Cours de Statistique

Olivier Maggioni

Avertissement

Ce document est conçu comme support de cours. Il ne possède ni la complétude ni l'exhaustivité d'un livre, voire d'un polycopié, qu'il ne saurait remplacer.

Chapitres

I Statistique Descriptive et Corrélative

II Probabilités

III Echantillonnage et estimations des paramètres

IV Tests Statistiques

V Séries Temporelles

Bibliographie

Statistique, cours et problèmes

Murray R. Spiegel, Série Schaum, McGraw-Hill, Paris 1993

Probabilités et statistiques pour Biologistes

Françoise Couty, Jean Debord, Daniel Fredon, Armand Colin, Paris 1990

Introduction

La Statistique : De quoi parle-t-on ?

La statistique peut être vue comme l'ensemble des méthodes et techniques permettant de traiter les données (informations chiffrées) associées à une situation ou un phénomène.

Cette démarche correspond à plusieurs objectifs, c'est pourquoi on subdivise la statistique en plusieurs domaines :

• Description d'une situation donnée (faire parler les chiffres).
C'est le cadre de la Statistique Descriptive.

• Mettre en évidence certaines relations.
On parle ici de statistique corrélative.

• Faire des prévisions à propos de phénomènes évoluant dans le temps.
Ce que l'on appelle les séries temporelles, ou chronologiques.

• D'induire des conclusions générales à partir de mesures faites sur un échantillon.
• De tester une hypothèse.
C'est l'objet de la statistique inférentielle.
Nous l'aborderons lors de la théorie des sondages (ou de l'échantillonnage).

En conséquence la statistique se révèle être un outil fondamental d'aide à la décision.

Objectifs du cours

• Acquérir une culture de base en statistique.

• Posséder le sens critique nécessaire à la compréhension de présentations ou travaux basés sur des études statistiques.

• Maîtriser les outils et techniques de base.

• Savoir choisir les outils adéquats pour le traitement des données, ceci en relation avec une problématique définie.

• Pouvoir utiliser de façon adéquate les logiciels statistiques.

I Statistique Descriptive et Corrélative

1.- Population, Echantillon, Variable Statistique, Effectifs, Fréquences, Variables Discrètes et Continues, Densité de fréquence, Histogramme, Fonction de répartition.

2.- Indicateurs de position : Moyenne, Mode, Médiane, Quantiles.

3.- Indicateurs de dispersion : Variance, Ecart-type, Intervalle Semi-interquartile.

4.- Autres indicateurs : Coefficients de Variations, Coefficient de Dissymétrie

5.- Corrélation et Régression linéaire : Distributions Conjointes, Marginales, Conditionnelles. Covariance, Coefficient de Corrélation, Droite de Régression. Variance expliquée et Résiduelle.

1.1.- Population, Échantillon, Variable Statistique

Définitions

• ***Population*** : ensemble d'unités statistiques.

Exemples :
- Tous les malades atteints de sclérose en plaque (où ? quand ?).
- Relevés pluviométriques quotidiens (population = jours).

• ***Echantillon***: sous-ensemble de la population.

En général nous n'avons pas accès à toute la population (recensement), d'où l'idée d'en extraire un sous-ensemble. Si on a une connaissance a priori, on peut parler d'échantillon représentatif (stratification).

• ***Variable statistique*** (ou caractère) : opération qui associe à chaque unité statistique une propriété, une modalité, un score.

• ***Observation*** : valeur prise par la variable sur une unité statistique.

• ***Données*** : sont constituées par l'ensemble des observations (tableaux, fichiers, données primaires).

Au sens mathématique du terme, une variable est une application de la population sur l'ensemble des scores. $X : P \to S$

Le fait que l'on note X une application peut être source de confusion. Cette notation devient cohérente dès que l'on parle de la distribution de la variable.

• On distingue les variables nominales (ou caractères qualitatifs) des variables numériques (ou caractères quantitatifs). Si on peut ordonner les modalités on parle aussi de variable ordinale. Les variables numériques se prêtent aux calculs (moyennes etc...), dans ce cas S est un ensemble numérique p.ex. S = IR.

Exemples

1.- Etat clinique : guéri, stationnaire, aggravé.

2.- Groupe sanguin.

3.- Relevés pluviométriques quotidiens (NE ;1999).

4.- Statistique médicale (OFS).
Codes diagnostics et d'interventions par patients, durée de séjour, régime d'assurance.

5.- Statistique administrative des établissements de santé (hôpitaux, cliniques, homes) (OFS).
Nombre de cas et nombre de journées par service, nombre de médecins d'infirmières etc…

Remarques

• Malgré la terminologie une population n'est pas nécessairement humaine.

• Attention aux fausses variables numériques (No de tél. AVS etc…).

• En général un relevé statistique fournit plusieurs variables que l'on peut voir comme un vecteur.

$$P \to IR^2$$

Par exemple à 2 variables : $$i \mapsto \begin{pmatrix} x_i \\ y_i \end{pmatrix}$$

• Une variable est dite discrète si elle peut prendre un nombre fini ou dénombrable (i.e. que l'on peut numéroter) de valeurs.

Dans ce qui suit nous nous intéresserons exclusivement aux variables numériques.

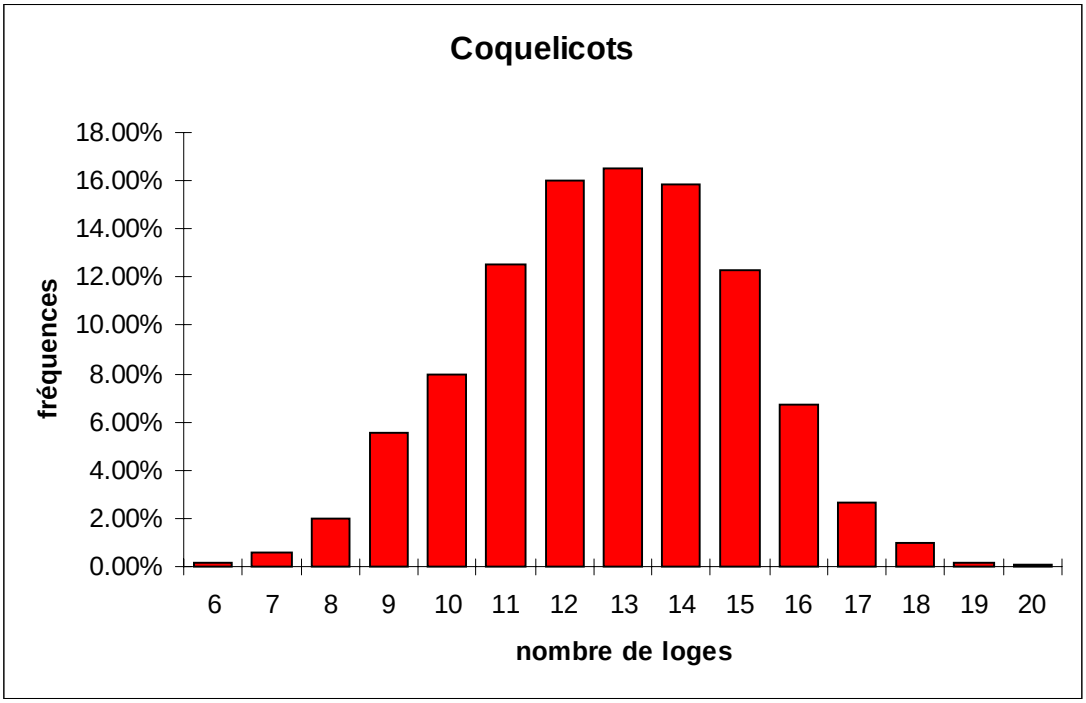
1.2 Effectifs et fréquences

Pour décrire la variable elle-même, il faut faire abstraction des unités statistiques, on regardera seulement combien d'unités ont obtenu chaque score. Ceci définit la distribution de la variable.

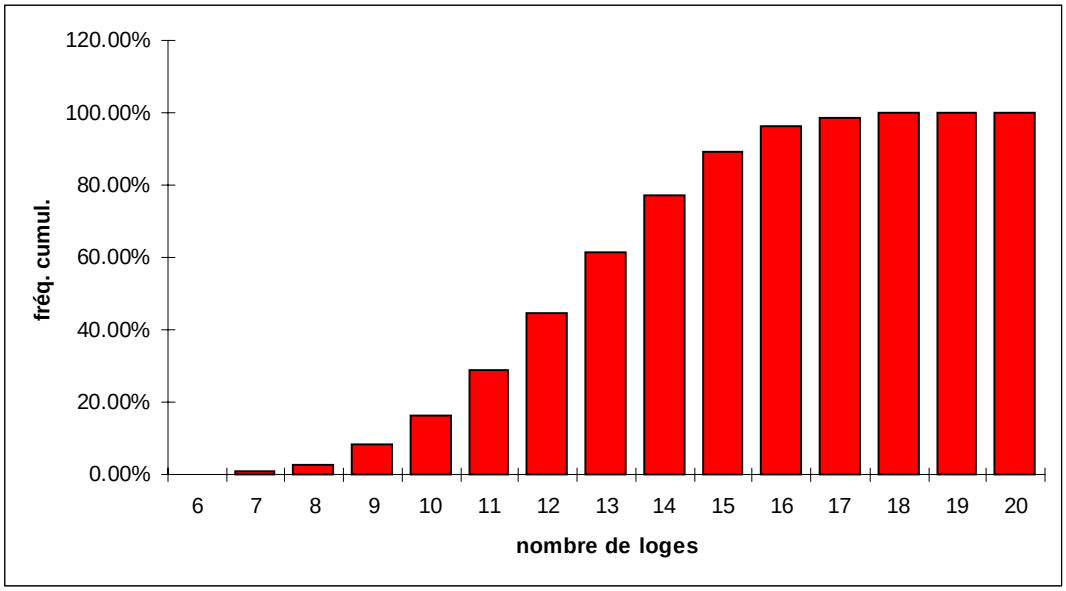
Exemple: nombre de loges capsulaires du coquelicot, (Biometrika, vol. 2. 1902)

Population 1905 coquelicots.

Nombre de loges	Nombre de coquelicots		
Scores x_k	Effectifs n_k	Fréquences f_k	fréquences cumulées
6	3	0.16%	0.16%
7	11	0.58%	0.73%
8	38	1.99%	2.73%
9	106	5.56%	8.29%
10	152	7.98%	16.27%
11	238	12.49%	28.77%
12	305	16.01%	44.78%
13	315	16.54%	61.31%
14	302	15.85%	77.17%
15	234	12.28%	89.45%
16	128	6.72%	96.17%
17	50	2.62%	98.79%
18	19	1.00%	99.79%
19	3	0.16%	99.95%
20	1	0.05%	100.00%
Total	**1905**	**100.00%**	



Représentations graphiques par des diagrammes en bâtons



Définitions

• L'***effectif*** d'un score est le nombre d'unités statistiques réalisant ce score.

• L'***effectif cumulé*** est donné par le nombre d'unités statistiques ayant un score inférieur ou égal.

$$n_k \uparrow = \sum_{j=1}^{k} n_j$$

• La ***fréquence*** d'un score est son effectif divisé par la taille de la population (ou effectif total)

$$f_k = \frac{n_k}{n}$$

• La ***fréquence cumulée*** est obtenue par la somme des fréquences des scores inférieurs ou égaux au score considéré.

$$f_k \uparrow = \sum_{j=1}^{k} f_j$$

Remarques :

• Un effectif en soi n'amène aucune information, il ne dit pas si le score a été réalisé souvent ou non. C'est pourquoi nous portons en général notre attention sur les fréquences.

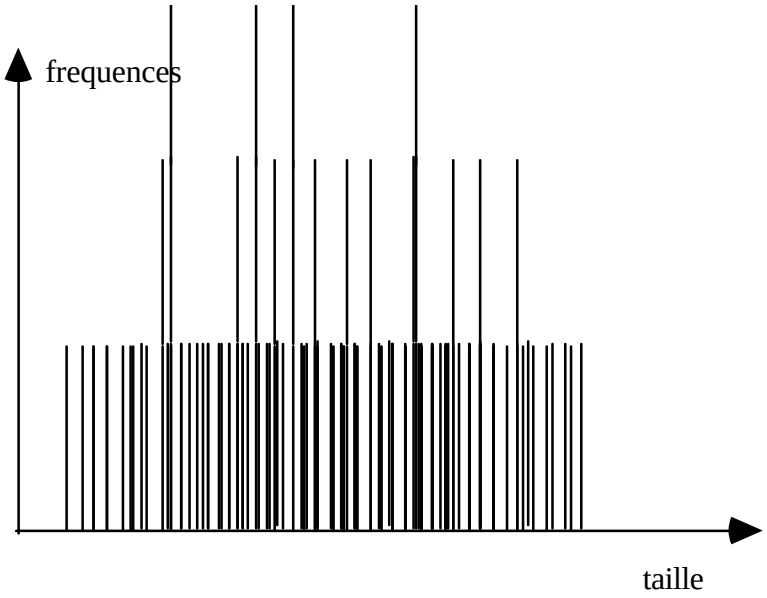
• Les fréquences (cumulées) quant à elles fournissent beaucoup d'information sur la série statistique. Dans l'exemple précédant elle nous permettent de voir directement que environ ¾ des coquelicots ont 14 loges ou moins.

• On représente graphiquement les fréquences (plus rarement les effectifs) à l'aide d'un diagramme en bâtons.

Ou par des camemberts (surtout dans le cas des variables nominales):

1.3 Variables discrètes et continues

On appelle ***variable discrète***, une variable qui ne peut prendre qu'un nombre fini ou dénombrable de valeurs, par exemple dans le cas du nombre de loges capsulaires les scores étaient donnés par les nombres {6 ; 7 ; 8 ; … ; 20}.
Si, en lieu et place de compter le nombre de loges capsulaires, nous avions mesuré la taille des coquelicots (au dixième de centimètre près), nous rendrions compte que toutes les valeurs comprises entre 0 et 50 cm pourraient potentiellement être atteintes. Dans ce cas on parle de variable continue. Comme représentation graphique le diagramme en bâton n'est pas adapté.



La raison étant qu'il est rare que deux coquelicots aient exactement la même taille.

Dans le cas des variable continues, il faut procéder à un regroupement en classes.

Définitions

Si $[a_k; b_k[$ désigne une classe (la k-ième), a_k et b_k sont appelés les bornes de la classe respectivement supérieure et inférieure.
Sa longueur b_k-a_k est appelé le ***diamètre*** de la classe (ou l'amplitude) noté δ.

$$\delta_k = b_k - a_k$$

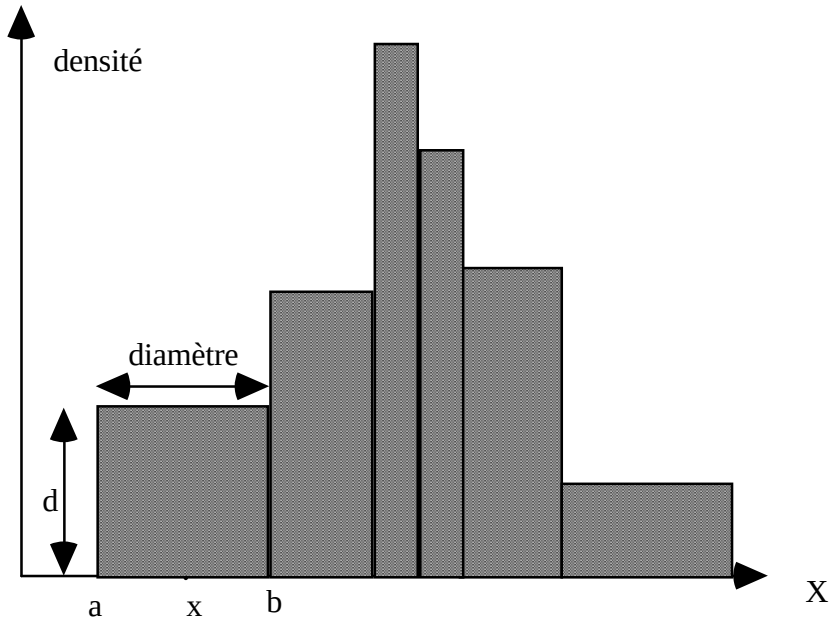
La moyenne des nombres a et b, le ***centre*** de la classe.

$$x_k = \frac{a_k + b_k}{2}$$

On parle alors d'effectifs de classe et de fréquence de classe, mais une nouvelle notion doit être introduite, la ***densité de fréquence***.
La densité de fréquence est la fréquence d'une classe divisée par son diamètre.

$$d_k = \frac{f_k}{\delta_k}$$

Dans le cas des variables continues, on représente graphiquement la densité de fréquence, c'est ce que l'on appelle un ***histogramme***.



Remarques

• Les classes doivent recouvrir tous les nombres compris entre la plus petite valeur que peut prendre la variable et la plus grande. Il ne peut donc pas y avoir d'espace entre la borne supérieure d'une classe et la borne inférieure de la suivante.

• Il faut distinguer les bornes apparentes des bornes effectives d'une classe.
Par exemple, dans le cas des âges, on trouve dans la littérature (journaux)

0 - 5
5 - 10
Alors que les années révolues correspondent aux bornes suivantes

[0; 6[
[6; 11[

• Il arrive que des variables discrètes (très étendues) soient traitées comme des variables continues. Par exemples si les scores sont des nombres d'individus, pouvant aller de 0 à 1'000. Dans ce cas, on groupera les scores en classes, 100 à 200 correspondra (par exemple) à la classe [99.5; 199.5[. C'est ce que l'on désigne habituellement par le terme de correction de continuité.

Exemple

chêne pédonculé

Pluviosite	Centre X	effectifs	frequences F
[700; 800[	750	10	1.55%
[800; 900[	850	85	13.18%
[900; 1000[	950	185	28.68%
[1000; 1100[	1050	122	18.91%
[1000; 1200[	1150	138	21.40%
[1200; 1300[	1250	43	6.67%
[1300; 1400[	1350	15	2.33%
[1400; 1500[	1450	12	1.86%
[1500; 1600[	1550	13	2.02%
[1600; 1700[	1650	10	1.55%
[1700; 1800[	1750	6	0.93%
[1800; 1900[	1850	5	0.78%
[1900; 2000[	1950	1	0.16%
Total		645	100.00%

Température	Centre X	effectifs	frequences F
[7; 8[	7.5	4	0.62%
[8; 9[	8.5	25	3.88%
[10; 11[	9.5	109	16.90%
[11; 12[	10.5	250	38.76%
[12; 13[	11.5	205	31.78%
[13; 14[	12.5	52	8.06%
Total		645	100.00%

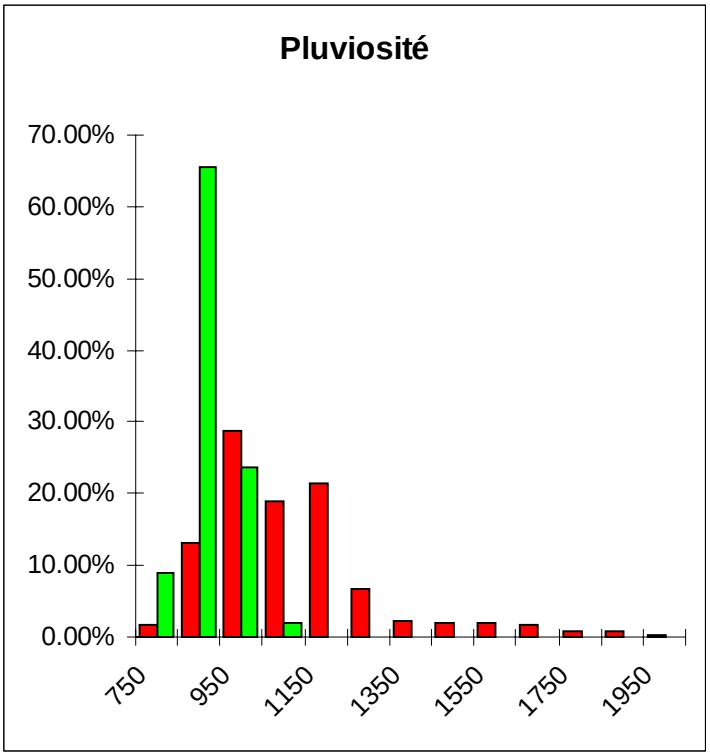
sols	effectifs	frequences F
acides	502	77.83%
calcaires	49	7.60%
montagneux	94	14.57%
Total	645	100.00%

chêne pubescent

Pluviosite	Centre X	effectifs	frequences F
[700; 800[	750	14	8.92%
[800; 900[	850	103	65.61%
[900; 1000[	950	37	23.57%
[1000; 1100[	1050	3	1.91%
Total		157	100.00%

Température	Centre X	effectifs	frequences F
[11; 12[	11.5	34	21.66%
[12; 13[	12.5	123	78.34%
Total		157	100.00%

sols	effectifs	frequences F
acides	23	14.65%
calcaires	134	85.35%
Total	157	100.00%

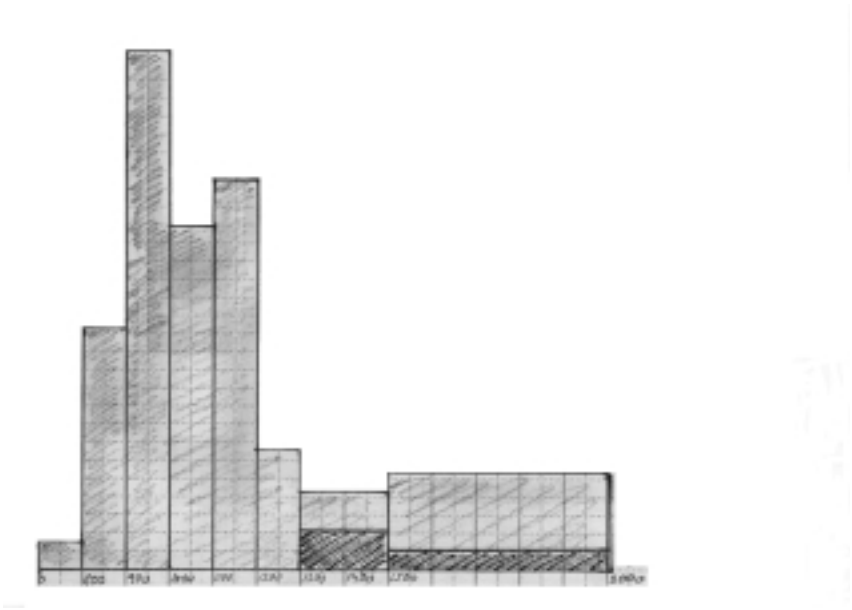


Attention, dans cet exemple toutes les classes ont le même diamètre.

Regroupons différemment, par exemple, la variable pluviosité, pour le chêne pédonculé :

Nouvelle répartition en classe		centre	diamètre	effectif	fréquence	densité	Freq.cumulées
700	800	750	100	10	1.55%	0.02%	1.55%
800	900	850	100	85	13.18%	0.13%	14.73%
900	1000	950	100	185	28.68%	0.29%	43.41%
1000	1100	1050	100	122	18.91%	0.19%	62.33%
1100	1200	1150	100	138	21.40%	0.21%	83.72%
1200	1300	1250	100	43	6.67%	0.07%	90.39%
1300	1500	1400	200	27	4.19%	0.02%	94.57%
1500	2000	1750	500	35	5.43%	0.01%	100.00%
			Total	645	100.00%	0.94%	

En représentant la fréquence (en gris) au lieu de la densité de fréquence (en noir), on surestime l'importance des classes ayant un plus grand diamètre.



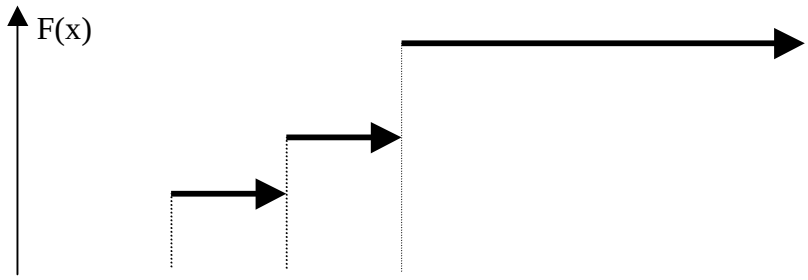
1.4 La fonction de répartition

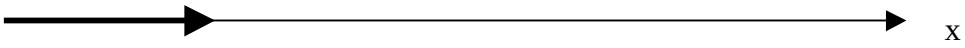
1.4.1 Cas discret

La fonction de répartition est une autre manière de décrire la distribution de la variable statistique. On associe à la variable statistique un fonction réelle définie comme :

$$F(x) = \text{Fréquence cumulée des scores} \bullet x$$

On obtient une fonction en escaliers calée sur le diagramme en bâton des fréquences cumulées. Il découle de la définition que cette fonction est continue à gauche.





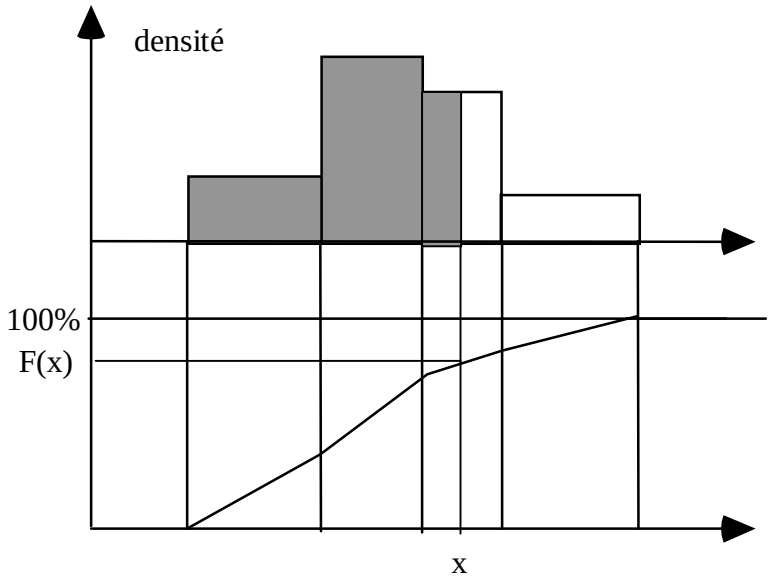
1.4.2 Cas continu

Il faut partir d'un regroupement en classes et représenter graphiquement à la fin de chaque classe (borne supérieure) la fréquence cumulée.

Rappelons que lors du regroupement, nous avons fait l'hypothèse que les scores sont uniformément distribués à l'intérieur des classes.

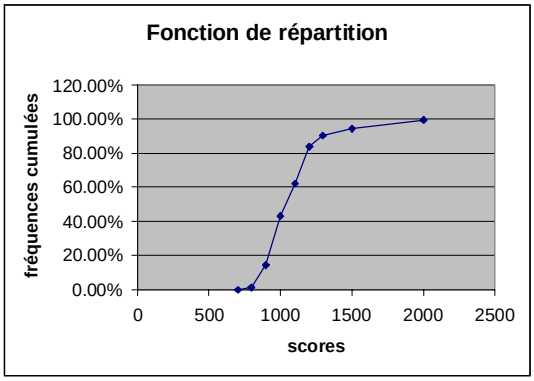
Ainsi en reliant ces points par des segments, on obtient la fonction de répartition de la V.S., qui peut s'interpréter de la manière suivante

F(x) = Fréquence cumulée des scores • x



Exemple : Reprenons la variable pluviosité, pour le chêne pédonculé

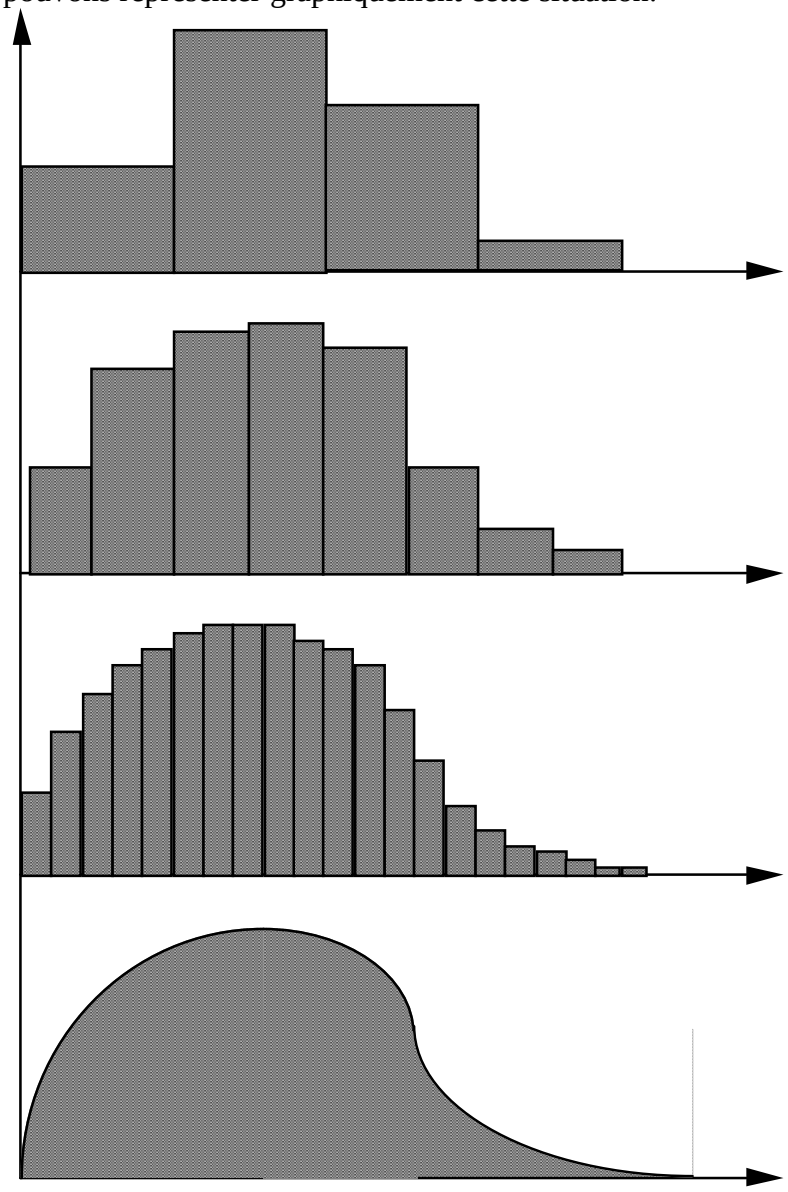
Borne sup	Freq. Cumul.
700	0.00%
800	1.55%
900	14.73%
1000	43.41%
1100	62.33%
1200	83.72%
1300	90.39%
1500	94.57%
2000	100.00%



1.5 Distribution théorique

Imaginons que nous disposions d'une population de taille infiniment grande et que nous puissions par là même diminuer les diamètres de nos classes jusqu'à des valeurs aussi petites que désiré. Alors nous faisons l'hypothèse que l'histogramme tend vers une distribution théorique qui n'est autre chose qu'une courbe.

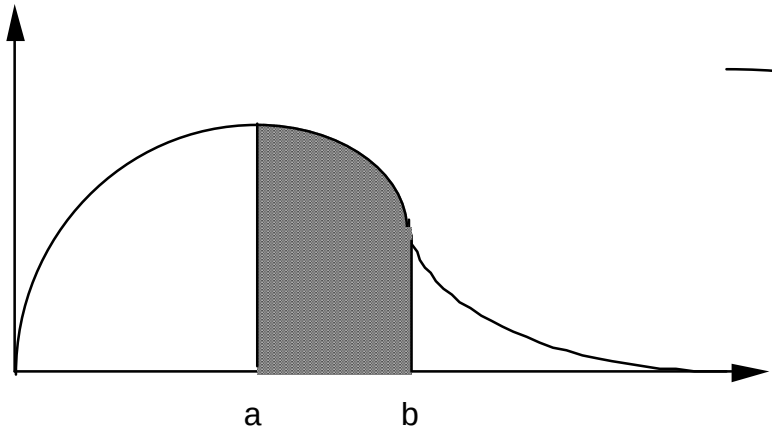
Nous pouvons représenter graphiquement cette situation:



Comment interpréter une distribution théorique, une fois que celle-ci a été identifiée?

• L'aire (ou surface) comprise entre deux valeurs *a* et *b*, représente la proportion de la population (fréquence) ayant un score compris entre *a* et *b*. Si f(*x*) désigne la densité de fréquence théorique, la fréquence de la classe [*a* ; *b*[est donnée par :

$$\int_a^b f(x)dx$$



Nous voyons ainsi qu'une condition nécessaire pour qu'une courbe puisse être une densité statistique est que l'aire comprise sous la courbe vaille 1.

$$\int_{-\infty}^{+\infty} f(x)dx = 1$$

Nous étudierons plusieurs densités théoriques, en particulier la loi normale, mais pour ce faire il nous faut introduire les principaux indicateurs de position et dispersion.

2 Indicateurs de position

Il s'agit ici de « compresser » au mieux l'information contenue dans la distribution de la variable par un nombre.

2.1 La moyenne

La notion de moyenne est bien connue de tout un chacun. La moyenne de *n*-nombres est donnée par

$$moyenne = \frac{x_1 + x_2 + \ldots x_n}{n} = \frac{\sum_{j=1}^{n} x_j}{n}$$

Dans le cas d'une variable statistique, cette formule est difficilement praticable, car elle nécessite de calculer la moyenne sur la population. C'est pourquoi il nous faut développer une formule équivalente, basée sur les scores et leurs fréquences.

Partons d'un exemple,

score	effectif	fréquence	effectif*score	fréquence*score
1	7	0.35	7	0.35
2	2	0.1	4	0.2
3	11	0.55	33	1.65
total	20	1	44	2.2

La moyenne peut donc s'obtenir en multipliant les scores par leurs effectifs, en sommant le tout et en le divisant par l'effectif total. Ceci revient à calculer la moyenne des scores pondérés par leurs fréquences.

$$moyenne = \frac{n_1 x_1 + n_2 x_2 + \ldots n_k x_k}{n} = \frac{\sum_{j=1}^{k} n_j x_j}{n} = \sum_{j=1}^{k} \frac{n_j}{n} x_j = \sum_{j=1}^{k} f_j x_j$$

On note la moyenne d'une variable statistique X, indifféremment

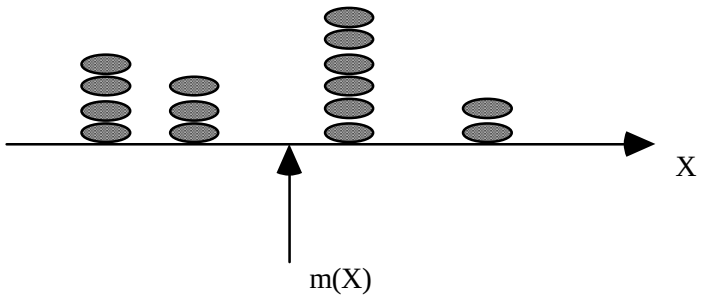
$$m = m(X) = m_X = \mu = \mu(X) = \mu_X$$

Dans le cas d'une variable continue (regroupement en classes), les calculs sont exactement les mêmes, il faut prendre les centres de classe comme valeurs des scores.

Exemple

classe	centre	fréquence	fréquence*centre
[0 ; 10[	5	23%	1.15
[10; 20[	15	46%	6.9
[20;50]	35	31%	10.85
	total	100%	18.9

Interprétation géométrique
Si à chaque unité statistique on associe un poids unitaire que l'on dispose sur un axe à la position de son score, la moyenne correspondra au centre de gravité du système.



Quelques propriétés liées à la moyenne

1.- $\sum_j f_j \cdot (x_j - \mu) = 0$

La somme des écarts à la moyenne vaut zéro.

2.- $\mu(aX + b) = a\mu(X) + b$

La moyenne est linéaire

3.- La moyenne minimise la fonction $G(z) = \sum_j f_j \cdot (x_j - z)^2$

2.2 La médiane

Grossièrement dit, la médiane est le score qui partage la population en deux parts égales.

Exemple

Salaires mensuels dans une petite entreprise de 5 salariés

(2'500.-, 3'200.-, 3'800.-, 4'500.-, 8'700.-)

moyenne = 4'540.-
médiane = 3'800.-

Modifions le dernier salaire à 22'500.-

moyenne = 7'300.-
la médiane quant à elle, n'a pas bougé. On dit que la médiane est un estimateur plus robuste que la moyenne (robustesse = résistance aux perturbations).
C'est un indicateur très utile quand les valeurs extrêmes sont peu fiables ou imprécises.

En ce qui concerne la médiane, nous sommes contraints à distinguer le cas discret du cas continu.

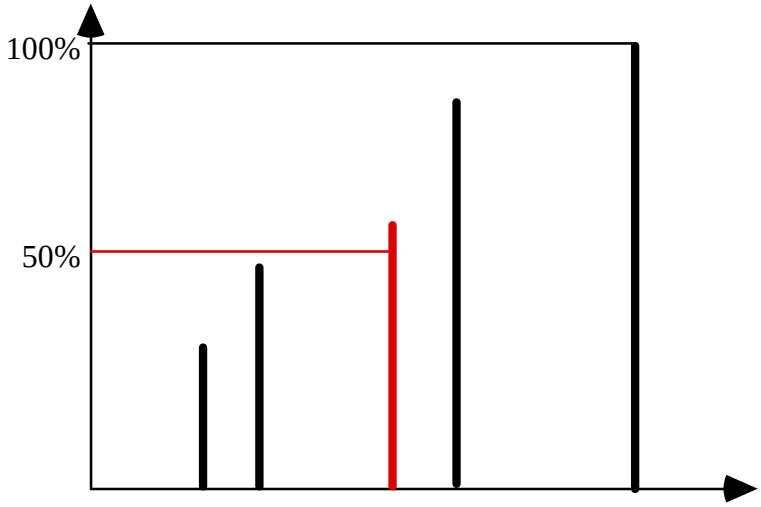
Définition (cas discret)

On appelle ***médiane***, toute valeur $\tilde{X}$ vérifiant les deux conditions

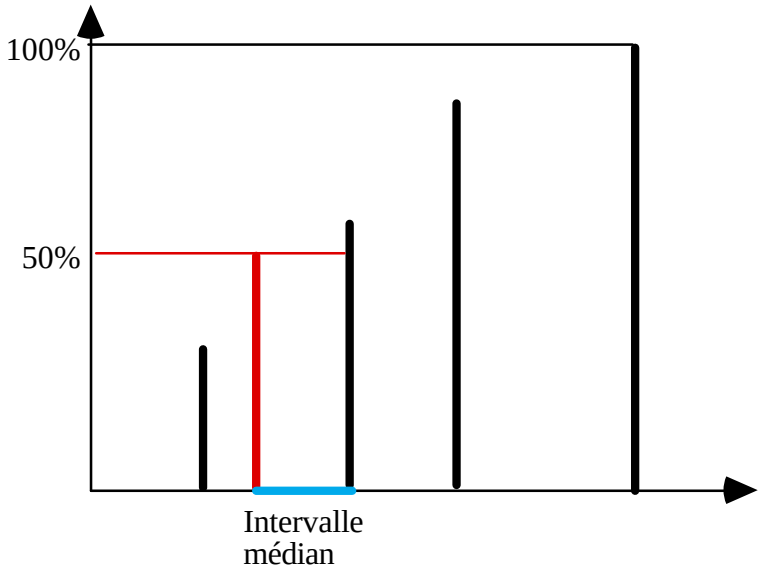
i) La moitié au plus de l'effectif total de la population à un score inférieur à cette valeur

ii) La moitié au plus de l'effectif total de la population à un score supérieur à cette valeur

Représentation graphique
Il est facile de représenter graphiquement la médiane à l'aide du diagramme en bâtons des fréquences cumulées.



Il se peut que la définition conduise à un intervalle médian, on en retient souvent le milieu comme valeur de la médiane.



Ceci arrive lorsqu'un score possède une fréquence cumulée de 50% exactement.

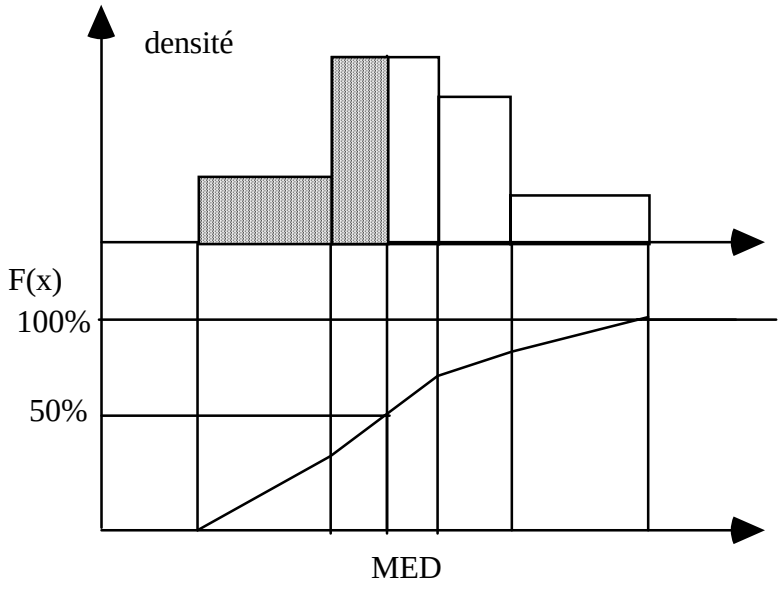
La médiane dans le cas continu

Il faut partir d'un regroupement en classes et représenter graphiquement à la fin de chaque classe (borne supérieure) la fréquence cumulée.

Rappelons que lors du regroupement, nous avons fait l'hypothèse que les scores sont uniformément distribués à l'intérieur des classes.

Ainsi en reliant ces points par des segments, on obtient la fonction de répartition de la V.S., qui peut s'interpréter de la manière suivante

F(x) = Fréquence des scores • x

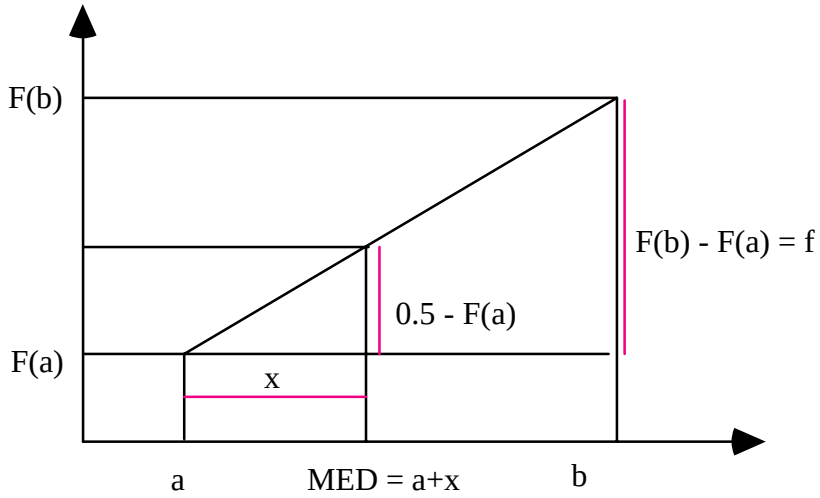


La médiane s'obtient donc comme l'image réciproque de 0,5, i.e. le score que la fonction de répartition envoie sur 0.5.

Détermination analytique de la médiane

1.- Déterminer la classe médiane [a; b[telle que F(a)• 50% et F(b) > 50%

2.- Calculer par règle de trois la position exacte de la médiane



$MED = a + x$ et x satisfait $\dfrac{x}{0.5 - F(a)} = \dfrac{\delta}{f}$

d'où $MED = a + \delta \cdot \dfrac{50\% - F(a)}{f}$

Considérons l'exemple - exercice suivant:

• compléter la table
• représenter l'histogramme
• représenter la fonction de répartition
• calculer mode et médiane

classe	diamètre δ_k	fréquence f_k	freq. cum.	densité
[0; 10[		10%		
[10; 15[		25%		
[15; 35[		40%		
[35; 50]		25%		

Quantiles
A partir de la fonction de répartition, nous avons déterminé la médiane en coupant l'intervalle [0; 1] en deux parts égales et en prenant l'image réciproque du point milieu. De la même manière il est possible de subdiviser l'intervalle [0; 1] en 4 parts égales, les points correspondants sont appelés les quartiles, (en 5 : les quintiles, en 10 les déciles, en 100 les centiles).

Au-delà de la médiane, c'est plus qu'un indicateur de position que l'on a à disposition, c'est une série de nombres qui nous permet de reconstituer la distribution
(de grossièrement pour les quartiles, à finement pour les centiles).

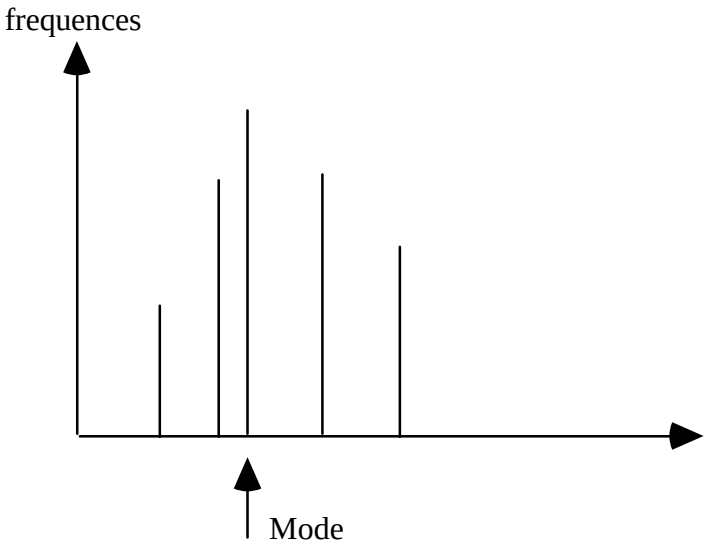
Diagramme de Tuckey ou boîte à moustache

2.3 Le mode

1.- Cas discret

Définition

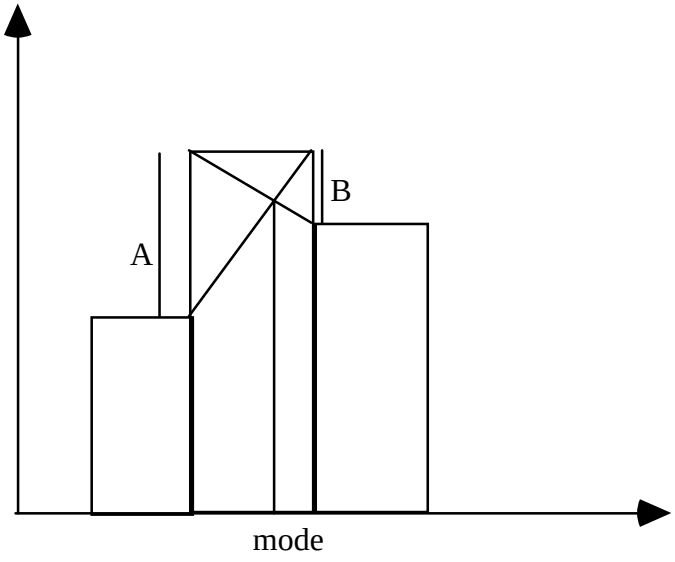
Le ***mode*** est le score ayant la plus haute fréquence (ou effectif)



2.- Cas continu

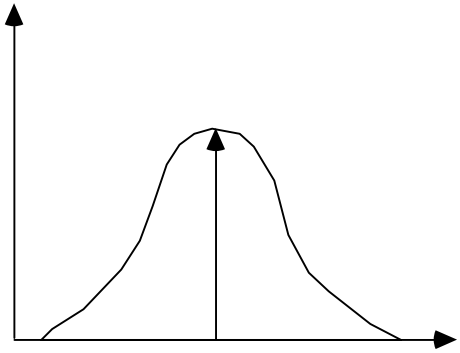
Définition
on appelle ***classe modale***, la classe ayant la plus haute densité de fréquence, et mode le centre de la cette classe.

Il possible de tenir compte de l'influence des premier voisins comme l'illustre la figure suivante:

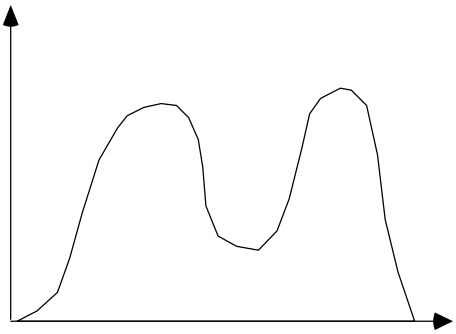


ratio = A/(A+B)

Dans le cas d'une distribution théorique, le mode est le maximum (ou les maxima) de la fonction densité.



mode



distribution bi-modale

3 Indicateurs de dispersion

L'idée étant de mesurer la dispersion de la distribution. Il y a trois manières de faire, qui correspondent à des buts différents.

• Sans référence à un indicateur de position, notion d'étendue.

• En référence à une valeur centrale (dispersion autour d'un indicateur de position).

• En indice relatif (coefficient de variation), dans un but de comparaison.

Définition

Étendue R (range) $R = x_n - x_1$

Attentions aux valeurs aberrantes
On élimine les "outliers" en considérant le 10 - 90 percentile range

$$R_{10\text{-}90} = C_{90} - C_{10}$$

Le $R_{10\text{-}90}$ correspond à une étendue où les données ont été nettoyées à l'aide d'un indicateur de position.

Dans le même ordre d'idée, on rencontre l'étendue inter-quartile.

Définition

L'Étendue inter quartile $EQ = Q_3 - Q_1$

L'intervalle semi-interquartile $DQ = EQ/2$

DQ est le pendant de l'écart-type, souvent utilisé lorsque l'on ne peut pas calculer la moyenne.

Ce sont des mesures de dispersion autour de la médiane. On procède de la même manière avec la moyenne.

Constat : La somme des écarts à la moyenne vaut 0

$$\sum_i n_i(\mu - x_i) = \mu \sum_i n_i - \sum_i n_i x_i = N(\mu - \sum_i \frac{n_i}{N} x_i) = N(\mu - \mu) = 0$$

Il est possible de palier à cette compensation des signes de deux manières:

1) En prenant la valeur absolue des écarts et en calculant leur moyenne, on obtient ainsi ***l'écart absolu moyen***.

$$E_{am} = \sum_i f_i |\mu - x_i|$$

2) Le traitement mathématique de la valeur absolue n'étant pas aisé, on lui préfère la mise au carré. On définit ainsi ***la variance***, comme étant la moyenne des carrés des écarts à la moyenne.

$$\sigma^2 = \sum_i f_i (\mu - x_i)^2$$

Pour des raisons d'unités et d'ordre de grandeur, on utilise ***l'écart-type*** qui n'est autre que la racine de la variance

$$\sigma = \sqrt{\sigma^2} = \sqrt{\sum_i f_i (\mu - x_i)^2}$$

Le coefficient de variation de l'écart-type

$$V_\sigma = \frac{\sigma}{\mu}$$

Ce n'est pas à l'aide de ces formules que l'on calcule la variance et l'écart-type, mais en appliquant le résultat suivant.

Théorème de Koenigs

$$\sigma^2 = \mu(X^2) - (\mu(X))^2$$

L'exemple suivant montre l'application de cette formule à l'aide d'un tableur.
La série statistique suivante représente le poids en Kg de 100 personnes.

Classes	effectifs
[58.5; 62.5[	5
[62.5; 65.5[	18
[65.5; 68.5[	42
[68.5; 74.5[	27
[74.5; 80.5[	8
Total	100

4 Autres indicateurs
4.1 Les coefficients de variation

Le coefficient de variation inter quartile

$$V_Q = \frac{DQ}{\tilde{X}}$$

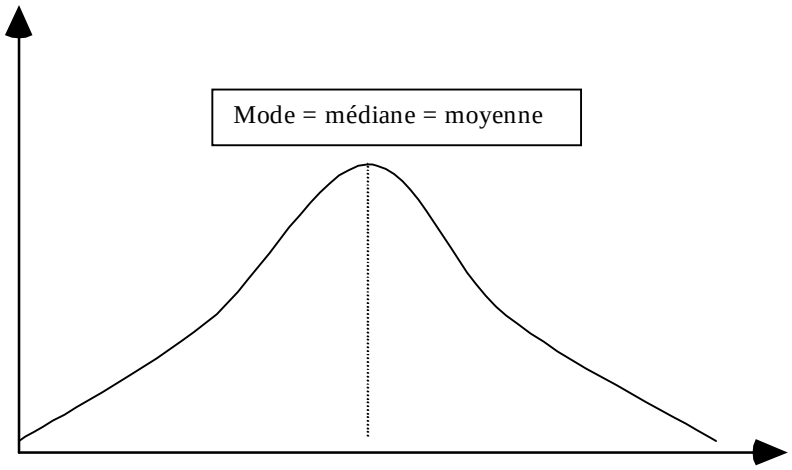
Le coefficient de variation de l'écart-type

$$V_\sigma = \frac{\sigma}{\mu}$$

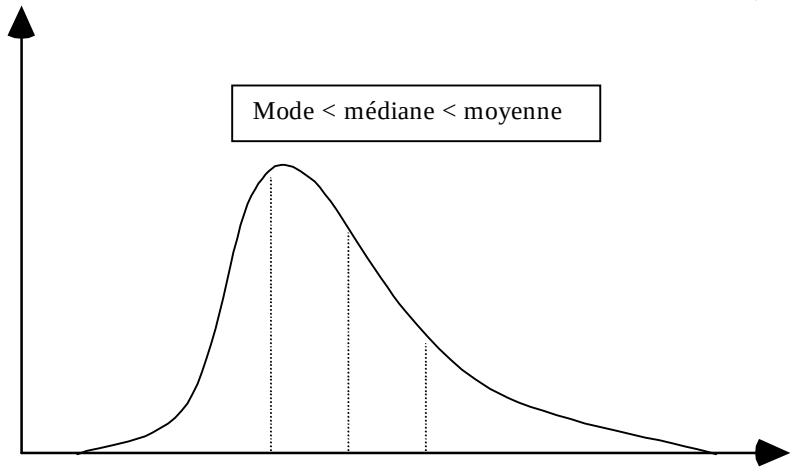
4.2 Les coefficients de dissymétrie

Voici 3 exemples de distribution d'une variable statistique.

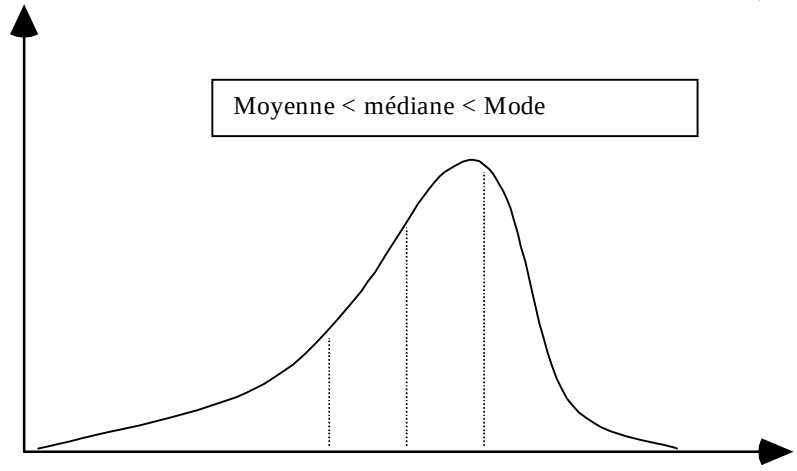
a)



b)



c)



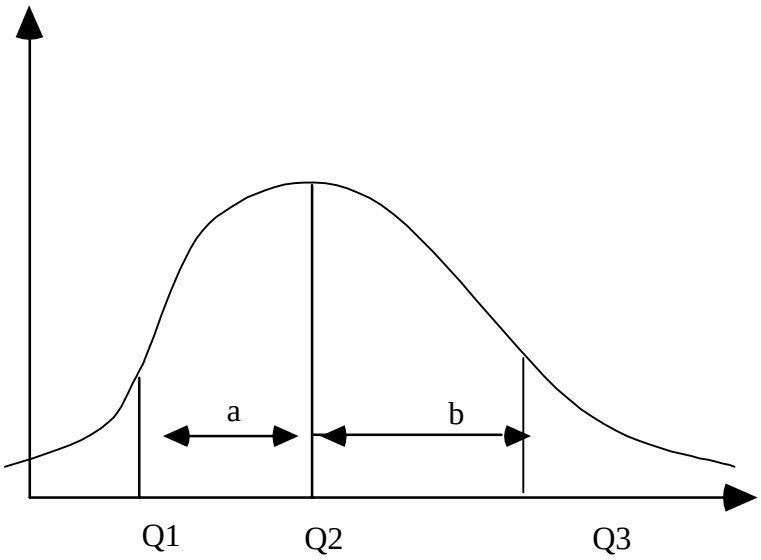
La distribution a) est dite symétrique, la moyenne la médiane et le mode sont confondus.
La distribution b) est dite biaisée à droite où positivement, à comprendre dans le sens d'une plus grande dispersion (ou étalée) à droite.
La distribution c) est dite biaisée à gauche où négativement, à comprendre dans le sens d'une plus grande dispersion (ou étalée) à gauche.
Il est existe plusieurs indicateurs permettant de rendre compte de cette situation.
Le plus utilisé est certainement le coefficient de dissymétrie de Pearson, qui se calcule facilement à partir de la moyenne, du mode et de l'écart-type.

$$DI_P = \frac{\mu - \text{mode}}{\sigma}$$

Le signe de cet indicateur correspond bien évidemment au signe du biais. Si on dispose des quartiles, on peut aussi utiliser le coefficient de dissymétrie inter-quartile

$$DI_Q = \frac{(Q_3 - Q_2) - (Q_2 - Q_1)}{Q_3 - Q_1}$$

Il s'interprète géométriquement à l'aide des distances inter-quartiles, *a* et *b*, comme le montre la figure ci-dessous.



$$DI_Q = \frac{b-a}{a+b} \begin{cases} > 0 \text{ si } b > a \\ = 0 \text{ si } b = a \\ < 0 \text{ si } b < a \end{cases}$$

5 Corrélation et Régression Linéaire

Nous allons nous occuper des liens qui peuvent exister entre deux variables définies sur la même population.

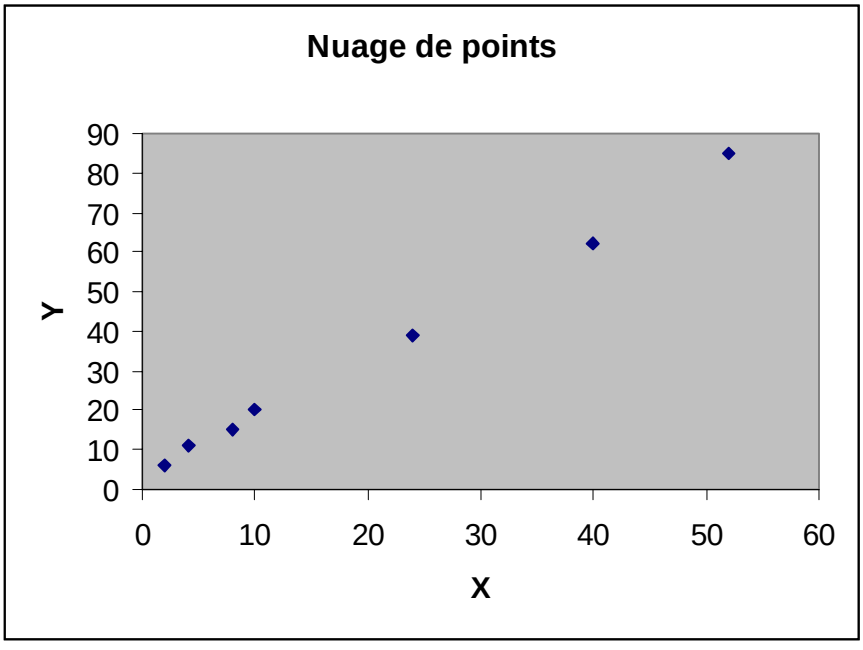
Exemple : Sur une population de feuilles,X représente le nombre de jours d'exposition au soleil et Y le nombre de stomates aérifères au millimètre carré.

X	Y
2	6
4	11
8	15
10	20
24	39
40	62
52	85

On devine le lien qui peut exister entre ces deux variables, il s'agit d'une hypothèse que nous souhaiterions analyser, le temps d'exposition influence le développement des stomates aérifères. Nous allons développer quelques outils qui nous permettront d'analyser ce genre de situation.

5.1 Nuage de points et tableau croisé

Dans l'exemple précédant, nous pouvons reporter sur un système d'axes les données conjointes (taux de change; nuitées) $(x_i;y_i)$. La représentation graphique, appelée nuage de points, montre une éventuelle tendance.



Si un score conjoint apparaît plusieurs fois, on peut soit décaler légèrement les points, soit augmenter proportionnellement à l'effectif la taille des points.

Les scores conjoints apparaissent le plus souvent avec des effectifs lorsque les variables sont données par regroupement en classe. On présente le plus souvent les données par un tableau croisé.

Exemple : Enquête sur les exploitations agricoles en France, 1981.
X âge du chef d'exploitation
Y surface agricole utilisée en ha

X/Y	0-1	1-2	2-5	5-10	10-20	20-35	35-50	50-100	>100	Total
<= 20	8	99	74	85	99	113	61	44	33	616
20-25	794	559	1120	1061	2672	4053	1611	1701	331	13'902
25-30	2185	945	2628	3325	6659	9177	5455	5350	1490	37'214
30-35	3407	2825	4687	6925	12226	16166	10810	10945	3202	71'193
35-40	5517	3791	6652	5958	12890	15983	9724	10560	3571	74'646
40-45	5981	4755	9049	10632	17502	21886	11897	14302	4213	100'217
45-50	9995	7716	16974	17290	29583	35665	18712	19555	5961	161'451
50-55	14436	11339	20880	23513	41323	45670	22253	23059	6272	208'745
55-60	14272	12519	21067	30981	52632	42998	21474	20276	5514	221'733
60-65	11541	11047	16738	17985	26750	20059	7656	7315	2510	121'601
65-70	13422	8442	15376	14116	12086	6922	2442	2342	751	75'899
70-75	12423	9389	13966	11847	8138	3917	1170	1101	564	62'515
>75	12865	7237	10124	7057	5790	2380	865	791	348	47'457
Total	106846	80663	139335	150775	228350	224989	114130	117341	34760	1'197'189

Le tableau définit la ***distribution conjointe***. Par projection, en considérant les totaux par lignes respectivement par colonne on obtient les distributions de X respectivement Y, on parle de ***distributions marginales***. Si on fixe la valeur d'une variable, par exemple X = [45 ; 50], la ligne correspondante fournit la ***distribution conditionnelle*** de Y.
Si les distributions conditionnelles de Y ou X sont toujours les mêmes (en fréquences et non en effectifs), on dit que les variables sont ***statistiquement indépendantes***.

5.2 Covariance et coefficient de corrélation

Nous avons vu que la variance d'une variable mesure sa dispersion. Nous voudrions mesurer l'écartement de deux variables. Pour ce faire, nous commençons par introduire la notion de covariance. Comme nous avons défini la variable X^2, utilisée dans le théorème de Koenigs $\sigma^2 = \mu(X^2) - (\mu(X))^2$, nous pouvons considérer la variable produit XY pour autant que les deux variables soient définies sur la même population. Alors la covariance étend la notion de variance prise au sens de la formule de Koenigs.

Définition

On désigne par ***covariance*** des variables X et Y le nombre

$$Cov(X;Y) = \mu(X \cdot Y) - \mu(X) \cdot \mu(Y)$$

remarques

• Si les variables sont indépendantes, on dit aussi non-corrélées, alors Cov(X;Y)=0
• $Cov(X;X) = \mu(X \cdot X) - \mu(X) \cdot \mu(X) = \sigma^2(X)$

On peut se demander quelle est la signification d'une variance grande ou petite. Malheureusement aucune car elle dépend des dispersions des variables *X* et Y. C'est pourquoi on introduit le coefficient de corrélation.

Définition

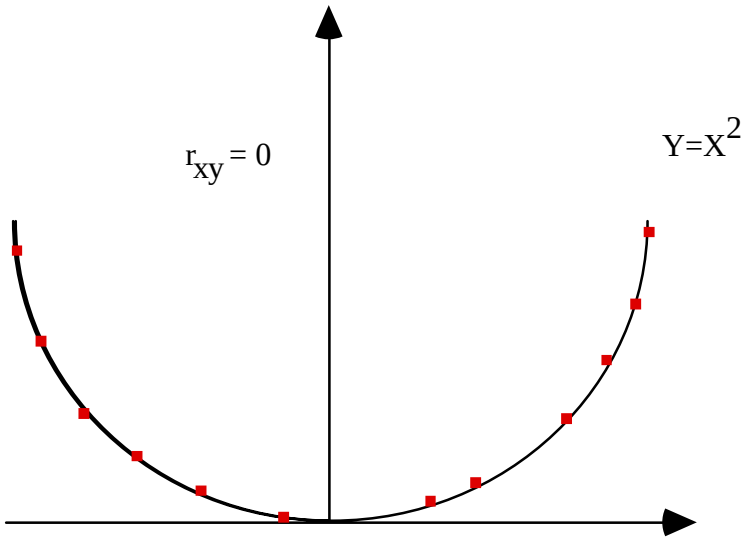
On appelle ***coefficient de corrélation*** des variables X et Y le nombre

$$r_{X;Y} = \frac{Cov(X;Y)}{\sigma_X \cdot \sigma_Y}$$

Le coefficient de corrélation est un nombre compris entre -1 et 1, qui mesure l'applatissement du nuage de points et son orientation. Ceci est représenté par le tableau suivant.

remarques

• Le coefficient de corrélation mesure une corrélation linéaire. r_{XY} peut être nul alors que la variable Y dépend fortement de X mais de façon non-linéaire. C'est pourquoi on ne devrait pas se passer d'une représentation en nuage de points.



• A l'inverse une forte corrélation ne doit pas être comprise comme une relation de causalité. Certaines variables n'ont aucune relation entre elles mais donnent lieu à des coefficients de corrélation proche de 1, ceci provient souvent du fait qu'elles sont elles mêmes influencées par une troisième variable (ou cause commune).

Exemple
Reprenons l'exemple des feuilles avec comme variable X les jours d'exposition et Y le nombre de stomates aérifères au millimètre carré.

	X	Y	x^2	Y^2	XY
	2	6	4	36	12
	4	11	16	121	44
	8	15	64	225	120
	10	20	100	400	200
	24	39	576	1521	936
	40	62	1600	3844	2480
	52	85	2704	7225	4420
Moyennes	20.00	34.00	723.43	1'910.29	1'173.14
	m(X)	m(Y)	m(X^2)	m(Y^2)	m(XY)

Cov(X;Y)	493.14
Var (X)	323.43
Ecart-type(X)	17.98
Var (Y)	754.29
Ecart-type(Y)	27.46
r(X;Y)	0.998

Exercice

Trouver dans les exemples (authentiques) suivants la cause commune.

1.- Grandeur des pieds et notes de dictées chez les 10 - 12 ans; r_{xy} proche de -1.

2.- Nombres de naissances et apparition des cigognes à Londres; r_{xy} proche de 1.

3.- Densité de nids de cigognes et taux de natalité r_{xy} proche de 1.

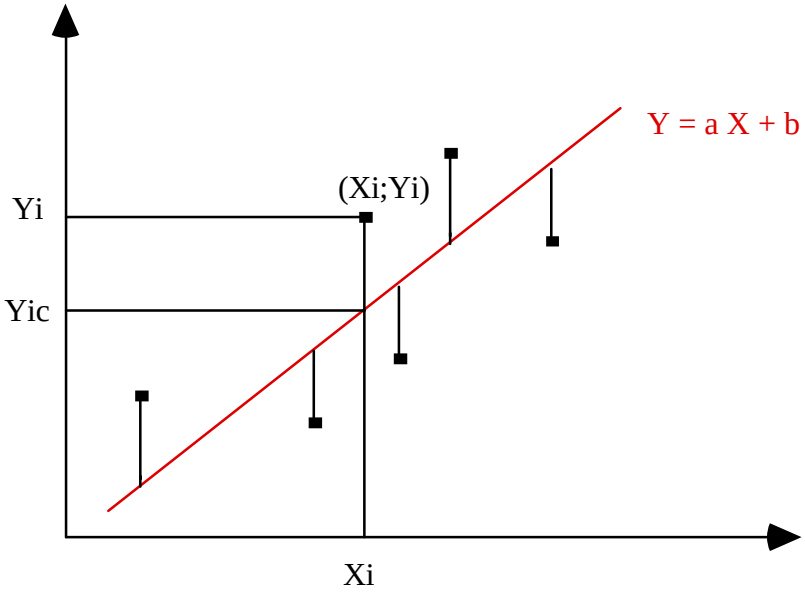
5.3 La droite de régression

Le coefficient de corrélation mesure la dépendance linéaire des variables. Si cette dépendance est bonne, on peut exprimer la variable *Y* comme fonction linéaire de *X*. C'est à dire que les valeurs y_i peuvent être remplacées par des valeurs calculées qui sont fonctions des x_i. Plus précisément

$$
\begin{aligned}
y_1 &= a\,x_1 + b \\
y_2 &= a\,x_2 + b \\
&\cdots\cdots \\
y_i &= a\,x_i + b \\
&\cdots\cdots \\
y_n &= a\,x_n + b
\end{aligned}
$$

Ce que l'on note $Y = aX + b$

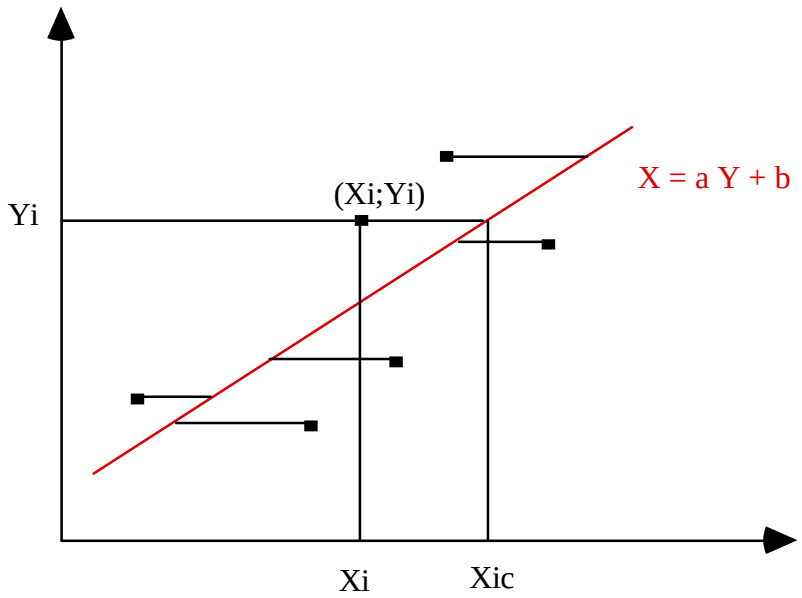
Il reste donc à déterminer les valeurs des paramètres *a* et *b*, qui désignent respectivement la pente et l'ordonnée à l'origine de la droite de régression.



On choisit alors la droite qui minimise la somme des carrés des distance entre les points y_i et les valeurs calculées correspondantes $y_{i\,c}$. (Méthode des moindres carrés).
Il est alors possible d'en déduire des formules pour a et b.

$$a = \frac{r_{xy}}{\sigma_x}\sigma_y = \frac{Cov(X;Y)}{\sigma_x^2}$$
$$b = \mu_Y - a\mu_X$$

On remarquera que ces expressions ne sont pas symétriques. En effet, si l'on veut exprimer X comme fonction de Y on obtiendra une autre droite, qui correspond à la minimisation des carrés de distances horizontales comme le montre la figure ci-dessous.



En général on régresse l'effet (Y) contre la cause (X). Cette relation de causalité ne provient pas de l'analyse statistique, mais bien de la connaissance que l'on a du phénomène considéré.

Une application intéressante de la droite de régression est l'outil de prévision que constitue cette dernière. Nous allons l'illustrer au travers de notre exemple fétiche.

Reprenons l'exemple des feuilles avec comme variable X les jours d'exposition et Y le nombre de stomates aérifères au millimètre carré.

Nous avions calculé les valeurs suivantes

Cov(X;Y)	493.14
Var (X)	323.43
Ecart-type(X)	17.98
Var (Y)	754.29
Ecart-type(Y)	27.46
r(X;Y)	0.998
m(X)=20	m(Y)=34

Calculons les paramètres *a* et *b* de la droite de régression. Nous régressons les nuitées (Y) contre les taux de change (X). On obtient

a = Cov(X;Y)/Var(X) = 1.525

b= m(Y) - a m(X) = 34 – 1.525 * 20 = 3.505

Ainsi la densité s'expriment comme y = 1.525 x + 3.505

Si pour une exposition de 45 jours on devrait prédire 1.525 *45 + 3.505= 72.118 stomates aérifères au millimètre carré.

On remarquera que si l'on ne souhaite pas connaître le coefficient de corrélation, on peut se passer du calcul de la variance de Y.

5.4 Régression et phénomènes non-linéaires

Bien que de nombreux phénomènes puissent s'exprimer raisonnablement par des corrélations linéaires, il arrive parfois que l'on soit confronté à des dépendances non-linéaires. Les plus courantes sont les dépendances quadratiques (voire polynomiales) et exponentielles. Pour les dépendances polynomiales il existe des formules analogues à celles que nous avons rencontrés dans le cas de la droite, appelées les équations normales, elles découlent aussi du principe des moindres carrés. Nous nous concentrerons sur les exponentielles.

Nous supposons que les variables *X* et *Y* sont reliées par une relation du type:

$Y = b \cdot a^X$ En prenant le logarithme de cette expression nous obtenons

$\log Y = \log(b \cdot a^X) = \log b + X \log a$

En effectuant les changements de variables $A = \log a$, $B = \log b$, $Z = \log Y$

nous nous retrouvons dans le cas d'une régression linéaire $Z = AX + B$.

Il faut bien être conscient que ceci ne correspond pas exactement à appliquer la méthode des moindres carrés sur le nuage de points original, mais sur celui que l'on a obtenu après un changement de variable qui ne respecte pas les distances (non isométrique). Ce qui revient à faire passer une droite selon les moindres carrés par le nuage de points représenté sur papier semi-logarithmique.

Exemple
Observation pendant 8 mois d'une population en extinction composée initialement de 200 individus.

modèle N(t)=a*exp(-k*t) ou aussi ln(N) = -k*t +ln(a)

	X : temps t	Y :ln(N)	N	X^2	Y^2	X*Y
	0	5.298	200	0	28.072	0.000
	1	5.193	180	1	26.967	5.193
	2	5.037	154	4	25.371	10.074
	3	4.942	140	9	24.420	14.825
	4	4.787	120	16	22.920	19.150
	5	4.718	112	25	22.264	23.592
	6	4.575	97	36	20.928	27.448
	7	4.431	84	49	19.632	31.016
	8	4.331	76	64	18.755	34.646
Moyennes	**4.000**	**4.812**	**129.222**	**22.667**	**23.259**	**18.438**

Var(X)	6.667
écart type(X)	2.582
Var(Y)	0.099
écart type(Y)	0.315
Cov(X;Y)	-0.812

r(X;Y)	-0.999
-k	-0.122
ln(a)	5.299

Estimations	
k=	0.122
t=	12
ln(N)=	3.839
N(12)=	**46**
t=	24
ln(N)=	2.378
N(24)=	**11**