TABLE DES MATIERES

INTRODUCTION 1

1-POPULATION-ECHANTILLON 2

2-DISTRIBUTION DE DONNEES 3

2-1 Distribution de données non groupées 3

2-2 Distribution de données groupées en classe 5

3-REPRESENTATION GRAPHIQUE D'UNE DISTRIBUTION 7

3-1 Distribution non groupée 7

3-2 Distribution groupée en classes 12

3-3 Premier classement des distributions 15

3-4 Autres méthodes de représentation graphique de distribution 17

4-PARAMETRES D'UNE SERIE STATISTIQUE 20

4-1 Emploi du signe Σ 20

4-2 Paramètres de position ou de tendance centrale 21

4-2-1 Le mode 21

4-2-2 La médiane 22

4-2-3 La moyenne arithmétique 24

1 Définition 24

2 Signification Physique de la moyenne arithmétique 25

3 Signification géométrique de la moyenne arithmétique 25

4 Changement de variable 27

5 Avantages et inconvénients de la moyenne arithm. 28

4-2-4 Que choisir ...? 29

4-2-5 La moyenne géométrique 30

4-2-6 La moyenne harmonique 30

4-3 Paramètres de dispersion 31

4-3-1 Introduction 31

4-3-2 Etendue, amplitude, intervalle de variation 32

4-3-3 Les quartiles 32

4-3-4 Intervalle interquartille 33

4-3-5 Ecart-moyen absolu 34

4-3-6 La variance, l'écart-type 35

1 Définition 35

2 Autre formule pour le calcul de σ^2 et de σ 36

3 Quelques propriétés de la variance et de l'écart-type 37

4 Changement de variable 38

5 Variable réduite 39
6 Estimation de l'écart-type de la population à partir de l'échantillon 41
4-3-7 Coefficient de variation 42
4-3-8 Usage des différents paramètres de dispersion 43
4-3-9 Inégalité de Bienaymé - Tchebicheff 43
5- La gaussienne 46
6- Série statistique double 47
6-1 Exemple 47
6-2 Tableau de corrélation 47
6-3 nuage de points 48
6-4 Corrélation 49
6-5 Corrélation linéaire 50
6-6 Méthode des moindres carrées 51
6-7 Coefficient de corrélation 55

Introduction

Toute expérience scientifique doit pouvoir être répétée et, dans les mêmes conditions, conduire aux mêmes résultats. Toute proposition, toute loi et, de manière générale, tout modèle mathématique qui prétend décrire ou expliquer le réel ne peut être contredit par un seul fait de la réalité observée. Les modèles mathématiques sont donc constamment remis en question et n'ont donc jamais un caractère définitif.

Mais, en sciences dites humaines, un fait isolé est rarement reproductible. Cependant, les événements qui se rattachent aux sciences humaines sont souvent reproductibles dans des ensembles bien définis (POPULATION) quand on les obtient par l'observation de groupes assez nombreux (ECHANTILLON). L'analyse statistique permet d'évaluer la reproductibilité des faits observés, la vraissemblance des propositions, la conformité des hypothèses.

La statistique est la science qui se propose de rassembler, d'ordonner, de représenter, d'étudier pour en tirer des conclusions, les données numériques se rapportant à des phénomènes collectifs.

La partie des statistiques qui se propose de rassembler, d'ordonner, de représenter les données s'appelle la statistique descriptive.

La partie qui s'occupe de tirer les conclusions et à laquelle le calcul des probabilités sert de base, s'appelle statistique ~~descriptive~~. mathématique

1~ Population - échantillon

L'ensemble des éléments auxquels se rapporte une recherche statistique s'appelle une POPULATION.

Exemples: Les élèves d'une classe, les petits pois d'une mêmes espèce, les revenus imposables en Belgique, la vitesse d'un corps lors d'une expérience de physique, le prix d'un disque dans différents points de vente...

Considérons comme population, l'ensemble des Belges. On se pose par exemple la question suivante: Quel est l'âge moyen du Belge?

Une première méthode consiste à relever l'âge de tous les Belges et d'en faire la moyenne. Une seconde méthode consiste à relever l'âge de quelques milliers de Belges "pris au hasard" et d'en faire la moyenne.

La première méthode est peu commode et sa précision est illusoire (naissance et décès durant le relèvement). La seconde méthode est précise dans la mesure où

1°- Les Belges sont vraiment pris au hasard dans l'ensemble de la population.

2°- Le nombre d'âges relevés est grand.

Définition: Echantillon: ensemble d'éléments à propos desquels on a effectivement recueilli des données.

Population : ensemble d'éléments parmi lesquels on aurait pu choisir l'échantillon c-à-d l'ensemble des éléments qui possèdent la caractéristique que l'on veut observer.

Un ECHANTILLON est représentatif d'une population pour un caractère s'il n'y a aucune raison de penser que la valeur de ce caractère puisse différer dans l'échantillon et dans la population. Le problème se pose du choix d'un échantillon valable.

2 - Distribution des données

Lorsqu'on collecte des données, il faut les représenter d'une façon claire avant de pouvoir les interpréter.

2-1 DISTRIBUTION DE DONNES NON GROUPEES

Exemple 1: Considérons les résultats obtenus sur 10 à un travail fait par une classe de 25 élèves. On obtient les résultats suivants:

2 4 5 2 7 6 3 4 5 5 5 5 4 2 4 3 5 2 5 6 6 6 7 9 4

Exemple 2: Prélèvement d'un échantillon de cosses de petits pois d'une variété déterminée tirée de la récolte 79 à l'institut horticole de Gembloux.
Nombre de petits pois par cosse:

2	4	5	6	2	4	9	3	1	7	8	6	8	7	10	8	12	10	9	
6	7	4	3	9	7	5	6	4	5	5	7	4	8	6	11	5	6	5	
5	7	6	8	5	3	4	2	3	0	9	1	4	0	8	6	6	5	5	
6	5	4	5	3	6	7	6	4	6	6	8	12	10	7	5	5	3	2	
5	6	4	7	6	8	4	3	5	4	6	14	11	9	7	3	1	2	0	
5	5	6	7	6	7	8	10	7	5	3	5	6	13	4	6	2	8	9	
11	12	9	0	2	3	6	8	7	2	3	5	7	6	8	9	10	12	8	
0	2	3	4	6	7	8	7	4	5	7	6	10	9	9	8	4	5	4	
5	6	7	8	8	5	5	5	6	4	7	7	9	12	3	2	2	4	5	
2	1	3	5	4	7	8	8	7	8	8	6	6	5	6	3	4	6	6	
6	6	7	8	6	6	5	6	4	6	6	6	9	7	4	8	5	6	0	
4	6	6	7	5	5	8	6	6	5	6	6	6	9	9					

Nous pouvons classer les résultats sous forme d'un tableau ordonné:

Exemple 1:

résultat de l'élève x_i	nombre d' apparitions n_i
2	4
3	2
4	5
5	7
6	4
7	2
8	0
9	1
	$n=\sum_{i=1}^{} n_i = 25$

Exemple 2:

Nombre de petits pois par cosse x_i	Nombre d' apparitions n_i
0	6
1	4
2	12
3	15
4	24
5	35
6	48
7	26
8	24
9	14
IO	6
II	3
12	5
13	1
14	1
	$n=\sum_{i=1}^{15} n_i = 224$

x_i sont les <u>valeurs observées</u> et n_i, l'effectif ou <u>fréquence absolue</u> d'apparition de la valeur observée x_i

(Dans l'exemple 1, $x_4 = 5$ et $n_4 = 7$ signifie qu'il y a 7 élèves qui ont obtenu la cote 5 sur 10)

S'il y a p valeurs observées et que l'effectif de l'échantillon est n, on a

$$\sum_{i=1}^{p} n_i = n$$

Pour de tels tableaux, on définit

1°) $\boxed{f_i = \frac{n_i}{n}}$ qui est la <u>fréquence relative</u> de la valeur observée x_i.

Dans l'exemple 1, $f_2 = 0{,}08$ signifie qu'il y a 8% des élèves qui ont obtenus la cote 3 sur 10.

on a $\boxed{\sum_{i=1}^{p} f_i = 1}$

2°) $\boxed{\varphi_k = \sum_{i=1}^{k} f_i = \frac{1}{n} \sum_{i=1}^{k} n_i}$ qui est une fréquence relative cumulée.

Dans l'exemple 1, φ_4 =0,72 signifie qu'il y a 72 % des élèves qui ont obtenu une Cote inférieure ou égale à 5.

Exemple 1:

x_i	n_i	f_i	φ_i
2	4	0.16	0.16
3	2	0.08	0.24
4	5	0.20	0.44
5	7	0.28	0.72
6	4	0.16	0.88
7	2	0.08	0.96
8	0	0	0.96
9	1	0.04	1
	25	1	

Exemple 2:

x_i	n_i	f_i	φ_i
0	6	0.026	0.026
1	4	0.017	0.044
2	12	0.053	0.098
3	15	0.066	0.165
4	24	0.107	0.272
5	35	0.156	0.428
6	48	0.214	0.642
7	26	0.116	0.758
8	24	0.107	0.866
9	14	0.062	0.928
10	6	0.026	0.955
11	3	0.013	0.968
12	5	0.022	0.991
13	1	0.004	0.995
14	1	0.004	0.999
	224	0.999	

2-2 DISTRIBUTION DE DONNEES GROUPEES EN CLASSE

Exemple 3: Un service social désire se faire une estimation du type de clientèle qui entreprend des démarches auprès de lui.

Pour un tel exemple, un tableau non groupé serait volumineux. On groupe alors les effectifs en classes. (Ce sera le cas pour toute variable continue.)

Chaque classe sera représentée par une valeur centrale X_i qui représentera dans les calculs les valeurs observées de la classe.

Pour chaque classe, on devra connaître l'étendue ou amplitude de la classe.Celle-ci vaut la différence des valeurs extrêmes de cette classe.

On définit ici comme en 2-1, la notion de fréquence absolue (n_i), de fréquence relative (f_i) et de fréquence relative cumulée (φ_k) ,

$$f_i = \frac{n_i}{n} \qquad \left(\sum_{i=1}^{p} f_i = 1 \right)$$

$$\varphi_k = \sum_{i=1}^{k} f_i = \frac{1}{n} \sum_{i=1}^{k} n_i$$

classes des âges de la clientèle x_i	centre de la classe X_i	n_i	f_i	φ_i
20]	10	5	0.005	0.005
]20 , 25]	22.5	19	0.019	0.024
]25 , 30]	27.5	37	0.037	0.061
]30 , 40]	35	72	0.073	0.134
]40 , 50]	45	114	0.115	0.249
]50 , 60]	55	151	0.152	0.401
]60 , 65]	62.5	177	0.178	0.579
]65 , 70]	67.5	200	0.201	0.780
]70	80	218	0.220	1
		993	1	

Exemple 4: Mesure de la capacité thoracique d'un échantillon de 50 êtres humains.

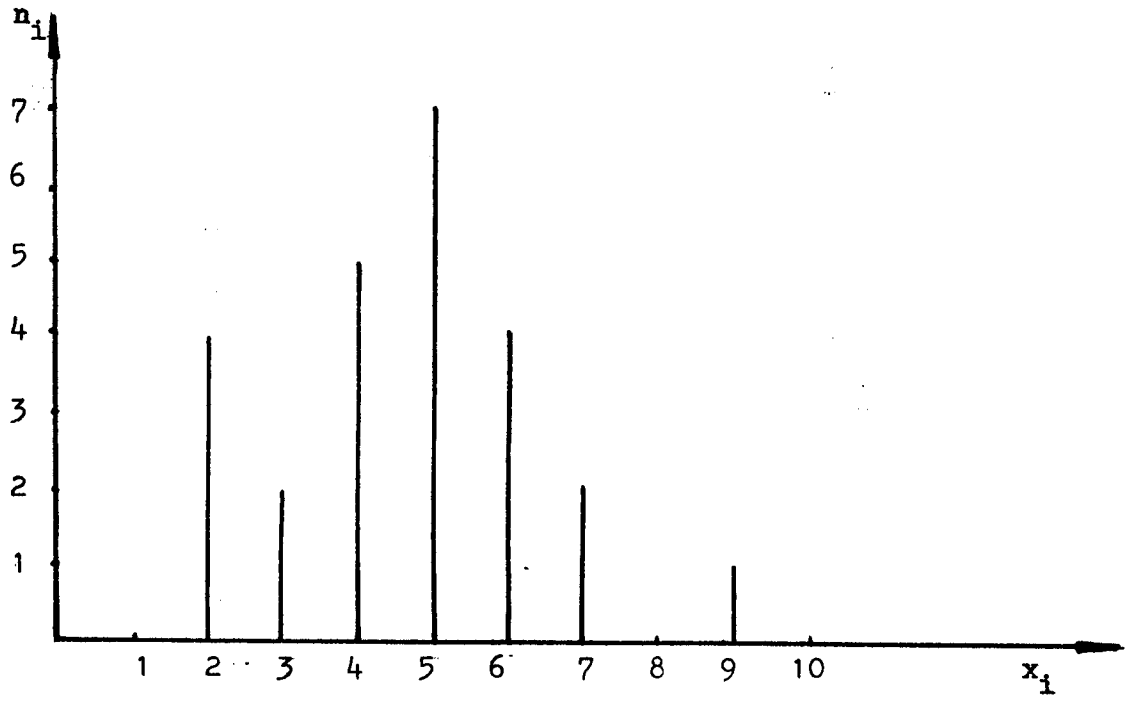
classes	X_i	n_i	f_i	φ_i
2400 cm^3 -2699 cm^3	2550	3	0.06	0.06
2700 -2999	2850	1	0.02	0.08
3000 -3299	3150	6	0.12	0.20
3300 -3599	3450	12	0.24	0.44
3600 -3899	3750	14	0.28	0.72
3900 -4199	4050	8	0.16	0.88
4200 -4499	4350	4	0.08	0.96
4500 -4799	4650	1	0.02	0.98
4800 -5099	4950	0	0	0.98
5100 -5399	5250	1	0.02	1
		50	1	

3- Représentation graphique d'une distribution

3-1 DISTRIBUTION NON GROUPEE

Les graphiques les plus souvent utilisés dans le cas de distibutions non groupéees sont
- le diagramme en bâtonnets
- le polygone des fréquences relatives ou absolues
- le diagramme des fréquences relatives cumulées

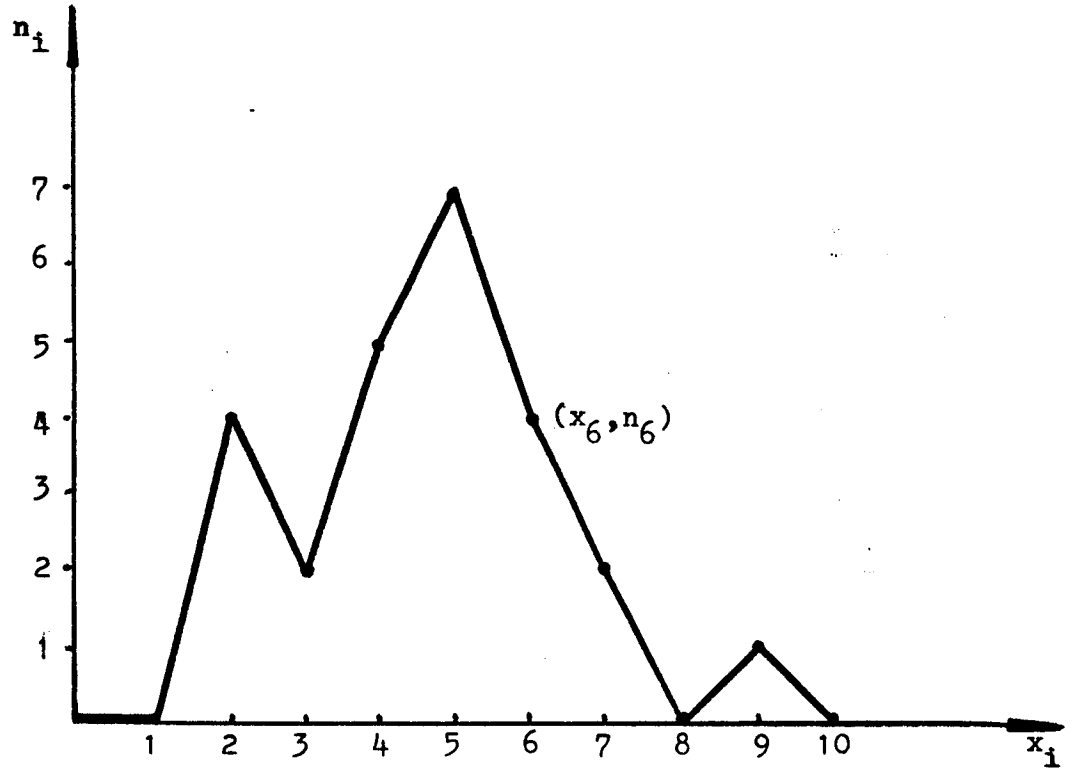
Exemple 1 : diagramme en bâtonnets.



Le bâtonnet correspondant à la valeur x_i de la valeur observée a une longueur n_i. La somme des longueurs des bâtonnets vaut donc n.

Si on porte en ordonnée les fréquences relatives (f_i) à la place des fréquences absolues (n_i) , on obtient le diagramme en bâtonnets des fréquences relatives et dans ce cas la somme des longueurs des bâtonnets vaudra 1.

Polygone des fréquences absolues:
Polygone obtenu en joignant les sommets des bâtonnets du diagramme précédent. On rejoint donc les points (x_i, n_i) par des segments de droite.



Polygone des fréquences relatives

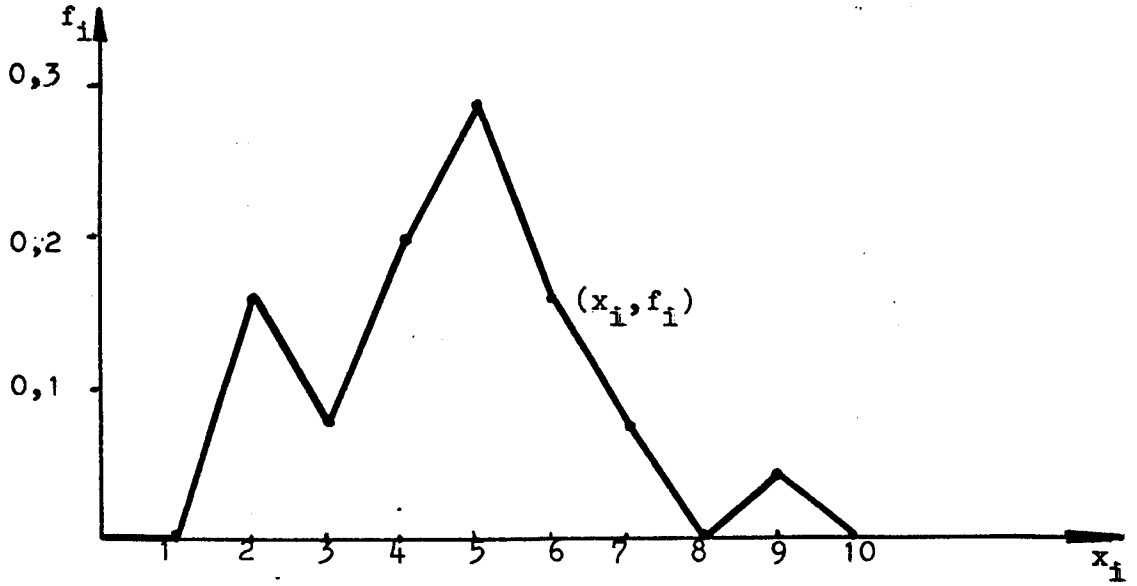
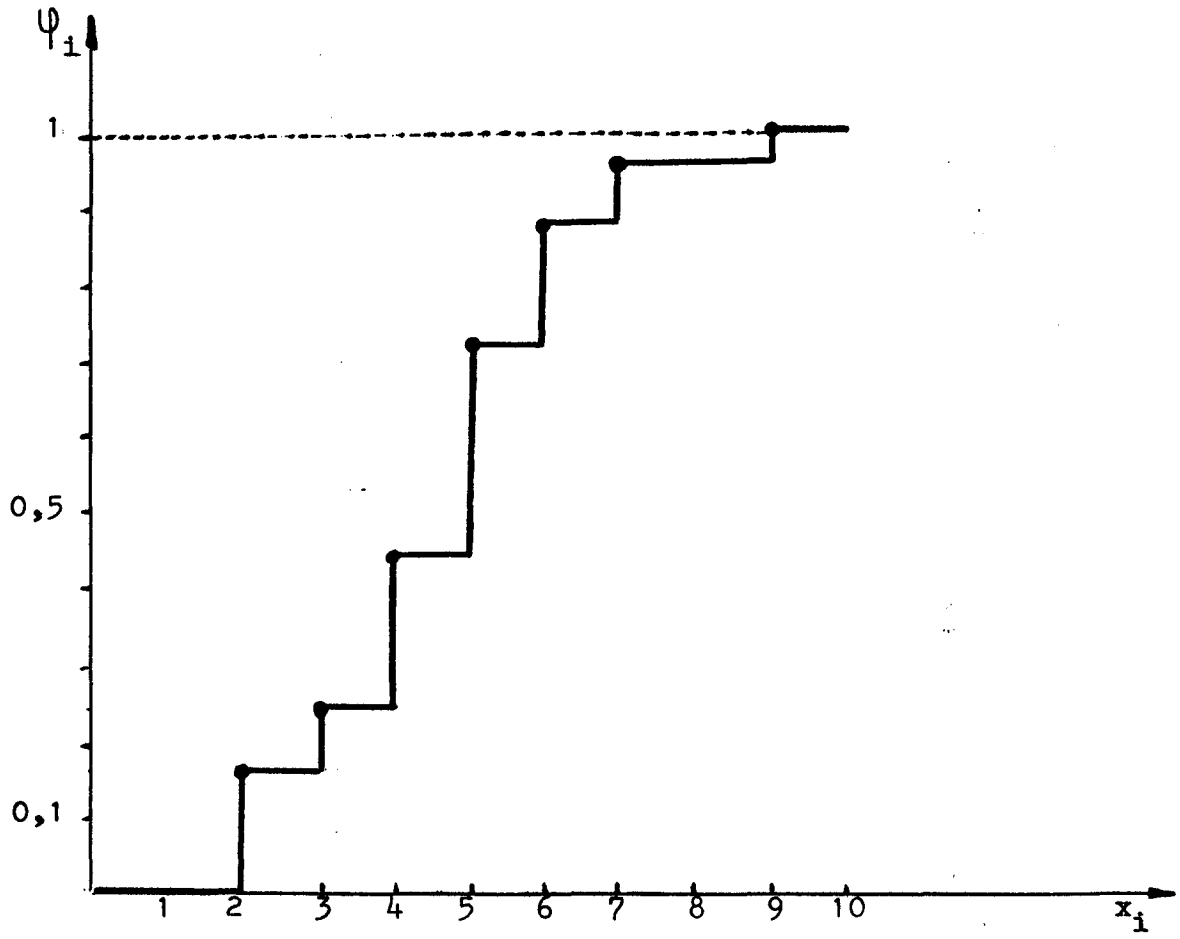


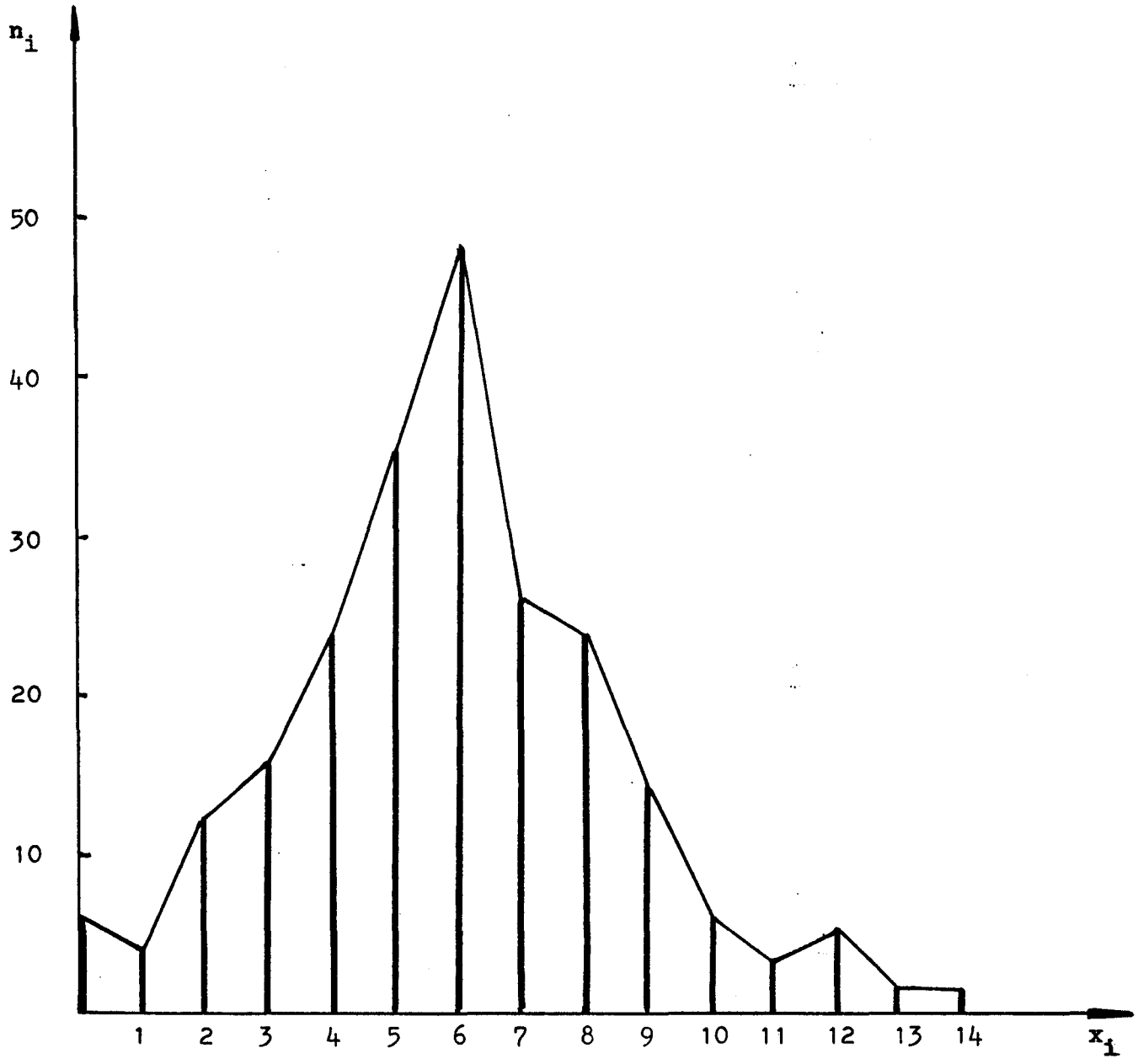
Diagramme des fréquences cumulées relatives:

On représente les points (x_i, φ_i). On trace un segment horizontal à partir de chaque point du graphique, vers la droite et ce jusqu'à arriver à l'aplomb de la valeur suivante de la variable. On obtient ainsi une fonction en escalier dont on dessine les contremarches.



Exemple 2

Diagramme en bâtonnets et polygone des fréquences absolues du nombre de petits pois par cosse.



Polygone des fréquences relatives du nombre de petits pois par cosse:

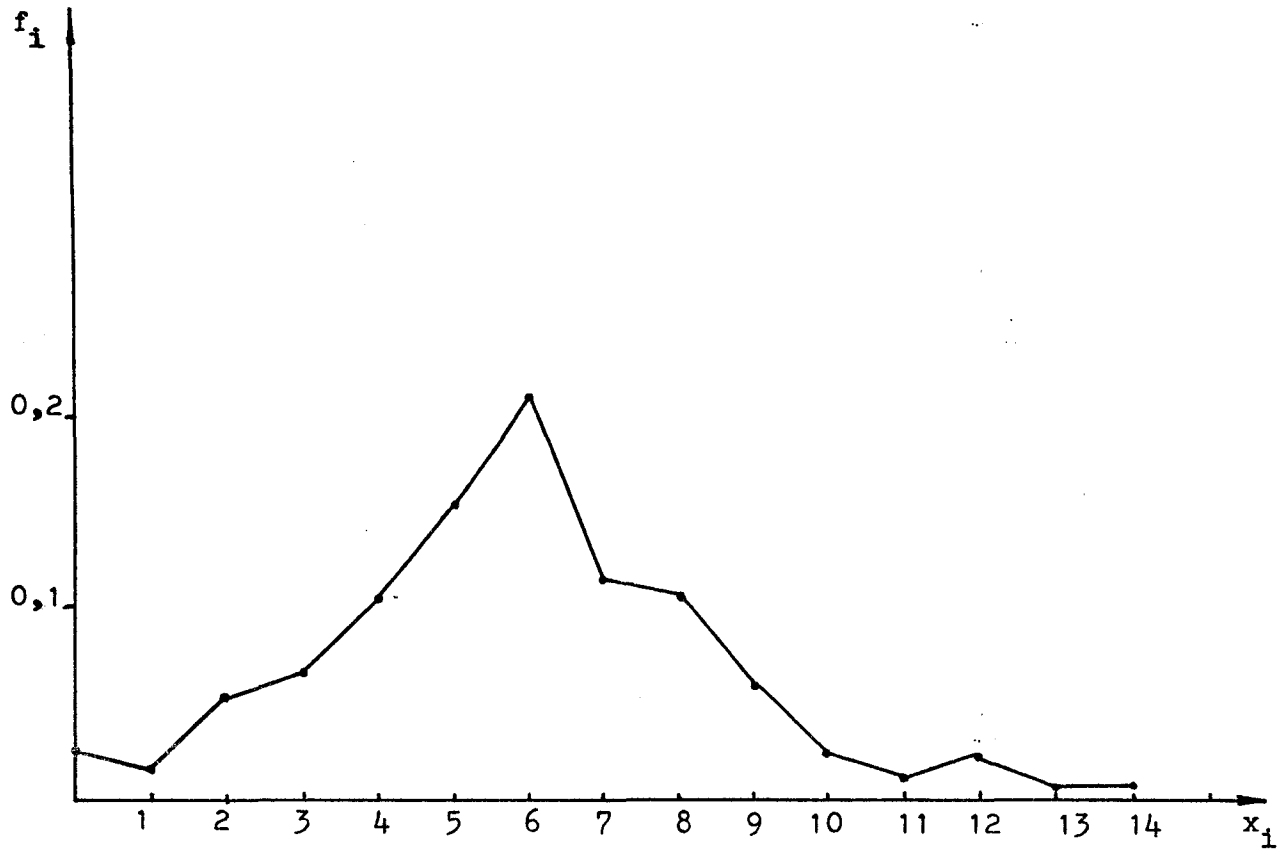
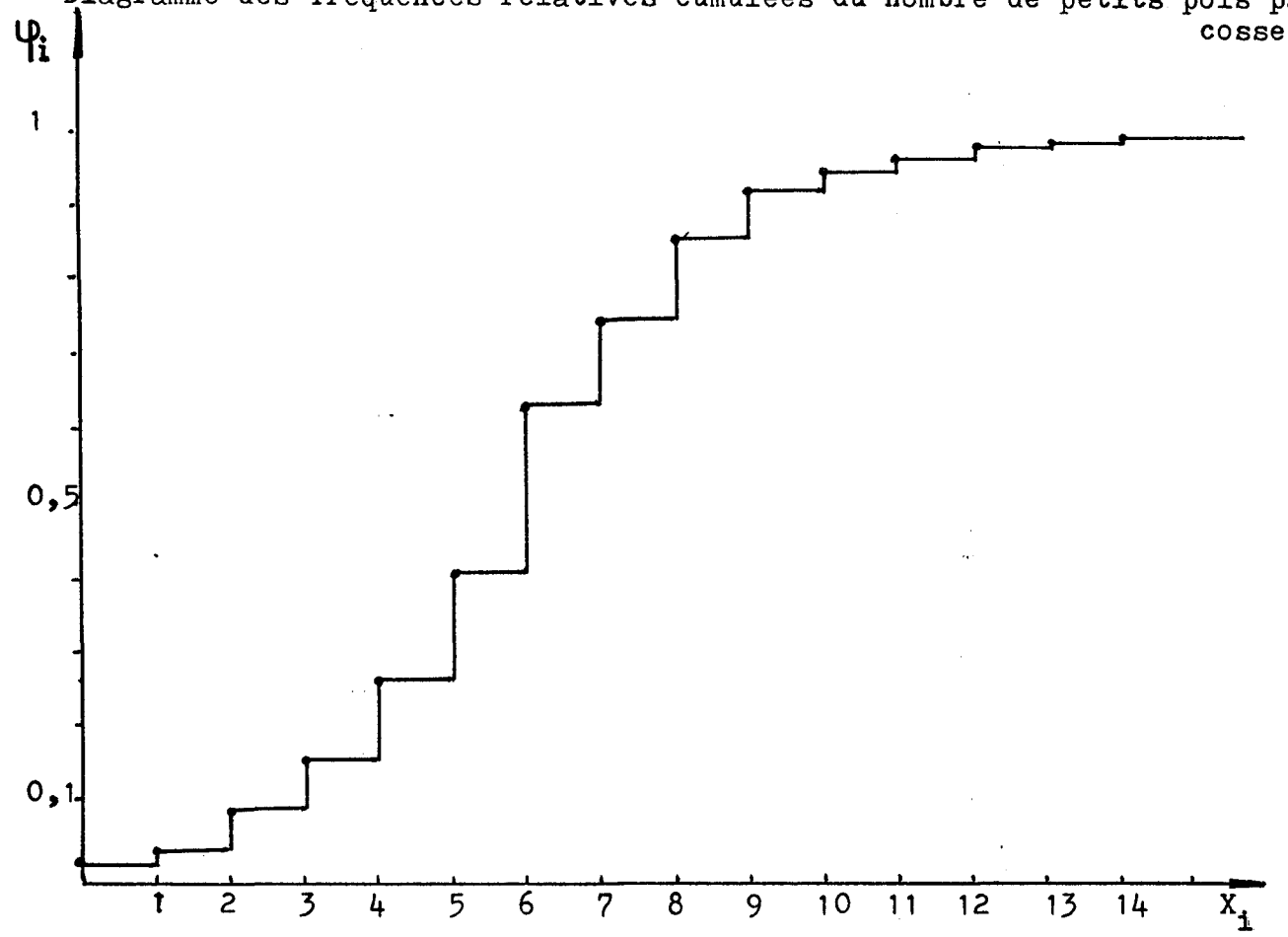


Diagramme des fréquences relatives cumulées du nombre de petits pois par cosse:

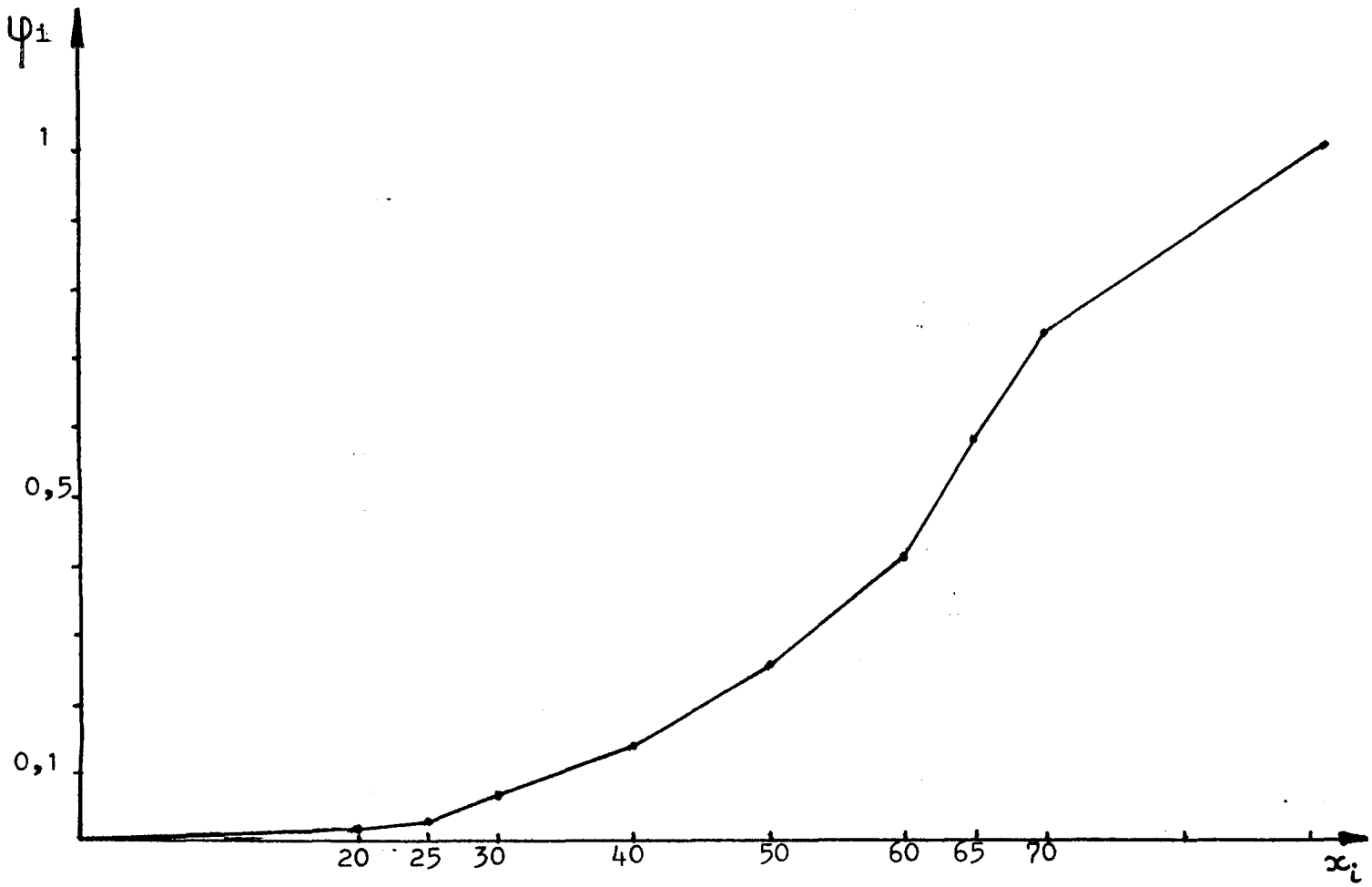


3-2 DISTRIBUTIONS GROUPEES EN CLASSES

Les graphiques les plus souvent utilisés pour les distributions groupées en classes sont - Les polygones des fréquences relatives cumulées
- Les histogrammes

Exemple 3 Polygone des fréquences relatives cumulées

On porte sur le graphique, les points (b_i, φ_i) où b_i est la borne supérieure de la classe i. On rejoint ensuite tous les points obtenus par des segments obliques.



Histogramme

Un histogramme est une représentation graphique au moyen de rectangles dont la base correspond à l'intervalle de la classe et dont l' aire a une mesure proportionnelle à la fréquence relative ou absolue de la classe.

A- l'intervalle de classe est le même pour toutes les classes

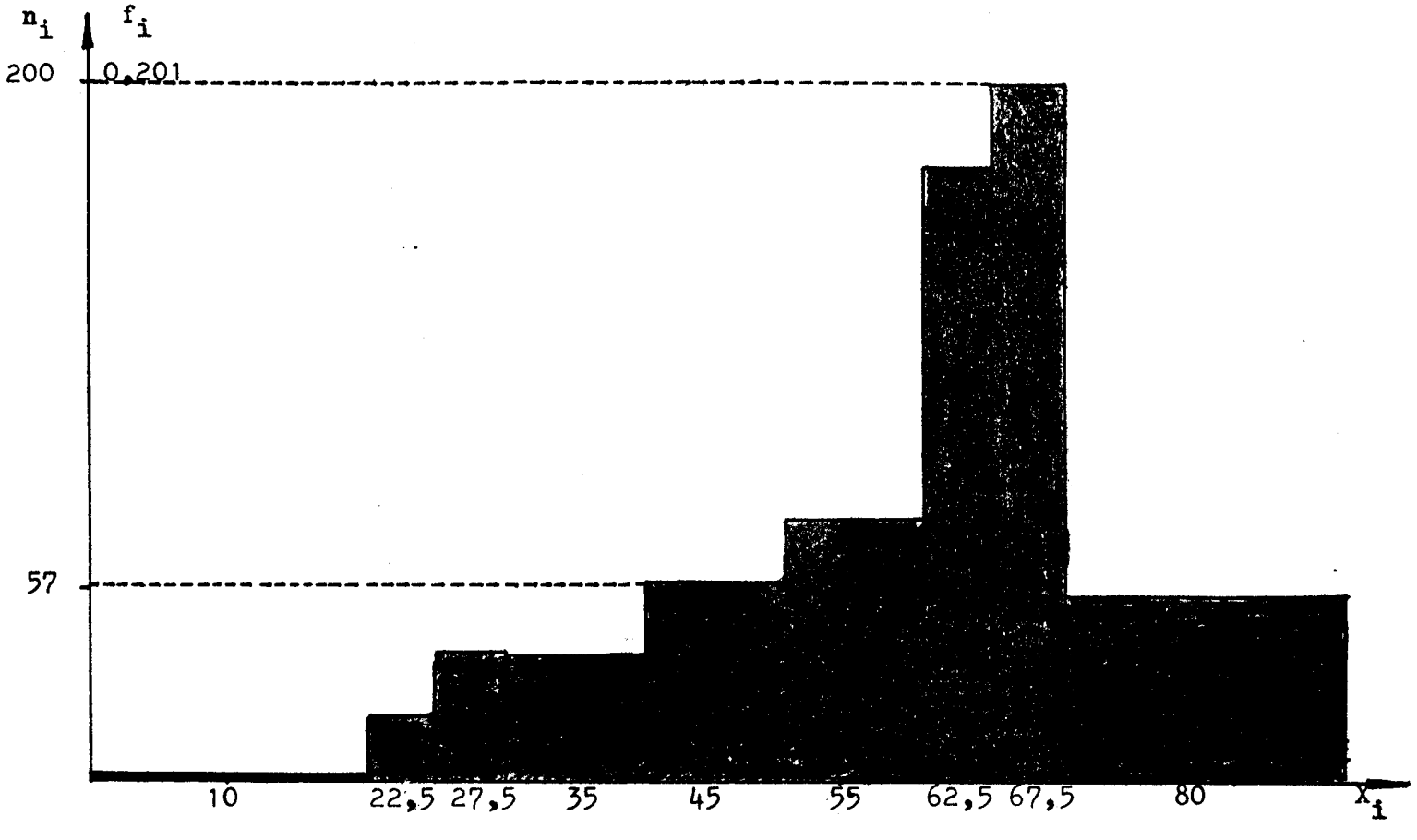
Dans ce cas, toutes les classes sont représentées par des rectangles ayant des bases égales et dont les hauteurs ont des mesures proportionnelles aux fréquences absolues (et donc aussi aux fréquences relatives).

Ce sera le cas dans l'exemple 4.

B- L'intervalle de classe est variable

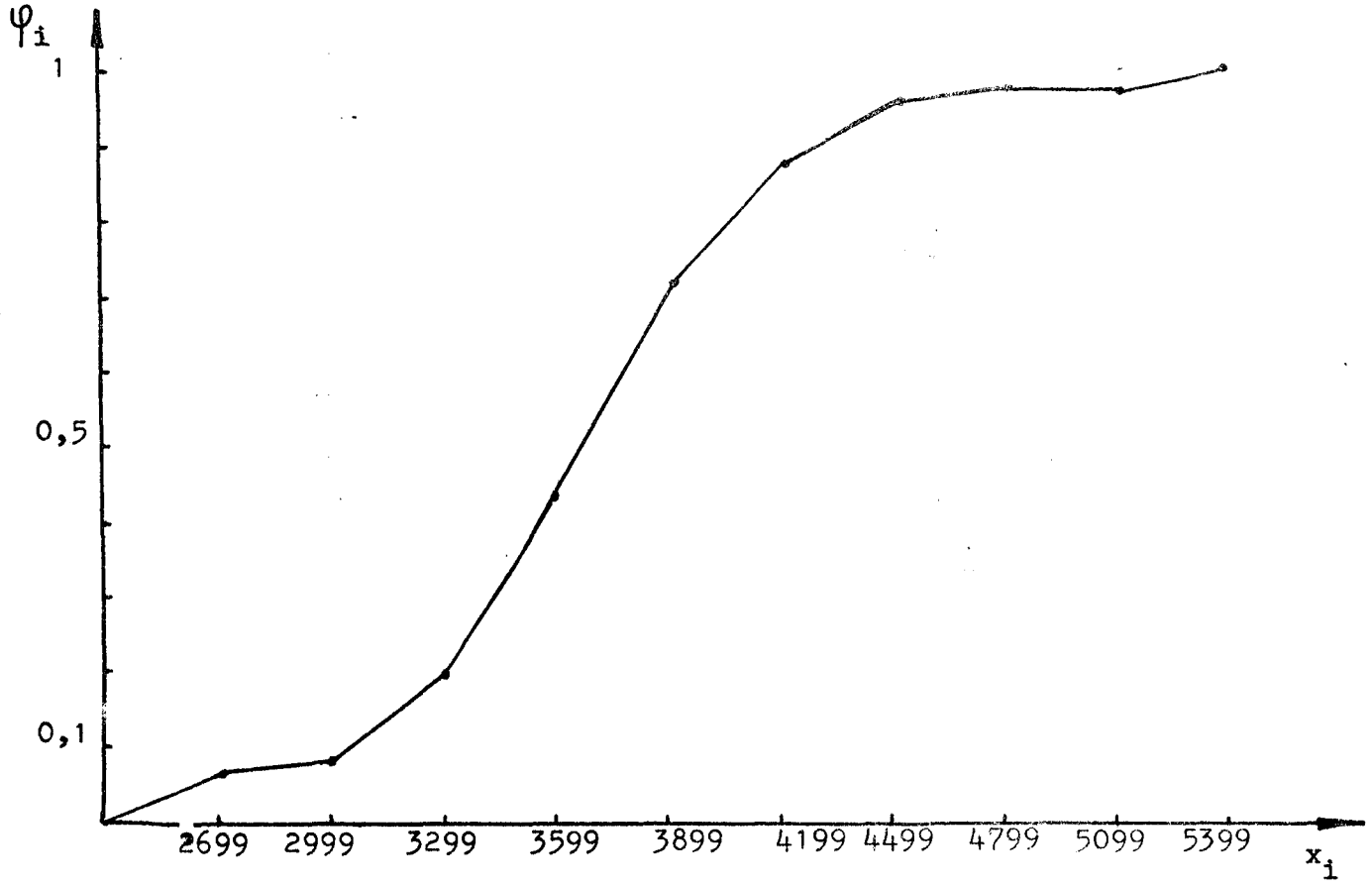
Puisque c'est l'AIRE du rectangle qui doit dépeindre la fréquence, ce serait alors une erreur de prendre une hauteur du rectangle proportionnelle à cette fréquence. La hauteur doit maintenant dépendre aussi de l'intervalle ou amplitude de chaque classe. Souvent, une certaine amplitude de classe se répète un grand nombre de fois et peut être considérée comme l'amplitude normale, tandis que certaines classes ont une autre amplitude, qui est en général un multiple de l'amplitude normale.

Dans ce cas, si l'amplitude d'une certaine classe est le double (le triple, le quadruple,...) de l'amplitude normale, on prendra comme hauteur du rectangle correspondant, à une certaine échelle, la moitié (le tiers, le quatrième,...) de la fréquence.

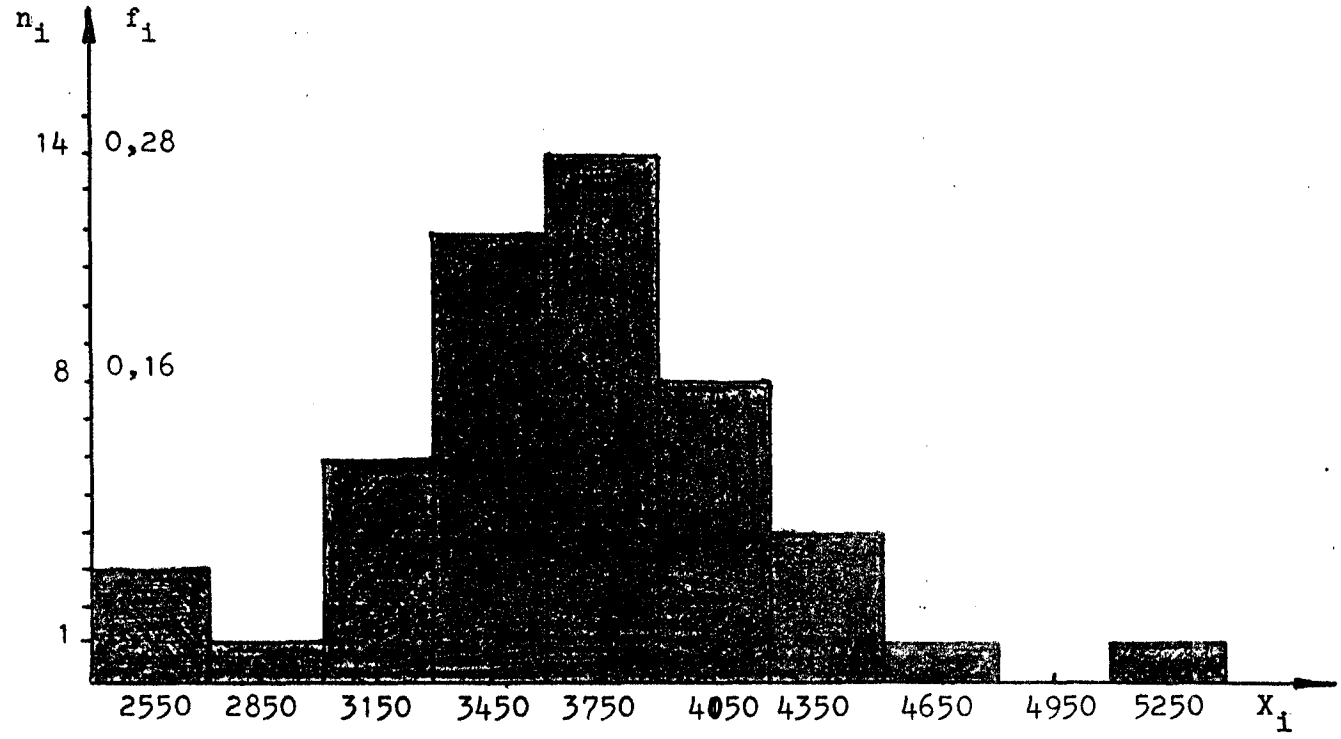


Exemple 4:

Polygone des fréquences relatives cumulées de la capacité thoracique



Histogramme



3-3 PREMIER CLASSEMENT DES DISTRIBUTIONS

On peut classer les distributions d'après l'allure de leur histogramme.

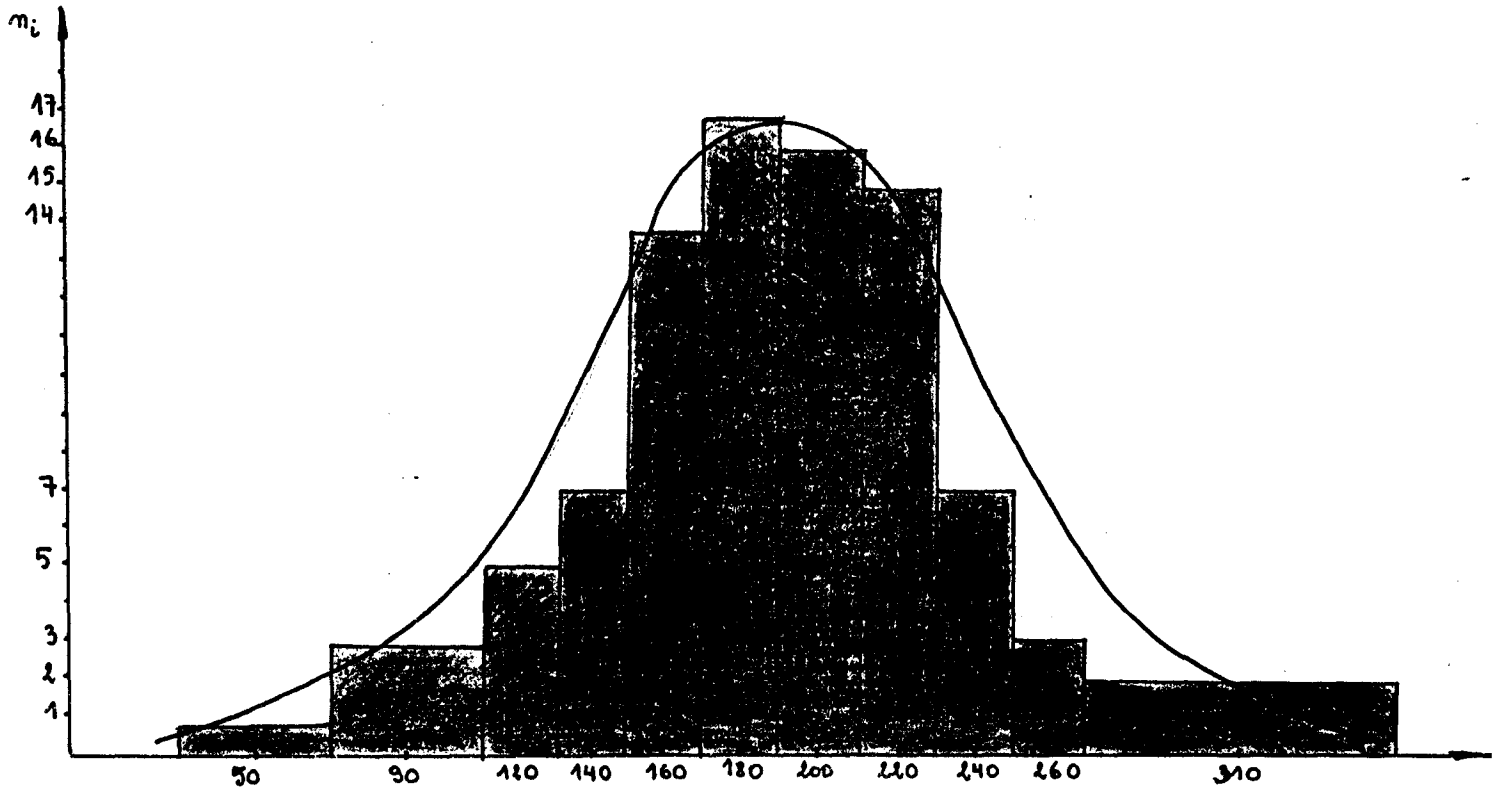
1- Distribution en cloche

L'histogramme peut s'ajuster à une courbe de Gauss (voir cours de 6^{e}) dont il présente les principaux caractère: forme en cloche, symétrie, concentration des valeurs observées autour de la "moyenne".

Exemple 5:

Répartition de 10.000 arbres fruitiers d'après le nombre de fruits produits en une saison. Le nombre des arbres est exprimé en pourcentage pour simplifier les calculs.

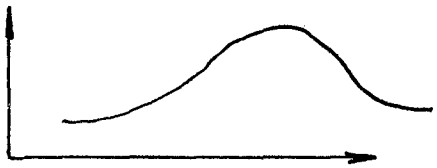
fruits x_i	n_i	X_i
30<70	2	50
70<110	6	90
110<130	5	120
130<150	7	140
150<170	14	160
170<190	17	180
190<210	16	200
210<230	15	220
230<250	7	240
250<270	3	260
270<350	8	310
	100	



2- Distribution dissymétrique

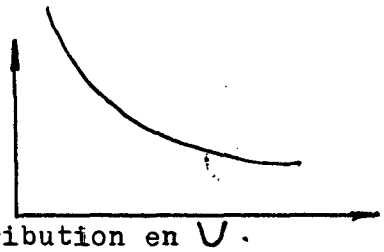
dissymétrie vers la gauche

(voir exemple 4)



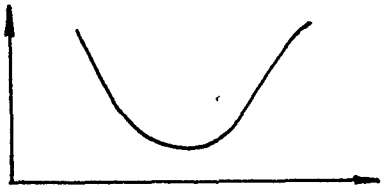
vers la droite

3- Disribution en L .



4- Distribution en J .

5- Distribution en U .

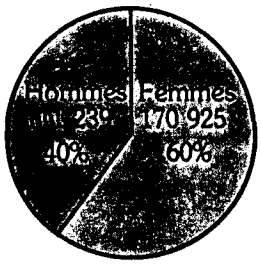


3-4 QUELQUES AUTRES METHODES DE REPRESENTATION GRAPHIQUE DE DISTRIBUTIONS

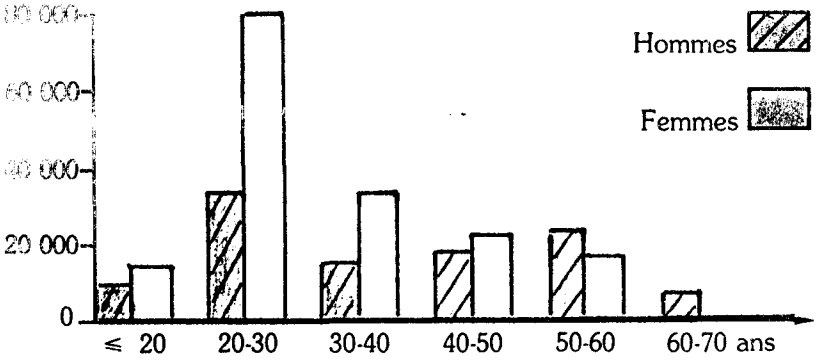
1- Le diagramme par secteurs

Le pays comptait 282 164 chômeurs complets (moyenne de 1978), soit ± 7,4 % de la population active.

Leur répartition par sexe



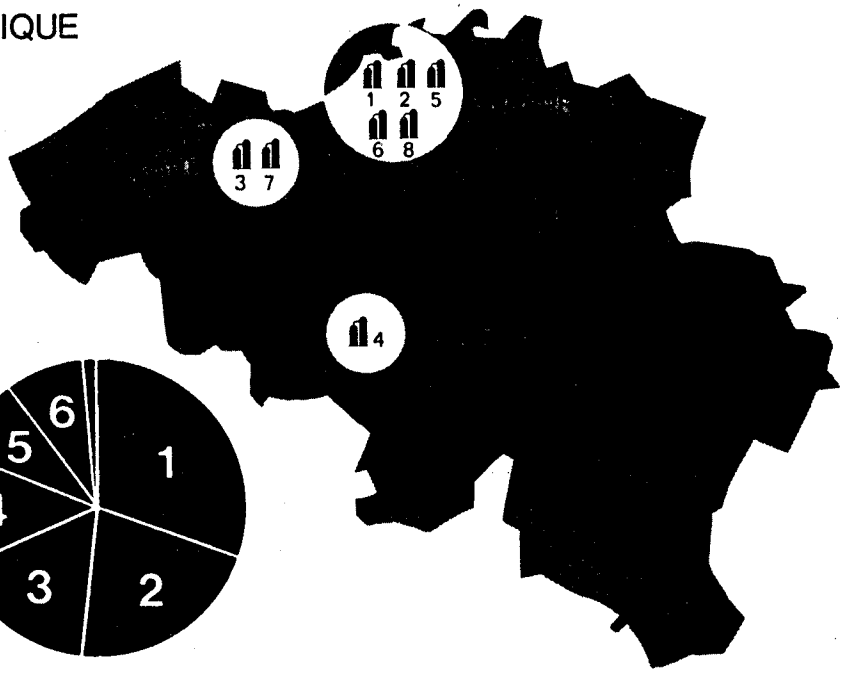
Leur répartition par âge et par sexe



Sources : L'Economie belge en 1978 - pp. 25 à 27
INS - Annuaire statistique de la Belgique 1978 - p. 6

Au 31 juillet 1979, le pays comptait 288 801 chômeurs complets

LES RAFFINERIES DE BELGIQUE ET LEUR CAPACITE ANNUELLE THEORIQUE DE RAFFINAGE - 1977 (en milliers de tonnes)



1. SIBP : 17 000
2. Esso : 12 000
3. Texaco Belgium : 9 370
4. Chevron Oil Belgium : 7 000
5. Albatros : 5 000
6. RBP : 5 000
7. Belgian Shell: 544
8. Anglo-Belges des Pétroles : 40

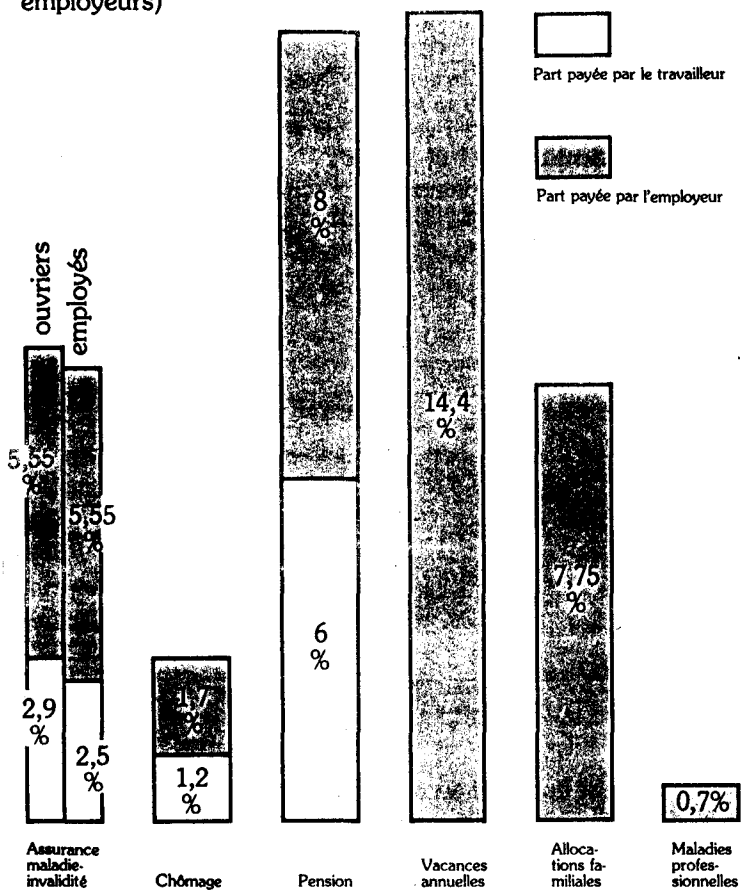
Total : 55 954

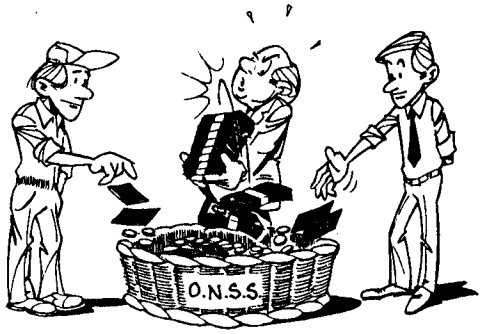
Statistiques de la Fédération Pétrolière Belge - 1968 à 1977 page 9

2 -Le diagramme par bandes ou tuyaux d'orgue

QUI FINANCE LES ASSURANCES SOCIALES ?

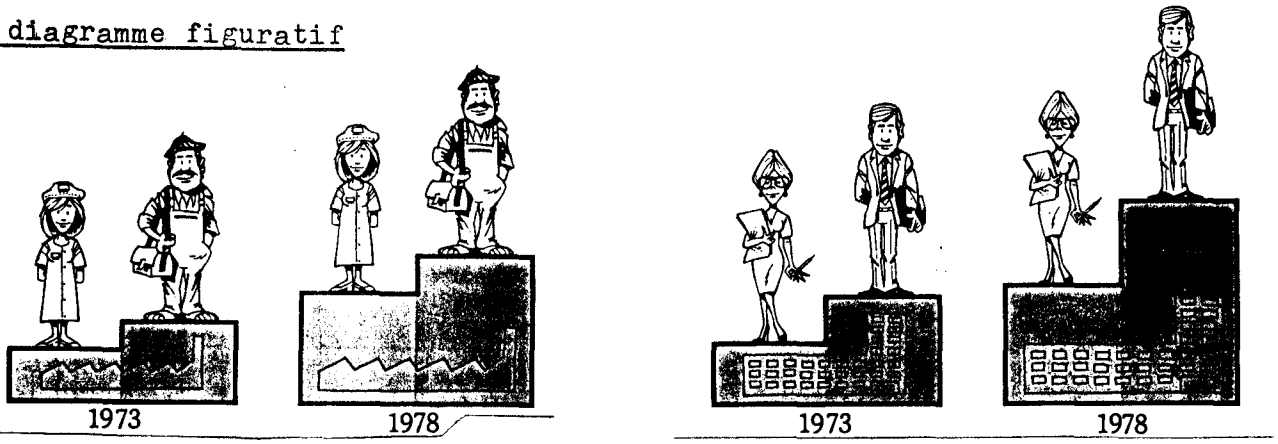
Répartition des cotisations (la part des salariés et celle des employeurs)





Source : Aperçu de la Sécurité sociale en Belgique - 1977 - p. 36

3- Le diagramme figuratif



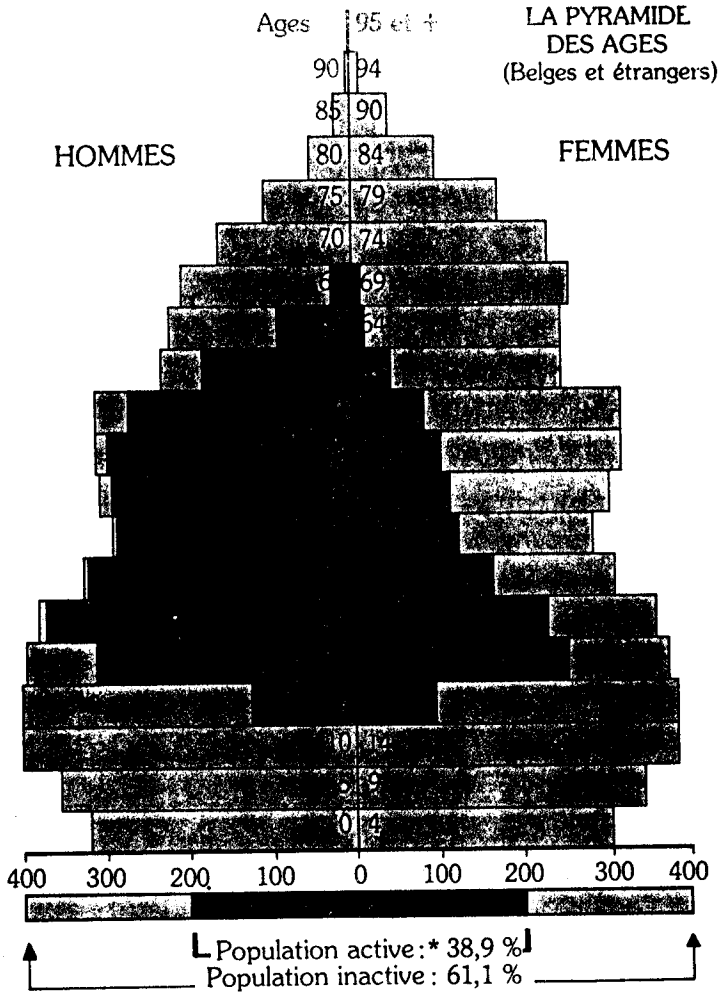
Gain horaire moyen de l'ouvrier et de l'employé dans l'industrie.

4- Par diagramme en pyramide

QUELLE EST LA POPULATION DE BELGIQUE ?

Le nombre total des **habitants de Belgique** s'élève à **9 823 000**.
Parmi ceux-ci on distingue : (chiffres de 1977)

- selon la nationalité	8 972 000 Belges	soit 91,5 %
	851 000 étrangers	soit 8,5 %
- selon le sexe	4 808 000 hommes	soit 49,- %
	5 015 000 femmes	soit 51,- %
- selon leur âge	2 144 000 jeunes (moins de 15 ans)	
	6 302 000 adultes (de 15 à 64 ans)	
	1 377 000 personnes âgées (plus de 65 ans)	



4- Paramètres d'une série statistique

Nous avons jusqu'ici décrit une série statistique par des graphiques. Dans ce qui suit, notre but sera de la caractériser par des nombres. On peut cataloguer ceux-ci en deux groupes:

- Les paramètres de position qui caractérisent la tendance centrale de la série.

Exemples: Le mode, la médiane, les moyennes ...

- Les paramètres de dispersion qui donnent une idée de l'étalement de la distribution de la série.

Exemples: L'amplitude ou étendue, la variance, les écarts, les quartiles, le coefficient de variation...

4-1 EMPLOI DU SIGNE $\sum$.

$$\sum_{i=1}^{K} x_i = x_1 + x_2 + x_3 + \ldots + x_K$$

$$\sum_{i=1}^{k} x_i^2 = x_1^2 + x_2^2 + x_3^2 + x_4^2 + \ldots + x_k^2$$

$$\sum_{i=1}^{k} a.x_i = a\sum_{i=1}^{k} x_i$$

$$\left(\sum_{i=1}^{k} x_i\right)^2 = (x_1 + x_2 + \ldots + x_k)^2$$

$$\sum_{i=1}^{k} x_i y_i = x_1 y_1 + x_2 y_2 + \ldots + x_k y_k$$

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

$$\sum_{i=1}^{k} (x_i + a) = \sum_{i=1}^{k} x_i + k.a$$

4-2 PARAMETRES DE POSITION OU DE TENDANCE CENTRALE

4-2-1 Le mode

Le mode est la valeur de la variable à laquelle correspond la fréquence la plus élevée . (Le mode est donc la valeur " à la mode ").

Pour une distribution non groupée, le mode est la valeur de la variable à laquelle correspond le point le plus élevé du diagramme en bâtonnets ou du polygone des fréquences.

Pour une distribution groupée en classes, le mode est la classe qui a la fréquence la plus élevée. On parle alors de classe modale. La classe modale se trouve aisément sur l'histogramme de la distribution.

Si on a des distributions dont la table des fréquences a deux (ou plusieurs) valeurs localement maximales, on parlera de distibutions bimodales (ou plurimodales). Ces distributions auront alors deux (ou plusieurs) modes ou classes modales .

Exemple 1 : Modes= 2 et 5
Exemple 2 : Mode = 6
Exemple 3 : Classe modale = 67,5
Exemple 4 : Classes modales = 2550 et 3750
Exemple 5 : Classe modale = 180

Avantages du mode comme paramètre de position:

- Il est aisé à déterminer et facilement interprétable
- Il n'est pas influencé par les valeurs occasionnelles ou aberrantes

Désavantages du mode comme paramètre de position:

- Il ne se prête pas aux traitements algébriques
- Il ne tient pas compte de toutes les données.

4-2-2 La médiane

La médiane d'une série statistique est la valeur de la variable qui satisfait à la condition que le nombre de valeurs de la série qui la dépassent aussi bien que le nombre de valeurs de la série qui lui sont inférieures sont, toutes les deux ,au plus égales à la moitié de l'effectif total de la série.

Médiane d'un distribution non groupée

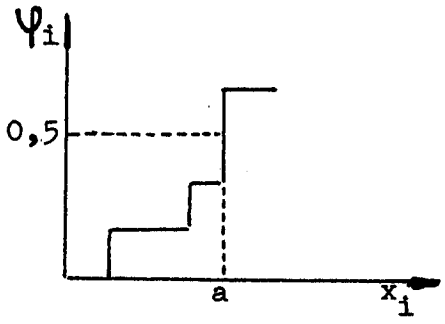
Si l'effectif (n) de la série est impair, la médiane est l'élément occupant le milieu de la série statistique ordonnée.

Si l'effectif de la série est pair, la médiane est par convention, la moyenne arithmétique des deux valeurs de la série répondant à la définition.

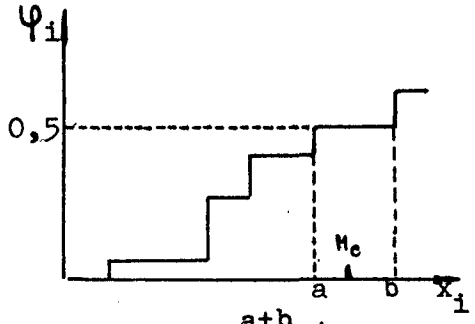
Exemple: $\{x_i\} = \{3,4, 4, 5, 6, 8, 8, 8, 10\}$ Me = 6

$\{x_i\} = \{5, 6, 8, 10, 12, 13\}$ $Me = \frac{8 + 10}{2} = 9$

Graphiquement, la médiane est la valeur de la variable correspondant à la fréquence 0,5 dans le diagramme des fréquences relatives cumulées.



Me = a



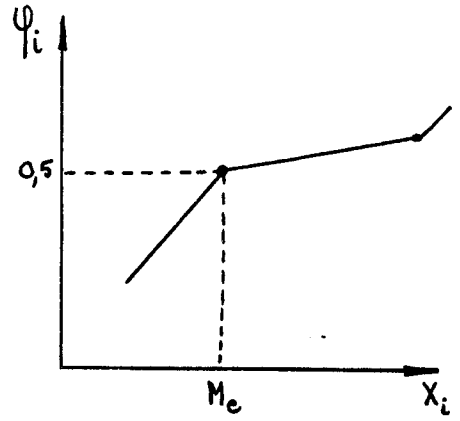
$Me = \frac{a+b}{2}$ (par convention)

Médiane d'une distribution groupée en classes

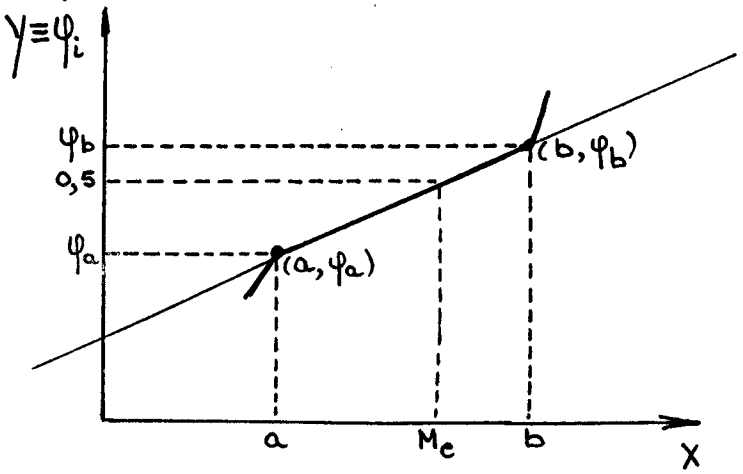
La médiane est l'abscisse correspondant à la verticale qui partage l'histogramme de la distribution en deux parties d'aires égales.

Dans le polygone des fréquences relatives cumulées, la médiane sera la valeur de la variable qui a comme ordonnée 0,5. Généralement on doit faire une interpolation linéaire qui suppose que les bornes de la classe médiane sont reliés par une droite. De cette manière, on obtient une estimation de la médiane:

1°)- Me est borne d'une classe: Pas de problème!



2)



$$D \equiv \frac{y-\varphi_a}{\varphi_b-\varphi_a} = \frac{x-a}{b-a} \quad (1)$$

Si $y = 0,5$; $X = M_e$

$$(1) \Rightarrow \frac{0,5-\varphi_a}{\varphi_b-\varphi_a} = \frac{M_e-a}{b-a}$$

$$\Rightarrow \boxed{M_e = \frac{0,5-\varphi_a}{\varphi_b-\varphi_a} \cdot (b-a) + a}$$

où a est la borne inférieure de la classe médiane

- φ_a est la somme des fréquences relatives de toutes les classes inférieures à la classe médiane
- $\varphi_b - \varphi_a$ est l'effectif de la classe médiane $= f_b$
- b-a est l'amplitude de la classe médiane

Cette interpolation est valable si la classe modale est de faible amplitude et si la répartition des variables y est uniforme.

Exemple 1 : Me = 5

Exemple 2 : Me = 6

Exemple 3 : $Me = \frac{0.5 - 0.401}{0.579 - 0.401} \cdot (65 - 60) + 60 = 62,8$

Exemple 4 : $Me = \frac{0.5 - 0.44}{0.72 - 0.44} \cdot (3899 - 3599) + 3599 = 3663$

Avantages de la médiane comme paramètre de position:

- Elle est bien définie, facilement interprétable et aisée à déterminer
- Elle n'est pas influencée par les cas aberrants

Inconvénients de la médiane comme paramètre de position:

- Elle ne se prête pas à des traitements algébriques
- Les fluctuations dues au hasard entre les médianes de différents échantillons extraits de la même population sont assez larges.

4-2-3 La moyenne arithmétique $\overline{X}$

1- La moyenne arithmétique d'une série de nombres est le quotient de la somme des éléments de cette série par l'effectif de la série.

- Série statistique donnée en un tableau brut, d'effectif n:

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

- Série statistique donnée en un tableau ordonné et recensé en p classes:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{p} n_i x_i = \sum_{i=1}^{p} f_i \cdot x_i$$

- Série statistique ayant une distribution de données groupées en classes:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{p} n_i X_i = \sum_{i=1}^{p} f_i \cdot X_i$$

On suppose donc que toutes les variables observées de la classe ont même valeur que celle du centre de la classe. Il est à noter que pour les distributions ayant une classe ouverte, la moyenne arithmétique n'existe pas! C'est le cas de l'exemple numéro 3. Dans cet exemple, on peut néanmoins considérer que 10 est le centre de la classe 20] et que 80 est le centre de la classe]70. Mais c'est omettre systématiquement toutes les personnes de plus de 90 ans.

Exemple 1: $\overline{x} = 4{,}64$

Exemple 2: $\overline{x} = 5{,}8$

Exemple 3: $\overline{x} = 59{,}9$ (contestable)

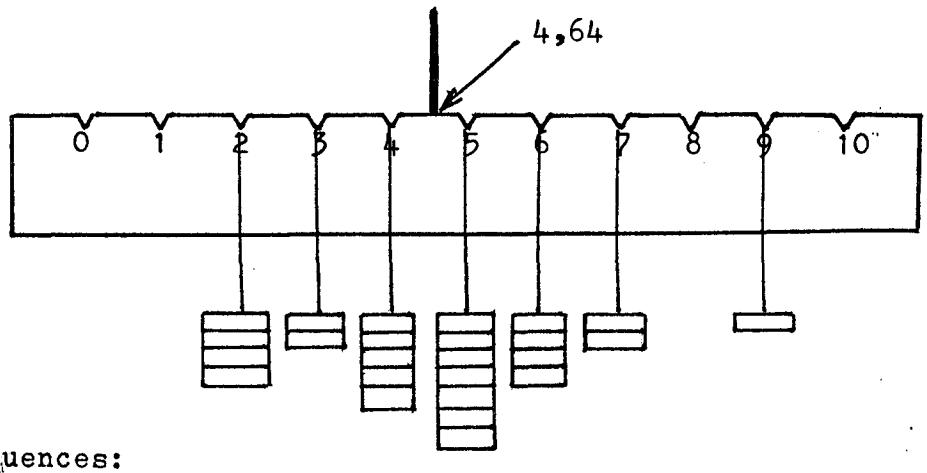
Exemple 4: $\overline{x} = 3660$

Exemple 5: $\overline{x} = 189{,}6$

2- Signification physique de la moyenne arithmétique

On représente l'échelle des valeurs observées par une barre de poids négligeable munie de crans qui correspondent aux différentes valeurs.

Exemple 1:



fréquences:

0 0 4 2 5 7 4 2 0 1 0

On représente les fréquences par des poids qui leur sont proportionnels, par exemple 7 N pour la fréquence 7.

On constate que la barre est en équilibre au point qui représente la moyenne arithmétique. Le théorème des moments est en effet vérifié:

$$2{,}64 \cdot 4 + 1{,}64 \cdot 2 + 0{,}64 \cdot 5 = 17{,}04$$

et $$0{,}36 \cdot 7 + 1{,}36 \cdot 4 + 2{,}36 \cdot 2 + 4{,}36 \cdot 1 = 17{,}04$$

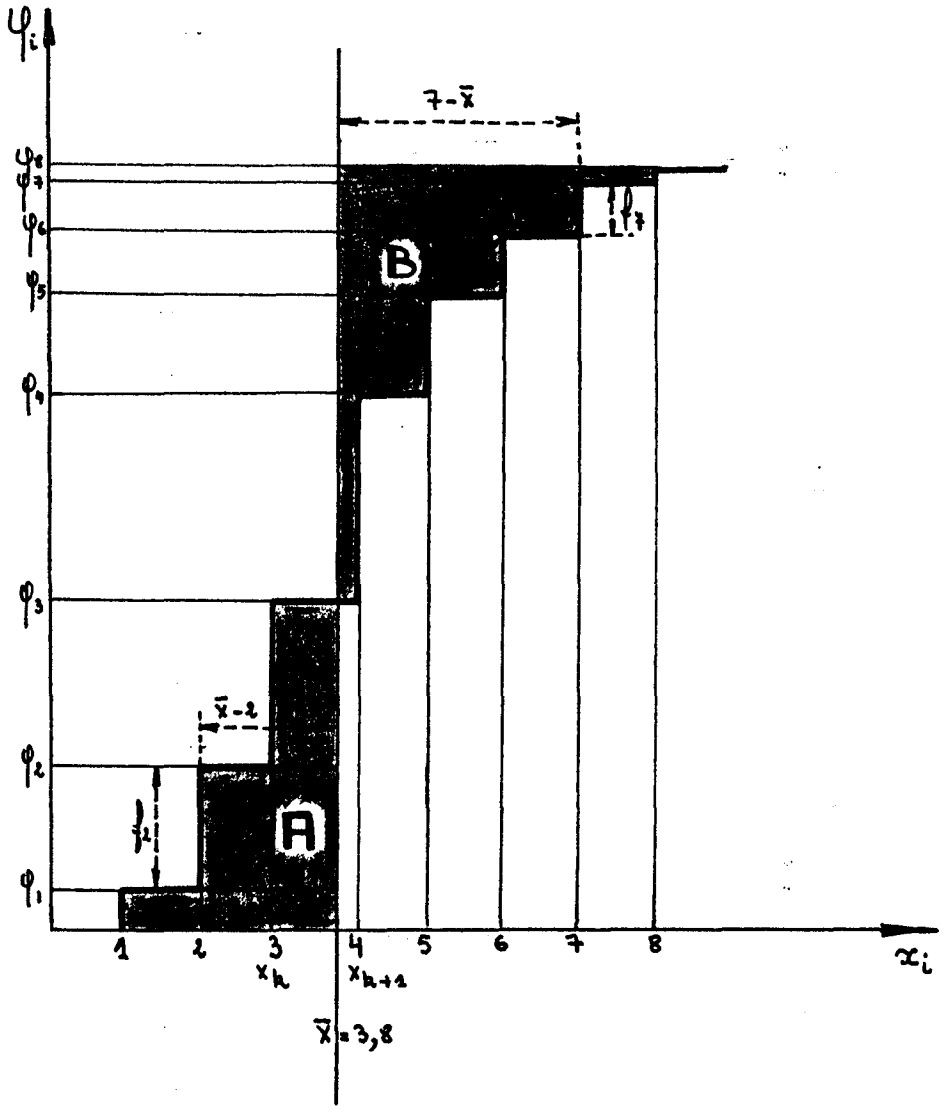
La moyenne arithmétique constitue donc le centre de masse des données lorsque celles-ci sont matérialisées par des objets pesants.

3- Signification géométrique de la moyenne arithmétique.

La verticale élevée de $\overline{x}$ dans le diagramme des fréquences relatives cumulées délimite deux aires égales:

Exemple 6: Distribution des logements d'une ville suivant le nombre de pièces habitables

x_i	f_i	φ_i
1	0,05	0,05
2	0,16	0,21
3	0,28	0,49
4	0,22	0,71
5	0,13	0,84
6	0,08	0,92
7	0,06	0,98
8	0,02	1



Aire de A = aire de B

Démonstration: Si $x_k < \bar{x} < x_{k+1}$

$$\text{Aire de A} = (\bar{x}-x_1).f_1 + (\bar{x}-x_2).f_2 + \ldots + (\bar{x}-x_k).f_k$$

$$= \bar{x}\sum_{i=1}^{k} f_i - \sum_{i=1}^{k} x_i f_i$$

$$\text{Aire de B} = (x_{k+1}-\bar{x}).f_{k+1} + (x_{k+2}-\bar{x}).f_{k+2} + \ldots + (x_p-\bar{x}).f_p$$

$$= \sum_{i=k+1}^{p} x_i f_i - \bar{x}\sum_{i=k+1}^{p} f_i$$

$$= \sum_{i=k+1}^{p} x_i f_i - \bar{x}.(1-\sum_{i=1}^{k} f_i)$$

$$= \sum_{i=k+1}^{p} x_i f_i - \bar{x} + \bar{x}\sum_{i=1}^{k} f_i$$

$$= \sum_{i=k+1}^{p} x_i f_i - \sum_{i=1}^{p} x_i f_i + \bar{x}\sum_{i=1}^{k} f_i$$

$$= -\sum_{i=1}^{k} x_i f_i + \bar{x}\sum_{i=1}^{k} f_i$$

$$= \text{aire de A}$$

Une démonstration analogue montre qu'on a la même propriété pour le polygone des fréquences relatives cumulées.

4- Changement de variable

Pour faciliter les calculs on peut être amené à effectuer un changement de variable du type

$$\boxed{x'_i = a.x_i + b} \qquad a, b \in \mathbb{R}$$

Le tableau d'effectifs (x_i, n_i) sera alors remplacé par un tableau (x'_i, n_i). Si $\bar{x}$ est la moyenne de la première série et $\bar{x}'$ celle de la série après changement de variable, on a:

$$\boxed{\bar{x}' = a.\bar{x} + b}$$

Démonstration: $\overline{x'} = \sum_{i=1}^{p} x'_i f_i = \sum_{i=1}^{p} (ax_i + b) f_i$

$$= a \sum_{i=1}^{p} f_i x_i + b \sum_{i=1}^{p} f_i$$

$$= a\overline{x} + b$$

Dans l'exemple 4, si on effectue le changement de variable

$$x'_i = \frac{x_i - 2550}{300}$$

la moyenne arithmétique est nettement plus aisée à calculer.

x_i	x'_i	n_i	$x'_i n_i$
2550	0	3	0
2850	1	1	1
3150	2	6	12
3450	3	12	36
3750	4	14	56
4050	5	8	40
4350	6	4	24
4650	7	1	7
4950	8	0	0
5250	9	1	9
		50	185

$$\overline{X'} = \frac{1}{50} \sum_{i=1}^{p} x'_i n_i = \frac{185}{50} = 3,70$$

$$3,70 = \frac{\overline{X} - 2550}{300} \Rightarrow \overline{x} = 300 \cdot 3,70 + 2550 = 3660$$

5- Avantages et inconvénients de la moyenne arithmétique

Avantages:

- Elle est aisée à calculer, bien définie et facilement interprétable.
- Elle se prête bien aux traitements algébriques
- Elle met en jeu les valeurs de toutes les données
- elle est la même dans les échantillons extraits d'une même population, aux fluctuations dues au hasard près.

Inconvénients

- Elle est fort influencée par les données extrêmes, surtout si le nombre de données n'est pas grand. (Un élève qui remet une copie blanche fait artificiellement diminuer la moyenne de la classe)
- Elle perd sa signification quand certaines données sont indifférenciées (Un zéro peut être obtenu pour différentes raisons)
- L'interprétation de la moyenne est moins évidente quand la distribution n'est pas symétrique.

4-2-4 Que choisir ... mode, médiane, moyenne arithmétique?

Le mode, la médiane et la moyenne arithmétique sont les paramètres de tendance centrale les plus souvent utilisés. Ils sont confondus dans les distributions unimodales symétriques. Dans les autres cas, pour décider lequel parmi ceux-ci peut le mieux représenter la distribution, il faut d'abord bien se poser le problème.

1- Peut-on compenser ce qu'il manque par ce qu'il y a en trop?

le déficit	par le gain
les lacunes	par les excédents

si oui, on recherche $\bar{x}$.

Exemple : Un représentant de commerce qui demande à son employeur une somme forfaitaire mensuelle pour ses déplacements en voiture.

2- Faut-il réduire au maximum les écarts?

On a alors recourt à la médiane ou au mode:

- A la médiane, si c'est la valeur de la différence qui importe
  ex.: Si on veut fabriquer un emballage valable pour différents articles et que chaque cm de carton coûte 1F, on choisira la médiane comme emballage standard.
- Au mode, si c'est le nombre de cas différents qui importe.
  ex.: Même exemple que ci-dessus , mais on doit prendre en considération que chaque retouche ou manipulation supplémentaire quelle qu'elle soit coûte 5 F.

D'autre part, on utilisera la médiane plutôt que la moyenne arithmétique dans le cas où la plupart des valeurs observées sont assez voisines, quelques-unes seulement s'en écartant beaucoup .

Exemple: Dans une entreprise qui compte 1000 ouvriers dont les salaires sont assez voisins et une vingtaine de techniciens dont le salaire est nettement supérieur, on voit que ces derniers n'influenceront pas la médiane des salaires alors qu'ils modifieront beaucoup la moyenne arithmétique.

Le caractère sommaire des paramètres de position est corrigé par les paramètres de dispersion qui permettent d'estimer à quel point n'importe quel élément de l'ensemble s'écarte en valeur de la mesure de tendance centrale choisie.

Il existe d'autres moyennes qui sont plus rarement utilisées par le statisticien :

4-2-5 La moyenne géométrique

La moyenne géométrique de n nombres est la racine n^e du produit de ces n nombres

$$m_g = \sqrt[n]{x_1 x_2 x_3 \dots x_n}$$

la moyenne géométrique est surtout utilisée lorsqu'on a affaire à des valeurs qui croissent fort.

Exemple: En 1950, il y avait dans notre pays 270000 voitures environ
En 1970, il y en avait environ 2 millions!

$$\overline{x} = (270000 + 2000000) / 2 = 1135000$$

$$m_g = \sqrt{270000 \cdot 2000000} = 734000$$

Quel est le meilleur résultat sachant que la valeur véritable pour 1960 était de 753136?

4-2-6 La moyenne harmonique

La moyenne harmonique de n nombres est l'inverse de la moyenne arithmétique de l'inverse de ces n nombres.

$$m_h = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \frac{1}{x_3} + \dots + \frac{1}{x_n}}$$

4-3 PARAMETRES DE DISPERSION

Il est à noter qu'à partir de maintenant, si on parle de moyenne, il s'agira de la moyenne arithmétique.

4-3-1 Exemple 7 :

Points obtenus par les élèves de deux classes différentes à un même contrôle.

Classe A

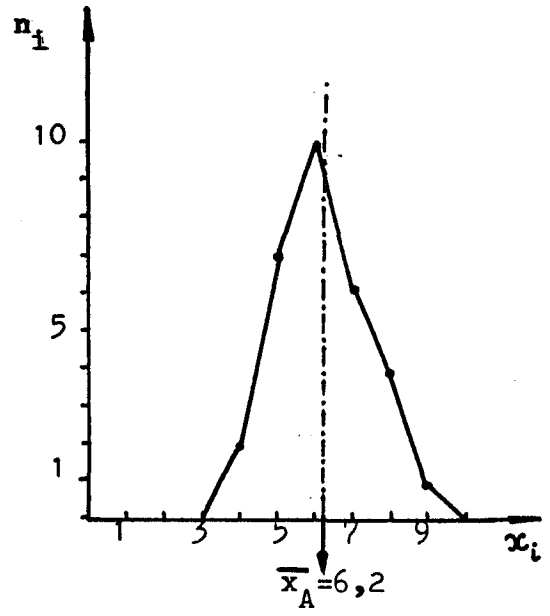
x_i	n_i	$n_i x_i$	φ_i
4	2	8	0,07
5	7	35	0,30
6	10	60	0,63
7	6	42	0,83
8	4	32	0,96
9	1	9	0,99
	30	186	

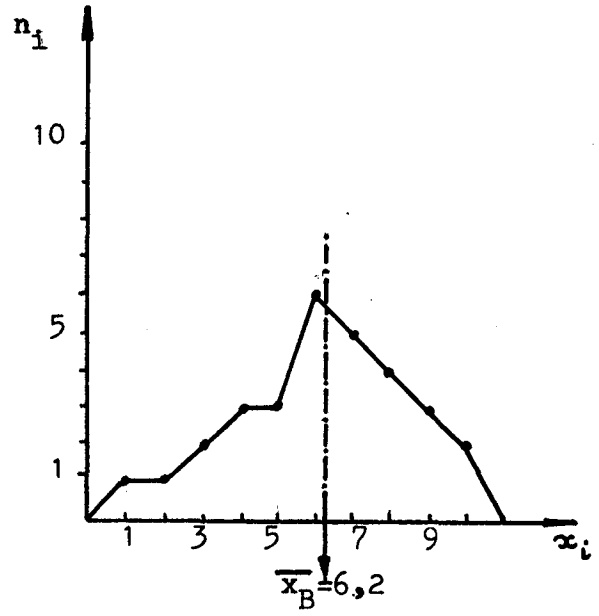
Classe B

x_i	n_i	$n_i x_i$	φ_i
1	1	1	0,03
2	1	2	0,06
3	2	6	0,12
4	3	12	0,22
5	3	15	0,32
6	6	36	0,52
7	5	35	0,69
8	4	32	0,82
9	3	27	0,92
10	2	20	0,99
	30	186	

$$\overline{x_B} = \overline{x_A} = \frac{186}{30} = 6,2$$

Les deux classes ont même effectif. La moyenne arithmétique est la même et pourtant, la différence entre les deux classes est grande. Dans la classe A, les résultats sont centrés autour de la moyenne, tandis qu'ils sont plus dispersés dans la classe B.





4-3-2 Etendue, amplitude ou intervalle de variation

L'étendue, l'amplitude ou l'intervalle de variation d'une distribution est la différence entre la plus grande et la plus petite des valeurs observées.

Dans l'exemple 7: Classe A: étendue = 9-4 = 5

Classe B: étendue = 10-1 =9

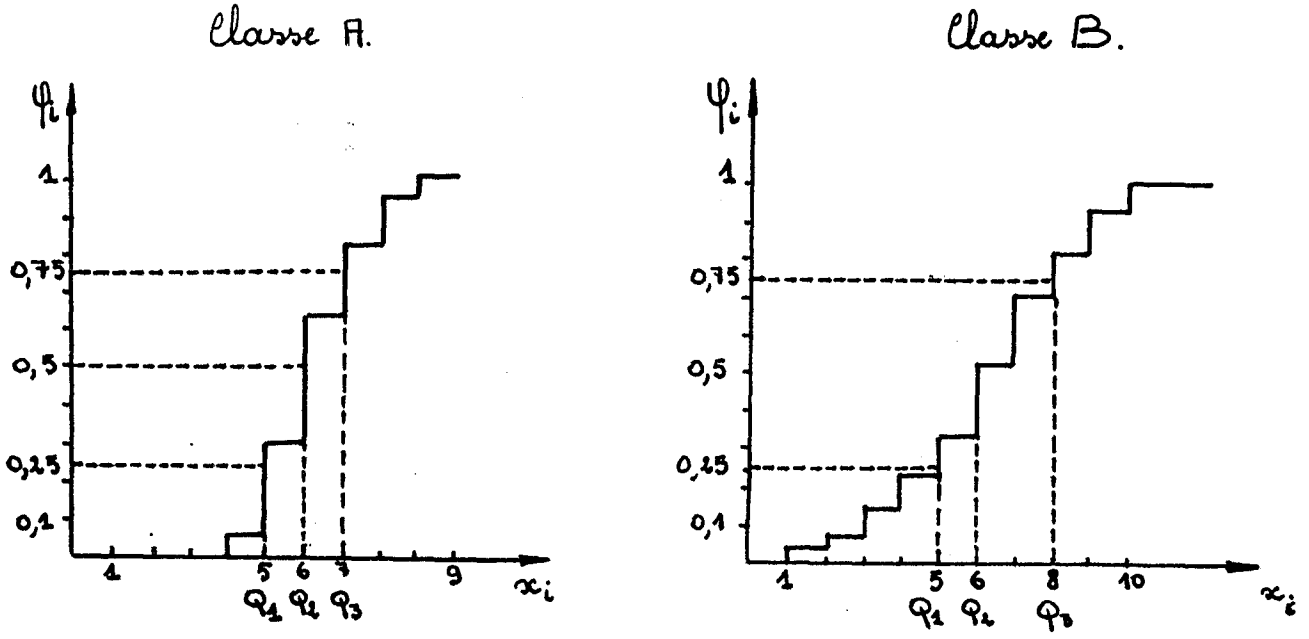
Ce qui signifie que la classe A est plus homogène que la classe B.

Ce paramètre n'est pas très significatif car il ne tient compte que de deux données. De plus dans les distributions groupées en classe, les valeurs extrêmes X_i sont souvent non définies; dans ce cas, l'étendue de la distribution n'a pas beaucoup de sens.

4-3-3 les quartiles Q_1, $M_e=Q_2$, Q_3

Les quartiles sont des paramètres qui divisent la série statistique en 4 parties égales. Un quart, au plus, des valeurs observées doivent être inférieures à Q_1 et les 3/4 ,au plus, des valeurs observées doivent lui être supérieures. Q_2 coïncide avec la médiane. Au plus les 3/4 des valeurs observées doivent être inférieures à Q_3 et au plus le quart des valeurs observées doivent lui être supérieures.

Dans le cas des distributions non groupées en classes, les quartiles Q_1, Q_2, Q_3 sont les valeurs des variables correspondant aux fréquences 0,25 , 0,5 , 0,75 dans le diagramme des fréquences relatives cumulées de la distribution



Dans le cas des distributions <u>groupées</u> en classes, Q_1, Q_2, Q_3 sont les abscisses des points qui ont respectivement 0,25 , 0,5 ,0,75 pour ordonnée dans le polygone des fréquences relatives cumulées. Pour trouver lesquartiles, on détermine d'abord les classes]a,b] dans lesquelles ils sont situés et on détermine leur valeur par interpolation linéaire comme dans le cas de la médiane page 23

$$Q_1 = \frac{0,25 - \varphi_a}{\varphi_b - \varphi_a} \cdot (b-a) + a \qquad \text{si } 0,25 \in]a,b]$$

$$Q_2 = M_e = \frac{0,5 - \varphi_c}{\varphi_d - \varphi_c} \cdot (d-c) + c \qquad \text{si } 0,5 \in]c,d]$$

$$Q_3 = \frac{0,75 - \varphi_k}{\varphi_l - \varphi_k} \cdot (l-k) + k \qquad \text{si } 0,75 \in]k,l]$$

Exemple 1: Q_1= , Q_2= M_e= , Q_3=
Exemple 2: Q_1= , Q_2= M_e= , Q_3=
Exemple 3: Q_1= , Q_2= M_e= , Q_3=
Exemple 4: Q_1= , Q_2= M_e= , Q_3=
Exemple 5: Q_1= , Q_2= M_e= , Q_3=
Exemple 6: Q_1= , Q_2= M_e= , Q_3=

4-3-4 Intervalle interquartille = $Q_3 - Q_1$

Cette valeur fournit une caractéristique de la dispersion qui n'est pas influencée par les valeurs extrêmes de la série. L'intervalle interquartille comprend la moitié des valeurs observées de la distribution qui se situent au centre de la distribution.

Exemple 1 : $Q_3 - Q_1$ =
Exemple 2 : $Q_3 - Q_1$ =
Exemple 3 : $Q_3 - Q_1$ =
Exemple 4 : $Q_3 - Q_1$ =
Exemple 5 : $Q_3 - Q_1$ =
Exemple 6 : $Q_3 - Q_1$ =
Exemple 7 : Classe A, $Q_3 - Q_1$ = Classe B, $Q_3 - Q_1$ =

4-3-5 Ecart-moyen absolu

On pourrait songer à caractériser la dispersion d'une distribution par la moyenne arithmétique de la somme des écarts de chaque valeur observée par rapport à la moyenne. Mais

$$\frac{1}{n}\sum_{i=1}^{p} n_i.(x_i-\bar{x}) = \frac{1}{n}\sum_{i=1}^{p} n_i x_i - \bar{x} = 0 \text{ !!!}$$

On évite l'annulation des écarts négatifs par les écarts positifs au moyen de valeurs absolues:

L'<u>écart-moyen absolu</u> d'une distribution est la moyenne arithmétique des valeurs absolues des différences entre la moyenne et les valeurs observées de la série $\frac{1}{n}\sum_{i=1}^{p} n_i.|x_i-\bar{x}|$

Exemple 7:

Classe A

x_i	n_i	$\lvert x_i-\bar{x}\rvert$	$n_i.\lvert x_i-\bar{x}\rvert$
4	2	2,2	4,4
5	7	1,2	8,4
6	10	0,2	2
7	6	0,8	4,8
8	4	1,8	7,2
9	1	2,8	2,8
	30		29,6

E.M.A. $= \frac{29,6}{30} = 0;99$

Classe B

x_i	n_i	$\lvert x_i-\bar{x}\rvert$	$n_i.\lvert x_i-\bar{x}\rvert$
1	1	5,2	5,2
2	1	4,2	4,2
3	2	3,2	6,4
4	3	2,2	6,6
5	3	1,2	3,6
6	6	0,2	I,2
7	5	0,8	4
8	4	I,8	7,2
9	3	2,8	8,4
10	2	3,8	7,6
	30		54,4

E.M.A. $= \frac{54,4}{30} = 1,81$

L'écart-moyen absolu présente comme inconvénient de donner le même "poids" à toutes les valeurs, et en particulier aux valeurs extrêmes.

Exemple 1: Ecart- moyen absolu =

Exemple 2: Ecart-moyen absolu =

Exemple 3: Ecart-moyen absolu =

Exemple 4: Ecart-moyen absolu =

Exemple 5: Ecart-moyen absolu =

Exemple 6: Ecart-moyen absolu =

L'écart-moyen n'est employé que lorsque les écarts sont très petits. Par exemple le physicien qui fait différentes mesures d'une même grandeur lors d'une expérience prendra en général la moyenne arithmétique des valeurs observées comme valeur de la grandeur mesurée et l'écart-moyen absolu comme erreur absolue de cette grandeur.

4-3-6 La variance - L'écart-type

On a vu qu'un inconvénient de l'écart-moyen absolu était de donner la même importance à toutes les valeurs observées. Pour une étude de la dispersion d'une distribution, il est préférable de valoriser les valeurs extrêmes des valeurs observées. Ceci peut se faire en prenant le carré des écarts par rapport à la moyenne. De cette manière si une valeur est 3 fois plus distante qu'une autre de la moyenne, elle interviendra 9 fois plus dans le paramètre de dispersion.

1- Définition: La variance (σ^2) d'une série statistique est la moyenne arithmétique des carrés des écarts par rapport à la moyenne de toutes les valeurs de la série.

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{p} n_i.(x_i-\bar{x})^2 = \sum_{i=1}^{p} f_i.(x_i-\bar{x})^2$$

L'ecart-type est la racine carrée de la variance

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{p} n_i.(x_i-\bar{x})^2} = \sqrt{\sum_{i=1}^{p} f_i.(x_i-\bar{x})^2}$$

La variance et l'écart-type seront d'autant plus grands que la distribution est dispersée.

Exemple 7:

Classe A

x_i	n_i	$n_i x_i$	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	$n_i (x_i - \bar{x})^2$
4	2	8	-2,2	4,84	9,68
5	7	35	-1,2	I,44	10,08
6	10	60	-0,2	0,04	0,4
7	6	42	0,8	0,64	3,84
8	4	32	I,8	3,24	12,96
9	1	9	2,8	7,84	7,84
	30	186			44,8

$\bar{x}_A = \frac{186}{30} = 6,2$

$\sigma_A^2 = \frac{44,8}{30} = 1,5$

$\sigma_A = I,2$

Classe B

x_i	n_i	$n_i x_i$	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	$n_i (x_i - \bar{x})^2$
1	1	1	-5,2	27,04	27,04
2	1	2	-4,2	17,64	17,64
3	2	6	-3,2	10,24	20,48
4	3	12	-2,2	4,84	14,52
5	3	15	-1,2	1,44	4,32
6	6	36	-0,2	0,04	0,24
7	5	35	0,8	0,64	3,2
8	4	32	1,8	3,24	12,96
9	3	27	2,8	7,84	23,52
10	2	20	3,8	14,44	28,88
	30	186			152,8

$\bar{x}_B = \frac{186}{30} = 6,2$

$\sigma_B^2 = \frac{152,8}{30} = 5,1$

$\sigma_B = 2,3$

2- Autre formule pour le calcul de la variance et de l'écart-type

$$\sigma^2 = \frac{1}{n} \sum_{i=1}^{p} n_i x_i^2 - \bar{x}^2 \qquad = \sum_{i=1}^{p} f_i x_i^2 - \bar{x}^2$$

Démonstration:

$$\sigma^2 = \sum_{i=1}^{p} f_i \cdot (x_i - \bar{x})^2$$

$$= \sum_{i=1}^{p} f_i \cdot (x_i^2 - 2x_i\bar{x} + \bar{x}^2)$$

$$= \sum_{i=1}^{p} f_i x_i^2 - 2\bar{x} \sum_{i=1}^{p} f_i x_i + \bar{x}^2 \sum_{i=1}^{p} f_i$$

$$= \sum_{i=1}^{p} f_i x_i^2 - 2\bar{x}.\bar{x} + \bar{x}^2$$

$$= \sum_{i=1}^{p} f_i x_i^2 - \bar{x}^2$$

Exemple 7: Classe A

x_i	x_i^2	n_i	$n_i x_i^2$
4	16	2	32
5	25	7	175
6	36	10	360
7	49	6	294
8	64	4	256
9	81	1	81
		30	1198

$$\sigma_A^2 = \frac{1198}{30} - (6,2)^2 = 1,5$$

$$\sigma_A = 1,2$$

Exemple 1: σ^2 = , σ =

Exemple 2: σ^2 = , σ =

Exemple 3: σ^2 = , σ =

Exemple 4: σ^2 = , σ =

Exemple 5: σ^2 = , σ =

Exemple 6: σ^2 = , σ =

3- Quelques propriétés de la variance et de l'écart-type

Les dimensions de la variance sont égales au carré de la dimension des données. L'écart-type a la même unité que les données.

Dans un échantillon où l'écart-type est inférieur à 15 % de la moyenne on peut considérer que la dispersion est faible. Dans un échantillon où l'écart-type est supérieur à 30 % de la moyenne, on peut considérer que la dispersion est forte. Ces nombres n'ont aucune valeur par eux-mêmes et sont donnés à titre purement indicatif . Le jugement doit être accordé à chaque situation. Dans l'exemple 7, par exemple,

$$\bar{x}_A = \bar{x}_B = 6,2 \qquad \sigma_A = 1,2 \qquad \sigma_B = 2,3$$

$$\frac{15}{100}.\ 6,2 = 0,93 \ , \quad \frac{30}{100}.\ 6,2 = 1,86$$

$$0,93 < \sigma_A < 1,86 \quad \text{et} \quad \sigma_B > 1,86$$

La classe A adonc une dispersion qui n'est ni faible, ni forte . La classeB a une dispersion forte.

La variance et l'écart-type ne dépendent pas de la taille des séries statistiques. Si deux groupes de même nature ont même dispersion, ils ont même variance et même écart-type, même si le nombre de données diffère d'un groupe à l'autre.

La variance est plus influencée par les données extrêmes que par les données centrales. Il faut donc se méfier des comparaisons de variance d'échantillons dont les distributions ont des formes très différentes.

4- Changement de variable

Pour faciliter les calculs, on peut être amené à effectuer un changement de variable du type

$$x'_i = a.x_i + b$$ (voir page 27)

Si σ est l'écart-type de la première série et σ' celui de la série après changement de variable , on a

$$\sigma'^2 = a^2.\sigma^2 \qquad \sigma' = |a|.\sigma$$

Démonstration:

$$\sigma'^2 = \sum_{i=1}^{p} f_i.(x'_i - \bar{x}')^2$$

$$= \sum_{i=1}^{p} f_i.\left[(ax_i + b) - (a\bar{x} + b) \right]^2$$

$$= \sum_{i=1}^{p} f_i.\ a^2.\ (x_i - \bar{x})^2$$

$$= a^2.\sigma^2$$

Exemple 4: On effectue le changement de variable $X'_i = \dfrac{X_i - 2550}{300} = \dfrac{1}{300} X_i - \dfrac{2550}{300}$

X_i	X'_i	n_i	$n_i X'_i$	X'^2_i	$n_i X'^2_i$
2550	0	3	0	0	0
2850	1	1	1	1	1
3150	2	6	12	4	24
3450	3	12	36	9	108
3750	4	14	56	16	224
4050	5	8	40	25	200
4350	6	4	24	36	144
4650	7	1	7	49	49
4950	8	0	0	64	0
5250	9	1	9	81	81
		50	185		831

$\overline{x'} = \frac{185}{50} = 3,7$

$\sigma'^2 = \frac{831}{50} - 3,7^2 = 2,9$ et $\sigma'^2 = (\frac{1}{300})^2 \cdot \sigma^2$

$\sigma^2 = 2,9 \cdot (300)^2 = 261\ 10^3$

$\sigma = 511$

5- Variable réduite

Pour comparer deux séries statistiques de même nature, on compare leurs moyennes qui caractérisent la tendance centrale et leurs écarts-type qui caractérisent la dispersion des distributions des séries.

Ce travail de comparaison est simplifié si on soumet la variable au préalable et dans chacune des séries à un changement de variable qui est tel que les nouvelles moyennes valent 0 et les nouveaux écarts-type 1.

Ce changement de variable est

$$x'_i = \frac{x_i - \overline{x}}{\sigma}$$

On a bien $\overline{x}' = \frac{1}{\sigma}.\overline{x} - \frac{\overline{x}}{\sigma} = 0$ et $\sigma' = \frac{1}{\sigma}.\sigma = 1$

Exemple: Un élève a 12/20 en math, la moyenne de la classe étant de 10/20 et sa ~~variance~~ écart-type de 2,4 . Il a 11/20 en français, la moyenne de la classe étant de 10/20 et sa ~~variance~~ écart-type de 1,1.En quelle matière l'élève sera-t-il le mieux classé?

En math, $x'_m = \frac{12 - 10}{2,4} = 0,83$

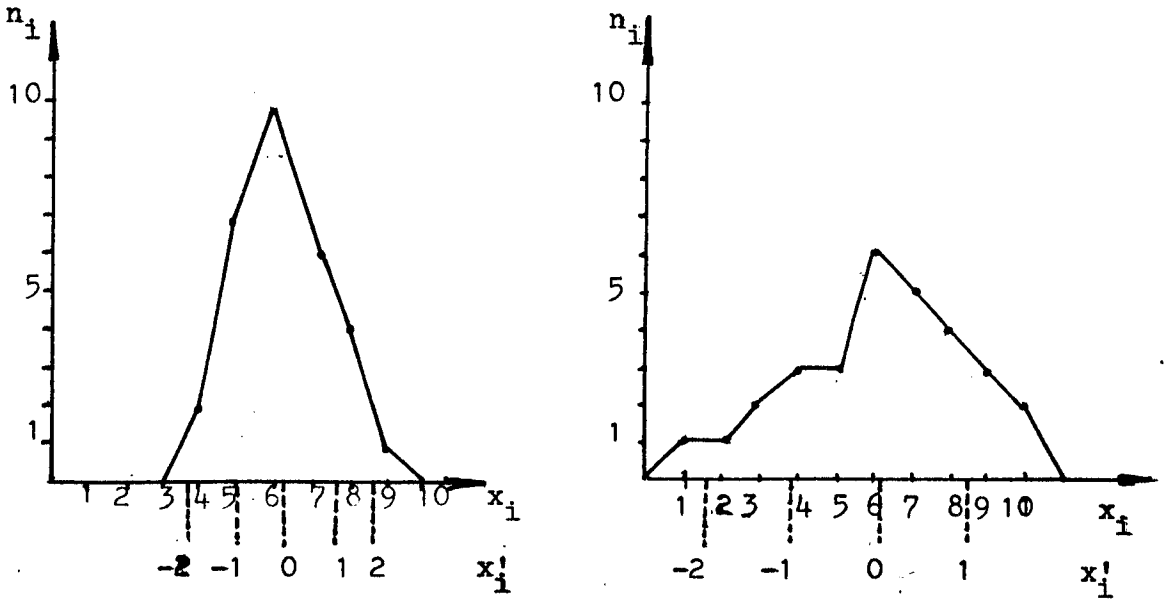
En français, $x'_f = \frac{11 - 10}{1,1} = 0,91$

L'élève sera donc mieux classé en français qu'en math.

Remarquons que si x'_i prend les valeurs ...-3, -2, -1, 0, 1, 2, 3, ...
x_i prend les valeurs ...$\overline{x}-3\sigma$, $\overline{x}-2\sigma$, $\overline{x}-\sigma$, $\overline{x}$, $\overline{x}+\sigma$, $\overline{x}+2\sigma$, $\overline{x}+3\sigma$, ...
et reprenons l'exemple 7:

Classe A $\bar{x}_A = 6,2$ $\sigma_A = 1,2$

ClasseB $\bar{x}_B = 6,2$ $\sigma_B = 1,3$



Variable réduite x'_i	...	-2	-1	0	1	2	...
x_i(A)		3,8	5	6,2	7,4	8,6	
x_i(B)		1,6	3,9	6,2	8,5	10,8	

Un 8,5 dans la classe B, "vaut" donc un 7,4 dans la classe A.

6- Estimation de l'écart-type de la population àpartir de l'échantillon

Si x est la moyenne de l'échantillon $x_1, x_2, x_3, \ldots x_n$ et si $\overline{x}'$ est la vraie moyenne de la population, pour calculer l'écart-type de l'échantillon nous avons travaillé sur les écarts

$$d_i = x_i - \overline{x}$$

mais pour avoir une bonne estimation de l'écart-type de l'échantillon, nous devrions travailler sur les écarts

$$d'_i = x_i - \overline{x}'$$

On a $d'_i - d_i = \overline{x} - \overline{x}'$ (1)

$$\sum_{i=1}^{n} (d'_i - d_i) = n(\overline{x} - \overline{x}')$$

$$\sum_{i=1}^{n} d'_i = n(\overline{x} - \overline{x}') \qquad \text{car} \sum_{i=1}^{n} d_i = 0$$

$$\frac{1}{n} \sum_{i=1}^{n} d'_i = \overline{x} - \overline{x}'$$

(1) $\rightarrow$ $d_i = d'_i - \frac{1}{n} \sum_{i=1}^{n} d'_i$

$$d_i^2 = d_i'^2 - \frac{2}{n} d'_i \sum_{i=1}^{n} d'_i + \frac{1}{n^2} \left(\sum_{i=1}^{n} d'_i\right)^2$$

$$\sum_{i=1}^{n} d_i^2 = \sum_{i=1}^{n} d_i'^2 - \frac{2}{n} \sum_{i=1}^{n} d'_i \sum_{i=1}^{n} d'_i + \frac{n}{n^2} \left(\sum_{i=1}^{n} d'_i\right)^2$$

$$= \sum_{i=1}^{n} d_i'^2 - \frac{1}{n} \left(\sum_{i=1}^{n} d'_i\right)^2$$

$$= \sum_{i=1}^{n} d_i'^2 - \frac{1}{n} \left(d_1'^2 + \ldots + d_n'^2 + \underbrace{2d'_1 d'_2 + \ldots + 2d'_{n-1} d'_n}_{\approx 0} \right)$$

$$\simeq \sum_{i=1}^{n} d_i'^2 - \frac{1}{n} \sum_{i=1}^{n} d_i'^2$$

(On émet l'hypothèse qu'il y a autant d'écarts positifs que d'écarts négatifs et que les écarts sont répartis de façon égale)

$$\simeq \frac{n-1}{n} \sum_{i=1}^{n} d_i'^2$$

$$\sum_{i=1}^{n} d_i'^2 \approx \frac{n}{n-1} \sum_{i=1}^{n} d_i^2$$

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

$$\Rightarrow \quad \sigma^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x}')^2 \simeq \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

$$\boxed{\sigma \simeq \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2} \simeq \sqrt{\frac{1}{n-1} \sum_{i=1}^{p} n_i (x_i - \bar{x})^2}}$$

(Correction de Bessel)

Il serait en fait plus correct de toujours diviser par n-1 plutôt que par n dans le calcul de σ^2 et de σ. La différence est peu importante si n est grand.

4-3-7 Coefficient de variation

$$v = \frac{\sigma}{\bar{x}} \qquad (\bar{x} \neq 0)$$

Ce paramètre ne dépend pas de l'unité de mesure choisie et est souvent exprimé en %.

$$v = \frac{100\sigma}{\bar{x}} \quad \%$$

L'avantage de ce paramètre est de pouvoir comparer la dispersion de différentes séries statistiques dont les effectifs sont très différents.

En se référant à la page 37, si $\bar{x}$ est positif, on peut dire que si $\sigma < \frac{15}{100} \bar{x}$, c-à-d si $v < 15\%$, alors la dispersion de la distribution est <u>faible</u> et que si $\sigma > \frac{30}{100} \cdot \bar{x}$ c-à-d si $v > 30$ %, alors la dispersion de la distribution est <u>forte</u>.

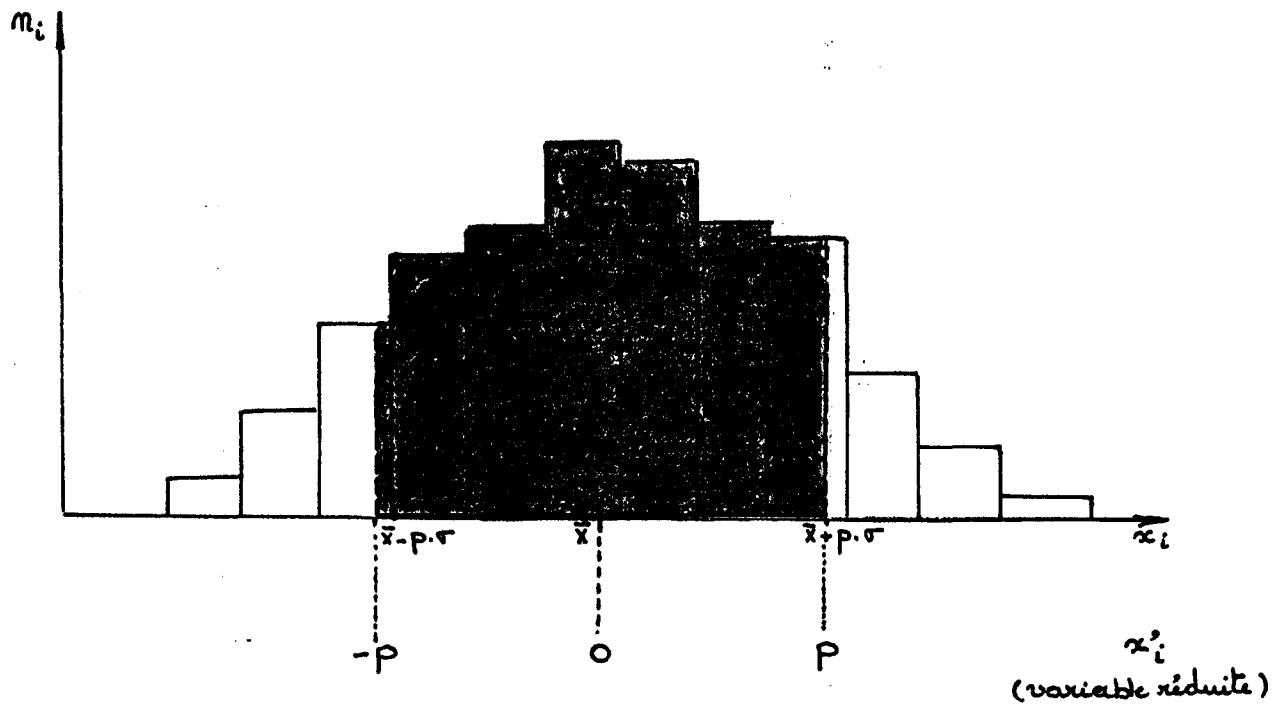
4-3-8 Usage des différents paramètres de la dispersion

L'étendue est peu utilisée car elle ne dépend que des deuxvaleurs extrêmes des valeurs observées.

L'écart-moyen absolu est utilisé lorsque tous les écarts par rapport à la moyenne sont très petits.

La variance, l'écart-type et le coefficient de variation sont le plus souvent utilisés; néanmoins pour les distributions disymétriques, on recourt plutôt aux quartiles et à l'écart interquartile.

4-3-9 Inégalité de Bienaymé - Tchebicheff



Soit k la fréquence absolue de <u>toutes</u> les valeurs observées correspondant à la région foncée.

$$k < n$$

Pour ces k observations que nous noterons $x_1, x_2, x_3 \ldots x_k$, on a

$$|\bar{x} - x_i| \leqslant \sigma.p \qquad i \in \{1,2 \ldots, k\}$$

Pour les n-k autres observations que nous noterons x_{k+1}, $x_{k+2} \ldots x_n$

$$|\bar{x} - x_{k+i}| > p.\sigma \qquad i \in \{1,2, \ldots, n-k\} \qquad (2)$$

D'autre part

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

$$n.\sigma^2 = (\bar{x}-x_1)^2 + (\bar{x}-x_2)^2 + \ldots + (\bar{x}-x_k)^2 + (\bar{x}-x_{k+1})^2 + \ldots + (\bar{x}-x_n)^2$$

$$n.\sigma^2 > (\bar{x}-x_{k+1})^2 + \ldots + (\bar{x}-x_n)^2$$

(2) $$n.\sigma^2 > p^2\sigma^2 + p^2\sigma^2 + \ldots + p^2\sigma^2$$

$$n.\sigma^2 > (n-k)p^2\sigma^2$$

$$n > (n-k)p^2 \qquad (\text{car } \sigma^2 \neq 0)$$

$$\frac{1}{p^2} > \frac{n-k}{n} \qquad (\text{car } n \text{ et } p^2 > 0)$$

$$\frac{1}{p^2} > 1 - \frac{k}{n}$$

$$\boxed{\frac{k}{n} > 1 - \frac{1}{p^2}}$$

La fréquence relative cumulée correspondant aux observations dont l'écart à la moyenne est au plus égal à p écarts-types est supérieure à $1 - \frac{1}{p^2}$

Exemple:

Distribution de 1526 manoeuvres agricoles suivant la durée hebdomadaire de travail

Durée en heures x_i	n_i	X_i	$X'_i = \frac{X_i - 52,5}{5}$	$n_i X'_i$	X'^2_i	$n_i X'^2_i$
40]	10	37,5	-3	-30	9	90
]40,45]	87	42,5	-2	-174	4	348
]45,50]	473	47,5	-1	-473	1	473
]50,55]	457	52,5	0	0	0	0
]55,60]	111	57,5	1	111	1	111
]60,65]	304	62,5	2	608	4	1216
]65,70]	54	67,5	3	162	9	486
]70	30	72,5	4	120	16	480
	1526			-677+1001 = 324		3204

$\overline{X'} = \frac{324}{1526} = 0,212...$

$\overline{X} = 53,6$

$\sigma'^2 = \frac{3204}{1526} - 0,212^2 = 2,0547...$

$\sigma^2 = 51,3675...$

$\sigma = 7,2$

Appliquons Bienaymé-Tchebicheff avec p=2

$[\overline{x} - 2\sigma, \overline{x} + 2\sigma] = [53,6 - 2.7,2 ; 53,6 + 2.7,2] = [39,2 ; 68,0]$

et $1 - \frac{1}{p^2} = 1 - \frac{1}{4} = 0,75$

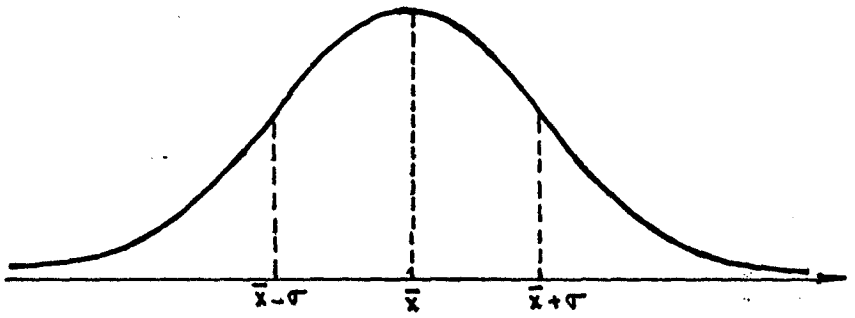
Il y a donc 75 % des manoeuvres qui ont une durée hebdomadaire de travail domprise entre 39,2 heures et 68,0 heures.

5~ La Gaussienne

Une étude complète de la distribution de Gauss (distribution normale, distribution en cloche) ne peut être faite ici puisque la connaissance de la notion d'intégrale est nécessaire pour faire cette étude.

La distribution normale étant la plus courante, il serait dommage de ne pas en dire quelques mots.

Les phénomènes qui se distribuent selon une courbe en "cloche" sont ceux qui sont dûs au hasard ou tout au moins, qui sont le résultat d'un grand nombre de facteurs, agissant tous indépendamment les uns des autres



La courbe de Gauss est symétrique et a une forte concentration des observations autour de la moyenne: 68 % des observations sont comprises entre $\bar{x} - \sigma$ et $\bar{x} + \sigma$ et 95 % des observations sont comprises entre $\bar{x} - 2\sigma$ et $\bar{x} + 2\sigma$.

Dans l'exemple 5: $\bar{x} = 189,6$ (pg 15)

$\sigma = 56,4$

68 arbres sur 100 doivent donc fournir entre 133 et 246 fruits

95 arbres sur 100 doivent donc fournir entre 77 et 302 fruits, pour que la loi soit gaussienne. Ces résultats sont voisins de ceux fournis par la série. On peut donc penser que cette production est "normale".

6~ Série statistique double

6-1 On peut étudier une populationsous deux caractères.

Exemple: Etude de l'ensemble des propriétés rurales envisagée sous un premier caractère "nombre de personnes actives" et sous un second caractère " surface cultivable".

Une enquête donne les résultats suivants

(a,b) : a= nombre de personnes actives de la propriété

b= surface cultivable en ha de la propriété

(1,20)	(2,42)	(1,25)	(3,48)	(2,38)	(1,35)	(2,50)	(2,35)
(2,35)	(2,40)	(1,26)	(3,52)	(3,48)	(1,30)	(2,27)	(2,36)
(1,22)	(3,39)	(1,30)	(5,70)	(5,68)	(2,36)	(2,32)	(2,42)
(3,45)	(4,65)	(2,35)	(1,25)	(1,30)	(2,38)	(3,55)	(3,52)
(4,60)	(4,70)	(4,58)	(1,26)	(1,26)	(2,45)	(3,56)	(2,30)

On range d'abord les couples par ordre de valeurs croissantes du premier élément . On obtient ainsi 5 classes. A l'intérieur de chaque classe on ordonne ensuite les couples d'après le second caractère.

(1,20)	(2,27)	(3,39)	(4,58)	(5,68)
(1,22)	(2,30)	(3,45)	(4,60)	(5,70)
(1,25)	(2,32)	(3,48)	(4,65)	
(1,25)	(2,35)	(3,48)	(4,70)	
(1,26)	(2,35)	(3,52)		
(1,26)	(2,35)	(3,52)		
(1,26)	(2,36)	(3,55)		
(1,30)	(2,36)	(3,56)		
(1,30)	(2,38)			
(1,30)	(2,38)			
(1,35)	(2,40)			
	(2,42)			
	(2,42)			
	(2,45)			
	(2,50)			

6-2 Tableau de corrélation

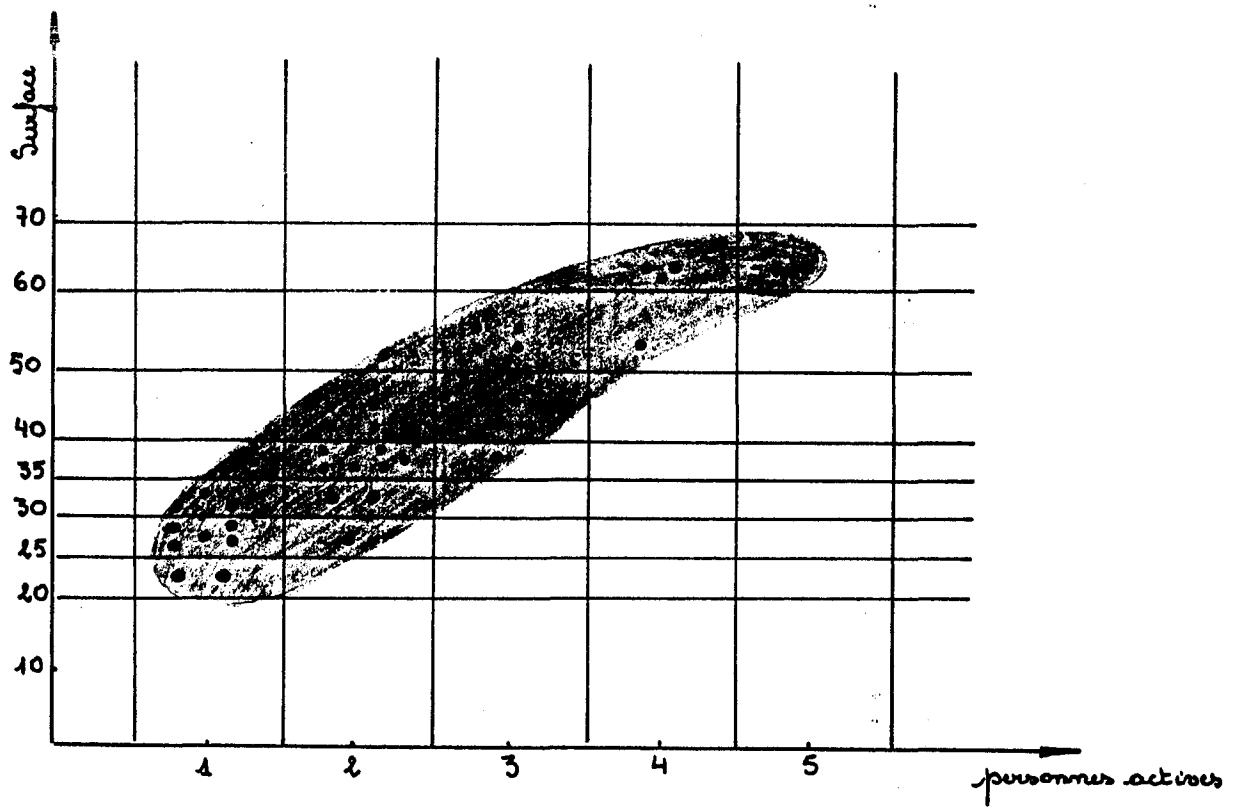
On forme ensuite un tableau en formant des classes à partir du second caractère.

Chaque colonne et chaque ligne constitue une série statistique simple (à un caractère) . La dernière ligne et la dernière colonne sont appelées distributions marginales.

surfaces \ personnes actives	1	2	3	4	5	Total
20 < 25	2					2
25 < 30	5	1				6
30 < 35	3	2				5
35 < 40	1	7	1			9
40 < 50		4	3			7
50 < 60		1	4	1		6
60 < 70				3	2	5
Total	11	15	8	4	2	40

6-3 Nuage de points

Cette série double peut être représentée par un nuage de points

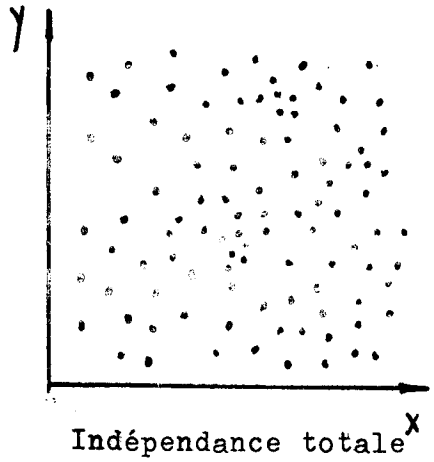


6-4 Corrélation

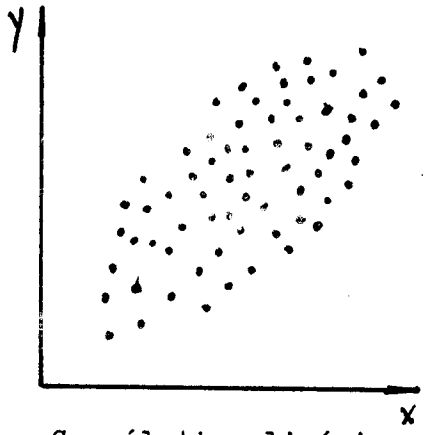
La forme du nuage nous donne de précieux renseignements sur la "corrélation" entre les deux caractères:

On dit qu'il y a corrélation entre deux caractères des éléments d'une même série lorsque les variations de l'un sont dépendantes des variations de l'autre. Cette dépendance peut aller de la liaison formelle (relation mathématique entre les deux caractères) à l'indépendance complète.

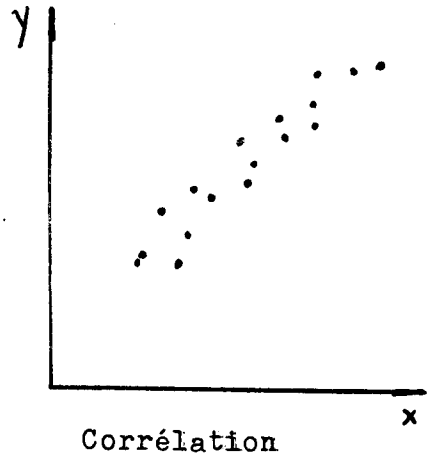
Exemples:



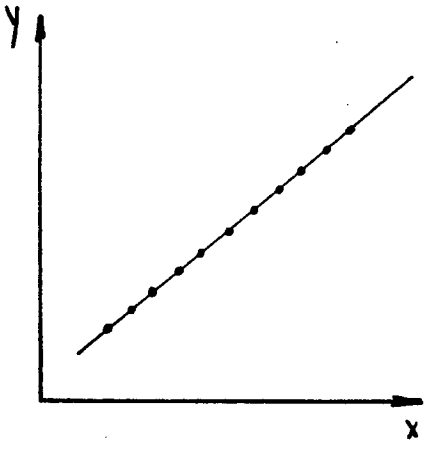
Indépendance totale



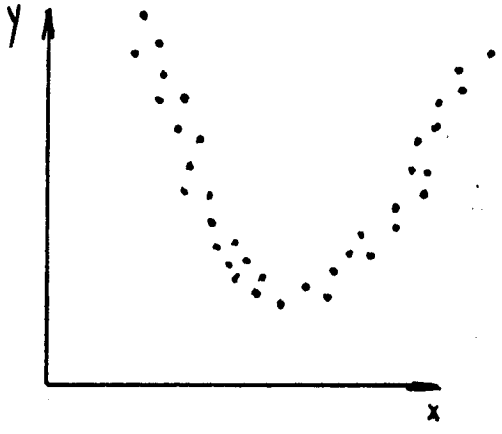
Corrélation linéaire large



Corrélation linéaire serrée

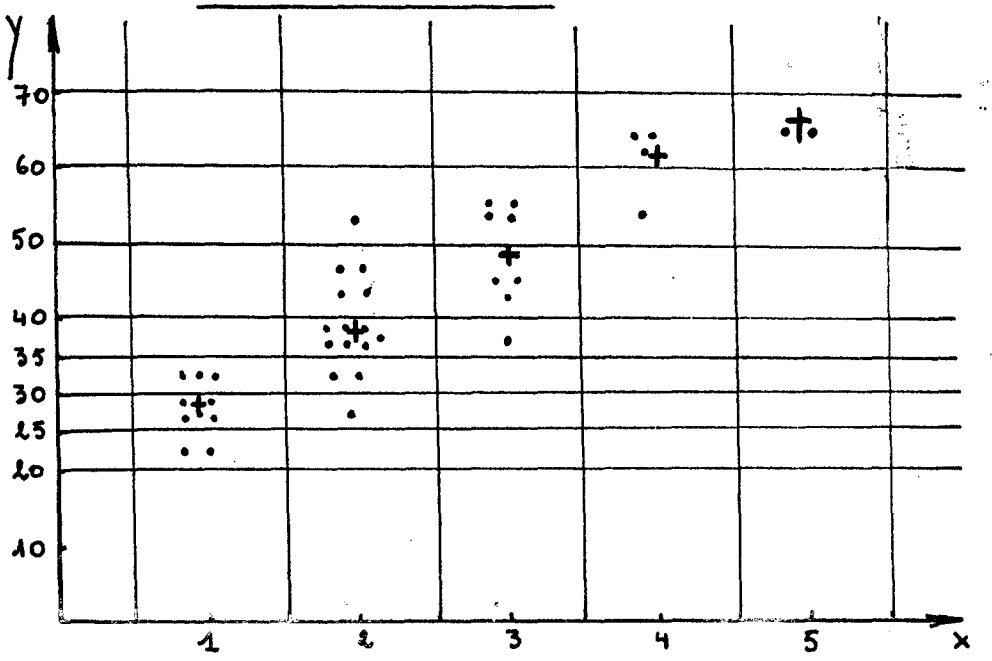


Corrélation linéaire parfaite

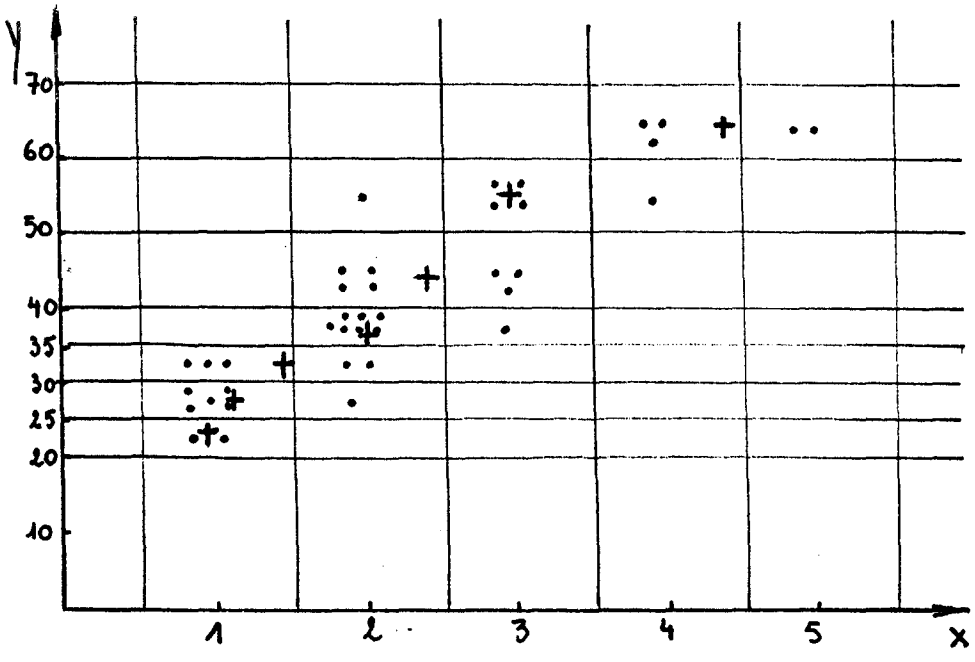


Corrélation parabolique

6-5 Corrélation linéaire



Valeurs centrales des colonnes



Valeurs centrales des lignes

Marquons d'une croix les valeurs centales des bandes parallèles à l'axe des ordonnées et les valeurs centrales des bandes parallèles à l'axe des abscisses. Dans les deux cas, les croix adoptent une direction linéaire.

Dans le premier cas, si nous ajustons cette série de croix à une droite, l'équation de cette droite $y = ax + b$ donnera la surface

cultivable des fermes en fonction du nombre de personnes actives. La série de croix donne en effet par leurs coordonnées une valeur approximative de y en fonction de x. La droite obtenue est appelée droite de régression de y en x. (On prédit y à partir de x)

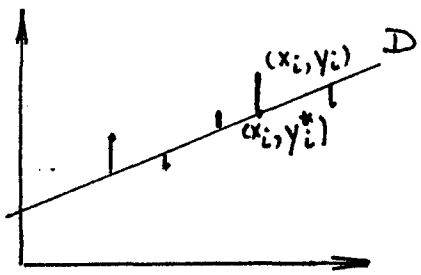
Dans le second cas, si nous ajustons cette série de croix à une droite, l'équation de cette droite x= a'y + b' donnera le nombre de personnes actives de la ferme en fonction de la surface cultivable. La série de croix donne en effet par leurs coordonnées une valeur approximative de x en fonction de y. La droite obtenue est appelée droite de régression de x en y. (on prédit x à partir de y)

Quelle est l'équation de ces droites?

6-6 Méthode des moindres carrés

Cette méthode consiste à trouver l'équation d'une droite telle que la somme des carrés des "distances" des points observés à la droite d'ajustement soit minimale.

Cas 1:



1°-La droite recherchée passe par le point $(\overline{x},\overline{y})$ où

$\overline{x}$ est la moyenne des abscisses des points(moyenne de la ligne marginale)

$\overline{y}$ est la moyenne des ordonnées des points (moyennede la colonne marginale)

$$D \equiv y - \overline{y} = a.(x - \overline{x})$$

2° - Calcul du coefficient directeur de D

Il faut que $\sum_{i=1}^{n} (y_i - y_i^*)^2$ soit minimale

$$\sum_{i=1}^{n}(y_i-y_i^*)^2 = \sum_{i=1}^{n}(y_i - a.(x_i-\bar{x}) - \bar{y})^2$$

$$= \sum_{i=1}^{n}(y_i^2 + a^2(x_i-\bar{x})^2 + \bar{y}^2 -2ay_i(x_i-\bar{x}) - 2y_i\bar{y} + 2a\bar{y}(x_i-\bar{x}))$$

Dérivons par rapport à a. Cette dérivée doit être nulle

$$2a\sum_{i=1}^{n}(x_i-\bar{x})^2 - 2\sum_{i=1}^{n} y_i(x_i-\bar{x}) + 2\bar{y}\sum_{i=1}^{n}(x_i-\bar{x}) = 0$$

$$a\sum_{i=1}^{n}(x_i-\bar{x})^2 = \sum_{i=1}^{n} y_i(x_i-\bar{x}) - \bar{y}\sum_{i=1}^{n}(x_i-\bar{x})$$

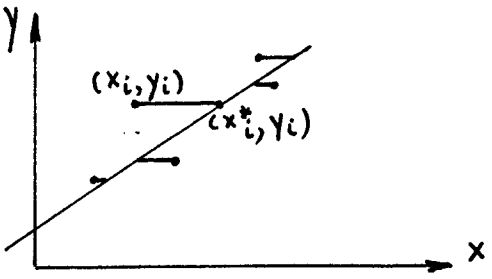
$$a = \frac{\sum_{i=1}^{n}(x_i-\bar{x}).(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$$

Droite de régression de y en x:

$$D \equiv y - \bar{y} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}.(x - \bar{x})$$

Cas 2: Par le même raisonnement on trouve,

Droite de régression de x en y:



$$D' \equiv x - \bar{x} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(y_i-\bar{y})^2}.(y - \bar{y})$$

surface (y_i) dans	valeurs centrale	personnes actives (x_i) 1	2	3	4	5	Totaux n_i	$y_i n_i$	$y_i - \bar{y}$	$(x_i - \bar{x}).a_i$	$(x_i - \bar{x}).a_i . (y_i - \bar{y})$	$(y_i - \bar{y})^2$	$(y_i - \bar{y})^2 . n_i$
20<25	22,5	2					2	45	-19,5	-2,55	49,725	380,25	760,5
25<30	27,5	5	1				6	165	-14,5	-6,65	96,425	210,25	1261,5
30<35	32,5	3	2				5	162,5	-9,5	-4,375	41,5625	90,25	451,25
35<40	37,5	1	7	1			9	337,5	-4,5	-2,475	11,1375	20,25	182,25
40<50	45		4	3			7	315	3	1,075	3,525	9	63
50<60	55		1	4	1		6	330	13	4,35	56,55	169	1014
60<70	65				3	2	5	325	23	10,625	244,375	529	2645
totaux n'_i		11	15	8	4	2	40	1680			503		6377,5
$x_i . n'_i$		11	30	24	16	10	91	$\bar{x} = 2,275$					
$x_i - \bar{x}$		-1,275	-0,275	0,725	1,725	2,725		$\bar{y} = 42$					
$(y_i - \bar{y}) . a_i$		-144,5	-40	56,5	82	46							
$(y_i - \bar{y}) . a_i (x_i - \bar{x})$		184,2375	11	40,9625	141,45	125,35	503						
$(x_i - \bar{x})^2$		1,625	0,075	0,525	2,975	7,425							
$(x_i - \bar{x})^2 . n'_i$		17,875	1,125	4,2	11,9	14,85	49,95						

vérification !

Explication de la ~~colonne~~ ligne $(y_i - \bar{y})a_i$

Considérons les points de la colonne 1(sur le graphique).
La somme des produits $(x_i - \bar{x})(y_i - \bar{y})$ relatifs à ces points est égale à:

$$2.(-1,275).(-19,5) + 5.(-1,275).(-14,5) + 3.(-1,275).(-9,5) + 1.(-1,275).(-4,5)$$

$$= (-1,275).\left[2.(-19,5) + 5.(-14,5) + 3.(-9,5) + 1.(-4,5)\right]$$

$$= (-1,275).\ (-144,5)$$

Le résultat de la quantité entre crochets est inscrit dans la première case de la ~~colonne~~ ligne $(y_i - \bar{y})a_i$.

Pour notre exemple:

Droite de régression de y en x:

$$D \equiv y - 42 = \frac{503}{49,95} . (x - 2,275)$$

$$D \equiv y = 10,07.x + 29,16$$

Droite de régression de x en y:

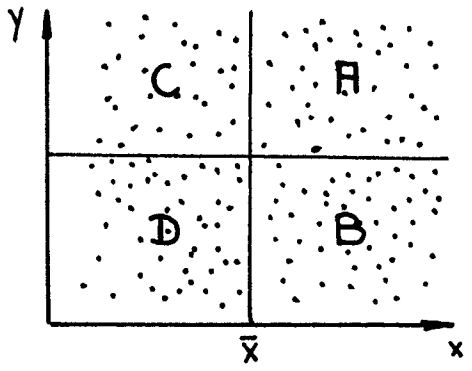
$$D' \equiv x - 2,275 = \frac{503}{6377,5} . (y - 42)$$

$$D' \equiv x = 0,078.y - 1$$

6-7 Coefficient de corrélation

Le degré de corrélation entre deux caractères d'une même population sera caractérisé par l'écartement entre les droites de régression.

Si les ~~droites~~ caractères sont liés par une liaison formelle, les deux droites seront confondues. Si les deux caractères sont indépendants, les deux droites seront respectivement parallèles aux axes car la valeur $\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$ serait sensiblement nulle; la somme des produits relatifs aux régions A et B annulant la somme des produits relatifs aux régions C et D;



Pour mesurer l'écartement des droites, nous calculerons la racine carrée du produit des "pentes" des droites de régression (! attention, ces pentes sont calculées dans des axes différents!!!)

$$\underline{\text{Coefficient de corrélation}}, r = \pm\sqrt{\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2} \cdot \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(y_i-\bar{y})^2}}$$

(L'emploi du signe + ou - sera expliqué dans les remarques qui suivent ce paragraphe)

Dans notre exemple, $r = \sqrt{10,07 \cdot 0,078} = 0,88$

Remarques:

1) On donne à r le signe de $\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})$.

Si $r > 0$, les deux droites sont croissantes et les valeurs des deux caractères sont proportionnelles.

Si $r < 0$, les deux droites sont décroissantes et lesvaleurs des deux caractères sont inversément proportionnelles.

2) Le coefficient de corrélation ne peut être utilisé que dans le cas de corrélation "linéaire". Pour les autres cas, on a recourt au rapport de corrélation de Pearson ou au coefficient de Spearman.

3) En dessous de r= 0,25 on a tendance à penser qu'il ny a aucune corrélation entre les deux caractères.

4) R ~ 1 : grande corrélation entre les deux caractères.

Quelques références

H. Breny	Calcul des probabilités et des statistiques	Dessain (1970)
L. D'Hainaut	Concepts et méthodes de la statistique	Labor-Nathan (1975)
D. Fèvre, Y. Pesez	Les mathématiques	Encyclopédies du savoir moderne (1975)
G. Herniaux	Cours de statistique	Masson (1971)
P. Jaffard	Statistique	Masson (1977)
P. Smets	Eléments de statistique médicale	Presses U.L.B. (1979)
J. Verlooy	Mathématique	Suite 45 (1972)
M. VioT	Statistiques	Institut horticole de Gembloux (1980)