Théorie de la corrélation

Pour un échantillon de N éléments, on mesure deux variables x et y. La dispersions suivant X et Y s'écrivent

$$var(X) = s_X^2 = \frac{\sum(X - \bar{X})^2}{N},$$
$$var(Y) = s_Y^2 = \frac{\sum(Y - \bar{Y})^2}{N}.$$

La droite de régression des moindres carrés

On peut tracer x en fonction de y ou y en fonction de x sur un *diagramme de dispersion*. La droite de régression des moindres carrés s'écrit

$$Y_{est} = a_0 + a_1 X, \qquad \text{ou} \qquad X_{est} = a_2 + a_3 Y,$$

avec

$$a_0 = \frac{(\sum Y)(\sum X^2) - (\sum X)(\sum XY)}{N \sum X^2 - (\sum X)^2}$$

$$a_2 = \frac{(\sum X)(\sum Y^2) - (\sum Y)(\sum XY)}{N \sum Y^2 - (\sum Y)^2}$$

$$a_1 = \frac{N \sum XY - (\sum X)(\sum Y)}{N \sum X^2 - (\sum X)^2} = \frac{\sum(X - \bar{X})(Y - \bar{Y})}{\sum(X - \bar{X})^2}$$

$$a_3 = \frac{N \sum XY - (\sum Y)(\sum Y)}{N \sum Y^2 - (\sum Y)^2} = \frac{\sum(X - \bar{X})(Y - \bar{Y})}{\sum(Y - \bar{Y})^2}$$

Ecart type de l'estimation

Une mesure de la dispersion de la droite de régression de Y sur X est

$$s_{Y,X} = \sqrt{\frac{\sum (Y - Y_{est})^2}{N}}.$$

Une mesure de la dispersion de la droite de régression de X sur Y est

$$s_{X,Y} = \sqrt{\frac{\sum (X - X_{est})^2}{N}}.$$

En général $s_{Y,X} \neq s_{X,Y}$.
En injectant la droite de régression des moindres carrés dans ces expressions, nous obtenons

$$s_{Y,X}^2 = \frac{\sum Y^2 - a_0 \sum Y - a_1 \sum XY}{N}.$$

Variations expliquée et non expliquée

La variation totale est la somme de la variation expliquée et de la variation non expliquée par la droite des moindres carrés.

$$\sum (Y - \bar{Y})^2 = \sum (Y - Y_{est})^2 + \sum (Y_{est} - \bar{Y})^2$$

Le coefficient de corrélation

Le coefficient de corrélation s'écrit

$$r = \pm \sqrt{\frac{\text{variation expliquée}}{\text{variation totale}}} = \pm \sqrt{\frac{\sum (Y_{est} - \bar{Y})^2}{\sum (Y - \bar{Y})^2}}.$$

Sans tenir compte du signe et en injectant s_Y dans l'équation ci-dessus, il est possble de montrer que

$$r = \sqrt{1 - \frac{s_{Y,X}^2}{s_Y^2}} \qquad \text{ou} \qquad s_{Y,X} = s_Y^2 \sqrt{1 - r^2}.$$

Formule du produit des moments du coefficient de corrélation linéaire

Si l'on suppose qu'il existe une relation linéaire entre X et Y, le coefficient de corrélation devient

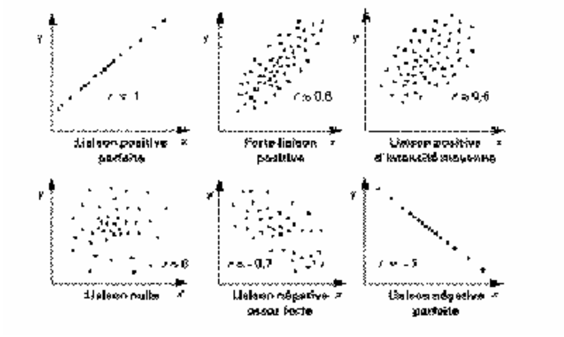
$$r = \frac{covar(X,Y)}{\sqrt{var(X)var(Y)}}$$

$$= \frac{covar(X,Y)}{\sqrt{s_X s_Y}}$$

$$= \frac{\sum(X-\bar{X})(Y-\bar{Y})}{\sqrt{(\sum(X-\bar{X})^2)(\sum(Y-\bar{Y})^2)}}$$

La nouvelle quantité $covar(X,Y)$ est appelée la covariance et s'écrit

$$covar(x,y) = \frac{1}{N}\sum(X-\bar{X})(Y-\bar{Y})$$



A2/ Test d'hypothèse sur la significativité de ρ

Les hypothèses testées :

H_0 : $\rho = 0$ ($r = 0$) indépendances des variables X et Y, pas de liens statistiques entre ces variables
H_1 : $\rho \neq 0$ ($r \neq 0$) les variables X et Y sont dépendantes

Les conditions d'applications (cf. Régression)

La statistique du test :

$$t_r = \frac{r\sqrt{N-2}}{\sqrt{1-r^2}}$$

t_r suit une loi de Student à *N-2* ddl donc

si $|\, t_r \,| > t_{seuil} \;\Rightarrow\;$ Rejet de H_0 (avec $t_{seuil} = t_{1-\alpha/2,N-2}$)

A3/ Test d'hypothèse sur $\rho = \rho_0 \neq 0$

H_0 : $\rho = \rho_0$.
H_1 : $\rho \neq \rho_0$.

$$Z_r = \frac{1}{2} Ln\left(\frac{1+r}{1-r}\right)$$

Z_r suit une loi normale de moyenne $\mu(Z_r)$ et de variance $\sigma(Z_r)$ avec :

$$\mu(Z_r) = \tfrac{1}{2} Ln\left(\tfrac{1+\rho}{1-\rho}\right) \quad \text{et} \quad \sigma(Z_r) = \tfrac{1}{n-3}$$

L'intervalle de confiance de Z_r à 1-α est $IC(1-\alpha) = \left[Z_r \pm U_{1-\alpha/2}\sqrt{\tfrac{1}{n-3}}\right]$, la transformation inverse permettra d'obtenir l'intervalle de confiance de ρ.

A4/ <u>Comparaison de 2 Coefficients de Corrélation</u>

$$Z_{r_1} = \tfrac{1}{2} Ln\left(\tfrac{1+r_1}{1-r_1}\right) \quad \text{et} \quad Z_{r_2} = \tfrac{1}{2} Ln\left(\tfrac{1+r_2}{1-r_2}\right) \quad \text{avec}$$

$$\mu(Z_{r_1}) = \tfrac{1}{2} Ln\left(\tfrac{1+\rho_1}{1-\rho_1}\right) \quad \text{et} \quad \sigma(Z_{r_1}) = \tfrac{1}{n_1-3}$$

$$\mu(Z_{r_2}) = \tfrac{1}{2} Ln\left(\tfrac{1+\rho_2}{1-\rho_2}\right) \quad \text{et} \quad \sigma(Z_{r_2}) = \tfrac{1}{n_2-3}$$

Les hypothèses testées :

$H_0 : \rho_1 = \rho_2 \ (Z_{r_1} = r_2)$
$H_1 : \rho_1 \neq \rho_2$

La statistique du test :

$$U = \frac{\left|Z_{r_1} - Z_{r_2}\right|}{\sqrt{1/(N_1-3) + 1/(N_2-3)}}$$

U suit une loi Normale donc

si $|U| > U_{1-\alpha/2}$ => Rejet de H_0