collection
arabisation et connaissances
dirigée par : Mustafa BENYAKLEF

ELEMENTS DE STATISTIQUE DESCRIPTIVE

EL KACIMI ALAOUI Aziz
MAITRE DE CONFERENCE
A L'UNIVERSITÉ DE LILLE III

7

Prix : 20 DHS

UNIVERSITÉ DE LILLE III

PREMIÈRE ANNÉE – HISTOIRE
TÉLÉ-ENSEIGNEMENT

par

AZIZ EL KACIMI

ÉLÉMENTS DE STATISTIQUE DESCRIPTIVE

Publié comme livre en version bilingue (Français et Arabe) aux
Éditions Maghrébines collection ***Arabisation et connaissances***

ANNÉE UNIVERSITAIRE 1983-1984

AVANT-PROPOS

Beaucoup de disciplines scientifiques, expérimentales (chimie, physique, biologie, médecine *etc.*) ou humaines (sociologie, économie, histoire, géographie *etc.*) ont de plus en plus besoin des méthodes statistiques et ne sauraient s'en passer la plupart du temps. Pas mal d'hypothèses scientifiques trouvent leurs origines dans des études statistiques et ceci motive fortement la mise en œuvre de leur vérification.

Un cours de géographie ou d'histoire quantitative a besoin d'être épaulé par un cours de statistiques (étude de la démographie, de la répartition des richesses, de l'évolution des prix à une certaine époque *etc.*). Ce cours se veut dans ce sens ; il constitue, sans prétention aucune, une introduction aux méthodes élémentaires de la statistique descriptive. J'ai essayé de multiplier les exemples et les exercices et d'en donner une solution détaillée (pour certains d'entre eux). J'espère qu'il pourra rendre service aux étdudiants des filières susmentionnées.

Toute critique portant aussi bien sur le contenu que sur la forme sera la bienvenue !

Lille, Avril 1984

Aziz El Kacimi

CONTENU

CHAPITRE I
GENERALITES SUR
LES SERIES STATISTIQUES

I-DEFINITIONS Toute étude statistique suppose la donnée d'une population.

Par exemple :

- La population de la France,
- L'ensemble des élèves d'un lycée,
- Une colonnie de micro-organismes.

Les élements de cette population sont appelés individus.

Toute propriété des individus de cette population est appelée caractère.

Un caractère peut être :

- Qualitatif (couleur,sexe,...)
- Quantitatif (Age,Taille,...).

Une étude statistique est en fait l'étude d'un caractère de la population considerée.Souvent le nombre d'individus est trés grand voire même infini.Ceci complique l'étude.Pour ce faire on se restreint à une partie S de cette population qu'on appelle échantillon et qu'on suppose assez representatif.

Dans la suite on supposera que le caractère est quantitatif.On a alors la :

I.I-Definition : Une série statistique est une correspondance qui à chaque individu associe une valeur de son caractère.

Exemple: Soit P la population de la France.La correspondance qui à chaque habitant associe son revenu definit une série statistique.

Une série statistique est souvent representée par un tablaau :

Individus	W_I	W_2	.....	W_k
Valeurs du caractère	X_I	X_2	.....	X_k

I.2.Effectifs total et partiel

-Le nombre d'individus d'un échantillon est appelé effectif total.

-Le nombre d'apparitions d'une valeur du caractère est appelé effectif partiel de cette valeur.

Si on note $x_I,\ldots,x_k$ les valeurs du caractère d'une série statistique avec les effectifs partiels respectifs $n_I,\ldots,n_k$,on peut representer cette série par un tableau du type suivant:

Valeurs	x_I	x_2	.....	x_k
Effectifs partiels	n_I	n_2	.....	n_k

Evidemment toutes ces séries statistiques peuvent être representées graphiquement.Nous allons donner brièvement les representations graphiques les plus utilsées.Nous nous limiterons à le faire sur des exemples.

I.3.Representations graphiques.

I.3.I.Diagramme en bâtons. La série statistique suivante donne le nombre d'enfants par ménage dans un échantillon de 282 ménages :

0	enfants	 52	ménages
I	-	 80	-
2	-	 60	-
3	-	 24	-
4	-	..... I8	-
5	-	..... I6	-
6	-	.....I4	-
7	-	.....I2	-
8	-	 6	-

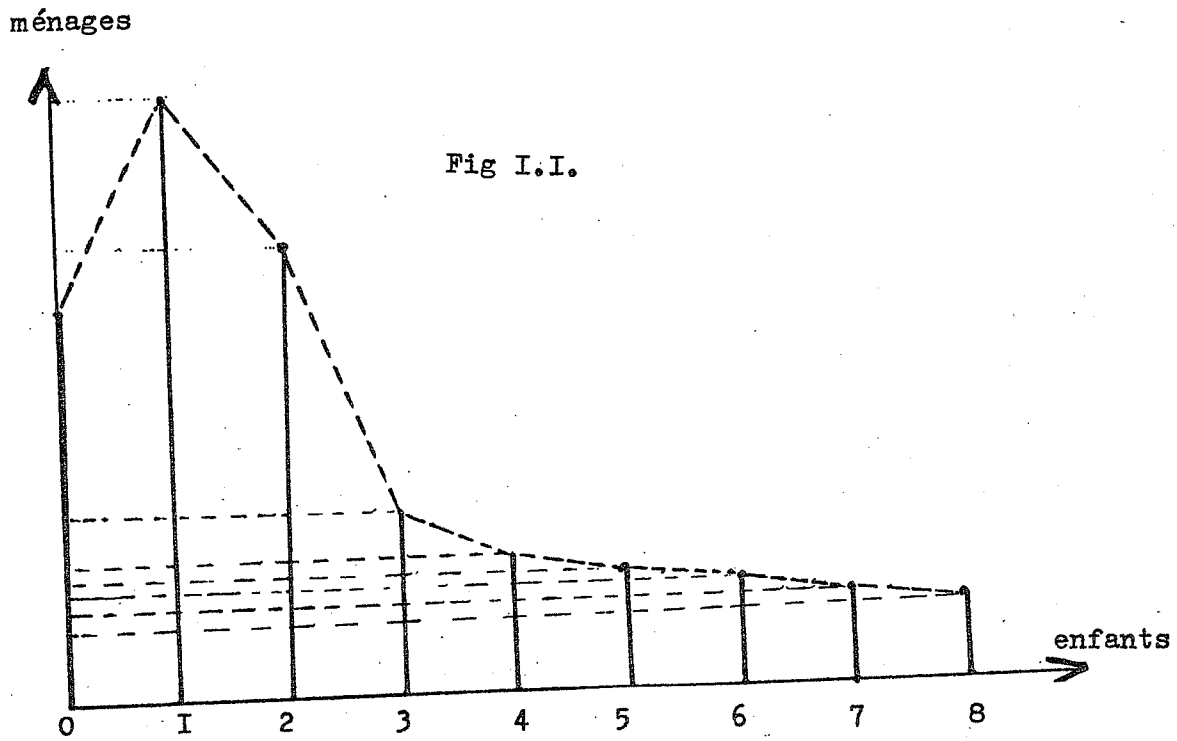


Fig I.I.

La courbe obtenue en joignant par des segments de droite les extrémités des batons est appelée le polygône des effectifs.

I.3.2.Diagpamme à secteurs:On construit un cercle et on partage le disque intérieur en secteurs dont les aires sont proportionnelles aux effectifs partiels.

Si on reprend la série statistique précédente on obtient le diagramme suivant :

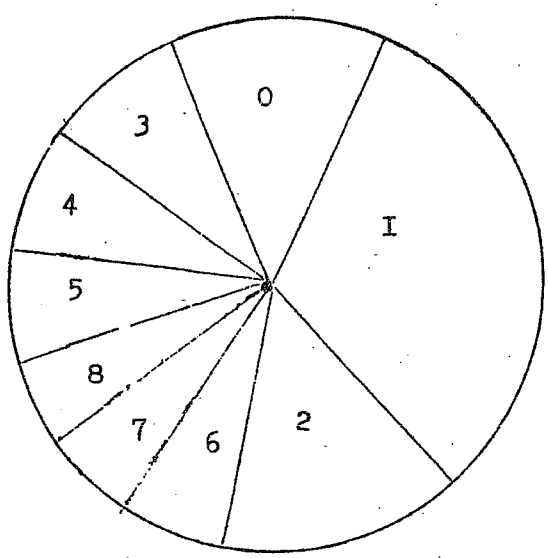


Fig I.2.

Les séries statistiques que l'on vient de considerer sont dites à valeurs isolées.Elles sont utilisées quand le nombre des valeurs du caractère n'est pas grand.Dans le cas contraire on introduit un autre type de séries statistiques :

2-SERIES CLASSEES. Commençons par un exemple.Supposons que l'on veuille étudier la distribution des salaires mensuels de I00 employés d'une grande entreprise.Le nombre étant relativement grand,on ne considerera pas chaque salaire isolément mais on constituera des intervalles de variation des salaires qu'on appelera des classes et qu'on poürra representer de la façon suivante :

Classes de salaires(en NF)	3500 à 4500	4500 à 5500	5500 à 8000	plus de 8000
Effectifs partiels	48	27	I7	8

2.I

D'une manière générale une série statistique classée peut se representer à l'aide d'un tableau du type suivant (qui est le même d'ailleurs que le tableau 2.I) :

Classes	$[x_I, x_2[$	$[x_2, x_3[$	.....	$[x_k, x_{k+I}[$
Effectifs partiels	n_I	n_2	.....	n_k

2.2

L'intervalle $[x_i, x_{i+I}[$ qu'on prendra généralement férmé à gauche et ouvert à droite est appelé classe et le nombre n_i correspondant (voir tableau 2.2) est l'effecfif de cette classe.

Les nombres :

$$e_i = x_{i+I} - x_i \quad \text{et} \quad c_i = \frac{x_i + x_{i+I}}{2}$$

sont appelés respectivement l'étendue et le centre de la classe $[x_i , x_{i+I}[$.

A une série classée est naturellement associée une série à valeurs isolées qui est la série des centres des differentes classes. Ainsi la série des centres associée à 2.I est donnée par le tableau suivant :

Centres des classes	4000	5000	6750	9250
Effectifs	48	27	I7	8

Remarque : La dernière classe de cet exemple n'a pas d'étendue. On convient généralement de prendre comme étendue celle de la classe qui prècède. Il arrive aussi que la première classe considérée n'ait pas d'étendue. Dans ce cas on prendra celle de la classe qui suit.

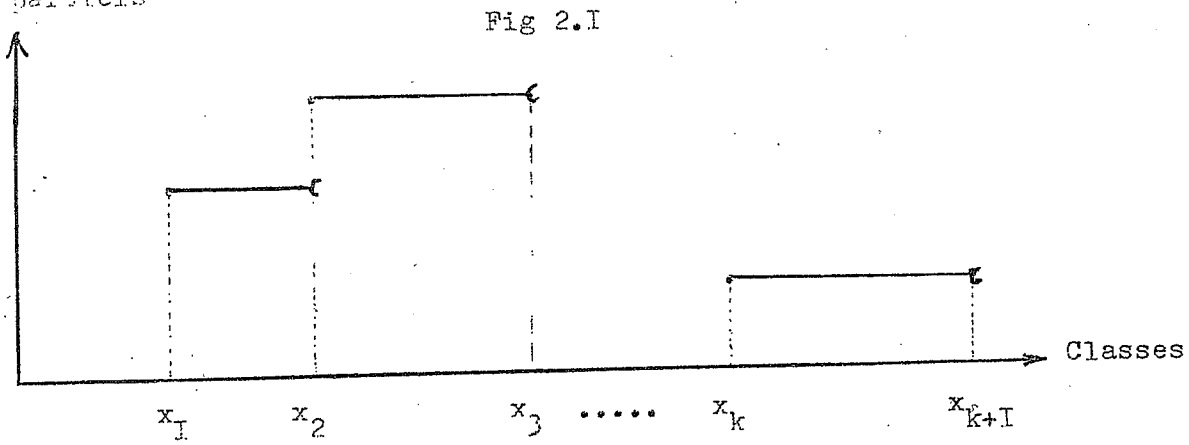
On verra que la série des centres d'une série classée est largement utilisée dans la pratique.

3-DIFFERENTES REPRESENTATIONS GRAPHIQUES D'UNE SERIE CLASSEE

Considèrons une série classée dont les classes sont $[x_i, x_{i+I}[$ et les effectifs partiels n_i où l'indice i varie de I à k.

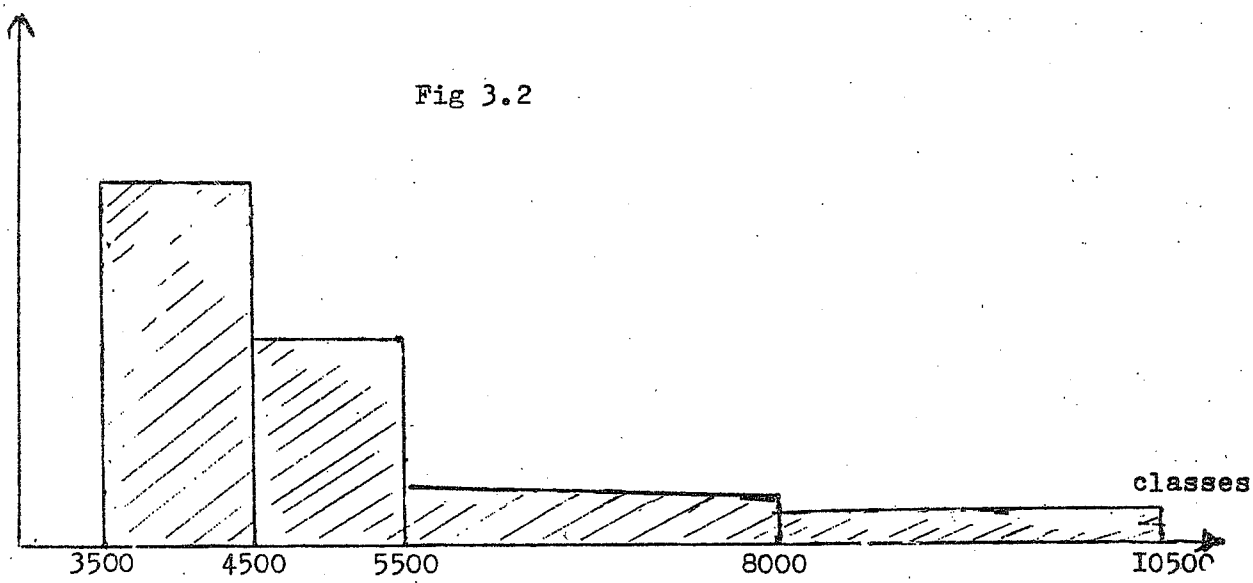
3.I-Diagramme en escalier. On trace un repère dont l'axe des abscisses repreaente les classes et l'axe des ordonnées les effectifs partiels.

Fig 2.I



3.2-Histogramme des effectifs. On considère toujours un repère et on trace des rectangles qui ont comme largeur l'étendue de la classe mesurée sur l'axe des abscisses et dont les aires sont proportionnelles aux effectifs partiels. Representons par exemple la serie classée 2.I par un histogramme des effectifs :

Fig 3.2



Important :Les hauteurs des differents rectangles ne sont pas proportionnelles aux effectifs partiels (sauf si les classes ont la même étendue)mais les aires doivent l'être.

Regardons par exemple comment on a construit les rectangles de l'histogramme ci-dessus.

Les 2 premières classes ont la même étendue,donc dire que les aires sont proportionnelles revient à dire que les hauteurs le sont.On peut donc choisir 48 mm pour la hauteur du I^{er} rectangle et 27 mm pour le second.Pour le 3ème on notera H sa hauteur.On a alors les relations de proportionnalité :

$$\frac{48 \cdot I000}{48} = \frac{H \cdot 2500}{I7}$$

D'où l'on tire :

$$H = 6,8 \text{ mm}$$

On fera la même chose pour constuire le dernier rectangle en comparant son aire à celle du 2ème par exemple.

2.3-Pyramide des âges. C'est une representation qui est souvent utilisée pour ullistrer la répartition des âges d'une population suivant le sexe. Elle se presente sous la forme suivante :

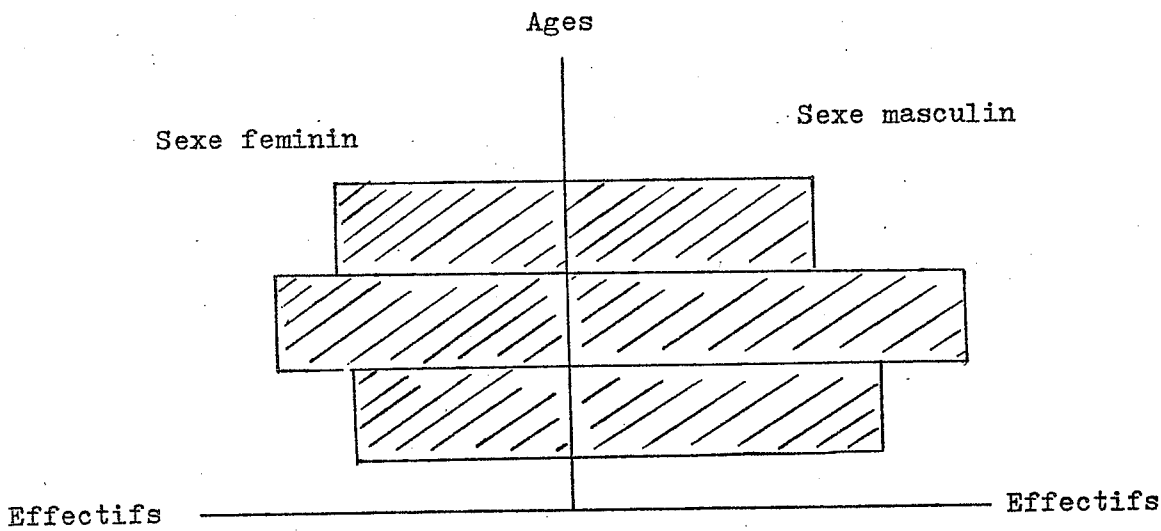


Fig 3.3

4-FREQUENCE.

4.I-Série à valeurs isolées .Soit la série à valeurs isolées:

$$x_I \quad x_2 \quad \dots \quad x_k$$

$$n_I \quad n_2 \quad \dots \quad n_k$$

où n_I , n_2 , ... , n_k sont les effectifs partiels respectivement des valeus x_I , x_2 , ... , x_k .

On appelle fréquence de la valeur x_i ,le nombre f_i defini par l'égalité:

$$f_i = \frac{n_i}{N} \quad \text{où } N = n_I + n_2 + \dots + n_k \text{ est l'effectif total.}$$

4.2-Série classée. Soit $[x_I , x_2[\quad \dots \quad [x_k , x_{k+I}[$

$$n_I \quad \dots \quad n_k$$

une série classée.On definit de la même manière la fréquence de la classe

$[x_i , x_{i+1}[$ qu'on note aussi f_i par l'égalité :

$$f_i = \frac{n_i}{N} .$$

On verifie aisement aussi bien pour une série à valeurs isolées que pour une série classée que la somme des fréquences est toujours égale à I. En effet on a :

$$f_I + f_2 + \dots + f_k = \frac{n_I}{N} + \frac{n_2}{N} + \dots + \frac{n_k}{N} = \frac{N}{N} = I .$$

On peut aussi obtenir une distribution des fréquences de la même manière qu'on a une distribution des effectifs. On la represente par un tableau :

Classes	$[x_I , x_2[$	...	$[x_k , x_{k+I}[$
Fréquences	f_I	...	f_k

auquel on peut associer un diagramme en escalier et un histogramme des fréquences.

Reprenons le tableau 2.I mais en considèrant cette fois-ci la distribution des fréquences. On obtient alors :

Classes de salaires (en NF)	3500 à 4500	4500 à 5500	5500 à 8000	plus de 8000
Fréquences	0,48	0,27	0,I7	0,08

5-EXERCICES.

5.I-Exercice resolu.Le tableau suivant donne le produit national brut par habitant de 50 pays africains (tiré de "Dossiers et Documents Le Monde" N° I08 du mois de Fev I984) :

PNB par habitant(en dollars)	Nombre de pays (Effectifs)
Moins de I000	39
I000 à 2000	7
2000 à 3000	2
3000 à 4000	2

Tab 5.I

Les PNB de 3 pays ne sont pas comptés : Celui de la Libye qui est de 8640 dollars et ceux de la Guinée équatoriale et de la Somalie qui ne sont pas connus.

a)Donner la distribution des fréquences.

b)Tracer l'histogramme des fréquences associé.

Solution.

a)Par definition la fréquence d'une classe est égale au rapport de l'effectif partiel de cette classe à l'effectif total.On calcule successivement ces rapports et on obtient le tableau :

Tab 5.2

PNB par habitant(en dollars)	Fréquences
Moins de I000	$\frac{39}{50} = 0,78$
I000 à 2000	$\frac{7}{50} = 0,I4$
2000 à 3000	$\frac{2}{50} = 0,04$
3000 à 4000	$\frac{2}{50} = 0,04$

b)Toutes les classes de cette série ont la même étendue et par conséquent on peut choisir les hauteurs des differents rectangles de l'histogramme des fréquences proportionnelles aux fréquences.On choisira I mm comme unité representant une fréquence de 0,0I.Ceci nous donne le diagramme :

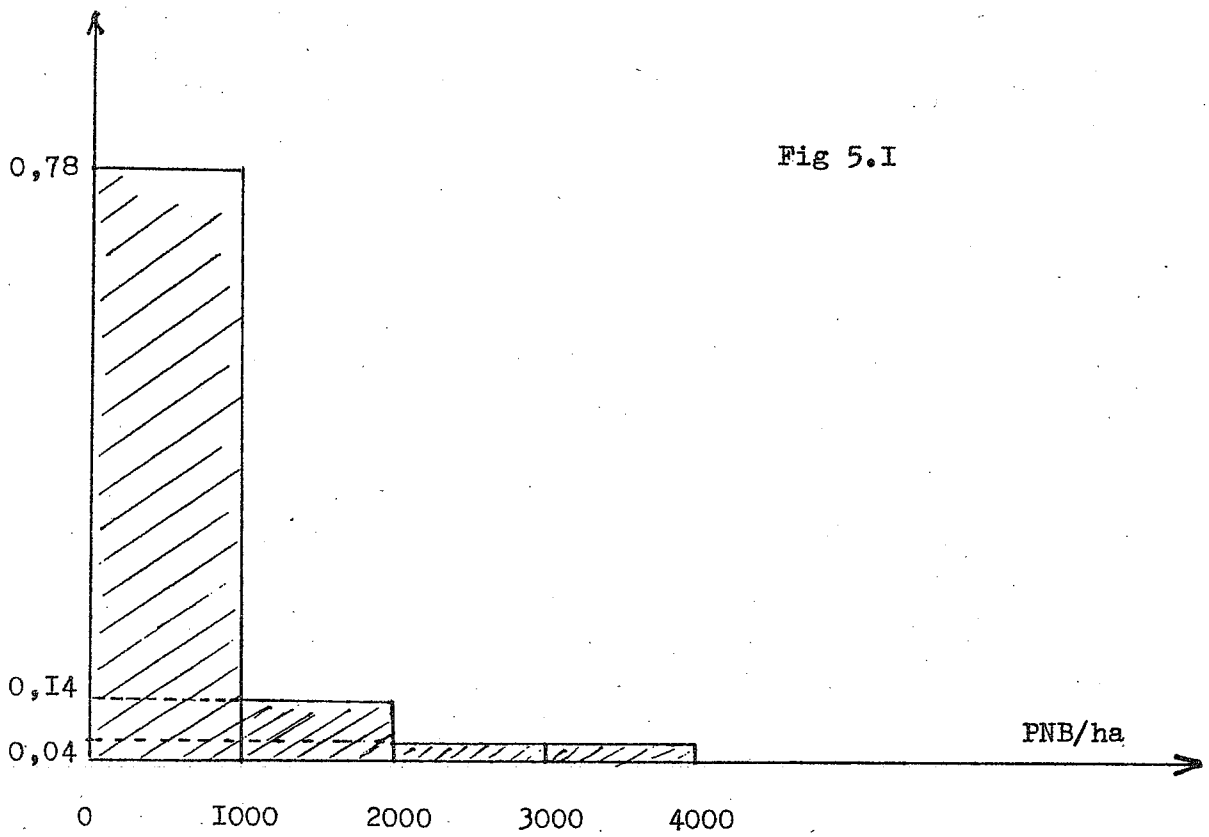


Fig 5.I

5.2-Exercices supplémentaires.

5.2.I-Soit la série des nombres :

I0,I0,II,I3,I3,I3,I7;20,25,26,33,33,33,33,46,53,56,59;

a)Mettre cette série sous forme d'une série classée dont les classes ont une étendue égale à I0.

b)Donner la distribution des effectifs et tracer l'histogramme associé.

c)Donner la distribution des fréquences et tracer l'histogramme associé.

5.5.2-Le tableau suivant donne la repartion de la population immigrée en France (aux alentours de l'année I974)tiré de "Les immigrés" CEDETIM

Age	Hommes	%	Femmes	%
moins de I9	366840	22,66	344330	32,92
De 20 à 64	II05760	68,32	555I20	53,08
65 et plus	I45740	9,02	I46220	I3,98

Tab 5.3

Tracer la pyramide des âges associée à cette statistique et interpreter les resultats.

6-EFFECTIFS ET FREQUENCES CUMULES.

Soit $\begin{matrix} x_I , \dots , x_k \\ n_I , \dots , n_k \end{matrix}$ une série à valeurs isolées telle que $x_I < x_2 < \dots < x_k$

On appelle effectif cumulé de la valeur x_i la somme des effectifs partiels de x_i et de toutes les valeurs de la série qui sont inferieures à x_i ,c'est à dire :

$$N_i = n_I + n_2 + \dots + n_i \quad (N_i \text{ denote cet effectif cumulé}) .$$

On appelle fréquence cumulée de la valeur x_i le nombre :

$$F_i = f_I + f_2 + \dots + f_i .$$

Pour une série classée on definit de la même manière l'effectif cumulé d'une classe (respectivement la fréquence cumulée) comme étant la somme des effectifs partiels (respectivement les fréquences) de cette classe et de toutes celles qui la précèdent.

Par exemple si on reprend le tableau 2.I on obtient les distributions des effectifs cumulés et fréquences cumulées.

Tab 6.I

Classes de salaires(en NF)	Effectifs cumulés
3500 à 4500	48
4500 à 5500	48 + 27 = 75
5500 à 8000	48 + 27 + I7 = 92
plus de 8000	48 +27 + I7 + 8 = I00

Classes de salaires (en NF)	Fréquences cumulées
3500 à 4500	0,48
4500 à 5500	0,75
5500 à 8000	0,92
plus de 8000	I

Tab 6.2

On peut remarquer que l'effectif cumulé de la Ière valeur ou de la Ière classe (la fréquence cumulée) est égal à l'effectif partiel de cette valeur ou de cette classe et que l'effectif cumulé (la fréquence cumulée) de la dernière valeur ou la dernière classe est égal à l'effectif total (à I).

Il existe aussi un histogramme des effectifs cumulés et un histogramme des frequences cumulées.Ils consistent en des rectangles dont les aires sont proportionnelles aux effectifs cumulés ou aux fréquences cumulées suivant le cas.

Representons par exemple celui de l'exercice 2.I.Commençons d'abord par donner la distribution des effectifs cumulés :

PNB par habitant(en dollars)	Effectifs cumulés
Moins de I000	39
I000 à 2000	46
2000 à 3000	48
3000 à 4000	50

Tab 6.3

Ceci nous permet donc de construire l'histogramme :

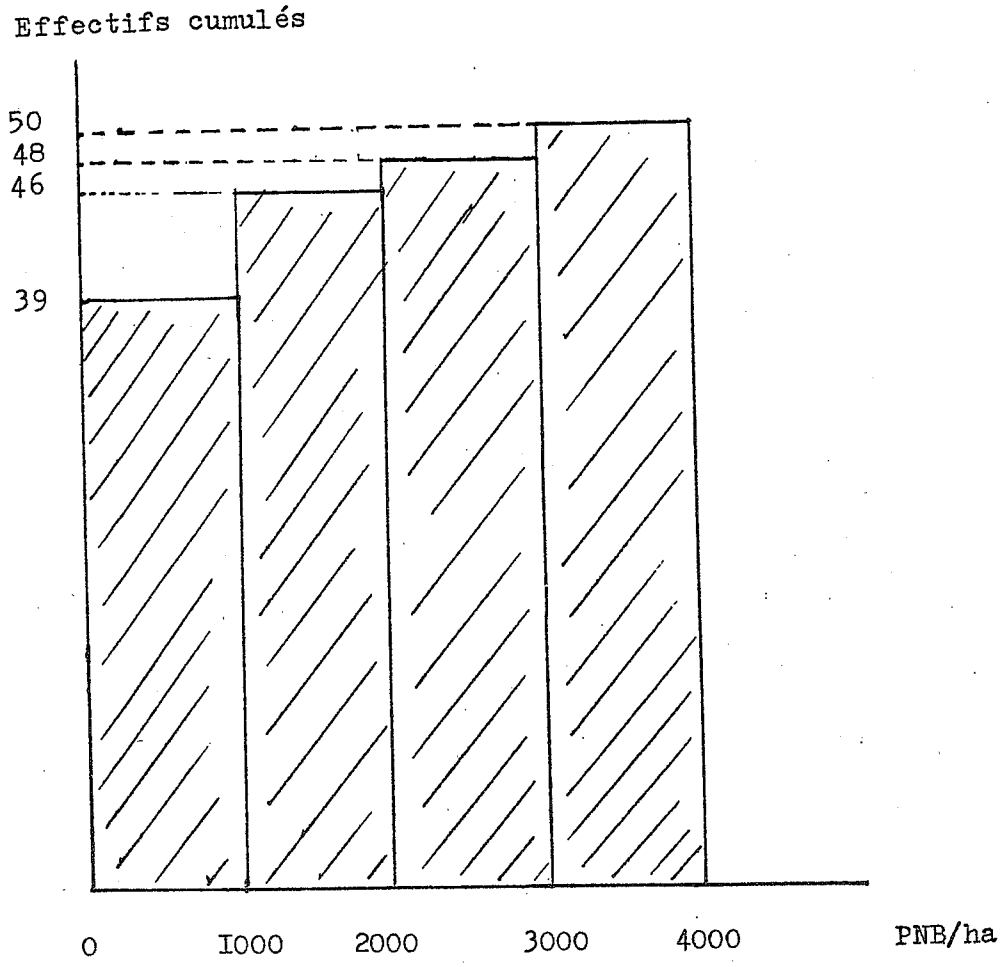


Fig 6.I

CHAPITRE II

VALEURS MOYENNES

Les valeurs moyennes sont des constantes naturellement associées à une série statistique et qui donnent une idée assez globale de la distribution des valeurs de cette série.

I.MEDIANE.

I.I-Cas d'une série isolée. Soit $x_I , \dots , x_k$ (I)

$n_I , \dots , n_k$

une série statistique à valeurs isolées. On supposera qu'on a $x_I < x_2 < \dots < x_k$. On peut alors écrire la série sous la forme $\underbrace{x_I,\dots,x_I}_{n_I \text{ fois}} , \underbrace{x_2,\dots,x_2}_{n_2 \text{ fois}} , \dots , \underbrace{x_k,\dots,x_k}_{n_k \text{ fois}}$.

Par exemple la série

23	34	45	47	68	(valeurs)
?	I	3	3	2	(effectifs)

peut être écrite sous la forme 23,23,34,45,45,45,47 ,47,47,68,68. avec des effectifs partiels tous égaux à I.

N'importe quelle série statistique à valeurs isolées du type (I) se ramène à une série du type $x_I \leqslant x_2 \leqslant \dots \leqslant x_k$ avec des effectifs partiels tous égauxà I.

Dans ce cas la Médiane sera par definition la valeur qui partage la série en 2 séries qui ont le même effectif total. Suivant que k est un nombre pair ou impair on va calculer explicitement la médiane.

i) k impair i.e. s'écrit comme suit : $k = 2r - I$ .La série se met alors sous la forme $x_I , \dots , x_{r-I} , x_r , x_{r+I} , \dots , x_{2r-I}$

la médiane est alors le nombre $M = x_r$.

ii) k pair. On peut l'écrire sous la forme $k = 2r$ .On écrit alors la série $x_I,\dots,x_r , x_{r+I} ,\dots,x_{2r}$. Dans ce cas pour la médiane on prend généralement $M = \frac{x_r + x_{r+I}}{2}$.

I.I.I-Exemples.

i)Soit la série 23,23,34,45,45,45,47,47,47,68,68.

Elle a II = 6.2 - I termes.Par consequent sa médiane est égale au $6^{ème}$ terme c.a.d M = 45.

ii)Soit la série 47,47,56,59 ,68,70,87,87.Elle a 8 = 2.4 termes.

Sa médiane est donc $M = \frac{59 + 68}{2} = 63,5$.

I.2-Cas d'une serie classée.Commençons par regarder les calculs sur un exemple.Soit la série classée donnée par le tableau suivant :

Classes	Effectifs partiels
I0 à 20	2
20 à 30	I
30 à 40	5
40 à 50	2

Tab I.I

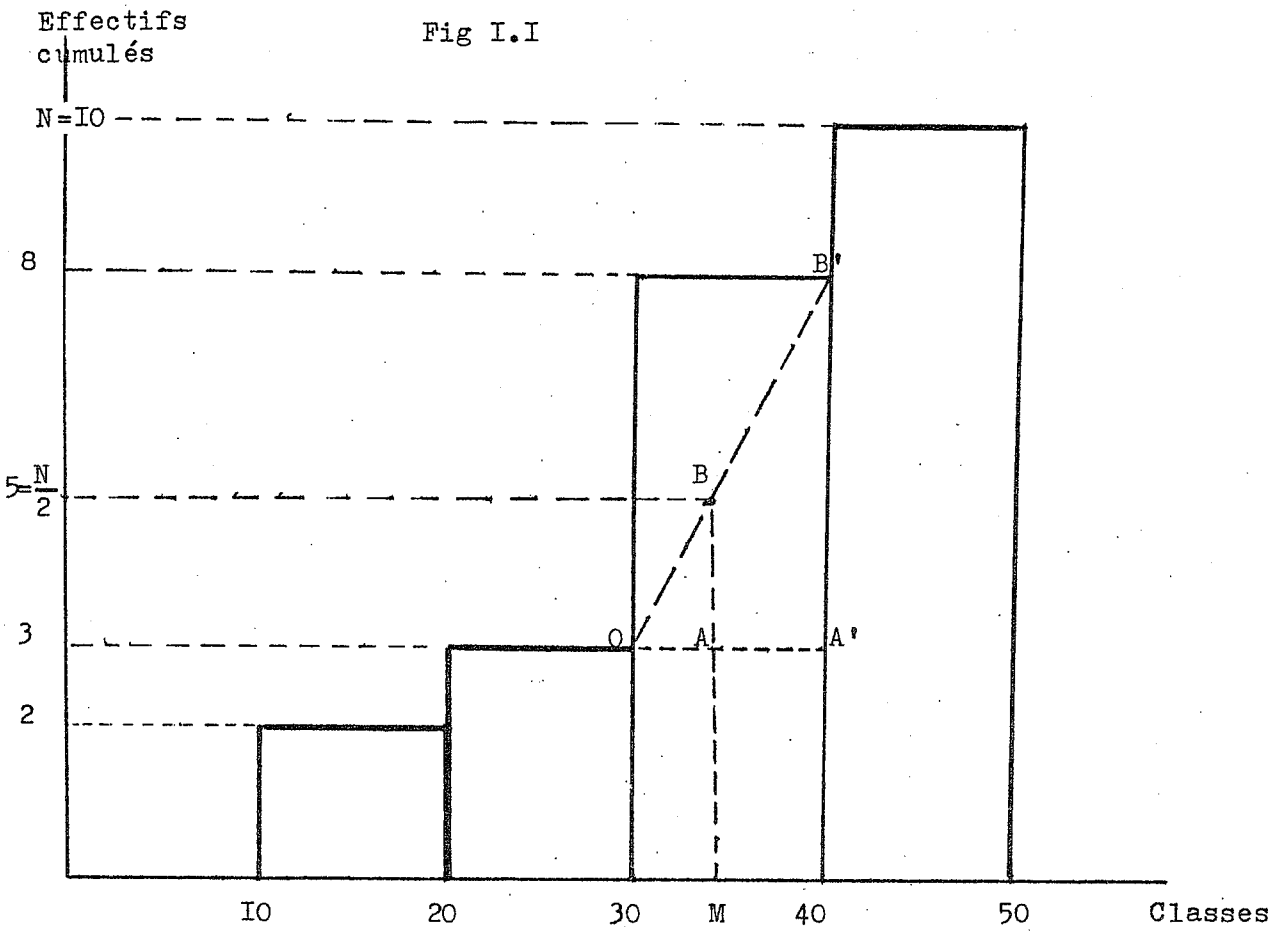
Dans ce cas où toutes les classes ont la même étendue on peut determiner la médiane graphiquement.Pour ce, donnons d'abord la distribution des effectifs cumulés de cette série :

Classes	Effectifs cumulés
I0 à 20	2
20 à 30	3
30 à 40	8
40 à 50	I0

Tab I.2

et traçons l'histogramme associé :

Fig I.I



Les deux I^{ères} ne donnent pas la moitié de l'effectif total et les 3 I^{ères} donnent un effectif superieur à cette moitié.La médiane est donc dans la classe $[30 , 40[$ .Pour la determiner on procède par interpolation proportionnelle.Ceci suppose que la distribution des valeurs de la série dans la classe $[30 , 40[$ est uniforme (car a priori il n'y a aucune raison pour qu'il y ait plus de valeurs proches de 30 que de 40).

Les triangles OAB et O'A'B' sont semblables.Ils ont donc leurs côtés homologues proportionnels.Ceci nous donne :

$$\frac{OA}{OA'} = \frac{AB}{A'B'} \qquad \text{c'est-à-dire } OA = OA' \cdot \frac{AB}{A'B'} .$$

D'autre part $OA' = IO$, $AB = 5 - 3 = 2$ et $A'B' = 8 - 3 = 5$.

En remplaçant les differentes expressions par leurs valeurs respectives on obtient $OA = 4$ et donc $M = 30 + 4 = 34$.

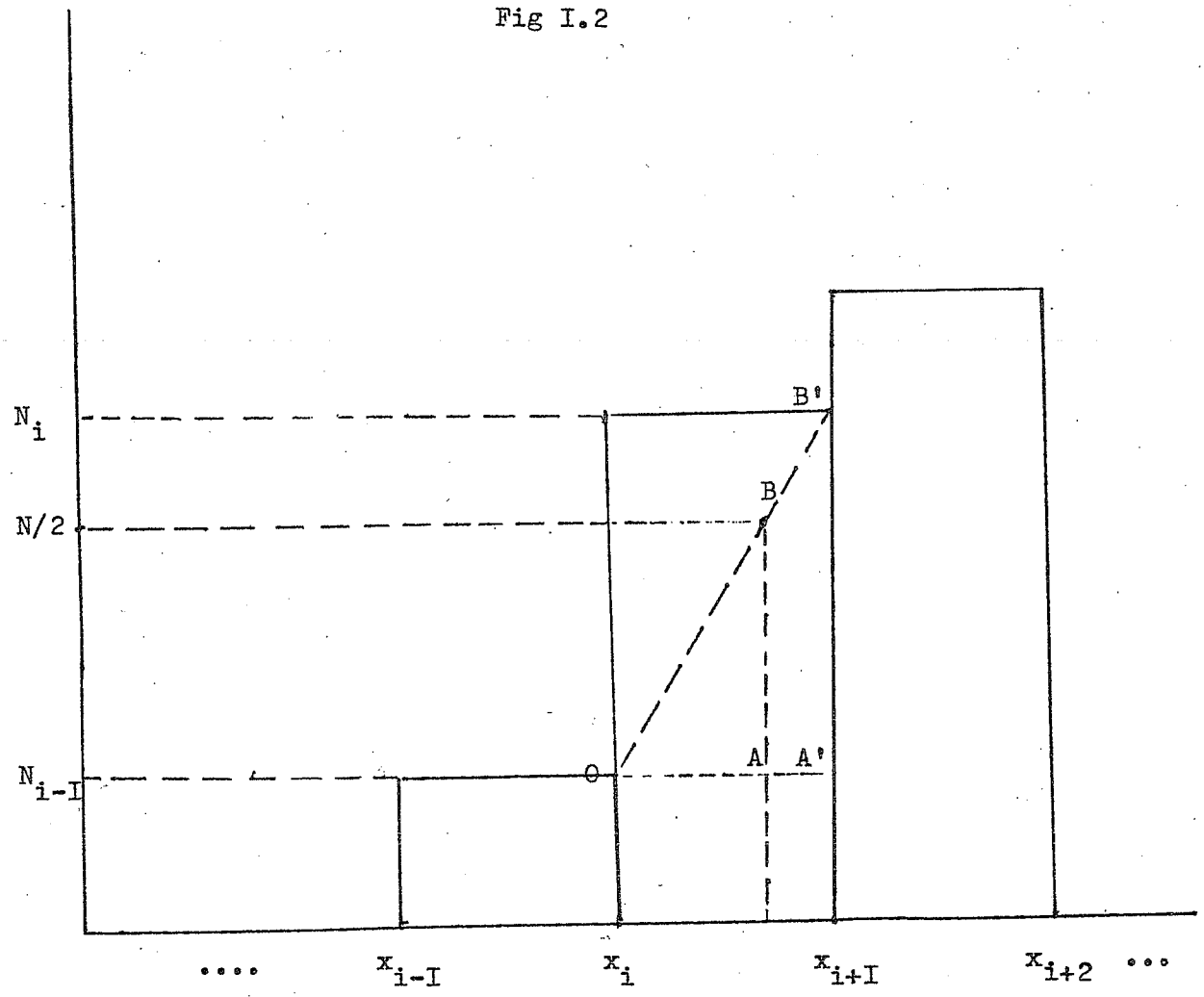
Si l'histogramme des effectifs cumulés est tracé d'une manière precise on determine la médiane en lisant directement l'abscisse du point B dont l'ordonnée est égale à N/2.

Dans le cas général d'une série statistique $[x_1 , x_2[\ldots [x_k , x_{k+1}[$ on supposera que les étendues sont toutes égales. $n_1 \quad \ldots \quad n_k$

On montre sans peine que la médiane appartient à un intervalle $[x_i , x_{i+1}[$. On note N_i l'effectif cumulé de cette classe et on trace l'histogramme des effectifs cumulés associé :

Effectifs cumulés

Fig I.2



Comme précédemment on a l'égalité : $\frac{OA}{OA'} = \frac{AB}{A'B'}$.

D'où l'on deduit $OA = OA' \cdot \frac{AB}{A'B'}$ .D'autre part on a :

$$OA' = e_i = x_{i+I} - x_i \quad , \quad A'B' = N_i - N_{i-I} \quad \text{et} \quad AB = N/2 - N_{i-I} \, .$$

On obtient finalement $OA = e_i \cdot \frac{N/2 - N_{i-I}}{N_i - N_{i-I}}$ et :

$$\boxed{M = x_i + e_i \cdot \frac{N/2 - N_{i-I}}{N_i - N_{i-I}}} \, . \qquad (I.I)$$

Remarque : On peut montrer facilement que cette formule donnant la médiane reste valable lorsque les étendues ne sont pas égales.Par contre la determination graphique directe ne l'est plus.En effet les hauteurs n'étant plus proportionnelles aux effectifs cumulés on ne peut plus repérer N/2.

I.3-Quantiles.On a vu que la médiane est une valeur qui partage la série en 2 séries qui ont le même effectif.De la même façon on définit les quartiles Q_I , Q_2 , Q_3 comme étant les valeurs qui partagent la série en 4 parties qui ont toutes le même effectif.Ainsi soit la série :

$$x_I \leqslant x_2 \leqslant \cdots \leqslant x_k$$ avec des effectifs partiels tous égaux a I (on a vu au I.I qu'on peut toujours se ramener à ce cas pour une série à valeurs isolées bien-sûr).On a alors :

$$\underbrace{x_I \cdots Q_I}_{25\%} \underbrace{\cdots Q_2}_{25\%} \underbrace{\cdots Q_3}_{25\%} \underbrace{\cdots x_k}_{25\%} .$$

Bien-sûr Q_2 n'est rien d'autre que la médiane.Q_I et Q_3 sont appelés

le premier et le troisième quartiles.Le nombre $Q_3 - Q_I$ <u>est l'inter-quartile</u>.Il donne une idée de la façon dont la série est étalée autour de la médiane.

Les <u>deciles</u> $Q_I, \ldots, Q_9$ sont des valeurs qui partagent la série en IO parties qui ont toutes le même effectif.Le nombre $Q_9 - Q_I$ est <u>l'interdecile</u> et joue le même rôle que l'interquartile.

De manière analogue on definit les <u>quantiles d'ordre p</u> comme étant les valeurs $Q_I, \ldots, Q_{p-I}$ qui divisent la série en p parties ayant toutes le même effectif.

Pour determiner gràphiquement les quantiles d'ordre p d'une série classée (ayant des classes de même étendue)on procède de la même façon que pour la mediane mais en choisissant les points d'ordonnées $\frac{N}{p} \ldots \frac{(p - I)N}{p}$.

On pourra à titre d'exercice calculer le I^{er} et le $2^{ème}$ quartiles de la série 2.I du chap I.

2-MOYENNE. Un étudiant passe un examen en 5 matières et obtient les notes suivantes :

	7	I3	I2	II	9,5
coefficients	2	2	I	3	2 .

Pour savoir son niveau il calcule évidemment la "moyenne" de ses notes :

$$\bar{X} = \frac{2.7 + 2.I3 + I.I2 + 3.II + 2.9,5}{2+2+I+3+2} = I0,4 .$$

On peut interpreter ces notes comme étant la série statistique à valeurs isolées :

	7	I3	I2	II	9,5	(valeurs)
	2	2	I	3	2	(effectifs partiels)

à laquelle on attache la valeur I0,4 et qui permet de dire si cet étudiant est ou non reçu à son examen.

Ceci se généralise au cas d'une série statistique à valeurs isolées quelconque.

Soit la série statistique à valeurs isolées:

$$\begin{matrix} x_I , \dots , x_k \\ n_I , \dots , n_k \end{matrix}$$

On appelle <u>moyenne arithmétique</u> de cette série le nombre :

$$(2.I) \qquad \bar{X} = \frac{n_I.x_I + \dots + n_k.x_k}{N}$$ où $N = n_I + \dots + n_k$ est l'effectif total de la série.

Avant de passer au cas d'une série classée on va donner quelques propriétés de la moyenne arithmétique.

— Considèrons une serie X $\begin{matrix} x_I \dots x_k \\ n_I \dots x_k \end{matrix}$ On obtient une nouvelle série Y en ajoutant à chaque terme de X la constante a, c'est à dire que Y est donnée par : $\begin{matrix} y_I = x_I + a , \dots , y_k = x_k + a. \\ n_I \quad , \dots , \quad n_k \end{matrix}$

Par definition la moyenne arithmétique de Y est le nombre :

$$\bar{Y} = \frac{n_I(x_I+a) + \ldots + n_k(x_k+a)}{N}$$

En developpant on obtient:

$$\bar{Y} = \frac{(n_I x_I + n_I a) + \ldots + (n_k x_k + n_k a)}{N}$$

$$= \frac{(n_I x_I + \ldots n_k x_k) + (n_I + \ldots + n_k)a}{N}$$

$$(2.2) \quad \bar{Y} = \frac{n_I x_I + \ldots n_k x_k}{N} + \frac{N}{N}a = \bar{X} + a \ .$$

La moyenne de la série Y est donc obtenue en ajoutant la constante a à celle de la série X.

On peut remarquer que la moyenne d'une série dont toutes les valeurs sont égales à une même constante est elle aussi égale à cette constante.La verification est immediate et est laissée au lecteur comme exercice.

—Soient X et Y deux séries : $x_I , \ldots , x_k$ et $n_I , \ldots , n_k$

$y_I , \ldots , y_k$ dont les valeurs respectives ont les mêmes effectifs partiels $n_I , \ldots , n_k$.On note Z = X+Y la somme de ces deux séries c'est à dire la série dont les termes sont les sommes respectives des termes de X et Y. On peut l'écrire donc : $z_I = x_I + y_I , \ldots , z_k = x_k + y_k$ avec $n_I , \ldots , n_k$

Calculons sa moyenne arithmétique.

$$\bar{Z} = \frac{n_I z_I + \ldots + n_k z_k}{N} = \frac{n_I(x_I + y_I) + \ldots + n_k(x_k + y_k)}{N}$$

On developpe :

$$\bar{Z} = \frac{(n_I x_I + \ldots + n_k x_k)}{N} + \frac{(n_I y_I + \ldots + n_k y_k)}{N}$$

c'est-à-dire :

(2.3) $$\boxed{\bar{Z} = \bar{X} + \bar{Y}}$$

La moyenne arithmétique de deux séries ayant les mêmes effectifs partiels est égale à la somme des moyennes des deux séries.

-Considèrons toujours une serie X $\begin{matrix} x_I & , \ldots , & x_k \\ n_I & , \ldots , & n_k \end{matrix}$ et multiplions toutes ses valeurs par une même constante a. Calculons la moyenne arithmétique de cette série. On a en notant $\bar{P}$ cette moyenne :

$$\bar{P} = \frac{n_I(ax_I) + \ldots + n_k(ax_k)}{N} \quad ;$$

on met a en facteur commun et on obtient:

(2.4) $$\bar{P} = \frac{n_I x_I + \ldots + n_k x_k}{N}, \quad a = a.\bar{X} \quad .$$

Si on multiplie toutes les valeurs d'une série par une même constante la moyenne de la série est aussi multipliée par cette constante pour donner la moyenne de la nouvelle série obtenue.

-En raison de la commutativité de l'addition des nombres reels la moyenne arithmétique d'une série ne dépend pas de l'ordre dans lequel on écrit ses termes.

Moyenne arithmétique d'une série classée.

Pour une série classée on définit sa moyenne comme étant celle de la série des centres associée. Plus précisément soit $\underset{n_I}{[x_I, x_2[} \ \ldots \ \underset{n_k}{[x_k, x_{k+I}[}$ et notons $c_i = \frac{x_i + x_{i+I}}{2}$ le centre de la classe $[x_i, x_{i+I}[$. On obtient ainsi la série à valeurs isolées : $\begin{matrix} c_I, \ \ldots \ , c_k \\ n_I, \ \ldots \ , n_k \end{matrix}$.

Par définition on pose :

$$(2.5) \quad \boxed{\bar{X} = \bar{C} = \frac{n_I c_I + \ldots + n_k c_k}{N}} \quad \text{avec} \quad N = n_I + \ldots + n_k .$$

Exemple.

Le tableau suivant donne une estimation de la population de 27 pays européens :

Population (en millions)	Nombre de pays
Moins de 5	7
5 à I0	7
I0 à I5	3
I5 à 20	2
20 à 25	2
25 à 40	2
40 à 55	I
55 à 70	3

Tab 2.I

Cette statistique est tirée de"Dossiers et Documents Le Monde" N° I08 du mois de Fev I984).

La population de l'URSS n'est pas comptée dans le tableau.Elle est de 270 millions.La compter nous oblige à introduire une classe d'une grande étendue relativement aux autres alors que ce n'est vraiment pas nécessaire pour comprendre la méthode de calcul.

Calculons la série des centres.On obtient alors la série à valeurs isolées :

2,5	7,5	I2,5	I7,5	22,5	32,5	47,5	62,5	(valeurs)
7	7	3	2	2	2	I	3	(effectifs)

En utilisant la formule (2.I) et en designant par $\bar{P}$ la population moyenne on aura :

$$\bar{P} = \frac{7.2,5 + 7.7,5 + 3.I2,5 + 2.I7,5 + 2.22,5 + 2.32,5 + I.47,5 + 3.62,5}{27}$$

ceci donne $\bar{P}$ = I8,06 Millions ha.

Le calcul de la moyenne arithmétique risque quelquefois d'être un peu lourd quand on le fait à la main.Pour ce on le ramène à celui d'une série plus simple en effectuant un changement de variable.

Soit $\begin{matrix} x_I & \cdots & x_k \\ n_I & \cdots & n_k \end{matrix}$ une série à valeurs isolées et posons pour tout i=I,...,k

$z_i = \frac{x_i - x_o}{e}$ où x_o et e sont des nombres reels quelconques mais e non nul.Ceci nous donne une nouvelle série Z $\begin{matrix} z_I & \cdots & z_k \\ n_I & \cdots & n_k \end{matrix}$ .De manière équivalente on a : $x_i = e.z_i + x_o$.

On peut donc considerer la série X comme la somme de 2 séries qui sont :

La serie Y $\begin{matrix} e.z_I & \cdots & e.z_k \\ n_I & \cdots & n_k \end{matrix}$ et la série dont tous les termes sont égaux à x_o .Cette dernière a pour moyenne arithmétique x_o.Pour Y on applique la formule (2.4) et on obtient $\bar{Y} = e.\bar{Z}$.D'où finalement :

(2.6) $\bar{X} = e.\bar{Z} + x_o$.

Comme exemple d'application on prend X = P (Tab 2.I).

On applique donc ce changement de variable à la série du Tab 2?I en prenant $x_o = 22,5$ et $e = 5$ par exemple .La série que l'on obtient sera

-4	-3	-2	-I	0	2	5	8	(valeurs)
7	7	3	2	2	2	I	3	(effectifs)

Cette série est obtenue en retranchant 22,5 à tous les termes de la série 2,5 7,5 I2,5 I7,5 22,5 32,5 47,5 62,5 et en divisant ensuite par 5.La formule (2.6) étant vraie pour tout x_o et tout e non nul on choisit évidemment ces deux valeurs de façon à simplifier au maximum la série.

La nouvelle série obtenue a donc pour moyenne :

$$\bar{Z} = \frac{7.(-4) +7.(-3) +3.(-2) +2.(-I) +2.0 +2.2 +I.5 +3.8}{27} = -0,89 .$$

D'ou $\bar{X} = 5.(-0,89) + 22,5 = I8,06$ .On retrouve ainsi le resultat obtenu par la méthode directe.

<u>Autre expression de la moyenne arithmétique.</u> Soit X une série statistique à valeurs isolées : $x_I,\ldots,x_k$; $n_I,\ldots,n_k$; $N = n_I + \ldots + n_k$.

On sait que la moyenne arithmetique de cette série s'écrit :

$$\bar{X} = \frac{n_I x_I + \ldots n_k x_k}{N} ,$$

ou encore :

$$\bar{X} = \frac{n_I}{N} x_I + \ldots + \frac{n_k}{N} x_k = f_I x_I + \ldots + f_k x_k .$$

$$\bar{X} = f_I x_I + \ldots + f_k x_k \quad . \quad (2.7)$$

AUTRES MOYENNES. Il existe d'autres moyennes que la moyenne arithmétique et qui sont d'un usage en statistiques. Donnons brièvement la definition de quelques-unes d'entre elles :

-Moyenne géométique. Soit $\begin{matrix} x_I , \dots , x_k \\ n_I , \dots , n_k \end{matrix}$ une série statistique à valeurs isolées.

Par definition la moyenne géométrique de cette série est le nombre :

$$G = \sqrt[N]{x^{n_I} . \dots . x^{n_k}} \qquad (2.8)$$ (on suppose x_i positif pour tout i)

Par exemple la moyenne géométrique de la série $\begin{matrix} 3 & II & I2 \\ 2 & I & 3 \end{matrix}$ (valeurs) (effectifs)

est donnée par :

$$G = \sqrt[6]{3^2 . II^I . I2^3} = \sqrt[6]{I7I072} = 7,45 .$$

Pour calculer la moyenne géométrique il est souvent plus commode de calculer LogG. Plus précisement on a :

$$LogG = Log \sqrt[N]{x_I^{n_I} . \dots . x_k^{n_k}}$$

$$LogG = \frac{n_I Logx_I + \dots + n_k Logx_k}{N} . \qquad (2.9)$$

Par exemple pour la série ci-dessus on a :

$$LogG = \frac{2.Log3 + I.LogII + 3.LogI2}{6}$$

$$= \frac{2.I,0986 + I.2,3978 + 3.2,4849}{6} = 2,0082.$$

En cherchant le nombre reel G dont le logarithme neperien vaut 2,0082 on trouve G = 7,4505. Ce qui correspond au resultat par le calcul direct. L'avantage de passer à une série de logarithmes est évidemment de ramener

le calcul d'une moyenne géométrique à celui d'une moyenne arithmetique.

-Moyenne harmonique.On se donne toujours une série :

$x_I , \dots , x_k$
$n_I , \dots , n_k$

La moyenne harmonique de cette série est le nombre H defini par la relation :

$$\frac{N}{H} = \frac{n_I}{x_I} + \dots + \frac{n_k}{x_k} \quad . \qquad (2.10)$$

On suppose bien-sûr que x_i est non nul pour tout $i = I, \dots, k$.

Le nombre H n'est pas défini d'une manière explicite; il est préférable de calculer d'abord le membre de droite.Par exemple pour la série :

3	II	I2	(valeurs)
2	I	3	(effectifs)

on a :

$$\frac{6}{H} = \frac{2}{3} + \frac{I}{II} + \frac{3}{I2} = \frac{88 + I^2 + 33}{I32} = \frac{I33}{I32} \quad .$$

D'où $H = \frac{I32.6}{I33}$,c'est-à-dire $H = 5{,}954$.

Evidemment dans le cas d'une série classée la moyenne géométrique et la moyenne harmonique sont celles de la série des centres associée.

3-MODE.On appelle mode d'une série statistique à valeurs isolées ou classe modale dans le cas d'une série classée,la valeur de la série,ou la classe,qui a le plus grand effectif.

Par exemple le mode de la série : 3 3 3 4 II I3 I3 I3 est la valeur 3 ou I3.

Cet exemple montre que certaines séries peuvent avoir plusieurs modes ou plusieurs classes modales si la série est classée.

EXERCICES

I) Soit la série à valeurs isolées x_I , x_2 (valeurs) x_I et x_2 positifs.
I I (effectifs) .

On designe par $\bar{X}$, G , H respectivement les moyennes arithmétique, géométrique et harmonique.

Demontrer la double inégalité :

$$H \leqslant G \leqslant \bar{X} \quad .$$

II)

On considère la statistique suivante :

I39	I7I	I34	I55	I44	I53	I75	I36	I49	I54
I37	I40	I42	I68	I52	I49	I48	I55	I68	I44
I34	I44	I37	I56	I53	I49	I69	I58	I47	I50
I52	I40	I62	I53	I77	I46	I52	I40	I45	I52
I5I	I45	I57	I56	I60	I70	I65	I58	I58	I6I

Tab 2.2

I°-Mettre cette série sous forme d'une série classée dont les classes ont une étendue égale à IO.

2°-Construire l'histogramme des effectifs cumulés .

3°-Calculer la médiane de cette série statistique. Calculer le I^{er} et le $3^{ème}$ quartiles.

4°-Calculer la moyenne arithmétique de cette série en effectuant un changement de variable (Prendre x_o=I55 et e = IO).

III) Tracer l'histogramme des effectifs de la série :

Classes	[10 , 20[	[20 , 40[	[40 , 70[
Effectifs partiels	6	20	15

Expliquer comment vous faites la construction des différents rectangles.

SOLUTIONS

I) Si $x_I = x_2 = a$ on a immediatement :

$$\bar{X} = \frac{x_I + x_2}{2} = \frac{a+a}{2} = a \quad , \quad G = \sqrt{x_I \cdot x_2} = \sqrt{a \cdot a} = a \quad \text{et}$$

$$\frac{2}{H} = \frac{I}{x_I} + \frac{I}{x_2} = \frac{I}{a} + \frac{I}{a} = \frac{2}{a} \quad \text{ce qui donne} \quad H = a \ .$$

D'où : $\bar{X} = G = H$.

Supposons x_I et x_2 distincts .

Pour comparer 2 réels positifs il suffit de comparer leurs carrés. On a successivement :

$$G^2 = x_I \cdot x_2 \quad , \quad \bar{X}^2 = \frac{(x_I + x_2)^2}{4} \quad \text{et de} \quad \frac{2}{H} = \frac{I}{x_I} + \frac{I}{x_2} \quad \text{on tire}$$

$$H = \frac{2x_I x_2}{x_I + x_2} \ .$$

Calculons $\bar{X}^2 - G^2 = \frac{(x_I + x_2)^2}{4} - x_I x_2 = \frac{(x_I + x_2)^2 - 4x_I x_2}{4}$

$$= \frac{(x_I^2 + x_2^2 + 2x_I x_2) - 4x_I x_2}{4} = \frac{x_I^2 + x_2^2 - 2x_I x_2}{4}$$

$$= \frac{(x_I - x_2)^2}{4}$$

Cette dernière quantité est toujours positive car c'est un carré.On en deduit $\bar{X}^2 - G^2 > 0$ et donc $\bar{X} > G$.

C.Q.F.D.

Calculons maintenant la difference :

$$G^2 - H^2 = x_I x_2 - \frac{4x_I^2 x_2^2}{(x_I+x_2)^2} = \frac{x_I x_2 (x_I+x_2)^2 - 4x_I^2 x_2^2}{(x_I+x_2)^2} .$$

On met $x_I x_2$ en facteur au numérateur;on obtient :

$$G^2 - H^2 = \frac{x_I x_2 ((x_I+x_2)^2 - 4x_I x_2)}{(x_I+x_2)^2} = \frac{x_I x_2 (x_I - x_2)^2}{(x_I+x_2)^2}$$

Ceci montre que $G^2 - H^2$ est toujours positif et donc $G > H$.

C.Q.F.D.

II)

I°-On remarque que les valeurs de cette série sont comprises entre I30 et I80.Nous allons donc constituer les classes de la façon suivante :

Classes	Effectifs partiels
[I30 , I40 [	6
[I40 , I50 [	I5
[I50 , I60 [	I8
[I60 , I70 [	7
[I70 , I80 [	4

Tab 2.4

Il y a 5 classes d'étendue égale à IO.

2°-On calcule d'abord les effectifs cumulés des différentes classes.On a le tableau suivant :

Classes	Effectifs cumulés
[I30 , I40[	6
[I40 , I50[	2I
[I50 , I60[	39
[I60 , I70[	46
[I70 , I80[	50

Tab 2.5

On trace l'histogramme correspondant en choisissant les hauteurs des rectangles proportionnelles aux effectifs cumulés puisque on est dans le cas où les classes ont la même étendue :

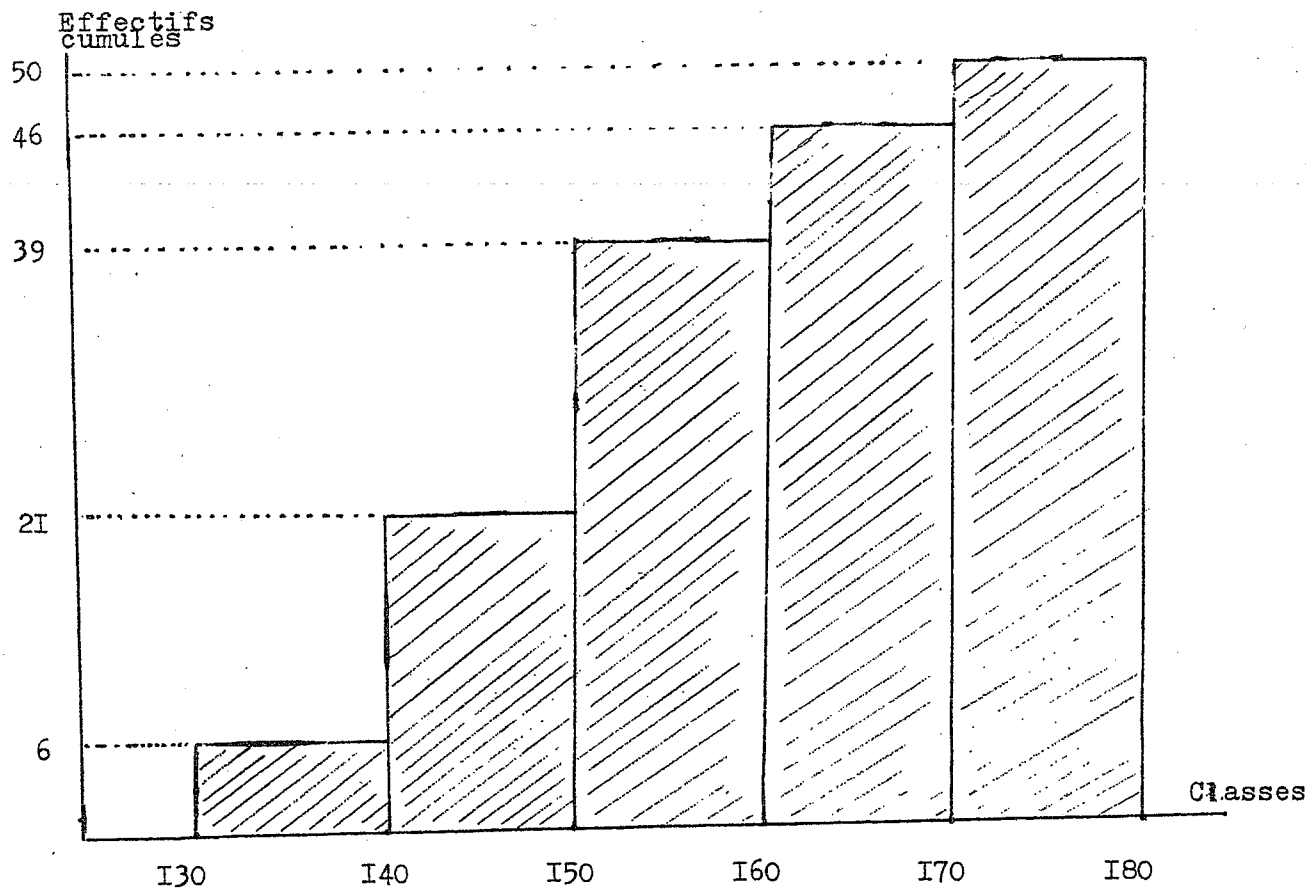


Fig 2.I

3°-L'effectif total de cette série estégal à 50.La moitié de l'effectif N/2 = 25 est dans l'itervalle associée à la classe $[I50 , I60[$ .Pour calculer la médiane M on procède comme au I.2 :

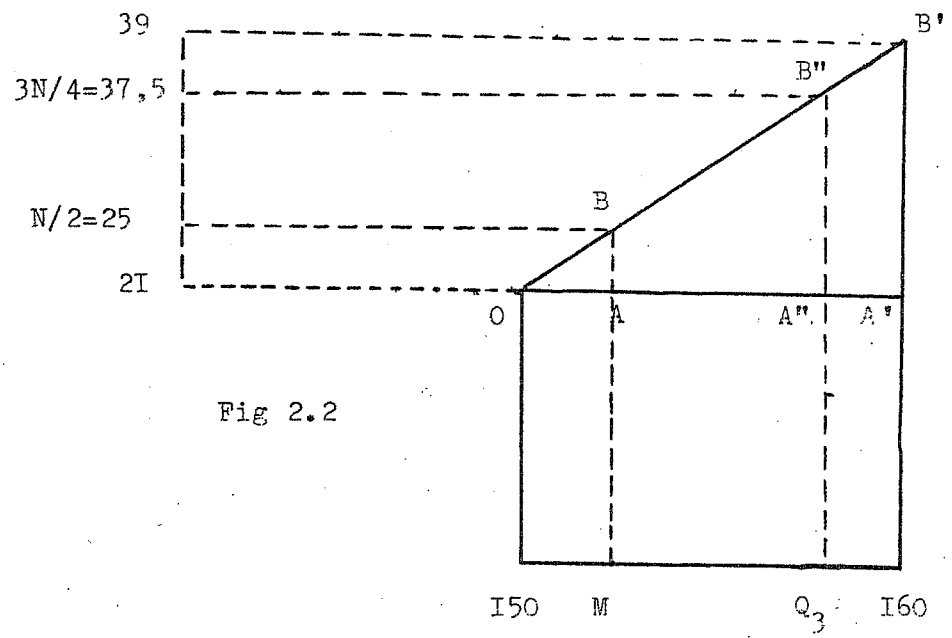


Fig 2.2

On a $\frac{OA}{OA'} = \frac{AB}{A'B'}$ c'est à dire $\frac{M - I50}{I60-I50} = \frac{25-2I}{39-2I}$ .D'où on deduit :

$$M = I50 + \frac{I0.4}{I8} = I50 + 2,22 = I52,22 .$$

Pour calculer le I^{er} et le $3^{ème}$ quartiles Q_I et Q_3 on considère le I/4 et les 3/4 de l'effectif total qui sont respectivement N/4 = I2,5 et 3N/4 = 37,5.Faisons le calcul par exemple de Q_3 (celui de Q_I se fait d'une manière analogue) :

Le nombre 3N/4 est dans l'intervalle de la classe $[I50 , I60[$.
En procèdant comme pour la médiane on obtient (voir Fig 2.2) :

$$\frac{OA''}{OA'} = \frac{A''B''}{A'B'} \quad \text{c'est à dire} \quad \frac{Q_3 - I50}{I60- I50} = \frac{37,5- 2I}{39 - 2I}$$

Ce qui donne $Q_3 = I50 + \frac{I0.I6,5}{I8} = I59,I6$.

Faire attention pour le calcul de Q_I car N/4 = I2,5 est dans l'intervalle associé à la classe $[I40 , I50[$.

4°-La série des centres est la série à valeurs isolées :

I35	I45	I55	I65	I75	(Valeurs)
6	I5	I8	7	4	(effectifs partiels)

Si on designe par x_i le terme général de cette série,on obtient en effectuant le changement de variable :

$$z_i = \frac{x_i - x_o}{e} = \frac{x_i - I55}{I0}$$

une nouvelle série à valeurs isolées :

-2	-I	0	I	2	(valeurs)
6	I5	I8	7	4	(effectifs partiels)

et qui a comme moyenne arithmétique :

$$\bar{Z} = \frac{6.(-2) + I5.(-I) + I8.0 + 7.I + 4.2}{50} = -0,24 .$$

On en deduit :

$$\bar{X} = e.\bar{Z} + x_o = I0.(-0,24) + I55 = I52,6 .$$

III)

Notons h_I, h_2 et h_3 les hauteurs des différents rectangles de l'histogramme des effectifs respectivement associés aux classes $[I0 , 20[$, $[20 , 40[$ et $[40 , 70[$ .Les aires de ces rectangles seront donc :

$$S_I = (20 - I0)h_I \quad , \quad S_2 = (40 - 20)h_2 \quad \text{et} \quad S_3 = (70 - 40)h_3 .$$

Les aires : S_I , S_2 et S_3 doivent être proportionnelles aux effectifs partiels 6 , 20 et I5 ,c'est à dire :

$$\frac{S_I}{6} = \frac{S_2}{20} = \frac{S_3}{I5} \quad \text{d'où on tire les relations :}$$

$$h_2 = 5h_I/3 \quad \text{et} \quad h_3 = 5h_I/6 .$$

Si donc on choisit h_I = 6 cm par exemple on doit prendre h_2 = IO cm et h_3 = 5 cm .On obtient finalement l'histogramme suivant :

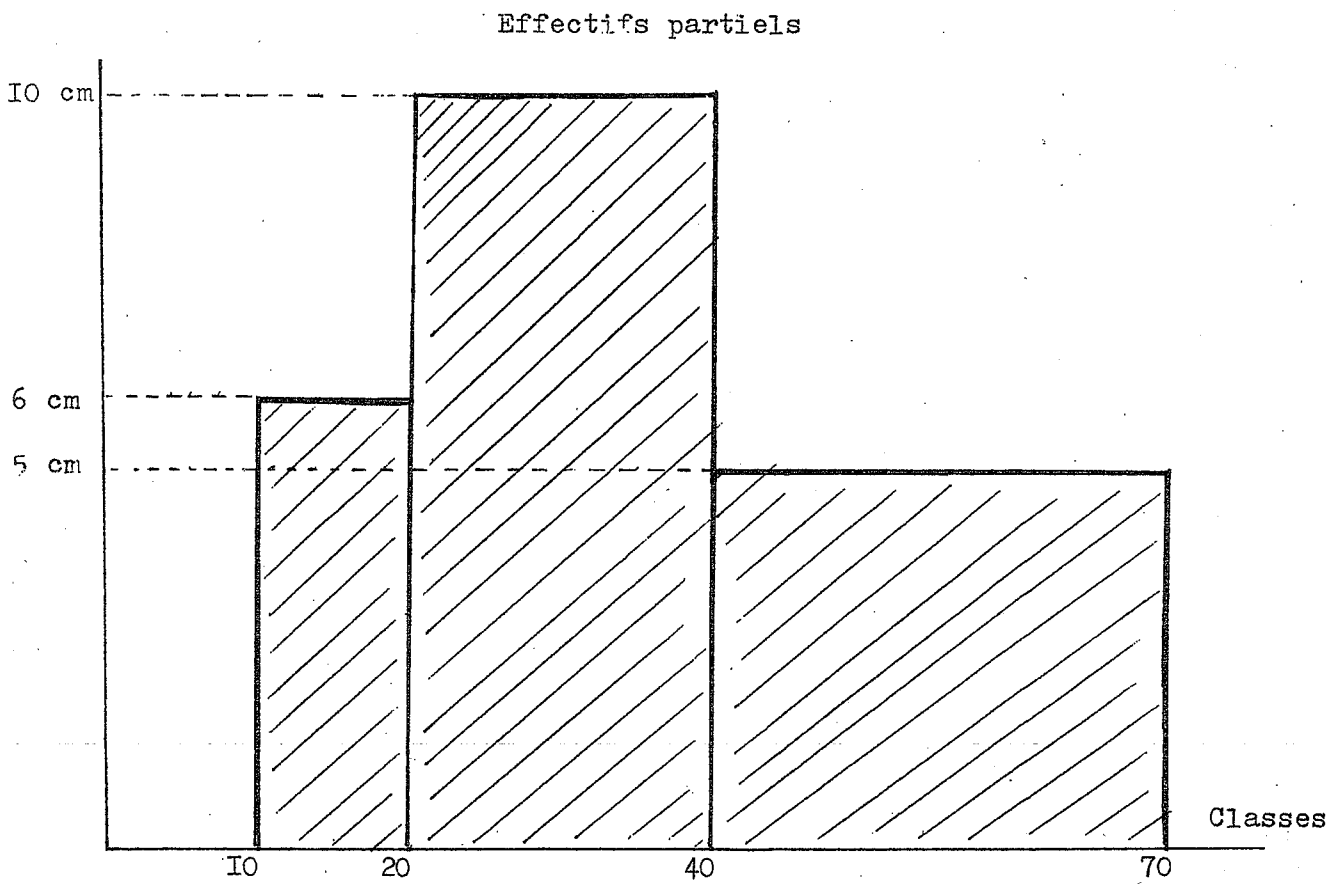


Fig 3.3

CHAPITRE III

PARAMETRES DE DISPERSION

Considèrons les séries statistiques suivantes :

X : 2 4 9 I3 I7

Y : 7 9 9 I0 I0 .

Les effectifs partiels sont tous égaux à I aussi bien pour X que pour Y. Il est facile de voir que ces deux séries ont la même moyenne arithmetique $\bar{X} = \bar{Y} = 9$ et la même médiane aussi $M_X = M_Y = 9$ .Si on intèrprète ces deux séries comme les notes de deux étudiants X et Y,on dira que Y est plus régulier que X.La moyenne et la médiane -plus généralement les valeurs moyennes-sont donc insuffisantes pour donner "certains renseignements" sur la série.Il faut introduire d'autres paramètres qui permettent de mesurer la dispersion d'une série autour de sa moyenne.C'est ce que nous allons faire dans ce chapitre.

I.VARIANCE.

Soit $X : \begin{matrix} x_I , \dots , x_k \\ n_I , \dots , n_k \end{matrix}$ une série à valeurs isolées. (I)

I.I-Definition.On appelle variance de X et on note Var(X) le nombre defini par :

$$\boxed{Var(X) = \frac{I}{N}(n_I(x_I - \bar{X})^2 + \dots + n_k(x_k - \bar{X})^2)} \qquad (I.I)$$

On remarque que $Var(X) \geqslant 0$ et ce n'est rien d'autre que la moyenne arithmétique de la série $\begin{matrix} (x_I - \bar{X})^2, \dots , (x_k - \bar{X})^2 \\ n_I , \dots , n_k \end{matrix}$.

I.2-Exemple.Calculons les variances des séries X et Y ci-dessus.On peut résumer les calculs dans le tableau suivant :

X	2	4	9	I3	I7
Effectifs	I	I	I	I	I
$x_i - \bar{X}$	-7	-5	0	4	8
$(x_i - \bar{X})^2$	49	25	0	I6	64
Y	7	9	9	I0	I0
Effectifs	I	I	I	I	I
$y_i - \bar{Y}$	-2	0	0	I	I
$(y_i - \bar{Y})^2$	4	0	0	I	I

On en deduit donc $\mathrm{Var}(X) = \frac{49 + 25 + 0 + I6 + 64}{5} = 30,8$.

$$\mathrm{Var}(Y) = \frac{4 + 0 + 0 + I + I}{5} = I,2 .$$

On voit donc sur cet exemple que plus la variance est grande, plus la série est dispersée autour de la moyenne.

Le cas où Var(X) est égale à 0 signifie que chaque terme de la série

$(x_I - \bar{X})^2$, ... , $(x_k - \bar{X})^2$

n_I , ... , n_k

est nul et donc pour tout i = I , ... ,k on a $x_i = \bar{X}$,c'est-à-dire que la série est constante.

I.3-Définition. Soit $\underset{n_I}{[x_I\ ,\ x_2[}\ ,\ \ldots\ ,\ \underset{n_k}{[x_k\ ,\ x_{k+I}[}$ une série classée.On appelle variance de cette série la variance de la série des centres des classes.

On peut donc toujours supposer que la série est à valeurs isolées tout le long de ce chapitre.

I.4-Quelques propriétés et règles de calcul.

I.4.I-Autre expression de la variance.

Par définition on a :

$$Var(X) = \frac{I}{N}\left(n_I(x_I - \bar{X})^2 + \ldots + n_k(x_k - \bar{X})^2\right) \quad .$$

On developpe chaque terme $(x_i - \bar{X})^2 = x_i^2 - 2\bar{X}x_i + \bar{X}^2$.On obtient :

$$Var(X) = \frac{(n_I x_I^2 + \ldots + n_k x_k^2) - 2\bar{X}(n_I x_I + \ldots + n_k x_k) + (n_I + \ldots + n_k)\bar{X}^2}{N}$$

$$= \frac{n_I x_I^2 + \ldots + n_k x_k^2}{N} - 2\bar{X}\,\frac{n_I x_I + \ldots + n_k x_k}{N} + \frac{N}{N}\bar{X}^2$$

Or on a :

$$\bar{X} = \frac{n_I x_I + \ldots + n_k x_k}{N} \quad \text{et} \quad \frac{n_I x_I^2 + \ldots + n_k x_k^2}{N} = \text{Moyenne de } X^2 \ .$$

D'où on deduit :

$$Var(X) = \text{Moyenne}(X^2) - 2\bar{X}^2 + \bar{X}^2 = \overline{X^2} - \bar{X}^2 \qquad (I.2)$$

Cette formule nous donne une façon de calculer la variance d'une série .

Exemple d'application.

Reprenons la série : X : 2 4 9 I3 I7 (valeurs)

I I I I I (effectifs)

La série X^2 est obtenue à partir de X en élevant toutes ses valeurs au carré :

X^2 : 4 I6 8I I69 289 (valeurs)

I I I I I (effectifs)

D'où :

$$\overline{X^2} = \frac{4 + I6 + 8I + I69 + 289}{5} = III,8$$

et par consequent

$$Var(X) = \overline{X^2} - \bar{X}^2 = III,8 - 8I = 30,8 .$$

Pour calculer la variance on pourra se servir aussi bien de la formule I.I que I.2;mais évidemment suivant le type de problèmes il faut choisir celle qui facilite le calcul.

I.4.2-Propriétés de la variance.

On considère toujours une série X du type (I) (voir p.35).Soit a un nombre reel.Calculons Var(X + a).Posons Z = X + a .On a :

Var(Z) = Moyenne arithmétique de $(Z - \bar{Z})^2$.

Or $\bar{Z}$ = Moyenne arithmétique de (X + a) = $\bar{X}$ + a d'après la formule (2.2) du Chap II (voir p.2I).D'où :

Var(Z) = Moyenne arithmétique de $(X + a - \bar{X} - a)^2$

= Moyenne arithmétique de $(X - \bar{X})^2$

= Var(X) d'après la remarque qui suit la definition I.I (voir p.35).On a donc :

$$\boxed{Var(X + a) = Var(X)} \qquad (I.3)$$

Posons maintenant P = aX .On a alors:

Var(P) = Var(aX) = Moyenne arithmétique de $(aX - \overline{aX})^2$.

Or $\overline{aX} = a\overline{X}$.D'où :

Moyenne arithmétique de $(aX - a\overline{X})^2$ = Moyenne arithmétique de la série $a^2(X - \overline{X})^2$.D'après la formule (2.4) (voir p.22) on a :

Moyenne arithmétique de $a^2(X - \overline{X})^2$ = a^2(Moyenne arithmétique de $(X - \overline{X})^2$ .Et finalement on obtient le resultat :

$$\boxed{Var(aX) = a^2 Var(X)} \qquad (I.4)$$

I.4.3-Calcul de la variance par changement de variable.

Soient x_o et $e \neq 0$ deux nombres reels .Posons pour tout i =I ,...,k

$z_i = \dfrac{x_i - x_o}{e}$ .Ceci est équivalent à $x_i = ez_i + x_o$.

Par (I.3) (p.38) on a : Var(eZ + a) = Var(eZ).Ce qui est aussi égal à e^2Var(Z) (formule I.4 ci-dessus).On a finalement:

$$\boxed{Var(X) = e^2 Var(Z)} \qquad (I.5)$$

Exemple d'application.

Calculons la variance de la série classée de l'exercice II p.28.Sa série des centres est donnée à la page.33 :

X	I35	I45	I55	I65	I75
	6	I5	I8	7	4

En effectuant le changement de variable :

$$z_i = \frac{x_i - I55}{I0}$$

on obtient une nouvelle série

Z	-2	-I	0	I	2	(valeurs)
	6	I5	I8	7	4	(effectifs)

On calcule donc la variance de Z en suivant les étapes rassemblées dans le tableau :

Valeurs	-2	-I	0	I	2
Effectifs	6	I5	I8	7	4
z_i^2	4	I	0	I	4

$$\text{Moyenne de } Z = \frac{6.(-2) + I5.(-I) + I8.0 + 7.I + 4.2}{50} = -0,24 .$$

$$\text{Moyenne de } Z^2 = \frac{6.4 + I5.I + I8.0 + 7.I + 4.4}{50} = I,24 .$$

D'où :

$$\text{Var}(Z) = \text{Moyenne de } Z^2 - \bar{Z}^2 = I,24 - 0,0576 = I,I824$$

et donc :

$$\text{Var}(X) = I0^2 \text{Var}(Z) = I00 . I,I824 = II8,24 .$$

2.ECART-TYPE.

Souvent pour mesurer la dispersion d'une série autour de sa moyenne on utilise une autre grandeur appelée écart-type et qui est définie par l'égalité :

$$\boxed{\sigma(X) = \sqrt{\text{Var}(X)}} \qquad (I.6) .$$

Par exemple pour les séries qu'on a considérées à la page 35 on a :

$\text{Var}(X) = 30,8$ et $\text{Var}(Y) = I,2$ et donc :

$$\sigma(X) = \sqrt{30,8} = 5,54$$

$$\sigma(Y) = \sqrt{I,2} = I,09 .$$

EXERCICE RESOLU

On considère les deux séries statistiques suivantes donnant la répartition en fonction de l'âge de la population espagnole en I960. Les données ont été obtenues à partir de deux échantillons de I0000 personnes de chaque sexe.

Classes	Hommes	Femmes
0 à 5	I04I	943
5 à I0	948	825
I0 à I5	903	833
I5 à 25	I565	I484
25 à 35	I600	I55I
35 à 45	I30I	I35I
45 à 55	I083	II5I
55 à 65	855	926
65 à 75	704	936

I°-Donner la distribution des effectifs cumulés et calculer l'âge médian masculin.

2°-Calculer l'âge moyen pour :

-Le sexe masculin,

-le sexe féminin.

3°-Calculer la variance et l'écart-type de chacune de ces deux séries.

SOLUTION.

I°-Rappelons que l'effectif cumulé d'une classe est égal à la somme de l'effectif de cette classe et ceux de toutes les classes qui la précèdent.On a ainsi :

Classes	Effectifs cumulés sexe masculin	Effectifs cumulés sexe feminin
0 à 5	I04I	943
5 à I0	I989	I768
I0 à I5	2892	260I
I5 à 25	4457	4085
25 à 35	6057	5636
35 à 45	7358	6987
45 à 55	844I	8I38
55 à 65	9296	9064
65 à 75	I0000	I0000

On voit donc que l'âge médian masculin tombe dans la classe $[25 , 35[$. Notons le M_h .Pour le calculer on peut utiliser la formule (I.I) (voir p.I8)

$$M_h = x_i + e_i \frac{N/2 - N_{i-I}}{N_i - N_{i-I}}$$

où dans le cas qui nous interesse on a :

$x_i = 25 \qquad e_i = 35 - 25 = I0$

$N_i = 6057 \qquad N_{i-I} = 4457 \qquad$ et $\qquad N/2 = 5000$

On obtient en repportant toutes ces valeurs dans la formule ci-dessus :

$$M_h = 25 + 10\,\frac{5000 - 4457}{6057 - 4457} = 28,39\ .$$

2°-Pour l'âge moyen masculin et l'âge moyen feminin on cherche d'abord les séries des centres qu'on notera X et Y.Ces deux séries X et Y vont avoir les mêmes valeurs puisque elles proviennent des mêmes classes.Par contre elles n'auront pas les mêmes effectifs partiels.

Classes	effectifs masculins	effectifs feminins	X	Y	$X^2 = Y^2$
0 à 5	1041	943	2,5	2,5	6,25
5 à 10	948	825	7,5	7,5	56,25
10 à 15	903	833	12,5	12,5	156,25
15 à 25	1565	1484	20	20	400
25 à 35	1600	1551	30	30	900
35 à 45	1301	1351	40	40	1600
45 à 55	1083	1151	50	50	2500
55 à 65	855	926	60	60	3600
65 à 75	704	936	70	70	4900

On aura alors :

$\bar{X} = 30,70$ et $\bar{Y} = 32,78$.

Le lecteur est invité à faire le calcul en détail.

3°-On sait que :

$$Var(X) = \overline{X^2} - \bar{X}^2 \quad \text{et} \quad Var(Y) = \overline{Y^2} - \bar{Y}^2\ .$$

En utilisant le tableau ci-dessus on trouve:

$\overline{X^2} = 1358,36$ et $\overline{Y^2} = 1513,10$ donc

$$Var(X) = I358,36 - (30,70)^2 = 4I5,87$$

$$Var(Y) = 1513,10 - (32,78)^2 = 438,57$$

On en deduit :

$$\sigma(X) = \sqrt{4I5,87} = 20,39$$

$$\sigma(Y) = \sqrt{438,57} = 20,94 .$$

On remarque qu'en moyenne la population feminine est un peu plus vieille que la population masculine. Mais les dispersions sont pratiquement les mêmes.

CHAPITRE IV

LES INDICES

On a besoin souvent de mesurer les variations dans le temps d'une ou plusieurs variables statistiques.Par exemple si le prix de l'essence est :

- P_o = 2,80 F à l'époque 0,
- P_I = 4,70 F à l'époque I,

on pose :

$$P_{I/O} = \frac{P_I}{P_o} = \frac{4,70}{2,80} = I,67 .$$

Ce rapport nous dit que le prix de l'essence à l'époque I est I,67 fois plus élevé qu'à l'époque 0.Il mesure donc les variations du prix de l'essence par rapport à l'époque 0 qu'on appelle époque de reférence.
Bien-sûr on peut faire la même chose en prenant l'époque I comme époque de reférence.Dans ce cas on obtient :

$$P_{O/I} = \frac{P_o}{P_I} = \frac{2,80}{4,70} = 0,59 .$$

Le nombre $P_{I/O}$ est appelé indice élémentaire .

D'une manière générale on considère une grandeur G qui varie dans le temps.On appelle alors indice élémentaire de G à l'époque I par rapport à l'époque 0 le nombre :

$$\boxed{G_{I/O} = \frac{G_I}{G_O}} \qquad (0.I)$$

Nous nous limiterons dans ce chapitre à l'étude des indices des prix et des quantités et qui permettent de mesurer l'évolution du coût de certains biens de consommation.

I. INDICES ELEMENTAIRES

I.I-Indices élémentaires des prix.

Considèrons un article et supposons que son prix est :

-P_0 à l'époque 0,

-P_I à l'époque I.

On pose :

$$P_{I/0} = \frac{P_I}{P_0} \qquad (I.I)$$

Le nombre $P_{I/0}$ est appelé l'indice élémentaire du prix de cet article à l'époque I par rapport à l'époque 0.

On definit de façon analogue :

$$P_{0/I} = \frac{P_0}{P_I} \qquad (I.2)$$

Cette fois-ci on a choisi l'époque I comme époque de reférence.On verifie facilement la relation :

$$P_{I/0} \cdot P_{0/I} = I \qquad (I.3)$$

Généralement on exprime cet indice en pourcentage.

I.I.I-Exemple. Le litre de lait coûtait :

-P_0 = 0,38 F en I950 (époque 0)

-P_I = 0,4I F en I958 (époque I).

On a alors :

$$P_{I/0} = \frac{\text{Prix du lait en I958}}{\text{Prix du lait en I950}} = \frac{P_I}{P_0} = \frac{0,4I}{0,38} = I,08$$ .Ce qui correspond à I08% .

I.I.2-Remarques

i)L'indice élémentaire du prix à une époque par rapport à elle-même est égal à I.En effet on a :

$$P_{0/0} = \frac{P_0}{P_0} = I \text{ ou } I00\% . \qquad (I.4)$$

ii)En considérant cette fois-ci les prix P_o,P_I et P_2 d'un article respectivement aux époques 0 , I et 2 on a :

$$P_{2/0} = \frac{P_2}{P_o} = \frac{P_I \cdot P_2}{P_I \cdot P_o} = \frac{P_I}{P_o} \cdot \frac{P_2}{P_I} = P_{I/0} \cdot P_{2/I} \quad . \qquad (I.5)$$

I.I.3-Exemple. Un article coûte :

- P_0 = I80 F en I973 (époque 0)
- P_I = 220 F en I975 (époque I)
- P_2 = 300 F en I980 (époque 2).

On a :

$$P_{I/0} = \frac{220}{I80} = I,22 \quad (I22\%) \quad \text{et} \quad P_{2/I} = \frac{300}{220} = I,36 \quad (I36\%) .$$

D'où :

$$P_{2/0} = P_{I/0} \cdot P_{2/I} = I,22.I,36 = I,65 \quad (I65\%) .$$

N.B

Le lecteur pourra aisément constater que pour multiplier les indices il évitera de prendre les expressions en pourcentage.

I.I.4-Exercice.

Le prix d'un article était P_I = 270 F en I982 et l'indice élémentaire du prix de cet article en I982 par rapport à I975 est I70%. Calculer son prix P_0 en I975.

I.2-Indice élémentaire des quantités.

On definit de manière analogue que ce qui précède l'indice élémentaire des quantités.

Si Q_0 designe la quantité d'un produit à l'époque de reférence 0 et Q_I à l'époque I on pose :

$$Q_{I/0} = \frac{Q_I}{Q_0} \qquad (I.6)$$

On vérifie facilement que cet indice possède les mêmes propriétés que l'indice élémentaire des prix.

Quand on veut voir comment évolue le coût de la vie dans le temps on calcule son indice.Pour cela on choisit plusieurs biens de consommation par exemple le pain,la viande,le lait,le sucre,les vêtements,le ticket de bus etc...

Dans ce paragraphe nous donnerons quelques exemples d'indices permettant de mesurer cette variation du coût de la vie.

On supposera qu'on a N articles - qu'on numérotera de I jusqu'à N - dont :

i)Les prix sont :

$P_0^I,\ldots,P_0^N$ à l'époque 0

$P_I^I,\ldots,P_I^N$ à l'époque I

ii)Les quantités sont :

$Q_0^I,\ldots,Q_0^N$ à l'époque 0

$Q_I^I,\ldots,Q_I^N$ à l'époque I

Le numero du bas indique l'époque.Celui du haut indique l'article.

2.INDICES DU COUT DE LA VIE.

2.I-Indice global des prix.

On definit l'indice global des prix de ces articles à l'époque I par rapport à l'époque 0 comme étant le rapport :

$$\boxed{I_{I/0} = \frac{\sum P_I^i}{\sum P_0^i}} \qquad (I.7) \quad .$$

Tout le long de ce paragraphe la sommation $\sum$ porte sur i =I ,...,N.

2.I.I-Exemple. Considèrons les prix des 4 produits suivants à 2 époques différentes :

Année	Viande Kg	Oeufs I2	Beurre 250 g	Vin Litre
I959	6	2,20	I,80	I
I962	I0	3	2,50	I,30

On calcule l'indice global des prix :

$$I_{I/0} = \frac{I0 + 3 + 2,50 + I,30}{6 + 2,20 + I,80 + I} = \frac{I6,80}{II} = I,52 \qquad (I52\%) \ .$$

2.2-Moyenne arithmétique des indices.

On calcule l'indice élémentaire $P^i_{I/0} = \frac{P^i_I}{P^i_0}$ du prix de chaque article et on fait leur moyenne arithmétique :

$$\overline{I}_{I/0} = \frac{I}{N} \sum P^i_{I/0} \qquad (I.8) \ .$$

Reprenons l'exemple 2.I.I et calculons l'indice élémentaire du prix de chacun des produits considérés :

$$P^I_{I/0} = \frac{I0}{6} = I,66 \quad (I66\%) \qquad P^2_{I/0} = \frac{3}{2,20} = I,36 \quad (I36\%)$$

$$P^3_{I/0} = \frac{2,50}{I,80} = I,38 \quad (I38\%) \qquad P^4_{I/0} = \frac{I,30}{I} = I,30 \quad (I30\%)$$

Ce qui donne en appliquant la formule (I.8) :

$$\overline{I}_{I/0} = \frac{I}{4}(I,66 + I,36 + I,38 + I,30) = I,42 \quad (I42\%) .$$

L'indice global des prix et la moyenne arithmétique des indices élémentaires presentent un inconvénient majeur : Ils ne tiennent pas compte de l'importance de chaque bien de consommation par rapport aux autres.Pour cela on introduit d'autres indices donnant des informations plus précises que celles que l'on peut obtenir à l'aide des indices que l'on vient de voir.

2.3-Indice pondéré.

Pour faire apparaître l'importance de chaque bien de consommation par rapport aux autres on pondère par sa quantité.On obtient alors les indices suivants :

i)Indice de Laspeyres qui se calcule en pondérant les prix par les quantités à l'époque de reférence :

$$I^{L}_{I/0} = \frac{\sum Q^{i}_{0} P^{i}_{I}}{\sum Q^{i}_{0} P^{i}_{0}} \qquad (I.9)$$

ii)Indice de Paasche.Il se calcule en pondérant les prix par les quantités à l'époque I :

$$I^{P}_{I/0} = \frac{\sum Q^{i}_{I} P^{i}_{I}}{\sum Q^{i}_{I} P^{i}_{0}} \qquad (I.I0)$$

Mais on peut très bien pondérer par la moyenne arithmétique de la quantité à l'époque 0 et la quantité à l'époque I.Dans ce cas on a :

iii)Indice de Marshall.

$$I^{M}_{I/O} = \frac{\sum (Q^{i}_{O} + Q^{i}_{I})P^{i}_{I}}{\sum (Q^{i}_{O} + Q^{i}_{I})P^{i}_{O}} \qquad (I.II) \ .$$

En faisant la moyenne géométrique des indices de Laspeyres et de Paasche on obtient un nouvel indice connu sous le nom de :

iv)Indice de Fisher.

$$I^{F}_{I/O} = \sqrt{I^{L}_{I/O} \cdot I^{P}_{I/O}} \qquad \text{i.e}$$

$$I^{F}_{I/O} = \sqrt{\frac{\sum Q^{i}_{O}P^{i}_{I}}{\sum Q^{i}_{O}P^{i}_{O}} \cdot \frac{\sum Q^{i}_{I}P^{i}_{I}}{\sum Q^{i}_{I}P^{i}_{O}}} \qquad (I.I2) \ .$$

Comme pour les indices élémentaires,l'indice de Fisher verifie la propriété de réversibilité du temps,c'est à dire :

$$I^{F}_{I/O} \cdot I^{F}_{O/I} = I \qquad (I.I3) \ .$$

3.EXERCICE RESOLU.

Le tableau suivant donne les prix et les quantités de 3 produits à 2 époques différentes : Tab 3.I

Produit (i)	I972 (époque O)		I980 (époque I)	
	Prix P^{i}_{O}	Quantités Q^{i}_{O}	P^{i}_{I}	Q^{i}_{I}
I	I2	6	I5	7
2	5	I3	8	II
3	I5	9	I3	I8

I°.Calculer l'indice global rendant compte de l'évolution de ces biens de consommation.

2°.Calculer les indices de :

-Laspeyres,

-Paasche,

-Marshall.

En deduire l'indice de Fisher.

On prendra I972 comme époque de reférence.

Solution.

I°.On calcule la somme $\sum P_0^i$ où i varie de I à 3.On obtient :

$$\sum P_0^i = I2 + 5 + I5 = 32$$

De même on a :

$$\sum P_I^i = I5 + 8 + I3 = 36$$

D'où, en appliquant la formule (I.7) :

$$I_{I/0} = \frac{36}{32} = I,I25 \quad (II2,5\%) .$$

2°.Pour traiter cette question on aura besoin de calculer les différentes expressions qui interviennent dans les formules (I.9) , (I.IO) et (I.II). On résumera tous ces calculs dans le tableau suivant :

Produit (i)	$Q_0^i P_0^i$	$Q_0^i P_I^i$	$Q_I^i P_0^i$	$Q_I^i P_I^i$	$(Q_0^i + Q_I^i) P_0^i$	$(Q_0^i + Q_I^i) P_I^i$
I	72	90	84	I05	I56	I95
2	65	I04	55	88	I20	I92
3	I35	II7	270	234	405	35I
Total	272	3II	409	427	68I	738

Tab 3.2

Ceci nous donne :

$$\text{-Indice de Laspeyres} = I^L_{I/0} = \frac{3II}{272} = I,I4 \quad (II4\%) .$$

$$\text{-Indice de Paasche} = I^P_{I/0} = \frac{427}{409} = I,04 \quad (I04\%) .$$

$$\text{-Indice de Marshall} = I^M_{I/0} = \frac{738}{68I} = I,08 \quad (I08\%) .$$

Par definition l'indice de Fisher est la moyenne géométrique des indices de Laspeyres et de Paasche,on obtient en appliquant la formule (I.I2) :

$$\text{-Indice de Fisher} = I^F_{I/0} = \sqrt{I,I4 . I,04} = I,09 \quad (I09\%) .$$

CHAPITRE V

LA CONCENTRATION

I. GENERALITES ET DEFINITIONS.

On considère une série statistique classée :

$$(I)\qquad \begin{array}{cccc} [a_1, a_2[& \cdots & [a_k, a_{k+1}[& \text{(classes)} \\ n_1 & \cdots & n_k & \text{(effectifs)} \end{array} \qquad \text{avec } a_i \geq 0 \text{ pour tout } i=1,\ldots,k+1$$

On note :

$x_i = \dfrac{a_i + a_{i+1}}{2}$ le centre de la classe $[a_i, a_{i+1}[$.

On suppose que le caractère mesuré par cette série est additif, c'est-à-dire que la somme de deux valeurs a un sens. Par exemple

-on peut additionner le revenu, le salaire, etc...

-on ne peut pas additionner l'âge, la taille, etc...

Pour chaque classe on peut considérer :

-son effectif partiel,

-l'importance du caractère (notion que l'on va preciser).

A ces deux points de vue on peut associer deux representations graphiques :

-l'histogramme des effectifs cumulés qu'on a dejà étudié,

-un histogramme qui décrit l'importance du caractère cumulé de chaque classe.

Ces deux representations nous permettent de calculer:

-la médiane (valeur qui partage la série en deux series ayant toutes les deux le même effectif total),

-une valeur qui partage la série en deux séries dont le caractère a la même importance aussi bien pour l'une que pour l'autre.

Cette dernière valeur est appelée la médiale.Son écart à la médiane est parfois utilisé pour faire l'analyse de la concentration.

I.I-Caractère.

On reprend la série classée (I).

I.I.I-Definition.

On appelle caractère de la classe $[a_i, a_{i+1}[$ le nombre $n_i x_i$. On peut dresser un tableau donnant le caractère de chaque classe ainsi que le caractère cumulé :

Classes	centre x_i	effectif n_i	caractère $n_i x_i$	caractère cumulé
$[a_1, a_2[$	x_1	n_1	$n_1 x_1$	$n_1 x_1$
$[a_2, a_3[$	x_2	n_2	$n_2 x_2$	$n_1 x_1 + n_2 x_2$
...	...	...	...	...
$[a_k, a_{k+1}[$	x_k	n_k	$n_k x_k$	$\sum_{i=1}^{k} n_i x_i$

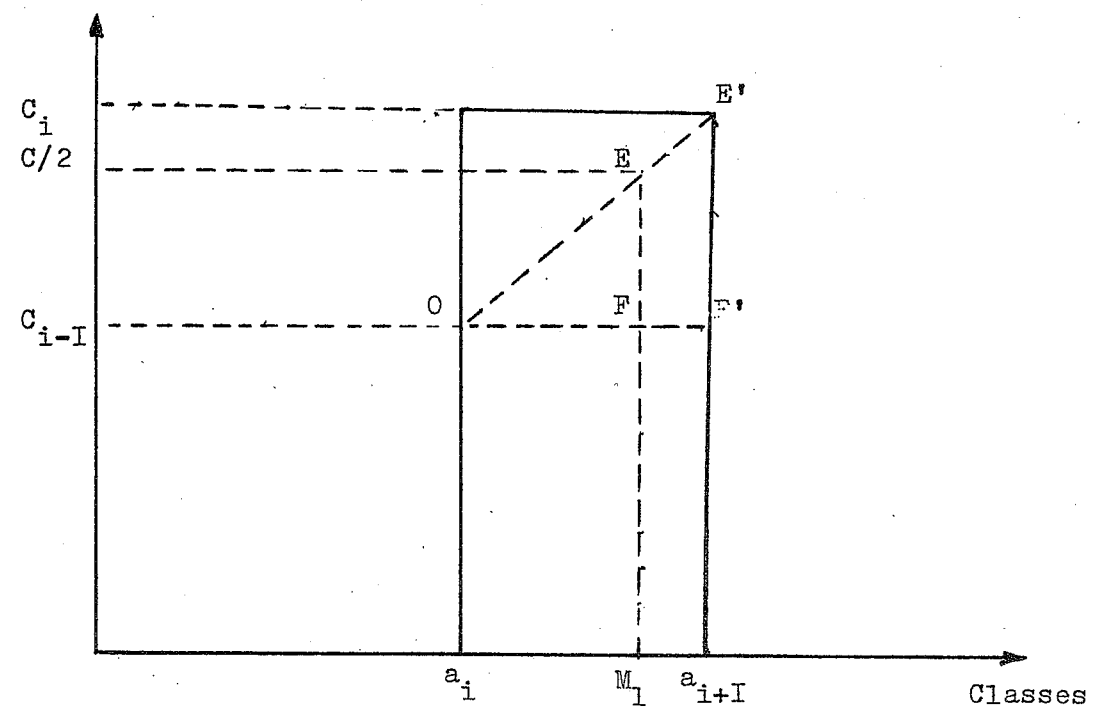
Tab I.I

I.2-Calcul de la médiale.

On calcule le caractère cumulé de chaque classe (voir Tab I.I) et on note C le caractère total.La médiale est alors la valeur qui partage la série en deux séries dont chacune a un caractère total égal à C/2.La médiale tombe dans une classe $[a_i, a_{i+1}[$.On procède par interpolation proportionnelle comme pour la médiane.

Caractère cumulé

Fig I.I



Pour tout i=I,...,k on note C_i le caractère cumulé de la classe $[a_i , a_{i+I}[$,c'est à dire :

$$C_i = \sum_{s=I}^{i} n_s x_s \qquad (I.I) \ .$$

On a alors les relations de proportionnalité :

$$\frac{OF}{OF'} = \frac{EF}{E'F'} \qquad \text{i.e} \quad \frac{M_1 - a_i}{a_{i+I} - a_i} = \frac{C/2 - C_{i-I}}{C_i - C_{i-I}}$$

D'où l'on deduit :

$$\boxed{M_1 = a_i + (a_{i+I} - a_i) \frac{C/2 - C_{i-I}}{C_i - C_{i-I}}} \qquad (I.2) \ .$$

I.3-Un exemple de calcul de la médiale.

On considère la statistique suivante donnant la production mondiale de blé en I982.(Source :Perspectives de l'alimentation, 26 mai F.A.O.,Rome).

Production en millions de tonnes	Nombre de pays
0 à I0	I9
I0 à 20	4
20 à 40	2
40 à 60	I
60 à 80	2

Tab I.2

Calculons le centre,le caractère et le caractère cumulé de chaque classe de cette série :

Classe	centre x_i	effectif n_i	caractère $n_i x_i$	caractère cumulé C_i
$[0,I0[$	5	I9	95	95
$[I0,20[$	I5	4	60	I55
$[20,40[$	30	2	60	2I5
$[40,60[$	50	I	50	265
$[60,80[$	70	2	I40	405

Tab I.3

Le caractère total est C = 405.Donc C/2 = 202,5.La médiale tombe dans la classe $[20\ ,\ 40[$ .On applique alors la formule (I.2) avec dans le cas qui nous interesse :

$a_i = 20$ $\qquad$ $a_{i+I} = 40$

$C_i = 215$ $\qquad$ $C_{i-I} = 155$.

On obtient :

$$M_1 = 20 + (40 - 20).\frac{202,5 - 155}{215 - 155} \quad \text{et finalement :}$$

$$\boxed{M_1 = 35,83}.$$

2.FONCTION ET COURBE DE CONCENTRATION.

On reprend les notations du paragraphe I ,c'est-à-dire on considère une série classée :

$[a_I, a_2[$ ... $[a_k, a_{k+I}[$ (Classes)

n_I ... n_k (Effectifs)

avec $N = n_I + \ldots + n_k$ et $x_i = \dfrac{a_i + a_{i+I}}{2}$ pour i=I,...,k .

On note f_i et F_i respectivement la fréquence et la fréquence cumulée de la classe $[a_i, a_{i+I}[$ i.e :

$$f_i = \frac{n_i}{N} \quad \text{et} \quad F_i = f_I + \ldots + f_i \qquad \text{pour tout } i = I,\ldots,k .$$

2.I.I-Definition.

La fonction de concentration est l'application qui à tout x_i associe le nombre :

$$\frac{\sum_{s=I}^{i} n_s x_s}{C}$$

2.I.2-Exemple.

Determinons la fonction de concentration dans le cas de l'exemple I.3.En utilisant le tableau I.3,on calcule la concentration pour tout x_i .On obtient :

Centre x_i	Fréquence cumulée F_i	caractère $n_i x_i$	caractère cumulé	concentration
5	0,67	95	95	0,23 (23%)
I5	0,82	60	I55	0,38 (38%)
30	0,89	60	2I5	0,53 (53%)
50	0,92	50	265	0,65 (65%)
70	I	I40	405	I (I00%)

Tab 2.I

2.2-Courbe de concentration.

2.2.I-Construction.

Considèrons la correspondance qui à la fréquence cumulée F_i de la classe $[a_i, a_{i+I}[$ associe la concentration :

$$\frac{\sum_{s=I}^{i} n_s x_s}{C}$$

et portons dans un repère orthonormé F_i en abscisse et $\sum_{s=I}^{i} n_s x_s / C$ en ordonnée.On obtient un ensemble de points à l'interieur du carré OABD de côté égal à I(voir Fig 2.I).

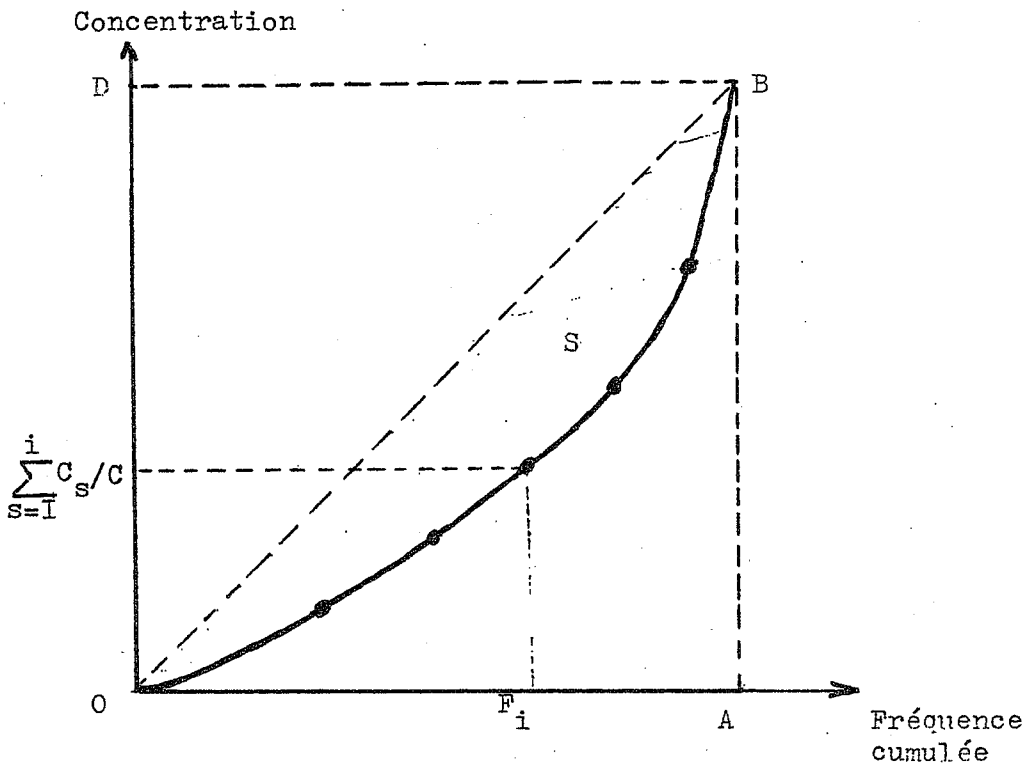


Fig 2.I

Si le nombre de classes est assez grand, ces points seront assez nombreux et on peut considérer qu'ils determinent pratiquement une courbe. On l'appelle la courbe de Lorentz

Le rapport de l'aire de la region du carré notée S à celle du triangle OAB est appelé Indice de concentration ou indice de Gini. On le note :

$$\nu = \frac{\text{Aire de S}}{\text{Aire de OAB}} \qquad (2.I) .$$

(Le caractère "ν" se lit "nu").

2.2.2-Remarques.

i)L'aire de S étant toujours inferieure ou égale à celle du triangle OAB,l'indice est donc un nombre compris entre 0 et I.

ii)Si $\mathcal{J} = 0$,l'aire de S est nulle et donc la courbe de Lorentz se confond avec le segment OB.Ceci signifie qu'un point quelconque sur cette courbe a son abscisse égale à son ordonnée.On en deduit donc que pour tout i=I,...,k on a :

$$F_i = \frac{\sum_{s=I}^{i} n_s x_s}{C} \qquad (2.2)$$

Or $F_i = f_I + \ldots + f_i$ et $\frac{\sum_{s=I}^{i} n_s x_s}{C} = \frac{n_I x_I + \ldots + n_i x_i}{C}$

On divise le numérateur et le dénominateur par $N = n_I + \ldots + n_k$.
On obtient alors en posant $A = \frac{C}{N}$:

$$\frac{\sum_{s=I}^{i} n_s x_s}{C} = \frac{\frac{n_I}{N} x_I + \ldots + \frac{n_i}{N} x_i}{A} = \frac{f_I x_I + \ldots + f_i x_i}{A}$$

La relation (2.2) devient alors :

$$f_I + \ldots + f_i = \frac{f_I x_I + \ldots + f_i x_i}{A} \quad \text{pour tout } i = I,\ldots,k ,$$

c'est-à-dire :

$$f_I = \frac{f_I x_I}{A} \qquad f_I + f_2 = \frac{f_I x_I + f_2 x_2}{A} \qquad \text{etc...}$$

La $I^{ère}$ égalité donne $f_I A = f_I x_I$ et en simplifiant par f_I on obtient $A = x_I$ .On remplace x_I par A dans la $2^{ème}$ égalité.Cette dernière devient alors :

$$f_I + f_2 = \frac{f_I A + f_2 x_2}{A} = \frac{f_I A}{A} + \frac{f_2 x_2}{A} = f_I + \frac{f_2 x_2}{A} \quad .$$

D'où : $f_2 = \frac{f_2 x_2}{A}$ .Ceci implique $f_2 A = f_2 x_2$ et donc $A = x_2$.

De la même manière on montre que $x_3 = x_4 = \ldots = x_k = A$.

On peut donc énoncer la proposition :

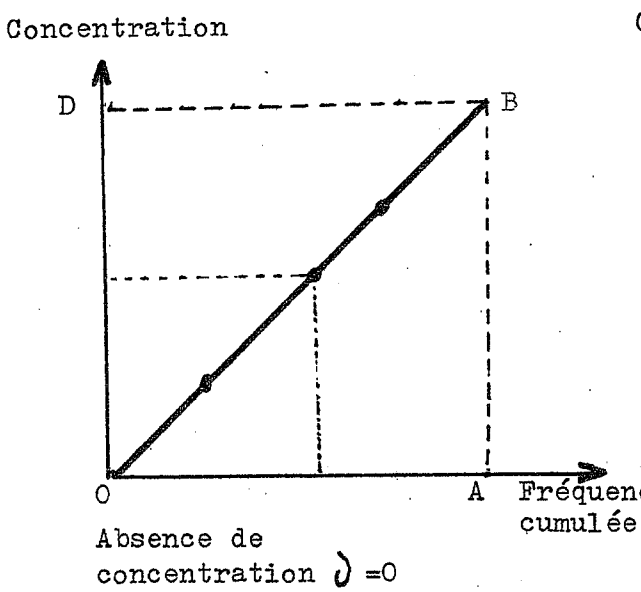
-Si l'indice de concentration est nul tous les termes de la série des centres sont égaux .

Dans ce cas on dit qu'il n'y a pas de concentration.

iii)Si $\mathcal{I} = I$ on dit qu'il y a concentration totale. Dans ce cas la courbe de Lorentz est soit le segment AB soit le segment OA. La série des centres $x_I, x_2, \ldots, x_k$ étant croissante,il en est de même de la série $F_I, F_2, \ldots, F_k$ des fréquences cumulées.On en deduit que pour F_i donné le nombre :

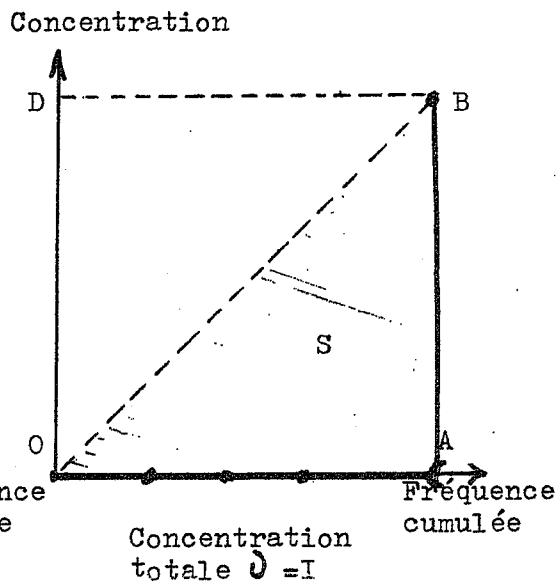
$$\frac{\sum_{s=I}^{i} n_s x_s}{C}$$ (qui n'est rien d'autre que $\frac{f_I x_I + \ldots + f_i x_i}{A}$ via la remarque ii)) est unique,et donc sur la courbe de Lorentz il n'y a qu'un seul point pour lequel l'abscisse soit égale à I,et c'est le point B.Le premier cas ne peut donc se presenter.La courbe de Lorentz est confondue avec le segment OA.Ce n'est donc pas une courbe continue puisque le point B n'est pas sur OA. Toutefois on peut considerer que la courbe de Lorentz est constituée par le segment OA et le segment AB.

Fig 2.2



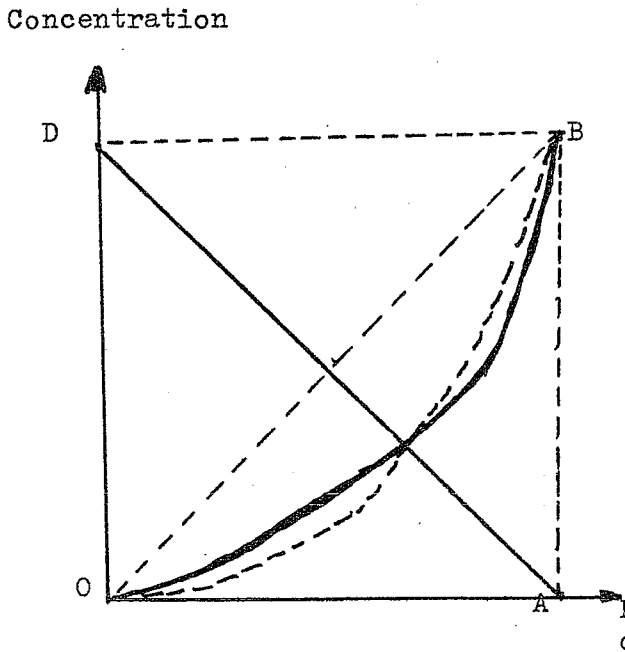
Absence de concentration $\mathcal{J} = 0$

Fig 2.3



Concentration totale $\mathcal{J} = I$

iv)Deux séries statistiques peuvent avoir le même indice de concentration.La Fig 2.4 donne un exemple de deux courbes de Lorentz qui délimitent avec le segment OB deux portions du carré OABD ayant des aires égales.



Ces deux courbes sont symétriques par rapport à l'axe AD.

Fig 2.4

2.2.3-Calcul de l'indice de Gini.

Considèrons toujours une série statistique de type (I).On a la série des fréquences cumulées $F_I, F_2, \ldots, F_k$;on lui rajoute une"fréquence" $F_o = 0$ pour la commodité du calcul.On pose :

$h_o = 0$ et $h_i = \dfrac{\sum_{s=I}^{i} n_s x_s}{C}$ la concentration de la valeur x_i .

On a vu que les points P_i dont les abscisses (respectivement les ordonnées) sont F_i (resp h_i) permettent de construire la courbe de Lorentz de cette série.

Pour calculer l'indice de Gini on doit calculer l'aire de S.Pour cela il suffit de calculer celle de S' (le complementaire de S dans le triangle OAB) .Si on construit la courbe de Lorentz en joignant les points P_i par des segments de droite,on voit que l'aire de S' est la somme des aires des trapèzes $P_i P_{i+I} P'_{i+I} P'_i$ .Or l'aire d'un trapèze de ce type est :

$$A_i = \frac{(P'_i P_i + P'_{i+I} P_{i+I}) . P'_i P'_{i+I}}{2} \qquad \text{c'est-à-dire}$$

$$A_i = \frac{(h_i + h_{i+I})(F_{i+I} - F_i)}{2} \quad . \text{Mais} \quad F_{i+I} - F_i = f_{i+I}$$

D'où :

$$\text{Aire de S} = 0{,}5 - \text{Aire de S'} = \frac{I}{2} - \sum_{i=0}^{k-I} A_i$$

et finalement :

$$\boxed{J = I - \sum_{i=0}^{k-I} (h_i + h_{i+I}) f_{i+I}} \qquad (2.3) .$$

Cette relation permet donc de calculer l'indice de Gini.

Concentration

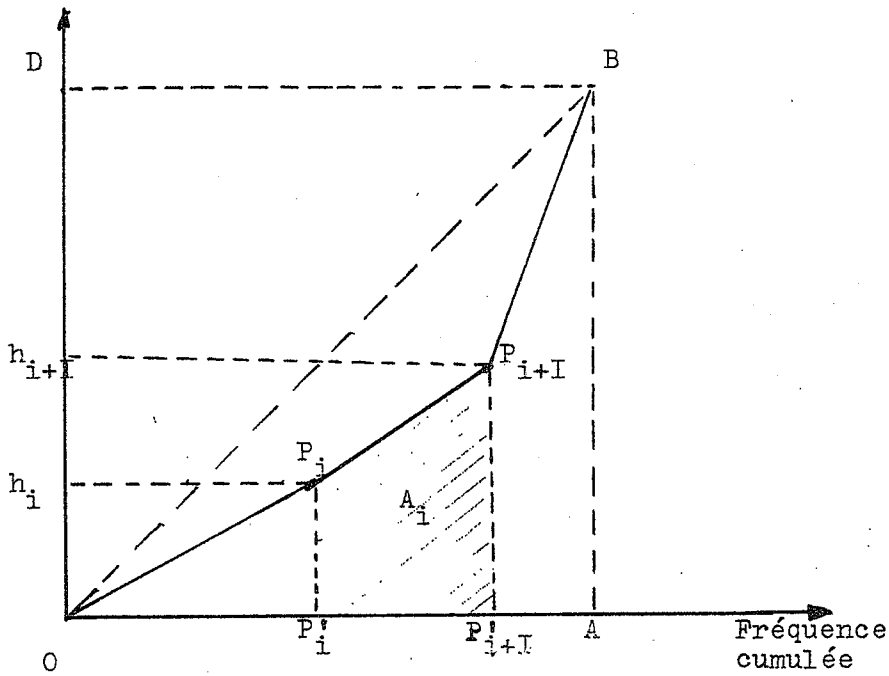


Fig 2.5

Dans la partie qui suit nous donnerons quelques exemples sur lesquels on fera l'analyse de la concentration en calculant les differents paramètres : Médiale ,fonction de concentration ,courbe de Lorentz et indice de Gini etc...

3.EXEMPLES.

3.I-Exemple où la concentration est presque totale.

Une famille constituée de IO personnes (4 filles,4 garçons, la mère et le père) s'est partagée une somme de I0000F d'une manière très inégale :

-50F pour chaque fille,

-I00F pour chaque garçon,

-400F pour la mère,

-9000F pour le père.

On peut considèrer qu'on a une série à valeurs isolées :

$x_I=50$ $x_2=I00$ $x_3=400$ $x_4=9000$ (Valeurs)

$n_I=4$ $n_2=4$ $n_3=I$ $n_4=I$ (Effectifs)

Calculons d'abord les différentes valeurs dont on aura besoin. On dresse le tableau suivant :

x_i	n_i	f_i	F_i	$f_i x_i$	$\sum_{s=I}^{i} f_s x_s$	Concentration notée h_i
50	4	0,4	0,4	20	20	0,02
I00	4	0,4	0,8	40	60	0,06
400	I	0,I	0,9	40	I00	0,I
9000	I	0,I	I	900	I000	I

Tab 3.I

i)Courbe de Lorentz

On porte les F_i en abscisses et les h_i en ordonnées.On obtient alors la courbe suivante :

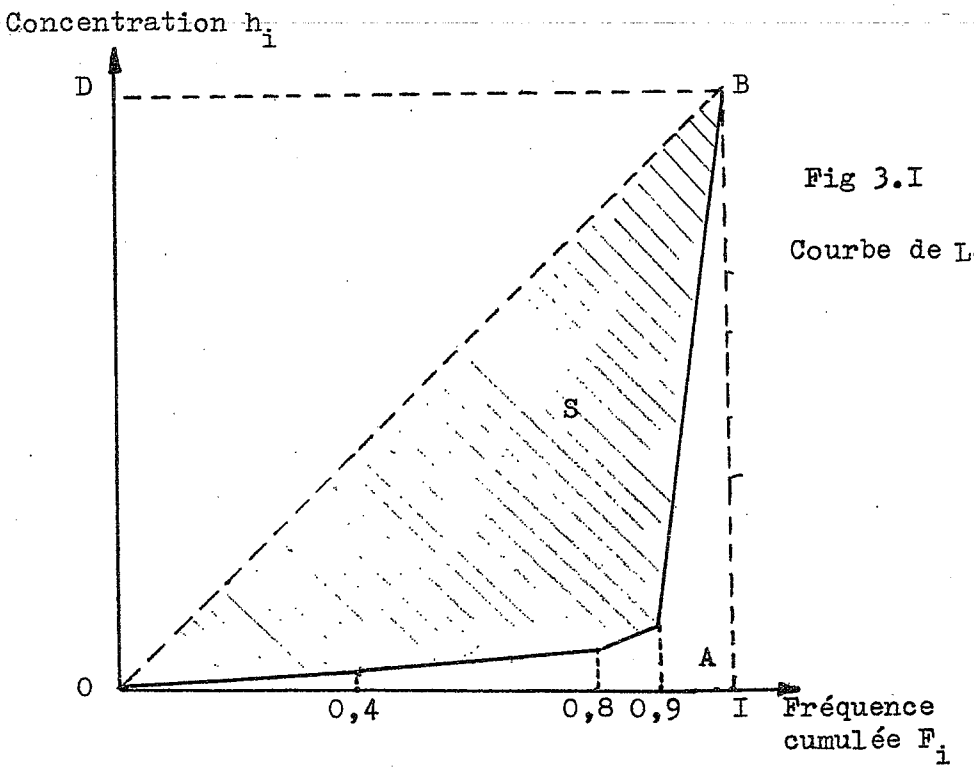


Fig 3.I

Courbe de Lorentz

ii)Indice de concentration de Gini.

Pour calculer cet indice on applique la formule (2.3) p.64.

On aura :

$$\mathcal{I} = I - (0,02.0,4 + 0,08.0,4 + 0,I6.0,I + I,I.0,I).$$

$$\mathcal{I} = 0,834.$$

L'indice $\mathcal{I}$ est voisin de I,la concentration est presque totale, chose à laquelle on peut s'attendre d'ailleurs puisque le père s'est accaparé pratiquement la totalité de la somme.

3.2-Autre exemple d'analyse de la concentration.

On reprend le tableau de l'exemple de la page.23 qui donne la répartition de la population de 27 pays européens.On le complète en calculant les différents paramètres qui permettent de faire l'analyse de la concentration.

Centre x_i	n_i	f_i	F_i	$n_i x_i$	C_i cumulé	Concentration h_i
2,5	7	0,26	0,26	I7,5	I7,5	0,036
7,5	7	0,26	0,52	52,5	70	0,I43
I2,5	3	0,II	0,63	37,5	I07,5	0,220
I7,5	2	0,07	0,70	35	I42,5	0,292
22,5	2	0,07	0,77	45	I87,5	0,384
32,5	2	0,07	0,84	65	252,5	0,5I7
47,5	I	0,05	0,89	47,5	300	0,6I5
62,5	3	0,II	I	I87,5	487,5	I

Tab 3.2

Dans cet exemple nous ferons l'analyse de la concentration de deux façons différentes :

-En calculant l'indice de Gini,

-en calculant l'écart entre la médiane et la médiale.

i)Calcul de la médiane.

L'effectif total est N=27 et donc N/2 =I3,5.La médiane tombe donc dans la classe $[5 , I0[$ .On applique alors la formule (I.I) p.I8 avec :

$a_i = 5$ (noté x_i dans la formule) $e_i = 5$; $N_i = I4$ et $N_{i-I} = 7$.

On obtient :

$$M_e = 5 + (I0 - 5).\frac{I3,5 - 7}{7}$$

$$M_e = 9,64 .$$

ii)Calcul de la médiale.

Le caractère total est C = 487,5.D'où C/2 = 243,5.De la même manière la médiale se calcule par la formule (I.2) p.56 avec :

$a_i = 25$. $a_{i+I} = 40$ $C_i = 252,5$ $C_{i-I} = I87,5$.

On a =

$$M_1 = 25 + (40 - 25).\frac{243,75 - I87,5}{65}$$

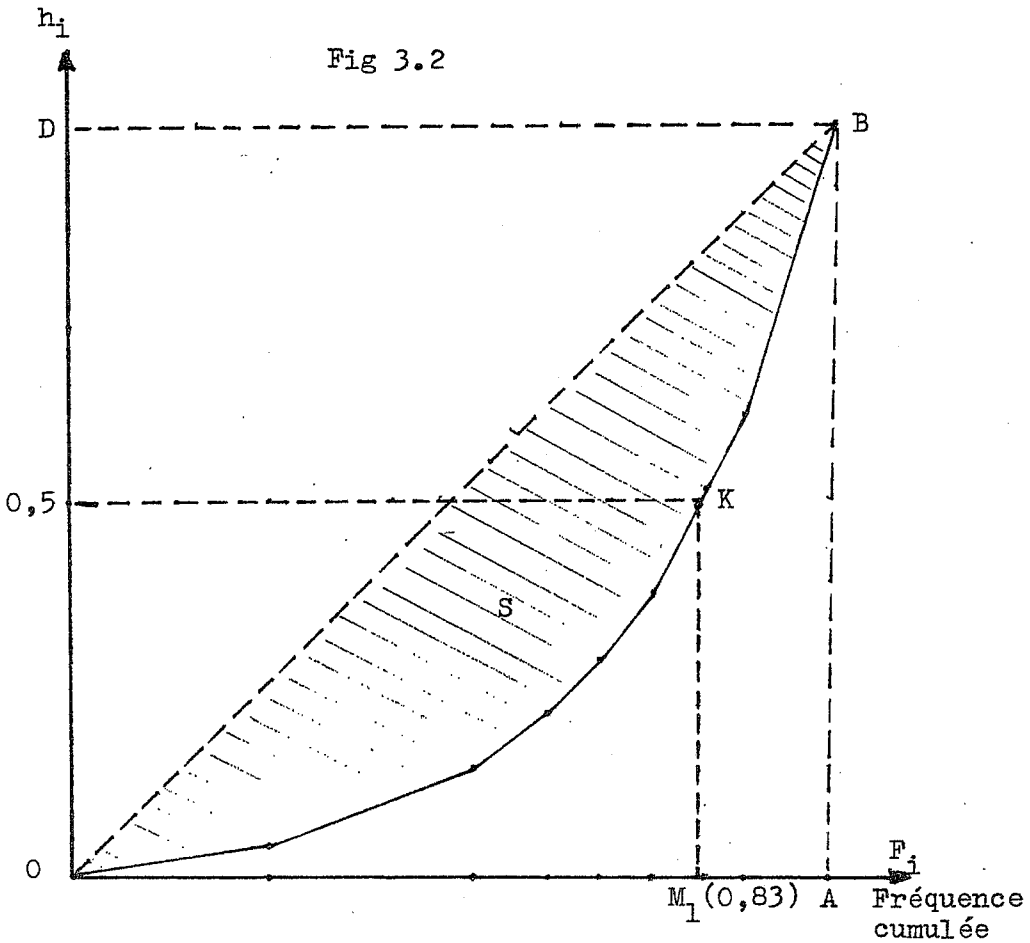
$M_1 = 37,98$. (la médiale tombe dans la classe $[25 , 40[$).

iii)Courbe de Lorentz.

On construit la courbe de Lorentz en portant les F_i en abscisses et les h_i en ordonnées (voir Tab 3.2).On obtient alors :

Concentration

Fig 3.2



iv)Indice de Gini.

On applique la formule (2.3) p.64 qui s'écrit dans le cas qui nous interesse ici :

$$\mathcal{I} = I - \left[h_I f_I + (h_I + h_2)f_2 + \ldots + (h_7 + h_8)f_8\right]$$

où :

h_I=0,036 h_2=0,I43 h_3=0,220 h_4=0,292

h_5=0,384 h_6=0,5I7 h_7=0,6I5 h_8=I et

f_I=0,26 f_2=0,26 f_3=0,II f_4=0,07

f_5=0,07 f_6=0,07 f_7=0,05 f_8=0,II .

On obtient finalement :

$$\mathcal{I} = I - 0,48 = 0,52 \quad (52\%) .$$

La concentration est donc moyenne.

v)Calcul de la médiale à l'aide de la courbe de Lorentz.

On peut remarquer que les concentrations sont obtenues en divisant les $\sum_{s=1}^{i} n_s x_s$ par le caractère total C.Il en resulte que les concentrations h_i sont respectivement proportionnelles à ces caractères cumulés.Donc la médiale sera"l'abscisse"du point de la courbe de Lorentz dont l'ordonnée est égale à 0,5.En fait elle sera donnée en "pourcentage".Il faudra donc reconvertir après. Pour le cas qui nous interesse on lit graphiquement l'abscisse du point K et on trouve 0,83.Or on sait que 0,77 correspond à 25 et 0,84 correspond à 40.En faisant une interpolation proportionnelle on trouve que 0,83 correspond à :

$$25 + 15.\frac{0,83 - 0,77}{0,84 - 0,77} = 37,85$$ .Ce qui est à peu près le resultat trouvé en ii).

vi)On peut aussi pour faire l'analyse de la concentration calculer l'écart entre la médiane et la médiale.On aura :

$$M_1 - M_e = 37,98 - 9,64 = 28,34 .$$

On voit donc que cet écart est relativement important "par rapport"à la plus grande valeur de la série.

vii)La courbe de concentration de Lorentz nous dit par exemple que 83% des pays européens (sans l'URSS) se partagent 50% de la population totale européenne (sans l'URSS évidemment).

CHAPITRE VI

SERIES DOUBLES

AJUSTEMENT CORRELATION

Il arrive quelquefois que deux séries statistiques mesurant deux caractères X et Y d'un échantillon d'une même population soient liées par une relation dans le sens que les valeurs de l'une peuvent être obtenues à partir de celles de l'autre à l'aide d'une correspondance.On dira que X et Y sont dépendantes.Nous essayerons dans ce chapitre de préciser le sens de cette "dependance" et de la calculer du moins approximativement.

Quitte à passer à la série des centres on pourra toujours supposer que les séries considérées sont à valeurs isolées.

I.SERIES DOUBLES.

I.I-Exemple.

On étudie le poids et la taille notés P et T de I00 élèves pris au hasard dans un lycée.On a obtenu les données suivantes :

P \ T	I50cm	I55	I60	I65	I70	
50 Kg	3	5	4	3	2	I7
55	5	7	7	8	7	34
60	6	6	4	6	8	30
65	4	5	3	2	5	I9
	I8	23	I8	I9	22	I00

Tab I.I

La case de la 3ème colonne et la 2ème ligne represente le nombre d'élèves qui ont un poids de 55 Kg et une taille de I60 cm c'est-à-dire 7.On note cette valeur n_{23}.C'est l'effectif partiel de la 2ème valeur du caractère poids et la 3ème valeur du caractère taille.

On peut s'interesser uniquement au caractère poids.Par exemple le nombre d'élèves ayant un poids égal à 55 Kg et une taille quelconque est 34.Ceci figure sur la dernière colonne.On note ce nombre $n_{2.}=34$.C'est l'effectif marginal de la 2ème valeur du caractère poids.Il s'obtient en faisant la somme de tous les termes de la 2ème ligne.On a : $n_{2.} = n_{2I} + n_{22} + \dots + n_{25}$.

De la même manière on peut se restreindre au caractère taille.La 2ème valeur 23 de la dernière ligne donne le nombre d'élèves mesurant I55 cm (avec un poids quelconque).De manière analogue $n_{.2}$ s'obtient en faisant la somme de tous les termes de la 2ème colonne:

$$n_{.2} = n_{I2} + n_{22} + \dots + n_{42} .$$

On dira que $n_{.2}$ est l'effectif marginal de la 2ème valeur du caractère taille.

I.2-Quelques définitions.

L'exemple que l'on vient de considérer est typique de ce qu'on appelle une série double ou à deux dimensions(parce qu'on étudie deux caractères sur un même échantillon).

I.2.I-Presentation.

On peut résumer tout dans un tableau semblable au Tab I.I :

X \ Y	y_I	...	y_j	...	y_1	Effectif marginal
x_I	n_{II}	...	n_{Ij}	...	n_{I1}	$n_{I.}$
...	...		...		...	...
x_i	n_{iI}	...	n_{ij}	...	n_{i1}	$n_{i.}$
...	...		...		...	...
x_k	n_{kI}	...	n_{kj}	...	n_{k1}	$n_{k.}$
Effectif marginal	$n_{.I}$	...	$n_{.j}$	...	$n_{.1}$	N

Tab I.2

où :

- $x_I,\ldots,x_k$ sont les valeurs du caractère X,

- $y_I,\ldots,y_1$ sont les valeurs du caractère Y

- n_{ij} est l'effectif partiel du couple (x_i,y_j) c'est-à-dire le nombre d'individus de l'échantillon pour lesquels les caractères X et Y valent respectivement x_i et y_j ,

- $n_{i.}$ (resp $n_{.j}$) l'effectif marginal de x_i (resp de y_j) i.e le nombre d'individus pour lesquels X (resp Y) vaut x_i (resp y_j) et qui est tel que :

$$n_{i.} = \sum_{j=I}^{1} n_{ij} \qquad (\text{resp } n_{.j} = \sum_{i=I}^{k} n_{ij}) \quad (I.I) ,$$

-et $$N = \sum_{i,j} n_{ij} = \sum_{i=I}^{k} n_{i.} = \sum_{j=I}^{1} n_{.j} \qquad (I.2)$$

est l'effectif total.

On gardera les notations de cette partie tout le long de ce chapitre.

I.2.2-Fréquences partielles et marginales.

On definit la fréquence partielle du couple (x_i, y_j) et la fréquence marginale de la valeur x_i (resp y_j) par les égalités :

$$f_{ij} = \frac{n_{ij}}{N} \quad \text{fréquence partielle} \qquad (I.3.I) ,$$

$$f_{i.} = \frac{n_{i.}}{N} \quad (\text{resp } f_{.j} = \frac{n_{.j}}{N}) \qquad (I.3.2) .$$

On vérifie aisément les relations :

$$f_{i.} = \sum_{j=I}^{l} f_{ij} \quad \text{et} \quad f_{.j} = \sum_{i=I}^{k} f_{ij} \qquad (I.4.I)$$

$$\sum_{i,j} f_{ij} = I \qquad (I.4.2) .$$

Calculons par exemple les fréquences partielles et marginales des couples de valeurs de la série double donnée par le Tab (I.I) :

T / P	I50	I55	I60	I65	I70	$f_{.j}$
50	0,03	0,05	0,04	0,03	0,02	0,I7
55	0,05	0,07	0,07	0,08	0,07	0,34
60	0,06	0,06	0,04	0,06	0,08	0,30
65	0,04	0,05	0,03	0,02	0,05	0,I9
$f_{i.}$	0,I8	0,23	0,I8	0,I9	0,22	I

Tab I.2

I.2.3-Effectifs et fréquences cumulés.

On appelle effectif cumulé du couple de valeurs (x_i,y_j) le nombre d'individus N_{ij} de l'échantillon pour lesquels le caractère X est inférieur ou égal à x_i et le caractère Y est inférieur ou égal à y_j.On l'exprime en fonction des effectifs partiels par la relation :

$$(I.5)N_{ij} = \sum_{s\leq i,\, t\leq j} n_{st} = (n_{II}+\ldots+n_{Ij})+(n_{2I}+\ldots+n_{2j})+\ldots+(n_{iI}+\ldots n_{ij}).$$

La fréquence partielle cumulée de (x_i,y_j) est définie comme suit :

$$F_{ij} = \frac{N_{ij}}{N} \qquad (I.6)\ .$$

Si $i = k$ (resp $j = 1$) on a :

$$N_{ij} = \sum_{t=I}^{j} n_{.t} \qquad (\text{resp } N_{ij} = \sum_{s=I}^{i} n_{s.})$$

et dans ce cas on parlera d'effectif marginal cumulé de y_j (resp de x_i).On le notera alors $N_{.j}$ (resp $N_{i.}$).

De même on a la fréquence cumulée d'une valeur x_i ou y_j qui se calcule de manière analogue.On se contentera de le faire sur l'exemple qui suit .

I.2.4-Considèrons le Tab I.2.Par exemple :

-La fréquence cumulée F_{22} de (55,I55) : On fait la somme des fréquences de tous les couples (P_i,T_j) pour lesquels $P_i \leqslant 55$ et $T_j \leqslant$ I55 .Ces couples sont (50,I50),(50,I55),(55,I50) et (55,I55).

Ce qui donne :

$$F_{22} = 0,03 + 0,05 + 0,05 + 0,07 = 0,2 .$$

-La fréquence marginale cumulée de la valeur $P_3 = 60$ est la somme des fréquences de tous les couples (P_i, T_j) pour lesquels $P_i \leqslant 60$, c'est-à-dire :

$$F_{3.} = 0,17 + 0,34 + 0,30 = 0,81 .$$

2. MOYENNES MARGINALES.

2.1-Definition.

Les moyennes marginales $\bar{X}$ et $\bar{Y}$ sont les moyennes arithmétiques de X et Y respectivement. Elles se calculent en fonction des valeurs de chacune des séries X et Y mais en considérant les effectifs marginaux:

$$\bar{X} = \frac{I}{N} \sum_{i=I}^{k} n_{i.} x_i = \sum_{i=I}^{k} f_{i.} x_i$$

$$\bar{Y} = \frac{I}{N} \sum_{j=I}^{l} n_{.j} y_j = \sum_{j=I}^{l} f_{.j} y_j$$

De la même manière on a :

$$Var(X) = \frac{I}{N} \sum_{i=I}^{k} n_{i.} (x_i - \bar{X})^2 = \sum_{i=I}^{k} f_{i.} (x_i - \bar{X})^2$$

$$Var(Y) = \frac{I}{N} \sum_{j=I}^{l} n_{.j} (y_j - \bar{Y})^2 = \sum_{j=I}^{l} f_{.j} (y_j - \bar{Y})^2 ,$$

$$\sigma(X) = \sqrt{Var(X)} \quad \text{et} \quad \sigma(Y) = \sqrt{Var(Y)} .$$

2.2-Exemple.

Les moyennes marginales de la série double de l'exmple I.I sont :

$$\overline{P} = \frac{17.50 + 34.55 + 30.60 + 19.65}{100} = 57,55$$

$$\overline{T} = \frac{18.150 + 23.155 + 18.160 + 19.165 + 22.170}{100} = 160,2 .$$

Le calcul des variances est laissé en exercice.

Dans la suite on se donnera une série double (X,Y) formée de deux séries simples X et Y dont les valeurs x_i et y_i sont numérotées de la même manière de 1 à N.

Nous allons étudier le lien qui peut exister entre X et Y.

3.AJUSTEMENT LINEAIRE.CORRELATION.

On peut representer chaque couple de valeurs (x_i,y_i) par un point M_i dans le plan.On obtient un ensemble de points qu'on appelle un nuage de points (Voir Fig 3.I).

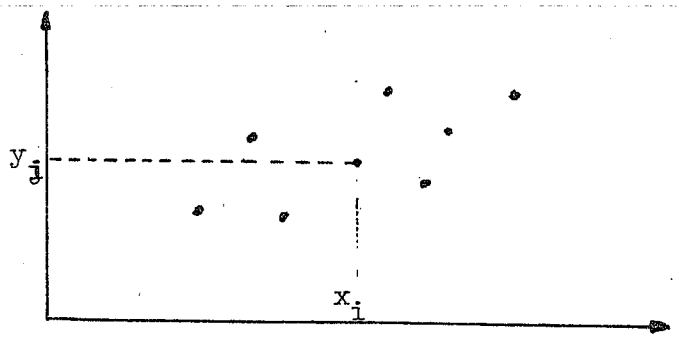


Fig 3.I

On se propose alors de chercher la courbe qui soit la plus proche possible de tous les points M_i .Lui imposer d'être très proche d'un point peut l'éloigner d'autres points.Il s'agit donc de trouver un "juste milieu".

La recherche d'une telle courbe est ce qu'on appelle l'ajustement de ce nuage de points.

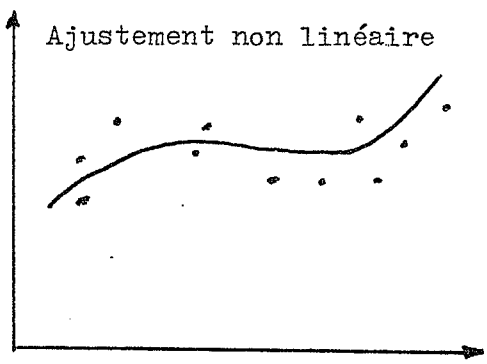


Fig 3.2

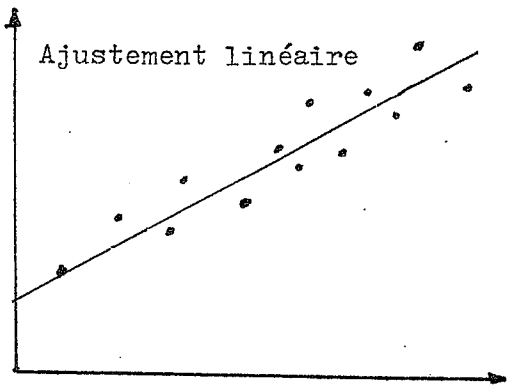


Fig 3.3

Le choix de la courbe n'est pas facile, encore moins son calcul. Nous nous contenterons de traiter le problème dans le cas d'une droite. On dira que l'ajustement est linéaire.

3.I-Droite de Mayer.

Considèrons les N points M_i dont les coordonnées sont x_i et y_i. On appelle point moyen du nuage defini par ces points le point G dont les coordonnées sont les moyennes arithmétiques $\overline{X}$ et $\overline{Y}$.

On peut ranger la série double S de telle sorte que la série des abscisses soit croissante (celle des ordonnées ne le sera pas forcément). On partage alors S en deux séries doubles S_I et S_2 ayant toutes les deux le même effectif total (plus ou moins un terme). On notera G_I et G_2 les points moyens associés à S_I et S_2 respectivement.

3.I.I-Definition.

La droite de Mayer associée à S est la droite definie par les points G_I et G_2 .

Cette droite permet d'ajuster le nuage de points associé à S.

On vérifie qu'elle passe par le point moyen G .

3.I.2-Exemple.

Considèrons la série double suivante :

$S = \{(2,3)\ (3,6)\ (3,4)\ (4,7)\ (4,5)\ (5,8)\ (6,5)\}$.

On partage S en deux séries S_I et S_2 en choisissant les 4 premiers termes pour S_I par exemple.Les points moyens de S_I et S_2 sont alors :

$G_I = (3,5)$ et $G_2 = (5,6)$.

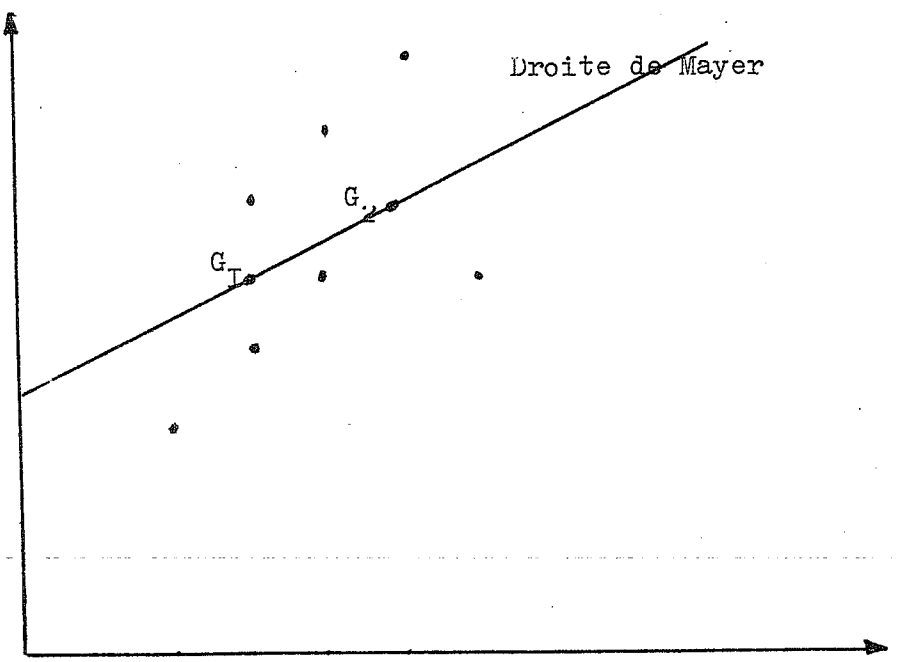


Fig 3.4

Cherchons l'équation de cette droite.On remarque que les points G_I et G_2 ont des abscisses distinctes.L'équation de la droite G_IG_2 est donc de la forme :

$Y = aX + b$ (3.I) .

Pour déterminer a et b il suffit d'exprimer le fait que cette droite passe par G_I et G_2,c'est-à-dire que les coordonnées de ces points vérifient cette équation i.e :

$$\begin{cases} 3a + b = 5 \\ 5a + b = 6 \end{cases} .$$

On en deduit :

$$a = I/2 \quad \text{et} \quad b = 7/2 .$$

Finalement $G_I G_2$ a pour équation : $Y = I/2(X + 7)$.

L'ajustement par la droite de Mayer n'est pas satisfaisant et est peu utilisé dans la pratique.Nous ajusterons plutôt par la droite des moindres carrés que nous allons introduire dans la partie qui suit.

3.2-Droite des moindres carrés.

3.2.I-Position du problème.

Considèrons les N points M_i dans le plan et notons D la droite cherchée.La distance de M_i à D est la mesure du segment $M_i H_i$ où H_i est le pied de la perpendiculaire à D passant par M_i (Voir Fig 3.5). Le calcul de cette distance est relativement compliqué.On calculera plutôt la distance "verticale" de M_i à D.Notons K_i le point de rencontre de D et de la droite d'équation $X = x_i$ et supposons que l'équation de D est de la forme :

$$Y = aX + b .$$

La distance de M_i à K_i est alors le nombre :

$$d_i = |ax_i + b - y_i| .$$

Nous chercherons donc la droite D de telle sorte que la somme :

$$U = \sum_{i=I}^{N} d_i^2$$

soit minimale.

Une telle droite D est appelée la droite des moindres carrés ou droite de regression de Y en X.

3.2.2-Existence et unicité de D.

Tout revient à montrer que a et b existent et sont uniques et rendent U minimale.

Reprenons la quantité :

$$U = \sum_{i=I}^{N} (ax_i + b - y_i)^2 .$$

On a d'abord :

$$(ax_i + b - y_i)^2 = x_i^2 a^2 - 2x_i(y_i - b)a + (y_i - b)^2$$

et par suite :

$$U = (\sum_{i=I}^{N} x_i^2)a - 2(\sum_{i=I}^{N} x_i(y_i - b))a + \sum_{i=I}^{N} (y_i - b)^2 .$$

On pose :

$$\lambda = \sum_{i=I}^{N} x_i^2 \qquad \beta = \sum_{i=I}^{N} x_i(y_i - b) \quad \text{et} \quad \gamma = \sum_{i=I}^{N} (y_i - b)^2$$

L'expression de U est alors un polynôme du second degré en a :

$$U = \lambda a^2 - 2\beta a + \gamma$$

avec λ positif .

Le nombre a cherché est la racine de la dérivée de U considérée comme fonction de a.On a :

$$U'(a) = 2\lambda a - 2\beta$$

et donc $U'(a) = 0$ si et seulement si :

$$a = \frac{\beta}{\lambda}$$ c'est-à-dire :

$$a = \frac{\sum_{i=I}^{N} x_i (y_i - b)}{\sum_{i=I}^{N} x_i^2} \qquad (3.3) \ .$$

Si on considère cette fois-ci U comme un polynôme du second degré en b, un calcul analogue à celui que l'on vient de faire donne :

$$b = \frac{I}{N} \left(\sum_{i=I}^{N} (y_i - a x_i) \right) = \frac{I}{N} \sum_{i=I}^{N} y_i - \frac{a}{N} \sum_{i=I}^{N} x_i$$

qui n'est rien d'autre que :

$$b = \overline{Y} - a\overline{X} \qquad (3.4) \ .$$

On remarque donc que le point moyen $G = (\overline{X}, \overline{Y})$ est sur la droite des moindres carrés.

Les équations (3.3) et (3.4) donnent finalement :

$$a = \frac{\sum_{i=I}^{N} (x_i - \overline{X})(y_i - \overline{Y})}{\sum_{i=I}^{N} (x_i - \overline{X})^2} \qquad (3.5)$$

$$b = \overline{Y} - a\overline{X}$$

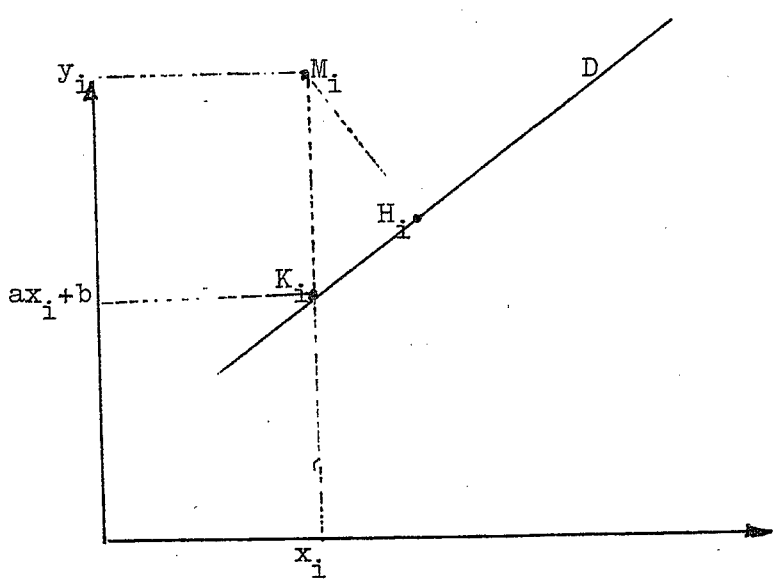


Fig 3.5

3.3-Covariance.

Divisons le numérateur et le dénominateur du second membre de (3.5) par N.On obtient alors :

$$a = \frac{\frac{I}{N}\sum_{i=I}^{N}(x_i - \bar{X})(y_i - \bar{Y})}{\frac{I}{N}\sum_{i=I}^{N}(x_i - \bar{X})^2} \qquad (3.6) \quad .$$

On voit que le dénominateur n'est rien d'autre que la variance de X.Le numérateur :

$$\frac{I}{N}\sum_{i=I}^{N}(x_i - \bar{X})(y_i - \bar{Y})$$

est un paramètre associé à (X,Y) et qu'on appelle la covariance de X et Y.On la note Cov(X,Y).

La formule (3.6) s'écrit alors :

$$a = \frac{Cov(X,Y)}{Var(X)} \qquad (3.7) \quad .$$

La covariance a le signe de a.Elle peut être positive ou négative contrairement à la variance.On peut aussi l'exprimer sous la forme suivante :

$$Cov(X,Y) = \frac{I}{N}\sum_{i=I}^{N} x_i y_i - \bar{X}.\bar{Y} \qquad (3.8) \quad .$$

3.4-Corrélation linéaire.

On appelle coéfficient de corrélation de X et Y le nombre :

$$r = \frac{Cov(X,Y)}{\sigma(X)\sigma(Y)} \qquad (3.9)$$

où $\sigma(X)$ (resp $\sigma(Y)$) est l'écart-type de X (resp de Y).On montre que r possède les propriètés suivantes (les détails et les démonstrations sont laissés au lecteur en exercice) :

i) r a toujours le signe de Cov(X,Y),

ii)r est compris entre -I et +I

iii) $|r| = I$ si et seulement si la droite des moindres carrés passent par tous les points M_i et dans ce cas on a :

$$y_i = ax_i + b \qquad \text{pour tout } i = I,\ldots,N .$$

Si r = 0 i.e Cov(X,Y) = 0 ,on dit que X et Y ne sont pas corrélées.

Nous allons reprendre les calculs que nous venons de faire sur un exemple simple.Ceci nous permettra de mieux comprendre les différentes étapes par lesquelles nous sommes passés.

4.EXEMPLE.

Considèrons la série double :

$(X,Y) = \{(2,3)\ (3,4)\ (4,8)\}$

Pour fixer les idées commençons par tracer cette droite n'importe comment pour le moment :

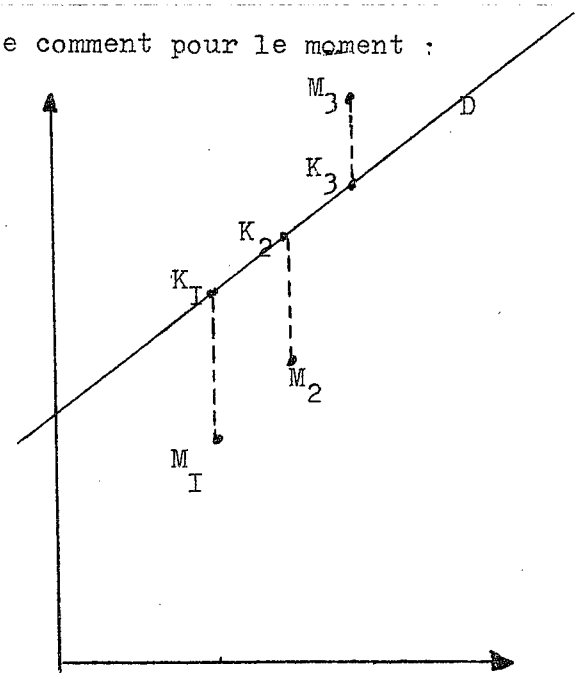


Fig 4.I

Cette série définit un nuage de points M_1 , M_2 et M_3. On calcule la somme des carrés des distances M_iK_i. On obtient :

$$U = (2a + b - 3)^2 + (3a + b - 4)^2 + (4a + b - 8)^2 .$$

On developpe U et on l'écrit sous forme d'un polynôme du second degré en a :

$$U = 29a^2 - 2(50 - 9b)a + (3 - b)^2 + (4 - b)^2 + (8 - b)^2$$

$$= 29a^2 - 2(50 - 9b)a + 3b^2 - 30b + 89 .$$

On calcule la dérivée de U par rapport à a :

$$U'(a) = 2(29a - (50 - 9b))$$

qui s'annule pour :

$$a = \frac{50 - 9b}{29} \qquad (4.I) \quad .$$

D'autre part on a :

$$\bar{X} = 3 \quad \text{et} \quad \bar{Y} = 5 \text{ et donc } \quad 5 = 3a + b \quad (4.2) .$$

Il suffit donc de résoudre le système linéaire en a et b formé par les équations (4.I) et (4.2) , c'est-à-dire :

$$\begin{cases} 29a + 9b = 50 \\ 3a + b = 5 \end{cases}$$

qui donne $a = 5/2$ et $b = -5/2$.

La droite des moindres carrés a donc pour équation :

$$Y = \frac{5}{2}(X - 1) .$$

Le lecteur pourra construire cette droite à titre d'exercice.

5.EXERCICE RESOLU.

Le tableau suivant indique la production interieure brute par secteur au Maroc de I970 à I976 (en Millions de DH) :

Année	Agriculture et Pêche X	Industrie et Mines Y
I970	3,72	3,37
I97I	3,95	3,53
I972	4,09	3,77
I973	3,65	4,08
I974	I3,39	4,44
I975	3,66	4,9I
I976	4,I2	5,43

Tab 5.I

Source : Secrétariat d'Etat chargé du Plan et du dévélopp-ement Régional.

On considère la P.I.B agricole et la P.I.B industrielle comme deux séries statistiques X et Y et on se propose d'étudier la corrélation qui peut exister entre elles.

I°-Calculer Var(X) , Var(Y) , Cov(X,Y) et le cöefficient de corrélation r de X et Y.

2°-Donner la droite de regression de Y en X.

SOLUTION.

On calcule les moyennes arithmétiques X et Y.On trouve :

$\overline{X} = 5,22$ et $\overline{Y} = 4,21$

et on complète le tableau 5.I de la façon suivante :

Année	x_i	y_i	x_i^2	y_i^2	$x_i y_i$
I970	3,72	3,37	I3,83	II,35	I2,53
I97I	3,95	3,53	I5,60	I2,46	I3,94
I972	4,09	3,77	I6,72	I4,2I	I5,4I
I973	3,65	4,08	I3,22	I6,64	I4,89
I974	I3,39	4,44	I79,29	I9,7I	65,74
I975	3,66	4,9I	I3,39	24,I0	I7,97
I976	4,I2	5,43	I6,97	29,48	22,37

Tab 5.2

I°- On en deduit :

$$\text{Var}(X) = \frac{I}{N}\sum_{i=I}^{N} x_i^2 - \overline{X}^2 = 38,45 - (5,22)^2 = 11,20 \quad ,$$

$$\text{Var}(Y) = \frac{I}{N}\sum_{i=I}^{N} y_i^2 - \overline{Y} = I8,28 - (4,21)^2 = 0,55$$

$$\text{Cov}(X,Y) = \frac{I}{N}\sum_{i=I}^{N} x_i y_i - \overline{X}.\overline{Y} = 23,26 - 21,97 = 1,29$$

et par conséquent :

$$r = \frac{Cov(X,Y)}{\sigma(X)\sigma(Y)} = \frac{1,29}{\sqrt{11,20 \cdot 0,55}} = 0,51 \quad .$$

La corrélation est assez apparente.

2°- Droite de regression de Y en X.

Elle est de la forme $Y = aX + b$ avec :

$$a = \frac{Cov(X,Y)}{Var(X)} = \frac{1,29}{11,20} = 0,11$$

et :

$$b = \bar{Y} - a\bar{X} = 4,21 - 0,11 \cdot 5,22 = 3,66$$

c'est-à-dire :

$$Y = 0,11x + 3,66.$$

CHAPITRE VII

SERIES CHRONOLOGIQUES

Considèrons le tableau suivant donnant le nombre d'émigrants en France des départements d'Outre-Mer (Martinique, Guadeloupe, Réunion et Guyane) de I962 à I97I.

Année	I962	63	64	65	66	67	68	69	70	7I
Total	I004	209I	4532	7006	78II	7962	75I4	8398	8807	9165

Tab 0.0

Source : Bilan d'activités I97I du BUMIDOM (Bureau pour le dévéloppement des Migrations Intéressant les Département d'Outre Mer).

On peut interpreter cette série comme une suite d'observations dans le temps.Par exemple on observe qu'en :

i)I964 il y a eu 4532 émigrants,

ii)I968 il y a eu 75I4 émigrants etc...

On dira qu'on a une série chronologique.

Nous allons introduire cette notion d'une façon plus précise et étudier ses propriétés et applications.

I.GENERALITES.

I.I-Définition.

On appelle série chronologique une série statistique dont les valeurs sont obtenues à partir d'observations ordonnées dans le temps.Par exemple :

-La production annuelle de blé ,

-Le taux annuel d'inflation,

-La production annuelle de pétrole etc...

sont des séries chronologiques.

On notera une telle série Y_t qui est la valeur prise par le caractère Y à l'instant t.Par exemple pour la série ci-dessus on a :

Pour t = I962	Y_t = I004	qu'on note Y_I
t = I963	Y_t = 209I	" Y_2
...	...	...
t = I97I	Y_t = 9I65	" Y_{IO}

Une série chronologique est une fonction du temps,donc peut être repesentée graphiquement par une "courbe" (s'il y a suffisamment de points) :

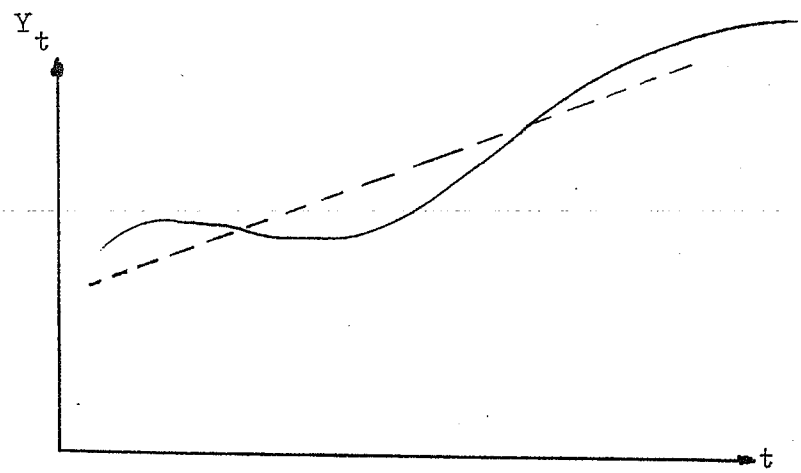


Fig I.1

L'analyse des séries chronologiques permet de prévoir certaines variations à différents niveaux.Ce qui justifie leur utilité et rend leur interêt incontestable.

I.2-Mouvements des séries chronologiques.

Ils sont au nombre de quatre.

I.2.I-Mouvement séculaire.

Il s'étend sur une période de temps assez longue.On s'intéresse beaucoup plus aux variations globales.Ceci est representé en trait interrompu sur la Fig I.I.On parlera alors de tendance séculaire.

Exemple : Série chronologique correspondant à la population mondiale de I800 à I980.

I.2.2-Mouvement saisonnier.

C'est un mouvement pendant une période relativement courte généralement d'une année.Dans la série ce mouvement correspond à des événements périodiques d'année en année.

Exemple : Nombre de vacanciers sur les routes au mois de juillet.

I.2.3-Mouvement cyclique.

C'est un cas intermédiaire entre le mouvement séculaire et le mouvement saisonnier.Son étendue est généralement de la décennie.

Exemple : Baisse de la production d'acier etc...

I.2.4-Mouvement aléatoire.

Ce sont des mouvements qui sont dûs au hasard et qu'on ne peut contrôler.Ils introduisent des"discontinuités" dans l'évolution d'un phénomène.

Exemple : Tremblement de terre,grève,resultat inattendu d'élections etc...

2. ANALYSE.

Analyser une série chronologique c'est étudier les mouvements qui la composent.

Nous presenterons dans ce paragraphe quelques méthodes qui permettent de faire cette analyse.

2.I-Moyennes mobiles.

On se donne une série chronologique $Y = (y_I, \ldots, y_N)$.

On appelle moyenne mobile d'ordre k, la suite des moyennes arithmétiques :

$$Y_1 = \frac{y_I + \ldots + y_k}{k} \qquad Y_2 = \frac{y_2 + \ldots + y_{k+I}}{k} \qquad Y_3 = \frac{y_3 + \ldots + y_{k+2}}{k}$$

$$\ldots \qquad Y_{N-k+I} = \frac{y_{N-k+I} + \ldots + y_N}{k} \quad .$$

On remarque donc que la moyenne mobile est constituée de N-k+I valeurs.

Reprenons la série chronologique donnée par le tableau (0.0) p.89. Elle est constituée de I0 termes et donc sa moyenne mobile d'ordre 4 aura 7 termes dont le premier et le deuxième sont :

$$Y_I = \frac{I004 + 209I + 4532 + 7006}{4} = 3658$$

$$Y_2 = \frac{209I + 4532 + 7006 + 78II}{4} = 5360$$

etc...

Quand les valeurs de la série chronologique sont obtenues par des observations annuelles, mensuelles ou sur une période quelconque on parlera de moyenne mobile sur k années, k mois ou k périodes.

L'intérêt de la moyenne mobile est de"régulariser" les valeurs de la série chronologique.

2.2-Estimation de la tendance séculaire.

Si on porte le temps en abscisse et Y en ordonnée on peut representer la série chronologique par un ensemble de points dans le plan.On pourra donc estimer la tendance en procèdant à un ajustement généralement linéaire par exemple par la méthode des moindres carrés,notion qui a déjà été introduite et étudiée au chapitre qui précède.

On peut aussi utiliser la moyenne mobile.

Regardons comment on procède concrètement sur un exemple.

Reprenons le tableau (0.0) .

t (temps)	1	2	3	4	5	6	7	8	9	IO
Y_t (observation au temps t)	I004	209I	4532	7006	78II	7962	75I4	8398	8807	9I65

Tab 2.I

Calculons la moyenne mobile d'ordre 3.On trouve une nouvelle série à 8 valeurs en remplaçant chaque y_t par :

$$\frac{y_{t-1} + y_t + y_{t+1}}{3} .$$

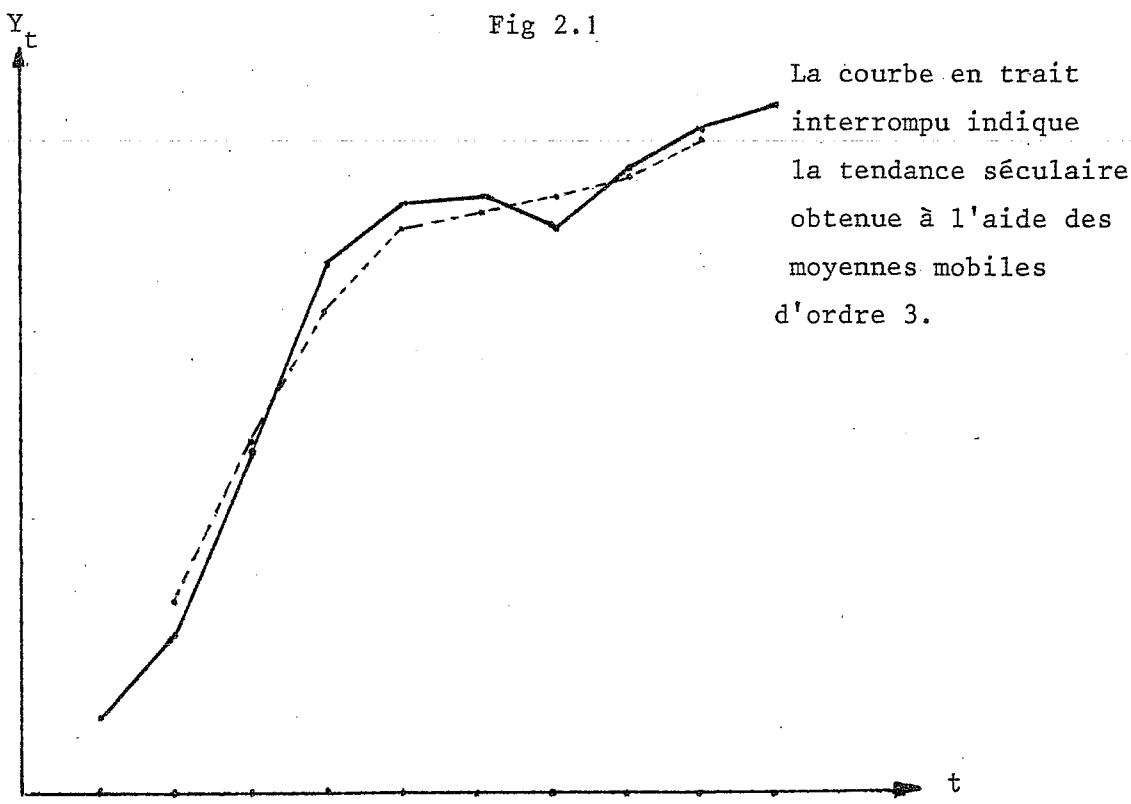
On obtient alors le tableau suivant :

t	Y_t	$\frac{y_{t-1}+ y_t +y_{t+1}}{3}$
1	1004	
2	2091	2542
3	4532	4543
4	7006	6449
5	7811	7593
6	7962	7762
7	7514	7958
8	8398	8239
9	8807	8790
10	9165	

Tab 2.2

Par exemple la première moyenne mobile 2542 est obtenue en faisant la somme des 3 premières valeurs de la colonne 2 et en divisant ensuite par 3.

Fig 2.1



La courbe en trait interrompu indique la tendance séculaire obtenue à l'aide des moyennes mobiles d'ordre 3.

On peut remarquer que dans l'exemple que l'on vient de traiter on aomis le premier et le dernier termes de la série. Si on avait utilisé la définition de la moyenne mobile telle qu'elle a été donnée en 2.I page 92 on aurait omis les deux dérnières valeurs de la série.Dans la pratique on peut prendre l'une ou l'autre des définitions de la moyenne mobile.Les deux sont pratiquement les "mêmes".

2.3-Coéfficient saisonnier.

On détermine la tendance séculaire à l'aide de la moyenne mobile ou la méthode des moindres carrés suivant le cas considèré (évidemment on choisira celle qui s'adapte mieux au problème). On obtient pour chaque valeur y_t une valeur ajustée qu'on notera y_t^a .Le coéfficient saisonnier à l'instant t est alors le rapport de la valeur observée y_t à la valeur ajustée y_t^a :

$$\boxed{\varrho_t = \frac{y_t}{y_t^a}} \qquad (2.2) .$$

2.4-Desaisonnalisation.

On divise chaque observation initiale par le coéfficient saisonnier moyen $\bar{\varrho}_t$ (moyenne des coéfficients saisonniers de la même époque calculée sur des années différentes).Plus précisément la valeur desaisonnalisée y_t^d de y_t est définie par :

$$\boxed{y_t^d = \frac{y_t}{\bar{\varrho}_t}} \qquad (2.3) .$$

3.EXERCICE RESOLU.

On suppose que les livraisons trimestrielles d'une certaine marchandise de I976 à I980 sont données par le tableau suivant (en milliers de tonnes) :

Année	I^{er}Trimestre	$2^{ème}$Tri	$3^{ème}$Tri	$4^{ème}$Tri
I976	I40	I42	I44	I4I
I977	I42	I47	I47	I49
I978	I50	I50	I52	I52
I979	I53	I55	I58	I58
I980	I60	I63	I65	I66

Tab 3.I

I°-Estimer la tendance séculaire par :

-La méthode des moyennes mobiles,

-La méthode des moyennes échelonnées (que l'on va introduire).

2°-Ajuster cette série chronologique par la droite des moindres carrés.

3°-Calculer les coéfficients saisonniers.

SOLUTION.

I°-Rappelons que la moyenne mobile d'ordre 3 est donnée par:

$$Y_I = \frac{y_I + y_2 + y_3}{3}, \quad Y_2 = \frac{y_2 + y_3 + y_4}{3} \quad \cdots$$

$$Y = \frac{y_{I8} + y_{I9} + y_{20}}{3}.$$

Dans le cas qui nous intersse ici on a N = 20.

Calculons par exemple Y_{II} : y_{II} est la valeur de Y au 3ème trimestre de l'année I978.On aura ainsi :

$$Y_{II} = \frac{I52 + I52 + I53}{3} = I52,33 .$$

On rassemble tous ces calculs dans le tableau 3.2.

t	y_t	Moyenne mobile
I	I40	I42
2	I42	I42,33
3	I44	I42,33
4	I4I	I43,33
5	I42	I45,33
6	I47	I47,66
7	I47	I48,66
8	I49	I49,66
9	I50	I50,66
I0	I50	I5I,33
II	I52	I52,33
I2	I52	I53,33
I3	I53	I55,33
I4	I55	I57
I5	I58	I58,66
I6	I58	I60,33
I7	I60	I62,66
I8	I63	I64,66
I9	I65	
20	I66	

Tab 3.2

Pour fixer un peu les idées traçons un graphique representant la série chronologique.

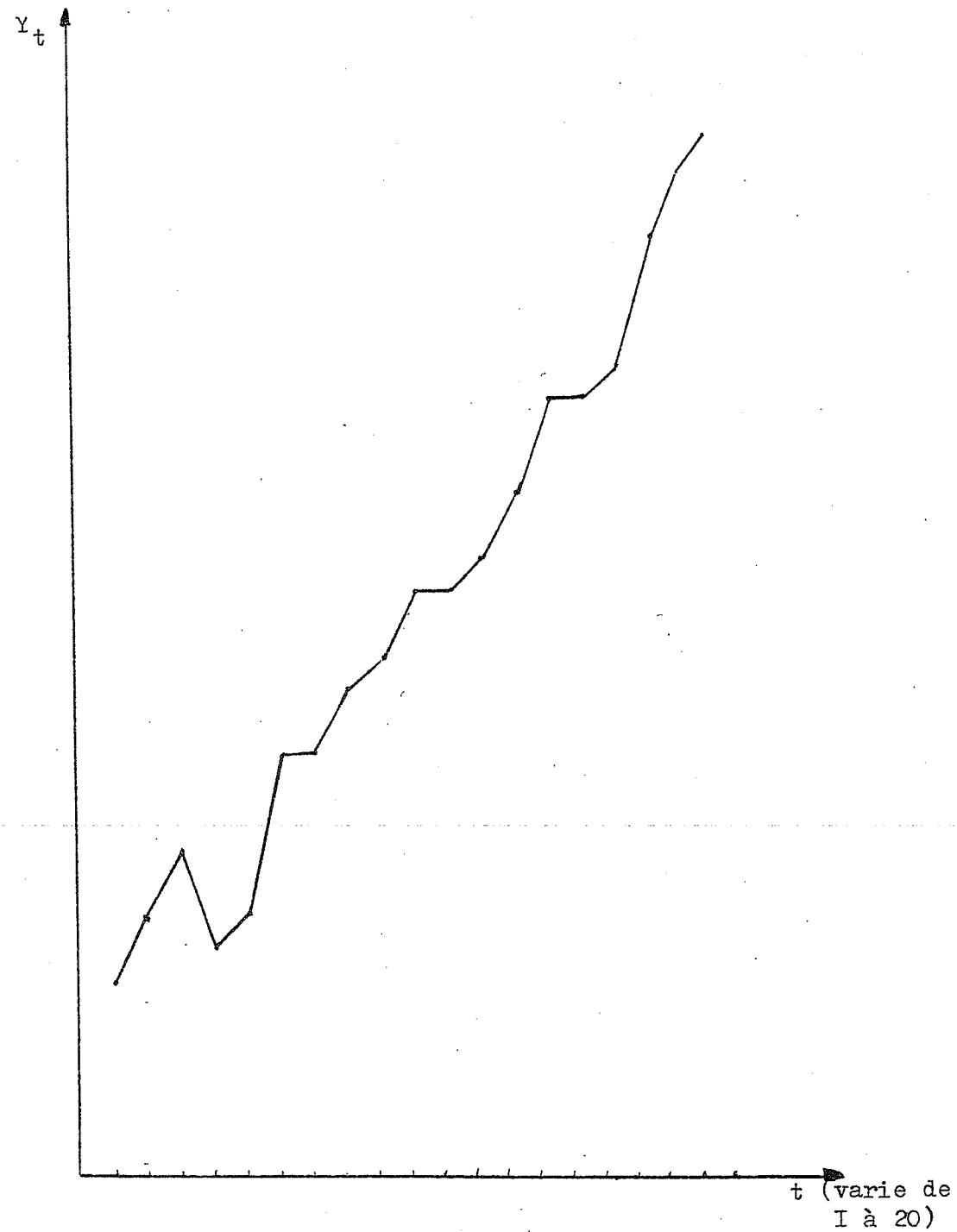


Fig 3.I

Les moyennes échelonnées se calculent en faisant la moyenne sur les quatres trimestres d'une même année.Le calcul est décrit ci-dessous :

t	y_t	Moyenne échelonnée
I976 I	I40	
2	I42	I4I,75
3	I44	
4	I4I	
I977 5	I42	
6	I47	146,25
7	I47	
8	I49	
I978 9	I50	
I0	I50	I5I
II	I52	
I2	I52	
I979 I3	I53	
I4	I55	
I5	I58	156
I6	I58	
I980 I7	I60	
I8	163	I63,50
I9	I65	
20	I66	

Tab 3.3

2°-Ajustement par la méthode des moindres carrés.

Cette droite des moindres carrés a une équation de la forme :

$Y=at + b$

où a et b sont donnés par les formules (3.5) et (3.4) page 82,

c'est à dire :

$$a = \frac{\sum_{t=1}^{N} ty_t - N\bar{t}\bar{Y}}{\sum_{t=1}^{N} t^2 - N\bar{t}^2} \qquad (3.1) \quad .$$

Ici t joue le rôle de X et prend les valeurs 1,...,20;

$$b = \bar{Y} - a\bar{t} \qquad (3.2)$$

Nous rassemblerons les différents calculs dans le tableau (3.4).

En utilisant ce tableau on calcule successivement :

$$\bar{t} = \frac{1+2+\ldots+20}{20} = 10,5 \qquad \text{et} \qquad \bar{t}^2 = 110,25$$

$$\sum_{t=1}^{20} ty_t = 32745 \qquad \sum_{t=1}^{20} t^2 = 2870$$

En portant toutes ces valeurs dans l'expression (3.1) on trouve :

$a = 1,33$ et donc $b = \bar{Y} - a\bar{t} = 151,70 - 1,33.10,5 = 137,73$.

L'équation de la droite des moindres carrés est finalement :

$$Y = \frac{133t + 13773}{100} \qquad (3.3) \quad .$$

On choisira de faire le calcul des valeurs ajustées en utilisant l'équation (3.3).On obtient le tableau (3.4) dans lequel on a aussi inseré les coéfficients saisonniers ϱ_t.

t	y_t	ty_t	t^2	y_t^a	ϱ_t (en %)
1	140	140	1	139,06	100,07
2	142	284	4	140,39	101,14
3	144	432	9	141,72	101,60
4	141	564	16	143,05	98,56
5	142	710	25	144,38	97,45
6	147	882	36	145,71	100,88
7	147	1029	49	147,09	99;93
8	149	1192	64	148,37	100,42
9	150	1350	81	149,70	100,20
10	150	1500	100	151,03	99,31
11	152	1672	121	152,36	99,76
12	152	1824	144	153,69	98,98
13	153	1989	169	155,02	98,69
14	155	2170	196	156,35	99,13
15	158	2370	225	157,68	100,20
16	158	2528	256	159,01	99,36
17	160	2720	289	160,34	99,78
18	163	2934	324	161,48	100,82
19	165	3135	361	163,00	101,22
20	166	3320	400	164,33	101,01

Tab 3.4

On constate que les coéfficients saisonniers sont très voisins de 1 (i.e de 100%).Ce qui signifie que la série chronologique étudiée est pratiquement la fonction linéaire du temps donnée par (3.3).

CHAPITRE VIII

EXERCICES RESOLUS

Dans ce chapitre nous donnons une correction détaillée pour certains exercices ayant été proposés à différentes sessions d'examen deStatistiques de l'Unité 01103 à l'Univærsité de Lille III.

Il est bien clair que ce chapitre ne remplira sa fonction que si l'étudiant manifeste au préalable sa volonté de resoudre chaque exercice avant de consulter sa solution.

Session de Juin 1974.

Exercice

On donne le tableau suivant traduisant une correspondance entre deux séries X et Y.

X	250	300	350	400	450	500	550	600	650	700	750	800	850	900	950	1000
Y	174	218	263	309	356	404	453	503	554	605	658	711	766	821	877	934

1°-Representer graghiquement cette correspondance.

2°-Dites pourquoi on peut considèrer comme valable un ajustement linéaire.

3°-Chercher la droite d'ajustement par les moindres carrés en donnant son équation.

Solution

Considèrons un repère constitué de deux axes réctangulaires dans lequel on portera les valeurs de X en abscisses et celles de Y en ordonnées.Chaque couple de valeurs (x,y) définit un point dans ce repère. Par exemple le point A a pour abscisse $x = 650$ et pour ordonnée $y = 554$. On obtient le nuage de points suivant :

1°-Representation graphique.

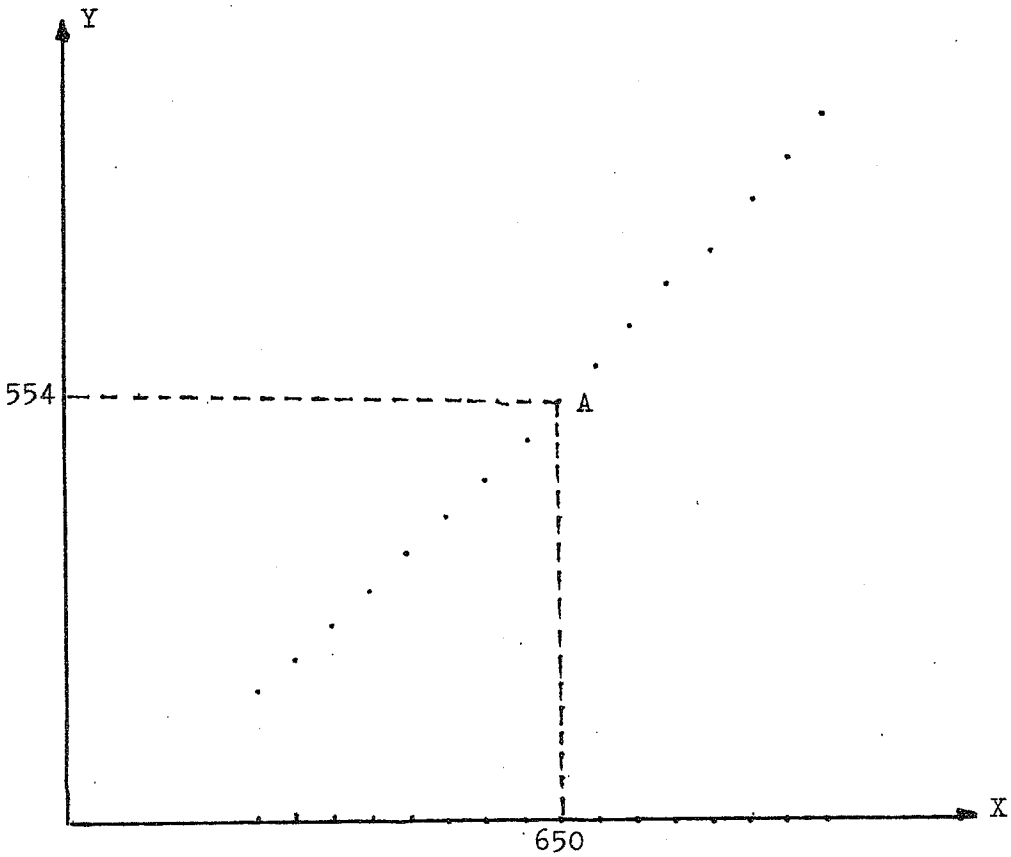


Fig 1.1

2°-On remarque que tous les points obtenus sont presque alignés.Ceci justifie fortement un ajustement linéaire que nous allons déterminer.

3°-Ajustement par la méthode des moindres carrés.

L'équation de la droite des moindres carrés est de la forme : $Y = aX + b$.Ici $a = \frac{Cov(X,Y)}{Var(X)}$ et $b = \overline{Y} - a\overline{X}$ (Voir p.82).

Pour déterminer a et b il suffit donc de calculer $\overline{X}$, $\overline{Y}$, Var(X) et Cov(X,Y).

Rappelons que :

$Var(X) = \overline{X^2} - \overline{X}^2$ et $Cov(X,Y) = \overline{XY} - \overline{X}.\overline{Y}$.

Nous résumons tous les calculs dans le tableau suivant :

X	Y	X^2	XY
250	174	62500	43500
300	218	90000	65400
350	263	122500	92050
400	309	160000	123600
450	356	202500	160200
500	404	250000	202000
550	453	302500	249150
600	503	360000	301800
650	554	422500	360100
700	605	490000	423500
750	658	562500	493500
800	711	640000	568800
850	766	722500	651100
900	821	810000	738900
950	877	902500	833150
1000	934	1000000	934000

En utilisant ce tableau on trouve :

$\overline{X} = 625$ et $\overline{Y} = 537$.

De la même manière on a $Var(X) = \overline{X^2} - \overline{X}^2 = 444375 - 390625 = 53750$

$Cov(X,Y) = \overline{XY} - \overline{X}.\overline{Y} = 390046 - 335625 = 54421.$

D'où $a = \frac{54421}{53750} = 1,01$

et $b = \overline{Y} - a\overline{X} = 537 - 1,01.625 = -95,81.$

Finalement on a l'équation de la droite des moindres carrés :

$$Y = 1,01.X - 95,81$$

Session d'Octobre 1974.

Exercice

On considère le tableau :

X	x_1	x_2	...	x_{19}	x_{20}
Y	y_1	y_2	...	y_{19}	y_{20}

En vue d'ajuster la correspondance ainsi définie à une droite on apréparé les calculs suivants :

Posant $z_i = x_i - 50$ et $w_i = y_i - 70$ on a obtenu :

$$\sum z_i = -41 \quad , \sum w_i = 39 \quad , \sum z_i w_i = 7153 \text{ et } \sum z_i^2 = 8369.$$

Les sommes $\sum$ étant prises pour i variant de 1 à 20.

1°-Calculer les moyennes de Z et W.En déduire celles de X et Y.

2°-Déterminer l'équation de la droite des moindres carrés donnant un ajustement linéaire entre X et Y.

Solution

1°-En utilisant la relation 2.6 p.25 avec e=1 on voit que :

$\overline{X} = \overline{Z} + 50$ et $\overline{Y} = \overline{W} + 70$.

Or : $\overline{Z} = \frac{\sum z_i}{20} = \frac{-41}{20} = -2,05$ et $\overline{W} = \frac{\sum w_i}{20} = \frac{39}{20} = 1,95$.

D'où : $\overline{X} = \overline{Z} + 50 = -2,05 + 50 = 47,95$

$\overline{Y} = \overline{W} + 70 = 1,95 + 70 = 71,95$.

2°-La droite des moindres carrés a pour équation :

$Y = aX + b$ avec $a = \frac{Cov(X,Y)}{Var(X)}$ et $b = \overline{Y} - a.\overline{X}$.

On sait que $Var(X) = Var(X + 50) = Var(Z) = \overline{Z^2} - \overline{Z}^2$.

Donc : $Var(X) = \frac{8369}{20} - (-2,05)^2 = 418,45 - 4,2025 = 414,25$

Avant de calculer Cov(X,Y) démontrons d'abord la proposition :

Si a et b sont des constantes réelles on a : $Cov(X+a,Y+b)=Cov(X,Y)$.

En effet on a :

$$Cov(X+a,Y+b)= \overline{(X+a)(Y+b)} - \overline{(X+a)}.\overline{(Y+b)} .$$

Or $(X+a)(Y+b) = XY + bX + aY + ab$ donc $\overline{(X+a)(Y+b)}=\overline{XY} + b\overline{X} + a\overline{Y} + ab$.

De la même façon on a :

$$\overline{(X+a)}.\overline{(Y+b)} =(\overline{X} +a).(\overline{Y} +b)$$

$$=\overline{X}.\overline{Y} + b.\overline{X} + a.\overline{Y} + ab.$$

On en déduit donc :

$Cov(X,Y) = \overline{XY} - \overline{X}.\overline{Y} = Cov(X+a,Y+b)$. Ce qui démontre la proposition.

En appliquant cette proposition à $X = Z + 50$ et $Y = W + 70$ on obtient :

$$Cov(X,Y) = Cov(Z,W) = \overline{ZW} - \overline{Z}.\overline{W}$$

$$= \frac{7153}{20} - (-2,05.1,95)$$

$$=357,65 -3,9975$$

$$353,66.$$

Finalement $a = \frac{353,66}{414,25} = 0,85$ et $b = \overline{Y} - a.\overline{X} = 71,95 - 0,85.47,95$

c'est-à-dire $b = 31,01$

Ce qui donne l'équation : $\boxed{Y = \frac{85X + 3101}{100}}$.

Session de Juin 1975.

Exercice 1

Pour la série classée suivante :

Classes	Effectifs partiels
[145,150[	51
[150,155[	95
[155,160[	131
[160,165[	152
[165,170[	120
[170,175[	77
[175,180[	48

1°-Calculer la médiane et la moyenne arithmétique.

2°-Construire l'histogramme des effectifs cumulés et déterminer graphiquement les valeurs des quartilles.

Solution

1°-Calculons les effectifs cumulés de chaque classe :

Classes	Effectifs cumulés
[145,150[	51
[150,155[	146
[155,160[	277
[160,165[	429
[165,170[	549
[170,175[	626
[175,180[	674

L'effectif total est N = 674 ; donc N/2 = 337.La médiane tombe donc dans la classe [160,165[.Pour la calculer on utilise alors la formule 1.1 p.18 et on obtient :

$$M = 160 + (165 - 160).\frac{337 - 277}{429 - 277} = 160 + 5.\frac{60}{152} = 161,97 .$$

Pour calculer la moyenne arithmétique on fait un changement de variable en posant : $z_i = \frac{c_i - 162,5}{5}$ où c_i est le centre de la classe $[x_i , x_{i+1}[$ .On obtient alors la série à valeurs isolées :

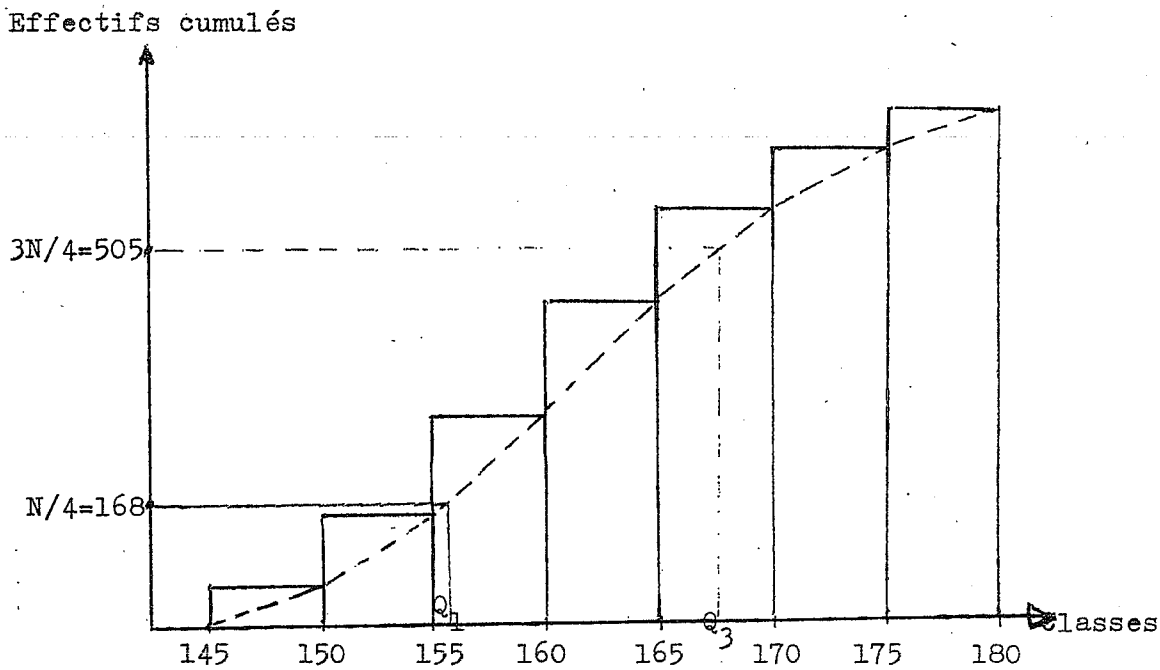
-3	-2	-1	0	1	2	3	(Valeurs)
51	95	131	152	120	77	48	(Effectifs).

Cette série a pour moyenne :

$$Z = \frac{-3.51+(-2).95+(-1).131+0.152+1.120+2.77+3.48}{674} = -0,83$$

Ce qui donne $X = 5.(-0,83) + 162,5 = 162,08$.

2°-Pour construire l'histogramme des effectifs cumulés on trace un repère sur lequel on porte les classes en abscisse et les effectifs cumulés en ordonnée.On obtient le diagramme suivant :



On lit alors sur le diagramme les valeurs des quartiles :

$Q_1 = 156$ et $Q_3 = 168$.

Exercice 2

Le tableau suivant represente la moyenne des prix de gros et la production de lait,de beurre et de fromage aux Etats-Unis pour les années 1949 et 1958.

En prenant 1949 comme année de reférence calculer :

1°-L'indice de Laspeyres pour l'année 1958,

2°-L'indice de Paasche pour l'année 1958 toujours.

Produits	Prix(en francs par Kg ou litre)		Quantités produites en millions de Kgs ou litres	
	1949	1958	1949	1958
Lait	0,40	0,41	9,6	10,4
Beurre	6,1	6	117	115
Fromage	3,4	3,9	78	83

Solution

Les indices de Laspeyres et de Paasche sont donnés respectivement par les formules (1.9) et (1.10) p.50.Dans le cas qui nous intéresse ici on a :

1°-Indice de Laspeyres :

$$\sum Q_0^i P_1^i = (9,6).(0,41)+(117).6+78.(3,9) = 1010,13$$

$$\sum Q_0^i P_0^i = (9,6).(0,40)+(117).(6,1)+78.(3,4) = 982,74.$$

On a donc $I^L_{1/0} = \frac{1010,13}{982,74} = 1,02$ (102%) .

De la même manière on a :

2°-Indice de Paasche :

$$\sum Q_1^i P_1^i = (10,4).(0,41)+115.6+83.(3,9) = 1017,96$$

$$\sum Q_1^i P_0^i = (10,4).(0,40)+115.(6,1)+83.(3,4) = 987,86 .$$

D'où :

$$I^P_{1/0} = \frac{1017,96}{987,86} = 1,03 \quad (103\%) .$$

Exercice 3

La série chronologique suivante indique pour les années 1948 à 1958 la production de houille grasse en millions de tonnes aux U.S.A.

Année	1948	1949	1950	1951	1952	1953	1954	1955	1956	1957	1958
Production	50	37	43	44	39	38	33	39	42	41	34

1°-Construire le tableau des moyennes mobiles sur 5 ans.

2°-Quel est l'inconvénient de la méthode des moyennes mobiles pour l'estimation de la tendance?

3°-Ajuster cette série par la méthode des moindres carrés.

4°-Utiliser cet ajustement pour estimer la production en 1962.

Solution

1°-Moyennes mobiles sur 5 ans.

Année	1948	1949	1950	1951	1952	1953	1954	1955	1956	1957	1958
Production	50	37	43	44	39	38	33	39	42	41	34
Moyennes mobiles			42,6	40,2	39,4	38,6	38,2	38,6	37,6		

2°-Parmi les inconvénients que presente la méthode des moyennes mobiles on peut citer :

-La réduction du nombre de termes dans la série.Par exemple la série initiale ci-dessus a 11 termes et celle obtenue par les moyennes mobiles n'en a que 7.Il y a un peu d'information qui se perd.

-Elle risque de donner lieu à des mouvements et des cycles nouveaux.

3°-Droite des moindres carrés.

On reécrit la série chronologique en considérant les années comme une série variant de 1 à 11.On obtient ainsi en designant par y_t la production à l'année t le tableau suivant :

t	1	2	3	4	5	6	7	8	9	10	11
y_t	50	37	43	44	39	38	33	39	42	41	34

La droite des moindres carrés a pour équation :

$Y_t = at + b$ avec :

$$a = \frac{\sum_{t=1}^{11} t y_t - N\bar{t}\bar{Y}_t}{\sum_{t=1}^{11} t^2 - N\bar{t}^2} \quad \text{et} \quad b = \bar{Y}_t - a\bar{t} .$$

En utilisant le tableau ci-dessus on trouve :

$$\sum_{t=1}^{11} t^2 = 506 \qquad \sum_{t=1}^{11} t y_t = 2557$$

$$\bar{t} = 6 \qquad \bar{t}^2 = 36 \qquad \text{et} \qquad \bar{Y}_t = 40 .$$

On en déduit : $a = \frac{2557 - 11.6.40}{506 - 11.36} = \frac{-83}{110} = -0,75$

et $b = 40 + 0,75.6 = 42,70$

Ce qui donne $Y_t = \frac{-75t + 4270}{100}$.

4°-L'année 1962 correspond à t = 15.D'où :

Production en 1962 = $y_{15} = \frac{-75.15 + 4270}{100} = 31,45$.

Session de Septembre 1975.

On considère la série statistique classée suivante :

Classes	Effectifs
[40 , 45[	8
[45 , 50[	11
[50 , 55[	31
[55 , 60[	61
[60 , 65[	54
[65 , 70[	58
[70 , 75[	43
[75 , 80[	25
[80 , 85[	17
[85 , 90[	7

1°-Déterminer la moyenne arithmétique $\bar{X}$ et l'écart-type σ.

2°-Quel est le pourcentage des valeurs comprises entre $\bar{X} - 2\sigma$ et $\bar{X} + 2\sigma$?

Solution

1°-La série des centres est :

C	42,5	47,5	52,5	57,5	62,5	67,5	72,5	77,5	82,5	87,5
	8	11	31	61	54	58	43	25	17	7

On pose : $z_i = \frac{c_i - 62,5}{5}$ .On obtient une nouvelle série :

Z	-4	-3	-2	-1	0	1	2	3	4	5	
	8	11	31	61	54	58	43	25	17	7	(Effectifs)

qui a pour moyenne arithmétique :

$$\bar{Z} = \frac{134}{315} = 0,42 \quad \text{et donc} \quad \bar{X} = 5\bar{Z} + 62,5 = 5.0,42 + 62,5 + 64,62.$$

On a d'autre part : $\sigma(X) = 5\,\sigma(Z)$.

Il suffit donc de calculer $\sigma(Z)$ lequel est :

$$\sigma(Z) = \sqrt{\overline{Z^2} - \bar{Z}^2} \quad .$$

On calcule $\overline{Z^2}$ et on trouve :

$$\overline{Z^2} = \frac{1252}{315} = 3,97 \ .$$

D'où : $\sigma(Z) = \sqrt{3,97 - (0,42)^2} = 1,94$

et $\sigma(X) = 5.1,94 = 9,73$.

2°-On a $\bar{X} - 2\sigma = 64,62 - 2.9,73 = 45$

$\bar{X} + 2\sigma = 64,62 + 2.9,73 = 84$

Commençons d'abord par calculer l'effectif des valeurs qui sont comprises entre 45 et 84. Il est égal à l'effectif des valeurs qui sont comprises entre 45 et 80 + l'effectif des valeurs qui sont comprises entre 80et 84 qui est égal à 4/5 fois celui de la classe $[80 , 85[$ .Ce qui donne en définitive 283 + 4/5(17)=297. Ce qui represente 94% de l'effectif total.

Session de Juin 1977.

Exercice

On considère la série à valeurs isolées suivante :

X	13	14	15	16	17	18	19	20	21	22
n_i	4	6	7	15	24	16	14	7	4	3

1°-Representer graphiquement cette série par un diagramme en bâtons et tracer le polygône des effectifs.

2°-Calculer la moyenne,la variance et l'écart-type de cette série.

2°-Par définition la moyenne arithmétique de cette série est :

$$\bar{X} = \frac{1}{N}\sum_{i=1}^{k} n_i x_i = \frac{1}{100}(4.13+6.14+ \ldots +4.21+3.22) = 17,33$$

On a $Var(X) = \overline{X^2} - (\bar{X})^2$.

Or :

$$X^2 = \frac{1}{100}(4.13^2+6.14^2+ \ldots + 4.21^2+3.22^2) = 304,57$$

D'où $Var(X) = 304,57 - (17,33)^2 = 4,24$ et donc :

$$\sigma(X) = \sqrt{Var(X)} = 2,05 .$$