Leçon n°8 : Séries statistiques à deux variables numériques. Nuage de points associé. Ajustement affine par la méthode des moindres carrés. Droite de régression. Applications. L'exposé pourra être illustré par un ou des exemples faisant appel à l'utilisation d'une calculatrice.

Rédigée par Cécile COURTOIS, le 20 août 2003.

Prérequis :

- Séries statistiques à une variable.
- L'inégalité de Cauchy-Schwarz

I Séries statistiques à deux variables.

I.1 Présentation du problème.

On a regroupé dans un tableau le temps passé par cinq élèves à travailler l'oral du CAPES lors de la dernière semaine et la note qu'ils ont obtenu.

Temps de travail x_i en heures	43	53	55	61	67
Note y_i sur 200	163	169	180	190	187

Il est intéressant de savoir s'il y a une relation de dépendance, autrement dit une corrélation entre ces deux caractères.

Dans ce cas, il sera encore plus intéressant de quantifier cette corrélation.

Ce sera l'objet de cette leçon.

Remarques :

a) Le jour de l'oral, il suffit de choisir des valeurs quelconques mais non aberrantes et notamment deux valeurs montrant que ce n'est pas le candidat qui a le plus travaillé qui a la meilleure note.

b) Au niveau du temps et de la présentation, il vaut mieux se contenter de recopier seulement le tableau et de donner les explications à l'oral, en se tournant vers le jury : c'est le début de la leçon et ils comprennent d'entrée que vous vous souciez d'eux.

I.2 Définitions.

Définition 1-1

Soit $n \geq 1$. Soient $(x_i)_{1 \leq i \leq n}$ et $(y_i)_{1 \leq i \leq n}$ deux séries statistiques à une variable. On dit alors que $(x_i, y_i)_{1 \leq i \leq n}$ est une série statistique à deux variables numériques.

Notations :

- On note $\bar{x}$ la moyenne de $(x_i)_{1 \leq i \leq n}$ et on a $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$.
- On note S_x l'écart type de $(x_i)_{1 \leq i \leq n}$ et on a $S_x = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$

Dans la suite de la leçon, on considère un entier naturel n non nul et une série statistique à deux variables $(x_i, y_i)_{1 \le i \le n}$.

Définition 1-2

Dans un repère orthogonal, on appelle nuage de points associé à la série $(x_i, y_i)_{1 \le i \le n}$ l'ensemble des points $M_i(x_i,y_i)$ pour i variant de 1 à n.

On appelle point moyen du nuage de points associé à cette série le point G de coordonnées $(\bar{x} ; \bar{y})$.

Remarque orale : G est l'isobarycentre des points du nuage.

On considère, dans la suite de la leçon, un repère orthogonal.

Définition 1-3

On appelle covariance de la série $(x_i, y_i)_{1 \le i \le n}$ le réel noté C_{xy} ou cov(x ;y) et défini par : $C_{xy}= \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$.

Remarque orale : La covariance permet de mesurer la dispersion des points du nuage autour du point moyen.

Proposition 1-4

(i) $C_{xy} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\bar{y}$.

(ii) $|C_{xy}| \le S_x S_y$ **avec égalité si et seulement si les points M_i du nuage sont alignés.**

Preuve :

(i) $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ donc $\sum_{i=1}^{n} x_i = n\bar{x}$.

De même, $\sum_{i=1}^{n} y_i = n\bar{y}$.

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

$$C_{xy} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\bar{y}.$$

(ii) $C_{xy}= \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$

D'après l'inégalité de Cauchy-Schwarz, on a :

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

D'où, en multipliant cette inégalité par $\frac{1}{n^2}$, on a : $C_{xy}^2 \leq \left[\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right]\left[\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2\right]$

ou encore : $C_{xy}^2 \leq S_x^2 S_y^2$.

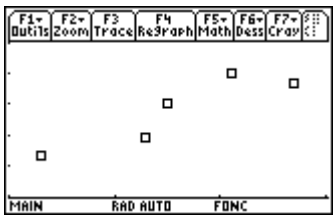
Donc $|C_{xy}| \leq S_x S_y$ puisque S_x et S_y sont par définition des réels positifs

.

On a l'égalité dans l'inégalité de Cauchy Schwarz si et seulement si il existe un couple réels $(\alpha ; \beta)$ tel que pour tout i compris entre 1 et n, $\alpha(x_i - \bar{x}) + \beta(y_i - \bar{y}) = 0$ c'est à dire si et seulement si les points du nuage sont alignés.

Exemple :

En considérant la série statistique définie dans le paragraphe I-1, on a C_{xy} = 75,76 et on peut tracer le nuage de points à l'aide de la calculatrice graphique :



II Ajustement affine par la méthode des moindres carrés.

Préambule :

Parfois, le nuage de points associé à une série statistique à deux variables a une forme « allongée » : il semble qu'on peut tracer une droite (et même plusieurs) autour de laquelle sont situés les points du nuage.

On dit que **chacune de ces droites réalise un ajustement affine du nuage**.

Il convient alors de se demander si une droite est « meilleure » qu'une autre et si oui, selon quel critère. C'est l'objet de ce paragraphe.

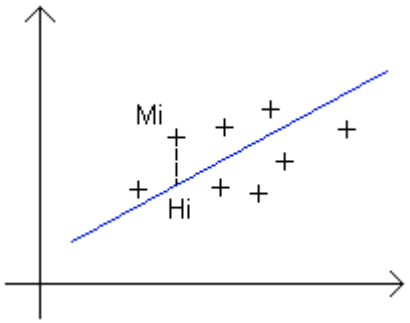
Remarque orale :

Si les points du nuage sont alignés, la réponse aux questions est triviale. On considère donc par la suite que les points du nuage ne sont pas alignés, c'est à dire que les valeurs des séries $(x_i)_{1 \leq i \leq n}$ et $(y_i)_{1 \leq i \leq n}$ sont distinctes deux à deux.

Remarque : Il faut expliquer tout ceci oralement, en se tournant vers le jury.

Principe de la méthode des moindres carrés :

La méthode des moindres carrés consiste à chercher s'il existe une droite Δ, et si oui, en déterminer une équation, réalisant un ajustement affine du nuage et minimisant la somme des carrés des distances M_iH_i (ce qui justifie la terminologie utilisée) où, pour i variant de 1 à n, Hi est le projeté du point M_i sur la droite Δ parallèlement à l'axe des ordonnées.



Cette recherche aboutit au résultat suivant :

Théorème- Définition 2-1

Il existe une unique droite Δ réalisant un ajustement affine du nuage de points $M_i(x_i ; y_i)$ par la méthode des moindres carrés. Cette droite est appelée droite de régression de y en x et a pour équation $y = a(x - \bar{x}) + \bar{y}$ où $a = \frac{C_{xy}}{S_x^2}$.

De plus, Δ passe par le point moyen G du nuage.

Preuve :

Soit Δ une droite d'équation y = ax + b où a et b sont deux réels. On désigne, pour i allant de 1 à n, par H_i le projeté de M_i parallèlement à (Oy) sur Δ.

$\sum_{i=1}^{n} M_iH_i^2 = \sum_{i=1}^{n} (y_i - ax_i - b)^2$ car la distance M_iH_i ne dépend que des ordonnées des points.

$$\sum_{i=1}^{n} M_iH_i^2 = \sum_{i=1}^{n} (y_i - ax_i)^2 - 2b\sum_{i=1}^{n} (y_i - ax_i) + nb^2$$

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

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

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

$$= n\left[S_y^2 - 2aC_{xy} + a^2S_x^2 + (b - \bar{y} + a\bar{x})^2\right]$$

$$= n\left[(b - \bar{y} + a\bar{x})^2 + \left(aS_x - \frac{C_{xy}}{S_x}\right)^2 + \frac{S_y^2S_x^2 - C_{xy}^2}{S_x^2}\right]$$

avec S_x et S_y non nul car les x_i et les y_i sont deux à deux distincts.

$\frac{S_y^2S_x^2 - C_{xy}^2}{S_x^2}$ est un nombre positif indépendant de a et de b.

Donc $\sum_{i=1}^{n} M_iH_i^2 \geq \frac{S_y^2S_x^2 - C_{xy}^2}{S_x^2}$ et on a l'égalité si et seulement si :

$$\begin{cases} b - \bar{y} + a\bar{x} = 0 \\ aS_x - \frac{C_{xy}}{S_x} = 0 \end{cases} \text{ ie } \begin{cases} b = \bar{y} - a\bar{x} \\ a = \frac{C_{xy}}{S_x^2} \end{cases}.$$

$G \in \Delta$ (vérification facile)

Remarques :

On peut définir la droite Δ' de régression de x en y : elle a pour équation, $x - \bar{x} = a'(y - \bar{y})$ où $a' = \frac{C_{xy}}{S_y^2}$.

La décision d'ajuster un nuage par une droite se prend jusqu'à présent à la seule vue du nuage de points.

Les statisticiens ont éprouvé le besoin de quantifier cette prise de décision, d'où la définition suivante.

Définition 2-2

On appelle coefficient de corrélation linéaire entre $(x_i)_{1\le i \le n}$ et $(y_i)_{1\le i \le n}$ le nombre réel noté r défini par $r = \frac{C_{xy}}{S_x S_y}$.

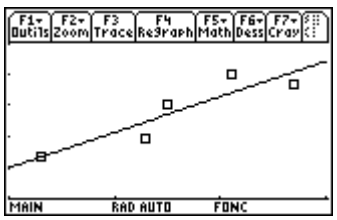
Remarques :

(i) $-1 \le r \le 1$. r est égal à 1 ou –1 si et seulement si les points du nuage sont alignés.

(ii) On dit que la corrélation entre $(x_i)_{1\le i \le n}$ et $(y_i)_{1\le i \le n}$ est très forte lorsque $|r| \ge \frac{\sqrt{3}}{2}$ et dans ce cas, on estime que le nuage de points est suffisamment allongé pour mettre en œuvre la méthode des moindres carrés.

Exemple :

En considérant la série statistique définie dans le paragraphe I-1, on a $r \approx 0{,}9$, ce qui justifie la mise en œuvre de la méthode des moindres carrés. La calculatrice graphique en permet un tracé rapide :

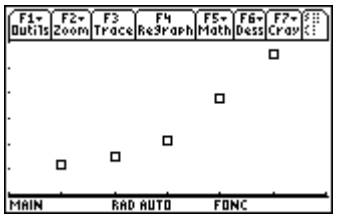


III Application.

Le déficit budgétaire en pourcentage du PIB de 1990 à 1994 est représenté dans le tableau ci-dessous.

Année x_i	1990	1991	1992	1993	1994
Déficit en % y_i	1,2	1,5	2,1	3,8	5,5

On représente rapidement le nuage de point associé à l'aide de la calculatrice :

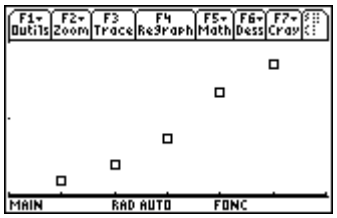


Il semble qu'on peut tracer une courbe ayant l'allure de celle de la fonction exponentielle et passant par « presque » tous les points.

On introduit donc la série statistique $(z_i)_{1\leq i \leq 5}$, définie, pour i allant de 1 à 5, par $z_i = \ln y_i$. On obtient les valeurs suivantes :

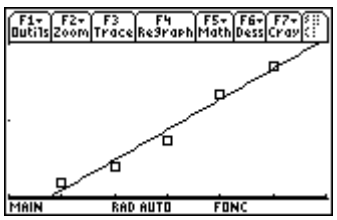
Année x_i	1990	1991	1992	1993	1994
$z_i = \ln y_i$	0,182	0,405	0,742	1,335	1,705

et le nuage de points suivant :

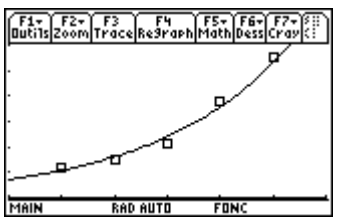


Pour cette nouvelle série statistique, on trouve $r \approx 0,9871$ ce qui justifie un ajustement affine.

La droite de régression de z en x a pour équation $z = a(x - 1992) + 0,873$ avec $a \approx 0,3974$.



On revient maintenant à la série statistique initiale. On a donc $y = e^{a(x-1992)+0,873}$, c'est-à-dire l'équation d'une fonction exponentielle : on a réalisé un ajustement exponentiel.

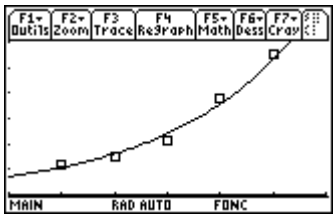


<u>Remarques :</u>

(i) Le jour de l'oral, choisissez une série statistique double issue de la vie économique et sociale pour montrer que cette leçon est destinée à la série ES. Pour en choisir les valeurs, il suffit de tracer le nuage de points de telle sorte qu'on puisse tracer une courbe exponentielle proche de ces points puis de lire les valeurs prises par la série que vous venez de définir.

Il se peut toutefois que par cette méthode, la théorie ne permet pas de justifier un ajustement affine sur la série $(x_i, z_i)_{1 \leq i \leq n}$. Dans ce cas, pas de panique, recommencez les calculs sur la série initiale et il est alors très probable que l'ajustement affine soit possible directement.

(ii) Certaines calculatrices (telles que la TI-89) réalise directement les ajustements exponentiels. Je vous donne la courbe obtenue, tracée sur le même graphique que la précédente :



Les courbes, sont à l'œil nu, confondues.

IV Compléments.

Voici quelques pistes pour l'entretien suivant l'exposé :

a) Il existe d'autres ajustements, selon la forme du nuage de points : l'ajustement logarithmique, l'ajustement par un polynôme de degré 2, 3 ou 4, l'ajustement par la fonction sinus. Je vous laisse deviner, par leur nom, à quoi ils correspondent…

b) Il existe également plusieurs méthodes d'ajustement affines, qui peuvent en particulier avoir leur place dans le dossier concernant les séries statistiques à deux variables.

L'ajustement affine par la droite de Mayer consiste à séparer le nuage de points en deux nuages de taille équivalente, à calculer les coordonnées des points moyens G_1 et G_2 de chaque nuage et à tracer la droite passant par G_1 et G_2.

L'ajustement par la droite des « extrêmes » consiste à ajuster le nuage par la droite (M_1M_n).

On peut demander aux élèves de Terminale ES, à l'aide d'un tableur, de comparer ces méthodes en calculant la somme des résidus (ie la somme des $M_iH_i^2$) et il apparaîtra, comme énoncé précédemment, que la droite de régression de y en x minimisera cette somme.

c) Le coefficient de corrélation linéaire n'est plus au programme de la classe de Terminale ES dans les nouveaux programmes de 2002. Désormais, les élèves prennent la décision d'ajuster un nuage de points à la seule vue de ce nuage. En effet, l'interprétation du coefficient de corrélation linéaire est jugée trop délicate en Terminale ES.

d) Lorsque C_{xy} est positif, les deux séries statistiques séries $(x_i)_{1\leq i \leq n}$ et $(y_i)_{1\leq i \leq n}$ sont classées par ordre croissant et lorsque C_{xy} est négatif, la série $(x_i)_{1\leq i \leq n}$ est classée par ordre croissant et la série $(y_i)_{1\leq i \leq n}$ est classée par ordre décroissant (simple constatations sur le graphique).

e) Il se peut que la théorie (ie le calcul du coefficient de corrélation linéaire) montre qu'il existe une très forte corrélation entre les deux séries étudiées mais les apparences sont parfois trompeuses. L'exemple classique est celui de l'enquête réalisée en Angleterre de 1924 à 1937, révélant que le coefficient de corrélation linéaire entre le nombre de permis délivrés chaque année pour l'installation d'un poste de radio et le nombre de malades mentaux dénombrés sur 10000 habitants était égal à 0,998, suggérant ainsi une relation quasiment fonctionnelle. Cela s'explique par le fait que ces deux séries sont corrélées à la série formée par les années durant lesquelles l'étude a été faite.

f) Quel est l'intérêt de réaliser un ajustement affine ? Il est double. Dans certains cas, il va permettre d'extrapoler : on va par exemple être en mesure de faire des prévisions lorsque qu'une des deux variables représente le temps. Inversement, l'ajustement peut permettre d'interpoler : si une année manque dans les données, mais est comprise entre deux dates données connues, on va pouvoir « deviner » la valeur prise par la deuxième variable.

g) *Lien avec les probabilités* : on considère les deux variables aléatoires X et Y prenant respectivement comme valeurs les valeurs des deux séries statistiques $(x_i)_{1\leq i \leq n}$ et $(y_i)_{1\leq i \leq n}$. On suppose de plus que pour i variant de 1 à n, $P(X = x_i)= \frac{1}{n} = P(Y = y_i)$.

Alors $E(X) = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}$ et $E(Y) = \bar{y}$.

On appelle alors <u>covariance de X et de Y</u> le nombre réel cov(X ;Y) défini par :

cov(X ;Y) = E((X-E(X))(Y-E(Y))

On a : cov(X ;Y) = E(XY) –E(X)E(Y) par linéarité de l'espérance

$\text{cov}(X ;Y) = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\bar{y} = C_{xy}$ ce qui justifie la terminologie employée.