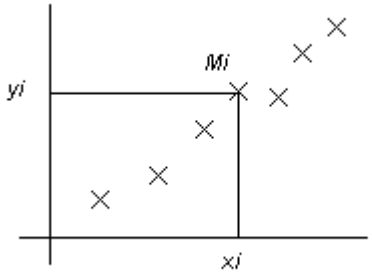
STATISTIQUES A DEUX VARIABLES - COURS

Lorsqu'une étude est faite sur un ensemble présentant deux caractères quantitatifs discrets, on obtient une série statistique double $(x_i ; y_i)$. L'économiste ou le gestionnaire cherchera s'il y a un lien de cause à effet entre ces deux caractères, et le statisticien quantifiera ce lien.

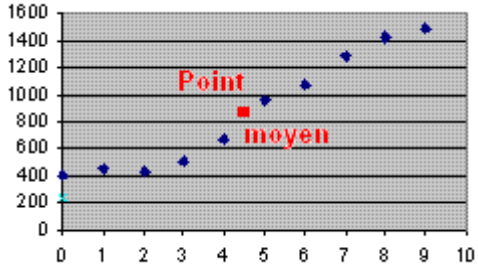
1) Nuage de points, covariance

Définitions :

Dans un repère orthogonal bien choisi, l'ensemble des points $M_i(x_i ; y_i)$ est appelé le **nuage de points**.



Si on note $\overline{x}$ la moyenne des valeurs x_i et $\overline{y}$ la moyenne des valeurs y_i, le point $G(\overline{x};\overline{y})$ est appelé **point moyen** de la série double.



Pour mesurer la dispersion des points d'un nuage par rapport au point moyen, on utilise :

Définition :

Si $(x_i ; y_i)$ est une série statistique double de *n* points, et si on note $\overline{x}$ la moyenne des valeurs x_i et $\overline{y}$ la moyenne des valeurs y_i, on définit la **Covariance** des deux variables *x* et *y* par :

$$Cov(x;y) = \frac{1}{n}\sum_{i=1}^{i=n}(x_i - \overline{x})(y_i - \overline{y})$$ (La covariance est la moyenne des produits des écarts de chacune des valeurs de la série par rapport à SA moyenne)

Théorème de Koenig :

De manière analogue à la variance, il existe une formule permettant de calculer directement la covariance de deux variables *x* et *y* :

$$Cov(x;y) = \frac{1}{n}\sum_{i=1}^{i=n} x_i y_i - \left(\frac{1}{n}\sum_{i=1}^{i=n} x_i\right)\left(\frac{1}{n}\sum_{i=1}^{i=n} y_i\right) = \overline{xy} - \overline{x}\times\overline{y}$$

où $\overline{xy}$ représente la moyenne des produits $x_i y_i$

Pour retenir : Covariance = Moyenne des produits – Produit des moyennes

jgcuaz@hotmail.com

Démonstration :

$$Cov(x;y)=\frac{1}{n}\left(\sum_{i=1}^{i=n}(x_i-\bar{x})(y_i-\bar{y})\right)=\frac{1}{n}\left(\sum_{i=1}^{i=n}x_iy_i-\bar{x}\sum_{i=1}^{i=n}y_i-\bar{y}\sum_{i=1}^{i=n}x_i+\sum_{i=1}^{i=n}\bar{x}\bar{y}\right)$$

$$=\frac{1}{n}\left(\sum_{i=1}^{i=n}x_iy_i-\bar{x}\times n\bar{y}-\bar{y}\times n\bar{x}+n\bar{x}\bar{y}\right)$$

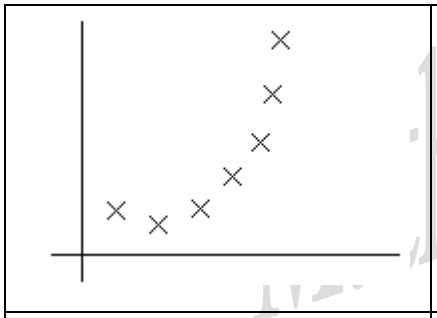
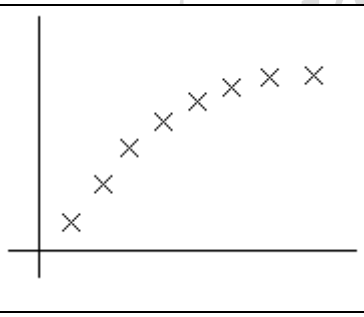
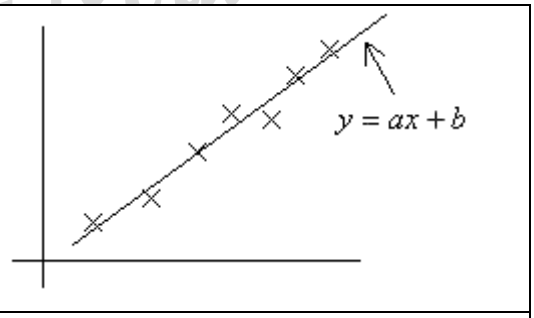
$$=\frac{1}{n}\sum_{i=1}^{i=n}x_iy_i-2\bar{x}\bar{y}+\bar{x}\bar{y}=\frac{1}{n}\sum_{i=1}^{i=n}x_iy_i-\bar{x}\bar{y}$$

Autrement dit, la Covariance est égale à la **moyenne du produit moins le produit des moyennes.**

$$Cov(x,y)=\underbrace{\overline{x\times y}}_{\text{moyenne du produit}}-\underbrace{\bar{x}\times\bar{y}}_{\text{produit des moyennes}}$$

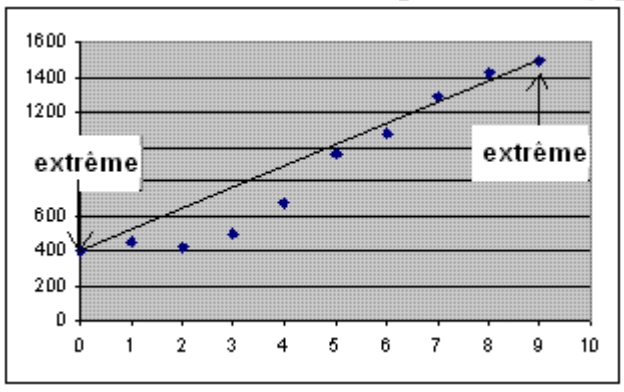
2) AJUSTEMENTS

Suivant la forme du nuage de points $(x_i;y_i)$, on peut essayer de trouver une fonction qui modélise le lien entre les deux caractères x et y, de telle façon que la courbe d'équation $y=f(x)$ passe le "plus près possible" du nuage de points

		
Forme « parabolique »	Forme « racine »	Forme « affine »

2-1) Ajustement par la droite des extrêmes

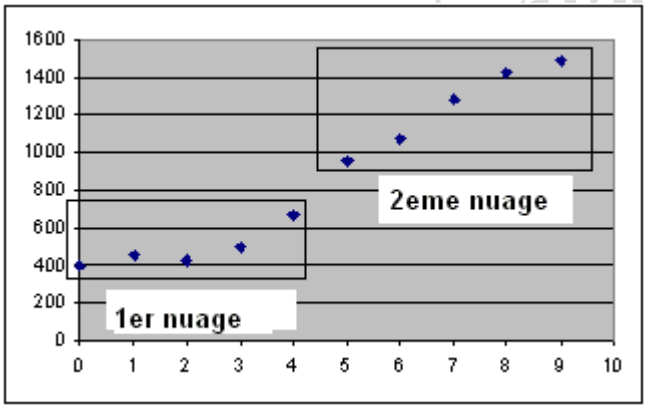
Cette méthode consiste à ajuster le nuage de points par la droite qui relie les deux points extrêmes du nuage (le premier et le dernier)

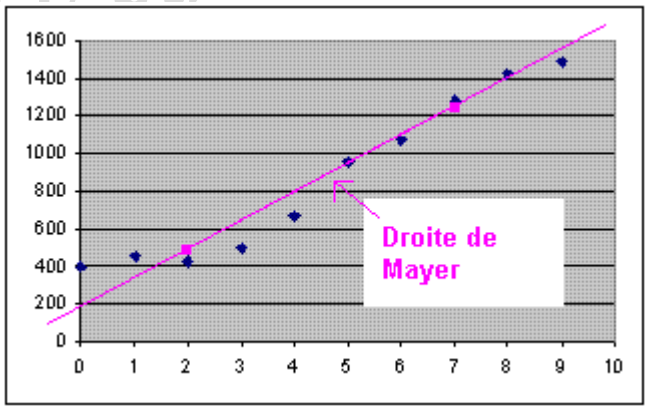


jgcuaz@hotmail.com

2-2) Ajustement par la méthode de Mayer

Cette méthode consiste à diviser le nuage en deux sous-nuages, de points moyens respectifs G_1 et G_2 et d'ajuster le nuage à l'aide de la droite (G_1G_2)



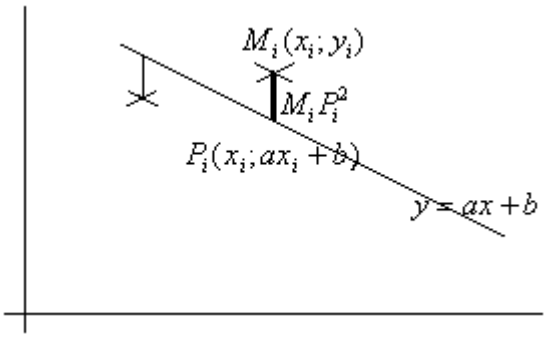


2-3) Ajustement affine par la méthode des moindres carrés

La méthode proposée définit ce que l'on entend par "passer le plus près possible".

On considère une série statistique double, représentée par le nuage de points $M_i(x_i;y_i)$.

On cherche une droite D d'équation $y=ax+b$ pour laquelle la somme des carrés $M_iP_i^2$ où les points P_i sont les projections des points M_i sur la droite soit minimale.



Autrement dit, on recherche les réels *a* et *b* tels que la somme $\sum(y_i-ax_i-b)^2$ soit minimale.

Cette somme est appelée **Somme des résidus** en *y*.

Théorème :

La droite d'équation $y=ax+b$ telle que $a=\dfrac{Cov(x;y)}{V(x)}$, et qui passe par le point moyen $G(\overline{x};\overline{y})$ est la droite qui rend minimale la somme des résidus en y $\sum(y_i-ax_i-b)^2$

Démonstration :

Notons $f(a,b)=\sum_{i=1}^{i=n}(y_i-ax_i-b)^2$ la somme des résidus en *y*. Alors :

jgcuaz@hotmail.com

$$f(a,b)=\sum_{i=1}^{i=n}\left(y_i-ax_i-b\right)^2=\sum_{i=1}^{i=n}b^2-\sum_{i=1}^{i=n}2b\left(y_i-ax_i\right)+\sum_{i=1}^{i=n}\left(y_i-ax_i\right)^2$$

$$=nb^2-2nb\left(\bar{y}-a\bar{x}\right)+\sum_{i=1}^{i=n}\left(y_i-ax_i\right)^2$$

$$=n\left[b^2-2b\left(\bar{y}-a\bar{x}\right)+\frac{1}{n}\sum_{i=1}^{i=n}\left(y_i-ax_i\right)^2\right]$$

$$=n\left[\left(b-\left(\bar{y}-a\bar{x}\right)\right)^2-\left(\bar{y}-a\bar{x}\right)^2+\frac{1}{n}\sum_{i=1}^{i=n}y_i^2-\frac{1}{n}\sum_{i=1}^{i=n}2ax_iy_i+\frac{1}{n}\sum_{i=1}^{i=n}a^2x_i^2\right]$$

$$=n\left[\left(b-\left(\bar{y}-a\bar{x}\right)\right)^2-\bar{y}^2+2a\overline{\overline{xy}}-a^2\bar{x}^2+\frac{1}{n}\sum_{i=1}^{i=n}y_i^2-\frac{1}{n}\sum_{i=1}^{i=n}2ax_iy_i+\frac{1}{n}\sum_{i=1}^{i=n}a^2x_i^2\right]$$

$$=n\left[\left(b-\left(\bar{y}-a\bar{x}\right)\right)^2+a^2\left(\frac{1}{n}\sum_{i=1}^{i=n}x_i^2-\bar{x}^2\right)-2a\left(\frac{1}{n}\sum_{i=1}^{i=n}x_iy_i-\overline{\overline{xy}}\right)+\frac{1}{n}\sum_{i=1}^{i=n}y_i^2-\bar{y}^2\right]$$

$$=n\left[\left(b-\left(\bar{y}-a\bar{x}\right)\right)^2+a^2V(x)-2aCov(x,y)+V(y)\right]$$

$$=n\left[\left(b-\left(\bar{y}-a\bar{x}\right)\right)^2+a^2\sigma(x)^2-2aCov(x,y)+\sigma(y)^2\right]$$

$$=n\left[\left(b-\left(\bar{y}-a\bar{x}\right)\right)^2+\left(a\sigma(x)-\frac{Cov(x;y)}{\sigma(x)}\right)^2-\left(\frac{Cov(x;y)}{\sigma(x)}\right)^2+\sigma(y)^2\right]$$

$$=n\left[\left(b-\left(\bar{y}-a\bar{x}\right)\right)^2+\left(a\sigma(x)-\frac{Cov(x;y)}{\sigma(x)}\right)^2+\frac{\sigma(x)^2\sigma(y)^2-Cov(x;y)^2}{\sigma(x)^2}\right]$$

Il est clair que $f(a,b)$ est alors minimum si les deux premiers carrés de la somme sont nuls, à savoir si $a=\frac{Cov(x;y)}{\sigma(x)^2}=\frac{Cov(x;y)}{V(x)}$ et si $\bar{y}=a\bar{x}+b$

Remarque : On peut également réaliser un ajustement affine de la variable *x* en fonction de *y*.

Alors l'équation de la droite de régression de *x* en fonction de *y* est donnée par $x=a'y+b'$

Avec $a'=\frac{Cov(x;y)}{\sigma(y)^2}=\frac{Cov(x;y)}{V(y)}$

3) Coefficient de corrélation linéaire

Définition :

On appelle coefficient de corrélation linéaire entre les variables *x* et *y* le rapport $r=\frac{Cov(x;y)}{\sigma_X\sigma_Y}$

Théorème :

jgcuaz@hotmail.com

1) Il a le même signe que les coefficients directeurs a et a' des droites de régression.

2) Son carré est le produit de ces coefficients : $r^2 = aa'$.

3) Le coefficient de corrélation linéaire r vérifie $-1 \leq r \leq 1$

4) Les points $M_i(x_i; y_i)$ sont alignés si et seulement si $r = 1$ *ou* $r = -1$

Démonstration :

1) On a $r = \dfrac{Cov(x;y)}{\sigma_X \sigma_Y} = \dfrac{Cov(x;y)}{(\sigma_X)^2} \times \dfrac{\sigma_X}{\sigma_Y} = a \times \dfrac{\sigma_X}{\sigma_Y}$

Comme $\sigma_X \geq 0$ et $\sigma_Y \geq 0$, a et r sont de même signe

2) Avec $a = \dfrac{Cov(x;y)}{V(x)}$ et $a' = \dfrac{Cov(x;y)}{V(y)}$, on a :

$$aa' = \frac{Cov(x;y)}{V(x)} \frac{Cov(x;y)}{V(y)} = \frac{Cov(x;y)^2}{V(x)V(y)} = \frac{Cov(x;y)^2}{\sigma(x)^2 \sigma(y)^2} = r^2$$

3) De plus $a^2 \times \dfrac{V(x)}{V(y)} = \dfrac{Cov(x;y)^2}{V(x)^2} \times \dfrac{V(x)}{V(y)} = \dfrac{Cov(x;y)^2}{V(x)V(y)} = r^2$

La droite D de régression de y en x par moindres carrés admet une équation de la forme $y = ax + b$. Comme

$G \in D$, $\bar{y} = a\bar{x} + b \Leftrightarrow b = \bar{y} - a\bar{x}$

La somme des résidus est donnée par :

$$f(a,b) = \sum_{i=1}^{i=n} (y_i - ax_i - b)^2 = \sum_{i=1}^{i=n} \left(y_i - a(x_i - \bar{x}) - \bar{y}\right)^2 = \sum_{i=1}^{i=n} \left(y_i - \bar{y} - a(x_i - \bar{x})\right)^2$$

$$= \sum_{i=1}^{i=n} \left[(y_i - \bar{y})^2 - 2a(x_i - \bar{x})(y_i - \bar{y}) + a^2 (x_i - \bar{x})^2\right]$$

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

$$= nV(y) - 2a \times nCov(x;y) + a^2 \times nV(x)$$

$$= n\left[V(y) - 2a \times aV(x) + a^2 \times V(x)\right] \quad car \quad a = \frac{Cov(x;y)}{V(x)} \Rightarrow Cov(x;y) = aV(x)$$

$$= n\left[V(y) - a^2 \times V(x)\right] = nV(y)\left[1 - a^2 \times \frac{V(x)}{V(y)}\right] = nV(y)\left(1 - r^2\right)$$

Cette somme étant positive, il en résulte que $1 - r^2 \geq 0 \Leftrightarrow -1 \leq r \leq 1$

Enfin M_i alignés $\Leftrightarrow 1 - r^2 = 0 \Leftrightarrow r = \pm 1$

jgcuaz@hotmail.com