I-Covariance :

Introduction (Rappelle variance)

En deuxième et en troisième année on a vu que la variance permet une mesure de l'écart à la moyenne des valeurs de la variable d'une série statistique simple. On peut se demander : existe-t-il un paramètre qui permet de mesurer la dispersion des points du nuage par rapport au point moyen dans le cas d'une série double ?

Définition :

Soit (X , Y), une série statistique double sur un échantillon de taille n.
On appelle covariance de (X , Y) le réel noté $cov(X\,,Y)$ défini par
$cov(X\,,Y) = \frac{1}{n}(\sum_{i=1}^{n} x_i.y_i) - \bar{X}.\bar{Y}$ où $(x_i\ ,y_i)$ est la valeur observée pour l'individu i si X et Y sont discrètes, ou bien le centre de la classe si l'une des variables est continue.

Conséquence : On a : $cov(X\,,Y) = cov(Y\,,X)$

Remarque :

La variance permet une mesure de l'écart à la moyenne des valeurs de la variable d'une série statistique simple :
1)La covariance permet une mesure de la dispersion des points du nuage par rapport au point moyen
2)La covariance est positive si X et Y ont tendance à varier dans le même sens
3)La covariance est négative si X et Y ont tendance à varier en sens contraire

Propriétés :

Soient $(x_i\ ,y_i)$ avec $1 \le i \le n$, une série statistique doubles, $\alpha \in IR$ et $\beta \in IR$ on a :

$$cov(x+\alpha\,,y+\beta) = cov(x\,,y) \text{ et } cov(\alpha x\,,\beta y) = \alpha.\beta cov(x,y)$$

$\boxed{Alpha\ 6}$ $\boxed{Alpha\ 9}$ $\boxed{Alpha:}$ *ou bien* $\left(\frac{1}{\underbrace{\boxed{alpha\ 0}}_{n}}\right)\underbrace{\boxed{alpha1}}_{\sum xy} - \underbrace{\boxed{alpha4}}_{\bar{X}} \times \underbrace{\boxed{alpha7}}_{\bar{Y}}$

Définition :

Soit (X , Y), une série statistique double sue un échantillon de taille n
Soit n_i le nombre de fois qu'apparaît le couple $(x_i\ ,y_i)$

$$cov(X\,,Y) = \frac{1}{n}\left(\sum_{j=1}^{q}\sum_{i=1}^{p} n_{ij}x_iy_i\right) - \bar{X}.\bar{Y}$$

II-Ajustement :

Introduction :

L'analyse d'un nuage de point $M_i(x_i\,,y_i)$ représentant une série statistique double $(x_i\,,y_i)$ peut conduire à la recherche d'une liaison entre les deux variables x et y . Cette liaison aide, entre autre, à faire des prévisions et à répondre à des questions parfois décisives.
Une question s'impose alors : peut-on trouver une formule mathématique qui exprime le lien entre les deux variables ? la réponse à cette question conduit à étudier le type de relation entre les deux variables (affine, polynomiale, homographique, logarithmique, exponentiel). On parle d'ajustement

Ajustement affine d'une série statistique double :

Méthode de Mayer :

Activité

Le tableau ci-dessus donne le relevé des valeurs d'une action (en DT) sur 10 jours consécutifs d'une bourse.

Jour X	1	2	3	4	5	6	7	8	9	10
Valeur Y	18.8	18.9	18.9	19.5	19.2	19	19.2	19.6	19.5	19.7

On note par le nuage N_1 des points associé à la série $(x_i\ , y_i)$ avec $1 \leq i \leq 5$ et N_2 le nuage des points restant.

1)Déterminer le point moyen G_1 de la première série

2)Déterminer le point moyen G_2 de la deuxième série

3)Déterminer l'équation de la droite (G_1G_2)

4)La droite (G_1G_2) passe telle par le point moyen G de la série totale ?

Réponse

$$\left.\begin{array}{l} 1)\ \overline{X}_1 = \dfrac{1+2+3+4+5}{5} = 3 \\ \overline{Y}_1 = \dfrac{18.8+..+19.2}{5} = 19.06 \end{array}\right\} G_1(3; 19.06)$$

$$\left.\begin{array}{l} 2)\ \overline{X}_2 = \dfrac{6+\cdots+10}{5} = 8 \\ \overline{Y}_2 = \dfrac{19+\cdots+19.7}{5} = 19.4 \end{array}\right\} G_2(8; 19.4)$$

3) L'équation réduite de la droite (G_1G_2) est de la forme $y = ax + b$

$$a = \frac{\overline{Y}_2 - \overline{Y}_1}{\overline{X}_2 - \overline{X}_1} = \frac{19.4 - 19.06}{8 - 3} = 0.068 \quad \text{et } b = \overline{Y}_1 - a\overline{X}_1 = 19.06 - 3 \times 0.068 = 18.856$$

$$(G_1G_2): y = 0.068x + 18.856$$

$$\left.\begin{array}{l} 4)\ \overline{X} = \dfrac{1+2\ldots+10}{5} = 5.5 \\ \overline{Y} = \dfrac{18.8+..+197}{5} = 19.23 \end{array}\right\} G(5.5; 19.23)$$

$$0.068 \times 5.5 + 18.856 = 19.23 \Longrightarrow G \in (G_1G_2)$$

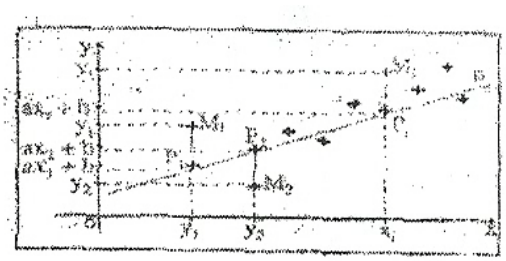
Définition :

Le principe de l'ajustement par la méthode de Mayer consiste à partager le nuage associé à une série $(x_i\ , y_i)$ en deux nuages dont le nombre de points diffère d'au plus un. On désigne par G_1 et G_2 les points moyens respectifs du premier et du deuxième nuage, la droite (G_1G_2) est appelée **droite de Mayer** on a $G \in (G_1G_2)$

Méthode d'ajustement par les moindres carrés :

Définition :

Le principe de l'ajustement par la méthode des **moindres carrés** consiste à déterminer les réels a et b tels que la somme $\sum_{i=1}^{n}(M_iH_i)^2$ soit minimale avec $M_i(x_i\ , y_i)$, $1 \leq i \leq n$



Le nuage de points d'une série statistique double, ainsi $D : y = ax + b$ et $H_i(x_i , y_i)$ le point de la droite D de même abscisse que M_i . **On admet qu'une telle droite existe et qu'elle est unique. On l'appelle droite de régression de y en x .**

Théorème :

La droite de régression de y en x dans un repère orthogonal associée à la série statistique double (X , Y) est la droite qui passe par le point moyen $G(\bar{X} , \bar{Y})$ et de coefficient directeur le réel $\underbrace{a = \frac{cov(X ,Y)}{V(X)}}_{\boxed{Alpha\ b}}$

Définition :

Soit (X , Y) une série statistique double sur un échantillon de taille n

1)La droite d'équation $y = \frac{cov(X,Y)}{V(X)} . (x - \bar{X}) + \bar{Y}$ est appelée droite des moindres carrés de Y en X, ou droite de régression de Y en X.

$$y = \frac{cov(X,Y)}{V(X)} . (x - \bar{X}) + \bar{Y} = a\,x + b$$

$$\underbrace{a = \frac{cov(X ,Y)}{V(X)}}_{\boxed{Alpha\ )}} \quad et \quad \underbrace{b = \bar{Y} - a\bar{X}}_{\boxed{Alpha\ (}}$$

Très important : dans les calculatrices Sharp a et b sont inverser

2)La droite d'équation $x = \frac{cov(X,Y)}{V(Y)} . (y - \bar{Y}) + \bar{X}$ est appelé droite des moindres carrées de X en Y, ou droite de régression de X en Y.

1)Coefficient de corrélation linéaire :

On peut toujours au vu des formules précédentes construire une droite de régression. Mais parfois cette dernière n'est d'aucune efficacité G, dans la mesure où les prédictions que l'on fait à partir de cette droite ne sont pas raisonnables. C'est le cas lorsqu'il n'existe pas réellement de corrélation entre les deux variables. Pour savoir si a est pertinent d'ajuster un nuage de point par les moindres carrés, on calcule un réel appelé coefficient de corrélation linéaire

Définition :

Soit (X , Y) une série statistique double. On appelle coefficient de corrélation linéaire le réel noté r(X , Y) défini par : $r(X,Y) = \frac{cov(X,Y)}{\sigma_X . \sigma_Y}$

Remarque :

1)On a : $-1 \leq r(X,Y) \leq 1$
2)Si $|r(X,Y)| = 1$ alors il y a une dépendance totale, l'une est une fonction affine de l'autre.
3)Si $|r(X,Y)| \in [0 ; 0{,}7]$ alors la corrélation entre X et Y est faible.
4)Si $|r(X,Y)| \in]0{,}7 ; 0{,}95]$ alors la corrélation entre X et Y est forte.
5)Si $|r(X,Y)| \in]0{,}95 ; 1]$ alors la corrélation entre X et Y est très forte.

4)Exemples d'ajustements non affines d'une série double :

(ajustement logarithmique)

Le tableau ci-dessous donne la production de pétrole de 1987 à 1997 suivant L'OPEP , x : le rang de l'année et y : la production (en millions de tonnes). On pose $X = \ln(x)$, les valeurs arrondies à 10^{-2} près

Année	1987	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997

Y : production	944	1065	1137	1232	1231	1297	1332	1333	1368	1408	1423
Rang : X	1	2	3	4	5	6	7	8	9	10	11
$X = \ln(x_i)$	0	0.69	1.1	1.39	1.61	1.79	1.95	2.08	2.2	2.3	2.4

1)a- Donner une équation de la droite de régression de Y en X de la série double (X , Y) sous la forme $y = \alpha x + \beta$ avec α et β arrondis à l'unité
b- En déduire une relation entre x et la production y : $y = f(x)$
2)a- Dans un même repère orthogonal, placer le nuage de points $M_i(x_i , y_i)$ et représenter la fonction f définie sur $[1 , +\infty[$
b- A l'aide de cet ajustement, donner une estimation de la production de pétrole en 2015, si cette politique se poursuit

(ajustement exponentiel)

Le tableau suivant donne l'effet de la pollution sur la population piscicole d'une rivière de 2006 et 2011Soit un repère orthogonal, on pose $Z = \ln(Y)$, les valeurs arrondies de Z à 10^{-2} près

Année	2006	2007	2008	2009	2010	2011
X : (Rang)	1	2	3	4	5	6
Y : (Nombre de poissons)	951.3	106.7	96.5	63.2	21	9.4
$Z = \ln(Y)$	6.86	4.67	4.57	4.15	3	2.24

1)Représenter le nuage de points $M_i(x_i , \ln(y_1))$, dans ce repère
2)a) Calculer le coefficient de corrélation de (X , Z) et justifier que l'on peut procéder à un ajustement affine par les moindres carrés.
b)Donner une équation de la droite de régression de Z en X, sous la forme $Z = \alpha X + \beta$, en arrondissant α et β au centième
c)En déduire en utilisant l'égalité $Z = \ln(Y)$, un ajustement exponentiel de Y en X sous la forme $Y = A.e^{B.X}$
3)On suppose que l'évolution de cette population se poursuit sur le même modèle
a)A partir de quelle année cette population sera-t-elle inférieur à 1000 ?
b)Donner une estimation de la population de cette rivière en l'an 2014 ?