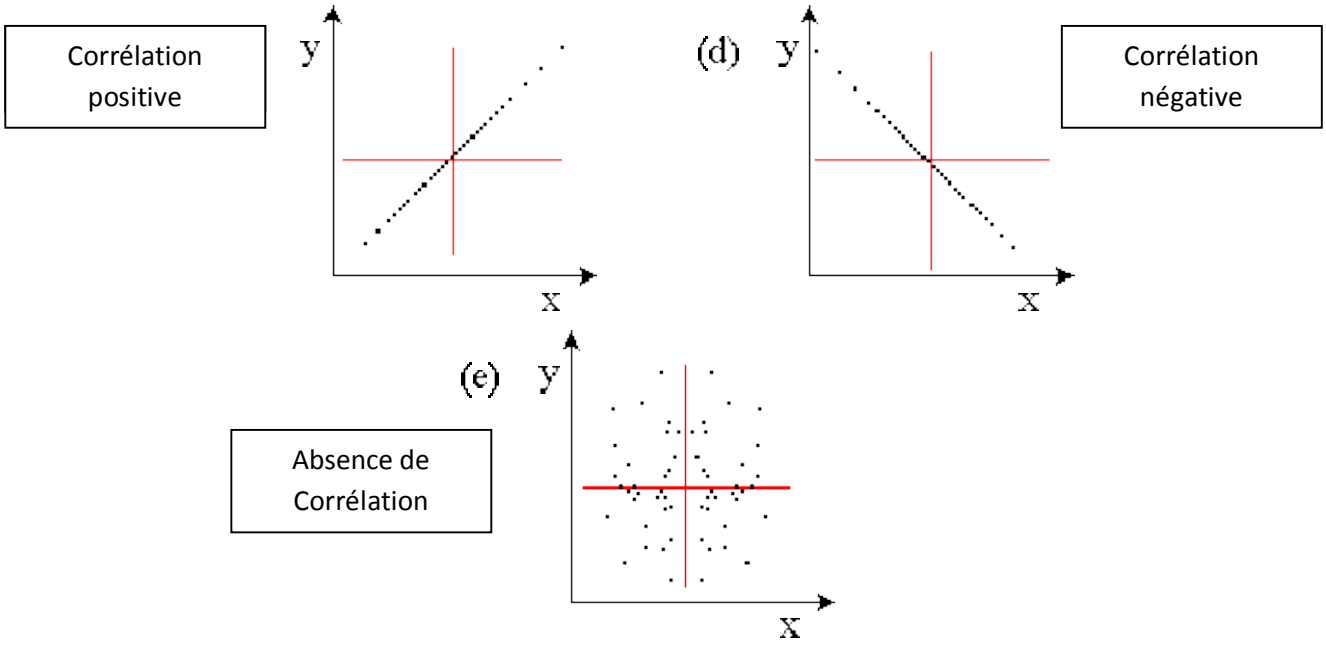
1 Introduction

Dans l'étude statistique d'une population qui porte simultanément sur deux variables quantitatives comme le rendement de blé et la dose de fertilisation, le poids et la taille des animaux d'élevage, il semble exister une relation entre les deux variables, en ce sens si on connait l'un, on doit pouvoir connaitre l'autre. On dit qu'il existe une liaison fonctionnelle ou corrélation entre les deux variables.

L'objectif est d'étudier cette liaison (cette corrélation) sous une forme mathématique, à l'aide d'une équation qui relie les variables (Méthodes de régression)

- Liaison entre 2 variables : régression simple
- Liaison entre plusieurs variables : régression multiple

Les liaisons les plus faciles à étudier sont les liaisons linéaires. Ces liaisons se présentent sous 3 formes représentées par les figures ci-dessous :



2 Coefficient de corrélation linéaire r

Le coefficient de corrélation r, chiffre l'intensité de liaison entre les deux aléatoires X et Y dans le cas d'une liaison linéaire. Soient dans une population P, deux variables aléatoires quantitatives X et Y, dont on désire étudier la liaison. L'intensité de liaison entre X et Y est chiffrée par le coefficient de corrélation linéaire r. l'étude suppose, cependant que les valeurs (xi,yi)soient distribuées selon une loi normale à deux dimensions . la dispersion des points par rapport au point moyen $(\bar{x}, \bar{y})$ se calcul par covariance qui a pour expression

$$\sigma_{xy} = \frac{\sum(xi-\bar{x})(yi-\bar{y})}{n}$$

Cette covariance rapportée aux écart-type respectifs de x et y détermine le coefficient de corrélation r

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

$$r= \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

$$\sigma_{xy} = \frac{\sum xiyi}{n} - \bar{x}\,\bar{y}$$

Le coefficient de corrélation r, varie entre -1 et + 1 :

- Quand r tend vers +1, on dit qu'il y a corrélation positive (si x augmente, y augmente) entre les deux variables ;
- Quand r tend vers -1, on dit qu'il y a une corrélation négative (si x augmente, y diminue) entre les deux variables ;
- Quand r tend vers 0, il n' y a pas de corrélation entre les deux variables

3 Liaison entre deux variables : la droite de régression (régression linéaire)

Dans la régression linéaire, on cherche à déterminer la liaison mathématique (l'équation liant les deux variables) entre les deux variables X et Y. pour déterminer l'équation liant les deux variables, on doit en premier lieu réunir les deux couples (xi,yi) correspondant à ces variables X et Y.

Supposons par exemple, que X et Y désignent respectivement la taille et le poids des vaches laitières. Alors un échantillon de n vaches (individus) révélerait les tailles x_1, x_2 ,x_3.....x_n et les poids y_1, y_2 , y_3……y_n

X : variable explicative

Y : variable expliquée

L'étape suivante, consiste à placer les points (x_1, y_1), (x_2, y_2), (x_3, y_3) ,…… (x_n, y_n) dans un système d'axes rectangulaires. L'ensemble des points obtenus est appelé Diagramme de dispersion (nuages de points).

A partir du diagramme de dispersion, on peut souvent représenter une courbe continue approchant les données, une telle courbe est appelée : Courbe d'ajustement. Le type le plus simple de courbe d'ajustement est la droite dont l'équation

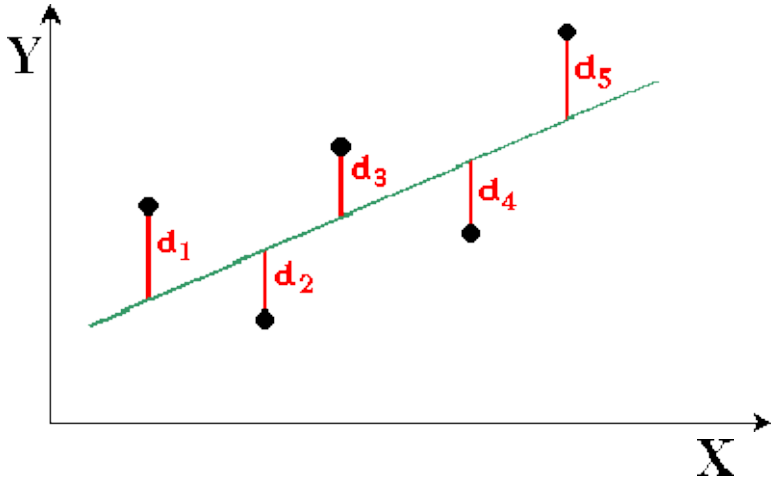
y = a x+b.

a : la pente de la droite

b : l'ordonnée à l'origine (la valeur de y quand x=0)

L'ajustement linéaire propose de remplacer le nuage de points par une droite. Mais sur le nuage de point, on peut tracer plusieurs droites à conditions que ces droites doivent passées par le point $(\bar{x}, \bar{y})$ appelé Barycentre, centre de gravité du nuage de points. Mais quelle est la meilleure droite qui ajuste bien le nuage de points ?

La meilleure droite est celle qui permet d'avoir la différence entre les valeurs yi et les valeurs correspondantes déterminées à partir de la courbe est minimale que possible



$$\sum d^2 = \text{minimum}$$

La meilleure droite est telle que $d1^2+d_2^2+d_3^2+d_4^2+d_5^2=0$; la courbe qui vérifie la propriété $d1^2+d_2^2+d_3^2+d_4^2+d_5^2$ est minimale, est la courbe des moindres carrées. La droite des moindres carrées est l'équation y= ax+b, qui permet de déterminer a et b tel que la somme des carrés des écarts entre les valeurs de yi et les valeurs correspondantes sur la droite y soit minimales : c'est la droite de régression de y par rapport à x, notée $D_Y(x)$.

L'objectif est donc de chercher une équation linéaire, qui peut résumer le nuage de points en une droite rectiligne. Autrement dit, trouver une droite passant le plus proche du nuage de points telle que la somme des carrés des écarts entre valeurs observées yi et valeurs estimées par le modèle (par l'équation) y soit le minimum possible, c'est la méthode des moindres carrés.

$$\sum(yi - y)^2 \sim soit\ minimum \text{ possible}, \sum(yi - axi - bi)^2 = 0$$

Les coefficients a et b peuvent être calculés à partir des formules suivantes :

$$a = \frac{(X_1 - \overline{X}).(Y_1 - \overline{Y}) + (X_2 - \overline{X}).(Y_2 - \overline{Y}) + \ldots + (X_n - \overline{X}).(Y_n - \overline{Y})}{(X_1 - \overline{X})^2 + (X_2 - \overline{X})^2 + \ldots + (X_n - \overline{X})^2}$$

$$a = \frac{\sum (X - \overline{X}).(Y - \overline{Y})}{\sum (X - \overline{X})^2}$$

L'ordonnée à l'origine :

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

Sachant que :

$$\overline{X} = \frac{1}{n} \sum X$$

$$\overline{Y} = \frac{1}{n} \sum Y$$

4 Le coefficient de détermination r^2

En élevant le coefficient de corrélation au carré, on obtient le coefficient de détermination r^2, exprimé en %, il indique la part de la variation de la variable y expliquée par la relation entre les variables X et Y. si tous les points ne sont pas situés sur la droite de régression, c'est que d'autres facteurs influent sur la variation de la variable Y : on dira que $(1-r^2)$% de la variation de la variable Y est attribuable à ces facteurs.