STATISTIQUES DESCRIPTIVES

STATISTIQUES DESCRIPTIVES

I Terminologie de base

1) La variable statistique (ou caractère) est le fait élémentaire dont on veut avoir une connaissance correcte, et qui fera ensuite l'objet d'une statistique. Un caractère peut être qualitatif ou quantitatif et dans ce dernier cas être continu ou discontinu.
On aboutira à une distribution statistique du caractère, donnant pour les différentes catégories, ou "classes" possibles le nombre d'individus rencontrés, ou "effectif" appartenant à cette classe.

Exemples :

opinion sur un produit donné		revenu annuel d'un ménage		nbre d'enfants par ménage	
classe	effectif	classe	effectif	classe	effectif
favorable neutre défavorable		moins d'un million [1 000 000, 2 000 000[2 000 000 et plus		0 1 2 3 4 5 et plus	

2) L'ensemble des individus faisant l'objet d'une observation est appelé la population ou référentiel statistique.
Le sous-ensemble de la population sur lequel la variable a été effectivement mesurée est appelé l'échantillon.

3)

Terminologie fondamentale de la statistique

Exemples / *Terminologie*	***Exemple 1.*** *Étude des montants hors taxe des 1 500 factures établies au mois de mars, par prélèvement de 100 d'entre elles au hasard.*	***Exemple 2.*** *Étude du stock de 1 000 boites de mouchoirs en fonction des modèles (homme, femme, enfant).*	***Exemple 3.*** *Détermination du nombre d'erreurs de frappe d'un pool dactylographique, grâce à un relevé, 10 jours consécutifs, des nombres d'erreurs faites journellement.*	***Exemple 4.*** *Établissement du budget de la prochaine campagne publicitaire, grâce à la relation existant entre le chiffre d'affaires et les dépenses publicitaires des 16 semestres précédents.*
Variable statistique	Le montant H.T. constitue la variable Cette variable est quantitative, continue et simple.	Le modèle (homme, femme, enfant) constitue la variable. Cette variable est qualitative et simple.	Le nombre d'erreurs de frappe faites journellement constitue la variable. Cette variable est quantitative, discontinue et simple.	Le couple « chiffre d'affaires - dépenses » constitue la variable. Cette variable est quantitative, continue et complexe.
Unité statistique individu	Chaque saisie du montant H.T. se fait sur une facture. La facture constitue l'unité statistique.	Chaque modèle concerne une boite. La boite est l'unité statistique.	On enregistre les fautes commises dans une journée. Le jour est l'unité statistique.	Le couple « chiffre d'affaires - dépenses » est noté semestre par semestre. Le semestre est l'unité statistique.
Échantillon	Les 100 factures constituent l'échantillon. La taille de l'échantillon est de 100.	L'étude porte sur les 1 000 boites. Il n'y a pas d'échantillon, il s'agit d'un recensement.	Le relevé s'effectue sur 10 jours. L'échantillon est constitué par ces 10 jours. La taille de l'échantillon est donc de 10.	L'échantillon est constitué par les 16 semestres. La taille de l'échantillon est donc de 16.
Population	Les 1 500 factures du mois de mars représentent la population. Selon le degré de précision recherché, on aurait pu faire un recensement.	Les 1 000 boîtes constituent la population.	Si les 10 jours ont été tirés au hasard pendant une durée d'un mois. La population est donc 30/31 jours.	La population est constituée de l'ensemble des 16 semestres. L'échantillon se confond ici avec la population.

4)

Désignons par x_i une valeur quelconque de la variable statistique et par n_i son effectif. La fréquence d'une valeur x_i est le rapport de l'effectif correspondant à l'effectif total. Ce rapport est noté f_i. $(1 \leq i \leq m)$

$$f_i = \frac{n_i}{\sum_{i=1}^{m} n_i} = \frac{n_i}{N}$$ où N est l'effectif total

Exemple : relevé des notes de 1 à 6 dans une classe de 28 élèves.

NOTE	EFFECTIF	FREQUENCE
1	1	0,0357143
2	2	0,0714286
3	6	0,2142857
4	10	0,3571429
5	5	0,1785714
6	4	0,1428571

La somme des fréquences est toujours 1.

5) Effectif simple n_i et effectif cumulé N_i
Fréquence simple f_i et fréquence cumulée F_i

L'effectif cumulé croissant indique combien d'unités de la population sont caractérisés par une valeur inférieure à ...
L'effectif cumulé décroissant indique combien d'unités de la population sont caractérisées par une valeur supérieure à ...

NOTE	EFFECTIF	EFFECTIF CROISSANT	CUMULE DECROISS.	FREQUENCE	FREQUENCE CROISSANTE	CUMULEE DECROISS.
1	1	1	28	0,03571	0,03571	1,00000
2	2	3	27	0,07143	0,10714	0,96429
3	6	9	25	0,21429	0,32143	0,89286
4	10	19	19	0,35714	0,67857	0,67857
5	5	24	9	0,17857	0,85714	0,32143
6	4	28	4	0,14286	1,00000	0,14286

II Graphiques des effectifs

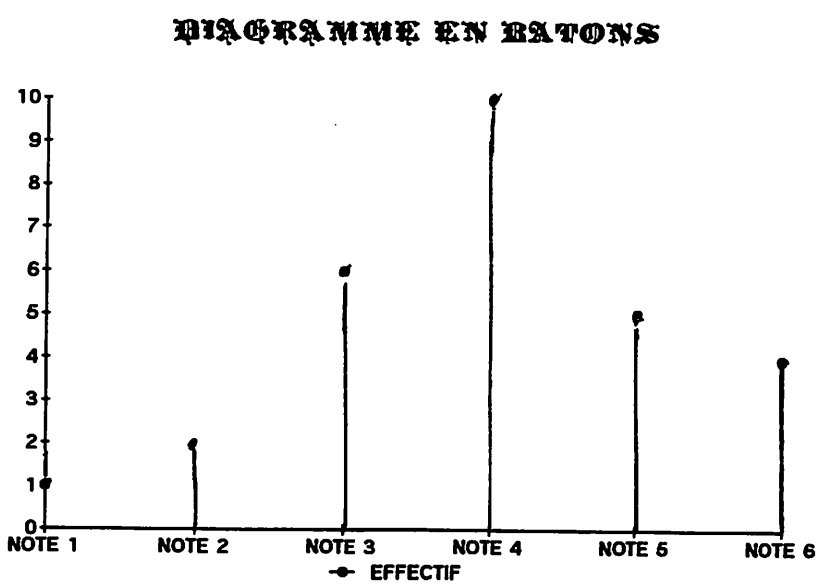
Fonction de distribution

1) Variable statistique discontinue

* graphique en bâtons

Sur l'axe des abscisses, on reporte des points correspondants aux valeurs isolées. Au-dessus de chaque point on trace un trait ou un bâton dont la longueur est proportionnelle à l'effectif correspondant.

Exemple: des notes dans une classe (I-F)



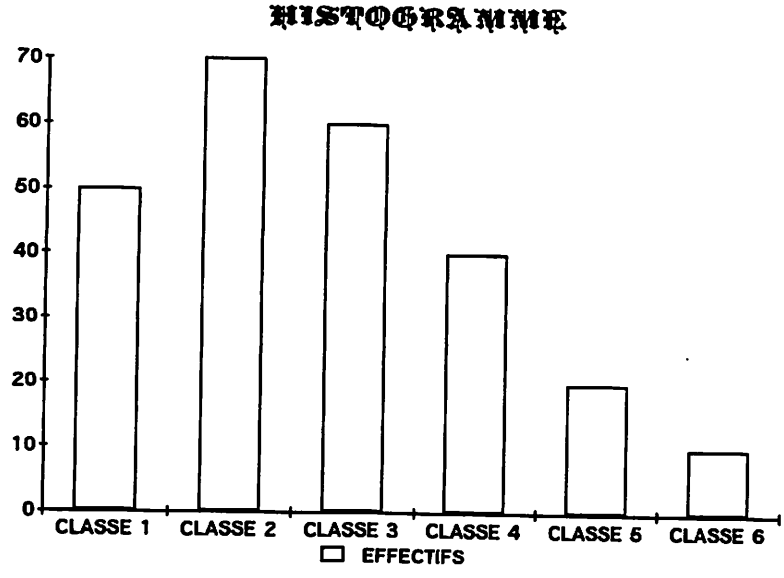
2) Variable statistique continue

* Histogramme

Pour chaque classe, on élève un rectangle ayant une base proportionnelle à l'intervalle de classe et une hauteur proportionnelle à l'effectif simple. Dans ce cas ce sont les surfaces, et non les hauteurs, qui sont proportionnelles aux effectifs.

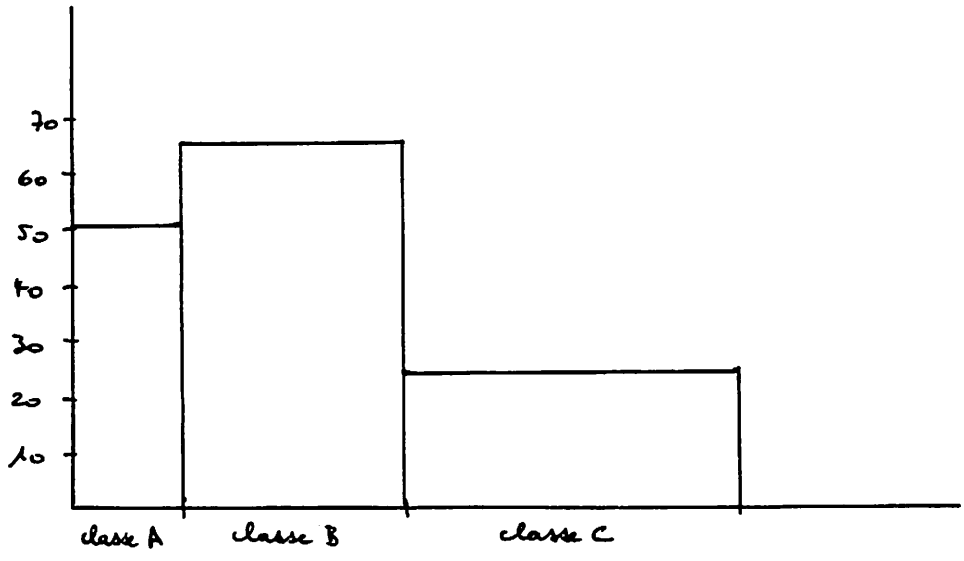
exemple: distribution des cours de l'action d'une société pendant une année

cours de l'action		effectifs
classe	x_i	n_i
1	1000 à 1499	50
2	1500 à 1999	70
3	2000 à 2499	60
4	2500 à 2999	40
5	3000 à 3499	20
6	3500 à 3999	10



Remarque : Pour des classes d'amplitudes inégales, il faut respecter la proportionnalité des surfaces :

cours de l'action		effectifs
classe	x_i	n_i
A	1000 à 1499	50
B	1500 à 2499	130
C	2500 à 3999	70

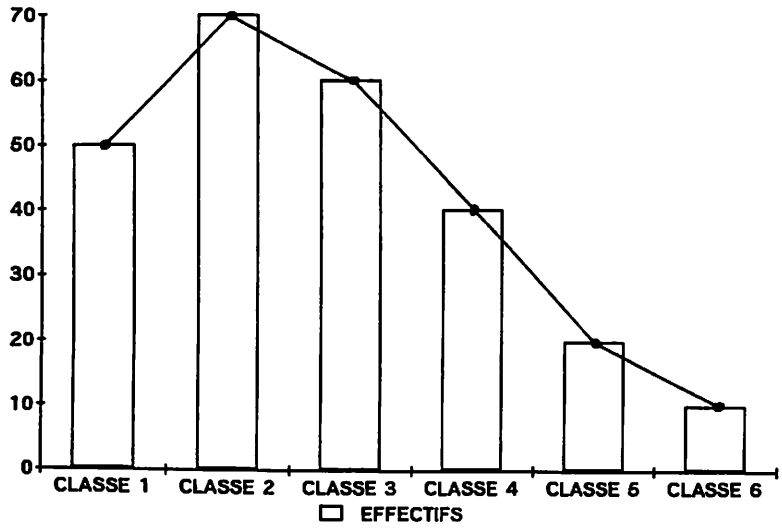


* Polygone des effectifs

Le polygone des effectifs est obtenu en joignant par des segments de droite les milieux des bases supérieures des rectangles.

Exemple : Cours de l'action d'une société de la page précédente, réparti en 6 classes de même amplitude.

POLYGONE DES EFFECTIFS



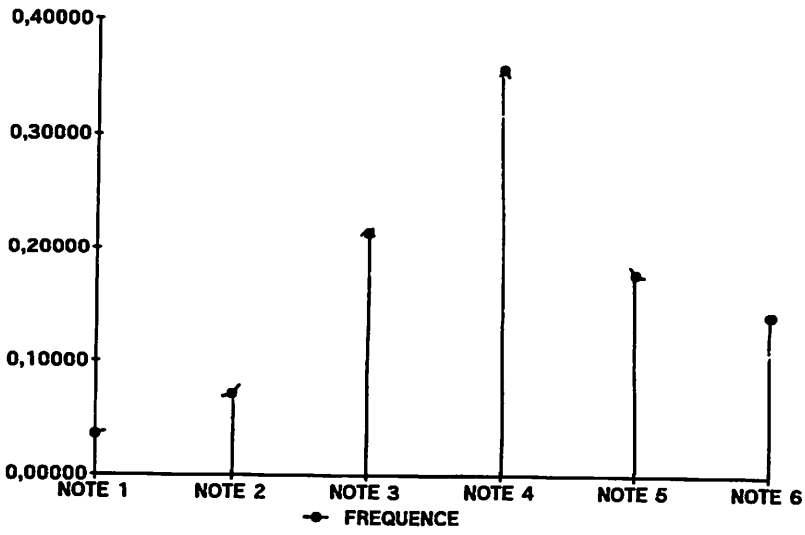
III Graphiques des fréquences

1) Variable statistique discontinue

* Graphique en bâtons

Exemple des notes dans une classe (I-4)

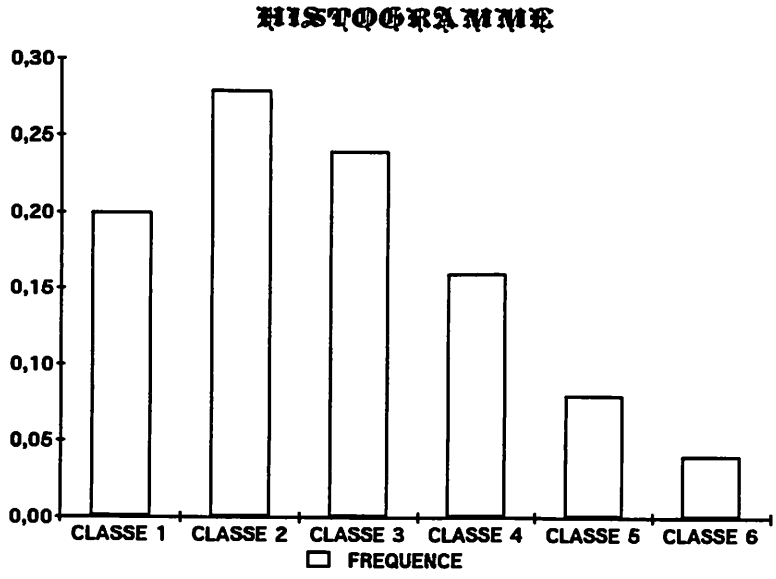
DIAGRAMME EN BATONS



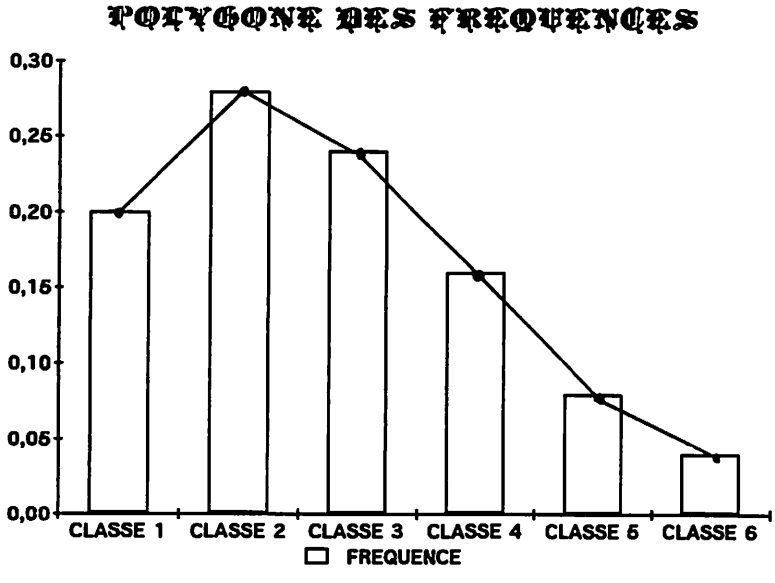
2) Variable statique continue

* histogramme

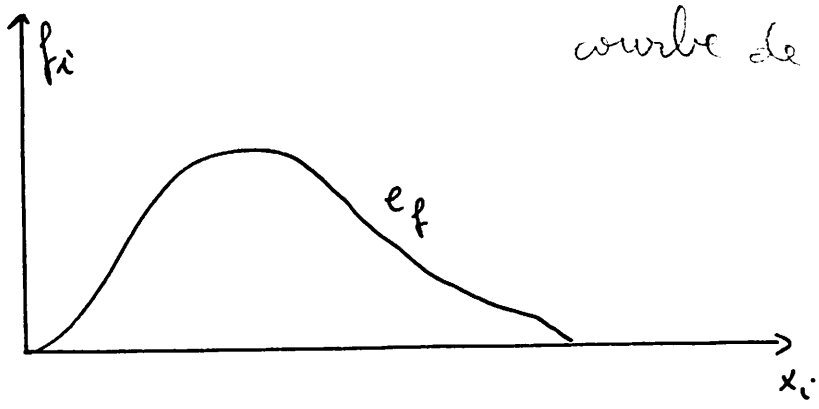
Exemple de la distribution des ventes de l'action d'une société pendant une année (II - 2)



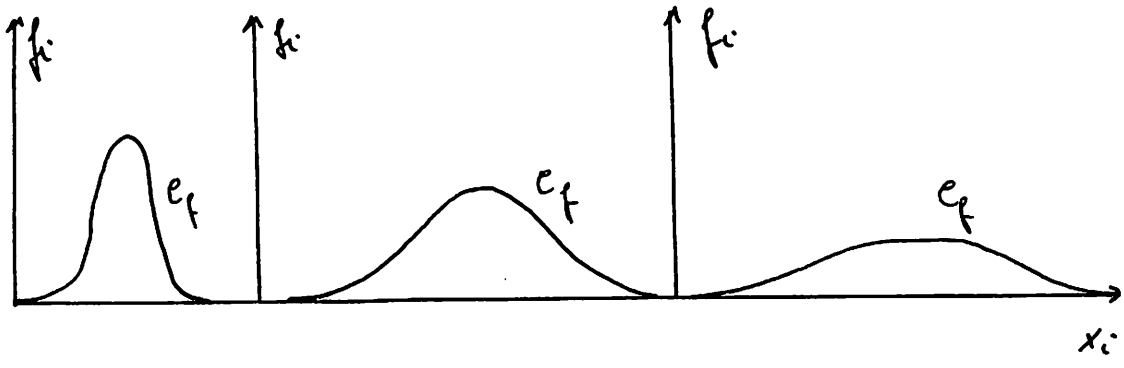
* polygone des fréquences



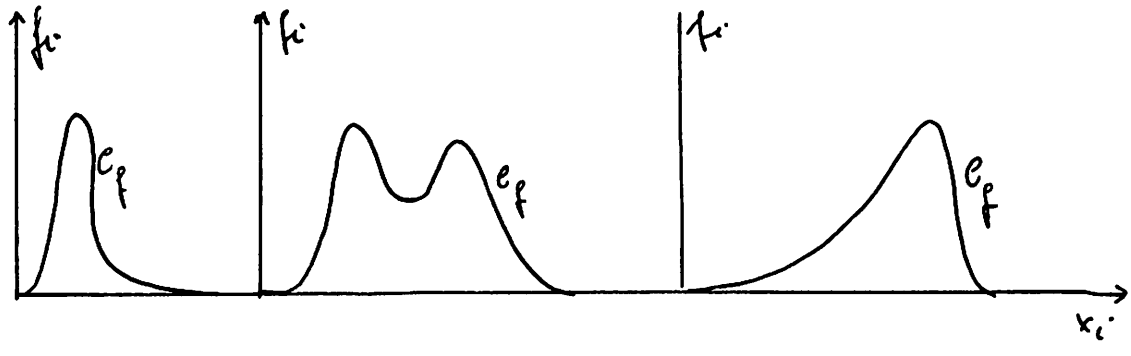
3) Lorsque l'intervalle des classes est très petit et les données nombreuses, la ligne brisée tend à devenir une courbe, appelée courbe des fréquences. La fonction y relative est notée f (fonction de distribution)



distributions symétriques :



distributions asymétriques :



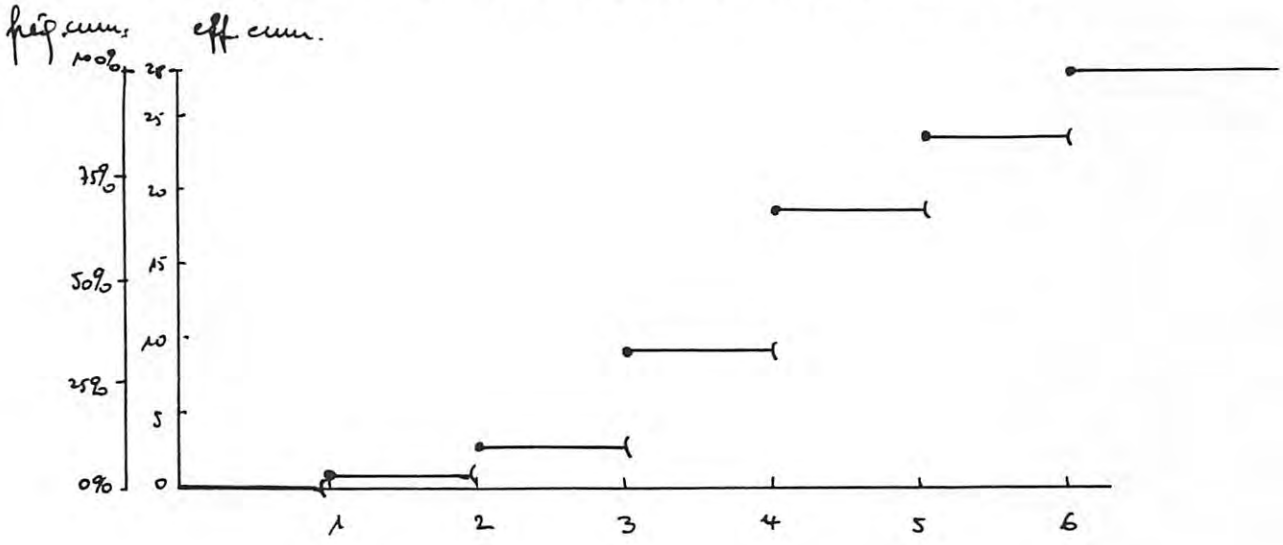
10

Graphiques des effectifs cumulés
Graphiques des fréquences cumulées
Fonction de répartition.

---

1) Variable statistique discontinue

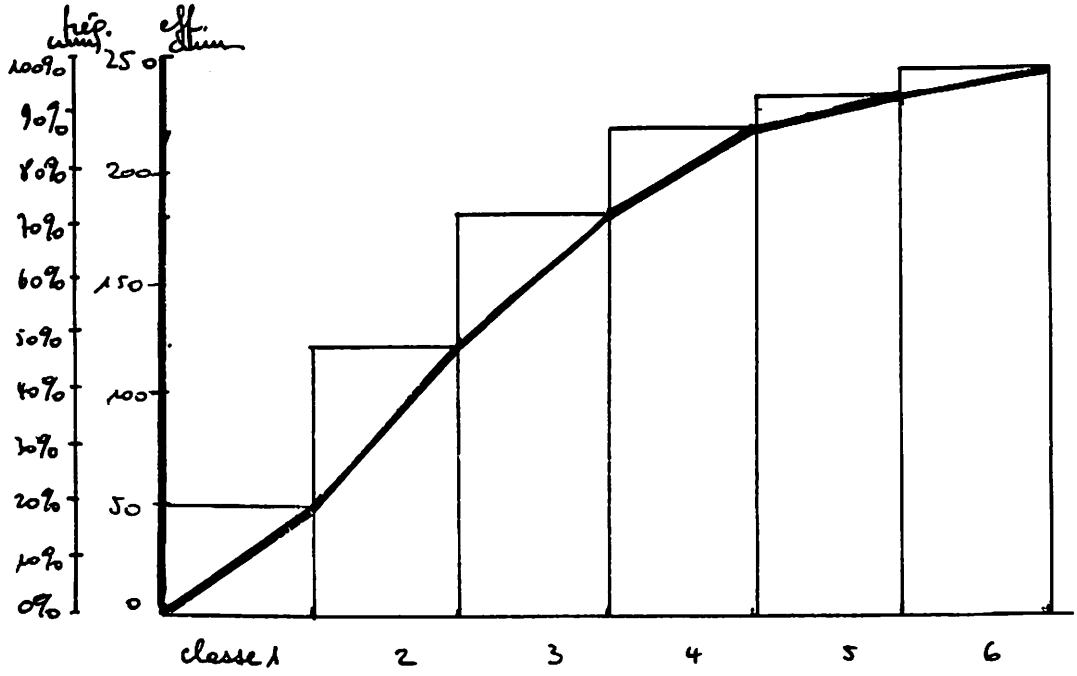
Exemple des notes dans une classe (I-5)



Cette courbe s'appelle la courbe des effectifs cumulés, resp. des fréq. cumulées. Elle permet par simple lecture graphique de connaître le nombre de notes inférieures ou égales à une valeur donnée.

2) Variable statistique continue

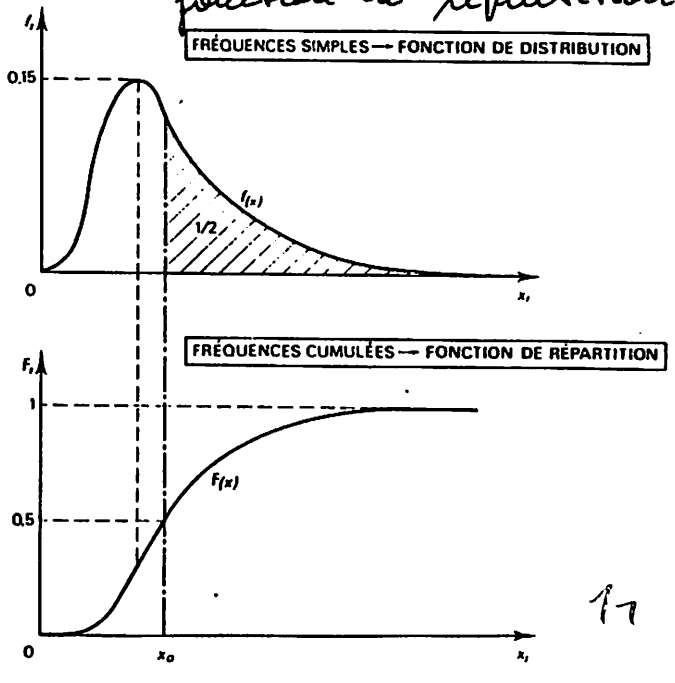
Exemple des cours de l'action d'une société pendant une année. (II-2)



polygone des effectifs (fréquences) cumulés.

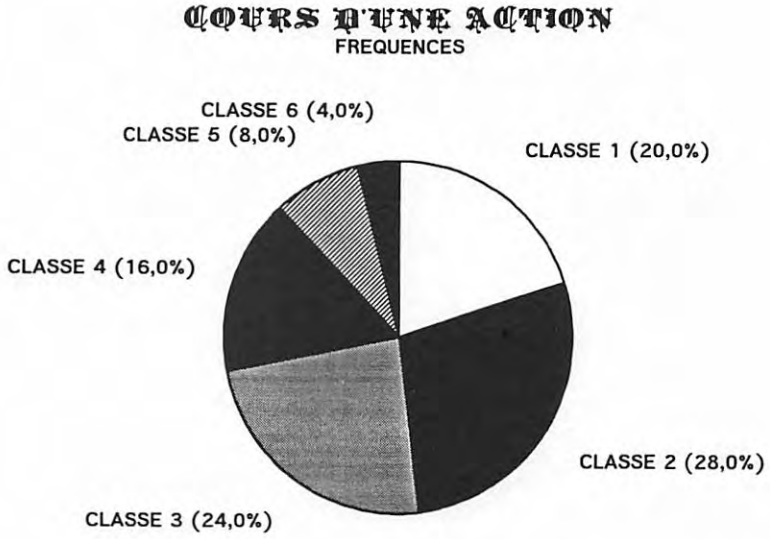
Il se construit en portant . les points correspondants à chaque classe à la limite supérieure de l'intervalle de classe. On suppose que, dans chaque classe, les effectifs sont répartis de manière régulière.

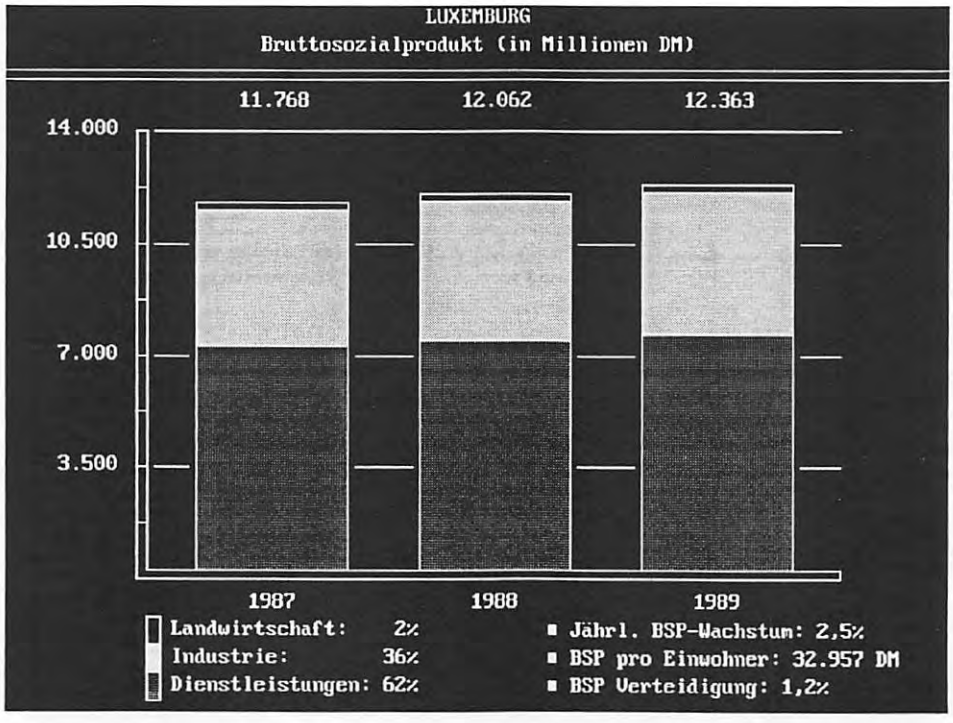
3) Lorsque l'intervalle des classes est très petit et les données nombreuses, la ligne brisée des fréquences cumulées tend à devenir une courbe, représentative de la fonction de répartition et notée F.



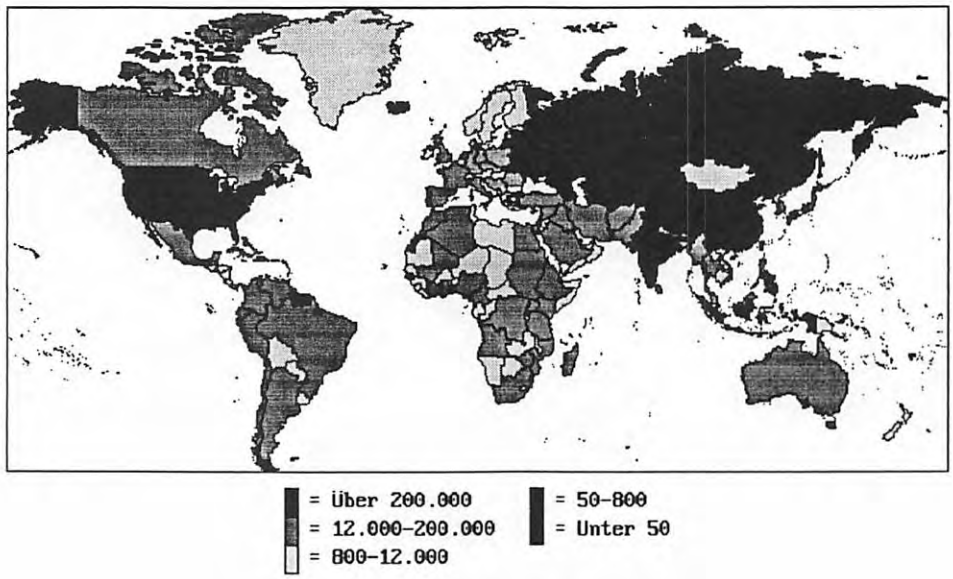
V

Représentations graphiques diverses

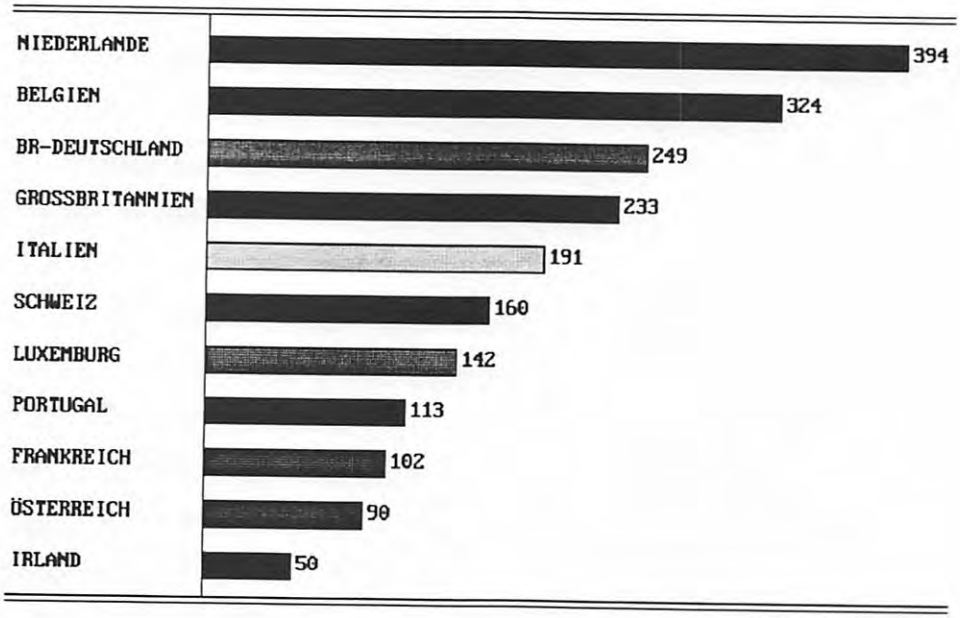




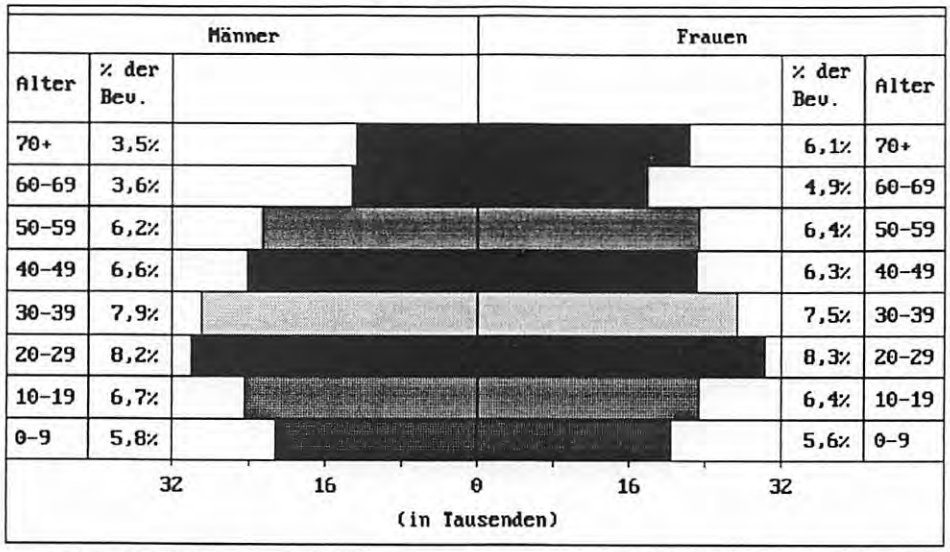
BEVÖLKERUNG 2000
(in Tausenden)



BEVÖLKERUNGSDICHTE
(pro km²)

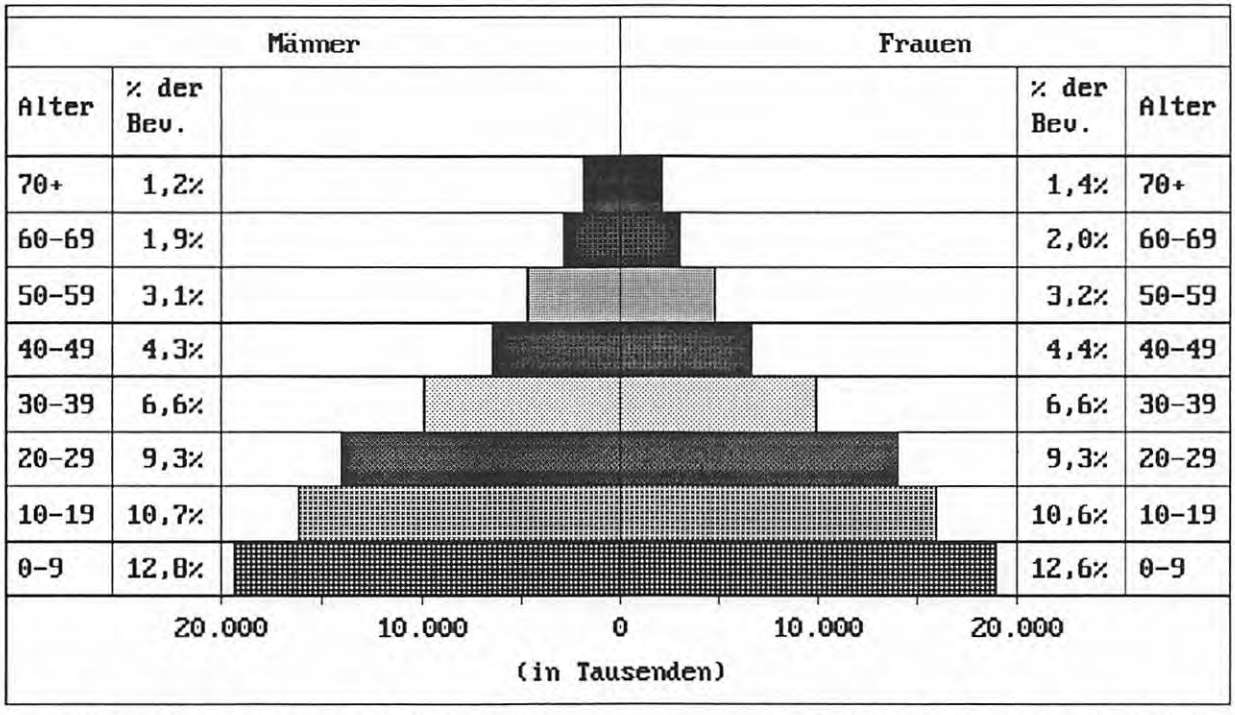


LUXEMBURG
Altersverteilung



- Gesamtbevölkerung: 366.000
- Männliche Gesamtbevölkerung: 178.000
- Weibliche Gesamtbevölkerung: 188.000
- Alphabetisierungsrate: 100%
- Urbanisierung: 78,0%

BRASILIEN
Altersverteilung



- Gesamtbevölkerung: 150.685.000
- Männliche Gesamtbevölkerung: 75.192.000
- Weibliche Gesamtbevölkerung: 75.493.000
- Alphabetisierungsrate: 76%
- Urbanisierung: 70,8%

VI Caractéristiques de tendance centrale et de position

1) Le mode

C'est la valeur de la variable correspondant à l'effectif ou à la fréquence maximale

exemple d'une variable discontinue

NOTE	EFFECTIF
0	0
1	1
2	2
3	0
4	2
5	0
6	1
7	2
mode → 8	5
9	2
10	4
11	3
12	3
13	1
14	2
15	1
16	3
17	0
18	2
19	0
20	1

notes de 1 à 20 obtenues par les 35 élèves d'une classe

← Cette série admet une valeur à effectif maximal. Elle est unimodale.

Le diagramme en bâtons y relatif permet aussi de déterminer le mode.

Si la série possède deux valeurs admettant des effectifs maximaux égaux, elle est appelée série bimodale.

	NOTE	EFFECTIF	
	0	0	
	1	1	
	2	2	
	3	0	
	4	2	
	5	0	
	6	1	
	7	2	
mode →	8	5	<—
	9	2	
	10	4	
	11	1	
	12	1	
mode →	13	5	
	14	2	
	15	1	
	16	3	
	17	0	
	18	2	
	19	0	
	20	1	

autres notes de 1 à 20 obtenues par les 35 élèves d'une classe

elle série admet deux valeurs à effectifs maximaux. Elle est bimodale.

Exemple d'une variable continue

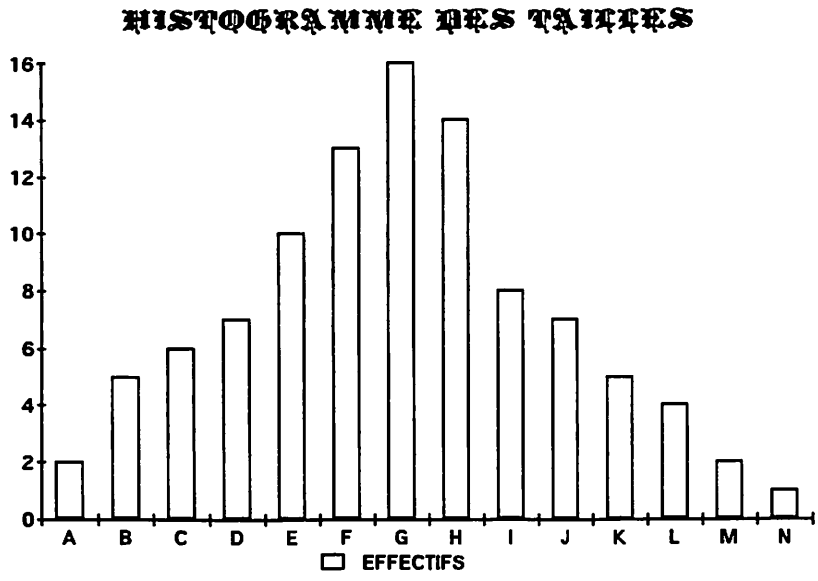
tailles de 100 élèves d'un lycée

TAILLES EN CM	137-140	141-144	145-148	149-152	153-156	157-160	161-164	165-168	169-172	173-176	177-180	181-184	185-188	189-192
CLASSES	A	B	C	D	E	F	G	H	I	J	K	L	M	N
EFFECTIFS	2	5	6	7	10	13	16	14	8	7	5	4	2	1

↑

Le mode est ici la classe G représentant des tailles allant de 161 à 164 ou plutôt de 160,5 à 164,49 cm.

L'histogramme des tailles nous donne le même renseignement.



2) La médiane

La médiane est la valeur de la variable qui partage l'effectif en deux parties égales.

Dans le cas d'une variable discontinue, deux cas peuvent se présenter :

* la série statistique possède un nombre impair de termes :

x_i	n_i
x_1	4
x_2	6
x_3	8
x_4	10
x_5	13
total	41

Le 21.ème représente la médiane. Il s'agit de x_4.

* la série statistique possède un nombre pair de termes :

x_i	n_i
x_1	4
x_2	5
x_3	6
x_4	7
x_5	8
total	30

Dans ce cas on peut choisir pour médiane tout élément de $]x_3, x_4[$.
On prend souvent le milieu $\frac{x_3+x_4}{2}$.

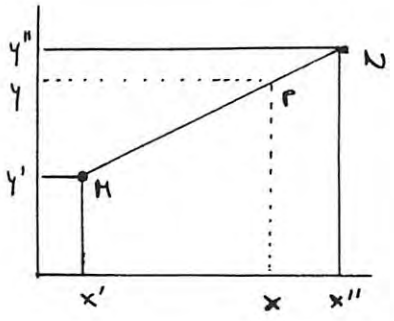
Dans le cas d'une <u>variable continue</u>, la valeur de la médiane peut se déterminer, soit par le calcul, soit graphiquement.

En reprenant l'exemple des tailles de 100 élèves d'un lycée (VI - 1)

TAILLES EN CM	137-140	141-144	145-148	149-152	153-156	157-160	161-164	165-168	169-172	173-176	177-180	181-184	185-188	189-192
CLASSES	A	B	C	D	E	F	G	H	I	J	K	L	M	N
EFFECTIFS	2	5	6	7	10	13	16	14	8	7	5	4	2	1
FREQUENCES	0,02	0,05	0,06	0,07	0,1	0,13	0,16	0,14	0,08	0,07	0,05	0,04	0,02	0,01
FREQ. CUMULEES	0,02	0,07	0,13	0,2	0,3	0,43	0,59	0,73	0,81	0,88	0,93	0,97	0,99	1

La classe médiane est la classe G. On procède ensuite par interpolation linéaire en supposant l'effectif uniformément réparti dans la classe. On obtient ainsi une valeur approchée du réel x tel que $F(x) = 1/2$.

Rappel sur l'interpolation linéaire :



$$P \in (MN) \iff \begin{vmatrix} x-x' & x''-x' \\ y-y' & y''-y' \end{vmatrix} = 0$$

$$\iff (y''-y')(x-x') - (x''-x')(y-y') = 0$$

y étant donné (0,5 dans nos exemples sur les médianes) cherchons x :

$$(y''-y')(x-x') = (x''-x')(y-y')$$

$$x-x' = \frac{(x''-x')}{y''-y'} \cdot (y-y')$$

$$\boxed{x = x' + (y-y') \cdot \frac{x''-x'}{y''-y'}}$$

Dans notre exemple : $\begin{cases} x' = 160,5 \\ y' = 0,43 \end{cases}$ $\begin{cases} x'' = 164,5 \\ y'' = 0,59 \end{cases}$

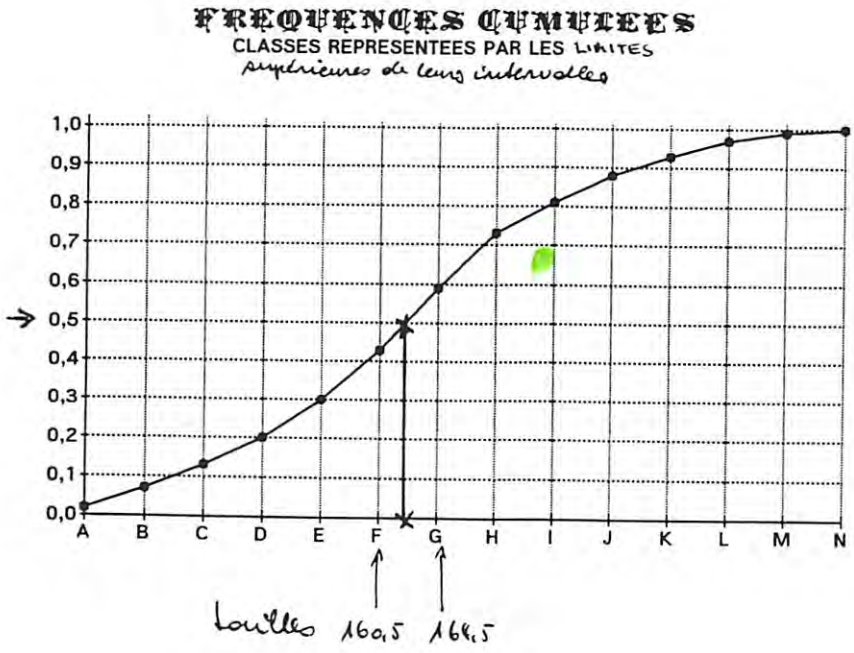
$$x = 160,5 + (0,5-0,43) \cdot \frac{164,5-160,5}{0,59-0,43}$$

$$= 160,5 + 0,07 \cdot \frac{4}{0,16}$$

$$= 160,5 + 1,75$$

$$= 162,25 \text{ cm}.$$

Nous retrouvons une valeur approchée de la médiane à l'aide du polygone des fréquences cumulées en y recherchant l'abscisse du point ayant pour ordonnée 0,5.



3) Les quartiles

Les quartiles sont les valeurs de la variable qui partagent l'effectif, rangé par ordre croissant, en quatre parties égales. Il existe trois quartiles : Q_1, Q_2, Q_3 correspondant aux fréquences cumulées respectives : 0,25 ; 0,50 ; 0,75. Q_2 est la médiane.

En reprenant l'exemple des tailles de 100 élèves on constate que Q_1 est dans la classe E, Q_2 est la médiane, Q_3 est dans la classe I.

Par interpolation linéaire on détermine Q_1 et puis Q_3 :

Q_1 : $\begin{cases} x' = 152,5 \\ y' = 0,2 \end{cases}$ $\begin{cases} x'' = 156,5 \\ y'' = 0,3 \end{cases}$

$$x = x' + (y - y') \cdot \frac{x'' - x'}{y'' - y'} = 152,5 + (0,25 - 0,2) \cdot \frac{156,5 - 152,5}{0,3 - 0,2}$$

$$= 152,5 + 0,05 \cdot \frac{4}{0,1} = 154,5$$

4) Les déciles et les centiles

Les déciles (les centiles) sont les valeurs de la variable qui partagent l'effectif, rangé par ordre croissant, en 10 (resp. 100) parties égales. Il existe 9 déciles D_1 à D_9 (99 centiles C_1 à C_{99}).
D_i correspond à une fréquence cumulée de $i.10\,\%$.
C_i correspond à une fréquence cumulée de $i\,\%$.

Pour le calcul et la détermination graphique, on se reporte aux quartiles.

5) Moyenne arithmétique

a) Dans l'exemple des notes de 1 à 20 obtenues par les 35 élèves d'une classe on a:

(1)

NOTE	EFFECTIF	NOTExEFFEC
0	0	0
1	1	1
2	2	4
3	0	0
4	2	8
5	0	0
6	1	6
7	2	14
8	5	40
9	2	18
10	4	40
11	3	33
12	3	36
13	1	13
14	2	28
15	1	15
16	3	48
17	0	0
18	2	36
19	0	0
20	1	20
MOYENNE:		10,285714

La moyenne vaut :

$$\bar{x} = \frac{\sum_{i=1}^{n} n_i x_i}{\sum_{i=1}^{n} n_i = N \text{ effectif total = constante}} \quad (1)$$

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

$$= \sum_{i=1}^{n} f_i x_i \quad (2)$$

f_i : fréquence

(c)

NOTE	EFFECTIF	NOTExEFFECTIF	FREQUENCE	NOTExFREQUENCE
0	0	0	0,00000	0,00000
1	1	1	0,02857	0,02857
2	2	4	0,05714	0,11429
3	0	0	0,00000	0,00000
4	2	8	0,05714	0,22857
5	0	0	0,00000	0,00000
6	1	6	0,02857	0,17143
7	2	14	0,05714	0,40000
8	5	40	0,14286	1,14286
9	2	18	0,05714	0,51429
10	4	40	0,11429	1,14286
11	3	33	0,08571	0,94286
12	3	36	0,08571	1,02857
13	1	13	0,02857	0,37143
14	2	28	0,05714	0,80000
15	1	15	0,02857	0,42857
16	3	48	0,08571	1,37143
17	0	0	0,00000	0,00000
18	2	36	0,05714	1,02857
19	0	0	0,00000	0,00000
20	1	20	0,02857	0,57143
MOYENNE:		10,285714		10,285714

b) Il est souvent commode d'effectuer un changement d'origine et/ou d'échelle, pour calculer la moyenne.

On pose $y_i = a x_i + b$

$$\bar{y} = \sum_{i=1}^{M} f_i y_i = \sum_{i=1}^{M} f_i (a x_i + b) = \sum_{i=1}^{M} a f_i x_i + \sum_{i=1}^{M} f_i b$$

$$= a \sum_{i=1}^{M} f_i x_i + b \underbrace{\sum_{i=1}^{M} f_i}_{\text{somme des fréquences} = 1} = a \bar{x} + b$$

Dans l'exemple ci-dessus, en posant $y_i = x_i - 10$, où 10 est la médiane on obtient:

$\bar{y} = 0,285714$

d'où: $\bar{x} = \bar{y} + 10 = 10,285714$

NOTE xi	EFFECTIF ni	FREQUENCE fi	yi	fiXyi
0	0	0,00000	-10	0,00000
1	1	0,02857	-9	-0,25714
2	2	0,05714	-8	-0,45714
3	0	0,00000	-7	0,00000
4	2	0,05714	-6	-0,34286
5	0	0,00000	-5	0,00000
6	1	0,02857	-4	-0,11429
7	2	0,05714	-3	-0,17143
8	5	0,14286	-2	-0,28571
9	2	0,05714	-1	-0,05714
10	4	0,11429	0	0,00000
11	3	0,08571	1	0,08571
12	3	0,08571	2	0,17143
13	1	0,02857	3	0,08571
14	2	0,05714	4	0,22857
15	1	0,02857	5	0,14286
16	3	0,08571	6	0,51429
17	0	0,00000	7	0,00000
18	2	0,05714	8	0,45714
19	0	0,00000	9	0,00000
20	1	0,02857	10	0,28571

$\bar{y} = 0,285714$

c) Lorsque la statistique porte sur des classes, on convient de considérer le milieu de chaque classe, affecté de la fréquence de cette classe.

TAILLES EN CM	137-140	141-144	145-148	149-152	153-156	157-160	161-164	165-168	169-172	173-176	177-180	181-184	185-188	189-192
CLASSES	A	B	C	D	E	F	G	H	I	J	K	L	M	N
MILIEU xi	138,5	142,5	146,5	150,5	154,5	158,5	162,5	166,5	170,5	174,5	178,5	182,5	186,5	190,5
EFFECTIFS ni	2	5	6	7	10	13	16	14	8	7	5	4	2	1
FREQUENCES fi	0,02	0,05	0,06	0,07	0,1	0,13	0,16	0,14	0,08	0,07	0,05	0,04	0,02	0,01
fi X xi	2,77	7,125	8,79	10,535	15,45	20,605	26	23,31	13,64	12,215	8,925	7,3	3,73	1,905
MOYENNE	162,3													

6) Autres moyennes

a) moyenne géométrique

$$G = \prod_{i=1}^{n} x_i^{f_i} = \sqrt[N]{\prod_{i=1}^{n} x_i^{n_i}}$$ où N est l'effectif total

Exemple : Supposons que pendant 4 années les prix augmentent de la façon suivante :

Année	1	2	3	4
augmentation	4%	5%	4%	3%

Chacune des augmentations est calculée sur l'année précédente. Calculons l'augmentation sur les 4 ans. Un produit qui valait au départ 1 € évolue de la manière suivante :

Année	1	2	3	4
prix du produit à la fin de l'année	1,04	1,092	1,13568	1,1697504

L'augmentation moyenne des prix est donnée par :

$$1 \cdot x^4 = 1 \cdot 1{,}04 \cdot 1{,}05 \cdot 1{,}04 \cdot 1{,}03$$

$$x^4 = 1{,}04 \cdot 1{,}05 \cdot 1{,}04 \cdot 1{,}03$$

$$x = \sqrt[4]{1{,}04 \cdot 1{,}05 \cdot 1{,}04 \cdot 1{,}03}$$

x est la moyenne géom. des indices annuels 1,04 ; 1,05 ; 1,04 ; 1,03.

$$x = \sqrt[4]{1{,}1697504}$$

$$x \approx 1{,}039975961$$

L'augmentation moyenne des prix est de 3,9975961 % environ.

b) **Moyenne harmonique**

On appelle moyenne harmonique de la série statistique le réel H défini par :

$$\frac{1}{H} = \sum_{i=1}^{n} \frac{f_i}{x_i} = \frac{1}{N} \sum_{i=1}^{n} \frac{n_i}{x_i}$$

Exemple :

Une voiture parcourt la distance de A à B à la vitesse moyenne de 75 km/h et la distance de B à A à la vitesse moyenne de 90 km/h. Quelle est la vitesse moyenne sur le trajet aller-retour ?

$V = \frac{d}{t}$; $V_a = 75 = \frac{d}{t}$ à l'aller

$V_r = 90 = \frac{d}{t'}$ au retour.

La moyenne aller-retour est :

$$V = \frac{2d}{t+t'} \Leftrightarrow \frac{1}{V} = \frac{t+t'}{2d} = \frac{1}{2}\left(\frac{t}{d} + \frac{t'}{d}\right)$$

$$\frac{1}{V} = \frac{1}{2}\left(\frac{1}{75} + \frac{1}{90}\right)$$

C'est la moyenne harmonique des vitesses à l'aller et au retour.

$$\frac{1}{V} = \frac{1}{2} \cdot 0,2\overline{7} = 0,1\overline{3}$$

$$V = 81,\overline{81} \text{ km/h}.$$

c) **Moyenne quadratique**

$$Q = \sqrt{\sum_{i=1}^{n} f_i x_i^2}$$

d) On peut démontrer qu'on a toujours :
$H \leq G \leq \bar{x} \leq Q$. De plus, si deux des quatre nombres sont égaux, alors ils sont tous égaux.

V II

Caractéristiques de dispersion

1) L'étendue

L'étendue d'une série statistique est la différence entre la plus grande et la plus petite valeur du critère.
Dans l'ex. des tailles des 100 élèves l'étendue vaut : 191,5 - 134,5 = 56 cm
L'étendue est très sensible aux variations des valeurs extrêmes qui sont souvent peu représentatives.

2) L'intervalle interquartile

L'intervalle interquartile d'une série statistique est égal à la différence $Q_3 - Q_1$.
Dans l'exemple des tailles de 100 élèves on a :
$Q_3 - Q_1 = 169,5 - 154,5 = 15$ cm

On trouve dans cet intervalle 50 % des observations centrées autour de la médiane. Plus l'intervalle est réduit, plus la concentration autour des valeurs centrales est forte.

3) L'intervalle interdécile

L'intervalle interdécile d'une série statistique est égal à $D_9 - D_1$. Dans cet intervalle se trouvent 80 % des observations. Il ne tient pratiquement pas compte des termes extrêmes.
Dans l'exemple des tailles de 100 élèves on a :

D_1 est dans la classe C :

$$x = 144,5 + (0,1 - 0,07) \cdot \frac{154,5 - 144,5}{0,13 - 0,07} = 146,5 \text{ cm}$$

D_9 est dans la classe K :

$$x = 176,5 + (0,9 - 0,88) \cdot \frac{160,5 - 176,5}{0,93 - 0,88} = 178,1 \text{ cm}$$

$D_9 - D_1 = 178,1 - 146,5 = 31,6$ cm

4) **L'écart absolu moyen**

L'écart absolu moyen est la moyenne arithmétique des écarts, en valeur absolue, à la moyenne arithmétique.

$$E_{\bar{x}} = \frac{\sum_{i=1}^{M} |x_i - \bar{x}| \cdot m_i}{\sum_{i=1}^{M} m} = \frac{\sum_{i=1}^{M} |x_i - \bar{x}| \, m_i}{N} \qquad (1)$$

$$E_{\bar{x}} = \sum_{i=1}^{M} f_i \, |x_i - \bar{x}| \qquad (2)$$

(1) Exemple des tailles des élèves

TAILLES EN CM	137-140	141-144	145-148	149-152	153-156	157-160	161-164	165-168	169-172	173-176	177-180	181-184	185-188	189-192
CLASSES	A	B	C	D	E	F	G	H	I	J	K	L	M	N
MILIEU xi	138,5	142,5	146,5	150,5	154,5	158,5	162,5	166,5	170,5	174,5	178,5	182,5	186,5	190,5
EFFECTIFS ni	2	5	6	7	10	13	16	14	8	7	5	4	2	1
MOYENNE $\bar{x}$	162,3													
$\lvert xi-\bar{x}\rvert$	23,8	19,8	15,8	11,8	7,8	3,8	0,2	4,2	8,2	12,2	16,2	20,2	24,2	28,2
$ni\lvert xi-\bar{x}\rvert$	47,6	99	94,8	82,6	78	49,4	3,2	58,8	65,6	85,4	81	80,8	48,4	28,2
$\Sigma ni\lvert xi-\bar{x}\rvert$	902,8													
ECART ABS.MOYEN	9,028													

(2) Exemple des notes de 35 élèves

NOTE xi	EFFECTIF ni	FREQUENCE fi	xiXfi	fiX$\lvert$xi-$\bar{x}\rvert$
0	0	0,00000	0,00000	0,00000
1	1	0,02857	0,02857	0,26531
2	2	0,05714	0,11429	0,47347
3	0	0,00000	0,00000	0,00000
4	2	0,05714	0,22857	0,35918
5	0	0,00000	0,00000	0,00000
6	1	0,02857	0,17143	0,12245
7	2	0,05714	0,40000	0,18776
8	5	0,14286	1,14286	0,32653
9	2	0,05714	0,51429	0,07347
10	4	0,11429	1,14286	0,03265
11	3	0,08571	0,94286	0,06122
12	3	0,08571	1,02857	0,14694
13	1	0,02857	0,37143	0,07755
14	2	0,05714	0,80000	0,21224
15	1	0,02857	0,42857	0,13469
16	3	0,08571	1,37143	0,48980
17	0	0,00000	0,00000	0,00000
18	2	0,05714	1,02857	0,44082
19	0	0,00000	0,00000	0,00000
20	1	0,02857	0,57143	0,27755
MOYENNE: $\bar{x}$			0,28571	
ECART ABSOLU MOYEN: $E_{\bar{x}}$				3,68163

5) La variance, l'écart-type

L'écart absolu moyen est exprimé en valeur absolue pour éviter la compensation due aux signes des différents termes. En effet:

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

$$= \bar{x} - \bar{x} = 0.$$

Pour éviter cette compensation il y a un 2. moyen qui consiste à élever les écarts au carré.

La variance d'une série statistique est la moyenne arithmétique des carrés des écarts $(x_i - \bar{x})^2$.

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

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

$$\sigma = \sqrt{V}$$

La détermination de V et σ peut se faire de façon directe ou à l'aide du théorème de König:

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

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

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

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

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

Calcul direct sur l'exemple des notes de 35 élèves :

NOTE x_i	EFFECTIF n_i	$n_i X x_i$	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	$n_i X (x_i - \bar{x})^2$
0	0	0	-10,28571	105,79592	0,00000
1	1	1	-9,28571	86,22449	86,22449
2	2	4	-8,28571	68,65306	137,30612
3	0	0	-7,28571	53,08163	0,00000
4	2	8	-6,28571	39,51020	79,02041
5	0	0	-5,28571	27,93878	0,00000
6	1	6	-4,28571	18,36735	18,36735
7	2	14	-3,28571	10,79592	21,59184
8	5	40	-2,28571	5,22449	26,12245
9	2	18	-1,28571	1,65306	3,30612
10	4	40	-0,28571	0,08163	0,32653
11	3	33	0,71429	0,51020	1,53061
12	3	36	1,71429	2,93878	8,81633
13	1	13	2,71429	7,36735	7,36735
14	2	28	3,71429	13,79592	27,59184
15	1	15	4,71429	22,22449	22,22449
16	3	48	5,71429	32,65306	97,95918
17	0	0	6,71429	45,08163	0,00000
18	2	36	7,71429	59,51020	119,02041
19	0	0	8,71429	75,93878	0,00000
20	1	20	9,71429	94,36735	94,36735

MOYENNE: $\bar{x}$ = 10,28571

VARIANCE: V = 21,46122

ECART-TYPE: σ = 4,63263

Calcul à l'aide du théorème de König :

NOTE x_i	EFFECTIF n_i	$n_i X x_i$	$n_i X x_i^2$
0	0	0	0
1	1	1	1
2	2	4	8
3	0	0	0
4	2	8	32
5	0	0	0
6	1	6	36
7	2	14	98
8	5	40	320
9	2	18	162
10	4	40	400
11	3	33	363
12	3	36	432
13	1	13	169
14	2	28	392
15	1	15	225
16	3	48	768
17	0	0	0
18	2	36	648
19	0	0	0
20	1	20	400

MOYENNE: 10,28571
CARRE DE LA MOYENNE: 105,79592

VARIANCE: 21,46122

ECART-TYPE: 4,63263

Remarque : en posant $y_i = ax_i + b$ on a :

$$V_y = \sum_{i=1}^{n} f_i (y_i - \bar{y})^2$$

$$= \sum_{i=1}^{n} f_i (ax_i + b - a\bar{x} - b)^2$$

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

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

$$V_y = a^2 V_x$$

$$\sigma_y = |a| \sigma_x$$

Refaire le calcul de la variance et de l'écart-type pour l'exemple des notes des 35 élèves en posant $y_i = x_i - 10$.

Commentaires

L'écart-type (et la variance) indique comment en moyenne, les valeurs de la variable sont groupées autour de la tendance centrale. Si l'écart-type est faible, les valeurs de la série sont peu dispersées.

VIII Statistique double

1) Dans l'observation d'une population nous allons nous intéresser à deux caractères au lieu d'un seul

Exemple : Soit X la production de fonte, Y la production d'acier en France, Italie, Grande-Bretagne et en Allemagne :

	D	F	GB	I
X	21,2	15,9	17,6	3,5
Y	37,3	19,8	26,7	9,8

Tout comme pour une statistique simple, on peut regrouper les données suivant les valeurs des x_i et y_i.

X / Y	3,5	15,9	17,6	21,2	Totaux
9,8	1	0	0	5	6
19,8	2	1	0	0	3
26,7	0	0	1	0	1
37,3	0	3	0	1	4
Totaux	3	4	1	6	4

N = 16

Soit en général :

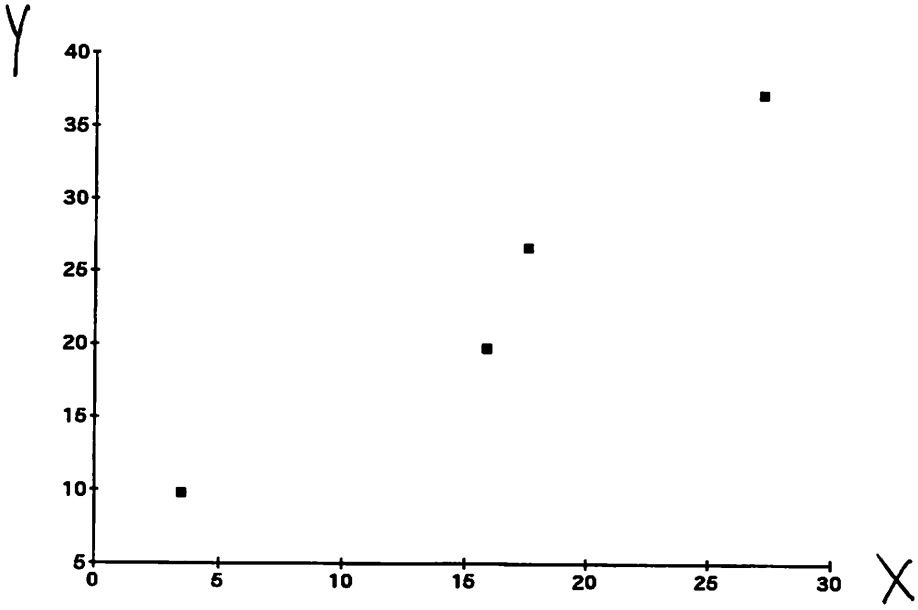
Y \ X	x_1 ...	x_i ...	x_M	Total
y_1				
y_j	...	n_{ij}	...	$\sum_i n_{ij} = n_{\cdot j}$
y_P				
Total		$\sum_j n_{ij} = n_{i\cdot}$		

32

On peut définir la fréquence $f_{ij} = \frac{n_{ij}}{N}$ où N est l'effectif total.

$N = \sum_{i,j} n_{ij}$ et $\sum_{i,j} f_{ij} = 1$.

Représentation graphique :



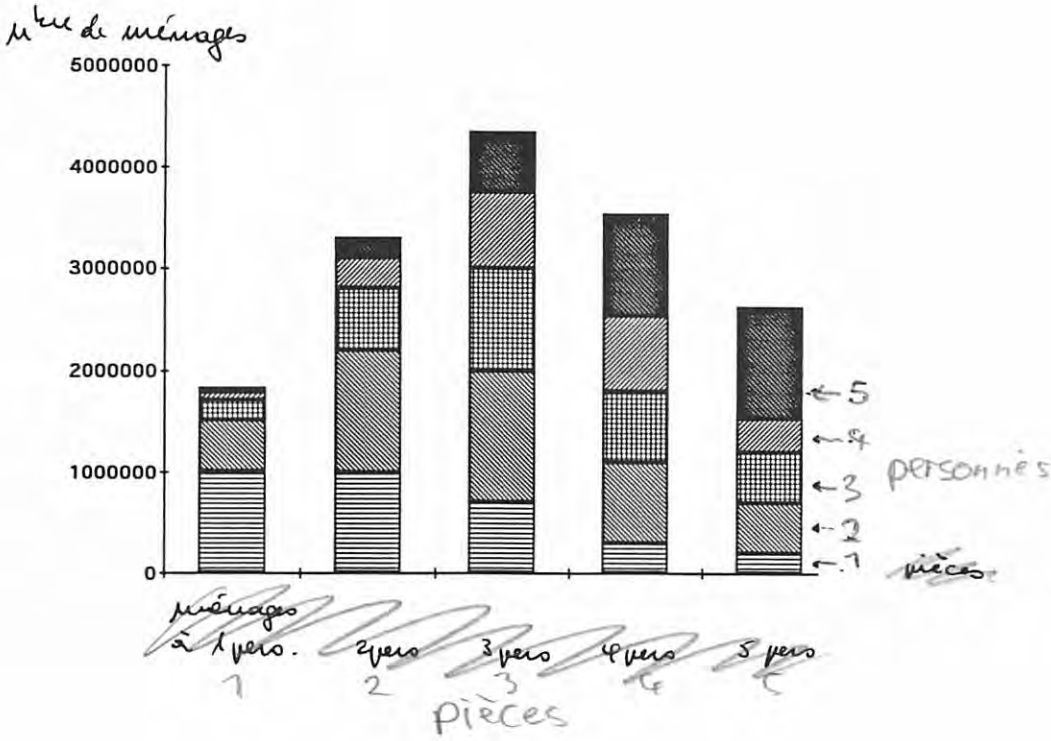
2) Soit une Statistique double donnée par x_i, y_i et les f_{ij}. On appelle statistique marginale du caractère X la statistique donnée par x_i et les $f_{i\cdot} = \sum_j f_{ij}$. On appelle statistique marginale du caractère Y la statistique donnée par y_i et les $f_{\cdot j} \sum_i f_{ij}$. Les f_{xi} et f_{yj} sont appelées fréquences marginales.

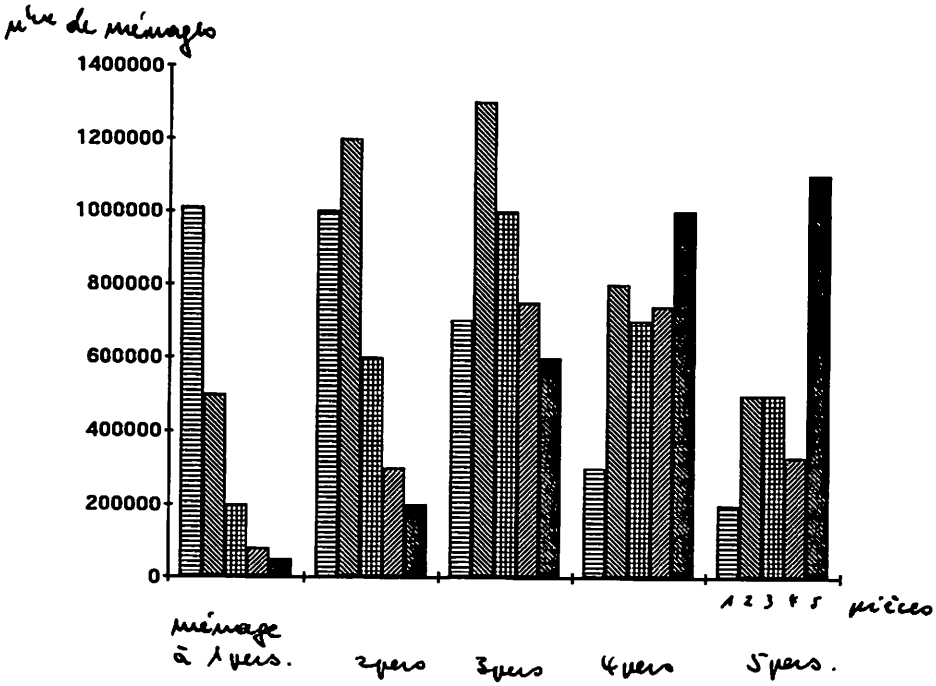
Exemple :

Statistique concernant les ménages suivant le nombre de pièces habitées (X) et le nombre de personnes du ménage (Y).

Y Nombre de personnes / X Nombre de pièces	x_1 1	x_2 2	x_3 3	x_4 4	x_5 5	Total
y_1 1	1 010 000	1 000 000	700 000	300 000	200 000	3 210 000
y_2 2	500 000	1 200 000	1 300 000	800 000	500 000	4 300 000
y_3 3	200 000	600 000	1 000 000	700 000	500 000	3 000 000
y_4 4	80 000	300 000	750 000	740 000	330 000	2 200 000
y_5 5	50 000	200 000	600 000	1 000 000	1 100 000	2 950 000
Total	1 840 000	3 300 000	4 350 000	3 540 000	2 630 000	15 660 000

Représentations graphiques :





Statistique marginale concernant le caractère X

x_i	$\mu_{i\cdot}$	$f_{i\cdot}$
1	1840.000	0,11750
2	3.300.000	0,21073
3	4.350.000	0,27778
4	3.540.000	0,22605
5	2.630.000	0,16794

total: 1

3) Valeurs caractéristiques d'une statistique double.

a) On appelle covariance des caractères X et Y le réel noté cov (x,y) égal à

$$\sum_{i,j} f_{ij}(x_i - \bar{x})(y_j - \bar{y})$$

Remarque :

Si l'effectif est faible et si on connaît les caractères X et Y pour chaque individu, on peut exprimer la covariance plus simplement par

$$\text{cov}(x,y) = \frac{1}{N} \cdot \sum_i (x_i - \bar{x})(y_i - \bar{y})$$

(cf exemple de la partie 1)

Propriétés :

* $\text{cov}(x,y)$

$$= \sum_{i,j} f_{ij}\, x_i y_j - \bar{x} \sum_{i,j} f_{ij}\, y_j - \bar{y} \sum_{i,j} f_{ij}\, x_i + \bar{x}\bar{y} \sum_{i,j} f_{ij}.$$

$$= \sum_{i,j} f_{ij}\, x_i y_j - \bar{x} \sum_j \left(\sum_i f_{ij} \right) y_j - \bar{y} \sum_i \left(\sum_j f_{ij} \right) x_i + \bar{x}\bar{y}$$

$$= \sum_{i,j} f_{ij}\, x_i y_j - \bar{x} \cdot \sum_j f_{\cdot j}\, y_j - \bar{y} \cdot \sum_i f_{i\cdot}\, x_i + \bar{x}\bar{y}$$

$$= \sum_{i,j} f_{ij}\, x_i y_j - \bar{x}\bar{y} - \bar{y} \cdot \bar{x} + \bar{x}\bar{y}$$

$$= \sum_{i,j} f_{ij}\, x_i y_j - \bar{x}\bar{y}.$$

$$= \frac{1}{N} \sum_{i,j} \mu_{ij}\, x_i y_j - \bar{x}\bar{y}$$

* En posant : $x'_i = a x_i + b$
$y'_i = c y_i + d$

on a :

$cov(x', y') =$

$$= \sum_{i,j} f_{ij} (a x_i + b - a\bar{x} - b)(c y_i + d - c\bar{y} - d)$$

$$= \sum_{i,j} f_{ij}\, a(x_i - \bar{x}) \cdot c(y_i - \bar{y})$$

$$= ac \sum_{i,j} f_{ij} (x_i - \bar{x})(y_i - \bar{y})$$

$$= ac \cdot cov(x, y)$$

Remarque :
Lorsqu'une statistique double est donnée par des classes, on travaillera sur les centres des classes.

b) On appelle coefficient de corrélation linéaire le réel noté $\rho(x,y)$ égal à : $\dfrac{cov(x,y)}{\sigma_x \cdot \sigma_y}$

$$\rho(x,y) = \frac{cov(x,y)}{\sigma_x \cdot \sigma_y}$$

c) Calcul pratique de la covariance et du coefficient de corrélation linéaire sur l'exemple de la partie 2)

Statistique concernant les ménages suivant le nombre de pièces habitées (X) et le nombre de personnes du ménage (Y)

y \ x	1	2	3	4	5	$n_{i,j}$	$n_{i,j}\cdot y_j$	$n_{\cdot,j}\cdot (y_j)^2$	$\sum n_{ij}\cdot x_i$	$(\sum n_{ij}\cdot x_i)\cdot y_j$	TOTAL $1\le i\le 5$
1	1010000	1000000	700000	300000	200000	3210000	3210000	3210000	7310000	7310000	
2	500000	1200000	1300000	800000	500000	4300000	8600000	17200000	12500000	25000000	
3	200000	600000	1000000	700000	500000	3000000	9000000	27000000	9700000	29100000	
4	80000	300000	750000	740000	330000	2200000	8800000	35200000	7540000	30160000	
5	50000	200000	600000	1000000	1100000	2950000	14750000	73750000	11750000	58750000	
$n_{i,\cdot}$	1840000	3300000	4350000	3540000	2630000						15660000
$n_{i,\cdot}\cdot x_i$	1840000	6600000	13050000	14160000	13150000						48800000
$n_{i,\cdot}\cdot (x_i)^2$	1840000	13200000	39150000	56640000	65750000						176580000
$\sum n_{ij}\cdot y_j$	3180000	7400000	12300000	11960000	9520000						44360000
$(\sum n_{ij}\cdot y_j)\cdot x_i$	3180000	14800000	36900000	47840000	47600000						150320000
TOTAL $1\le j\le 5$						15660000	44360000	156360000	48800000	150320000	

$\bar{x} = \frac{4880}{1566} \simeq 3,11622$

$V(x) \simeq 1,56504$

$\sigma_x \simeq 1,25102$

$\bar{y} = \frac{4436}{1566} \simeq 2,83269$

$\sum_{j=1}^{5} \mu_{\cdot j}\cdot y_j^2 = \dots$

$V(y) = \frac{156360000}{15660000} - \bar{y}^2$

$\simeq 1,96051$

$\sigma_y \simeq 1,400\,18$

$\text{cov}(x,y) = \frac{1}{N}\sum_{i,j} \mu_{ij}\, x_i y_j - \bar{x}\bar{y} = \frac{150.320.000}{15660.000} - \bar{x}\bar{y} \simeq 0,77168$

$\rho(x,y) = \frac{\text{cov}(x,y)}{\sigma_x \sigma_y} = 0,44054$

Nous allons revenir à l'interprétation de ces résultats dans la prochaine partie.

IX Ajustement

1) Nuage de points

a) Sur une population donnée, on étudie deux variables quantitatives X et Y qui prennent les valeurs $x_1, \dots, x_N$ et $y_1, \dots, y_N$. Pour chaque individu i de la population on a un couple de valeurs (x_i, y_i). Par rapport aux calculs des parties précédentes on a : $\mu_{i\bullet} = 1$, $\mu_{\bullet j} = 1$, $\mu_{ij} = 1$ pour $i = j$ et $\mu_{ij} = 0$ pour $i \neq j$. (cf. 1. en. des Stat. doubles)

Les principales formules deviennent :

$$\bar{x} = \frac{\sum_{i=1}^{N} x_i}{N} \; ; \quad \bar{y} = \frac{\sum_{i=1}^{N} y_i}{N} \; ;$$

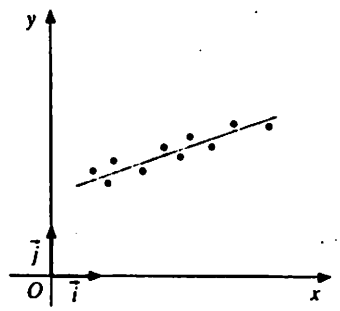
$$V(x) = \frac{\sum_{i=1}^{N} (x_i - \bar{x})^2}{N} \; ; \quad V(y) = \frac{\sum_{i=1}^{N} (y_i - \bar{y})^2}{N}$$

$$= \frac{\sum_{i=1}^{N} x_i^2}{N} - \bar{x}^2 \; ; \quad = \frac{\sum_{i=1}^{N} y_i^2}{N} - \bar{y}^2$$

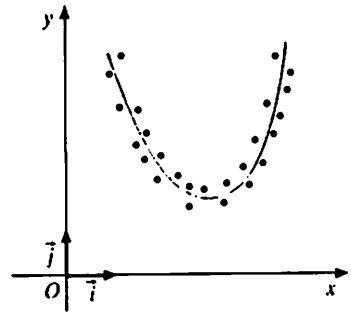
$$\mathrm{Cov}(x,y) = \frac{\sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y})}{N} = \frac{\sum_{i=1}^{N} x_i y_i}{N} - \bar{x}\,\bar{y}$$

b) Les nuages de points peuvent avoir différentes formes :

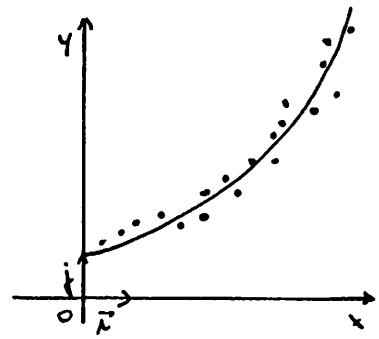
* les points sont répartis autour d'une droite. X et Y sont liés en gros par une relation du type $Y = aX + b$,



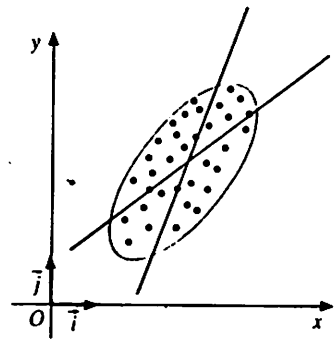
* les points sont répartis autour d'une parabole; on cherche une relation du type $y = aX^2 + bX + c$,



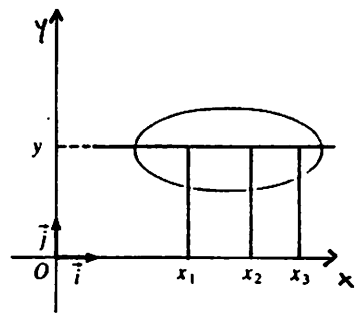
* les points sont répartis autour d'une courbe comme ; on cherche une relation du type $y = f(x)$,



* le nuage a la forme d'une ellipse ; c'est dans ce cas où on ajustera le nuage par deux droites dites de régression,



* si les axes de l'ellipse sont parallèles aux axes, les variables X et Y sont à priori indépendantes.



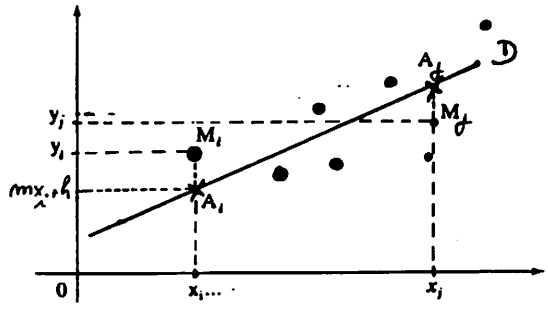
<u>Remarque</u> : Dans le cas où le nuage de points laisse entrevoir une dépendance entre X et Y de la forme $y = f(x)$, on fera un ajustement à l'aide de la courbe C représentative de f. L'intérêt d'un tel ajustement est de permettre d'effectuer des prévisions "dans un domaine raisonnable" (estimation des naissances dans un proche avenir, estimation de

la croissance d'une entreprise ...). Il suffira pour cela de dire que la courbe C est une bonne approximation de la série statistique en une valeur choisie.

2) Ajustement affine par la méthode des moindres carrés.

a) Droite de régression de y en x

Cette méthode consiste à trouver une droite D d'équation $y = mx + h$ telle que la somme des carrés des distances $M_i A_i$ $(1 \leq i \leq N)$ c.à.d. en termes mathématiques: $S = \sum_{i=1}^{N} d(M_i, A_i)^2$ soit minimum. Attention: il ne s'agit pas des distances des M_i $(1 \leq i \leq N)$ à D !



$$S = \sum_{i=1}^{N} d(M_i, A_i)^2 = \sum_{i=1}^{N} (y_i - m x_i - h)^2$$

Il faut trouver m et h tel que S soit minimal.
Les conditions à remplir sont:

$$\underbrace{\frac{\delta S}{\delta m} = 0 \text{ et } \frac{\delta S}{\delta h} = 0}_{①} \text{ et } \underbrace{\frac{\delta^2 S}{\delta m^2} \cdot \frac{\delta^2 S}{\delta h^2} - \left(\frac{\delta^2 S}{\delta m \delta h}\right)^2 > 0}_{②} \text{ et } \underbrace{\frac{\delta^2 S}{\delta m^2} > 0}_{③}$$

① • $\frac{\delta S}{\delta m} = +2 \sum_{i=1}^{N} (y_i - m x_i - h)(-x_i)$

$$= -2 \sum_{i=1}^{N} y_i x_i + 2m \sum_{i=1}^{N} x_i^2 + 2h \sum_{i=1}^{N} x_i$$

$$= -2N \left(\frac{\sum_{i=1}^{N} x_i y_i}{N} - m \frac{\sum_{i=1}^{N} x_i^2}{N} - h \frac{\sum_{i=1}^{N} x_i}{N} \right)$$

$$= -2N \left(\operatorname{cov}(x,y) + \bar{x}\bar{y} - m V(x) - m \bar{x}^2 - h \bar{x} \right)$$

$$\frac{\delta S}{\delta m} = 0 \iff \left(V(x) + \bar{x}^2\right) m + \bar{x}\, h = \operatorname{cov}(x,y) + \bar{x}\,\bar{y} \quad (I)$$

- $$\frac{\delta S}{\delta h} = -2 \sum_{i=1}^{N} \left(y_i - m x_i - h\right)$$

$$= -2N \left(\frac{\sum_{i=1}^{N} y_i}{N} - m \frac{\sum_{i=1}^{N} x_i}{N} - h \frac{\sum_{i=1}^{N} 1}{N} \right)$$

$$= -2N \left(\bar{y} - m\bar{x} - h \right)$$

$$\frac{\delta S}{\delta h} = 0 \iff \bar{x}\, m + h = \bar{y} \quad (II)$$

- Résolvons le système (I, II)

D'après (II) : $h = \bar{y} - m\bar{x}$

dans (I) :

$$\left(V(x) + \bar{x}^2\right) m + \bar{x}\bar{y} - m\bar{x}^2 = \operatorname{cov}(x,y) + \bar{x}\bar{y}$$

$$V(x)\, m = \operatorname{cov}(x,y)$$

$$m = \frac{\operatorname{cov}(x,y)}{V(x)}$$

$$h = \bar{y} - \frac{\operatorname{cov}(x,y)}{V(x)} \cdot \bar{x}$$

(2) $$\frac{\delta^2 S}{\delta m^2} = +2N\left(V(x) + \bar{x}^2\right)$$

$$\frac{\delta^2 S}{\delta h^2} = 2N$$

$$\frac{\delta^2 S}{\delta m\, \delta h} = 2N\bar{x}$$

- $$\frac{\delta^2 S}{\delta m^2} \cdot \frac{\delta^2 S}{\delta h^2} - \left(\frac{\delta^2 S}{\delta m \cdot \delta h}\right)^2 = 4N^2\left(V(x) + \bar{x}^2 - \bar{x}^2\right)$$

$$= 4N^2 \cdot V(x) > 0$$

③ $\frac{\delta^2 S}{\delta m^2} = 2N(V(x) + \bar{x}^2) > 0$

Donc toutes les conditions sont remplies.
L'équation de la droite de régression de y en x est :

$$y = mx + h = \frac{cov(x,y)}{V(x)} \cdot x + \bar{y} - \frac{cov(x,y)}{V(x)} \cdot \bar{x}$$

D : $y - \bar{y} = \frac{cov(x,y)}{V(x)} (x - \bar{x})$

Cette droite a pour pente $\frac{cov(x,y)}{V(x)}$ et passe par le point de coordonnées $(\bar{x}, \bar{y})$.

Pour le calcul pratique remarquons que :

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

b) Exemple :

i	1	2	3	4	5	6	7
x_i	15	26	38	52	72	91	100
y_i	382	359	333	310	277	241	219

$\bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i = \frac{1}{7}(15 + 26 + 38 + 52 + 72 + 91 + 100) \approx 56,28571$

$\bar{y} = \frac{1}{N} \sum_{i=1}^{N} y_i = \frac{1}{7}(382 + 359 + 333 + 310 + 277 + 241 + 219) = 303$

$\sum_{i=1}^{N} x_i y_i = 15 \cdot 382 + 26 \cdot 359 + 38 \cdot 333 + 52 \cdot 310 + 72 \cdot 277 + 91 \cdot 241 + 100 \cdot 219$
$= 107613$

$\sum_{i=1}^{N} x_i^2 = 15^2 + 26^2 + 38^2 + 52^2 + 72^2 + 91^2 + 100^2 = 28514$

$$\frac{cov(x,y)}{V(x)} = \frac{107\,613 - 7 \cdot 56,285714\ldots \cdot 303}{28\,514 - 7 \cdot (56,285714\ldots)^2}$$

$$= \frac{-1769}{6\,337,42857\ldots} = -1,85706$$

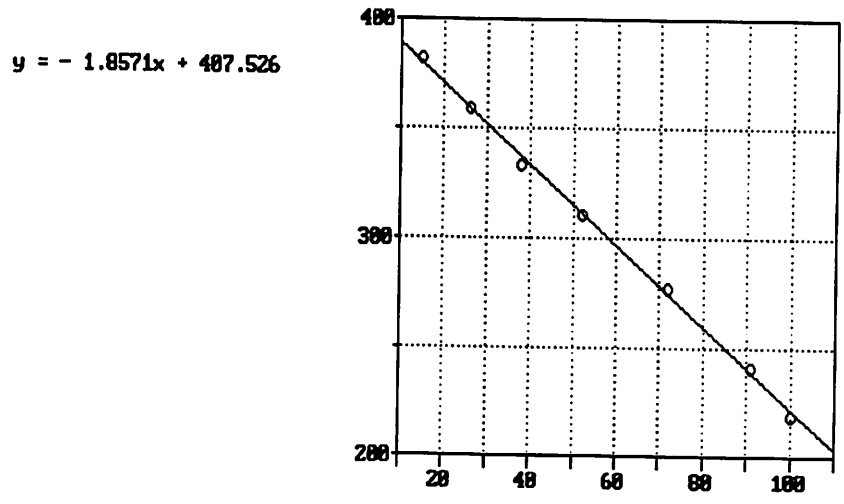
$y - \bar{y} = m(x - \bar{x})$

$y = -1,85706\ x + 1,857062 \cdot 56,28571 + 303 = 407,52608$

$y = -1,85706\ x + 407,52608$

Eq. de la droite de régression de y en x.

xi	15	26	38	52	72	91	100	moyenne des xi	56,285714
yi	382	359	333	310	277	241	219	moyenne des yi	303
xiyi	5730	9334	12654	16120	19944	21931	21900	somme des xiyi	107613
xi^2	225	676	1444	2704	5184	8281	10000	somme des xi^2	28514
								pente m	-1,857062
								valeur de h	407,52608
								covariance	-1681,286
								variance de X	905,34694



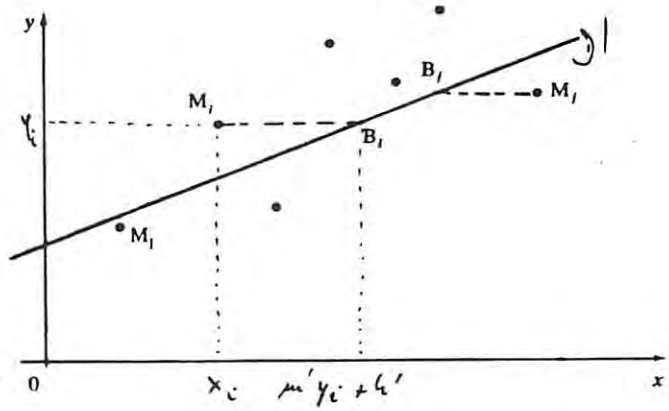
Ayant déterminé la droite de régression de y p.r. à x, il est possible d'utiliser l'équation obtenue pour estimer la valeur de y qui devrait correspondre à une valeur de x donnée.

• si $x = 45$ on peut estimer que
$y = -1{,}85706 \cdot 45 + 407{,}52608$
$y = 323{,}95838$; il s'agit d'une interpolation.

• si $x = 250$ on peut estimer que
$y = -1{,}85706 \cdot 250 + 407{,}52608$
$= -56{,}73892$; il s'agit d'une extrapolation.
L'extrapolation est à manipuler avec prudence, car il arrive que l'approximation $y = mx + h$ ne soit valable que dans un certain domaine.

c) Droite de régression de x en y

Nous cherchons une droite D' d'éq. :
$x = m'y + h'$ telle que
$S' = \sum_{i=1}^{N} d(M_i, B_i)^2$
$= \sum_{i=1}^{n} (x_i - m'y_i - h')^2$



soit minimal.
D'après des calculs analogues à ceux de 2 a) on trouve :

$$D' : \quad x - \bar{x} = \frac{\mathrm{cov}(x,y)}{V(y)} \cdot (y - \bar{y})$$

Ceci permet d'estimer x à partir de y observé.
Cette droite passe par le point de coordonnées $(\bar{x}, \bar{y})$ et a pour pente : $\dfrac{V(y)}{\mathrm{cov}(x,y)}$.

D et D' se coupent au point de coordonnées $(\bar{x}, \bar{y})$.
et leurs pentes ont même signe, celui de $\mathrm{cov}(x,y)$

d) Exemple (données de 26)

x_i	15	26	38	52	72	91	100	moyenne des x_i	56,285714
y_i	382	359	333	310	277	241	219	moyenne des y_i	303
x_iy_i	5730	9334	12654	16120	19944	21931	21900	somme des x_iy_i	107613
x_i^2	225	676	1444	2704	5184	8281	10000	somme des x_i^2	28514
y_i^2	145924	128881	110889	96100	76729	58081	47961	somme des y_i^2	664565

régression de y en x	valeur de m	-1,857062
	valeur de h	407,52608
régression de x en y	valeur de m'	-0,537348
	valeur de h'	219,10222
	covariance	-1681,286
	variance de X	905,34694
	variance de Y	3128,8571
coefficient de corrélation linéaire $\rho(x,y)$		-0,998944

Equation de D' : $x = -0,537348\,y + 219,10222$

x	10	100
y_D	388,96	221,82
$y_{D'}$	389,14	221,65

En réalité les droites sont pour ainsi dire superposées.

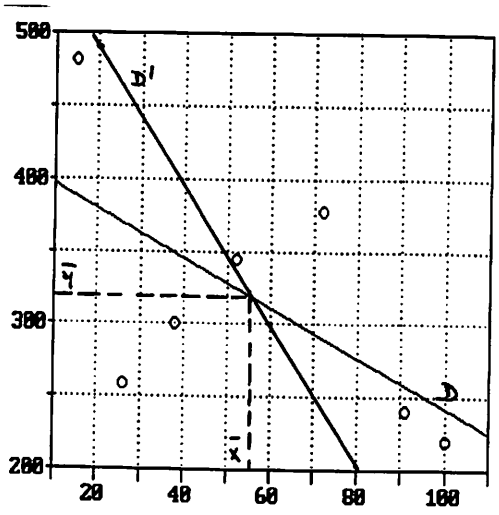
e) Exemple (données de 25 modifiées)

x_i	15	26	38	52	72	91	100	moyenne des x_i	56,285714
y_i	482	259	300	345	377	241	219	moyenne des y_i	317,57143
x_iy_i	7230	6734	11400	17940	27144	21931	21900	somme des x_iy_i	114279
x_i^2	225	676	1444	2704	5184	8281	10000	somme des x_i^2	28514
y_i^2	232324	67081	90000	119025	142129	58081	47961	somme des y_i^2	756601

D : $y = -1,711127x + 413,88341$

D' : $x = -0,214143y + 124,29143$

régression de y en x	valeur de m	-1,711127
	valeur de h	413,88341
régression de x en y	valeur de m'	-0,214143
	valeur de h'	124,29143
	covariance	-1549,163
	variance de X	905,34694
	variance de Y	7234,2449
coefficient de corrélation linéaire $\rho(x,y)$		-0,605331



f) Cas dans lequel les deux droites de régression coïncident

Comme D et D' passent toutes les deux par le point de coordonnées $(\bar{x}, \bar{y})$, on sait que D et D' coïncident ssi elles ont même pente ; c.à.d.

$$m = \frac{1}{m'} \iff \frac{cov(x,y)}{V(x)} = \frac{V(y)}{cov(x,y)}$$

$$\iff \frac{(cov(x,y))^2}{V(x) \cdot V(y)} = 1$$

$$\iff \rho(x,y)^2 = 1$$

$$\iff \rho(x,y) = 1 \text{ ou } \rho(x,y) = -1$$

g) Encadrement de $\rho(x,y)$

- Dans 2a) nous avons établi que la droite de régression de y en x qui minimisait $S = \sum_{i=1}^{N} (y_i - mx_i - h)^2$ avait pour équation

$$y - \bar{y} = m(x - \bar{x}) \quad \text{où} \quad m = \frac{cov(x,y)}{V(x)}$$

Ainsi :

$$S = \sum_{i=1}^{N} (y_i - m x_i - h)^2$$

$$= \sum_{i=1}^{N} [y_i - (m x_i + h)]^2$$

$$= \sum_{i=1}^{N} [y_i - m(x_i - \bar{x}) - \bar{y}]^2$$

$$= \sum_{i=1}^{N} [(y_i - \bar{y}) - m(x_i - \bar{x})]^2$$

$$= m^2 \sum_{i=1}^{N} (x_i - \bar{x})^2 - 2m \sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y}) + \sum_{i=1}^{N} (y_i - \bar{y})^2$$

$$= N [V(x) \cdot m^2 - 2 \operatorname{cov}(x,y)\, m + V(y)]$$

Comme $S \geq 0$, l'équation du second degré en m a un discriminant $\Delta \leq 0$

$$\Delta \leq 0 \iff 4[\operatorname{cov}(x,y)]^2 - 4 V(x) V(y) \leq 0$$

$$[\operatorname{cov}(x,y)]^2 \leq \underbrace{V(x) V(y)}_{\text{positif}}$$

$$\frac{[\operatorname{cov}(x,y)]^2}{V(x) \cdot V(y)} \leq 1$$

$$\left[\frac{\operatorname{cov}(x,y)}{\sigma_x \sigma_y}\right]^2 \leq 1$$

$$-1 \leq \frac{\operatorname{cov}(x,y)}{\sigma_x \sigma_y} \leq 1$$

$$-1 \leq \rho(x,y) \leq 1$$

h) Angle des droites de régression

D'après ce qui précède on a :

$$S = N\left[V(x)\,\mu^2 - 2\,\mathrm{cov}(x,y)\cdot\mu + V(y) \right]$$

$$= N\left[V(x)\cdot\frac{[\mathrm{cov}(x,y)]^2}{V(x)^2} - 2\,\mathrm{cov}(x,y)\cdot\frac{\mathrm{cov}(x,y)}{V(x)} + V(y) \right]$$

$$= N\left[\frac{[\mathrm{cov}(x,y)]^2}{V(x)} - 2\frac{[\mathrm{cov}(x,y)]^2}{V(x)} + V(y) \right]$$

$$= N\,V(y)\left[1 - \frac{[\mathrm{cov}(x,y)]^2}{V(x)V(y)} \right]$$

$$= N\,V(y)\left[1 - [\rho(x,y)]^2 \right]$$

et

$$S' = N\,V(x)\left[1 - [\rho(x,y)]^2 \right]$$

Ainsi :

si $|\rho| = 1$: tous les points sont alignés sur $D = D'$

la corrélation entre x et y est fonctionnelle et affine

si $|\rho|$ est voisin de 1 : S est alors très faible ; les points sont très proches de la droite D et de la droite D' qui elles font un angle très faible (cf 2 d) ; on prend en général $\frac{\sqrt{3}}{2} \leq |\rho| < 1$

la corrélation linéaire est très forte

si $|\rho|$ est proche de 0 : S se rapproche de son maximum $N\,V(y)$; les droites de régression font un grand angle (cf 2e)

la corrélation linéaire est faible.

i) Commentaires

Pour l'exemple 2d) les x_i et y_i représentent les valeurs obtenues lors d'une expérience en physique. y_i est la tension aux bornes d'un générateur qui débite dans un circuit. On a la loi : $y_i = E - R x_i$ où y_i est exprimé en volts, la force électro-motrice E en volts, la résistance interne R du générateur en ohms, l'intensité x_i du courant en ampères.
Dans toute étude statistique d'un phénomène, c'est le spécialiste de la discipline concernée qui doit interpréter les résultats. D'une corrélation observée entre les variables on ne peut pas conclure à une liaison physique objective et encore bien moins à une dépendance causale.

Un exemple classique est l'affirmation que le poids du nouveau-né croît avec l'âge de la mère.

Dans le tableau qui suit figurent l'âge de la mère (X) et le poids de l'enfant (Y) pour un échantillon de 200 naissances, présentés avec un groupement à 2 dimensions en classes d'âge de 2 ans (repré. par leurs clubs) et en classes de poids de 200 g (repré. par leurs centres).

Pour effectuer les calculs on a posé :

$$\begin{cases} x'_i = \dfrac{x_i - 24{,}5}{2} = \dfrac{1}{2} x_i - 12{,}25 \\ y'_j = \dfrac{y_j - 3100}{200} = \dfrac{1}{200} y_j - 15{,}5 \end{cases}$$

Tous les calculs du tableau sont effectués sur les x'_i et y'_j.
Si nous voulions dresser une liste de tous les couples $(x_i ; y_i)$, afin de rester fidèles à la notation adoptée dans cette partie, nous devrions mentionner certains couples plusieurs fois. (20,5 ; 3100) est mentionné 9 fois p. ex.

Âge de la mère (x) et poids de l'enfant (y) pour 200 naissances

y \ x	x'	16,5	18,5	20,5	22,5	24,5	26,5	28,5	30,5	32,5	34,5	36,5	38,5	40,5	42,5	44,5	$n_{\cdot j}$	y'_j	$n_{\cdot j} y'_j$	$n_{\cdot j} y'^2_j$	$\sum_i n_{ij} x'_i$	$(\sum_i n_{ij} x'_i) y'_j$	total
	y'	— 4	— 3	— 2	— 1	0	1	2	3	4	5	6	7	8	9	10							
2 100	— 5						3										3	— 5	— 15	75	3	— 15	
2 300	— 4		2		1				1	1							5	— 4	— 20	80	0	0	
2 500	— 3		2	2	1	2		1				1					9	— 3	— 27	81	— 3	9	
2 700	— 2			2	3	2	1	1	1	2	1						13	— 2	— 26	52	12	— 24	
2 900	— 1		2	3	3	4	2	1	1								16	— 1	— 16	16	— 8	8	
3 100	0	2	1	9	2	1	3	2	4	3		2		1			30	0	0	0	20	0	
3 300	1		2	9	7	6	4	6	1	1	1	3			1		41	1	41	41	24	24	
3 500	2	1	5	2	4	3	6	4	3	3		2	2	1			36	2	72	144	42	84	
3 700	3			3	3	2	2	3		1		1	1		1		17	3	51	153	25	75	
3 900	4		1	1		2	1		1	2	4				1	1	14	4	56	224	46	184	
4 100	5			1		3	1			1		2	1				9	5	45	225	22	110	
4 300	6					2		1			1						4	6	24	144	7	42	
4 500	7					1					1						2	7	14	98	5	35	
4 700	8																0	8	0	0	0	0	
4 900	9																0	9	0	0	0	0	
5 100	10									1							1	10	10	100	4	40	
$n_{i\cdot}$		3	15	32	24	28	23	19	12	15	8	11	4	2	3	1							200
x'_i		— 4	— 3	— 2	— 1	0	1	2	3	4	5	6	7	8	9	10							
$n_{i\cdot} x'_i$		— 12	— 45	— 64	— 24	0	23	38	36	60	40	66	28	16	27	10							199
$n_{i\cdot} x'^2_i$		48	135	128	24	0	23	76	108	240	200	396	196	128	243	100							2 045
$\sum_j n_{ij} y'_j$		2	0	18	8	46	12	23	4	25	28	17	12	2	8	4							209
$(\sum_j n_{ij} y'_j)\, x'_i$		— 8	0	— 36	— 8	0	12	46	12	100	140	102	84	16	72	40							572
Total																	200		209	1 433	199	572	

$\bar{x}' = \frac{199}{200} = 0,995 \quad \Rightarrow \quad \bar{x} = 2\bar{x}' + 24,5 = 26,49$

$\bar{y}' = \frac{209}{200} = 1,045 \quad \Rightarrow \quad \bar{y} = 200\,\bar{y}' + 3100 = 3309$

$V(x') = \frac{2045}{200} - \bar{x}'^2 = 9,238975 \Rightarrow V(x) = 4 \cdot 9,238975 = 36,9399$

$\sigma_x = 6,07782$

$V(y') = \frac{1433}{200} - \bar{y}'^2 = 6,072975 \Rightarrow V(y) = 40000 \cdot 6,072975$

$= 242\,919$

$\sigma_y = 492,86813$

$\text{cov}(x', y') = \frac{1}{N} \sum_{i,j} n_{ij}\, x'_i \cdot y'_j - \bar{x}'\,\bar{y}'$

$= \frac{572}{200} - 0995 \cdot 1,045$

$= 1,820225$

57

$$cov(x,y) = \frac{1}{ac}\, cov(x',y')$$

$$= \frac{1}{\frac{1}{2}\cdot\frac{1}{200}}\, cov(x',y')$$

$$= 400 \cdot cov(x',y')$$

$$= 728,09$$

$$\rho(x,y) = \frac{cov(x,y)}{\sigma_x \cdot \sigma_y} = \frac{728,09}{6,07782 \cdot 92,86813}$$

$$\simeq 0,243056 \quad \text{(corrélation faible)}$$

Pour la droite de régression de y en x on a :

$$a = \frac{cov(x,y)}{V(x)} \simeq 19,71$$
$$b = \bar{y} - a\bar{x} \simeq 2786,88$$

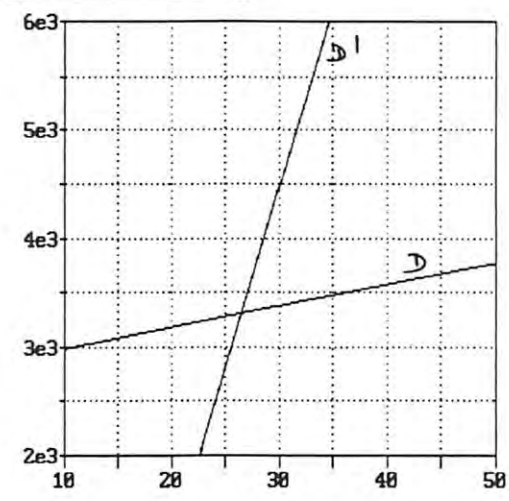
$\Big\}$ D : $y = 19,71x + 2786,88$

Pour la droite de régression de x en y on a :

$$a' = \frac{cov(x,y)}{V(y)} \simeq 2,9973 \cdot 10^{-3}$$
$$b' = \bar{x} - a'\bar{y} \simeq 16,57$$

$\Big\}$ D' : $x = 0,0029973y + 16,57$

$y = 333,64x - 5529,09$

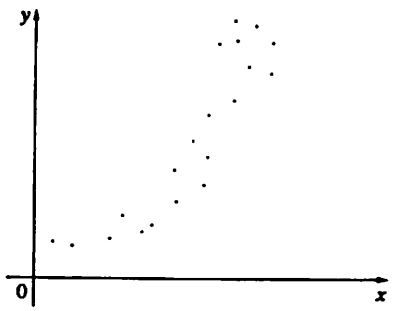


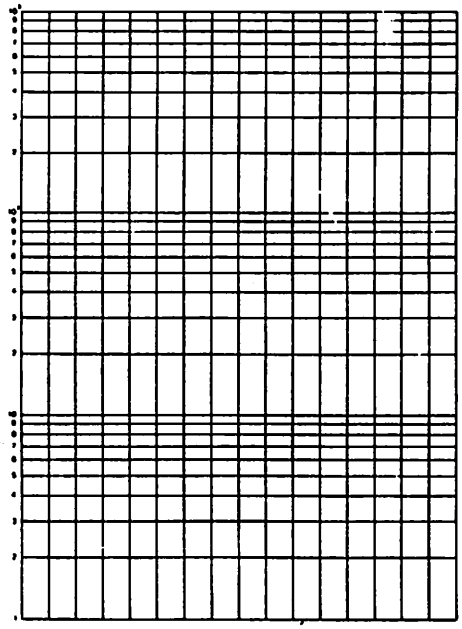
Il y a donc une faible corrélation entre le poids du nouveau-né et l'âge de la mère ; le poids augmentant avec l'âge.

Toutefois une analyse médicale plus fine montre que si l'on ne prend que des nouveaux-nés de même rang de naissance (1er enfant p. ex.), on ne trouve pas ce lien entre le poids et l'âge. La réalité est que le poids croît avec le rang de la naissance, lequel est évidemment en relation avec l'âge de la mère. Il en résulte la corrélation que nous venons d'étudier et qui n'est pas apparente.

3) Ajustement exponentiel et ajustement par une fonction puissance

a) Sur l'exemple suivant l'ajustement linéaire donne des résultats médiocres. Une fonction du type $y = \lambda \cdot a^x$ ($a > 0, \lambda > 0$) est plus apte à approcher la relation qui lie X et Y. Pour la représentation graphique on utilise du papier semi-logarithmique. Les abscisses sont les x_i et les ordonnées les $\log y_i$ (logarithme décimal).





$y_i = \lambda a^{x_i}$

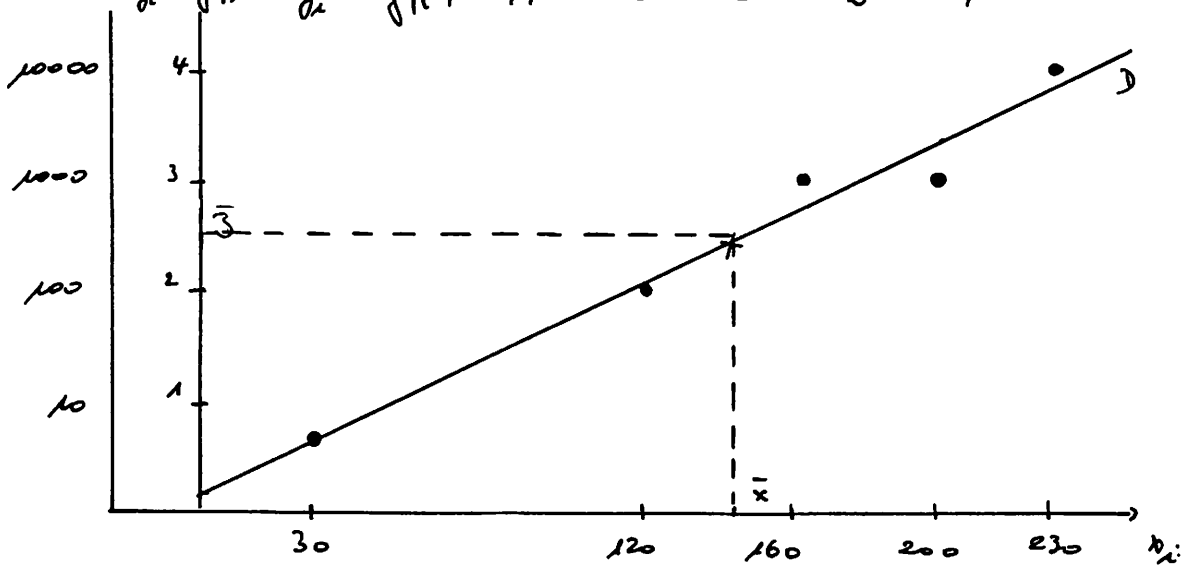
$\log y_i = \log \lambda + x_i \log a$

$\log y_i = (\log a) \cdot x_i + \log \lambda$

On dit qu'on a procédé à un ajustement exponentiel de la statistique.

Exemple :

x_i	30	120	160	200	230
y_i	5	100	1000	1000	10000
$z_i = \log y_i$	0,7	2	3	3	4



A l'aide de la méthode des moindres carrés de la partie 2) on trouve :

$\bar{x} = 148$; $\bar{z} = 2,54$

$V(x) = 4856$; $V(z) = 1,25$; $\text{cov}(X, Z) = 76,310488$

$\left.\begin{array}{l} m = 0,015771 \\ p = 0,214402 \end{array}\right\}$ éq. de la droite de régression de z en x : $z = 0,015771\,x + 0,214402$

c.à.d. : $\log y = 0,015771\,x + 0,214402$

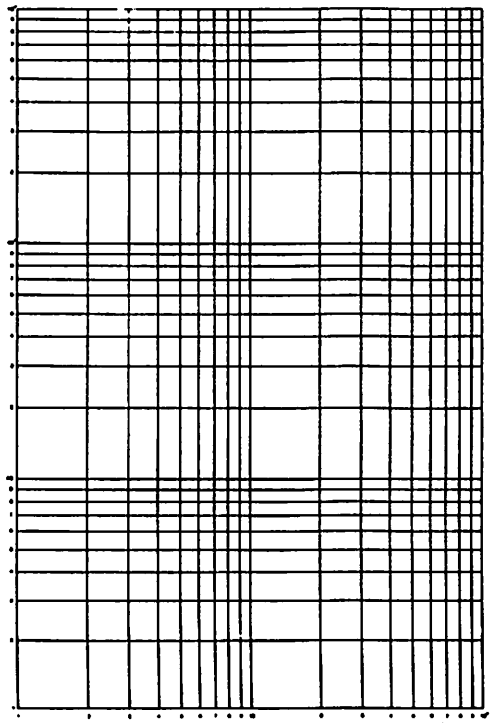
$y = 10^{0,214402} \cdot (10^{0,015771})^x$

$y = 1,63689 \cdot 1,03684^x$

La fonction exponentielle d'ajustement est donc

$y = 1,63689 \cdot 1,03684^x$

b) Nous venons de constater que sur papier semi-logarithmique le nuage de l'exemple précédent a une allure à peu près rectiligne.
Dans d'autres cas nous obtenons cette allure à peu près rectiligne du nuage en utilisant du papier logarithmique où les abscisses sont les $\log x_i$ et les ordonnées les $\log y_i$.



Alors on obtient une relation du type

$\log y = a \log x + b$

$y = x^a \cdot 10^b$

$y = 10^b \cdot x^a$

$y = \lambda \cdot x^a$

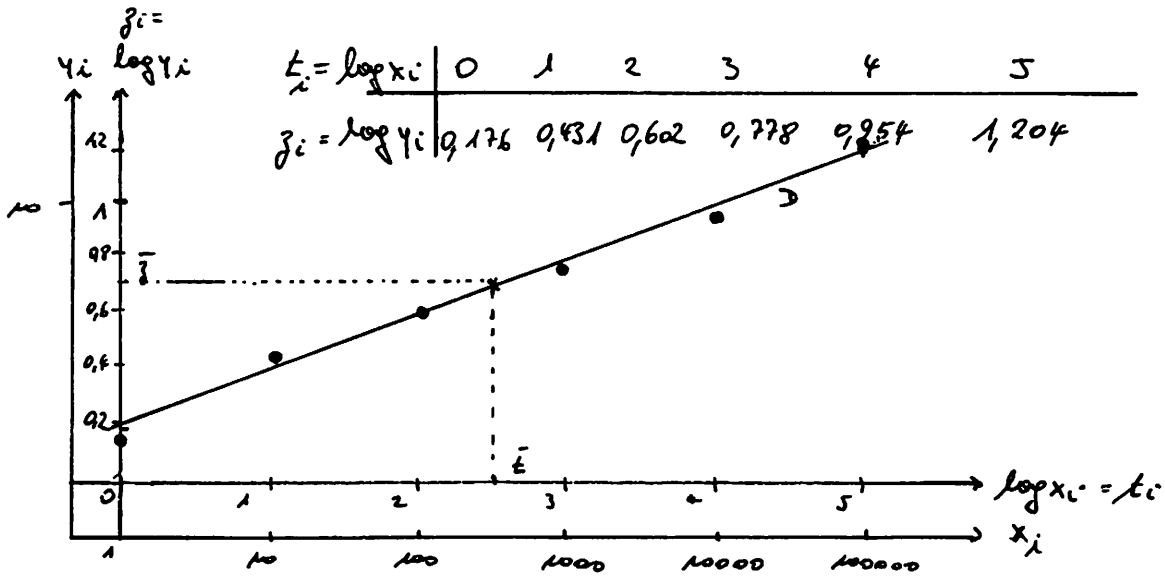
$(\lambda > 0$ et $a > 0)$

Il s'agit d'un ajustement par une fonction puissance.

Exemple :

x_i	1	10	100	1000	10000	100000
y_i	1,5	2,7	4	6	9	16

$t_i = \log x_i$	0	1	2	3	4	5
$z_i = \log y_i$	0,176	0,431	0,602	0,778	0,954	1,204



A l'aide de la méthode des moindres carrés de la partie 2) on trouve :

$\bar{t} = 2{,}5$ $\quad \bar{z} = 0{,}691$

$V(t) = 2{,}91667$ $\quad V(z) = 0{,}11345$ $\quad \text{cov}(t,z) = 0{,}57375$

$m = 0{,}19671$
$h = 0{,}19621$
} ép. de la droite de régression de z en t : $z = 0{,}19671\, t + 0{,}19621$

D'où : $\log y = 0{,}19671 \log x + 0{,}19621$

$y = 10^{0{,}19621} \cdot x^{0{,}19671}$

D : $\underline{y = 1{,}57293 \cdot x^{0{,}19671}}$

4) Ajustements effectués par ordinateur

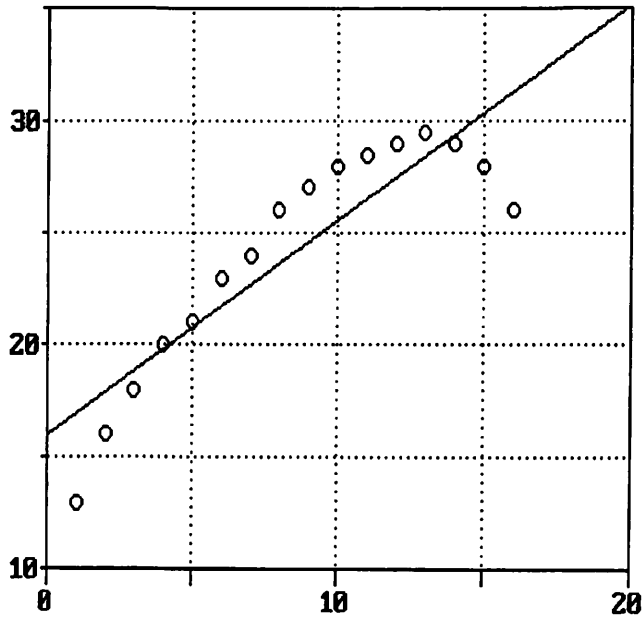
a) Les cinq ajustements suivants ont été faits à partir des données suivantes :

x_i	y_i	x_i	y_i	x_i	y_i
1/	13	16/	26	31/	18
2/	16	17/	25	32/	20
3/	18	18/	22	33/	22
4/	20	19/	20	34/	24
5/	21	20/	18	35/	25
6/	23	21/	17	36/	27
7/	24	22/	15	37/	28
8/	26	23/	14	38/	30
9/	27	24/	13	39/	31
10/	28	25/	12.5	40/	32
11/	28.5	26/	12	41/	30
12/	29	27/	12.5	42/	28
13/	29.5	28/	14	43/	26
14/	29	29/	15	44/	25
15/	28	30/	16	45/	22

Lineare Regression :

y =.9618x + 15.95

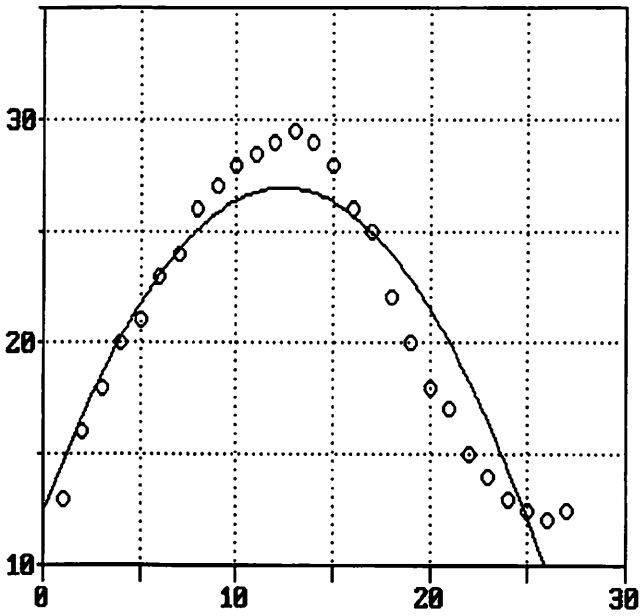
mit den 16 ersten Werten der Tabelle



Polynom-Regression :

y = - .0944x^2 + 2.3459x
+ 12.3812

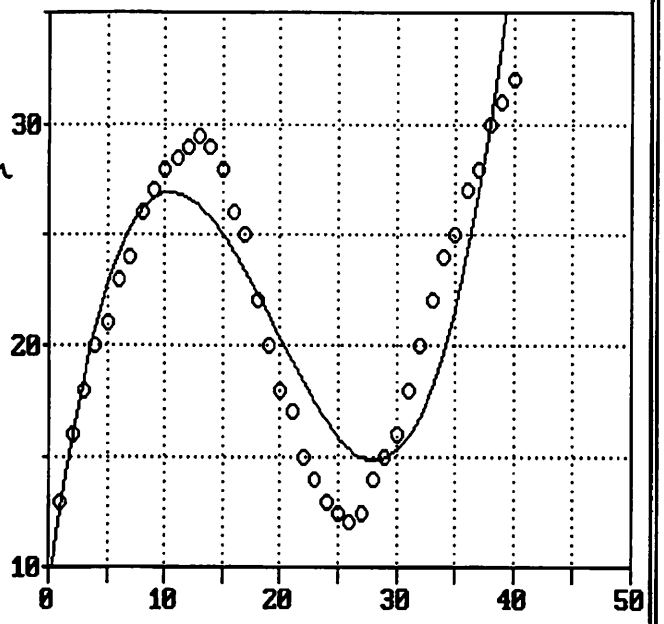
mit den 27 ersten Werten der Tabelle



Polynom-Regression :

y =.0044x^3 - .2527x^2
+ 3.8677x + 9.1513

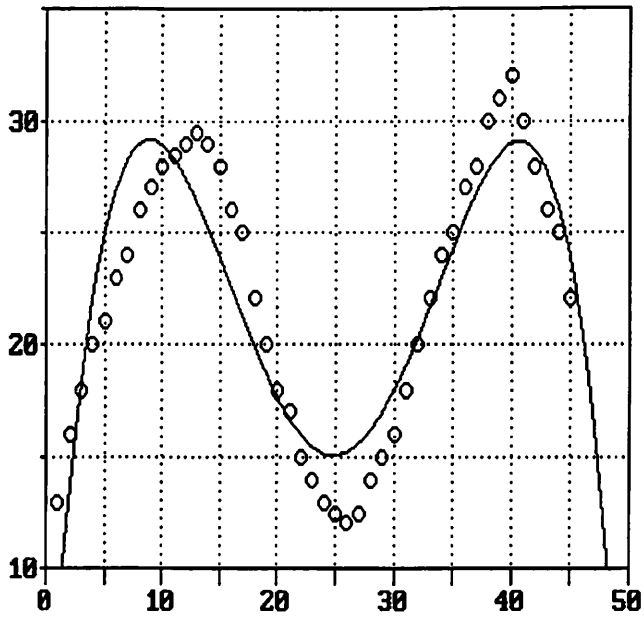
mit den 40 ersten Werten der Tabelle



Polynom-Regression :

$y = -.0002x^4 + .0218x^3 - .7003x^2 + 7.8937x + .3851$

mit allen Werten der Tabelle

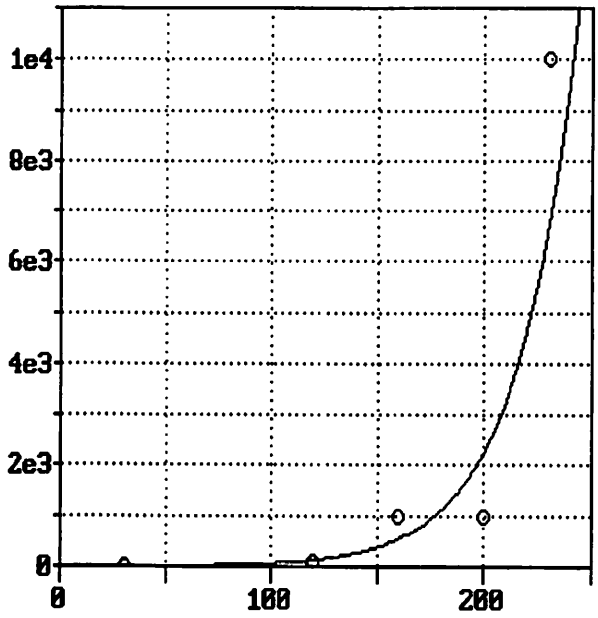


b)

Exponentielle Regression :

$y = 1.6376 \cdot 1.0368\ ^x$

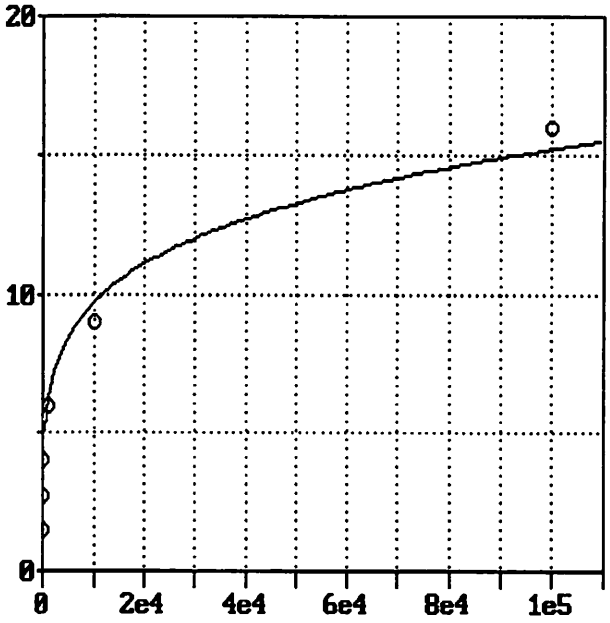
Beispiel 3 a)



Geometrische Regression :

$y = 1.5821 \cdot x^{.1967}$

Beispiel 3 b)



X

Les nombres indices

1) Les indices simples

époque	prix par kg de la matière X
t_0	3000 fr
t_1	3240 fr

Augmentation du prix : 8% ; c.à.d. à un prix de 100 fr à l'époque t_0, correspond un prix de 108 fr à l'époque t_1.

On appelle indice simple d'un phénomène mesurable, le rapport des valeurs P_i et P_0 prises par le phénomène aux époques t_i et t_0 (année de base).

$$\frac{I}{i/0} = \frac{P_i}{P_0}$$

2) Les indices synthétiques

Les indices simples ne prennent en compte que la variation d'une seule composante d'un phénomène (le prix par kg de la matière X p.ex.)

Les indices synthétiques combinent l'évolution simultanée de plusieurs phénomènes, chacun d'eux étant mesuré par un indice simple.

a) Indice des prix de Laspeyres

Exemple :

mois	article	prix	quantité
1	A	115	4
	B	200	6
	C	325	11
2	A	118	5
	B	204	5
	C	331	18

Pour calculer l'indice des prix de Laspeyres on fait le rapport de la somme des nouveaux prix pondérés par les anciennes quantités par la somme des anciens prix pondérés par les anciennes quantités.

$$I L_{2/1} = \frac{118 \cdot 4 + 204 \cdot 6 + 331 \cdot 11}{115 \cdot 4 + 200 \cdot 6 + 325 \cdot 11} = \frac{472 + 1224 + 3641}{460 + 1200 + 3575}$$

$$= \frac{5337}{5235} = 1{,}019\,484\,2\ldots$$

En prenant pour base 100 le 1. mois, on arrive le 2. mois à 101,948 42... . Plus généralement :

mois	article	prix	quantité	valeur
0	$A_1 \cdots A_k \cdots A_m$	$p_{10}, \ldots, p_{k0}, \ldots p_{m0}$	$q_{10}, \ldots, q_{k0}, \ldots q_{m0}$	$v_{10} \cdots v_{k0} \cdots v_{m0}$
⋮	⋮	⋮	⋮	⋮
i	$A_1 \cdots A_k \cdots A_m$	$p_{1i}, \ldots, p_{ki}, \ldots p_{mi}$	$q_{1i}, \ldots, q_{ki}, \ldots q_{mi}$	$v_{1i} \cdots, v_{ki}, \ldots v_{mi}$ $= p_{ki} \cdot q_{ki}$
⋮	⋮	⋮	⋮	

$$I_{L_{i/0}} = \frac{\sum_{k=1}^{M} p_{ki}\, q_{k0}}{\sum_{k=1}^{M} p_{k0}\, q_{k0}}$$

$$= \frac{\sum_{k=1}^{M} \left[\frac{p_{ki}}{p_{k0}} \cdot (p_{k0}\, q_{k0}) \right]}{\sum_{k=1}^{M} p_{k0}\, q_{k0}}$$

$$= \frac{\sum_{k=1}^{M} I_{k\,i/0}\, v_{k0}}{\sum_{k=1}^{M} v_{k0}}$$

$I_{L_{i/0}}$ est la moyenne arithmétique des indices simples des prix, pondérés par les valeurs du mois initial.

b) Indice des prix de Paasche

Pour calculer l'indice des prix de Paasche on fait le rapport de la somme des nouveaux prix pondérés par les nouvelles quantités par la somme des anciens prix pondérés par les nouvelles quantités.

$$I_{P_{2/1}} = \frac{118 \cdot 5 + 204 \cdot 5 + 331 \cdot 18}{115 \cdot 5 + 200 \cdot 5 + 325 \cdot 18} = \frac{590 + 1020 + 5958}{575 + 1000 + 5850} = \frac{7568}{7425}$$

$$= 1,01925925\ldots$$

En prenant pour base 100 le 1. mois on arrive à 101,92592... le 2. mois.

En général :

$$IP_{i/0} = \frac{\sum_{k=1}^{M} P_{ki}\, Q_{ki}}{\sum_{k=1}^{M} P_{k0}\, Q_{ki}}$$

$$= \frac{\sum_{k=1}^{M} V_{ki}}{\sum_{k=1}^{M} \frac{P_{k0}}{P_{ki}} \cdot P_{ki} \cdot Q_{ki}}$$

$$= \frac{\sum_{k=1}^{M} V_{ki}}{\sum_{k=1}^{M} \frac{1}{I_{k\,i/0}} \cdot V_{ki}}$$

$$\frac{1}{IP_{i/0}} = \frac{\sum_{k=1}^{M} \frac{1}{I_{k\,i/0}} V_{ki}}{\sum_{k=1}^{M} V_{ki}}$$

$IP_{i/0}$ est la moyenne harmonique des indices simples des prix, pondérés par les valeurs du i^{e} mois.

XI

Séries chronologiques

1) Une série chronologique est une suite d'observations chiffrées dans le temps.
La succession des résultats observés résulte de trois composantes :
- la tendance ou trend ; c'est le reflet du mouvement de longue période,
- les variations saisonnières sont des fluctuations liées à la période étudiée,
- les variations accidentelles de caractère imprévisible ; elles modifient ponctuellement la série chronologique

2) L'analyse de la tendance peut se faire à l'aide de la méthode d'ajustement des moindres carrés (3e exemple), mais aussi à l'aide de la méthode de la moyenne mobile décrite ci-dessous.

3) La correction des variations saisonnières se fait de deux manières différentes, suivant que les variations saisonnières
- semblent constantes dans le temps (1)
- semblent proportionnelles à la tendance générale. (2)

a) Dans le premier cas l'ajustement peut se faire par la méthode des moyennes :

* Dans le premier exemple nous considérons la moyenne des chiffres d'affaires par 2 mois et la moyenne des mois de janv ; de fév ; ... de déc et nous en déduisons la déviation saisonnière p. r. à la moyenne.

** Dans la suite du premier exemple nous groupons les mois trois par trois pour constituer une

nouvelle série chronologique à huit termes dont les y'_i sont la moyenne des trois termes y_{i-1}, y_i, y_{i+1}

x'_i	x_i	y_i	y'_i
2	1	65	76,6
	2	73	
	3	92	

Cette méthode est celle des <u>moyennes échelonnées</u>.

*** Dans le deuxième exemple nous utilisons la méthode des <u>moyennes mobiles</u>.
Pour "lisser" les irrégularités d'une série chronologique on peut remplacer y_i par la moyenne
* de y_{i-1}, (y_i), y_{i+1} (moy : mob. trimestrielle)
* de $\frac{1}{2} y_{i-3}$, y_{i-2}, y_{i-1}, (y_i), y_{i+1}, y_{i+2}, $\frac{1}{2} y_{i+3}$
(moy. mob. semestrielle)
* de ... pr ex. moy. mob. annuelle.

Ce procédé permet de garder un plus grand nombre de points moyens que la méthode précédente et d'éliminer les variations saisonnières. L'exemple 2 où le graphique mensuel et les graphiques des moyennes mobiles trim, sem et annuelles sont sur une même figure permet d'apprécier la valeur de cette méthode.

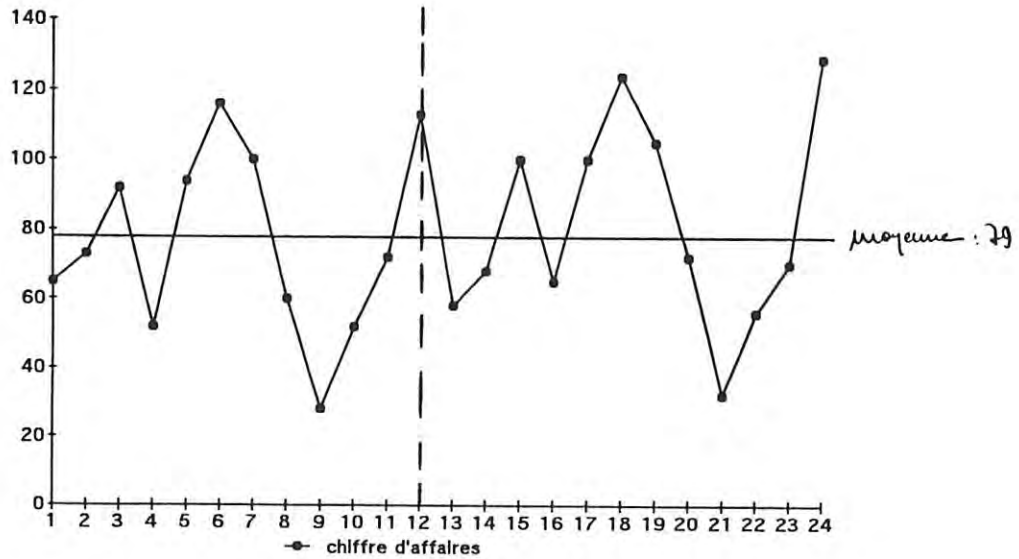
b) Dans le <u>deuxième cas</u> l'ajustement se fait à l'aide de la méthode des moindres carrés.

CORRECTION DES VARIATIONS SAISONNIERES

PREMIER EXEMPLE

Vente mensuelle d'un magasin d'alimentation en dizaines de milliers francs:

mois x_i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
chiffre d'affaires y_i	65	73	92	52	94	116	100	60	28	52	72	113	58	68	100	65	100	124	105	72	32	56	70	129



L'allure de la courbe permet d'affirmer que le mouvement saisonnier est périodique.

Nous calculons la moyenne des y_i sur 24 mois et à tout x_i nous associons la moyenne des deux y_i de janvier, de février, ... et nous déterminons l'écart saisonnier p.r. à cette moyenne.

CORRECTION DES VARIATIONS SAISONNIERES

PREMIER EXEMPLE

Vente mensuelle d'un magasin d'alimentation en dizaines de milliers francs:

mois x_i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
chiffre d'affaires y_i	65	73	92	52	94	116	100	60	28	52	72	113	58	68	100	65	100	124	105	72	32	56	70	129
moyenne de deux mois identiques	61,5	70,5	96	58,5	97	120	102,5	66	30	54	71	121	29	34	50	32,5	50	62	52,5	36	16	28	35	64,5
moyenne générale	79	79	79	79	79	79	79	79	79	79	79	79	79	79	79	79	79	79	79	79	79	79	79	79
déviation saisonnière à la moyenne	-17,5	-8,5	17	-20,5	18	41	23,5	-13	-49	-25	-8	42	-50	-45	-29	-46,5	-29	-17	-26,5	-43	-63	-51	-44	-14,5

CORRECTION DES VARIATIONS SAISONNIERES

DEUXIEME EXEMPLE moyennes mobiles

Vente mensuelle d'un magasin d'alimentation en dizaines de milliers francs:

mois xi	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
chiffre d'affaires yi	65	73	92	52	94	116	100	60	28	52	72	113	58	68	100	65	100	124	105	72	32	56	70	129
moyenne trimestrielle (*) mobile		76,67	72,33	79,33	87,33	103,3	92	62,67	46,67	50,67	79	81	79,67	75,33	77,67	88,33	96,33	109,7	100,3	69,67	53,33	52,67	85	
moyenne semestrielle (**) mobile				84,92	86,75	80,33	75	73,17	71,08	67,33	64,5	71,17	78,25	81,67	84,92	89,75	94	88,67	82,25	79	76,92			
moyenne annuelle (***) mobile							76,13	75,63	75,75	76,63	77,42	78	78,54	79,25	79,92	80,25	80,33	80,92						

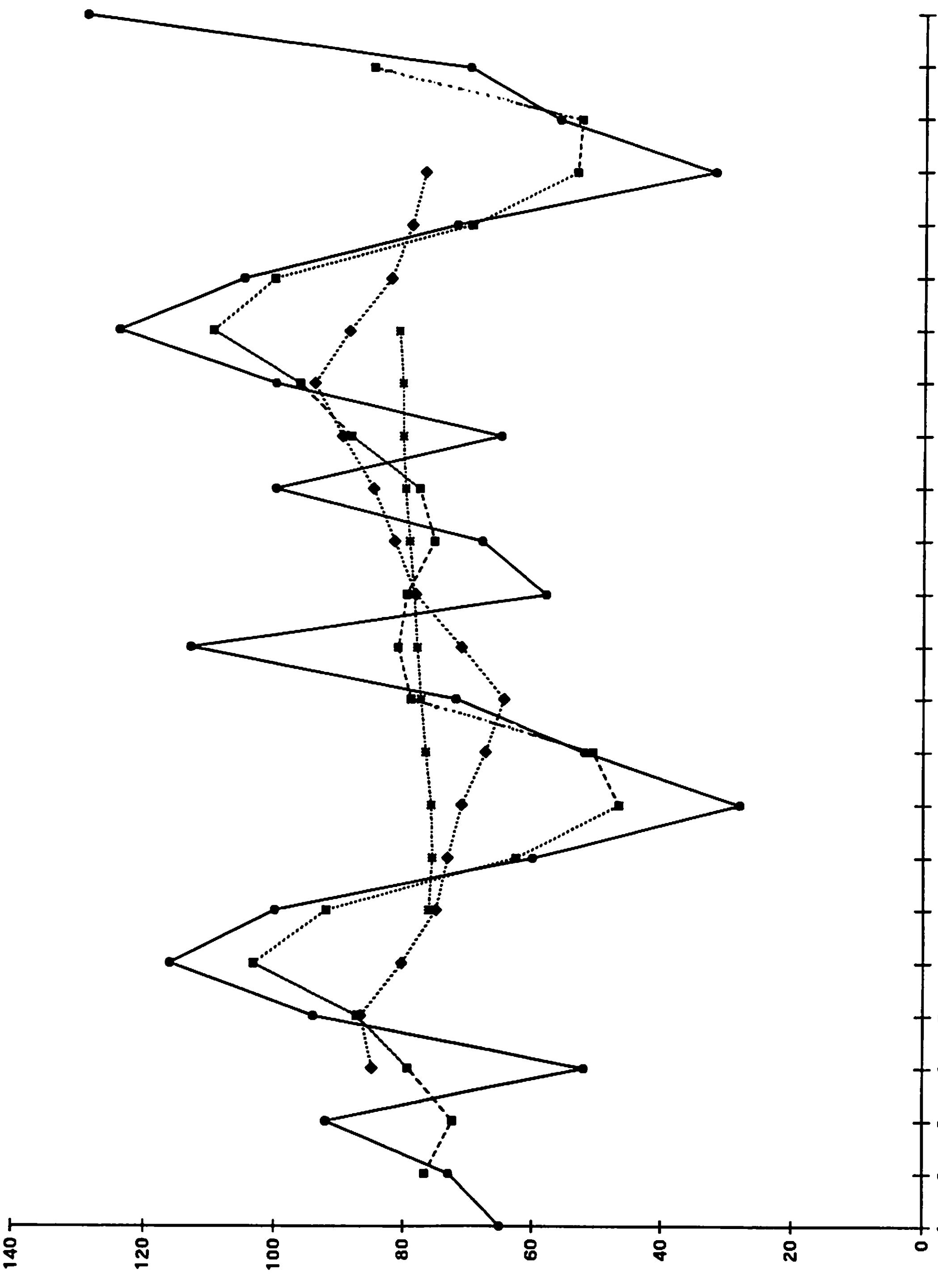
Représentation graphique de la page suivante : valeurs mens. y_i ●——●

(*) $\dfrac{y_{i-1} + (y_i) + y_{i-1}}{3}$ = moy. trim. mobile associée à y_i ■- - -■

(**) $\dfrac{(y_{i-3}/2) + y_{i-2} + y_{i-1} + (y_i) + y_{i+1} + y_{i+2} + (y_{i+3}/2)}{6}$ = moy. sem. mobile associée à x_i ◆.....◆

(***) $\dfrac{(y_{i-6}/2) + \sum_{k=i-5}^{i-1} y_k + (y_i) + \sum_{k=i+1}^{i+5} y_k + (y_{i+6}/2)}{12}$ = moy. ann. mobile associée à x_i *——*

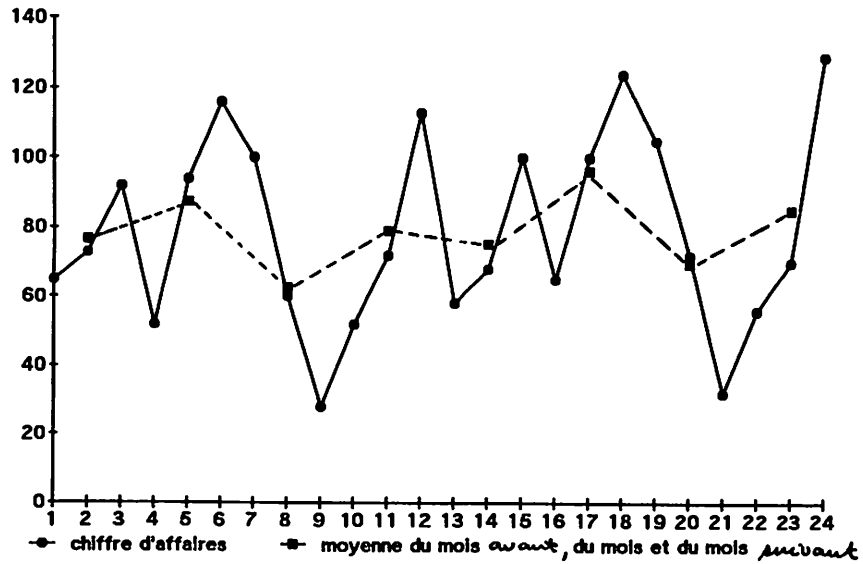
no 2 3 a)



0
20
40
60
80
100
120
140
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24

1. exemple, avec moyennes échelonnées

mois xi	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
chiffre d'affaires yi	65	73	92	52	94	116	100	60	28	52	72	113	58	68	100	65	100	124	105	72	32	56	70	129
moyenne du mois avant, du mois et du mois suiv.		76,67			87,33			62,67			79			75,33			96,33			69,67			85	



CORRECTION DES VARIATIONS SAISONNIERES

TROISIEME EXEMPLE

Vente mensuelle d'un magasin d'alimentation en dizaines de milliers francs:

mois xi	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
chiffre d'affaires yi	180,0	195,0	188,0	222,0	211,0	180,0	222,0	250,0	210,0	185,0	217,0	260,0	230,0	250,0	240,0	270,0	265,0	233,0	275,0	305,0	265,0	240,0	270,0	314,0

AJUSTEMENT:yi=4,3657xi+181,971

chiffre d'affaires ajusté	186,3	190,7	195,1	199,4	203,8	208,2	212,5	216,9	221,3	225,6	230,0	234,4	238,7	243,1	247,5	251,8	256,2	260,6	264,9	269,3	273,7	278,0	282,4	286,7
coefficient saisonnier(*)	0,966	1,023	0,964	1,113	1,035	0,865	1,045	1,153	0,949	0,82	0,944	1,109	0,963	1,028	0,97	1,072	1,034	0,894	1,038	1,133	0,968	0,863	0,956	1,095

(*) =chiffre d'affaires/chiffre d'affaires ajusté