UNIVERSITE DE LA MANOUBA

ECOLE SUPERIEURE DE COMMERCE DE TUNIS

Année universitaire 2012/2013

Résumé de cours de Statistiques descriptives.

Elaboré par *Foued Ben said*

Plan

Chapitre I : Analyse descriptive uni variée :

I. Vocabulaire statistique :

1. Statistique descriptive et statistiques inferentielles :

a-Statistique descriptive : c'est l'ensemble des outils qui permettent de résumer l'information contenue dans une base de données en utilisant des tableaux, des graphiques et des paramètres numériques. La statistique descriptive consiste à recueillir, synthétiser et résumer les données.

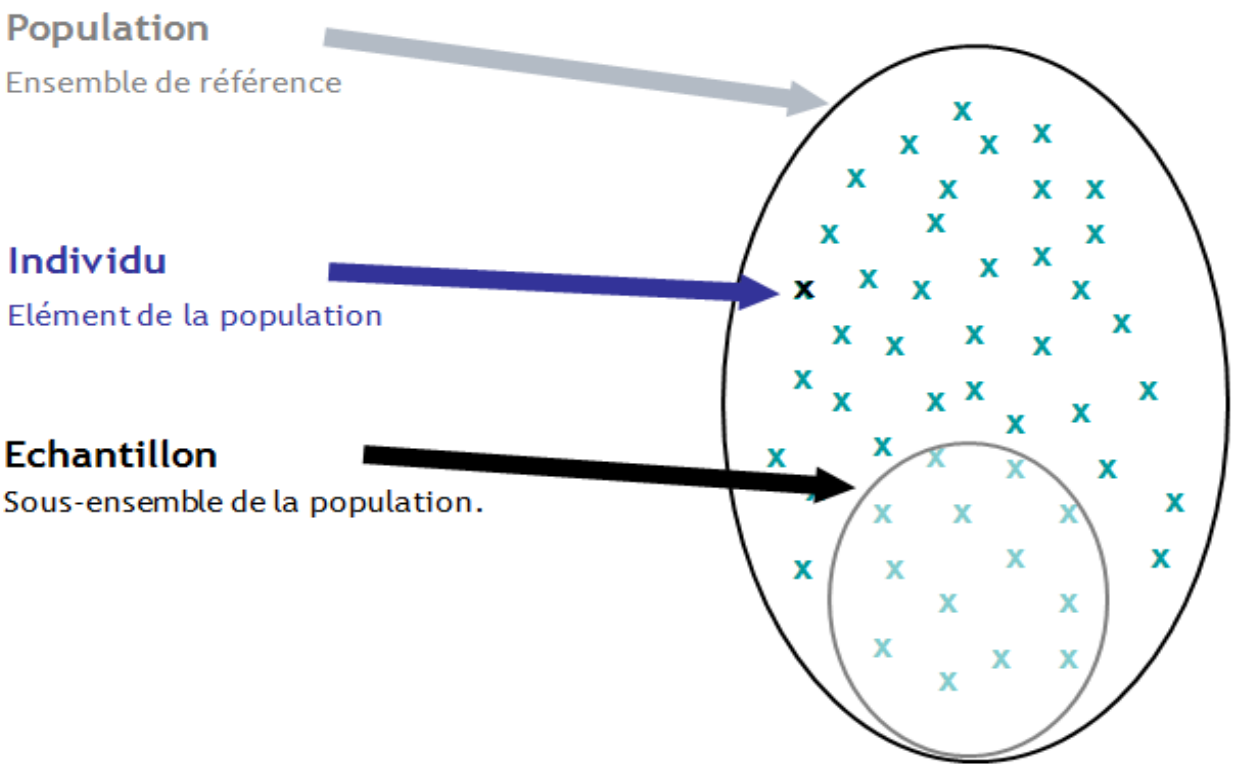
b-Statistique inférentielle : c'est l'ensemble des techniques qui permettent de généraliser des conclusions relatives à la population entière, à partir d'un sous-ensemble tiré selon une méthode scientifique de cette population. La statistique inférentielle se base sur la théorie des probabilités pour infirmer ou confirmer les hypothèses imposées sur les paramètres de la population.

2- Population, Echantillon et Individu :

a- individu : l'unité statistique faisant l'objet d'une observation (exemples : les banques, les pays …) ; cette unité statistique est l'entité abstraite qui représente un consommateur, un logement ou un produit.

b- population : l'ensemble des individus ou des unités statistiques qui font l'objet d'une étude (exemples : ensemble des habitants d'un pays, l'ensemble des navires d'une flotte navale) ; dans la plupart des études l'observation de tout les individus de la population pourrait être difficile et trop couteuse, dans ce cas on peut sélectionner un sous ensemble représentatif de cette population appelé : échantillon.

c- échantillon : un sous-ensemble tiré de la population mère dont les individus sont concernés par une étude. Le choix de l'échantillon se fait en respectant certaines règles qui permettent d'assurer la représentativité de l'échantillon par rapport à la population mère.



3- caractère statistique et différents types du caractère : le caractère est le phénomène étudié en statistique, il représente l'objet de l'observation statistique auprès des individus. L'âge des enquêtés constitue un caractère, le revenu du ménage et sa localisation géographique constituent des caractères statistiques.

a- les modalités : les modalités sont les différentes positions que peut prendre un caractère, ces modalités se caractérisent par leurs unités de mesure et leur ordre ou l'orientation. L'orientation signifie qu'on peut classer les modalités selon un ordre quelconque. On classe le caractère selon la signification de l'orientation et l'unité de mesure.

b-caractère qualitatif ou nominal : un caractère qualitatif ou variable qualitative est une variable qui possèdes des modalités sans unité de mesure ni orientation. Exemple : la région géographique, la nationalité…

c-caractère quantitatif discret ou ordinal : les modalités d'un caractère quantitatif discret son mesurables et peuvent êtres ordonnées. Les modalités sont finies et dénombrables et elles sont en général des entiers naturels. Exemple : le nombre des pièces d'un logement.

c-caractère quantitatif continu ou métrique : les modalités d'un caractère quantitatif continu son mesurables et peuvent êtres ordonnées. Les modalités sont infinies et leur représentation nécessite le recours à des intervalles ou classes. Exemple : le revenu du chef du ménage.

d-série de données statistiques : l'ensemble des modalités observées auprès des individus constitue une série de données statistiques. Cette série est l'objet de l'analyse descriptive, qui à pour objectif de la résumer par des tableaux des graphiques et des indicateurs.

II-Représentation des caractères statistiques :

1- Tableau statistique **:**

Le tableau statistique permet de résumer la série statistique en faisant un regroupement des individus associés aux modalités auxquelles ils appartiennent. La représentation générale d'un tableau statistique est la suivante :

a- Caractère qualitatif

Modalités	Effectifs (fréquences absolues)
m_1	n_1
m_2	n_2
m_3	n_3
m_4	n_4
total	n

Chaque tableau doit être illustré par un titre et une source.

Exemple :

Motivation	$\mathbf{n_i}$
Balnéaire	29 539 440
Ville	958 335
Circuit Sahara	1 059 135
Total	31 556 910

Répartition des nuitées selon la motivation (source ONT 2009)

b- Caractère discret :

la représentation d'un caractère quantitatif discret par un tableau :

X_i	Effectifs (fréquences absolues)
X_1	n_1
X_2	n_2
X_3	n_3
.	.
X_k	N_k
total	n

Titre et source

Exemple :

La répartition des logements selon le nombre des pièces :

modalités	Effectifs (fréquences absolues
0	4
1	5
2	9
3	3
4	7
5	2
total	30

c- Caractère quantitatif continu :

étant données que les modalités du caractère quantitatif sont infini donc on doit les regrouper dans des classes pour les représenter dans un tableau :

classes	Effectifs (fréquences absolues)
$[e_1 ; e_2[$	n_1
$[e_2 ; e3[$	n_2
$[e_i ; e_{i+1}[$	n_3
.	.
$[e_{k-1} ; e_k[$	N_k
total	n

Exemple :

modalités	effectifs
[0;4[	20
[4;6[	60
[6;8[	90
[8;10[	100
[10;12[	70
[12;14[	70
[14;16[	40
[16;20]	20
total	470

La répartition des étudiants selon les notes obtenues en statistique

Les classes peuvent être construites avec des amplitudes inégales, et le nombre total des classes peut être approximé par $\sqrt{n}$.

d- la notion de fréquence relative :

on calcule pour les effectifs absolus les fréquences relatives :

$f_i = \frac{n_i}{n}$ et $\sum f_i = 1$

Qui représentent les parts de l'effectif de chaque modalité n_i dans l'effectif total n

classes	effectifs	fréquence relative
[0;500[	366	0,674
[500 ; 1 000[	92	0,169
[1 000 ; 2 000[	43	0,079
[2 000 ; 5 000[	25	0,046
[5 000 ; 10 000[	8	0,015
[10 000 ; 50 000[	8	0,015
50 000 et +	1	0,002
total	543	1,00

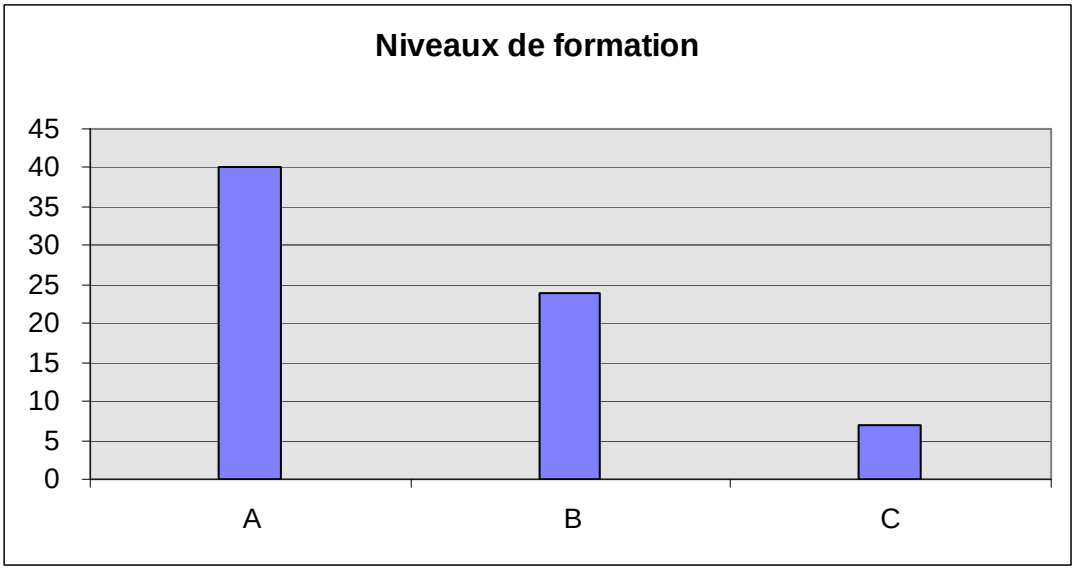
2- représentation graphique :

a- le diagramme en tuyaux d'orgue :

La représentation graphique d'un caractère qualitatif peut être réalisée par un diagramme en tuyaux d'orgue, le diagramme représente un ensemble de rectangles de largeurs égales et les hauteurs sont proportionnelles aux effectifs (fréquences).

modalité	*Effectif*	*Fréquence*
A	40	0,56
B	24	0,34
C	7	0,10
TOTAL	**71**	**1**

La répartition des étudiants selon les niveaux de formation



b- le diagramme en secteurs :

On peut représenter graphiquement un caractère qualitatif par un diagramme en secteurs ou diagramme en « camembert » :

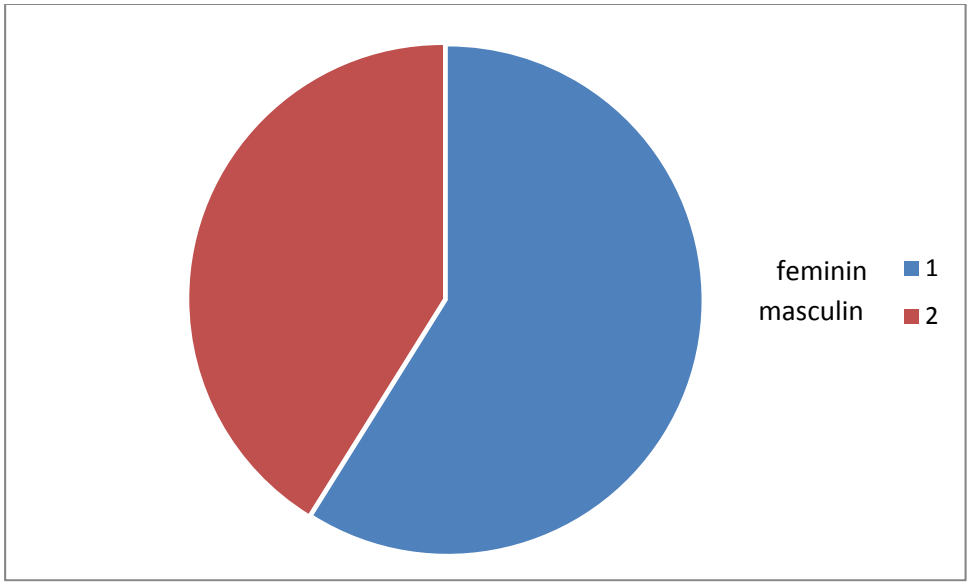
le principe de construction de ce diagramme est basé sur le fait que l'angle de chaque secteur est proportionnel à la fréquence relative des individus de chaque modalité ;

$\alpha_i = f_i \times 360$

Exemple :

La répartition des étudiants selon le sexeTapez une équation ici. :

modalités	effectif	fréquences
feminin	53	0,6625
masculin	37	0,4625
total	80	1



c- le diagramme en bâtons :

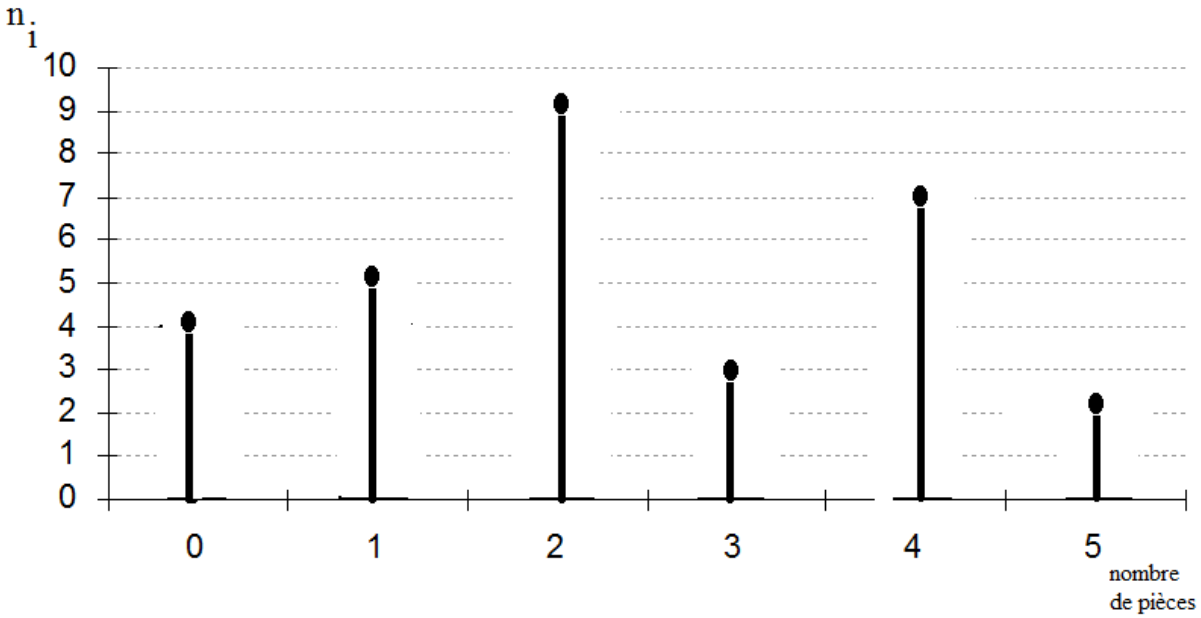
le diagramme en bâton est un diagramme qui permet de représenter graphiquement un caractère quantitatif discret.

Les modalités de la variable sont portées sur l'axe des abscisses et les fréquences absolues ou relatives sont portées sur l'axe des ordonnées.

Le principe de construction de ce diagramme est basé sur le fait q'à partir de chaque modalité on trace un segment de droite à extrémité « ronde», et la hauteur de chaque segment est proportionnelle aux fréquences.

Exemple : la répartition des logements selon le nombre des pièces.

modalités	Effectifs (fréquences absolues)
0	4
1	5
2	9
3	3
4	7
5	2
total	30

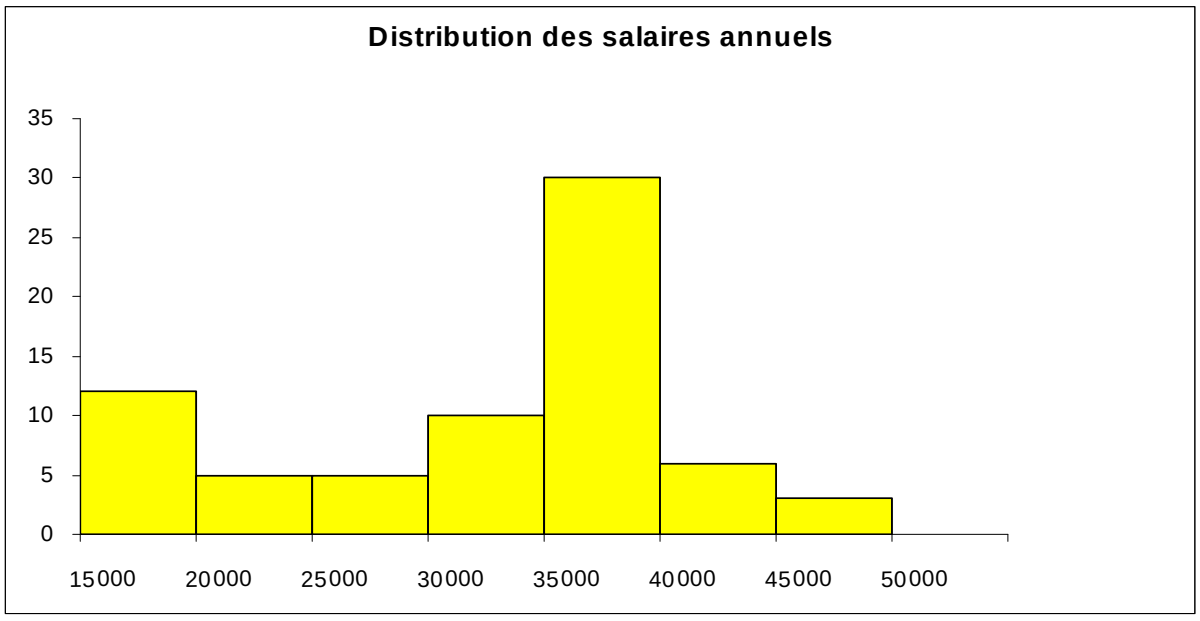


d- Histogramme et polygone des fréquences :

L'histogramme des fréquences est un graphique qui permet de représenter un caractère quantitatif continu, il est constitué de rectangles juxtaposés dont les surfaces sont proportionnelles aux fréquences.

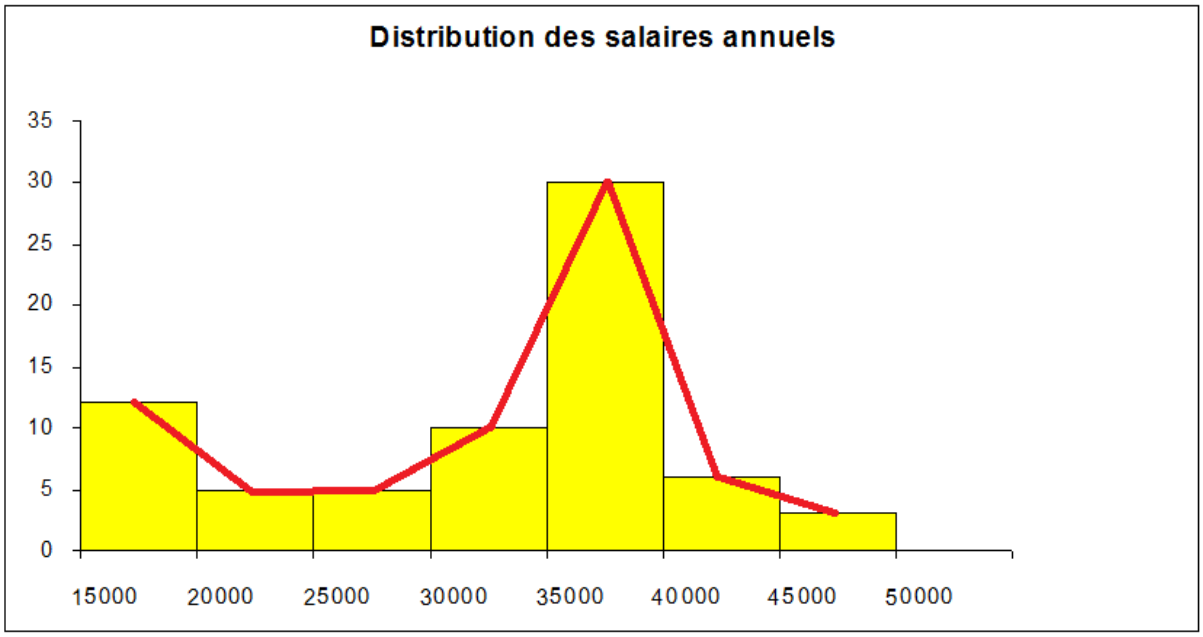
Exemple : la répartition des employés selon les salaires annuels :

modalités	Effectifs	fréquences
[15000,20000[	12	0,169
[20000, 25000[	5	0 ,070
[25000,30000[	5	0,070
[30000, 35000[	10	0,141
[35000,40000[	30	0,423
[40000,45000[	6	0,085
[45000,50000[	3	0,420
total	71	1



Représentation graphique de la répartition des employés selon les salaires

Le polygone des fréquences est une courbe qui relie les sommets des rectangles d'un histogramme d'un caractère quantitatif continu.



Lorsque les amplitudes des classes sont inégales la construction d'un histogramme des fréquences, basée sur le principe de proportionnalité entre surfaces et fréquences, nécessite le recours à des corrections pour respecter ce principe. On corrige les inégalités des amplitudes en se référant à une amplitude de référence qui permet de corriger ses fréquences.

classe	n_i	a_i	d_i	n_i^c
[100-150[	120	50	2,4	240
[150-250[	340	100	3,4	340
[250-300[	200	50	4	400
[300-400[	160	100	1,6	160
[400-500[	120	100	1,2	120
[500-700[	60	200	0,3	30
TOTAL	1000			

Répartition des employés selon les salaires mensuels.

a_i= l'amplitude de la classe

$d_i = \frac{n_i}{a_i}$ représente la densité des individus dans chaque classe.

3- Fonction de répartition d'un caractère quantitatif :

La fonction de répartition est une fonction qui permet de calculer la proportion des individus ayant une modalité inferieur à une modalité donnée.

Définition

: la fonction de répartition du caractère X est définie ainsi :

$F : \mathbb{R} \rightarrow [0,1]$

$F(x) \rightarrow p(X \leq x)$ la proportion des individus ayant des modalités $\leq$ à x

a- les fréquences cumulées

Pour calculer la fonction de répartition d'un caractère on doit calculer les fréquences cumulées $F_i = f_1 + f_2 + \ldots + f_i$

X (modalités)	Effectif (N_i)	Fréquence relative (f_i)	Fréquence relative cumulée (F_i)
x_1	n_1	f_1	$F_{1=} f_1$
x_2	n_2	f_2	$F_2 = f_1 + f_2$
…	…	…	…
x_k	n_k	f_k	$F_k = 1$

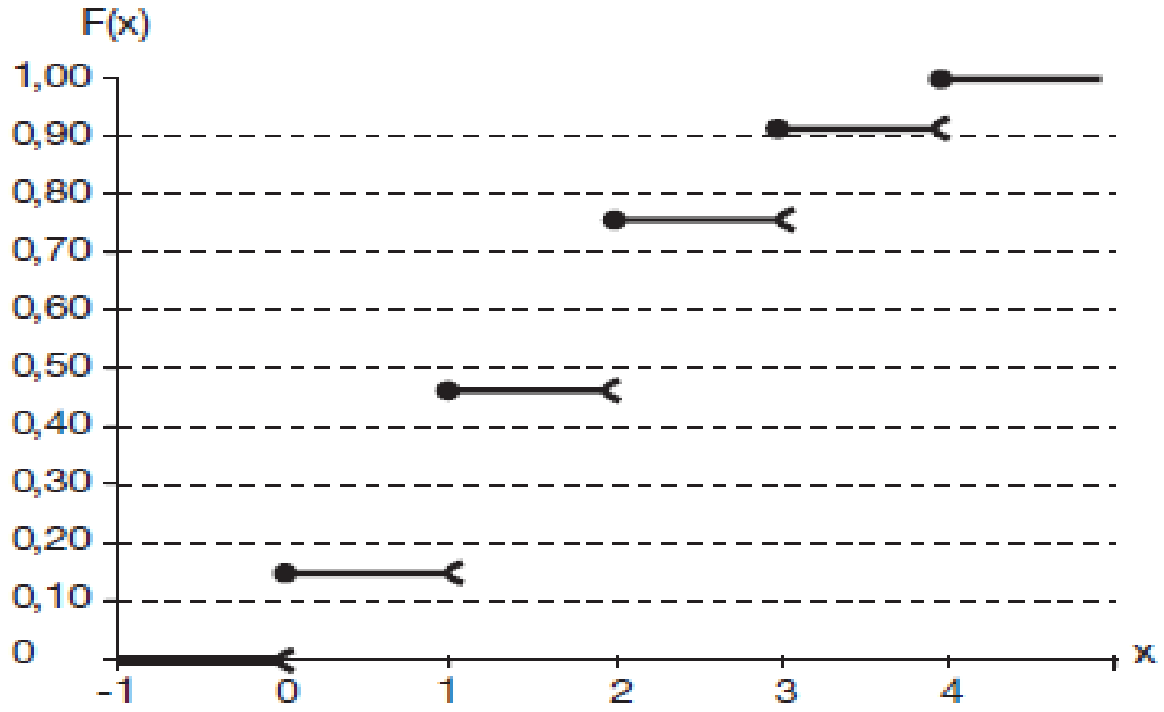
b- la représentation graphique de la fonction de répartition d'un caractère discret :

Considérons la répartition de 1500 ménages selon le nombre d'enfant, le tableau de répartition est présenté ainsi :

x_i	f_i	F_i
0	0,1726	0,1726
1	0,3047	0,4773
2	0,2849	0,7622
3	0,1480	0,9101
4	0,0899	1
TOTAL	1	

Répartition des ménages selon le nombre d'enfants

La représentation graphique de la fonction de répartition doit passer par une courbe en escalier :

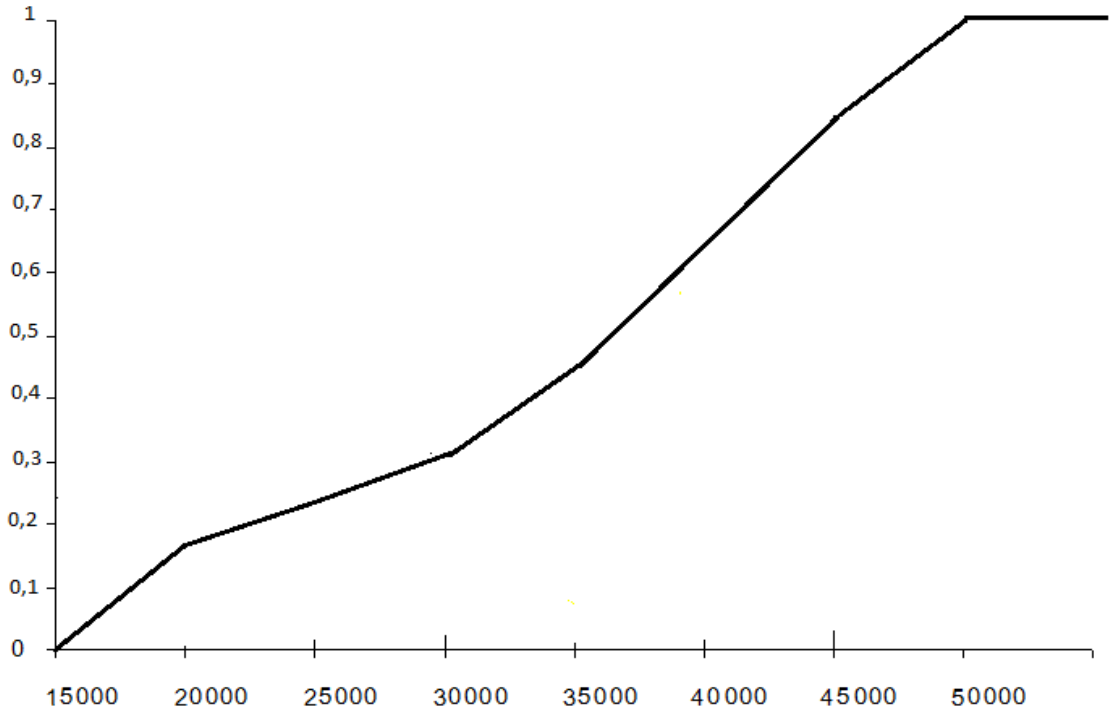


c- la représentation graphique de la fonction de répartition d'un caractère continu :

considérons la répartition des salariés selon les salaires :

modalités	Effectifs	fréquences	F_i cumulées
[15000,20000[	12	0,169	0,169
[20000, 25000[	5	0 ,070	0,239
[25000,30000[	5	0,070	0,310
[30000, 35000[	10	0,141	0,451
[35000,40000[	30	0,423	0,873
[40000,45000[	6	0,085	0,958
[45000,50000[	3	0,420	1
total	71	1	

La représentation graphique est réalisée selon le principe d'une interpolation linéaire des salaires dans chaque classe.



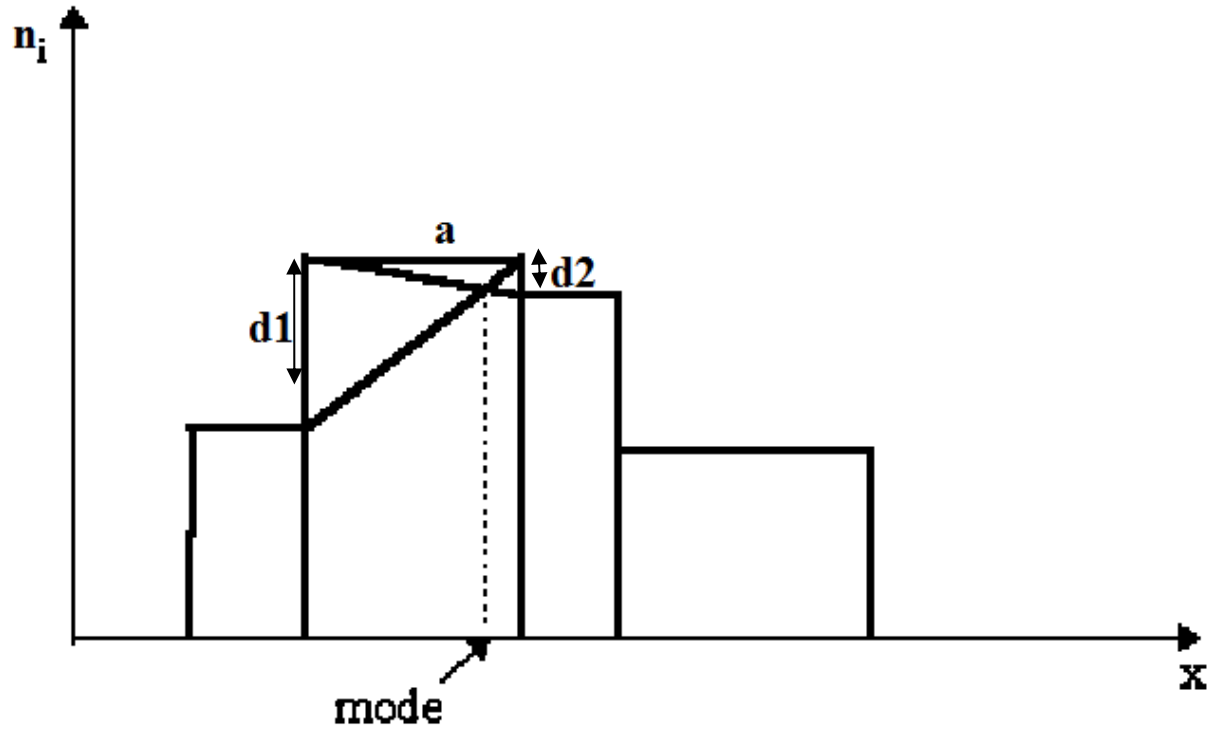
III-les indicateurs de position et de tendance centrale:

Ces indicateurs sont des paramètres calculés à partir de la série statistique dans le but de donner un résumé interprétable et exhaustif de l'information contenue dans cette série.

1- Le mode :

Le mode correspond à la modalité la plus fréquente. Pour un caractère continu pour le quel les données sont groupées en classes, la classe modale correspond à celle associée à l'effectif (corrigé) le plus élevé ou graphiquement au plus haut rectangle de l'histogramme.

Dans ce cas le mode est calculé à partir du centre de la classe modale selon la méthode suivante :

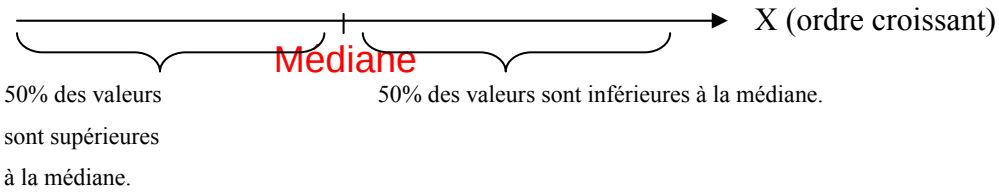


Si le mode appartient à la classe $[e_i ; e_{i+1}[$ alors :

$$M_O = e_i \times (\frac{d_1}{d_1 + d_2} \times a_i)$$

2- La médiane :

La médiane est la modalité qui divise la série des données statistiques en deux parties égales après avoir ranger ces données en ordre croissant (ou décroissant).



a- Cas d'un caractère discret :

Lorsqu'on possède la série des données brutes et distribution (non groupée), on doit ranger les n observations en ordre croissant.

Si n est **impair,** la médiane est la $\left(\frac{N+1}{2}\right)^{ième}$ observation.

Si n est **pair,** la médiane est habituellement définie comme étant le point milieu entre la $\left(\frac{N}{2}\right)^{ième}$ et la $\left(\frac{N}{2}+1\right)^{ième}$ observation.

b- cas d'un caractère continue :

la médiane est la modalité *x* tel que

$F(M_e) = P(X \leq M_e)=0,5$

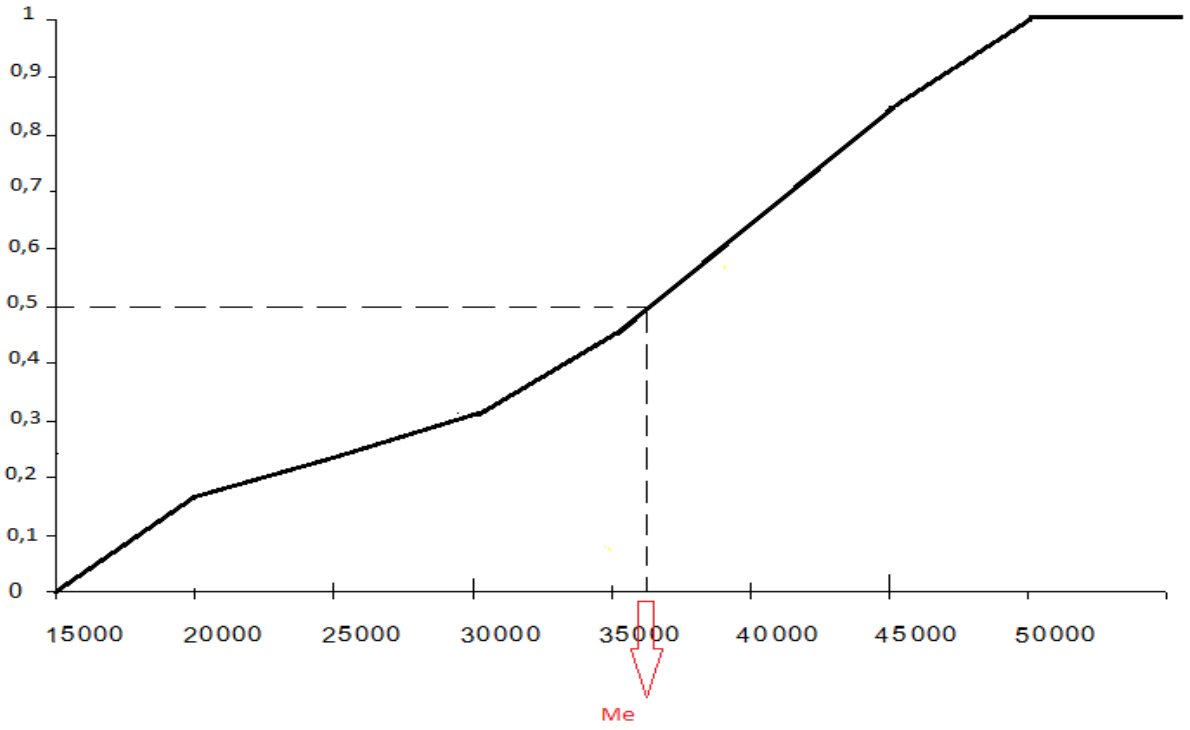
Pour calculer la médiane on doit déterminer la classe médiane à partir des fréquences cumulées croissant, puis on calcule la valeur ponctuelle de la médiane selon l'hypothèse de l'uniformité de la répartition des individus à l'intérieur de la classe médiane.

$M_e \in [e_i, e_{i+1}[$

$$M_e = e_i + (\frac{0{,}5 - F_{i-1}}{F_i - F_{i-1}} \times a_i)$$

La médiane se caractérise par le fait que sa valeur n'est pas influencée par les observations aberrantes ou les observations extrêmes.

Exemple



Me ϵ [35000,40000[

$$M_e = 35000 + (\frac{0{,}5 - 0{,}451}{0{,}873 - 0{,}451} \times 5000)$$

50% des salariés possèdent un salaire inferieur à la médiane.

3- les quantiles :

Les quantiles sont des indicateurs qui divisent la distribution en quatre parties égales. le premier quantile est indicateur noté Q_1 tel que

$F(Q_1) = P(X \leq Q_1)=0,25$

Si $Q_1 \in [e_i,e_{i+1}[$

$$Q_1 = e_i + (\frac{0,25 - F_{i-1}}{F_i - F_{i-1}} \times a_i)$$

Le troisième quantile est noté Q_3

$F(Q_3) = P(X \leq Q_3)=0,75$

Si $Q_3 \in [e_i,e_{i+1}[$

Alors

$$Q_3 = e_i + (\frac{0,75 - F_{i-1}}{F_i - F_{i-1}} \times a_i)$$

3- les moyennes :

La moyenne est un indicateur de tendance centrale qui permet de déterminer le centre de la distribution, la moyenne arithmétique est la moyenne est la plus utilisée, mais il existe d'autres types de moyennes utilisées dans le calcul de la tendance centrale de distributions statistiques telles que la moyenne géométrique et la moyenne quadratique.

a- La moyenne arithmétique :

La moyenne arithmétique est la somme de toutes les données observées divisées par le nombre des individus de l'échantillon.

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \text{Ou bien} \qquad \frac{x_1 + x_2 + \ldots + x_n}{n}$$

Si les données sont présentées dans un tableau statistique dans le quel chaque modalité est associée à fréquence absolue ou relative alors on calcule la moyenne arithmétique pondérée ainsi :

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{k} n_i x_i \qquad \text{ou} \qquad \bar{x} = \sum_{i=1}^{k} f_i x_i$$

Exemple : calcule du nombre de pièce moyen à partir de la distribution des logements selon le nombre des pièces :

Xi	ni	ni xi
0	4	0
1	5	5
2	9	18
3	3	9
4	7	28
5	2	10
total	30	70

$\bar{x} = \frac{\sum_{i=1}^{6} n_i x_i}{n} = \frac{70}{30} = 2{,}3$ le nombre de pièces moyen par logement est égale à 2

Dans le cas d'un tableau d'un caractère continu on remplace X_i par le centre de la classe $[e_i ; e_{i+1}[$ noté C_i

$$C_i = \frac{e_i + e_{i+1}}{2} \; et \; dans \; ce \; cas \; \bar{x} = \frac{\sum_{i=1}^{k} n_i C_i}{n}$$

Calcule de salaire moyen

classe	n_i	Ci	ni Ci
[100-150[	120	125	15000
[150-250[	340	200,00	68000
[250-300[	200	275,00	55000
[300-400[	160	300,00	48000
[400-500[	120	450,00	54000
[500-700[	60	600,00	36000
TOTAL	1000		276000

$$\bar{x} = \frac{\sum_{i=1}^{k} n_i C_i}{n} = 276$$

La moyenne arithmétique correspond au centre d'inertie ou centre de gravité de la distribution puisqu'elle vérifie toujours cette égalité :

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

La moyenne arithmétique est un paramètre qui peut être influencé par les observations extrêmes ou aberrantes.

b-la moyenne géométrique :

La moyenne géométrique d'une série statistique brute est définie ainsi :

$$\overline{x_g} = \sqrt[n]{\prod_{i=1}^{n} x_i}$$

Pour des données groupées la moyenne géométrique pondérée est calculée ainsi :

$$\overline{x}_g = \sqrt[n]{x_1^{n_1} \times x_2^{n_2} \times \ldots \times x_k^{n_k}}$$

$$\ln(\overline{x_g}) = \frac{1}{n}\sum_{i=1}^{n} n_i \ln(x_i)$$

Ce type de moyenne est surtout utilisé pour calculer des pourcentages moyens.

r étant un taux d'accroissement, $1+r_i$ est appelé coefficient multiplicateur e l'année i; et le coefficient multiplicateur moyen est alors égal à la moyenne géométrique des coefficients multiplicateurs annuels.

Exemple

année	capital	rendement	valeur
1	100	30%	130
2	130	10%	143

La moyenne arithmétique des rendements = (30+10)/2=20% cette moyenne ne donne pas une valeur finale de 143 après deux années. 143≠ 100 x $(1,2)^2$.

Le rendement moyen est une moyenne géométrique et il égale à 19,58%=$\sqrt{1,3 \times 1,1}$-1

c- La moyenne harmonique :

La moyenne harmonique est la moyenne de l'inverse de la variable x, ou bien l'inverse de la moyenne arithmétique, elle est calculée ainsi pour des données brutes:

$$\overline{x}_h^{-1} = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{x_i}$$

Pour des données groupées la moyenne harmonique est égale à :

$$\overline{x}_h^{-1} = \frac{1}{n}\sum_{i=1}^{k}\frac{n_i}{x_i}$$

La moyenne harmonique permet de calculer la moyenne des grandeurs obtenues à partir d'un rapport de deux variables tels que le taux de change, l'indice du prix le taux de chômage…

Exemple

Un routier conduit sa moto à 80km/h pendant 120km puis à 120km/h pendant encore 120km, Quelle a été sa vitesse moyenne durant son trajet?

La vitesse moyenne est égale à $\frac{2}{\frac{1}{80}+\frac{1}{120}} = 96\text{km/h}$

La moyenne arithmétique est égale à 100km/h

d- la moyenne quadratique :

La moyenne quadratique permet de calculer la moyenne des carrés des caractères, pour une série de données brute elle est calculée ainsi :

$$\overline{x}_q^{2} = \frac{1}{n}\sum_{i=1}^{n} x_i^2$$

Lorsque les données sont présentées dans un tableau statistique alors :

$$\overline{x}_q^{2} = \frac{1}{n}\sum_{i=1}^{k} n_i x_i^2$$

L'ensemble des moyennes calculées pour un caractère doivent vérifier l'inégalité suivante :

$$\min x_i \leq \overline{x_h} \leq \overline{x_g} \leq \overline{x} \leq \overline{x_q} \leq \max x_i$$

IV-les paramètres de dispersion, de forme et de concentration:

1- Les paramètres de dispersion :

Pour analyser une distribution on peut utiliser en plus des indicateurs de tendance centrale, telles que la médiane ou la moyenne, d'autres indicateurs qui permettent de mesurer la dispersion ou l'éparpillement de la série dans le but de bien décrire la distribution d'une variable.

Par exemple, les deux séries d'observations suivantes :

-20, -10, 0, 10, 20

-2000, -1000, 0, 1000, 2000

Possèdent la même moyenne et la même médiane (0) mais se diffèrent selon un autre indicateur qui mesure l'écart des ses observations par rapport à la valeur centrale. On va présenter dans cette partie les mesures de dispersion les plus utilisées : l'étendue, l'écart interquartile, la variance, l'écart-type et le coefficient de variation.

a- L'étendue:

L'étendue est un paramètre qui mesure l'écart entre la valeur la plus élevée et la valeur la plus faible de la distribution :

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

b- l'écart interquartile :

l'intervalle interquartile est l'intervalle $[Q_1 ; Q_3[$, cet intervalle contient 50% des observations.

L'écart interquartile est l'amplitude de l'intervalle interquartile :

$$EIQ = Q_3 - Q_1$$

L'écart interquartile est un indicateur qui a l'avantage d'écarter les observations extrêmes.

c- l'écart type :

L'écart type est l'indicateur de dispersion le plus utilisé et le plus simple à interpréter. Il permet de comparer les distributions dont la tendance centrale est identique. Il donne la variation moyenne de la distribution autour de la moyenne arithmétique. Pour calculer l'écart type on doit d'abord calculer la variance de X qui est égale à la somme des carrés des écarts à la moyenne divisée par l'effectif n, par la suite l'écart-type est égal à la racine de la variance.

La variance de X est calculée ainsi :

Pour des données brutes la variance est égale à :

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

Le développement de cette formule permet de donner une formule plus simple à manipuler dans le calcul pratique de la variance.

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

Lorsque les données sont groupées alors :

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

$$\sigma_x^2 = \frac{1}{n}\sum_{i=1}^{k} n_i x_i^2 - \bar{x}^2$$

Exemple de calcule de la variance pour un caractère discret :

Xi	n_i	$n_i x_i$	X_i^2	$n_i X_i^2$
0	4	0	0	0
1	5	5	1	5
2	9	18	4	36
3	3	9	9	27
4	7	28	16	112
5	2	10	25	50
total	30	70		230

La répartition des logements selon le nombre des pièces

$$\sigma_x^2 = \frac{1}{n}\sum_{i=1}^{k} n_i x_i^2 - \bar{x}^{-2} = \frac{230}{30} - (2{,}333)^2 = 2{,}223$$

$\sigma_x = 1{,}44$

calcule de la variance d'un caractère continu :

classe	n_i	Ci	ni Ci	C_i^2	$n_i C_i^2$
[100-150[	120	125	15000	15625	1875000
[150-250[	340	200	68000	40000	13600000
[250-300[	200	275	55000	75625	15125000
[300-400[	160	300	48000	90000	14400000
[400-500[	120	450	54000	202500	24300000
[500-700[	60	600	36000	360000	21600000
TOTAL	1000		276000		90900000

La répartition des salariés selon le salaire mensuel

$$\sigma_x^2 = \frac{1}{n}\sum_{i=1}^{k} n_i x_i^2 - \bar{x}^{-2} = \frac{90900000}{1000} - (276)^2 = 14724$$

d- Le coefficient de variation :

Lorsqu'on veut comparer la dispersion ou l'étalement de deux séries d'observations qui n'ont pas le même ordre de grandeur ou qui portent sur des variables différentes, on ne peut pas utiliser directement les écarts types. Le coefficient de variation se définit comme le rapport de l'écart type divisé par la moyenne, exprimé en pourcentage.

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

2- les paramètres de forme:

a- **Le coefficient d'asymétrie:**

*- Le coefficient de Pearson:

Le coefficient d'asymétrie de **Pearson** fait intervenir le mode M_o , il est définie par:

$$P = \frac{\bar{x} - M_o}{\sigma_X}$$

*- Le coefficient d'asymétrie de **Yule:**

Il fait intervenir la médiane et les quartiles, il est défini par:

$$Y = \frac{Q_1 + Q_3 - 2M_e}{2(Q_3 - Q_1)}$$

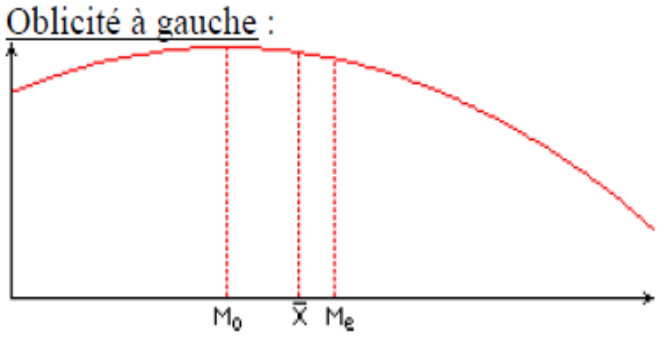
*- Le coefficient d'asymétrie de **Fisher :**

Il fait intervenir les moments centrés, il est défini par:

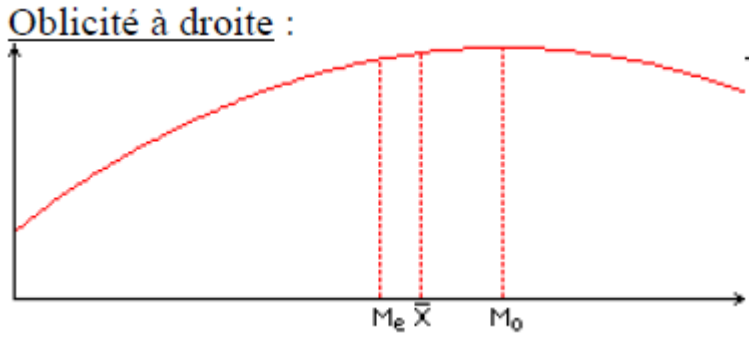
$$F = \frac{\frac{1}{n}\sum (x_i - \bar{x})^3}{\sigma_x^3}$$

*- Interprétation :

Lorsque le coefficient d'asymétrie est positif, la distribution est plus étalée à droite : on dit qu'il y a **oblicité à gauche.**



Lorsque le coefficient d'asymétrie est négatif, la distribution est plus étalée à gauche : on dit qu'il y a **oblicité à droite.**



b- le coefficient d'aplatissement:

*- le coefficient de Pearson

$$P = \frac{\frac{1}{n}\sum (x_i - \bar{x})^4}{\sigma_x^4}$$

*- Le coefficient d'aplatissement de Yule:

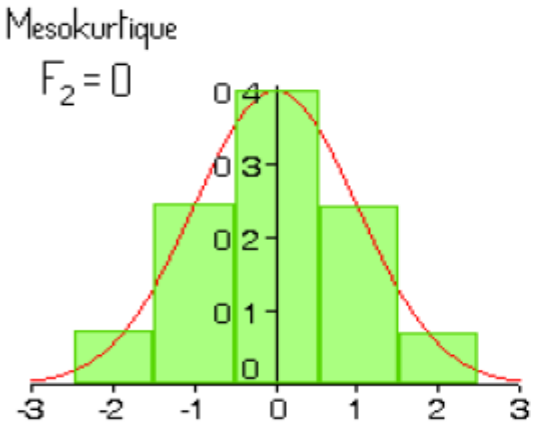
$$Y_2 = \frac{\frac{1}{n}\sum (x_i - \bar{x})^4}{\sigma_x^4} - 3$$

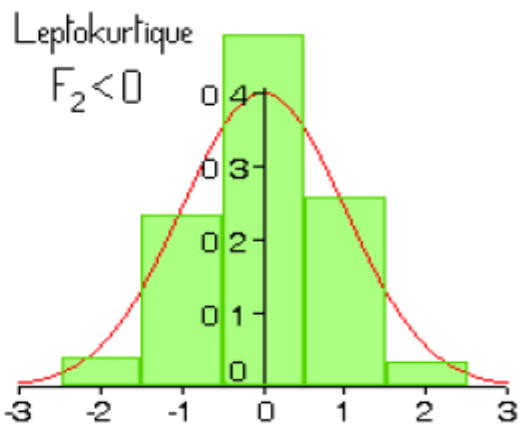
*- Interprétation:

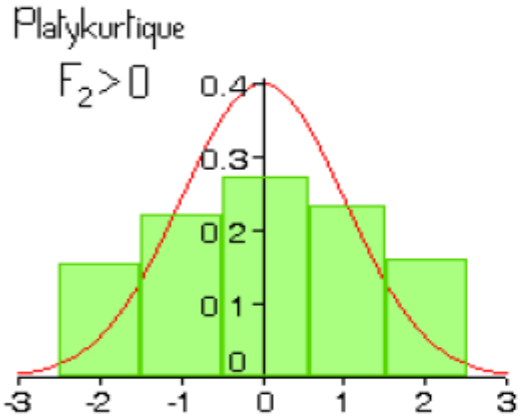
Si F 2 est égal à 0, le polygone statistique de la variable réduite a le même aplatissement qu'unecourbe en cloche, on dit que la variable est mésokurtique.

Si F 2 est > 0, le polygone statistique de la variable réduite est moins aplati qu'une courbe en cloche, on dit que la variable est leptokurtique.

Si F 2 est < 0, le polygone statistique de la variable réduite est plus aplati qu'une courbe en cloche, on dit que la variable est platykurtique.







4- concentration et indice de Gini:

Pour calculer la concentration on doit calculer la masse salariale versée à chaque catégorie de salaire groupée en classe, la part de chaque classe dans la masse salariale totale est égale à :

$$q_i = \frac{n_i x_i}{\sum_i^k n_i x_i}$$

$$Q_i = \frac{\sum_{i=1}^{i} n_i x_i}{\sum_i^k n_i x_i}$$

Qi représente la part salariale cumulée

classe	n_i	Ci	ni Ci	q_i	Q_i
[100-150[	120	125	15000	0,054	0,054
[150-250[	340	200	68000	0,246	0,301
[250-300[	200	275	55000	0,199	0,500
[300-400[	160	300	48000	0,174	0,674
[400-500[	120	450	54000	0,196	0,870
[500-700[	60	600	36000	0,130	1
TOTAL	1000		276000	1	

La Médiale, est le salaire tel que la moitié de la masse salariale a servi à payer ceux qui touchent un salaire inferieur à la Médiale.

Si la Médiale ϵ [e_{i-1} , e_i[

$$Mle = e_{i-1} + (\frac{0,5 - Q_{i-1}}{Q_i - Q_{i-1}} \times a_i)$$

Si la Médiane ϵ [e_{i-1} , e_i[

f_i	F_i
0,120	0,120
0,340	0,460
0,200	0,660
0,160	0,820
0,120	0,940
0,060	1,000
1,000	

$$M_e = e_{i-1} + (\frac{0,5 - F_{i-1}}{F_i - F_{i-1}} \times a_i)$$

$$M_e = 250 + (\frac{0,5 - 0,460}{0,2} \times 50) = 260$$

Le ratio de concentration = $\frac{\Delta M}{Etendue}$

ΔM=Mle - M_e

Si Δm = 0 alors la médiale est égale à la médiane c'est-à-dire 50% de la masse salariale est accaparée par 50% des salariés dans ce cas la répartition est parfaitement égalitaire.

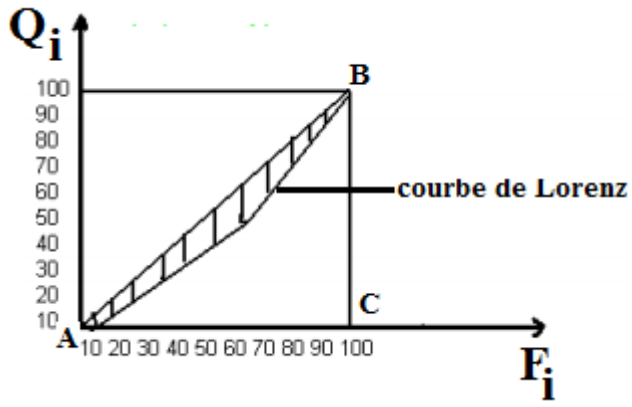
* Si Δm ≠ 0 cela indique qu'il y a une concentration

* Si Δm est faible par rapport à l'intervalle de variation la concentration est faible

* Si Δm est important par rapport à l'intervalle da variation alors concentration est forte.

a- Courbe de Lorenz :

La courbe de Lorenz est une courbe qui relie par des segments de droites les points d'un plan portant en abscisses les fréquences cumulées en%, et en ordonnées Qi :



La première bissectrice représente la courbe d'une répartition parfaitement égalitaires, plus la courbe de Lorenz est proche à cette bissectrice plus la concentration est faible et plus elle s'éloigne plus la concentration est forte.

b- Indice de Gini :

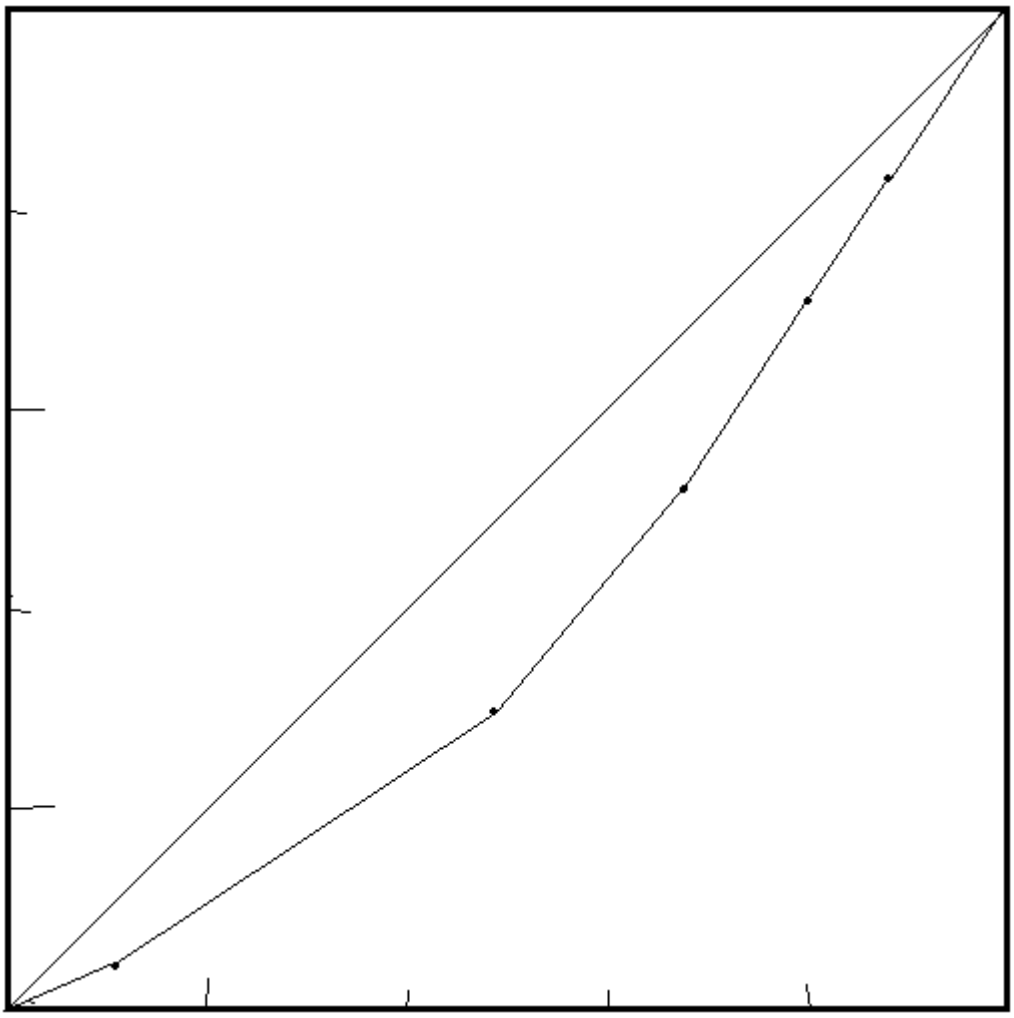
L'indice de Gini est un indice qui permet de mesurer la concentration d'une distribution statistique. Cet Indice représente le double de la surface comprise entre la courbe de Lorenz et la diagonale A B.

$$IG= 1- \Sigma\ fi*(Qi+Qi-1)$$

Exemple :

classe	n_i	Ci	ni Ci	q_i	Q_i	f_i	F_i	(Qi+Qi-1)	fi*(Qi+Qi-1)
[100-150[	120	125	15000	0,053	0,053	0,12	0,12	0,053	0,00633803
[150-250[	340	200	68000	0,239	0,292	0,34	0,46	0,345	0,11732394

[250-200[	200	275	55000	0,194	0,486	0,2	0,66	0,778	0,1556338
[300-400[	160	350	56000	0,197	0,683	0,16	0,82	1,169	0,18704225
[400-500[	120	450	54000	0,190	0,873	0,12	0,94	1,556	0,18676056
[500-700[	60	600	36000	0,127	1	0,06	1	1,873	0,11239437
TOTAL	1000		284000	1		1			0,76549296



IG=1-0,76549296 = 0,235

Donc la concentration est faible.

Chapitre II : les distributions statistiques à deux dimensions et ajustement linéaire:

I-distribution jointe, marginale et conditionnelle

1-le tableau de contingence:

X est une variable statistique pouvant prendre K modalités $x_1, \ldots, x_K$ et Y est une variable statistique pouvant prendre L modalités $y_1, \ldots, y_L$. On construit le tableau de contingence :

X \ Y	y1	. . .	yl	. . .	yL	Total
x1	n11	. . .	n1l	. . .	n1L	n1.
. . .	. . .		. . .		. . .	. . .
xk	nk1	. . .	nkl	. . .	nkL	nk.
. . .	. . .		. . .		. . .	. . .
xK	nK1	. . .	nKl	. . .	nKL	nK.
Total	n.1	. . .	n.l	. . .	n.L	n..

$$n_{k.} = \sum_{l=1}^{L} n_{kl}$$

$$n_{.l} = \sum_{k=1}^{K} n_{kl}$$

$$n_{..} = \sum_{k=1}^{K} \sum_{l=1}^{L} n_{kl}$$

nb de pièces	superficie				
	[10-30[	[30-50[	[50 -70[	[70-80[	TOTAL
1	3,00	1,00	0,00	0,00	4,00
2	1,000	14,00	3,00	0,00	18,00
3	0,00	1,00	7,00	4,00	12,00
4	0,00	0,00	10,00	7,00	17,00
total	4,00	16,00	20,00	11,00	51,00

2- distribution jointe:

$$f_{kl} = p(X = k; Y = l) = \frac{n_{kl}}{n..}$$

	superficie				
	10_30	**30-50**	**50 -70**	**70-80**	**TOTAL**
1	**0.06**	**0.02**	**0.00**	**0.00**	**0.08**
2	**0.020**	**0.27**	**0.06**	**0.00**	**0.35**
3	**0.00**	**0.02**	**0.14**	**0.08**	**0.24**
4	**0.00**	**0.00**	**0.20**	**0.14**	**0.33**
total	**0.08**	**0.31**	**0.39**	**0.22**	**1.00**

X \ Y	y1	. . .	yl	. . .	yL	Total
x1	f11	. . .	f1l	. . .	f1L	f1.
. . .	. . .		. . .		. . .	. . .
xk	fk1	. . .	fkl	. . .	fkL	fk.
. . .	. . .		. . .		. . .	. . .
xK	fK1	. . .	fKl	. . .	fKL	fK.
Total	f.1	. . .	f.l	. . .	f.L	f..

3- distribution marginale:

On appelle distribution marginale des fréquences (des effectifs) la distribution des fréquences (effectifs) obtenue dans la marge d'un tableau de contingence, en ajoutant les fréquences (effectifs) ligne par ligne, ou colonne par colonne.

$$f_{k.} = p(X = k) = \frac{n_{k.}}{n..}$$

$$f_{.l} = p(Y = l) = \frac{n_{.l}}{n..}$$

4- distribution conditionnelle:

$$f_{k/l} = p(X = k/Y = l) = \frac{n_{kl}}{n_l}$$

	superficie				
	10_30	30-50	50 -70	70-80	TOTAL
1	0,750	0,250	0,000	0,000	1,000
2	0,056	0,778	0,167	0,000	1,000
3	0,000	0,083	0,583	0,333	1,000
4	0,000	0,000	0,588	0,412	1,000

	superficie			
	10_30	30-50	50 -70	70-80
1	0,75	0,0625	0	0
2	0,25	0,875	0,15	0
3	0	0,0625	0,35	0,36363636
4	0	0	0,5	0,63636364
total	1	1	1	1

5-Variance marginale et conditionnelle

$$\bar{x} = \frac{1}{n_{..}} \sum_{i=1}^{k} n_{i.} X_i$$

$$V(x) = \frac{1}{n_{..}} \sum_{i=1}^{k} n_{i.} X_i^2 - \bar{x}^2$$

nb de pièces	superficie						
	[10-30[	[30-50[	[50 -70[	[70-80[	TOTAL	$\Sigma n_{i.} X_i$	$\Sigma n_{i.} X_i^2$
1	3	1	0	0	4	4	4
2	1	14	3	0	18	36	72
3	0	1	7	4	12	36	108
4	0	0	10	7	17	68	272
total	4	16	20	11	51	144	456

$$\bar{x} = \frac{1}{n_{..}} \sum_{i=1}^{k} n_{i.} X_i = 2{,}82$$

$$V(x) = \frac{1}{n_{..}} \sum_{i=1}^{k} n_{i.} X_i^2 - \bar{x}^2 = 0{,}988$$

$$\bar{y} = \frac{1}{n_{..}} \sum_{j=1}^{l} n_{.j} C_j$$

$$V(y) = \frac{1}{n_{..}} \sum_{j=1}^{l} n_{.j} C_j^2 - \bar{y}^2$$

nb de pièces	superficie				
	[10-30[	[30-50[	[50 -70[	[70-80[	TOTAL
1	3	1	0	0	4
2	1	14	3	0	18
3	0	1	7	4	12
4	0	0	10	7	17
total	4	16	20	11	51
C_j	20	40	60	75	
$\Sigma n_{.j} C_j$	80	640	1200	825	2745
$\Sigma n_{.j} C_j^2$	1600	25600	72000	61875	161075

variance conditionnelle:

$$\overline{x_{i/j}} = \frac{1}{n_{.j}} \sum_{i=1}^{k} n_{ij} X_i$$

$$V(x_{i/j}) = \frac{1}{n_{.j}} \sum_{i=1}^{k} n_{ij} X_i^2 - \overline{x}^2_{i/j}$$

$$\overline{y_{j/i}} = \frac{1}{n_{i.}} \sum_{j=1}^{l} n_{ij} C_J$$

$$V(Y_{j/i}) = \frac{1}{n_{i.}} \sum_{j=1}^{l} n_{ij} C_j^2 - \overline{y}^2_{j/i}$$

6- covariance et coefficient de corrélation :

a- Covariance entre deux variables

$$COV(X,Y) = \frac{1}{n_{..}} \sum_{i=1}^{k} \sum_{j=1}^{l} n_{ij} (X_i - \overline{X})(Y_j - \overline{Y})$$

$$COV(X,Y) = \frac{1}{n_{..}} \sum_{i=1}^{k} \sum_{j=1}^{l} n_{ij} X_i Y_j - \overline{X}\,\overline{Y}$$

nb de pièces							
	[10-30[	[30-50[	[50 -70[	[70-80[	TOTAL	Σ nij Yj	Σ nij YjXi
C_i	20	40	60	75			
1	3	1	0	0	4	100	100
2	1	14	3	0	18	760	1520
3	0	1	7	4	12	760	2280
4	0	0	10	7	17	1125	4500
total	4	16	20	11	51		8400

COV (X;Y)=(8400/51)-(2,82*53,82)=12,93

b- le Coefficient de Corrélation :

$$r_{xy} = \frac{COV(X;Y)}{\sigma_x \quad \sigma_y}$$

$$r_{xy} = \frac{\frac{1}{n_{..}} \sum_{i=1}^{k} \sum_{j=1}^{l} n_{ij} X_i Y_j - \overline{X}\,\overline{Y}}{\sqrt{\frac{1}{n_{..}} \sum_{i=1}^{k} n_{i.} X_i - \overline{X}^2} \quad \sqrt{\frac{1}{n_{..}} \sum_{j=1}^{l} n_{.j} Y_j - \overline{Y}^2}}$$

Interprétation

$$-1 \le r_{xy} \le 1$$

$$\left| r_{xy} \right| \le 1$$

$si \quad r_{xy} \to 1$ forte corrélation positive

$si \quad r_{xy} \to -1$ forte corrélation négative

$si \quad r_{xy} \to 0$ faible corrélation

$si \quad r_{xy} = 1$ parfaite corrélation positive

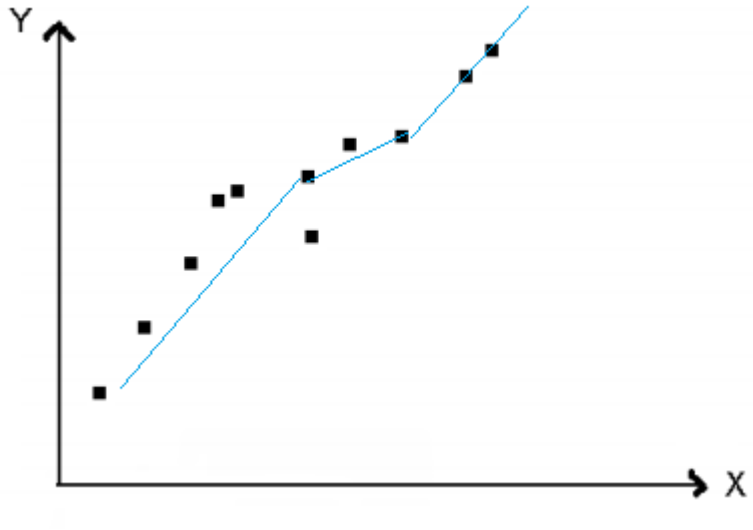
$si \quad r_{xy} = 0$ absence de corrélation

II – l'ajustement linéaire :

1- la courbe de régression :

la représentation graphique de la distribution d'un couple de variables quantitatives (X,Y) se fait dans un plan à deux axes. La variable X ou bien la variable « exogène » ou indépendante est portée sur l'axe des abscisses, et la variable Y ou variable « endogène » ou dépendante est portée sur l'axe des ordonnées. Le graphique obtenu représente un nuage de points dont le centre de gravité est alors le point de coordonnées $(\bar{x},\bar{y})$.

À partir du nuage de points, on peut visualiser le comportement moyen d'une des variables en fonction des valeurs de l'autre, pour cela, on trace les courbes de régression



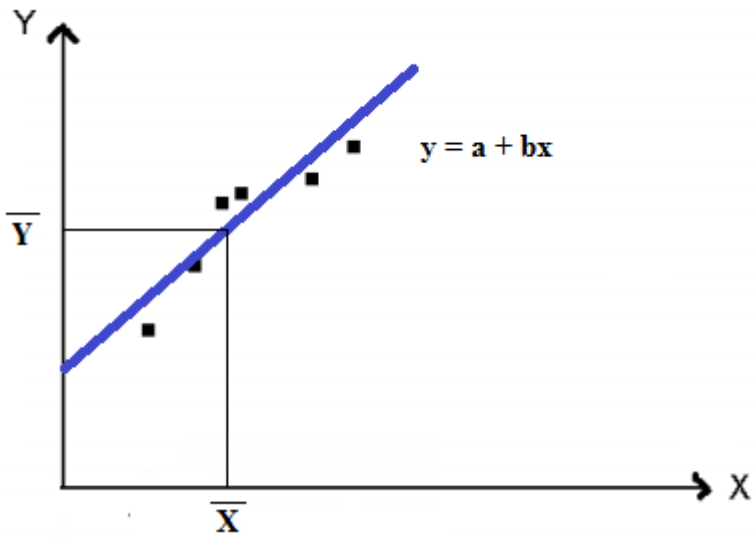
La courbe de régression de Y en X est la courbe joignant les points de coordonnées $\left(x_i, \bar{Y}/_{X=x_i}\right)$ .dans la pratique la courbe de régression ne permet de faires des extrapolations ou des prévisions mais elle peut servir à donner des informations sur la nature et forme de la relation entre les deux X et Y.

Pour mieux analyser la nature de la relation entre les deux caractères étudiés on peut ajuster le nuage de point obtenu par une droite linéaire appelé droite de régression ou droite des moindres carrés.

2- ajustement linéaire et droite des moindres carrés :

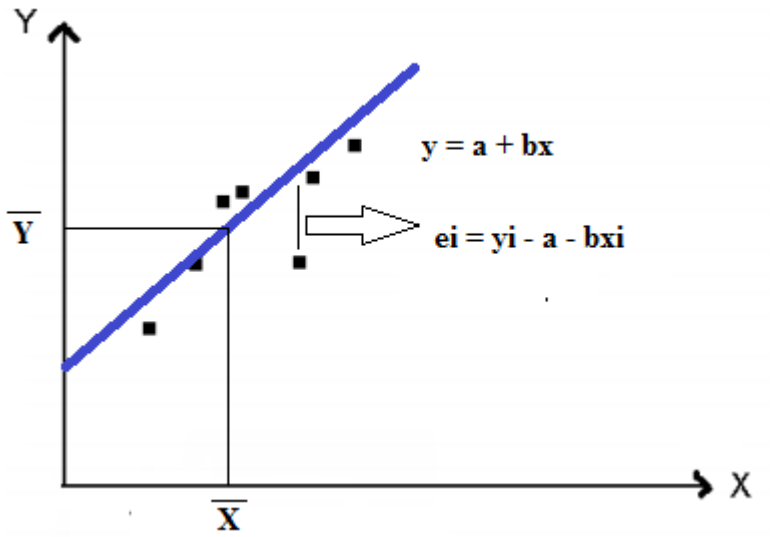
Le statisticien cherche à déterminer une relation déterministe entre X et Y dans le but de réaliser des extrapolations et des prévisions sur les valeurs moyennes de Y sachant les valeurs de X.

L'ajustement linéaire du nuage des points consiste à ajuster ce nuage par une droite linéaire qui permet de déterminer une approximation de la relation entre X et Y par une relation linéaire simple.



La meilleur technique qui permet d'ajuster ce nuage de point par une droite linéaire est la technique des moindres carrés. Cette technique consiste à déterminer les paramètres d'une droite qui minimise la somme des carrés des écarts entre les points sur la droite el les points observées sur le nuage.

Cette technique est appelée la technique des moindres carrés ordinaires qui permet de calculer les valeurs des coefficients de la droite d'ajustement à partir des valeurs observées des variables X et Y.



L'écart entre la valeur observée et la valeur théorique est égale à

$$e_i = y_i - a - bx_i$$

La technique des moindres carrés consiste à minimiser

$$\min_{\text{a et b}} \sum_{i=1}^{n} (e_i = y_i - a - bx_i)^2$$

Les condition de 1ère ordre sont :

$$\frac{\partial \sum e_i^2}{\partial a} = -2 \sum_{i=1}^{n} (y_i - a - bx_i) \quad = 0$$

$$\frac{\partial \sum e_i^2}{\partial b} = -2 \sum_{i=1}^{n} x_i (y_i - a - bx_i) \quad = 0$$

La résolution de ces deux équations donne la valeur estimée de a notée $\hat{a}$ et la valeur estimée de b notée $\hat{b}$

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

$$\hat{b} = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}$$

La valeur théorique de la variable y et calculer à partir de l'équation de la droite de régression :

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

Les résultats de la technique des moindres carrés doivent vérifier les propriétés suivantes :

$$\sum_{i=1}^{n} \hat{y}_i = \bar{y}$$

$\sum_{i=1}^{n} \hat{e}_i$=0

Exemple

Dans le but d'analyser le lien entre le prix d'une voiture d'un modèle quelconque et son âge, un statisticien dispose des données suivantes :

i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
X	2	12	13	7	6	3	12	10	9	7	4	2	10	6	3
Y	22	2	4	14	15	19	7	8	10	11	16	18	11	12	21

Le calcule des paramètres de la droite de régression nécessite le calcul de quelques grandeurs :

i	X_i	Y_i	$X_i Y_i$	X_i^2	$\hat{y_i}$	$\hat{e_i}$	Y_i^2
1	2	22	44	4	20,2740	1,7260	484
2	12	2	24	144	5,2640	-3,2640	4
3	13	4	52	169	3,7630	0,2370	16
4	7	14	98	49	12,7690	1,2310	196
5	6	15	90	36	14,2700	0,7300	225

6	3	19	57	9	18,7730	0,2270	361
7	12	7	84	144	5,2640	1,7360	49
8	10	8	80	100	8,2660	-0,2660	64
9	9	10	90	81	9,7670	0,2330	100
10	7	11	77	49	12,7690	-1,7690	121
11	4	16	64	16	17,2720	-1,2720	256
12	2	18	36	4	20,2740	-2,2740	324
13	10	11	110	100	8,2660	2,7340	121
14	6	12	72	36	14,2700	-2,2700	144
15	3	21	63	9	18,7730	2,227	441
	7,06666667	12,6666667	1041	950		-0,034	2906

$\bar{x}$= 7,066 $\bar{y}$ = 12,66 $\sum_{i=1}^{n} x_i y_i$=1041 $\sum_{i=1}^{n} x_i^2$=950

$\hat{b}$= (1041-15*(7,066*12,666))/(950-(15*(7,066)^2))= -1,4

$\hat{a}$ = 12,66+1,4(7,066)= 2,77

La droite de régression $\widehat{y_i} = 2{,}77 - 1{,}4\ x_i$

3- analyse de la variance et qualité d'ajustement :

Dans le but d'analyser la qualité d'ajustement et déterminer le pouvoir explicatif de la droite de régression à ajuster la relation entre les variables X et Y, on doit décomposer la variance totale ou bien la variance expliquées par la variable endogène Y en deux types de variabilité ou variances, une expliquées par la droite de régression ou bien variance explicative et un résiduelle expliquée par le terme d'erreur ou l'écart e_i.

La décomposition de la variance est présentée dans l'équation de l'analyse de la variance :

La somme des carrés totale= somme des carrés expliquée+ somme des carrés résiduelle.

SCT=SCE+SCR

$$SCT = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2$$

$$SCE = \sum_{i=1}^{n} \hat{y}_i^2 - n\bar{y}^2 = \hat{b}^2(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2)$$

$$SCE = \sum_{i=1}^{n} \hat{e}_i^2$$

L'équation de l'analyse de la variance permet de calculer un indicateur d'évaluation de la qualité de l'ajustement appelé le coefficient de détermination :

$$R^2 = \frac{SCE}{SCT} = 1 - \frac{SCR}{SCT}$$

$0 < R^2 < 1$

Si R^2 tend vers zéro alors l'ajustement est de mauvaise qualité

Si R^2 tend vers 1 alors l'ajustement est de bonne qualité.

Bibliographie :

Grais B . (1992) : *statistiques descriptives*coll economie Module, DUNOD.

Goldfarb **B**, Pardoux C **(1999) «** *Introduction à la méthode statistique* **» 2**e **edition, DUNOD**

Exercice 1

La distribution des salaires horaires de 250 employés d'une entreprise.

Salaire horaire (en dinars)	**Effectifs n_i**
[8 ; 8.4 [	**10**
[8.4 ; 8.8 [	**30**
[8.8 ; 9 [	**60**
[9 ; 9.2 [	**72**
[9.2 ; 9.6 [	**40**
[9.6 ; 10.2[	**24**
[10.2 ; 11 [	**14**
Total	**250**

1) Calculer le premier quartile $\mathbf{Q_1}$ et le troisième quartile $\mathbf{Q_3}$. Interpréter ces deux paramètres en utilisant l'intervalle interquartile.
2) Calculer le premier décile $\mathbf{d_1}$ et le neuvième décile $\mathbf{d_9}$. Interpréter chaque paramètre.
3) Déterminer le nombre d'employés ayant un salaire horaire compris entre le premier quartile $\mathbf{Q_1}$ et le neuvième décile $\mathbf{d_9}$.
4) Calculer la moyenne et la variance.

Exercice 2

La distribution statistique suivante porte sur un ensemble de parcelles de terrain classées d'après leur superficie exprimée en hectares.

Surface des parcelles en hectares	**Nombre de parcelles**
[0 ; 10 [	**16**
[10 ; 20 [	**30**
[20 ; 40 [	**18**
[40 ; 70 [	**10**
[70 ; 100 [	**6**
Total	**80**

1) Déterminer la médiale **Mle**. Interpréter.

2) Calculer l'indice de Gini. Que peut-on conclure ?

Exercice 3

La répartition de 600 ménages selon la taille des ménages et le nombre des pièces occupées est décrite par le tableau suivant.

Nombre de pièces y_j Nombre de personnes x_i	1	2	3	4	5 ou plus	
1	42	30	35	13	5	125
2	17	43	60	33	24	177
3	11	22	45	29	13	120
4	6	15	20	17	18	76
5 ou plus	4	10	10	48	30	102
Total	80	120	170	140	90	600

1) Donner la valeur et expliquer le sens de l'effectif $\mathbf{n_{33}}$ et l'effectif $\mathbf{n_{45}}$**.**
2) **Population ? Caractère ? Nature ?**
3) Donner toutes les distributions marginales de X et Y.
4) Donner la distribution conditionnelle de X sachant Y= 3.
5) Donner la distribution conditionnelle de Y sachant X= 2.
6) Calculer fréquence marginale…….

Exercice 4

Soit la distribution suivante de la longueur de 100 arbres plantés par la municipalité de Manouba :

Diamètre	n_i
[12, 15[	20
[15, 20[	25
[20, 23[	44
[23, 25[	11
Total	100

1)

1) Calculer la moyenne de la longueur de ces arbres.
2) Calculer le premier et le second quartile et interpréter.
3) Calculer le troisième décile.
4) Calculer la variance et l'écart-type.

Exercice 5

Lors d'une période de sécheresse, un agriculteur relève la quantité totale d'eau (en m^3) utilisée par son exploitation depuis le premier jour et donne le résultat suivant :

Nombre de jours écoulés : xi	1	3	5	8	10
Volume utilisé (en m^3) : yi	2.25	4.3	8	17.5	27

1) Donner l'équation de la droite de régression de y en x obtenue par la méthode des moindres carrés sous la forme y = ax + b, en présentant la méthode ainsi que tous les calculs nécessaires.
2) Calculez le coefficient de corrélation linéaire.

Exercice 6

Dans un gouvernorat du Nord-Ouest, on a voulu étudier la répartition des terrains agricoles entre les 200 habitants de cette région. On a obtenu les résultats suivants des superficies exprimées en hectares.

Superficies en hectares	n_i
[0-1[	80
[1-5[	60
[5-10[	20
[10-50[	20
[50-100[	16
[100-300[	4
Total	200

5) Calculer la moyenne et la variance.
6) Tracer la courbe cumulative.
7) Calculer la médiane, le premier quartile Q_1. Interpréter chaque paramètre.
8) Donner sans calcul le 9ème décile d_9. Interpréter.
9) Calculer la médiale. Interpréter.
10) Tracer la courbe de concentration et calculer l'indice de Gini. Que peut-on conclure ?

Exercice 7

Le service du personnel d'une grande entreprise nous a fourni la distribution statistique des cadres supérieurs suivant la rémunération mensuelle (**X**) et l'âge (**Y**).

Age / Salaire	Moins de 25 ans	[25 ;30[	[30 ;35[	[35 ;40[	[40 ;45[	[45 ;50[	[50 ;55[	55 ans et plus
Moins de 800	207	121	38	17	10	2	7	3
[800 ;900[	302	461	513	103	86	6	10	2
[900 ;1000[	18	526	682	567	613	431	105	60
[1000 ;1200[	0	111	342	298	416	480	226	37
[1200 ;1500[	0	1	3	182	227	263	98	18
[1500 ;2000[	0	0	0	18	22	13	12	5
2000 et plus	0	0	0	1	14	6	7	5

1) Quelle est la valeur de l'effectif n_{33}. Donner une interprétation.
2) Donner les distributions marginales de **X** (salaire) et **Y** (âge).
3) Quelle est la proportion des employés percevant un salaire compris entre 900 et 1000 dinars.
4) Donner la distribution conditionnelle de **Y** pour les employés qui ont un salaire entre 1000 et 1200 dinars.
5) Donner la distribution conditionnelle de **X** pour les salariés dont l'âge est supérieur à 40 ans.
6) Les deux variables **X** et **Y** sont-elles indépendants ?