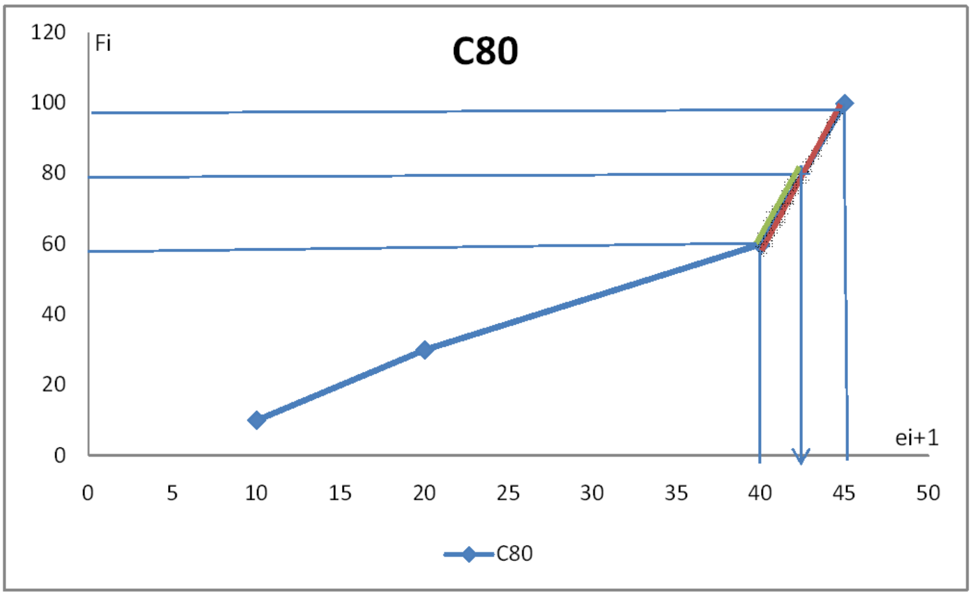
$(C_{80}-40)/(b—a)=(80-F(a))/(F(b)-F(a))$

Par conséquent on déduit C_{80} :

$C_{80}=a+(b-a)(80-F(a))/(F(b)-F(a))=40+5.(80-60)/40=42.50$

GRAPHE

Figure 17 Graphe C_{80} : cas continu



1.2.2.5 Ecart absolu moyen

Soit $X = \{(xi, ni)\}_{1 \leq i \leq p}$ une vs réelle.

L'**écart absolu moyen** de X *est* la moyenne arithmétique des valeurs absolues des écarts de X à sa moyenne : $e_a(\bar{x}) = 1/N \sum_i ni \,|\, xi - \bar{x} \,|$

L'écart absolu moyen de X par rapport à sa médiane (ou par rapport à un nombre réel a quelconque) est défini par $e_a(\mathrm{Me}) = 1/N \sum_i ni \,|\, xi - x_e |$

On peut démontrer que l'écart absolu moyen par rapport à un nombre réel a est minimum lorsque a est égal à la moyenne de X.

N.B. : L'écart absolu moyen présente un inconvénient majeur : il ne se prête pas facilement aux calculs algébriques, à cause de la valeur absolue.

. EXEMPLE

. VS discrète

Soit la distribution suivante, calculer l'écart absolu par rapport à la moyenne ? puis calculer l'écart absolu par rapport à la médiane ?

Tableau 51 EX.24 Ecart absolu

Xi	ni
5	10
10	8
15	9
20	7
30	8
40	3
	45

. Réponse

Tableau 52 Solution EX.24

Xi	ni	ni Xi	ni ↑	$ni\left\|Xi - \overline{X}\right\|$	$ni\left\|Xi - Me\right\|$
5	10	50	10	120	100
10	8	80	18	56	40
15	9	135	27	18	0
20	7	140	34	21	35
30	8	240	42	104	120
40	3	120	45	69	75
	45	765		388	370

$$Me \Leftrightarrow \mathrm{n}\uparrow(\mathrm{Me}) \geq \frac{n+1}{2} = 23$$

$$Me = 15$$

$$Ea\ (\mathrm{Me}) = 1/n \sum_i ni\ \left|Xi - \mathrm{Me}\right| = 8.22$$

$$\overline{X} = \sum_i \frac{ni\ \mathrm{Xi}}{n} = \frac{765}{45} = 17$$

$$Ea\ (\overline{\mathrm{X}}) = 1/n \sum_i ni\ \left|Xi - \overline{\mathrm{X}}\right| = 8.62$$

. VS continue

Soit la distribution suivante, calculer l'écart absolu par rapport à la moyenne ? puis calculer l'écart absolu par rapport à la médiane ?

Tableau 53 EX.25 Ecart absolu

ei	ei+1	ni
5	10	10
10	15	18
15	20	40
20	30	20
30	40	8
40	45	4

. Solution

Tableau 54 Solution EX.25

e_i	e_{i+1}	cl_i	n_i	c_i	$\lvert c_i - c_e \rvert$	$n_i \lvert c_i - c_e \rvert$	ni $\lvert Ci - \bar{C} \rvert$
5	10	[5-10[	10	7.5	10.25	102.5	120
10	15	[10-15[	18	12.5	5.25	94.5	126
15	20	[15-20[	40	17.5	0.25	10	80
20	30	[20-30[	20	25	7.25	145	110
30	40	[30-40[	8	35	17.25	138	124
40	45	[40-45[	4	42.5	24.75	99	92
			100			589	652

$$L'e.a.(Ce) = 1/n \sum_i ni \ \lvert \mathrm{Ci} - \mathrm{Ce} \rvert = 5.89$$

$$L'e.a.(\bar{C}) = 1/n \sum_i ni \ \lvert \mathrm{Ci} - \bar{\mathrm{C}} \rvert = 6.52$$

1.2.2.6 Variance et écart-type

Soit $X = \{(xi, ni)\}_{1 \leq i \leq p}$ une vs réelle.

La **variance** de X est la moyenne arithmétique des carrés des écarts de X à sa moyenne :

$$s^2(x) = 1/N \sum_i n_i (x - \bar{x})^2$$

L'écart **type** = $\sqrt{s^2(X)} = s(x)$

Formule de König de la variance:

$$s^2(x) = 1/N \sum_i n_i x_i^{\,2} - \bar{x}^2 = m_2 - m_1$$

Propriétés:

$s^2(a \pm bx) = b^2\, s^2(x)$

$s^2(a) = 0$

. EXEMPLE

. VS discrète

Tableau 55 Calcul Algébrique de la Variance d'une vs discrète

Xi	ni	niXi	niXi²

Soit la VS discrète suivante, calculer σ² puis σ ?

Tableau 56 EX.26 Variance

Xi	ni
5	2
10	3
20	1
8	2
	8

. Réponse

Tableau 57 Solution EX.26

Xi	ni	ni Xi	ni X²
5	2	10	50
10	3	30	300
20	1	20	400
8	2	16	128
	8	76	878

$m_1 =$ 9.5 109.75 $= m_2$

$m_1^2 =$ 90.25

$$\sigma^2 = \frac{1}{n}\sum ni(X_i - \overline{X})^2 = \frac{1}{n}\sum_i niX_i^2 - \overline{X}^2$$

$$\sigma^2 = m_2 - m_1^2 = 109.75 - 90.25 = 19.50 \quad \Rightarrow \quad \sigma = 4.42$$

. VS continue

Tableau 58 Calcul Algébrique de la Variance d'une vs continue

cl_i	n_i	c_i	n_ic_i	$n_ic_i^2$
$[e_0, e_1[$				
$[e_1, e_2[$				
$[e_i, e_{i+1}[$				

Soit la VS continue suivante, calculer σ^2 puis σ ?

Tableau 59 EX.27 Variance

cl_i	fi%
[0-10[	10
[10-20[	40
[20-30[	40
[30-50[	10

. Solution

Tableau 60 Solution EX.27

cl_i	fi%	fi	ci	fi Ci	Ci²	fi Ci²
[0-10[	10	0.1	5	0.5	25	2.5
[10-20[	40	0.4	15	6	225	90
[20-30[	40	0.4	25	10	625	250
[30-50[	10	0.1	40	4	1600	160

$$\sigma^2 = \frac{1}{n}\sum ni(C_i - \overline{C})^2 = \frac{1}{n}\sum_i niC_i^2 - \overline{C}^2$$

$$\sigma^2 = m_2 - m_1^2 = 502.5 - 20.5^2 = 82.25 \Rightarrow \sigma = 9.07$$

Calcul pratique de la variance

Formule de *KOENIG :* $v(x) = m_2 - m_1^2 = \frac{1}{n}\sum_i n_i x_i^2 - \overline{x}^2$

1.2.2.7 Coefficient de variation

cv = $s(x)/\overline{x}$

CV est un nombre sans dimension qui permet de comparer deux vs de natures différentes. On remarquera que, au signe près, c'est l'écart type de la vs $x/\overline{x}$.

. EXEMPLE

Soient les caractéristiques suivantes de deux distributions statistiques :

$$\sigma_1^2 = 4.16 \quad \overline{X}_1 = 16$$

$$\sigma_2^2 = 7.84 \quad \overline{X}_2 = 12$$

Comparer ces deux distributions du point de vue dispersion ?

. Solution

Parce que les moyennes des deux distributions sont différentes, la comparaison se fait par l'intermédiaire des coefficients de variations :

$$CV_1 = \frac{\sigma_1}{\overline{X}_1} = 0.13 \text{ (ou 13\%)} \quad CV_2 = \frac{\sigma_2}{\overline{X}_2} = 0.23 \text{ (ou 23\%)}$$

Parce que $CV_1 < CV_2$: la distribution ' 1 ' est meilleure que la distribution '2' de point vue dispersion.

1.2.2.8 Moments

Soit x une vs réelle.

Le **moment simple d'ordre *r*** *de x* : $m_r = 1/N \sum_i x_i^r$

$r = 0 : m0 = 1.$

$r = 1 : m1 = \bar{x}$

$r = 2 : m2 = 1/N \sum_i x_i^2$

Le **moment centré d'ordre *r***: $\mu_r = 1/N \sum_i (x_i - \bar{x})^r$

$r = 0$: µ0 = 1.

$r = 1$: µ1 = 0.

$r = 2 : \mu 2 = s^2(X) = m2 - m1$

. EXEMPLE

. **VS discrète**

. Série numérique

Soit la série numérique suivante, calculer les moments simples m_2, m_3 et m_4 puis les moments centrés μ_1, μ_2 et μ_3 ?

Xi = 2, 4, 8, 6

. Solution

Tableau 61 Solution Moments

Xi	X^2i	x_i^3	x_i^4	μ_1	μ_2	μ_3
2	4	8	16	-3	9	-27
4	16	64	256	-1	1	-1
8	64	512	4096	3	9	27
6	36	216	1296	1	1	1
	120	800	5664	0	20	0
	30	200	1416	0	5	0

$m_2 = 30$

$m_3 = 200$

$m_4 = 1416$

$\mu_1 = 0$

$\mu_2 = 5$

$\mu_3 = 0$

. Série classée

Soit la distribution suivante, calculer les moments simples m_2, m_3 et m_4 puis les moments centrés μ_2, μ_3 et μ_4 ?

Tableau 62 EX.28 Moments

Xi	ni
5	2
10	3
20	1
8	2

. Solution

Tableau 63 Solution EX.28

Xi	ni	ni Xi	ni X^2	ni X^3	ni X^4	$(X_i - \overline{X})$	$n_i(X_i - \overline{X})^2$	$n_i(X_i - \overline{X})^3$	$n_i(X_i - \overline{X})^4$
5	2	10	50	250	1250	-4.5	40.5	-182.5	820.125
10	3	30	300	3000	30000	0.5	0.75	0.375	0.1875
20	1	20	400	8000	160000	10.5	110.25	1157.625	12155.0625
8	2	16	128	1024	8192	-1.5	4.5	-6.75	10.125
	8	76	878	12274	199442		156	968.75	12985.5
		9.5	109.8	1534	24930		19.5	121.09375	1623.1875

$$m_2 = \frac{\sum_i niXi}{n} = \frac{878}{8} = 109.8 \qquad m_3 = \frac{\sum_i niX_i^3}{n} = \frac{12274}{8} = 1534 \qquad m_4 = \frac{\sum_i niX_i^4}{n} = \frac{199442}{8} = 24930$$

$$\mu 2=\frac{\sum_i ni\left(Xi-\overline{X}\right)^2}{n}=\frac{156}{8}=19.5 \qquad \mu 3=\frac{\sum_i ni\left(Xi-\overline{X}\right)^3}{n}=\frac{968.75}{8}=121.1 \qquad \mu 4=\frac{\sum_i ni\left(Xi-\overline{X}\right)^4}{n}=\frac{12985.5}{8}=1623.2$$

. **VS continue**

Soit la VS continue suivante, calculer les moments simples m_1, m_2 et m_3 puis les moments centrés μ_1, μ_2 puis μ_3?

Tableau 64 EX.29 Moments

ei	ei+1	ni
0	10	1
10	20	2
20	30	4
30	40	3
		10

. Solution

Tableau 65 Solution EX.29

ei	ei+1	ni	Ci	ni Ci	ni C²i	$n_i \; C_i^3$	µ1	µ2	µ3
0	10	1	5	5	25	125	-19	361	-6859
10	20	2	15	30	450	6750	-18	162	-1458
20	30	4	25	100	2500	62500	4	4	4
30	40	3	35	105	3675	128625	33	363	3993
		10		240	6650	198000	0	890	-4320
				24	665	19800	0	89	-432

$m_1 = 24$

$m_2 = 665$

$m_3 = 19800$

$\mu_1 = 0$

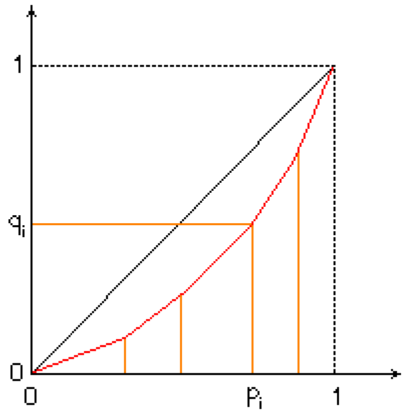
$\mu 2 = 89$

$\mu_3 = - 432$

1.2.3 PARAMETRES DE CONCENTRATION

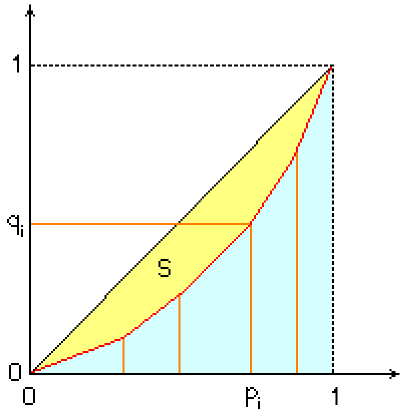
1.2.3.1 Courbe de Lorenz

Figure 18 Courbe de Lorenz



1.2.3.2 Indice de Gini

Figure 19 Indice de Gini



L'indice de Gini permet de mesurer les inégalités scolaires, les inégalités de statut, les inégalités de salaires, etc.

1.2.3.3 Médiale

La **médiale** d'une variable statistique *X* est la valeur de *X* qui partage la masse globale en deux parties égales.

Figure 20 Médiale

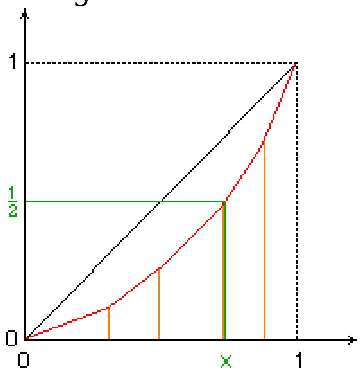
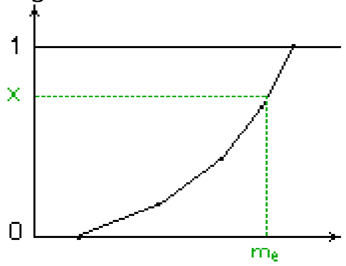


Figure 21 Médiane



. EXEMPLE

. VS discrète

Soit la distribution de 25 entreprises selon le chiffre d'affaires, calculer la médiane 'Me', la médiale 'M_l' puis le coefficient de concentration 'Ic' ?

Tableau 66 EX.30 Ic

Xi	ni
2	10
3	6
4	4
5	3
7	2
	25

Solution

Tableau 67 Solution EX.30

Xi	ni	ni↑	niXi	niXi↑	fi	Fi	(niXi)r	Mi	Fi-Fi-1	Mi+Mi-1	*
2	10	10	20	20	0.40	0.4	0.24	0.24	0.40	0.24	0.10
3	6	16	18	38	0.24	0.64	0.22	0.46	0.24	0.70	0.17
4	4	20	16	54	0.16	0.80	0.19	0.65	0.16	1.11	0.18
5	3	23	15	69	0.12	0.92	0.18	0.83	0.12	1.48	0.18
7	2	25	14	83	0.08	1.00	0.17	1.00	0.08	1.83	0.15
	25		83								0.76

$$Me \Leftrightarrow n\uparrow(Me) \geq \frac{n}{2}$$

$$Me \Leftrightarrow n\uparrow(Me) \geq 12.5 \Rightarrow Me = \text{ 3 u.m.}$$

$$Ml \Leftrightarrow n\uparrow(Ml) \geq \frac{ni \text{ ci}}{2}$$

$$Ml \Leftrightarrow n\uparrow(Ml) \geq \text{ 41.5 } \Rightarrow \text{ Ml} = \text{ 4 u.m}$$

(Fi-Fi-1) (Mi+Mi-1) = 0.76

Ic = 1 – 0.76 = 0.24 : La concentration du chiffre d’affaires est faible.

. VS continue

Soit la distribution de 36 ouvriers d’une entreprise selon le salaire perçu. Calculer la médiane ‘Me’, la médiale ‘M_l’ puis le coefficient de concentration ‘Ic’ ?

Tableau 68 EX.31 Ic

ei	ei+1	Ci	ni
1	1.4	1.2	10
1.4	1.6	1.5	12
1.6	2.4	**2**	6
2.4	2.6	2.5	3
2.6	3.4	3	3
3.4	4.6	4	2
			36

.Solution

Tableau 69 Solution1 EX.31

ei	ei+1	Ci	ni	ni↑	niCi	niCi↑	fi	fi↑	(nici)r	(nici)r↑
1	1.4	1.2	10	10	12	12	0.28	0.28	0.18	0.18
1.4	1.6	1.5	12	22	18	30	0.33	0.61	0.27	0.45
1.6	2.4	2	6	28	12	42	0.17	0.78	0.18	0.63
2.4	2.6	2.5	3	31	7.5	49.5	0.08	0.86	0.11	0.74
2.6	3.4	3	3	34	9	58.5	0.08	0.95	0.14	0.88
3.4	4.6	4	2	36	8	66.5	0.06	1.00	0.12	1.00
			36		66.5					

$$CLe \Leftrightarrow n\uparrow(CLe) \geq \frac{n}{2}$$

$$CLe \Leftrightarrow n\uparrow(CLe) \geq 18 \Rightarrow CLe = [1.4\ 1.6[$$

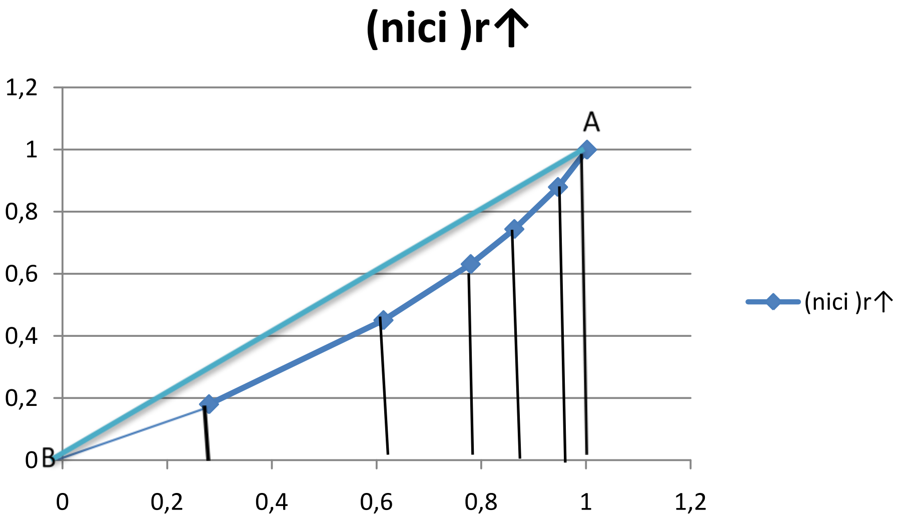
$$Me = 1.4 + 0.2 * \frac{18-10}{12} = 1.53 \text{ u.m.}$$

$$CL_{ml} \Leftrightarrow n\uparrow(CL_{ml}) \geq \frac{ni\ \text{ci}}{2}$$

$$CL_{ml} \Leftrightarrow n\uparrow(CL_{ml}) \geq 33.25 \Rightarrow CL_{ml} = [1.6\ 2.4[$$

$$Ml = 1.6 + 0.8 * \frac{33.25-30}{12} = 1.82 \text{ u.m.}$$

Figure 22 Coefficient de concentration Ic



. Coefficient de concentration ' Ic '

. Méthode 1

$$C = \frac{surface \text{ S}}{surface \text{ du triangle 0AB}}$$

$$Surface \text{ du triangle 0AB} = \frac{0.28*0.18}{2} + \frac{(0.18+0.45)*(0.61-0.28)}{2} + \frac{(0.45+0.63)*(0.78-0.61)}{2}$$
$$\frac{(0.63+0.74)*(0.86-0.78)}{2} + \frac{(0.74+0.88)*(0.95-0.86)}{2} + \frac{(0.88+1)*(1-0.95)}{2} = 0.0252 +$$
$$0.010395 + 0.0918 + 0.0548 + 0.0729 + \ 0.047 = 0.39565$$

$$Surface \text{ S} = \frac{1}{2} - 0.39565 = 0.10435$$

$$C = \frac{0.10435}{0.5} = 0.2087$$

. Méthode 2

Tableau 70 Solution2 EX.31

ei	ei+1	Ci	ni	ni↑	niCi	niCi↑	fi	Fi	(nici)r	Mi	Fi-Fi-1	Mi+Mi-1	*
1	1.4	1.2	10	10	12	12	0.28	0.28	0.18	0.18	0.28	0.18	0.05
1.4	1.6	1.5	12	22	18	30	0.33	0.61	0.27	0.45	0.33	0.63	0.21
1.6	2.4	2	6	28	12	42	0.17	0.78	0.18	0.63	0.17	1.08	0.18
2.4	2.6	2.5	3	31	7.5	49.5	0.08	0.86	0.11	0.744	0.08	1.38	0.11
2.6	3.4	3	3	34	9	58.5	0.08	0.95	0.14	0.879	0.08	1.62	0.14
3.4	4.6	4	2	36	8	66.5	0.06	1.00	0.12	1.00	0.06	1.88	0.10
			36		66.5								0.80

(Fi-Fi-1) (Mi+Mi-1)=
0.28*0.18+0.63*0.33+0.17*1.08+0.08*1.37+0.09*1.62+0.05*1.88=0.0504+0.2
079+0.1836+0.1096+0.1458+0.094=0.7913

Ic = 1-0.7913=0.2087 : Nous concluons que la concentration des salaires est faible

1.2.4 PARAMETRES DE FORME

1.2.4.1 Coefficient d'asymétrie (skewness)

PEARSON : $P = \frac{x - m_0}{s(x)}$

YULE : $Y = \dfrac{Q1 + Q3 - 2m_e}{2(Q3 - Q1)}$

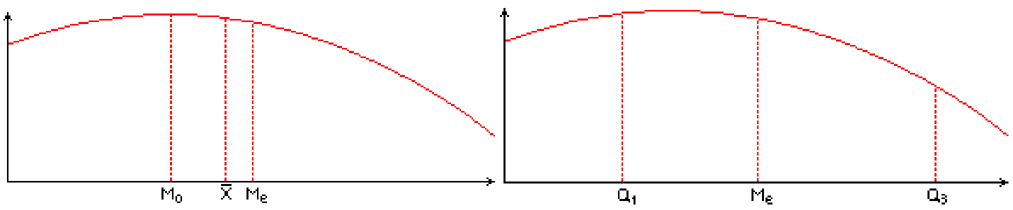
FISHER : $F_1 = \dfrac{\mu_3}{s^3}$

$F_1 > 0$ courbe étalée à droite

$F_1 < 0$ courbe étalée à gauche

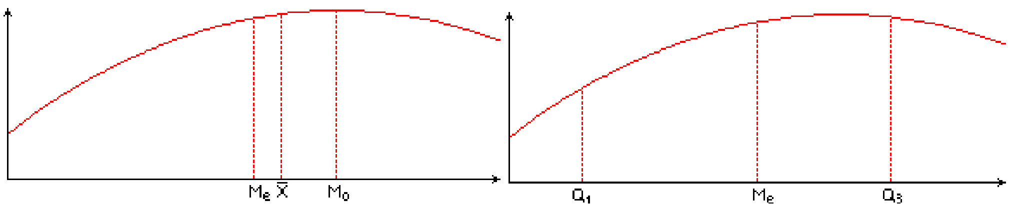
$F_1 > 0$

Figure 23 Skewness



$F_1 < 0$

Figure 24 Kurtosis



1.2.4.2 Coefficient d'aplatissement (kurtosis)

$$\mathbf{F_2} = \frac{\mu_4}{s^4}$$

F2 > 3 courbe moins aplatie que N(0,1)

F2 < 3 courbe plus aplatie que N(0,1)

Figure 25   KURTOSIS < 0

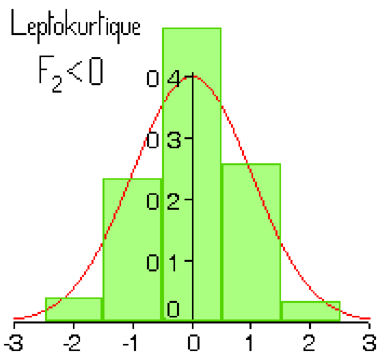
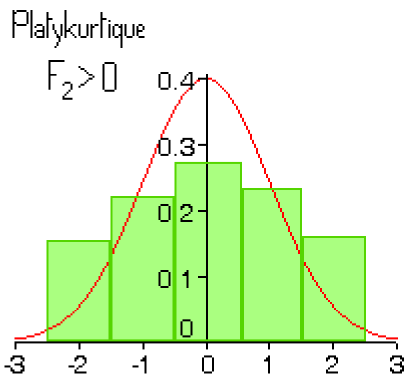


Figure 26 KURTOSIS > 0



. EXEMPLE

. VS discrète

Soit la distribution de 40 élèves selon la note obtenue en math (sur 10), tester l'asymétrie et l'aplatissement au moyen des paramètres de forme de Fisher F_1 et F_2 ?

Tableau 71 EX.32 F_1, F_2

Notes	nombre d'élèves
4	1
5	3
6	6
7	15
8	12
9	3
	40

. Solution

Tableau 72 Solution EX.32

Xi	ni	ni xi	ni x²i	$ni\,(Xi - X)^3$	$ni(Xi - X)^4$
4	1	4	16	-29.22	89.99
5	3	15	75	-27.00	56.15
6	6	36	216	-7.56	8.16
7	15	105	735	-0.01	0.00
8	12	96	768	9.34	8.60
9	3	27	243	21.23	40.77
	40	283	2053	-33.20	203.67
		7.08	51.33	-0.83	5.09

$$\sigma = \sqrt{m_2 - m_1^2} = \sqrt{51.33 - 7.08^2} = \sqrt{1.20} = 1.1$$

$$\sigma^3 = 1.33$$

$$\sigma^4 = 1.46$$

$$\mu_3 = 1/n \sum_i ni\ (Xi - X)^3 = 1/40 * -33.20 = -0.83$$

$$F_1 = \frac{\mu_3}{\sigma^3} = \frac{-0.83}{1.33} = -0.62$$

F1= - 0.62 < 0 : La distribution est ainsi asymétrique à gauche.

$$\mu_4 = 1/n \sum_i ni\ (Xi - X)^4 = 1/40 * 203.67 = 5.09$$

$$F_2 = \frac{\mu_4}{\sigma^4} - 3 = \frac{5.09}{1.46} - 3 = 0.49$$

$F_2 > 0$: la distribution est platykurtique.

. VS continue

Soit la distribution suivante d'une vs continue, tester l'asymétrie et l'aplatissement au moyen des paramètres de forme de Fisher F_1 et F_2 ?

Tableau 73 EX.33 F_1, F_2

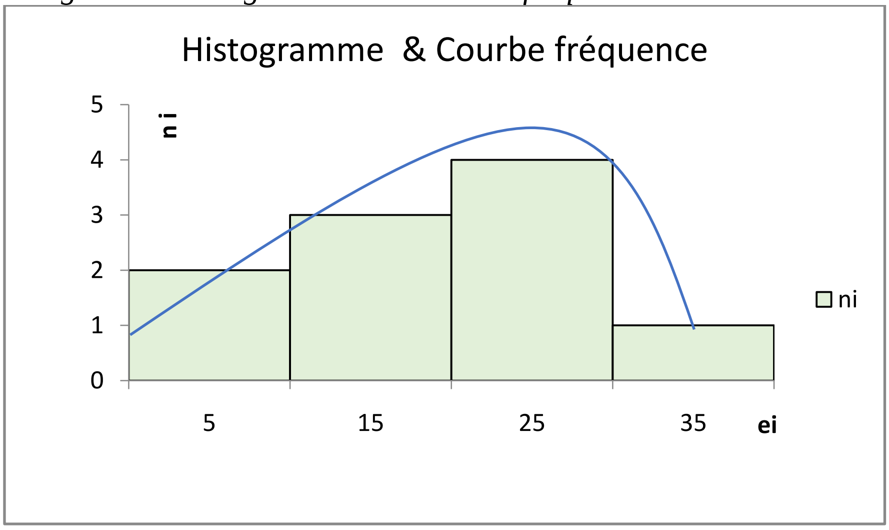
cl_i	ni
[0-10[	2
[10-20[	3
[20-30[	4
[30-40[	1
	10

. Solution

Tableau 74 Solution EX.33

cl_i	ni	Ci	niCi	$(C-\overline{C})$	$(C-\overline{C})^3$	$(C-\overline{C})^4$	$ni(C-\overline{C})^3$	$ni(C-\overline{C})^4$	niCi²
[0-10[	2	5	10	-14	-2744	38416	-5488	76832	50
[10-20[	3	15	45	-4	-64	256	-192	768	675
[20-30[	4	25	100	6	216	1296	864	5184	2500
[30-40[	1	35	35	16	4096	65536	4096	65536	1225
	10		190				-720	148320	**4450**
			19						

Figure 27 Histogramme et courbe de fréquence



$m_2 = \dfrac{4450}{10} = 445$ $\quad$ $m_1 = \dfrac{190}{10} = 19$ $\quad$ $\sigma^2 = m_2 - m_1^2 = 445 - 19^2 = 84$ $\quad$ $\sigma = 9.16$

$\sigma^3 = 768.57$ $\quad$ $\sigma^4 = 7056$

$\mu_3 = -72$ $\quad$ $F1 = \dfrac{\mu 3}{\sigma 3} = \dfrac{-72}{768.57} = -0.09$

$\mu_4 = 14832$ $\quad$ $F2 = \dfrac{\mu 4}{\sigma^4} - 3 = \dfrac{14832}{7056} - 3 = 2.10 = -0.90$

F1= -0.9 < 0 : la courbe est asymétrique à gauche
F2= -0.9 <0 : la courbe est platykurtique.

CH2. STATISTIQUE BIVARIEE

Soit une population Ω de taille n décrite simultanément par rapport à deux caractères x et y respectivement de k et p modalités.

2.1 Distribution conjointe

L'effectif partiel d'un couple (x_i,y_j) est le nombre n_{ij}= card($x=x_i \cap y=y_j$).

La distribution des effectifs est l'application qui à tout couple (x_i,y_j) associe son effectif partiel n_{ij} : $(x_i, y_j) \rightarrow n_{ij}$

La distribution marginale d'effectif du couple (x_i, y_j) est 1 distribution d'effectif de x_i d'une part et une distribution d'effectif de y_j d'autre part :

$X_i \rightarrow$ card $(x=x_i , \forall y_j)$

$Y_j \rightarrow$ card$(y=y_j , \forall x_i)$

Tableau de contingence

C'est un tableau à double entrée représentant la distribution conjointe du couple (Xi, Yj)

Tableau 75 Tableau de cotingence

	y_1	y_j	y_p	
x_1	n_{11}	n_{1j}	n_{1p}	$n_{1.}$
x_i	n_{i1}	n_{ij}	n_{ip}	ni.
x_k	n_{k1}	n_{kj}	n_{kp}	$n_{k.}$
	$n_{.1}$	n.j	$n_{.p}$	n..= n

n_{ij} : effectif des éléments ayant le caractère i et le caractère j

$n_{i.}$:effectif des éléments ayant le caractère i $= \sum_j n_{ij}$

$n_{.j}$: effectif des éléments ayant le caractère j $= \sum_i n_{ij}$

$$n.. = \sum_i \sum_j n_{ij} = \sum_i n_{i.} = \sum_j n_{.j} = n$$

Partant de la définition de la fréquence $f_{ij} = {nij}/{n..}$; $\sum_i \sum_j fij = 1$, le tableau des fréquences s'ensuit :

Tableau 76 Tableau des Fréquences conjointes

	y_1	y_j	y_p	
x_1	f_{11}	f_{1j}	f_{1p}	$f_{1.}$
x_i	f_{i1}	f_{ij}	f_{ip}	fi.
x_k	f_{k1}	f_{kj}	f_{kp}	$f_{k.}$
	$f_{.1}$	f.j	$f_{.p}$	1

. EXEMPLE

Soit la distribution de 20 travailleurs selon le salaire X et l'âge Y.

Tableau 77 EX.34 Tableau des fréquences conjointes

X \ Y	25	35	45	55	$n_{i.}$
3	3	1	1	0	5
5	1	5	0	0	6
7	0	1	3	0	4
9	0	0	1	2	3
11	0	0	2	0	2
$n_{.j}$	4	7	7	2	20

1.Déterminer et interpréter n_{22}, n_{32}?

2. Déterminer et interpréter f_{23}, f_{34} et f_{44} ?

Solution

1.$n_{22}=5$ 5 ouvriers ont un salaire de 5 DA et un âge de 35 ans

$n_{32}=1$ 1 ouvrier a un salaire de 7 DA et un âge de 35 ans

2.

X \ Y	25	35	45	55	$f_{i.}$
3	0.15	0.05	0.05	0	0.25
5	0.05	0.25	0	0	0.3
7	0	0.05	0.15	0	0.2
9	0	0	0.05	0.1	0.15
11	0	0	0.1	0	0.1
$f_{.j}$	0.2	0.35	0.35	0.1	1

$f_{21}=0.05$: 5% des ouvriers ont un salaire de 5 DA et l'âge de 25 ans.

$f_{44}=0.1$: 10% des ouvriers ont un salaire de 11 DA et l'âge de 55 ans.

2.2 Distributions marginales

fi.= $ni./nij$ **;** $\sum_i fi. = 1$ **f.j=** $n.j/nij$ **;** $\sum_j f.j = 1$

Tableau 78 Tableau des Fréquences marginales des Xi

Modalités du caract. xi	Effectif marginal	Fréquence marginale
x_1	$n_{1.}$	$f_{1.}$
x_i	$n_{i.}$	$f_{i.}$
x_k	$n_{k.}$	$f_{k.}$
$\sum$	$n_{..}$	1

Tableau 79 Tableau des Fréquences marginales des Yj

Modalités du caract. yi	Effectif marginal	Fréquence marginale
y_1	$n_{.1}$	$f_{.1}$
y_i	$n_{.2}$	$f_{.2}$
y_p	$n_{.p}$	$f_{.p}$
$\sum$	$n_{..}$	1

Caractéristiques *des distributions marginales*

$$\overline{\overline{x}} = 1/n_{..} \sum_i n_{i.} x_i = \sum_i f_{i.} x_i \qquad \overline{\overline{y}} = 1/n_{..} \sum_i n_{.j} y_j = \sum_i f_{.j} y_j$$

$$v(x) = 1/n_{..} \sum_i n_{i.} x_i^2 - \overline{\overline{x}}^2 \qquad v(y) = 1/n \sum_i n y_j^2 - \overline{\overline{y}}^2$$

Selon la formule de définition :

$$\mathbf{v(x)} = 1/n_{..} \sum_i ni.(xi - \overline{\overline{x}})^2 = \sum_i fi.(xi - \overline{\overline{x}})^2 \qquad \mathbf{v(y)} = 1/n_{..} \sum_j n_{.j}(y_j - \overline{\overline{y}})^2 = \sum_j f_{.j}(y_j - \overline{\overline{y}})^2$$

. EXEMPLE

1. Etablir les tableaux des effectifs marginaux et des fréquences marginales de xi

   et de yj?

   Déterminer et interpréter $n_2.$, $f_2.$, $n._{4,}$ $f._4$, n.. et f..?

2. Calculer les moyennes marginales de xi et d'yj ?

3. Calculer les variances marginales de xi et d'yj ?

1a. Distribution marginale de X

X	$n_{i.}$	fi.
3	5	0.25
5	6	0.3
7	4	0.2
9	3	0.15
11	2	0.1
n..	20	1

1b. Distribution marginale de Y

Y	25	35	45	55	n..
$n_{.j}$	4	7	7	2	20
f.j	0.2	0.35	0.35	0.1	1

$n_2.$= 6

f2. = 0.3

$n._4$= 2

f.4= 0.1

f..=1

n..= 20

2. $\overline{\overline{X}} = \sum_{i=1}^{5} f_{i\cdot} \times x_i = \dfrac{5\times3+6\times5+4\times7+3\times9+2\times11}{20} = 6.10$ DA

$\overline{\overline{Y}} = \sum_{j=1}^{4} f_{\cdot j} \times y_j = \dfrac{4\times25+7\times35+7\times45+2\times55}{20} = 38.5$ ANS

3.

$\sigma_x^2 = \sum_{i=1}^{5} f_{i\cdot}\left(x_i - \overline{x}\right)^2 = \left(\sum_{i=1}^{5} f_{i\cdot}\left(x_i\right)^2\right) - \left(\overline{x}\right)^2 = \dfrac{5\times3^2+6\times5^2+4\times7^2+3\times9^2+2\times11^2}{20} - (6.10)^2 = 6.59 DA^2 \quad \sigma = 2.5 \; ans$

$\sigma_y^2 = \sum_{i=1}^{4} f_{\cdot j}\left(y_j - \overline{y}\right)^2 = (0.2\ 25^2 + 0.35\ 35^2 + 0.35\ 45^2 + 0.1\ 55^2) - (38.5)^2 = 82.75 \text{ da}^2 \quad \sigma = 9 \text{ ans}$

2.3 Distributions conditionnelles

$\mathbf{f}_i^{\,j} = {nij}/{n.j}$ $\quad \sum_i f_i^{\,j} = \mathbf{1}$ $\quad \mathbf{f}_j^{\,i} = {nij}/{ni.}$ $\quad \sum_j f_j^{\,i} = \mathbf{1}$

Tableau 80 Tableau des Fréquences conditionnelles Xi

Modalités caract. xi	effectif	Fréquence conditionnelle
x_1	n_{1j}	$f_1^{\,j}$
x_i	n_{ij}	$f_i^{\,j}$
x_k	n_{kj}	$f_k^{\,j}$
$\sum$	$n_{.j}$	1

Tableau 81 Tableau des fréquences conditionnelles Yj

Modalités caract. y_j	effectif	Fréquence conditionnelle
y_1	n_{i1}	$f_1^{\,i}$
y_j	n_{i2}	$f_2^{\,i}$
Y_p	n_{ip}	$f_p^{\,i}$
$\sum$	$n_{i.}$	1

$\dfrac{nij}{n..} = \dfrac{ni.}{n..}\,\dfrac{nij}{ni.} = \dfrac{n.j}{n..}\,\dfrac{nij}{n.j} \Leftrightarrow$ $\mathbf{f_{ij} = f_{i.}\, f_j^{\,i} = f_{.j}\, f_i^{\,j}}$ *p(ab)=p(a) p(b/a)*

Caractéristiques *des distributions conditionnelles*

$\bar{x}_j = \frac{1}{n_{.j}} \sum_i n_{ij} x_i = \sum_i f_i^{\,j}\, \mathbf{x}_i$

$\bar{y}_i = \frac{1}{n_{i.}} \sum_j n_{ij} y_j = \sum_i f_j^{\,i}\, \mathbf{y}_j$

$v_j(x) = \frac{1}{n_{.j}} \sum_i n_{ij} x_i^2 - \bar{x}_j^{\,2}$

$v_i(y) = \frac{1}{n_{i.}} \sum_j n_{ij} y_j^2 - \bar{y}_i^{\,2}$

Selon la formule de définition :

$$v_j(x) = \frac{1}{n_{.j}} \sum_i n_{ij} (x_i - \bar{x}_j)^2 = \sum_i f_i^{\,j} (x_i - x_j)^2$$

$$v_i(y) = \frac{1}{n_{i.}} \sum_j n_{ij} (y_j - \bar{y}_i)^2 = \sum_i f_j^{\,i} (y_j - y_i)^2$$

. Relation entre $\bar{\bar{x}}$ et $\bar{x}_j$; v(x) et v_j(x)

$\bar{\bar{x}} = \Sigma_j\, f_{.j}\, \bar{x}_j$

$\bar{\bar{y}} = \Sigma_i\, f_{i.}\, \bar{y}_i$

v(x)=$\sum_j f_{.j} v_j(x) + \sum_j f_{.j} (\bar{x}_j - \bar{\bar{x}})^2$

v(y)=$\sum_i f_{i.} v_i(y) + \sum_i f_{i.} (\bar{y}_i - \bar{\bar{y}})^2$

. *EXEMPLE*

1. Calculer les moyennes conditionnelles de xi et d'yj ?

2. Calculer les variances conditionnelles de xi et d'yj ?

Solution

1a. Distribution conditionnelle de X sachant Y=35ans

X	n_{i2}	$f_i^2 = \frac{n_{i2}}{n_{.2}}$
3	1	$\frac{1}{7}$
5	5	$\frac{5}{7}$
7	1	$\frac{1}{7}$
9	0	0
11	0	0
$\sum$	7	1

$$\bar{x}_2 = \sum_{i=1}^{5} f_i^{\,2} x_i = \frac{1\times3+5\times5+1\times7}{7} = 5 \text{ DA}$$

1b. *Distribution conditionnelle de Y sachant X=3 DA*

Y	n_{ij}	$f_j^1 = \frac{n_{1j}}{n_{1\cdot}}$
25	3	$\frac{3}{5}$
35	1	$\frac{1}{5}$
45	1	$\frac{1}{5}$
55	0	0
$\sum$	5	1

$$\bar{y}_1 = \sum_{j=1}^{4} f_j^1 y_j = \frac{3\times25+1\times35+1\times45}{5} = 31 \text{ ans}$$

2a. $$V_2(\text{x}) = \sum_{i=1}^{5} f_i^{\,2} x_i^2 - \left(\bar{x}_2\right)^2 = \frac{1\times3^2+5\times5^2+1\times7^2}{7} - 5^2 = 1.143 \text{ DA}^2$$

2b. $$V_1(y) = \sum_{j=1}^{4} f_j^1 y_j^2 - \left(\bar{y}_1\right)^2 = \frac{3\times25^2+1\times35^2+1\times45^2}{5} - 31^2 = 64 \text{ ans}^2$$

2.4 Indépendance

x i y $\Leftrightarrow$ $\mathbf{f}_i^{\,j} = \mathbf{f}_{i.}$ *p(a/b) =p(a)*

x i y $\Leftrightarrow \forall$ i, i=$\overline{1,k}$ $f_i^1 = f_i^{\,j} = f_i^{\,k}$ $\Leftrightarrow \frac{ni1}{n.1} = \frac{nij}{n.j} = \frac{nik}{n.k} = \frac{ni1+nij+nik}{n.1+n.j+n.k} = \frac{ni.}{n..}$ cad $\frac{nij}{n.j} = \frac{ni.}{n..}$ ou

$f_i^{\,j}$ = fi. cad A / B_j = A

Considérons :

fij= fi. f^{i}_{j} = f.j f^{j}_{i} lorsque x i y, on a f^{j}_{i} =fi. d'où f^{i}_{j} =f.j et

fij = fi. f.j *p(ab)=p(a).p(b)*

cad fij= $\frac{ni.\, n.j}{n..^2}$ (les lignes (ou les colonnes) st proportionnelles entre elles)

x lié fonctionnellement à y **:** $\forall i, j \rightarrow$ i= θ (j) est unique

. EXEMPLE

X i Y ↔nij* n.. = ni. * n.j

Soit i= 3 j=2 : n32 * n..= n3. * n.2 ?

$1 \times 20 \neq 4 \times 7 \Rightarrow$ X n'est pas indépendant de Y

Coefficient de régression

Décomposition de v(x)

$$v(y) = \sum_i f_{i.} v_i(y) + \sum_i f_{i.} (\bar{y}_i - \bar{\bar{y}})^2$$

var.mgle = var.résiduelle + var.expliquée par l'hétérogéniété des var

var.totale = var.résiduelle + var. expliquée par la régression

2.5 Représentation graphique

i. Deux caractères qualitatifs : tuyaux d'orgues, graphes semi-circulaires ou

graphe en anneaux

Soit le tableau de contingence ci-dessous représentant la distribution des travailleurs selon leur statut et la branche économique, représenter graphiquement cette distribution ?

- tuyaux d'orgues

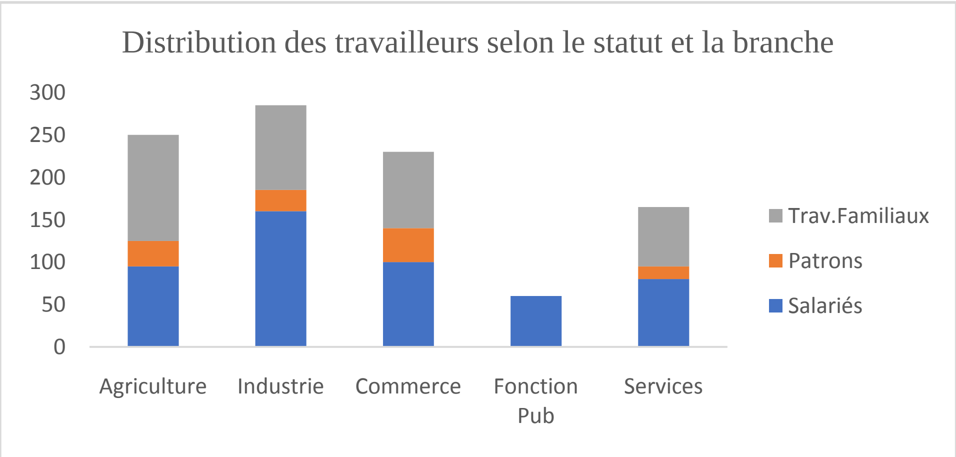
Tableau 82 EX.35 Tableau de contingence de 2 variablees qualitatives

	Salariés	Patrons	Trav.Familiaux	
Agriculture	95	30	125	250
Industrie	160	25	100	285
Commerce	100	40	90	230
Fonction Pub	60	0	0	60
Services	80	15	70	165
Total	495	110	385	990

Tableau 83 EX.36 Tableau de contingence en %

	Salariés	Patrons	Trav.Familiaux	
Agriculture	19%	27%	32%	25%
Industrie	32%	23%	26%	29%
Commerce	20%	36%	23%	23%
Fonction Pub	12%	0%	0%	6%
Services	16%	14%	18%	17%
Total	100%	100%	100%	100%

Figure 28 Graphe de deux variables qualitatives



- Graphique semi-circulaire

Figure 29 Graphe semi-circulaire relatif aux salariés

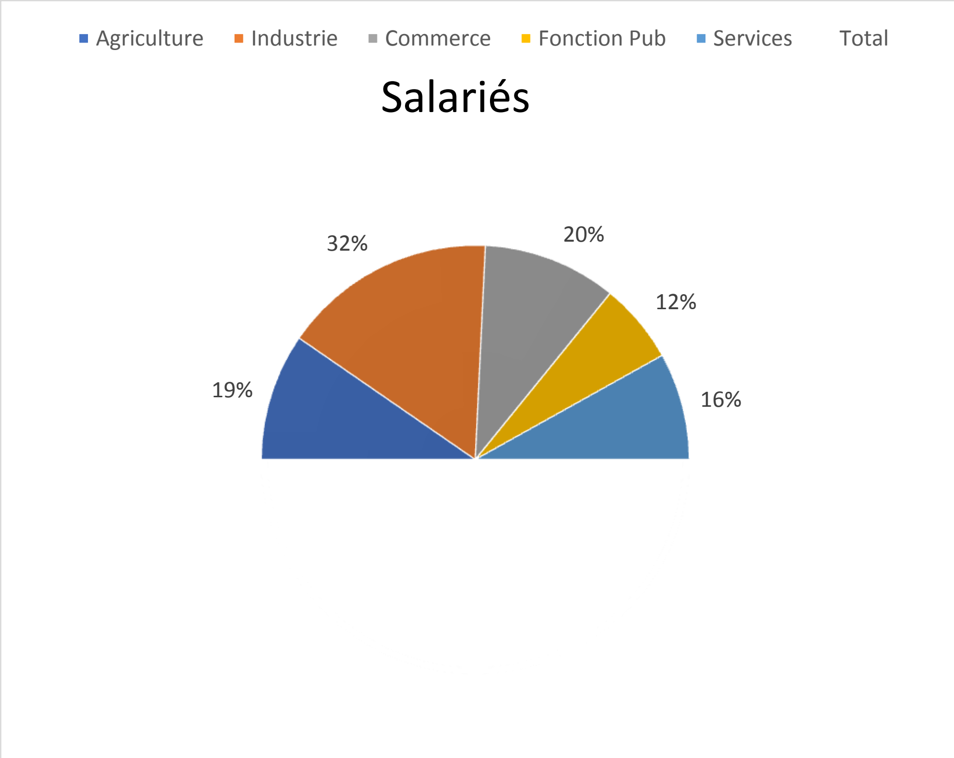
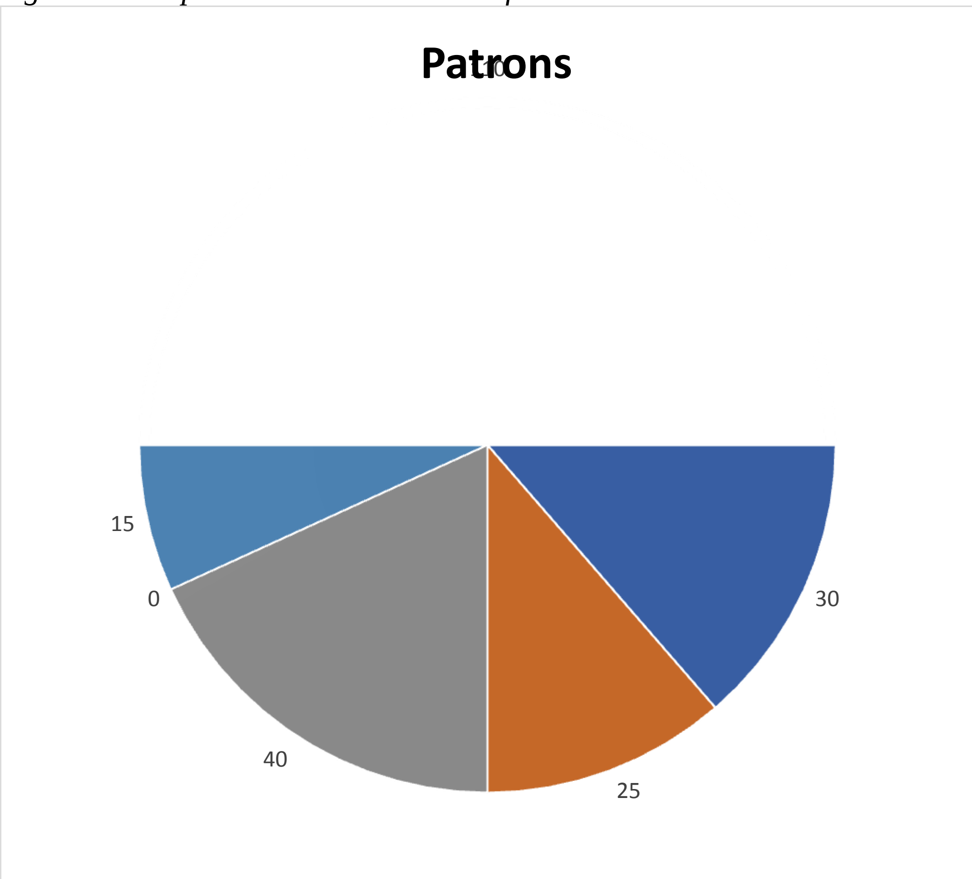


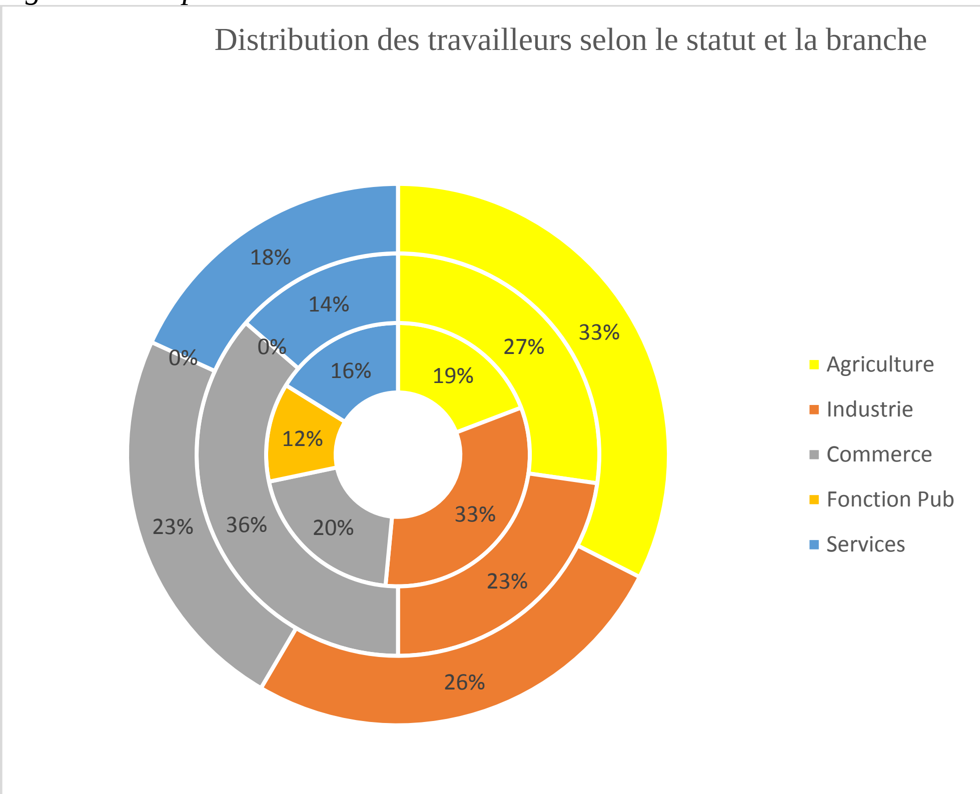
Figure 30 Graphe semi-circulaire relatif aux Patrons



- Graphe en Anneaux

Il constitue une variante du graphique circulaire.

Figure 31 Graphe en anneaux



ii. variables mixtes

-1 caractère qualitatif et 1 caractère quantitatif discret : diagramme en bâtons (1) ou tuyaux d'orgue(2)

(1) : en Y, 1modalité du caractère qualitatif ; en X, les modalités du caractère quantitatif discret

(2) : en X, les modalités du caractère quantitatif discret (base de chaque tuyau d'orgue) ; en Y, les modalités du caractère qualitatif (fi%)

. EXEMPLE

Soit la distribution suivante des logements selon le nombre de pièces et en fonction du statut d'occupation, donner la représentation graphique appropriée ?

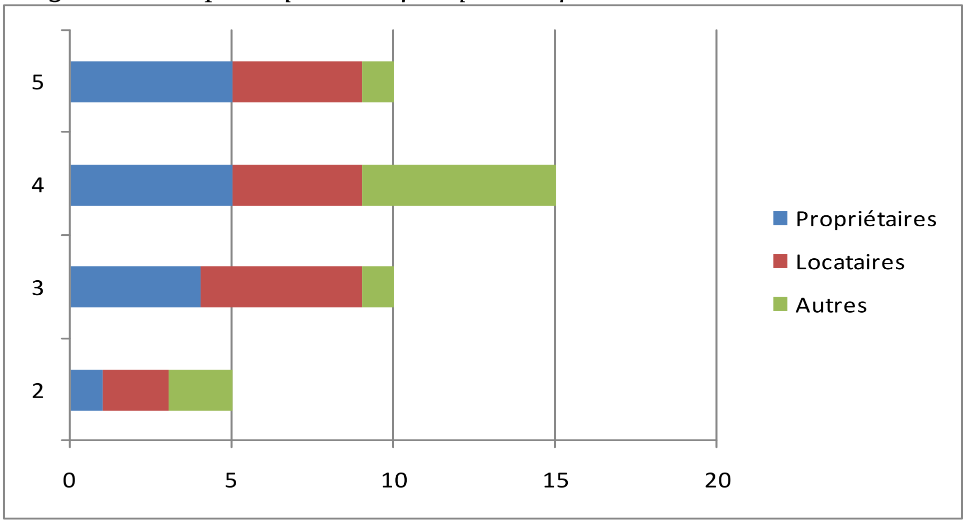
Tableau 84 EX.37 Tableau de contingence

	2	3	4	5
Propriétaires	1	4	5	5
Locataires	2	5	4	4
Autres	2	1	6	1
	5	10	15	10

Pour le graphe en tuyau d'orgue, on saisit en X les modalités du caractère

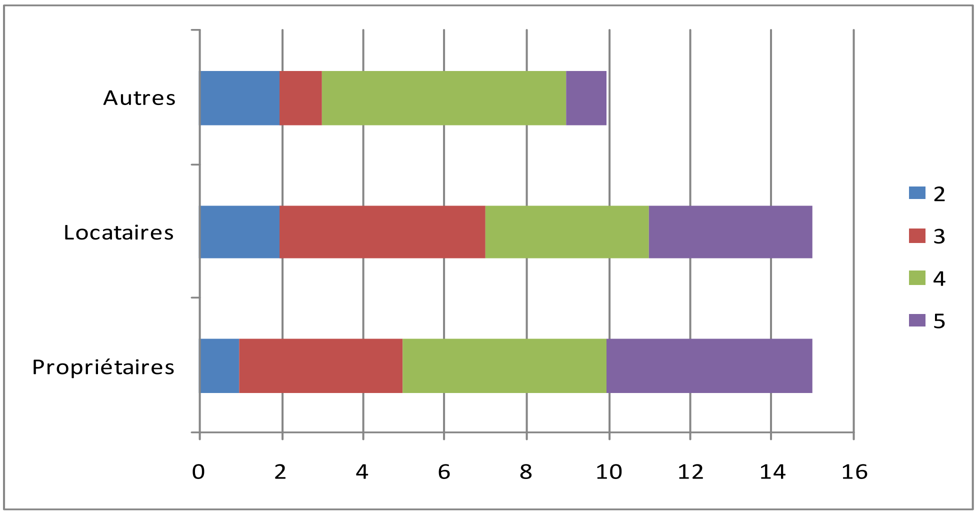
quantitatif et en Y les modalités du caractère qualitatif

Figure 32 Graphe1 quantitatif & qualitatif



Ou on trace en X les modalités du caractère qualitatif et en Y les modalités du caractère quantitatif.

Figure 33 Graphe2 qualitatif & quantitatif



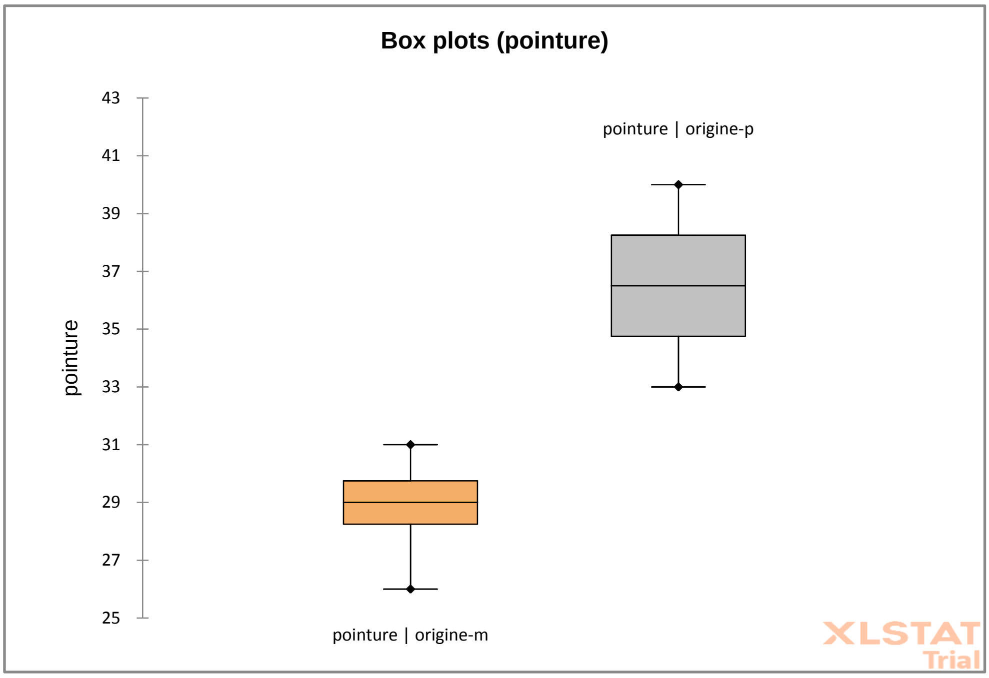
. On peut illustrer une table de contingence variable qualitative – variable quantitative par un diagramme Box plots (boîte à moustaches) : en abscisse on saisit les modalités de la variable qualitative et en ordonnée les valeurs de la variable quantitative.

Tableau 85 EX.38 Distribution des pointures en fonction de l'origine

origine	taille
m	26
p	40
m	31
m	29
p	33
m	30
m	29
m	28
m	29
m	33
m	32
m	31
.	.

Source : Wibinar Xlstat

Figure 34 Graphe Boîte à moustache



Source : Wibinar Xlstat

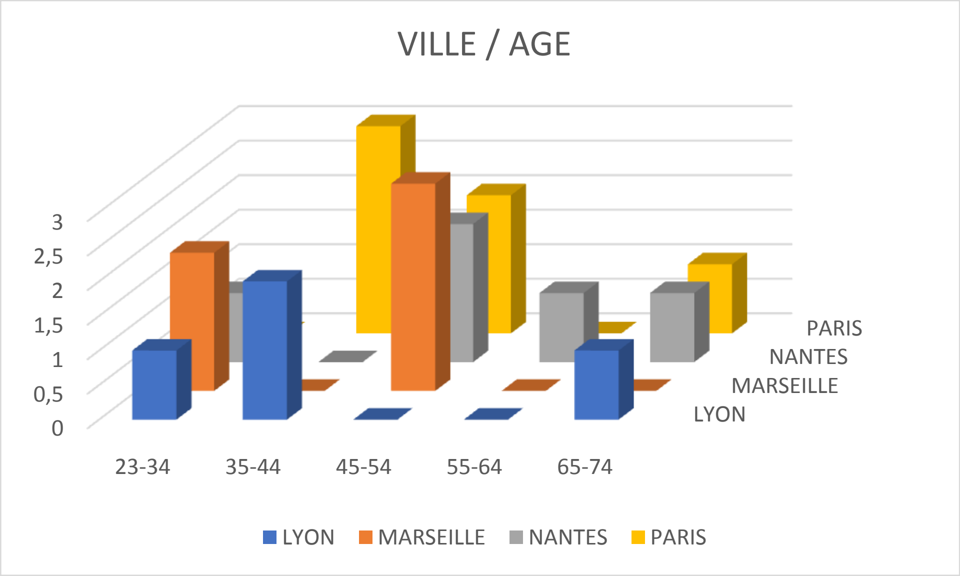
- *1 caractère qualitatif et 1caractère quantitatif continue* : tuyaux d'orgue

En X, les modalités du caractère quantitatif continue (base de chaque tuyau d'orgue) ; en Y, le caractère qualitatif (f_i%↑) ;

Tableau 86 EX.39 Tableau de contingence

AGE \ VILLE	LYON	MARSEILLE	NANTES	PARIS
23-34	1	2	1	0
35-44	2	0	0	3
45-54	0	3	2	2
55-64	0	0	1	0
65-74	1	0	1	1

Figure 35 Graphe qualitatif et quantitatif continue



iii. Variables quantitatives

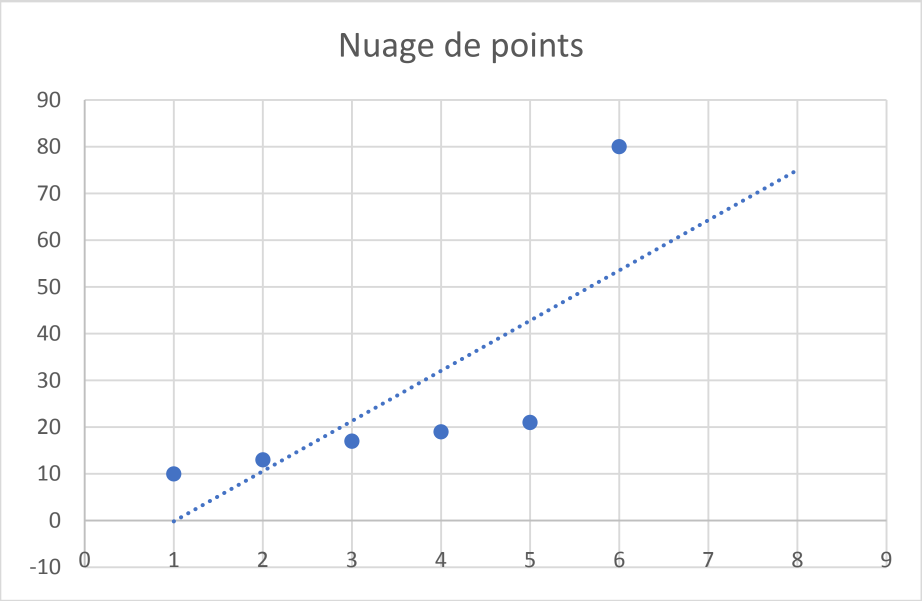
- 2 caractères discrets: nuage de points

Pour une variable quantitative, discrète ou continue, on peut utiliser une représentation par un **nuage de points** dans un plan. On peut remplacer chaque point par un ***cercle*** délimitant une *aire proportionnelle à l'effectif ou à la fréquence*.

Tableau 87 EX.40 Tableau de 2 variables quantitatives

X_i	Y_i
22	10
22.5	13
23	17
23.5	19
24	21
115	80

Figure 36 Graphe en nuage de points



- *2 caractères continus : stéréogramme*

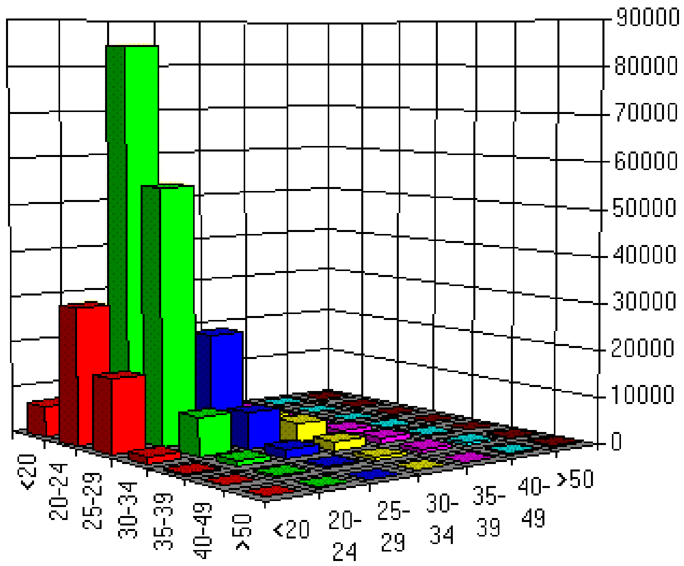
. *EXEMPLE*

Mariages célébrés en 1962, suivant l'âge des époux (1e colonne : âge de l'époux, 1e ligne : âge de l'épouse).

Tableau 88 EX.41 Tableau de contingence de 2 variables continues

	<20	20-25	25-30	30-35	35-40	40-50	>50
<20	6756	3051	180	15	3	3	0
20-25	29416	84556	13430	1205	168	50	10
25-30	15893	54978	22774	3890	651	113	14
30-35	1789	8289	7809	4111	1021	244	15
35-40	255	1304	1996	2078	1232	362	20
40-50	66	283	447	733	852	697	120
>50	6	46	59	83	145	336	472

Figure 37 Graphe d'1 distribution de 2 variables continues

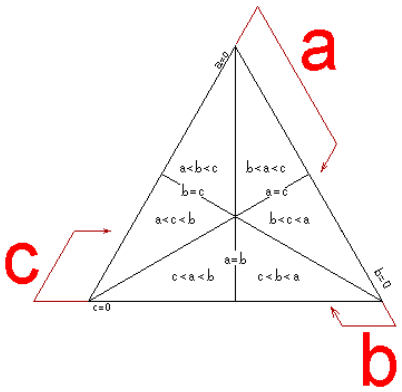


Le volume de chaque parallélépipède est proportionnel à l'effectif

iv. Autres représentations

D'autres représentations sont citées en littérature entre autres la représentation en étoile et la représentation triangulaire ci-dessous.

Figure 38 Graphe triangulaire



2.6 Covariance, corrélation et régression

Dans l'étude de la dépendance entre 2 variables quantitatives X et Y, on se pose trois questions fondamentales :

-X et Y sont-ils liés, comment mesurer cette liaison ?

-Trouver une fonction qui permet de déterminer Y à partir de X

-Estimer les paramètres de cette fonction ?

Évaluer la liaison entre X et Y de manière à mettre en évidence

-le « sens » de la liaison ;

- la « force » de la liaison

. Covariance

cov(x,y) mesure la dépendance ou non dépendance entre x et y:

cov(x,y)= m_{xy}=

$$\frac{\sum (x_i - \bar{x})(y - \bar{y})}{n} = \frac{\sum_i x_i y_i}{n} - \bar{x}\bar{y}.$$

$\boxed{\text{cov}(x,y)=0 \Leftrightarrow x, y \text{ indépendants}}$

.Coefficient de corrélation

Le coeff. de corrélation **r** mesure le pouvoir explicatif du modèle.

Il met en i/i la dispersion expliquée par le modèle et la dispersion totale du modèle

$$\boxed{\mathbf{r} = \frac{\text{cov}(x,y)}{\sigma_x \sigma_y} = \frac{\sum_i x_i y_i - n\bar{x}\bar{y}}{\sqrt{(\sum_i x^2{}_i - n\bar{x}^2)(\sum_i y^2{}_i - n\bar{y}^2)}}}$$

. EXEMPLE

Sur la base tableau 76 des fréquences conjointes repris ci-dessous, calculer la covariance [cov (X, Y)] entre le salaire et l’âge, la corrélation [r(X,Y)] et l’équation de la droite de régression linéaire Y=aX+b ?

Tableau 89 EX.42 Tableau1 de contingence

X \ Y	25	35	45	55	$n_{i.}$
3	3	1	1	0	5
5	1	5	0	0	6
7	0	1	3	0	4
9	0	0	1	2	3
11	0	0	2	0	2
$n_{.j}$	4	7	7	2	20

$$\operatorname{cov}(X,Y)=\frac{\sum_{i=1}^{5}\sum_{j=1}^{4} n_{ij}x_i y_j}{n_{..}}-\overline{xy}$$

$$\Leftrightarrow \operatorname{cov}(X,Y)=\frac{3\times 75+1\times 105+1\times 125+5\times 175+1\times 245+3\times 315+1\times 405+2\times 495}{20}=17.15 \quad (>0)$$

Ceci signifie que l'âge et le salaire ne sont pas indépendants.

. Corrélation

Le coefficient de corrélation linéaire s'évalue par :

$r(x,y)=\dfrac{\operatorname{cov}(X,Y)}{\sigma_x\sigma_y}=\dfrac{17.15}{\sqrt{6.59}-\sqrt{82.75}}=0.73$, Il exprime une forte relation positive entre l'âge et le salaire.

a. Régression linéaire

Si x et y sont liés, la régression consiste à trouver une fonction permettant de déterminer y à partir de x. La forme de la fonction peut être linéaire (y=ax+b) ou linéarisable (y=log x+b) ;

cas linéaire : y = ax+b

$$\boxed{\hat{a}=\frac{\operatorname{cov}(x,y)}{v(x)}=\frac{\sum_i x_i y_i - n\bar{x}\bar{y}}{\sum_i x^2_i - n\bar{x}^2} \quad ; \quad \hat{b}=\bar{y}-\hat{a}\bar{x}}\,.$$

Démonstration

$$\min\left(\sum_i e_i^2\right) = \min(yi-\hat{y})^2 = \min\sum_i (yi-\hat{a}x-\hat{b})^2$$

$$\frac{\partial\sum_i e_i^2}{\partial\hat{a}} = 0 \Leftrightarrow \frac{\partial\sum_i (yi-\hat{a}x-b)^2}{\partial\hat{a}} = \sum[2.\ (yi-\hat{a}x-\hat{b})\ .-x_i] = 0 \Leftrightarrow \hat{a}\sum x_i y_i - \hat{a}\sum x_i^2 - \hat{b}\sum x_i = 0$$

$$\Leftrightarrow \hat{a}\sum x_i y_i - \hat{a}\sum x_i^2 - (\overline{y}-\hat{a}\overline{x})\sum x_i = 0 \Leftrightarrow \hat{a}\sum x_i y_i - \hat{a}\sum x_i^2 - \overline{y}\sum x_i - \hat{a}\overline{x}\sum x_i = 0 \Leftrightarrow$$

$$\hat{a} = \frac{\sum_i x_i y_i - n\overline{x}\overline{y}}{\sum_i x^2{}_i - n\overline{x}^2} = \frac{\mathrm{cov}(x,y)}{v(x)}$$

$$\frac{\partial\sum_i e_i^2}{\partial\hat{b}} = 0 \Leftrightarrow \frac{\partial\sum_i (yi-\hat{a}x-b)^2}{\partial\hat{b}} = 0 \Leftrightarrow \sum[2.\ (yi-\hat{a}x-\hat{b})\ .-1] = 0 \Leftrightarrow \sum\ (yi-\hat{a}x-\hat{b}) = 0 \Leftrightarrow$$

$$\sum y_i - \hat{a}\sum x_i - n\hat{b} = 0 \Leftrightarrow \hat{b} = \overline{y} - \hat{a}\overline{x}$$

. EXEMPLE

$$\hat{a} = \boxed{\frac{\mathrm{cov}(x,y)}{v(x)} = \frac{17.15}{6.59} = 2.60}$$

$$\boxed{\hat{b} = \overline{y} - \hat{a}\overline{x} = 0.67 - 2.6 * 6.10 = -15.19}$$

L'équation de la droite de régression est alors : Y=2.60X – 15.19

b. Régression linéaire pondérée

Soit le tableau de contingence suivant, calculer la droite de régression ? :

Tableau 90 EX.43 Tableau 2 de contingence

	20	30	40
25	35	6	2
35	5	25	7
45	2	3	15

$$\hat{a} = \frac{\operatorname{cov}(x,y)}{v(x)} = \frac{\sum_i n_i x_i y_i - n\bar{x}\bar{y}}{\sum_i n_i x^2{}_i - n\bar{x}^2} = 4.48; \quad \hat{b} = \bar{y} - \hat{a}\bar{x} = 0.725$$

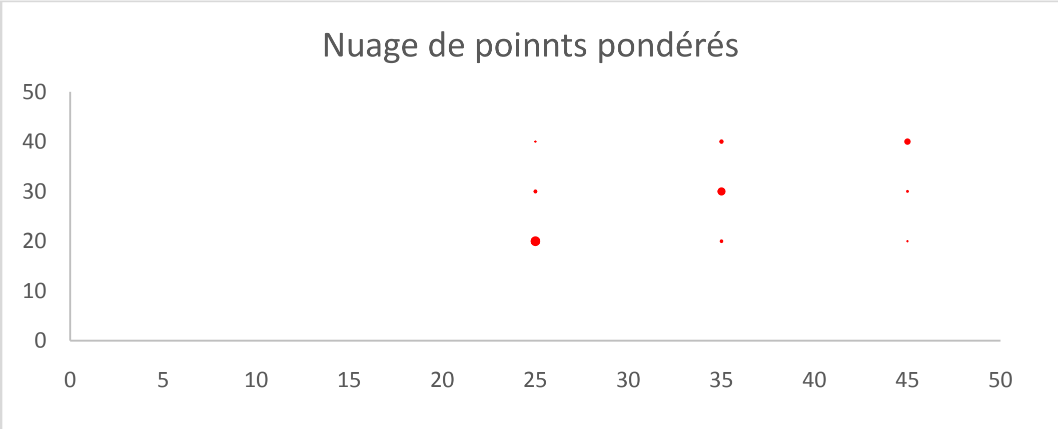
$$r = \frac{\operatorname{cov}(x,y)}{\sigma_x \sigma_y} = \frac{\sum_i n_i x_i y_i - n\bar{x}\bar{y}}{\sqrt{(\sum_i n_i x^2{}_i - n\bar{x}^2)(\sum_i n_i y^2{}_i - n\bar{y}^2)}}$$

Tableau 91 Solution Calcul de la droite de régression

X_i	Y_i	n_i	ni Xi Yi	Xi²	ni Xi²	ni Xi	ni Yi
25	20	35	17500	625	21875	875	700
35	20	5	3500	1225	6125	175	100
45	20	2	1800	2025	4050	90	40
25	30	6	4500	625	3750	150	180
35	30	25	26250	1225	30625	875	750
45	30	3	4050	2025	6075	135	90
25	40	2	2000	625	1250	50	80
35	40	7	9800	1225	8575	245	280
45	40	15	27000	2025	30375	675	600
315	270	100	96400		112700	3270	2820
3.15	2.7						
32.7	28.2						
1069.29							

La représentation graphique se fait par un nuage de points proportionnels.

Figure 39 Graphe de la régression pondérée



CH3. STATISTIQUE MULTIVARIEE

Soit une population Ω de taille n décrite simultanément par rapport à trois caractères ou plus x, y, z … respectivement de k, p, r … modalités.

3.1 ACP

3.1.1 Formalisation

$$\textit{Soit } \mathrm{X}=\left[\underline{X}_1, \ \ldots \ , \ \underline{X}_n\right] \quad \text{n points} \ / \quad \underline{\mathrm{X}}_{\mathrm{i}} \in R^d$$

$$\mathrm{Pr}\,\textit{ojetons } \mathrm{X} \text{ sur } \vec{\mathrm{U}} \ / \ \dim\vec{\mathrm{U}} < \mathrm{d} \text{ et var } (\mathrm{U}^t\mathrm{X})_{\max}$$

$$\text{var } (\mathrm{U}_i{}^t\mathrm{X})_{\max} \quad \equiv \quad \underset{\underline{u}_i}{Max}\left[\text{var } (\mathrm{U}_i{}^t\mathrm{X})\right]$$

$$\underset{\underline{u}_i}{Max}\left[\text{var } (\mathrm{U}_i{}^t\mathrm{X})\right]$$

$$\textit{Soit } \text{var}(X) = S_{d,d}$$

$$\text{var } (\mathrm{U}_i{}^t\mathrm{X}) = \mathrm{U}_i{}^t S \ U_i$$

$$\underset{\underline{u}_i}{Max}\left[\text{var } (\mathrm{U}_i{}^t\mathrm{X})\right] = \underset{\underline{u}_i}{Max}\left(\mathrm{U}_i{}^t S \ U_i\right)$$

1. méthode du Lagrangien

$$\underset{\underline{u}_i}{Max}\left[\text{var } (\mathrm{U}_i{}^t\mathrm{X})\right] = \underset{\underline{u}_i}{Max}\left(\mathrm{U}_i{}^t S \ U_i\right) \ \equiv \left| \underset{\underline{u}_i}{Max}\left(\mathrm{U}_i{}^t S \ U_i\right) \text{ si } \mathrm{U}_1^t U_1 = 1\right.$$

$$L(U_1, \lambda) = \mathrm{U}_i{}^t S \ U - \lambda(\mathrm{U}_1^t U_1 - 1)$$

$$\frac{\partial L}{\partial \underline{U}_1} = 2S\underline{U}_1 - 2\lambda U_1 = 0 \ \Leftrightarrow \ \mathrm{SU}_1 = \lambda_1 \underline{U}_1 \ \left| \mathrm{U}_1 : \textit{vecteur} \text{ propre}; \ \lambda_1 : \textit{valeur}\right.$$

$$\underline{U}^t \ S \ \underline{\mathrm{U}} = U^t \lambda \ U = \lambda U^t U = \lambda$$

S possède au plus d vecteurs propres et d valeurs propres

$$\left.\begin{array}{l} \lambda_1 > \lambda_2 > \lambda_3 > \ldots > \lambda_d \\ U_1 > U_2 > U_3 > \ldots > U_d \end{array}\right\} \mathrm{U}_1 \text{ de S correspond au vecteur propre maximum et}$$

constitue le première composante principale.

Si l'on choisit U_2 comme solution, U_2 correspond à $\lambda_2 \ / \ \lambda_2 < \lambda_1$ par suite on prend U_1 comme meilleur solution.

2. La méthode SVD

$X_{d,n} = U \Sigma V^T$

U = vecteur propre de XX^T,

$\Sigma = \begin{bmatrix} \lambda_1 & & \\ & \lambda_2 & \\ & & \lambda_d \end{bmatrix}$ valeurs propres de XX^T *ou* $X^T X$

V= vecteurs propres de $X^T X$

$\left[X_{d,n} - U_{d,n}\right] = \tilde{X}$ /

/ X= [ ] U=[ ] $\underline{\bar{X}} = \frac{1}{n}\sum_{i=1}^{n} \underline{X}_i$

$U \Sigma V^T = \tilde{X}$

$\tilde{X}\tilde{X}^T \begin{bmatrix} \lambda_1 & & \\ & \lambda_2 & \\ & & \lambda_d \end{bmatrix} \tilde{X}^T \tilde{X}$ / $\tilde{X}\tilde{X}^T = \mathrm{var}(X) = S$

Pour trouver U_i^* : 1.*centraliser* X; 2. faire SVD: colonnes ou PC(U_i)

Passage de R^d à R^p *puis* renvoie de l'image obtenue à R^d

$X_{d,n} = \left[\underline{X}_1, \ \ldots, \underline{X}_n\right]$

$Y_{1,1} = U_{1d}^T X_{d,1} \Rightarrow \widehat{X}_{d,1} = U_{d,1}\ Y_{1,1}$

$Y_{2,1} = U_{2,d}^T X_{d,1} \Rightarrow \widehat{X}_{d,1} = U_{d,2}\ Y_{2,1}$:

... = ... $\Rightarrow$ ... =...

L'image (obtenue dans R) renvoyée de R origine c_1 est meilleure car .....

3.*Méthode* Pr *atique* (Heuristique)

$X_{d,n} > XX'$ réduite $\Rightarrow \tilde{X}\ \tilde{X}' >$ orthogonalisation $\mathrm{U}_i > choix$ du nombre de U_i sur inertie expliquée gd $\dfrac{\lambda_1 + \lambda_{cti}}{\sum_i \lambda_i} = grande$ (90%) $(\sum_i \lambda_i = tr\ \mathrm{XX}') > normalisation\ \mathrm{U}_i > \lambda'_{p,n} = U'_{p,d} X_{d,n}$

-

. EXEMPLE 1

Un constructeur de piscine désire évaluer l'influence du revenu du propriétaire, de la surface de la propriété et de l'attitude envers le soleil dans la décision d'achat d'une piscine. Le service Marketing présente les données suivantes d'un échantillon de 24 clients potentiels.

Tableau 92 Exemple 1 d'ACP

i	R	S	SO
A	9,3	30,2	8
B	10,2	40,1	10
C	9,7	35,3	6
D	11,5	45,1	4
E	12	38	5
F	14,2	50,1	9
G	18,6	60,2	10
H	28,4	100,4	3
I	13,2	25,1	2
J	10,8	40,7	7
K	22,7	68,4	9
L	12,3	60,3	8
M	8,2	38,2	2
N	7,8	40,2	0
O	11,4	44,8	1
P	16,3	50,6	4
Q	12,4	42,5	8
R	11,5	60,3	0
S	6,8	39,7	6
T	10,4	35,4	4
U	14,2	42,6	3
V	11,6	38,4	2
X	8,4	30,2	4
Y	9,1	25,7	5

Source : Initiation à l'analyse des données Jean Lagarde p. 85

3.1.2 Exemple

Tableau 93 Exemple 2 d'ACP

i	P	Tmax	Tmin
1 Ajaccio	12,04	23,7	5,9
2 Brest	17,18	15,5	-1,8
3 Dunkerque	11,83	13,1	2,8
4 Nancy	6,23	13,5	-2,4
5 Nice	16,99	21,1	7,2
6 Toulouse	3,87	20,3	-0,90

Source : Pages J., Rennes

Cet exemple a été résolu par le logiciel Xlstat 2024.

0. Présentation des données centrées réduites

i	P	Tmax	Tmin
1 Ajaccio	0,14	1,44	1,09
2 Brest	1,17	-0,59	-0,96
3 Dunkerque	0,1	-1,18	0,27
4 Nancy	-1,03	-1,08	-1,12
5 Nice	1,13	0,8	1,44
6 Toulouse	-1,5	0,6	-0,72

1. Statistiques descriptives :

Variable	Observations	Obs. avec données manquantes	Obs. sans données manquantes	Minimum	Maximum	Moyenne	Ecart-type
P	6	0	6	-1,500	1,170	0,002	1,094
Tmax	6	0	6	-1,180	1,440	-0,002	1,094
Tmin	6	0	6	-1,120	1,440	0,000	1,098

2.Matrice de corrélation (Pearson (n)) :

Variables	P	Tmax	Tmin
P	**1**	0,085	0,486
Tmax	0,085	**1**	0,624
Tmin	0,486	0,624	**1**

3. Valeurs propres :

	F1	F2	F3
Valeur propre	1,833	0,917	0,249
Variabilité (%)	61,110	30,583	8,307
% cumulé	61,110	91,693	100,000

4. Vecteurs propres :

	F1	F2	F3
P	0,459	0,791	-0,406
Tmax	0,562	-0,612	-0,557
Tmin	0,688	-0,027	0,725

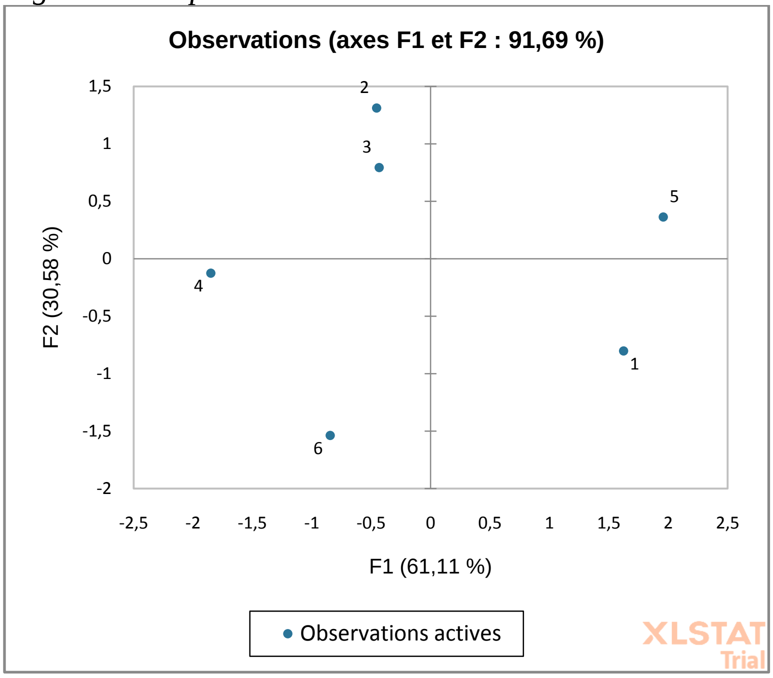
5. Coordonnées des variables :

	F1	F2	F3
P	0,621	0,757	-0,203
Tmax	0,761	-0,586	-0,278
Tmin	0,932	-0,026	0,362

6 Coordonnées des observations :

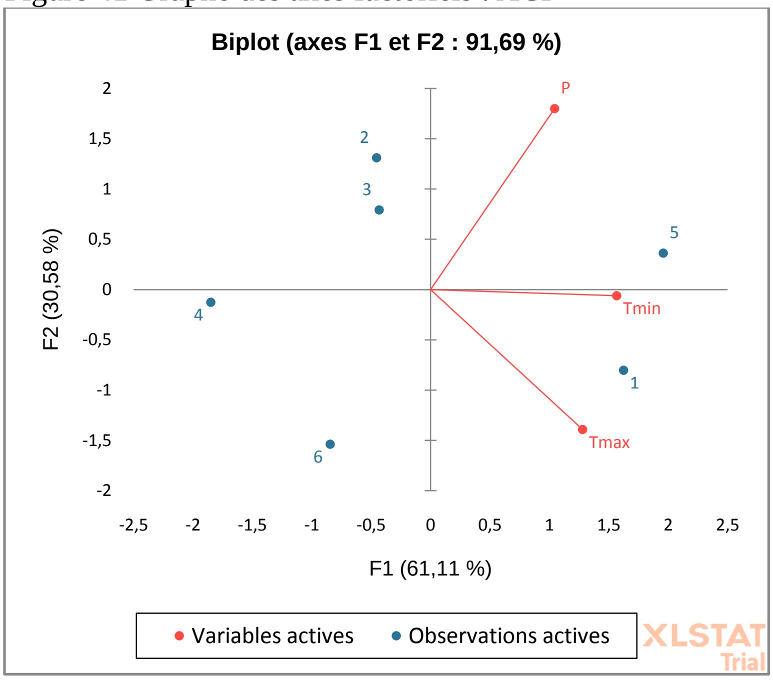
	F1	F2	F3
1	1,624	-0,803	-0,072
2	-0,454	1,311	-0,841
3	-0,433	0,792	0,812
4	-1,850	-0,126	0,210
5	1,958	0,363	0,136
6	-0,845	-1,538	-0,246

Figure 40 Graphe des observations : ACP



Source : Résultat Xlstat 2024

Figure 41 Graphe des axes factoriels : ACP



Source : Résultat Xlstat 2024

3.2 Régression Logistique

3.2.1 Formalisation

Le modèle « Logit binaire multiple» est un modèle statistique, à l'instar du modèle « Probit »,où la variable dépendante est qualitative nominale binaire et les variables indépendantes sont un mélange de variables quantitatives et catégorielles.

Soit y_i une variable nominale binaire $y_i \sim \beta(1, p_i)$ / $y_i \in \{0,1\}$, P(yi=1)=pi; P(yi=0)=1-pi.

Essayons de regresser y_i sur k prédicteurs $(x_{i1}, x_{i2}, ..., xi_k) = X_i^{'}$ à l'instar de la régression linéaire

$y_i = X_i^{'}\beta + u_i$. *Comme* il est impossible de définir la régression à partir de $p(y=+/x)=a_0 + a_1x_1 + . + a_kx_k$,

le premier terme étant dans $\{0,1\}$ alors que le second est dans R^k.

Une solution à ce problème passerait par la création d'une variable latente y^* *telle* que:

$y = 1 \Leftrightarrow y^* = X_i^{'}\ \beta + u_i > 0$. Il s'ensuit : $p_i = P(y_i = 1) = P(y^* > 0) = P(X_i^{'}\ \beta + u_i > 0) =$

$P(\ u_i > -X_i^{'}\ \beta) = 1 - F(-X_i^{'}\ \beta)$ c.a.d. $y_i \sim \beta(1, p_i)$ *avec* $p_i = 1 - F(-X_i^{'}\ \beta) = \dfrac{1}{1+e^{-X_i^{'}\beta}}$

$F(X_i^{'}\ \beta\ ) = \dfrac{1}{1+e^{-X_i^{'}\beta}}$ pour le Logit d'où $X_i^{'}\beta\ = \ln\left(\dfrac{p}{1-p}\right) = \ln\ (\textit{Odds}\ p)$.

$F(X_i^{'}\ \beta\ ) = \Phi(X_i^{'}\ \beta)\ = \dfrac{1}{\sqrt{2\pi}}\displaystyle\int_{-\infty}^{X_i^{'}\beta} e^{-z^2/2}dz$ pour le Probit d'où $X_i^{'}\beta = \Phi^{-1}$.

$$P(y = 1/x = x_o) = \frac{1}{1+e^{-X_i^{'}\beta}} = \frac{e^{X_i^{'}\beta}}{1+e^{X_i^{'}\beta}}$$

$$P(y = 0/x = x_o) = \frac{1}{1+e^{X_i^{'}\beta}}$$

$$f(yi/x = xi) = \left(\frac{e^{X_i^{'}\beta}}{1+e^{X_i^{'}\beta}}\right)^{yi} \left(\frac{1}{1+e^{X_i^{'}\beta}}\right)^{1-yi}$$

Par MLE, $\beta_{MLE} = \arg Max$

$$L(; \beta) = \prod_i \mathrm{P(y=+/x)}^{\mathrm{y_i}} \left[\mathrm{1\text{-}P(y=+/x)}^{1-\mathrm{y_i}} \right]$$

$$LL\ (\beta) = \sum Log\left[f(x_i\ ; \beta) \right] = \sum yi \left[\mathrm{X}'\beta\text{-Log}(1+e^{X_i'\beta}) \right] + (1-yi) - Log(1+e^{X_i'\beta})$$

$$= \sum \mathrm{yi}\ \beta'\ \mathrm{x_i}\ \text{- Log}(1+e^{X_i'\beta})$$

$i-Calcul$ du gradient $\underset{\beta}{\nabla}\mathrm{L}(\beta)$

$$\underset{\beta}{\nabla}\mathrm{L}(\beta) = \underset{\beta}{\nabla}\left[\sum_i yi\ x'\beta\ \text{- ln(1+e}^{x'\beta}) \right] = \left[\sum_i \underset{\beta}{\nabla}\ yi\ x'\beta\ \text{- ln(1+e}^{x'\beta}) \right] = \left[\sum_i \underset{\beta}{\nabla}(yi\ x_i\ '\beta)\ \text{-}\ \underset{\beta}{\nabla}[\text{ln(1+e}^{x'\beta})] \right]$$

$$= \sum_i yi\ \mathrm{xi\ \text{-}\ xi}\ \frac{\mathrm{e}^{x'\beta}}{1+\mathrm{e}^{x'\beta}} = \sum_i xi(yi - pi) = \mathrm{X}^T(\mathrm{Y}-\hat{Y}) \quad /\mathrm{X} = \begin{pmatrix} x_{11} & x_{12} & & x_{1k} \\ & & & \\ x_{n1} & x_{n2} & & x_{nk} \end{pmatrix}_{n,\mathrm{k}} \quad : \text{"design matrix ";}$$

$$(\mathrm{y}_i - \hat{\mathrm{y}}_i)_{n,1}$$

$ii-Calcul$ du Hessian $\underset{\beta\beta}{\nabla} L(\beta)$

$$\underset{\beta\beta}{\nabla}\mathrm{L}(\beta) = \underset{\beta}{\nabla}[\ \sum_i x_i(y_i - p_i)] = \sum_i \underset{\beta}{\nabla}[x_i(y_i - p_i)] = \sum_i \underset{\beta}{\nabla}(-p_i\ \mathrm{xi})] = \sum_i \underset{\beta}{\nabla}(-\frac{1}{1+e^{-x'\beta}}\ \mathrm{xi})] =$$

$$\sum_i -(-1)\left[\frac{1}{1+e^{-x'\beta}} \right]^2 (-\mathrm{xi})\ e^{-x'\beta}\ \mathrm{xi})] = -\sum_i \left[\frac{1}{1+e^{-x'\beta}} \right]\left[\frac{e^{-x'\beta}}{1+e^{-x'\beta}} \right](-\mathrm{xi})\ \ \mathrm{xi})] = -x^T p(1-p)x =$$

$$-X^T\mathrm{W\ X} \quad /\mathrm{W} = \begin{pmatrix} p_1(1-p_1) & 0 \\ 0 & p_n(1-p_n) \end{pmatrix};$$

iii- Calcul des paramètres optimaux β* par la méthode itérative de Newton-Raphson

$$\beta_{t+1} = \beta_t - H^{-1}g$$

$$\beta_{t+1} = \beta_t - \frac{\underset{\beta}{\nabla}\mathrm{L}(\beta)}{\underset{\beta\beta}{\nabla}\mathrm{L}(\beta)} \Leftrightarrow \beta_{t+1} = \beta_t + (\mathrm{X}^T\mathrm{WX})^{-1}X^T(\mathrm{Y}\text{-}\mathrm{Y}^t)$$

Remarque : si H n'est pas inversible,il faut appliquer Levenberg Marquardt method ou gradient descent

3.2.2 Exemple

Soient les données suivantes ; cherchons par la régression logistique un modèle de prédiction de la maladie cardiaque en fonction des variables age, tx max et angine : cœur =f(age, tx-max et angine) / cœur = variable binaire ; age et tx max=variables quantitatives; angine= variable binaire (Résolu par Xlstat 2024).

1. trouver $\hat{a}_1$ $\hat{a}_2$ $\hat{a}_3$?

Tableau 94 Exemple 1 Regression Logistique

age	tx_max	angine	cœur	y
50	126	1	présence	1
49	126	0	présence	1
46	144	0	présence	1
49	139	0	présence	1
62	154	1	présence	1
35	156	1	présence	1
67	160	0	absence	0
65	140	0	absence	0
47	143	0	absence	0
58	165	0	absence	0
57	163	1	absence	0
59	145	0	absence	0
44	175	0	absence	0
41	153	0	absence	0
54	152	0	absence	0
52	169	0	absence	0
57	168	1	absence	0
50	158	0	absence	0
44	170	0	absence	0
49	171	0	absence	0

Source : Rakotomala, Lyon

Statistiques descriptives (Données quantitatives) :

Variable	Observations	Obs. avec données	Obs. sans données	Minimum	Maximum	Moyenne	Ecart-type
age	20	0	20	35,000	67,000	51,750	8,162
tx_max	20	0	20	126,000	175,000	153,850	14,438

Statistiques descriptives (Données qualitatives) :

Variable	Modalités	Comptages	Effectifs	%
y	0	14	14	70,000
	1	6	6	30,000
Statistiques descriptives (Données quantitatives)				
angine	0	15	15	75,000
	1	5	5	25,000

Matrice de corrélation :

	age	tx_max	angine-0	angine-1	y
age	**1**	-0,025	-0,033	0,033	-0,267
tx_max	-0,025	**1**	0,018	-0,018	-0,606
angine-0	-0,033	0,018	**1**	-1,000	-0,378
angine-1	0,033	-0,018	-1,000	**1**	0,378
y	-0,267	-0,606	-0,378	0,378	**1**

Statistiques de multicolinéarité :

	age	tx_max	angine-0	angine-1
Tolérance	0,998	0,999	0,999	0,999
VIF	1,002	1,001	1,001	1,001

Régression de la variable y (Modalité témoin = 0) :

Coefficients d'ajustement (Variable y) :

Statistique	Indépendant	Complet
Observations	20	20
Somme des poids	20,000	20,000
DDL	19	16
-2 Log(Vraisemblance)	24,435	0,000
R²(McFadden)	0,000	1,000
R²(Cox and Snell)	0,000	0,705
R²(Nagelkerke)	0,000	0,901
AIC	26,435	8,000
SBC	27,430	11,983
Itérations	0	22

Test de l'hypothèse nulle H0 : Pr(y=1)=0,3 :

Statistique	DDL	Khi²	Pr > Khi²
-2 Log(Vraisemblance)	3	24,435	<0,0001
Score	3	11,764	0,008
Wald	3	0,000	1,000

Analyse de Type II (Variable y) :

Source	DDL	Khi² (Wald)	Pr > Wald	Khi² (LR)	Pr > LR
age	1	0,000	0,991	7,653	0,006
tx_max	1	0,000	0,991	19,702	<0,0001
angine	1	0,000	0,991	14,229	0,000

Test de Hosmer-Lemeshow (Variable y) :

Statistique	Khi²	DDL	Pr > Khi²
Statistique de Hosmer-Lemeshow	0,000	4	1,000

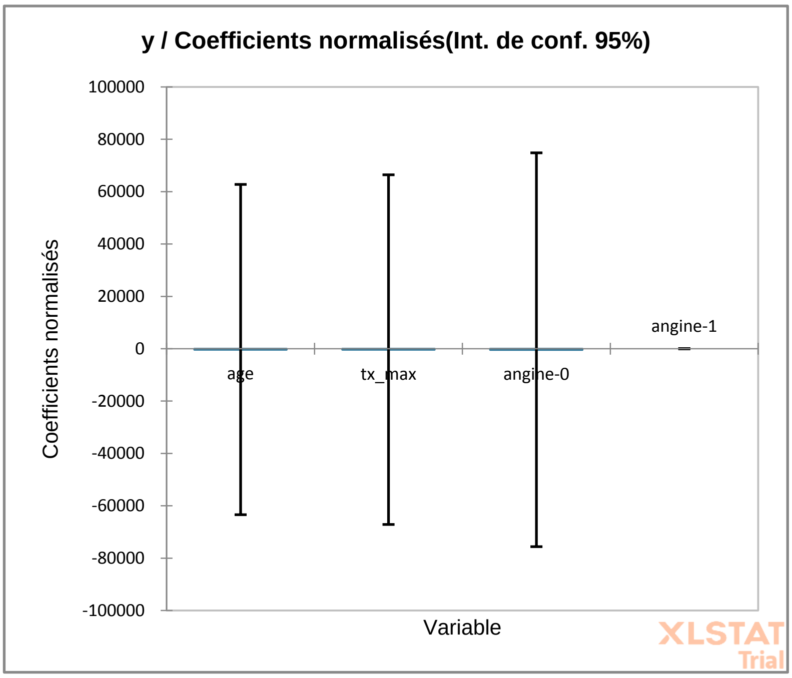
Paramètres du modèle (Variable y) :									
Source	Valeur	Erreur standard	Khi² de Wald	Pr > Khi²	Wald Borne inf. (95%)	Wald Borne sup. (95%)	Odds ratio	Odds ratio Borne inf. (95%)	Odds ratio Borne sup. (95%)
Constante	12160,951	1128151,818	0,000	0,991	##########	##########			
age	-78,190	7338,384	0,000	0,991	-14461,158	14304,778			
tx_max	-47,373	4391,898	0,000	0,991	-8655,335	8560,589			
angine-0	-1727,071	160751,109	0,000	0,991	-316793,455	313339,312			
angine-1	0,000	0,000							

. Equation du modèle (Variable y) :

Pr(y=1) = 1 / (1 + exp(-(12160,951272-78,190111*age-47,373017*tx_max-1727,071185*angine-0)))

Coefficients normalisés (Variable y) :

Source	Valeur	Erreur standard	Khi² de Wald	Pr > Khi²	Wald Borne inf. (95%)	Wald Borne sup. (95%)
age	-342,943	32186,233	0,000	0,991	-63426,799	62740,914
tx_max	-367,540	34074,208	0,000	0,991	-67151,761	66416,681
angine-0	-412,308	38376,500	0,000	0,991	-75628,866	74804,251
angine-1	0,000	0,000				



Prédictions et résidus (Variable y) :

Observation	y	Préd(y)	Pr(0)	Pr(1)
Obs1	1	1	0,000	1,000
Obs2	1	1	0,000	1,000
Obs3	1	1	0,000	1,000
Obs4	1	1	0,000	1,000
Obs5	1	1	0,000	1,000
Obs6	1	1	0,000	1,000
Obs7	0	0	1,000	0,000
Obs8	0	0	1,000	0,000
Obs9	0	0	1,000	0,000
Obs10	0	0	1,000	0,000
Obs11	0	0	1,000	0,000
Obs12	0	0	1,000	0,000
Obs13	0	0	1,000	0,000
Obs14	0	0	1,000	0,000
Obs15	0	0	1,000	0,000
Obs16	0	0	1,000	0,000
Obs17	0	0	1,000	0,000
Obs18	0	0	1,000	0,000
Obs19	0	0	1,000	0,000
Obs20	0	0	1,000	0,000

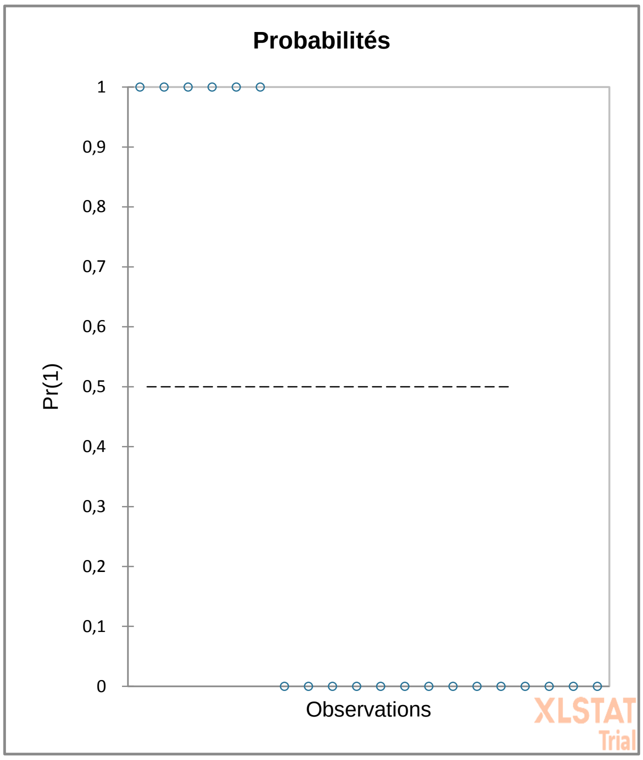
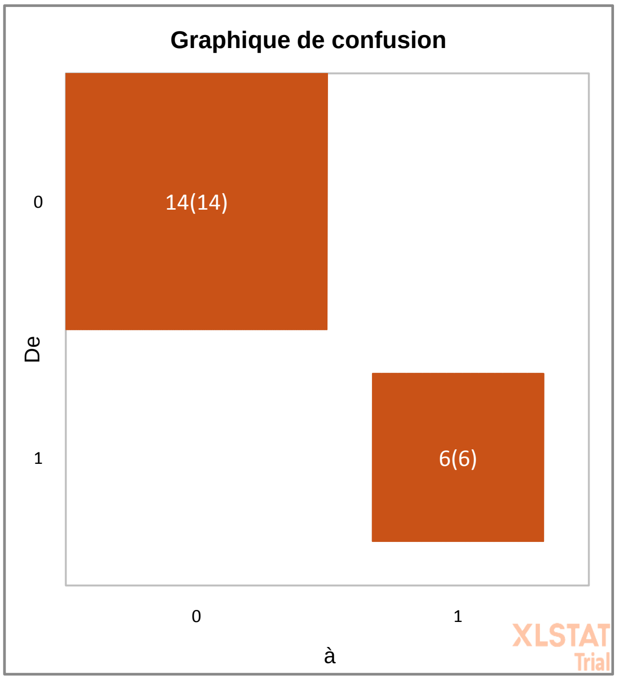
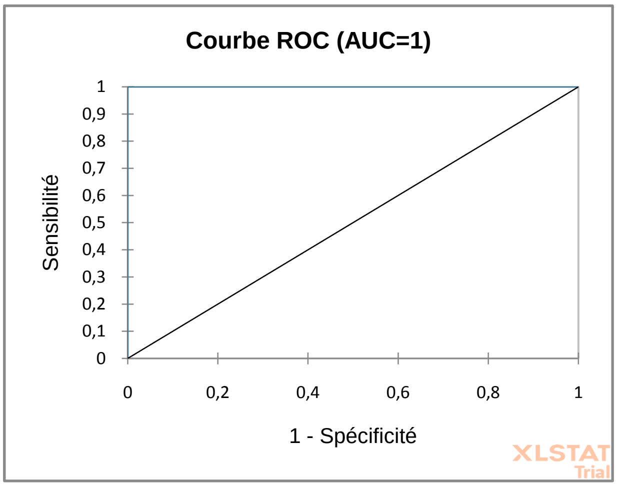


Tableau de classification pour l'échantillon d'apprentissage (Variable y) :

de \ Vers	0	1	Total	% correct
0	14	0	14	100,00%
1	0	6	6	100,00%
Total	14	6	20	100,00%





CH4. LES INDICES

4.1 Indices simples

Definition : Soit l'évolution temporelle d'1grandeur g : $g_0, g_1, .. g_t$ aux dates 0,1, ...t.

On appelle indice simple de g par rapport à la date t_0, le rapport $I_{t/0}(g) = \dfrac{g_t}{g_0}$

(la date 0 s'appelle date de référence, la date t date courante). Il permet de comparer l'évolution d'1 même grandeur (de plusieurs grandeurs différentes) sur 2 périodes.

4.1.1 Propriétés

Circularité (transférabilité)

$I_{t/0}(g) = I_{t/t'}(g) * I_{t'/0}(g)$ (1) puisque $g_t/g_0 = g_t/g_{t'} * g_{t'}/g_0$

$$(1) \Rightarrow I_{t/t'}(g) = \frac{I_{t/0}(g)}{I_{t'/0}(g)} \quad (2)$$

$$(2) \Rightarrow I_{t/t'}(g) = \frac{I_{t/1}(g)}{I_{t'/1}(g)} = \frac{I_{t/2}(g)}{I_{t'/2}(g)} = \ldots\ldots = \frac{I_{t/t-1}(g)}{I_{t'/t-1}(g)}$$

Ceci permet la comparaison entre 2 dates indépendamment de la date de référence.

Réversibilité

$$I_{0/t}(g) = \frac{g_0}{g_t} = \frac{1}{g_t/g_0} = \frac{1}{I_{t/0}(g)} \Rightarrow \boxed{I_{0/t}(g) = \frac{1}{I_{t/0}(g)}}$$

Enchaînement $\boxed{I_{t/0}(g) = I_{t/t-1}(g) * I_{t-1/t-2}(g) * I_{t-2/t-3}(g) * \ldots I_{1/0}(g)}$

L'indice simple est circulaire, réversible et enchaînable.

EXEMPLE 1

Le tableau suivant indique les prix unitaires d'un produit au cours de la période $2000-2003$.

Tableau 95 EX.44 Prix unitaires d'1 produit durant 2000-2003

AN	2000	2001	2002	2003
PRIX	120	130	125	135

1. *Calculer les indices* élémentaires des prix $I_{2001/2000}$? $I_{2002/2000}$? $I_{2003/2000}$
2. *Calculer l'indice des prix* $I_{2003/2002}$
3. *Calculer l'indice des prix* $I_{2002/2003}$?
4. *Calculer l'indice des prix* $I_{2003/2001}$?

. Solution

1. $I_{2001/2000} = \frac{130}{120}.100 = 108 \quad I_{2002/2000} = \frac{125}{120}.100 = 104 \quad I_{2003/2000} = \frac{135}{120}.100 = 112$

2. $I_{2003/2002} = I_{2003/2000} / I_{2002/2000} = \frac{112}{104}.100 = 107.7$ (transférabilité)

$I_{2002/2001} = I_{2002/2000} / I_{2001/2000} = \frac{104}{108} = 96.3$ (transférabilité)

3. $I_{2002/2003} = \frac{1}{I_{2003/2002}} = \frac{1}{107.7}.100 = 92.8$ (réversibilité)

4. $I_{2003/2001} = I_{2003/2002}.I_{2002/2001} = \frac{107.7*96.3}{100} = 104$ (enchaînement)

4.1.2 Opérations sur les indices

Addition

i. soit g_t une somme pondérée : $g_t = \sum_i a_i g_i$

$I_{t/0}(g) = \frac{g_t}{g_0} = \frac{\sum_i a_i g_i}{g_0} = \sum_i a_i \frac{g_i}{g_0}$ *l'indice d'1somme pondérée=la somme des indices pondérés*

ii. soit g_t et g_0 st des sommes pondérées: gt= $\sum_i a_i g_t^i$; $g_0 = \sum_i a^i g_0^i$; a coeff. cstts

$I_{t/0}(g) = \frac{\sum_i a_i g_t^i}{\sum_i a^i g_0^i} = \frac{\sum_i a_i g_0^i \frac{g_t^i}{g_0^i}}{\sum_i a^i g_0^i} = \sum_i w_i \frac{g_t^i}{g_0^i}$ *L'indice élémentaire d'1somme pondérée = moyenne arithmétique pondérée des indices élémentaires.*

iii. soit g_t et g_0 des sommes pondérées à coefficients variables :gt= $\sum_i a_t^i g_t^i$; $g_0 = \sum_i a_0^i g_0^i$; a coefficients variables.

$I_{t/0}(g) = \frac{g_t}{g_0} = \frac{\sum_i a_t^i g_t^i}{\sum_i a_0^i g_0^i} = \frac{\sum_i a_t^i g_0^i \frac{g_t^i}{g_0^i}}{\sum_i a_0^i g_0^i} = \frac{\sum_i a_t^i g_0^i I_{t/0}(g^i)}{\sum_i a_0^i g_0^i} = \sum_i w_i^t I_{t/0}(g^i)$

L'indice élémentaire d'1 somme pondérée = somme pondérée des indices.

MULTIPLICATION $\frac{a_t b_t}{a_o b_o} = \frac{a_t}{a_o}\frac{b_t}{b_o} \Leftrightarrow$ I $_{t/0}$(ab)= I $_{t/0}$(a)* I $_{t/0}$(b)

. EXEMPLE

Le prix du cuivre est passé de 0.212 dollars en 1950 à 0.258 dollars en 1958. Le cours du dollar est passé de 350 dinars algérien à 490. Ainsi, l'indice des prix du cuivre en dinars algérien est :

$$I_{1958/1950}(c_f) = I_{1958/1950}(d)\;\; I_{1958/1950}(c_d) =$$

$$= \frac{490}{350} * \frac{0.258}{0.212} = 1.7$$

DIVISION $\frac{a_t/b_t}{a_o/b_o} = \frac{a_t}{a_o}\Big/\frac{b_t}{b_o} \Leftrightarrow$ I $_{t/0}(a/b) = \frac{\text{I}_{t/o}(\text{a})}{\text{I}_{t/o}(\text{b})}$

. *EXEMPLE*

Le prix du litre de lait est passé de 36.8 en Algérie en 1950 algérien à 45.6 en 1958. Pour un citoyen américain, l'indice des prix du lait français est ainsi de :

$$I_{1958/1950}(L_d) = \frac{I_{1958/1950}(L_F)}{I_{1958/1950}(d)}$$

$$= \frac{\frac{45.6}{36.8}}{\frac{490}{350}} = \frac{1.24}{1.4} = 0.89$$

4.2 Indices synthétiques

4.2.1 Indices arithmétiques

Soient cette fois plusieurs produits X_j (j=1…n) dont on veut connaitre l'évolution globale au cours du temps i (i=1…m). X_{ij} dénote la valeur de la ième composante du produit X_j. x_{1j} est la valeur de base du produit Xj. On définit alors 3 indices arithmétiques : l'indice des moyennes, la moyenne des indices et la moyenne pondérée des indices.

Tableau 96 Indice arithmétique

L'indice des moyennes	*Le moyenne des indices*	*La moyenne pondérée des indices*
$I_{i/1} = 100.\dfrac{\sum_i x_{ij}}{\sum_j x_{1j}}$	$I_{i/1} = 100.\dfrac{1}{n}\sum_j \dfrac{x_{ij}}{x_{1j}}$	$I_{i/1} = \dfrac{\sum_i p_j (x_{ij}/x_{1j})}{\sum_j p_j}$

Propriétés : L'indice arithmétique n'est pas réversible, sa base n'est pas transférable.

. EXEMPLE

Soit la production de 3 produits (le blé, l'orge et le seigle) au cours des années 2000, 2001 et 2002. Si les pondérations associées au blé, à l'orge et au seigle sont respectivement de 100, 80 et 70. Calculer l'indice des moyennes, la moyenne des indices et la moyenne pondérée des indices ?

Tableau 97 EX.45 Prix de 3 produits durant 2000-2003

	Blé	Orge	Seigle
2000	5700	200	40
2001	6320	280	50
2002	6000	260	60

. Solution

1 .L'indice des moyennes

	Blé	*Orge*	*Seigle*	$\sum Xj$	$I_{i/1}$	base 100
2000	5700	200	40	5940	$\dfrac{5940}{5940} = 1$	100
2001	6320	280	50	6650	$\dfrac{6650}{5940} = 1.2$	120
2002	6000	260	60	6320	$\dfrac{6320}{5940} = 1.06$	106

2. Moyenne des indices

	Blé	*Orge*	*Seigle*	R_1	R_2	R_3	$1/3\sum_{j=1}^{3} R_j$	$I_{a/1}$
2000	5700	200	40	$\frac{5700}{5700}=1$	$\frac{200}{200}=1$	$\frac{40}{40}=1$	1	100
2001	6320	280	50	$\frac{6320}{5700}=1.11$	$\frac{280}{200}=1.4$	$\frac{50}{40}=1.25$	1.25	125
2002	6000	260	60	$\frac{6000}{5700}=1.05$	$\frac{260}{200}=1.3$	$\frac{60}{40}=1.5$	1.28	128

3. *Moyenne pondérée des indices*

	Blé	*Orge*	*Seigle*	R_1	R_2	R_3	p_1r_1	p_2r_2	p_3r_3	$\sum_{j=1}^{3} p_j rj / \sum p_j$	$I_{i/1}$
2000	5700	200	40	$\frac{5700}{5700}=1$	$\frac{200}{200}=1$	$\frac{40}{40}=1$	100	80	70	1	100
2001	6320	280	50	$\frac{6320}{5700}=1.11$	$\frac{280}{200}=1.4$	$\frac{50}{40}=1.25$	111	112	98	1.28	128
2002	6000	260	60	$\frac{6000}{5700}=1.05$	$\frac{260}{200}=1.3$	$\frac{60}{40}=1.5$	105	104	91	1.18	118

4.2.2 Indices géométriques

On utilise ici plutôt les moyennes géométriques simple et pondérée des indices au lieu des moyennes arithmétiques simples et pondérées.

Tableau 98 Indices géométriques

Indice géométrique	*Indice géométrique pondéré*
$I_{i/1}=100.\frac{[\prod_i x_{ij}]^{1/n}}{[\prod_i x_{1j}]^{1/n}}$	$I_{i/1}=[\prod_j (\frac{x_{ij}}{x_{1j}})^{\alpha_j}]^{1/\sum_j \alpha_j}+100$ $\log I_{i/1}=\frac{1}{\sum_j \alpha_j}\sum_i \alpha_j \log(\frac{x_{ij}}{x_{1j}})+2$

Propriétés : L'indice géométrique est réversible, sa base est transférable

$$G_{i/u} \,.\, G_{u/i} = \frac{[\prod_i x_{ij}]^{1/n}}{[\prod_i x_{uj}]^{1/n}} \cdot \frac{[\prod_i x_{uj}]^{1/n}}{[\prod_i x_{ij}]^{1/n}} = 1 \Rightarrow G_{u/i} = \frac{1}{G_{i/u}}$$

Soit $G_{i/x}$, *calculons* $G_{i/y}$: $G_{i/y} = G_{i/x}\ G_{x/y}$

$$\begin{cases} G_{i/y} = G_{i/x}\ G_{x/y} \\ G_{x/y} = \dfrac{1}{G_{y/x}} \end{cases} \Rightarrow G_{i/y} = \frac{G_{i/x}}{G_{y/x}}$$

. EXEMPLE

Soient les prix de 3 produits (la viande, le poisson et les œufs) au cours des années 2000, 2001 et 2002. Si les pondérations associées à la viande, au poisson et aux œufs sont respectivement de 0.6, 0.3 et 0.1. Calculer les indices géométriques et géométriques pondérés des différents produits ?

Tableau 99 EX.46 Prix de 3 produits durant 2000-2002

	Viande	Poisson	Œufs
2000	1700	400	200
2001	1800	500	250
2002	1900	660	300

Solution

a. Indice géométrique

	Viande	*Poisson*	*Œufs*	$\prod_{j=1}^{3} X_j$	$\left[\prod_{j=1}^{3} X_j\right]^{1/3}$	$\frac{I_{i/1}}{100}$	I_g
2000	1700	400	200	136000000	511.06	1	100
2001	1800	500	250	225000000	604.33	1.18	118
2002	1900	660	300	376200000	717.16	1.40	140

b. Indice géométrique pondéré

	Viande	*Poisson*	*Œufs*	$\prod_{j=1}^{3}\left(x_{ij}/x_{1j}\right)^{\alpha_j}$	$\left[\prod_{j=1}^{3}\left(x_{ij}/x_{1j}\right)^{\alpha_j}\right]^{1/\sum \alpha_j}$	I_{g_p}
2000	1700	400	200	1	1	100
2001	1800	500	250	1.1320	1.1320	113
2002	1900	660	300	1.2935	1.2935	129

4.2.3 Indices synthétiques Laspeyers, Paashe et Fisher

Définition : soit 1 grandeur G complexe : $G^1, G^2, \ldots G^i$ (Soit G le niveau des prix, Gi =prix de l'article i), l'indice élémentaire des constituants Gi est : $I_{t/o}(G)=\frac{G_t^i}{G_0^i}$.

Le problème est de synthétiser en un indice unique les indices des constituants G^i de G.

Soient G^1, G^2 ,…G^i : i biens de consommation.

Soient $G_{1/0}^1$, $G_{1/0}^2$,…$G_{1/0}^i$: les indices élémentaires y afférents de $1/0$

On veut synthétiser en 1 indice unique les indices élémentaires des différents biens pour connaître l'évolution de la consommation.

Les indices synthétiques Laspeyers, Paashe et Fisher emploient en plus des prix les quantités des produits.

Soit $w_0^i = \frac{p_o^i q_o^i}{\sum p_o^i q_o^i}$ l'importance relative de i dans la grandeur de l'époque t_o (période de base)

Soit $w_t^i = \dfrac{p_t^i q_t^i}{\sum_i p_t^i q_t^i}$:importance relative de i dans la grandeur de l'époque t_t (période courante)

$L_{1/0}(G^i) = \sum_i w_o^i \, I_{1/0}(G^i) = \sum_i w_o^i \dfrac{G_1^i}{G_o^i}$: moyenne arithmétique pondérée des indices élémentaires par coefficients w_o^i de date base

$P_{1/0}(G) = \dfrac{\sum_i w_t^i}{\sum_i w_t^i \dfrac{G_0}{G_1}} = \dfrac{1}{\sum_i w_t^i \dfrac{G_0}{G_1}}$: moyenne harmonique pondérée des indices simples par coefficients w_t^i de la date courante

$F_{1/0}(G^i) = \sqrt{L_{1/0}(G) P_{1/0}(G)}$: moyenne géométrique simple des indices de Paasche et Laspeyres

Tableau 100 Indice de Laspeyres $L_{t/0}(G)$

$L_{p\ 1/0}$:indice L des prix(p)	$L_{p\ 1/0} = \sum_i w_o^i \dfrac{p_1}{p_o} = \sum_i \dfrac{p_o q_o}{\sum_i p_o q_o} \dfrac{p_1}{p_o} = \dfrac{\sum_i p_1^i q_o^i}{\sum_i p_o^i q_o^i}$
$L_{q\ 1/0}$:indice L des quantités(q)	$L_{q\ 1/0} = \sum_i w_o^i \dfrac{q_1}{q_o} = \sum_i \dfrac{p_o q_o}{\sum_i p_o q_o} \dfrac{q_1}{q_o} = \dfrac{\sum_i p_o^i q_1^i}{\sum_i p_o^i q_o^i}$
$L_{pq\ 1/0}$ indice	$L_{pq\ 1/0} = \sum_i w_o^i \dfrac{p_1 q_1}{p_o q_o} = \sum_i \dfrac{p_o q_o}{\sum_i p_o q_o} \dfrac{p_1 q_1}{p_o q_O} = \dfrac{\sum_i p_1^i q_1^i}{\sum_i p_o^i q_o^i}$

. EXEMPLE

Soient 4 produits X_1, X_2, X_3 et X_4 dont les prix et les quantités sont donnés à deux périodes différentes 1 et t. Calculer les indices Lp, Lq et Lvg ?

Tableau 101 Prix de 4 produits durant périodes 1 et 2

	Période 1		Période t	
	P_{i1}	Q_{i1}	P_{it}	Q_{it}
X_1	12	6	15	7
X_2	5	13	8	11
X_3	15	9	13	18
X_4	8	10	10	9

Solution

	p_1	q_1	p_t	q_t	p_1q_1	p_tq_t	p_1q_t	p_tq_1
x_1	12	6	15	7	72	105	84	90
x_2	5	13	8	11	65	88	55	104
x_3	15	9	13	18	135	234	270	117
x_4	8	10	10	9	80	90	72	100
					352	517	481	411

$$L_p = \frac{p_tq_1}{p_1q_1} = \frac{411}{352}.100 = 116.7$$

$$L_q = \frac{p_1q_t}{p_1q_1} = \frac{481}{352}.100 = 136.6$$

$$L_{vg} = \frac{p_tq_t}{p_1q_1} = \frac{517}{352}.100 = 146.8$$

Indice de PAASCHE (P (G))

Tableau 102 Indice de Paasche $P_{t/0}$ (G)

$P_{p\ 1/0}$:indice P des prix(p)	$P_{p\ 1/0} = \dfrac{\sum w_t}{\sum w_t \dfrac{p_o}{p_t}} = \dfrac{1}{\sum \dfrac{p_1 q_1}{\sum p_1 q_1} \dfrac{p_o}{p_1}} = \dfrac{\sum_i p_1^i q_1^i}{\sum_i p_o^i q_1^i}$
$P_{q\ 1/0}$:indice P des quantités(q)	$P_{q\ 1/0} = \dfrac{\sum w_t}{\sum w_t \dfrac{q_o}{q_t}} = \dfrac{1}{\sum \dfrac{p_1 q_1}{\sum p_1 q_1} \dfrac{q_o}{q_1}} = \dfrac{\sum_i p_1^i q_1^i}{\sum_i p_1^i q_O^i}$
$P_{pq\ 1/0}$ indice	$P_{pq\ 1/0} = \dfrac{1}{\sum \dfrac{p_1 q_1}{\sum p_1 q_1} \dfrac{p_o q_o}{p_1 q_1}} = \dfrac{1}{\dfrac{\sum_i p_o q_o}{\sum_i p_1 q_1}} = L_{pq\ 1/0}$

Propriétés

Circularité

Aucun des trois indices ne possède la propriété de circularité.

a. Indice de Laspeyres

Le rapport des indices de Laspeyres aux dates 2 et 1 n’est pas l’indice de Laspeyres à la date 2 par rapport à la date 1.

$$\frac{L_{2/0}(G)}{L_{1/0}(G)} = \frac{\sum_i w_0^i \frac{G_2^i}{G_0^i}}{\sum_i w_0^i \frac{G_1^i}{G_0^i}} = \frac{\sum_i w_0^i \frac{G_1^i}{G_0^i} \frac{G_2^i}{G_1^i}}{\sum_i w_0^i \frac{G_1^i}{G_0^i}} = \frac{\sum_i w_0^i I_{1/0}(G^i)}{L_{1/0}(G)} I_{2/1}(G^i)$$

Tandis que

$$L_{2/1}(G) = \sum_i w_1^i I_{2/1}(G^i)$$

b. Indice de Paashe

$$\frac{P_{2/0}(G)}{P_{1/0}(G)} = \frac{\sum_i w_1^i \frac{G_0^i}{G_1^i}}{\sum_i w_2^i \frac{G_0^i}{G_2^i}}$$

Tandis que

$$P_{2/1}(G) = \sum_i I_{2/1}(G^i) = \frac{1}{\sum_i w_2^i \frac{G_1^i}{G_2^i}}$$

c. Indice de Fisher

$$F_{2/1}\left(G\right) = \frac{F_{2/0}\left(G\right)}{F_{1/0}\left(G\right)} \quad ?$$

$$F_{2/1}\left(G\right) = \sqrt{L_{2/1}(G)P_{2/1}(G))} = \frac{\sqrt{P_{2/0}(G)L_{2/0}(G))}}{\sqrt{P_{1/0}(G)L_{1/0}(G))}} \quad ?$$

$$L_{2/1}(G)\ P_{2/1}(G) \neq \frac{P_{2/0}(G)}{P_{1/0}(G)} \ \frac{L_{2/0}(G)}{L_{1/0}(G)}$$

car $L_{2/1}(G) \neq \dfrac{P_{2/0}(G)}{P_{1/0}(G)}$ $\quad (\textit{non circularité de } L)$

et $P_{2/1}(G) \neq \dfrac{L_{2/0}(G)}{L_{1/0}(G)}$ $\quad (\textit{non circularité de } P)$

Ainsi l'indice de Fisher ne possède pas à son tour la propriété de circularité.

Réversibilité

Les indices de Laspeyreset de Paashe ne possèdent pas la propriété de réversibilité contrairement à l'indice de Fisher.

a. Indice de Laspeyres

$$L_{0/1}(G) = \sum_i w_1^i \frac{G_0^i}{G_1^i} = \frac{1}{P_{1/0}(G)} \neq \frac{1}{L_{1/0}(G)}$$

b. Indice de Paashe

$$P_{0/1}(G) = \frac{1}{\sum_i w_0^i \dfrac{G_1^i}{G_0^i}} = \frac{1}{L_{1/0}(G)} \neq \frac{1}{P_{1/0}(G)}$$

c. Indice de Fisher

$$F_{0/1}(G) = \sqrt{L_{0/1}(G)P_{0/1}(G))} = \frac{1}{\sqrt{P_{1/0}(G)L_{1/0}(G))}} = \frac{1}{F_{1/0}(G)}$$

. EXERCICE

Soient 4 produits X_1, X_2, X_3 et X_4 dont les prix et les quantités sont donnés à deux périodes différentes t_1 et t.

Tableau 103 Prix de 4 produits durant périodes 1 et 2

	Période 1		Période t	
	P_{i1}	Q_{i1}	P_{it}	Q_{it}
X_1	12	6	15	7
X_2	5	13	8	11
X_3	15	9	13	18
X_4	8	10	10	9

1. Calculer les indices Pp, Pq ,Pvg , F_p , F_q?

Solution

	p_1	q_1	p_t	q_t	p_1q_1	p_tq_t	p_1q_t	p_tq_1
x_1	12	6	15	7	72	105	84	90
x_2	5	13	8	11	65	88	55	104
x_3	15	9	13	18	135	234	270	117
x_4	8	10	10	9	80	90	72	100

$$Pp=\frac{\sum p_tq_t}{\sum p_1q_t}=\frac{517}{481}.100= 107.48$$

$$P_q=\frac{\sum p_tq_t}{\sum p_tq_1}=\frac{517}{411}.100= 125.79$$

$$P_{vg}=\frac{\sum p_tq_t}{\sum p_1q_1}=\frac{517}{352}.100= 146.87$$

$$F_p=\sqrt{L_pP_p}=\sqrt{116.7\ \ *107.48}=111.9$$

$$F_q=\sqrt{L_qP_q}=\sqrt{136.6\quad *125.79}=131.08$$

BIBLIOGRAPHIE

Bailly P., & Carrère C., (2015). « Statistiques descriptives L'économie et les chiffres ». Presses Universitaires de Grenoble.

The Open University, (2015). "Descriptive Statistics ".

Chattamvelli R., & Shanmugam R., (2023)."Descriptive Statistics for Scientists and Engineers: application in R" . Springer

Bernstein S., (1998)."Elements of Statistics I Descriptive Statistics and Probability". McGraw-Hill

Goos P., *& Meintrup D. ,(2015)." Statistics with JMP_ Graphs, Descriptive Statisti*cs and Probability". Wiley (2015)

Julie Scott Jones, & John Goldring J.,(2022)."Exploratory and descriptive statistics". Sage Publications

William W.S. Wei, (2005). "Time Series Analysis _ Univariate and Multivariate Methods ". Addison Wesley.

Daniel J. Denis, (2016). "Applied Univariate, Bivariate, and Multivariate Statistics". John Wiley & Sons, Inc

Rebecca M. Warner , (2913)." Applied Statistics_ From Bivariate Through Multivariate Techniques".SAGE Publications

Husson F., Le S., & Pages J. (2010) "Exploratory Multivariate Analysis by Example Using R". Chapman & Hall _ CRC

Neil H. Timm, (2002)." Applied multivariate analysis (Springer Texts in Statistics)". Springer

Alvin C.Rencher, (2002)." Wiley series in probability and mathematical statistics - Methods of multivariate analysis" J. Wiley

Hardle et al., (2003). "Applied Multivariate Statistical Analysis", Springer

Härdle W., & Simar L., (2007) " Applied Multivariate Statistical Analysis". Springer

Daniel J. Denis ,(2019). "SPSS Data Analysis for Univariate, Bivariate and Multivariate Statistics". Wiley