REPUBLIQUE DU SENEGAL

UN PEUPLE - UN BUT - UNE FOI

MINISTERE DE L'ENSEIGNEMENT SUPERIEUR

SECRETARIAT D'ETAT A LA RECHERCHE SCIENTIFIQUE ET TECHNIQUE

INSTITUT SENEGALAIS DE RECHERCHES AGRICOLES (I.S.R.A.)

LABORATOIRE NATIONAL DE L'ELEVAGE ET DE RECHERCHES VETERINAIRES

METHODES STATISTIQUES

NOTIONS COMPLEMENTAIRES A L'USAGE DES ETUDIANTS DE L'EISMV

Par J.P. DENIS

REF. N° 046

LNERV/ZOOT/JPD/Février 1981

INTRODUCTION

La statistique descriptive a essentiellement pour but de présenter les données observées sous une forme telle que l'on puisse en prendre connaissance facilement.

Les observations considérées sont soit quantitatives, soit qualitatives. Les données quantitatives se divisent elles mêmes en dénombrements (ou comptages) et en mesures (ou mensurations>.

Dans le cas des dénombrements la caractéristique étudiée est une variable de nature discontinue ou discrète, ne pouvant prendre que des valeurs entières non négatives.

Dans le cas des mesures, la variable est de nature continue hauteur, poids...). Mais les données dont on dispose varient toujours d'une manière discontinue ; l'intervalle séparant deux valeurs consécutives pouvant être choisi par l'observateur.

Enfin les données qualitatives peuvent être assimilées au cas des variables discontinues, en supposant que les différentes variantes du caractère qualitatif soient rangées dans un ordre correspondant par exemple à la suite des nombres entiers psitifs (différentes couleurs par exemple).

A - Statistique descriptive à une dimension

Elle permet de conserver les données sous la forme de quelques paramètres ou valeurs typiques. Le calcul de ces paramètres constitue la réduction des données qui peut être utilement réalisée quel que soit leur volume.

1 - LES PARAMETRES DE POSITIONS

La moyenne arithmétique

La moyenne arithmétique qu'on appelle tout simplement moyenne est désignée par le symbole $\bar{x}$.

.../...

Elle est égale à la somme des valeurs observées x^1, x^2, ... xi, . . . xn, divisé par le nombre d'observations.

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} xi$$

Comme chaque valeur, xi doit être prise en considération autant de fois qu'elle a été observée, l'expression devient, dans le cas des distributions de fréquence

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{p} ni\ xi$$

x_1, x_2, ... xi, . ..xp étant alors les valeurs observées (distributions non groupées) ou les points centraux des classes (distributions groupées) et n_1, n_2, ... ni, . ..np les fréquences correspondantes.

La somme des écarts xi - $\bar{x}$ entre les valeurs observées et la moyenne est nulle, et c'est par rapport à la moyenne que la somme des carrés des écarts est la plus petite.

II - LES PARAMETRES DE DISPERSION

1 - La variance, l'écart type et le coefficient de variation

1/1 - La variance d'une série statistique ou d'une distribution de fréquence est la moyenne arithmétique des carrés des écarts par rapport à la moyenne :

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

$$\text{et } \sigma^2 = \frac{1}{n} \sum_{i=1}^{p} ni\ (xi - \bar{x})^2$$

1/2 - L'écart type est la racine carré de la variance.

.../...

1/3 - Le coefficient de variation est obtenu en exprimant l'écart type en valeur relative ou en pour cent de la moyenne, lorsque celle-ci est positive.

$$V = \frac{\sigma}{\bar{x}}$$

ou

$$V = \frac{100\ \sigma}{\bar{x}}$$

1/4 - Propriétés

La variance, l'écart type et le coefficient de variation sont nuls si et seulement si tous les écarts $xi - \bar{x}$ sont nuls, c'est-à-dire si toutes les valeurs observées sont égales entre elles et donc égales à leur moyenne.

$$x_1 = x_2 = \ldots = \bar{x}$$

Le coefficient de variation est totalement indépendant des unités de mesure utilisées. C'est un nombre pur, alors que l'écart type s'exprime dans les mêmes unités que les valeurs observées. L'un est un paramètre de dispersion relative, l'autre un indice de dispersion absolue.

La variance, l'écart type et le coefficient de variation ont des qualités comparables à celles de la moyenne. Le coefficient de variation, en particulier, permet de comparer la variabilité relative de plusieurs séries statistiques ou de plusieurs distributions de fréquence dont les ordres de grandeur sont très différents.

2 - Les moments

Les moments d'ordre k par rapport au point c sont définis comme suit respectivement pour les séries statistiques et pour les distributions de fréquences.

$$\frac{1}{n} \sum_{i=1}^{n} (xi - c)^k$$

et $$\frac{1}{n} \sum_{i=1}^{p} ni_1 (xi - c)^k$$

.../...

En pratique, on utilise presque exclusivement les moments par rapport à l'origine (c = o)

$$ak = \frac{1}{n} \sum_{i=1}^{n} xi^k \text{ ou } \frac{1}{n} \sum_{i=1}^{p} ni\, xi^k$$

et les moments par rapport à la moyenne (c = $\bar{x}$) ou moments centrés

$$mk = \frac{1}{n} \sum_{i=1}^{n} (xi - \bar{x})^k$$

$$\text{ou } \frac{1}{n} \sum_{i=1}^{p} ni\, (xi - \bar{x})^k$$

En particulier le moment d'ordre 1 par rapport à l'origine se confond avec la moyenne, le moment centré d'ordre 1 est toujours nul, et le moment centré d'ordre 2 se confond avec la variance.

$$a_1 = \bar{x} \quad m_1 = 0 \quad m2 = \sigma^2$$

B - Statistique descriptive à deux dimensions

Elle a essentiellement pour but de mettre en évidence les relations qui existent entre deux séries d'observations considérées simultanément.

L'étude simultanée de 2 séries d'observations fait intervenir les notions suivantes :

- la notion -- généralisée, de moment et la covariance
- les droites de régression au sens des moindres carrés
- le coefficient de corrélation et le coefficient de détermination.

La notion de corrélation concerne la netteté ou l'intensité de la relation existant entre les '2 séries de résultats, tandis que la notion de regression est liée à l'allure, supposée linéaire, de cette relation.

1 - LES MOMENTS ET LA COVARIANCE

· La généralisation à 2 dimensions de la notion de moment donne naissance aux expressions :

$$\frac{1}{n} \sum_{i=1}^{n} (xi - c)^k (yi - d)^l$$

$$\text{et } \frac{1}{n} \sum_{i=1}^{p} \sum_{p=1}^{q} nij \, (xi - c)^k (yi - d)^l$$

Elles représentent les moments d'ordre k en x et d'ordre 1 en y, par rapport à c pour x et d pour y,

Eh posant $c = \bar{x}$ et $d = \bar{y}$, on obtient les moments centrés m kl aussi appelés moments par rapport aux moyennes.

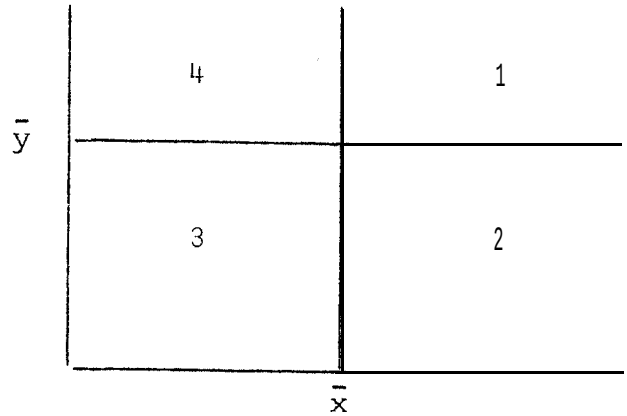
La notion généralisée de moment centré permet de définir la covariance (m11) qui caractérise simultanément les 2 séries d'observations.

$$\text{cov } (x, y) = m11 = \frac{1}{n} \sum_{i=1}^{n} (xi - \bar{x}) (yi - \bar{y})$$

$$\text{ou } \frac{1}{n} \sum_{i=1}^{p} \sum_{j=1}^{q} nij \, (xi - x) (yi - y)$$

La covariance est positive ou négative selon que la relation entre les 2 séries de données est croissante et décroissante, c'est-à-dire selon que les valeurs élevées d'une série correspondent dans l'ensemble, aux valeurs élevées ou aux valeurs peu élevées de l'autre.

Figure1



.../...

En effet en considérant les droites $x = \bar{x}$ et $y = \bar{y}$, qui divisent le plan (x, y) en 4 régions, on peut constater que les valeurs observées supérieures aux 2 moyennes ou inférieures aux 2 moyennes apportent une contribution positive à la covariance, les écarts par rapport aux moyennes étant de même signe alors que les valeurs observées supérieures à une moyenne et inférieures à l'autre lui apportent une contribution négative, les écarts par rapport aux moyennes étant de signes contraires.

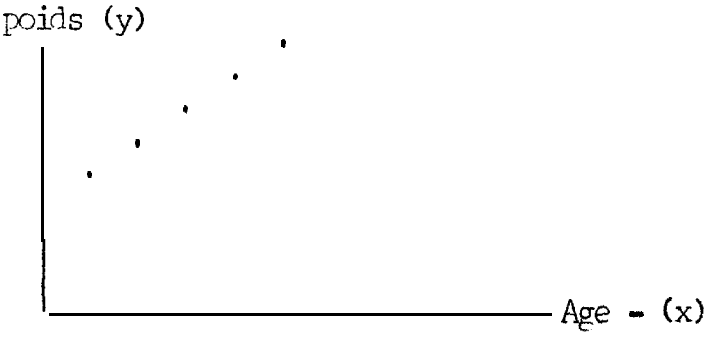
La covariance est toujours, en valeur absolue, inférieure ou égale au produit des écarts types.

$$\left| \mathbf{cov}\ (x, y) \right| \leq \sigma x\ \sigma y$$

II - LES DROITES DE REGRESSION AU SENS DES MOINDRES CARRES

Droite de régression de y en x

Le diagramme de régression de y en fonction de x (ou de y en x) est formé des points conditionnels (xi, yi). Il donne une idée de la façon dont varie en moyenne la variable y, dite dépendante, en fonction de la variable x, dite indépendante.



Lorsque le diagramme de régression est linéaire ou approximativement linéaire, on peut rechercher l'équation de la droite qui S'y ajuste le mieux. Cette droite de régression, dite aussi droite de régression de y en x, est généralement déterminée par la méthode des moindres carrés, c'est-à-dire de manière à rendre minimum la somme des carres des écarts entre les points observés et les points correspondants de la droite.

.../...

Si l'équation de la droite est $y = a + bx$ et si on dispose d'une série de valeurs observées (xi, yi) la somme des carrés des écarts à minimiser est :

$$\Sigma_{i=1}^{n} [yi - y(xi)]^2 = \Sigma_{i=1}^{n} (yi - a - bxi)^2$$

Les valeurs xi et yi étant connues, cette somme est fonction uniquement des paramètres a et b. Le minimum peut donc être déterminé en annulant les dérivées partielles par rapport à a et par rapport à b, ce qui permet d'écrire les équations normales qui indiquent que la somme algébrique des écarts entre les valeurs observées et les ordonnées correspondantes de la droite de regression est nulle : les écarts négatifs, relatifs aux points situés en dessous de la droite, compensent exactement les écarts positifs, relatifs aux points situés au dessus de la droite. On en déduit aussi que :

$$\bar{y} = a + b\bar{x}$$

$$\text{ou} \quad a = \bar{y} - b\bar{x}$$

c'est-à-dire que la droite de régression passe par le point moyen $(\bar{x}, \bar{y})$.

D'autre part, on peut écrire la droite de régression de y en x sous la forme :

$$y = \frac{\text{cov}(x, y)}{\sigma^2 x}(x - \bar{x}) + \bar{y}$$

$$\text{ou} \quad Y = byx\,(x - \bar{x}) + \bar{y}$$

Le coefficient byx étant le coefficient de régression de y en x.

Droite de régression de x en y

On peut définir par raison de symétrie la droite de régression de x en y en calculant le minimum de la somme des carrés parallèlement à l'axe des Abcisses x.

Cette droite es-l telle que

$$y = bxy\ (y - \bar{y}) + \bar{x}$$

ou bxy est le coefficient de régression de x en y

$$bxy = \frac{cov\ (x, y)}{\sigma^2 y}$$

Variance résiduel-le et écart type résiduel de y

On appelle résidus de y par rapport à x les écarts

$$yi - y\ (xi)$$

entre les points observés et les points correspondants de la droite de régression de y en x.

La variance résiduelle de y est la variance de ces résidus c'est-s-dire :

$$\sigma^2 y\ x = \frac{1}{n} \Sigma_{i=1}^{n} [yi - y\ (xi)]^2 \quad (1)$$

ou $$\sigma^2 y\ x = \frac{1}{n} \Sigma_{i=1}^{p} \Sigma_{j=1}^{q} nij\ [yi - y\ (xi)]^2$$

on peut écrire pour [(1) série statistique]

$$\sigma^2 y\ x = \sigma^2 y - cov^2\ (x, y/\ /\sigma^2 x$$

La variance résiduelle de y apparaît comme un indice de dispersion des points observés autour de la droite de régression de y en x. La quantité

$$\frac{cov^2\ (x, y)}{\sigma^2 x}$$

peut être considérée comme la part de la variance de y qui est "expliquée" par la régression de y en x, tandis que la variance résiduelle $\sigma^2 y\ x$ est la part de cette variance qui ne peut être ainsi expliquée.

L'écart type résiduel est la racine carrée de la variance résiduelle.

.../...

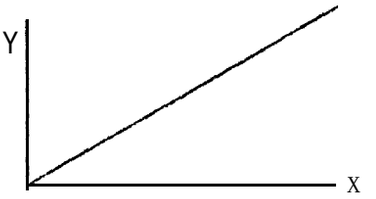
III - LE COEFFICIENT DE CORRELATION ET LE COEFFICIENT DE DETERMINATION

On a vu qur la covariance est toujours inférieure **ou égale en** valeur absolue, au produit des écarts types, Le coefficient de corrélation est le rapport de la covariance à cette valeur minimum. On le désigne par le symbole r ou r_{xy}

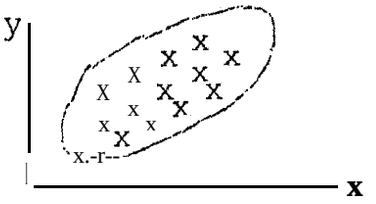
$$r = \frac{\text{cov } (x, y)}{\sigma x \ \sigma y}$$

On voit ainsi que le coefficient de corrélation possède le même signe que la covariance et qu'il est toujours compris entre - 1 et + 1.

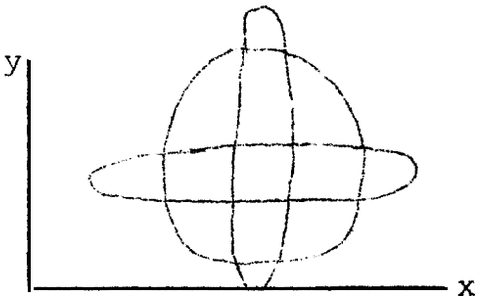
r = 1 si tous les points observés se trouvent sur une même droite de coefficient angulaire positif



o<r<1 Si le nuage des points est allongé parallèlement à une telle droite

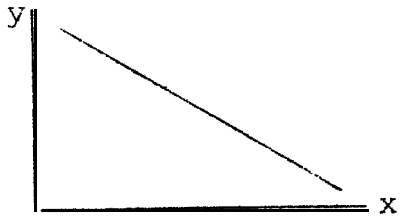


r = o si le nuage de points est allongé parallèlement à l'une des axes de coordonnéês ou s'il a une forme arrondie.

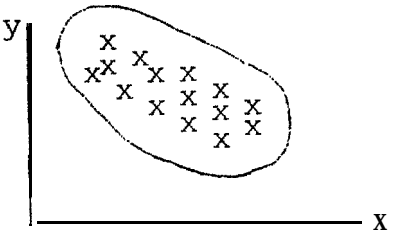


.../...

$r = -1$ si tous les points observes se trouvent sur une même droite de coefficient angulaire négatif.



$-1 < r < 0$ si le nuage de points est allongé parallèlement à une telle droite.



Le coefficient de corrélation mesure donc la netteté de la liaison existant entre les 2 séries d'observations, pour autant que cette liaison soit linéaire ou approximativement linéaire.

Dans le cas où s'applique la notion de régression de y en x, on peut en déduire que le carré du coefficient de corrélation est égal à la part de la variance de y qui est "expliquée" par la régression de y en x. Cette quantité est aussi appelée <u>coefficient de détermination</u>.

Il faut signaler que l'existence d'une corrélation même élevée entre 2 séries d'observations n'implique pas nécessairement l'existence d'une relation de cause à effet entre les 2 variables considérées.

<u>Calcul de la covariance et des paramètres dérivés</u>

La covariance d'une série statistique double peut s'écrire :

$$\text{cov}\,(x, y) = \frac{1}{n}\left[\sum_{i=1}^{n} x_i\, y_i - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)\right]$$

.../..

On peut désigner par SPE ou SPE xy la quantité mise entre crochets. [illegible] est égale à la somme des produits des écarts par rapport aux moyennes.

La détermination des variances et de la covariance nécessiter3 alors le calcul des sommes, des sommes des carrés et de la somme des produits.

$$\Sigma_{i=1}^{n} xi,\ \Sigma_{i=1}^{n} yi,\ \Sigma_{i=1}^{n} xi^2,\ \Sigma_{i=1}^{n} yi^2,\ \Sigma_{i=1}^{n} xi\ yi$$

Les paramètres dérivés de la covariance peuvent se calculer ainsi :

$$byx = \frac{SPE}{SCEx}$$

$$r = \frac{SPE}{\sqrt{SCEx\ SCEy}}$$

$$\sigma^2 y\,x = \frac{1}{n}\left[SCEy - \frac{SPE^2}{SCEx}\right]$$

La quantité mise entre crochets est à la somme des carrés des écarts résiduelle = SCEy.x

$$SPE = \Sigma_{i=1}^{n} xi\ yi - \frac{1}{n}\left(\Sigma_{i=1}^{n} xi\right)\left(\Sigma_{i=1}^{n} yi\right)$$

$$SCEx = \Sigma_{i=1}^{n} xi^2 - \frac{1}{n}\left(\Sigma_{i=1}^{n} xi\right)^2$$

$$SCEy = \Sigma_{i=1}^{n} yi^2 - \frac{1}{n}\left(\Sigma_{i=1}^{n} yi\right)^2$$

$$cov\ (xi\ yi) = \frac{SPE}{n.}$$