Sylvain ETIENNE
PLC1, groupe 1
etiennesy@wanadoo.fr

SERIES STATISTIQUES A DEUX VARIABLES NUMERIQUES. NUAGE DE POINTS ASSOCIE. AJUSTEMENT AFFINE PAR LA METHODE DES MOINDRES CARRES. DROITE DE REGRESSION. APPLICATIONS. L'EXPOSE POURRA ETRE ILLUSTRE PAR UN OU DES EXEMPLES FAISANT APPEL A LA CALCULATRICE.

Niveau : Complémentaire.

Pré-requis : Séries statistiques à une variable – Inégalité de Cauchy-Schwarz – Fonction exponentielle –

I INTRODUCTION.

Les études statistiques nous permettent, en général, d'analyser et de prévoir une tendance. Le but de cet exposé est de déterminer s'il existe un lien de dépendance entre deux caractères que nous étudions simultanément, ou d'un caractère que nous étudions à différentes dates. Nous allons donc, dans un premier temps, définir des séries statistiques à deux variables. Puis, nous étudierons la possibilité de faire un ajustement affine.

II SERIES STATISTIQUES A DEUX VARIABLES.

A) DEFINITIONS.

Définition 1 :

> Soit un entier $n \geq 1$. Soit deux séries statistiques à une variable $(x_i)_{1\leq i\leq n}$ et $(y_i)_{1\leq i\leq n}$. Alors, une série statistique à deux variables numériques est l'ensemble des couples $(x_i, y_i)_{1\leq i\leq n}$. La pondération de chaque couple $(x_i, y_i)_{1\leq i\leq n}$ est égale à 1.

Dans la suite de l'exposé, *n* désigne un entier supérieur ou égal à 1 et nous considérons une série statistique à deux variables $(x_i, y_i)_{1\leq i\leq n}$.

Exemple 1 :

Le tableau suivant donne le PNB ainsi que le nombre d'hôpitaux pour 1 million d'habitants dans quelques pays européens.

Pays	A	B	C	D	E	F	G	H
PNB, *x*, en euros par habitants	5 100	7 800	11 200	15 800	20 100	26 230	28 910	31 910
Nombre *y* d'hôpitaux par million d'habitants	620	1 080	1 550	2 100	3 000	3 800	4 200	4 400

L'ensemble des couples $(x_i, y_i)_{1\leq i\leq n}$ définie une série statistique à deux variables.

Définition 2 :

Le plan est muni d'un repère orthogonal. A chaque couple (x_i, y_i), nous associons le point M_i de coordonnées (x_i, y_i). L'ensemble des points M_i, $i \in [\![1, n]\!]$ est le nuage de points associé à la série statistique double (x_i, y_i).

Nous appelons point moyen du nuage le point G de coordonnées $(\overline{x}, \overline{y})$ avec $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ et $\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$.

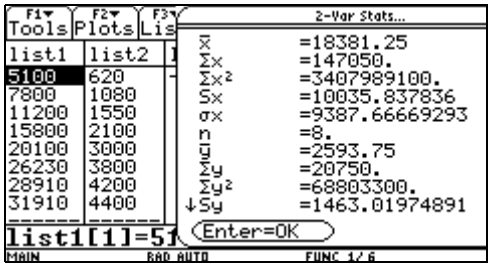
Remarque :

- G est l'isobarycentre des points du nuage.

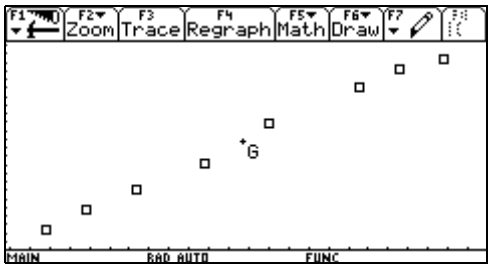
Nous nous plaçons, dans la suite de l'exposé, dans un repère orthogonal $(O, \vec{i}, \vec{j})$.

Revenons à l'exemple 1 :

Rentrons les données sur une T.I. Voyage 200. Nous pouvons en plus lui demander de nous sortir les données des séries à une variables x et y.
Elle nous donne :
$\overline{x} = 18381,25$ et $\overline{y} = 2593,75$.



Puis, créons un nuage de point (« scatter » en anglais !) en prenant pour x « list1 », et pour y « list2 ».
Nous avons de plus affiché le point G, isobarycentre des points du nuage.



Nous voudrions pouvoir mesurer la dispersion des points du nuage autour du point G. D'où la définition suivante :

Définition 3 :

Nous appelons covariance de la série statistique double (x_i, y_i) le nombre réel, noté $Cov(x, y)$, défini par :

$$Cov(x, y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}).$$

Revenons à l'exemple 1 :

Nous trouvons pour la covariance de l'exemple 1 :
$Cov(x, y) = 12814382,8125$.

B) PROPOSITION.

Proposition 1 :

-i- Théorème de Huyghens-König : $Cov(x,y)=\frac{1}{n}\sum_{i=1}^{n}x_i y_i-\bar{x}\,\bar{y}$.

-ii- Soit σ_x (resp. σ_y) l'écart-type de $(x_i)_{1\le i\le n}$ (resp. $(y_i)_{1\le i\le n}$), d'où $\sigma_x^2=\sum_{i=1}^{n}(x_i-\bar{x})^2$ (resp. $\sigma_y^2=\sum_{i=1}^{n}(y_i-\bar{y})^2$). Alors $|Cov(x,y)|\le\sigma_x\sigma_y$ et l'égalité a lieu si et seulement si les points du nuage sont alignés.

Démonstration :

-i- Par définition, nous avons : $\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i$ et $\bar{y}=\frac{1}{n}\sum_{i=1}^{n}y_i$. D'où,

$$\begin{aligned}Cov(x,y)&=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}),\\&=\frac{1}{n}\sum_{i=1}^{n}(x_iy_i-x_i\bar{y}-\bar{x}\,y_i+\bar{x}\,\bar{y}),\\&=\frac{1}{n}\left(\sum_{i=1}^{n}x_iy_i\right)-n\bar{x}\,\bar{y}-n\bar{x}\,\bar{y}+n\bar{x}\,\bar{y},\\Cov(x,y)&=\frac{1}{n}\sum_{i=1}^{n}(x_iy_i-\bar{x}\,\bar{y}).\end{aligned}$$

-ii- Appliquons l'inégalité de Cauchy-Schwarz à $Cov(x,y)^2$. Nous obtenons alors :

$$\begin{aligned}Cov(x,y)^2&=\frac{1}{n^2}\left(\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})\right)^2,\\&\le\frac{1}{n^2}\left(\sum_{i=1}^{n}(x_i-\bar{x})^2\right)\left(\sum_{i=1}^{n}(y_i-\bar{y})^2\right),\\&\le\left(\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2\right)\left(\frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2\right),\end{aligned}$$

i.e. : $Cov(x,y)^2\le\sigma_x^2\sigma_y^2$.

Or, par définition, l'écart-type est positif, donc : $|Cov(x,y)|\le\sigma_x\sigma_y$.

De plus, l'égalité de Cauchy-Schwarz a lieu si et seulement si $\exists(\alpha,\beta)\in\mathbb{R}^2$ tel que $\forall i\in[\![1,n]\!]$, $\alpha(x_i-\bar{x})+\beta(y_i-\bar{y})=0$, donc si et seulement si les points du nuage sont alignés. □

Sylvain ETIENNE
PLC1, groupe 1
etiennesy@wanadoo.fr

III AJUSTEMENT AFFINE PAR LA METHODE DES MOINDRES CARRES.

A) INTRODUCTION.

Nous cherchons une fonction *f* dont la courbe représentative passe « au plus près » des points du nuage. C'est le problème de l'ajustement.

Pour l'exemple 1, la forme « allongée » du nuage de points permet de penser qu'une droite convient pour ajuster le nuage. Nous parlons alors d'ajustement affine ou linéaire.

Il existe, cependant, d'autres ajustements : exponentiel, logarithme, polynomial…

Un tel ajustement permet alors de réaliser des estimations :

- Par interpolation, dans l'intervalle connu,
- Par extrapolation, au delà de cet intervalle.

Nous allons étudier, à présent, une méthode dans le cas de l'ajustement affine : la méthode des moindres carrés.

B) PRINCIPE.

Nous considérons un nuage de points $\left(M_i\left(x_i, y_i\right)\right)_{1\le i\le n}$. Soit (*D*) une droite d'équation $y = a\,x + b$.

Définition 4 :

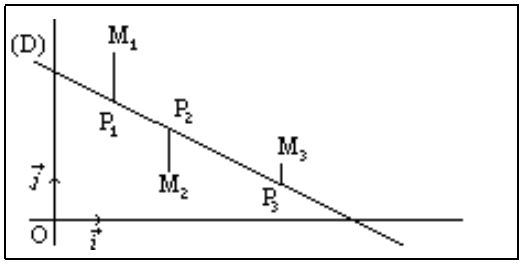
Nous appelons somme des résidus associée à la droite (*D*) le nombre réel *S* défini par :

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

Si P_i désigne le point d'abscisse x_i sur la droite (*D*), nous avons : $S = \sum_{i=1}^{n} M_i P_i^2$.

Définition 5 :

Nous appelons méthode des moindres carrés la méthode qui consiste à rechercher les coefficients *a* et *b* tels que la somme *S* soit minimale. Remarquons que *S* est une fonction des deux variables *a* et *b*.



C) DETERMINATION DES COEFFICIENTS.

Supposons *a* fixé et considérons *S* comme un polynôme du second degré en *b*. Il vient alors : $S = n b^2 - 2b\sum_{i=1}^{n}\left(y_i - a\,x_i\right) + \sum_{i=1}^{n} y_i^2$. Or, tout polynôme du second degré $\alpha x^2 + \beta x + \gamma$,

avec $\alpha > 0$, est minimum lorsque $x = -\frac{\beta}{2\alpha}$. Alors S est minimum lorsque $b = \frac{1}{n}\sum_{i=1}^{n}(y_i - a\,x_i)$, que nous pouvons écrire $b = \frac{1}{n}\sum_{i=1}^{n} y_i - \frac{a}{n}\sum_{i=1}^{n} x_i$. Soit : $b = \overline{y} - a\,\overline{x}$.

La droite recherchée doit donc vérifier l'équation $y = a\,x + \overline{y} - a\,\overline{x}$, ou encore : $y - \overline{y} = a(x - \overline{x})$.

Ainsi, parmi toutes les droites de coefficient directeur donné *a*, celle qui rend *S* minimale est celle qui passe par le point moyen $G(\overline{x}, \overline{y})$.

C'est une condition nécessaire.

Cherchons à présent le coefficient directeur.

Nous ne considérons désormais que les droites (*D*) qui passent par le point *G*. Afin de simplifier les écritures, plaçons-nous dans le repère $(G, \vec{i}, \vec{j})$: les droites (*D*) ont alors pour équations $Y = a\,X$ avec les formules de changement de repère définies par : $\forall i \in [\![1, n]\!]$, $X_i = x_i - \overline{x}$, et $Y_i = y_i - \overline{y}$.

Relativement au repère $(G, \vec{i}, \vec{j})$, la somme des résidus vaut :

$$S = \sum_{i=1}^{n}(Y_i - a\,X_i)^2,$$

$$= \sum_{i=1}^{n} Y_i^2 - 2a\sum_{i=1}^{n} Y_i X_i + a^2 \sum_{i=1}^{n} X_i^2.$$

Ce polynôme du second degré en *a* est minimum si et seulement si, $a = \frac{\sum_{i=1}^{n} Y_i X_i}{\sum_{i=1}^{n} X_i^2}$. Et en revenant aux séries statistiques initiales, nous avons :

$$a = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2}.$$

En conclusion, nous pouvons donc énoncer le théorème suivant :

Théorème 1 :

La droite d'équation $y = a\,x + b$ qui rend minimale la somme des résidus est la droite :

- qui passe par le point moyen $G(\overline{x}, \overline{y})$,
- qui a pour coefficient directeur $a = \frac{Cov(x, y)}{\sigma_x^2}$.

Cette droite, unique, s'appelle droite de régression de *y* par rapport à *x*.

Démonstration :

L'étude faite ci-dessus nous a donné des conditions nécessaires. Ces conditions sont évidemment suffisantes. □

Remarque :

- Il est possible de définir la droite de régression de x en y : elle passe, elle aussi, par le point moyen et elle a pour coefficient directeur $a' = \dfrac{Cov(x,y)}{\sigma_y^2}$.

D) COEFFICIENT DE CORRELATION.

Cependant, il nous faudrait un outil nous permettant de décider d'ajuster un nuage de points par une droite. C'est l'objet de la définition suivante.

Définition 5 :

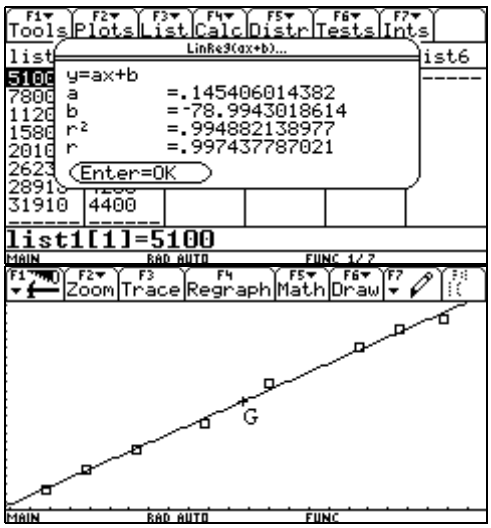
Nous appelons coefficient de corrélation linéaire le nombre réel, noté $r(x,y)$, tel que :

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

Remarques :

- $-1 \le r(x,y) \le 1$.
- $a\,a' = \left(r(x,y)\right)^2$.
- La corrélation entre $(x_i)_{1\le i\le n}$ et $(y_i)_{1\le i\le n}$ est dite forte si $r(x,y)^2 \ge \dfrac{3}{4}$, et dans ce cas, nous estimons qu'il est possible de mettre la méthode des moindres carrés en œuvre.

Retour à l'exemple 1 :
Demandons à la calculatrice de nous donner les informations relatives à une régression linéaire. Les résultats sont les suivants :
Remarque : après cette opération, elle nous donne même les résidus…



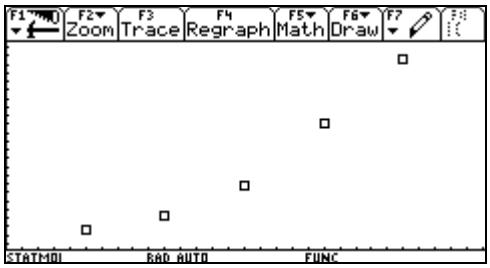
$r(x,y)^2 \approx 0{,}99 \ge \dfrac{3}{4}$, nous pouvons donc tracer la courbe correspondante :

IV APPLICATION.

Le tableau suivant donne l'évolution du nombre de passagers sur une ligne aérienne entre 1994 et 1998 :

Année	1994	1995	1996	1997	1998
Rang de l'année x_i	1	2	3	4	5
Nombre de passagers p_i	7 523	9 401	12 889	20 065	27 546

Représentons le nuage de points associé à l'aide de la calculatrice :

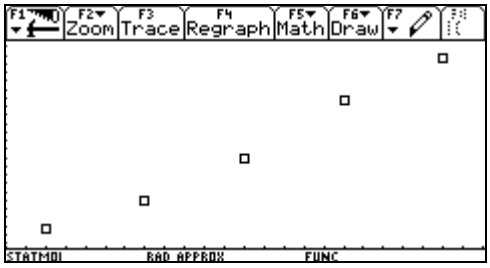


Il semble qu'une courbe exponentielle serait la meilleure courbe qui passerait « au plus près » des points du nuage.

Nous établissons alors une nouvelle série statistique définie par : $(q_i)_{1\le i\le n} = (\ln(p_i))_{1\le i\le n}$. Nous obtenons alors le tableau suivant :

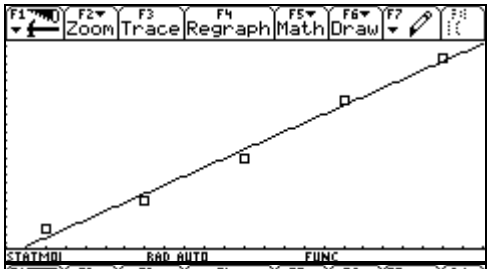
Rang de l'année x_i	1	2	3	4	5
$q_i = \ln(p_i)$	8,926	9,149	9,464	9,907	10,22

Représentons le nouveau nuage de points :

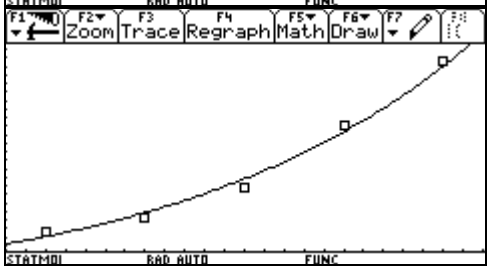


Nous trouvons un coefficient de corrélation $r(x,q)^2 \approx 0{,}9885$, ce qui justifie un ajustement affine.

Nous obtenons alors la droite de régression de q en x : $q = ax + b$ avec $a \approx 0{,}3354$ et $b \approx 8{,}5276$.



Revenons à la série initiale. Nous obtenons $p = e^{0{,}3354x+8.5276}$. Nous avons alors réalisé un ajustement exponentiel.



etiennesy@wanadoo.fr

V CONCLUSION.

L'étude des séries statistiques à deux variables permet de mettre en rapport deux caractères afin de pouvoir déterminer une valeur manquante ou de prévoir une tendance. Néanmoins, deux caractères peuvent avoir un très fort coefficient de corrélation sans pour autant être réellement lié. Un exemple est l'accroissement simultané des divorces dans les familles d'Italie du Nord et l'acquisition d'un ordinateur. La conclusion serait que les hommes préfèrent leur ordinateur à leur femme…