Exposé 8 : série statistique à deux variables numériques. Nuage de points associé. Ajustement affine par la méthode des moindres carrés. Droite de régression. Applications.

Prérequis[1] : -Série statistique à une variable numérique et ses éléments caractéristiques (variance, écart type, etc.)
-Equation de droite dans le plan affine euclidien
-Trinôme du second degré : forme canonique, minimum

Niveau : Terminale ES

Motivation : on envisage ici le cas où deux caractères à la fois sont observés (taille et âge d'un groupe d'enfants, superficie et rendement d'un ensemble de champs...). Il s'agit alors de séries statistiques doubles. Le problème qui se pose est alors de savoir s'il existe globalement une relation entre les variables statistiques ainsi définies. Si c'est le cas, le statisticien parle alors de corrélation entre les variables.

Contexte : nous étudions le cas de deux variables X et Y observés simultanément sur une même population de taille $n > 1$.

1 Série statistique à deux variables

1.1 Présentation du problème

Exemple : on sélectionne 10 personnes inscrites à un stage de formation. Avant le début de la formation, ces stagiaires subissent une épreuve A notée de 0 à 20. A l'issue du stage, une épreuve B identique à la première est notée aussi de 0 à 20. Les résultats sont rassemblés dans le tableau suivant. Quantifier, si elle existe, une corrélation entre les deux caractères.

Epreuve A	3	4	6	7	9	10	9	11	12	13
Epreuve B	8	9	10	13	15	14	13	16	13	19

1.2 Définition

Définition : soit $\Omega = \{w_1, ..., w_n\}$ une population de taille n. On dit que "deux variables" X et Y définissent sur Ω une série statistique double $(x_i, y_i)_{1 \leqslant i \leqslant n}$, avec $X(\omega) = x_i$ et $Y(\omega) = y_i$, lorsque :

1. $x_1 \leqslant x_2 \leqslant ... \leqslant x_n$ (ou $y_1 \leqslant y_2 \leqslant ... \leqslant y_n$, mais pas forcément les deux en même temps)
2. $X(\Omega)$ et $Y(\Omega)$ ne sont pas des singletons
3. les couples $(x_i, y_i)_{1 \leqslant i \leqslant n}$ sont deux à deux distincts

Conséquence : moyenne de $(x_i)_{1 \leqslant i \leqslant n}$: $\overline{x} = \dfrac{1}{n} \displaystyle\sum_{i=1}^{n} x_i$, variance de $(x_i)_{1 \leqslant i \leqslant n}$: $V(X) = \dfrac{1}{n} \displaystyle\sum_{i=1}^{n} (x_i - \overline{x})^2$

écart type de $(x_i)_{1 \leqslant i \leqslant n}$: $\sigma_x := \sqrt{V(X)} = \sqrt{\dfrac{1}{n} \displaystyle\sum_{i=1}^{n} (x_i - \overline{x})^2}$

Dans l'exemple : $\overline{x} = 8.4$, $\overline{y} = 13$, $\sigma_x = 3.1686$, $\sigma_y = 3.16228$ (calculatrice)

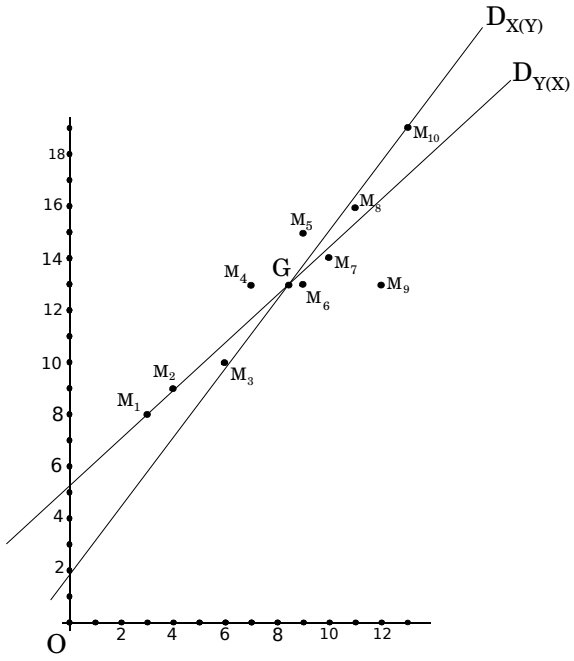
[1] Sources : Sabine, Blandine. Tapé par Gwendal Haudebourg, réalisé avec LaTeX. Mis à jour le 03/07/2007.

1.3 Représentation graphique : nuage de points

Définition : dans un repère othogonal $(O, \vec{i}, \vec{j})$, on appelle nuage de points de la série statistique $(x_i, y_i)_{1 \leqslant i \leqslant n}$ l'ensemble des points M_i de coordonnées (x_i, y_i). Le point G de coordonnées $(\overline{x}, \overline{y})$ est appelé point moyen du nuage.

remarque : G est l'isobarycentre du système de points $\{M_i\}_{1 \leqslant i \leqslant n}$, car $G = bar\{(M_1; 1), ..., (M_n; 1)\}$, donc $G(\frac{1}{n}\sum_{i=1}^{n} x_i; \frac{1}{n}\sum_{i=1}^{n} y_i) = (\overline{x}, \overline{y})$

Dans l'exemple : afficher le nuage de points associé à la calculatrice.



1.4 Paramètres d'une série statistique double

Définition : on appelle covariance du couple (X, Y) le réel noté C_{XY} ou σ_{XY}, ou encore $cov(X, Y)$:

$$C_{XY} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})$$

Propriétés :

1. $C_{XY} = \frac{1}{n}(\sum_{i=1}^{n} x_i y_i) - \overline{x}.\overline{y}$
2. $\forall (a, b, c, d) \in \mathbb{R}^4$, $cov(aX + b, cY + d) = a.cov(X, Y)$
3. $|C_{XY}| \leqslant \sigma_X.\sigma_Y$, avec égalité sisi les points $M_i(x_i, y_i)$ sont alignés

preuve : (3) par définition de la variance, $\forall \lambda \in \mathbb{R}$, $V(\lambda X + Y) \geqslant 0$. Or

$$V(\lambda X + Y) = \frac{1}{n}\sum_{i=1}^{n}(\lambda x_i + y_i - \lambda\overline{x} - \overline{y})^2 = \frac{1}{n}\sum_{i=1}^{n}[\lambda(x_i - \overline{x}) + (y_i - \overline{y})]^2 = \lambda^2 V(X) + 2\lambda C_{XY} + V(Y).$$

Comme $X(\Omega)$ n'est pas un singleton par hypothèse, $V(X)$ n'est jamais nul. Il s'agit donc d'un trinôme du second degré, positif ou nul pour tout réel λ, donc son discriminant est négatif ou nul (car ne change pas de signe), ie : $(C_{XY})^2 - V(X)V(Y) \leqslant 0$, d'où $|C_{XY}| \leqslant \sigma_X\sigma_Y$.
De plus, $(C_{XY})^2 = V(X)V(Y) \Leftrightarrow \Delta = 0 \Leftrightarrow \exists\lambda_0 \in \mathbb{R}$ tq. $V(\lambda_0 X + Y) = 0 \Leftrightarrow \exists\lambda_0$ tq.
$\frac{1}{n}\sum_{i=1}^{n}[\lambda_0(x_i - \overline{x}) + (y_i - \overline{y})]^2 = 0 \Leftrightarrow \forall i = 1...n,\ \lambda_0(x_i - \overline{x}) + (y_i - \overline{y}) = 0$, ce qui signifie que les points $M_i(x_i, y_i)$ appartiennent à la droite d'équation $\lambda_0(x - \overline{x}) + (y - \overline{y}) = 0$
Réciproquement, s'il existe une droite d'équation $y = ax + b$ telle que $\forall i = 1...n,\ y_i = ax_i + b$ (ie les points sont alignés), alors $\overline{y} = a\overline{x} + b$, et le calcul donne : $(C_{XY})^2 = a^2[V(X)]^2 = V(X)V(Y)$ □

Cela nous incite à introduire le coefficient de corrélation r_{XY}

1.5 Coefficient de corrélation linéaire

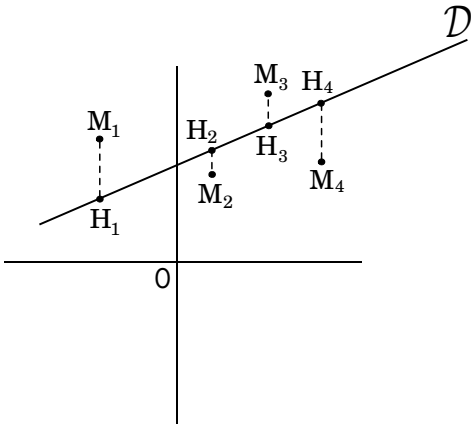
Selon notre hypothèse, σ_X et σ_Y ne sont pas nul. On peut donc poser : $r_{XY} = \frac{C_{XY}}{\sigma_X\sigma_Y}$. Ce coefficient r_{XY} est appelé le coefficient de corrélation linéaire des variables X et Y.

Proposition :
(i) $-1 \leqslant r_{XY} \leqslant 1$
(ii) Il y a égalité (ie r_{XY}=1 ou -1) sisi les points sont alignés.
(iii) le coefficient de corrélation r_{XY} est invariant par **transformation affine sur les variables**, ie : $r_{aX+b,cY+d} = r_{X,Y}$

2 Ajustement par la méthode des moindre carrés

Etant donné une série statistique à deux variables, existe t-il une "relation" entre les deux variables observées ? La réponse peut-être négative (on parlera alors de variables indépendantes). La réponse peut-être assez imprécise à priori : les deux variables augmentent en même temps, ou X augmente quand Y diminue. Nous allons voir en prenant appui sur l'analyse du nuage de points, comment établir une relation fonctionnelle entre les variables. Nous étudierons plus particulièrement le cas ou le nuage paraît se distribuer au voisinage d'une droite.

Principe de la méthode : le plan étant rapporté à un repère orthogonal $(O, \vec{i}, \vec{j})$, soit $(M_i)_{1\leqslant i\leqslant n}$ le nuage de points de coordonnées $(x_i, y_i)_{1\leqslant i\leqslant n}$. On cherche (si elle existe) une droite $D : y = ax + b$ qui ajuste l'ensemble des couples $(x_i, y_i)_{1\leqslant i\leqslant n}$, en minimisant la somme des carrés des distances M_iH_i, où $H_i = proj_{\|(Oy),D}(M_i)$ (donc $M_iH_i^2 = (y_i - ax_i - b)^2$). Autrement dit, on cherche des réels a, b pour lequel $\varphi(a,b) := \sum_{i=1}^{n}(y_i - ax_i - b)^2$ soit minimale.
Nous allons montrer qu'il existe une unique droite rendant minimale $\varphi(a,b)$. Remarquons que le minimum de $\varphi(a,b)$ sera d'autant plus petit que l'alignement des points M_i sera meilleur. A la limite, si les points M_i sont alignés, il existe une droite unique annulant $\varphi(a,b)$.



Théorème : Il existe une unique fonction affine $x \mapsto ax + b$ ajustant par la méthode des moindres carrés une série statistique double $(x_i, y_i)_{1 \leqslant i \leqslant}$. Ses coefficients sont donnés par les relations : $a = \dfrac{C_{XY}}{V(X)}$, $b = \overline{y} - a.\overline{x}$. De plus, cette droite passe par $G(\overline{x}, \overline{y})$.

preuve : $\varphi(a,b) = \sum_{i=1}^{n}(y_i - ax_i - b)^2 = nb^2 - 2b\sum_{i=1}^{n}(y_i - ax_i) + \sum_{i=1}^{n}(y_i - ax_i)^2$. Or $\sum_{i=1}^{n}(y_i - ax_i) = n(\overline{y} - a\overline{x})$,

donc $\varphi(a,b) = nb^2 - 2nb(\overline{y} - a\overline{x}) + \sum_{i=1}^{n}(y_i - ax_i)^2$

Ecrivons ce trinôme du second degré sous sa forme canonique[2] :

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

$$\varphi(a,b) = n[(b - (\overline{y} - a\overline{x}))^2 + (\frac{1}{n}\sum_{i=1}^{n}y_i^2 - \overline{y}^2) + a^2(\frac{1}{n}\sum_{i=1}^{n}x_i^2 - \overline{x}^2) - 2a(\frac{1}{n}\sum_{i=1}^{n}x_iy_i - \overline{x}.\overline{y})]$$

$$\varphi(a,b) = n[(b - (\overline{y} - a\overline{x}))^2 + a^2\sigma_X^2 - 2a.C_{XY} + \sigma_Y^2] = n[(b - (\overline{y} - a\overline{x}))^2 + (a\sigma_X - \frac{C_{XY}}{\sigma_X})^2 + \sigma_Y^2 - \frac{C_{XY}^2}{\sigma_X^2}]$$

$$\varphi(a,b) = n[(b - (\overline{y} - a\overline{x}))^2 + (a\sigma_X - \frac{C_{XY}}{\sigma_X})^2 + \frac{\sigma_X^2\sigma_Y^2 - C_{XY}^2}{\sigma_X^2}]$$

Il est clair que $\varphi(a,b)$ est minimal lorsque les deux premiers carrés sont nuls (le troisième terme ne dépendant pas de a et b), ce qui donne $a = \dfrac{C_{XY}}{\sigma_X^2}$ et $b = \overline{y} - a\overline{x}$. Le minimum de $\varphi(a,b)$ vaut donc

$$n\frac{\sigma_X^2\sigma_Y^2 - C_{XY}^2}{\sigma_X^2}$$

□

Définition : cette droite d'équation $y = ax + b$ ajustant le nuage de points par la méthode des moindres carrés, est appelée droite de régression (ou droite des moindres carrés) de Y par X, notée $D_{Y(X)}$.

Remarques :

1. $D_{Y(X)}$ passe par G car : $a.\overline{x} + b = a.\overline{x} + \overline{y} - a\overline{x} = \overline{y}$
2. On peut aussi chercher la droite de régression de X en Y (rôles inversés de X et Y), notée $D_{X(Y)}$ d'équation $x = a'.y + b'$ qui ajuste le nuage de points $(y_i, x_i)_{1 \leqslant i \leqslant n}$. En permutant X et Y, il vient : $a' = \frac{C_{XY}}{V(Y)}$ et $b' = \overline{x} - a'\overline{y}$

[2] on a besoin dans cette preuve de $\frac{1}{n}\sum_{i=1}^{n}x_i^2 - \overline{x}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$, cf. compléments

3. La droite $D_{X(Y)}$ passe aussi par le point moyen du nuage G
4. Si $C_{XY} = 0$ alors $a = a' = 0$. Les deux droites de régressions $D_{Y(X)}$ et $D_{X(Y)}$ sont respectivement paralèles à (Ox) et (Oy)
5. Les deux droites sont confondues sisi $a = \frac{1}{a'}$ (les deux droites passant par G, de coeff. directeur a et $\frac{1}{a'}$), ie lorsque les points sont alignés (les variables X et Y étant liés par une relation affine).
6. Le coefficient de corrélation linéaire permet d'estimer la corrélation entre les deux variables X et Y. Plus les points sont "étroitement alignés", plus la **valeur absolue** du coeff. e corrélation sera proche de 1. On admet que l'ajustement affine est pertinent si $|r_{XY}| \geqslant 0.75$

Dans l'exemple : $C_{XY} = 9$, $\sum xy = 1182$, $r_{XY} \backsimeq 0.898$. Tracer $D_{Y(X)}$ et $D_{X(Y)}$ à la calculatrice.

3 Applications

Dans certains cas, le nuage statistique laisse préssentir une relation fonctionnelle globale entre les caractères X et Y, mais il apparaît clairement que cette relation n'est pas affine.

3.1 Ajustement par une fonction exponentielle

Si les points $M_i(x_i, y_i)$ sont "proches" de la courbe définie par $y = \lambda a^x$, alors les points $P_i(x_i, \ln y)$ sont proches de la droite $y = \ln \lambda + x. \ln a$ (et réciproquement). La méthode consiste donc à construire le nuage de points $P_i(x_i, \ln y_i)$ et à chercher la droite de régression entre X et Y. Pour ce faire, on construit le nuage sur du papier gradué de façon arithmétique en abscisse et logarithmique en ordonnée (cela nous donne la droite cherchée).

Exemple : le tableau ci-dessous donne la production anuelle d'une usine de pâte à papier (en tonnes) en fonction de l'année :

Année	1996	1997	1998	1999	2000	2001	2002	2003
Production	325	351	382	432	478	538	708	930

On trace le nuage de points correspondant au tableau. Pour l'année i, on note p_i la production de pâte à papier et $l_i = \ln(p_i)$. On trace le nouveau nuage de points (i, l_i). En utilisant la calculatrice, on en déduit la droite d'ajustement des moindres carrés de l_i en i. On en déduit la fonction d'ajustement de la production en fonction de l'année. On peut ainsi prévoir la production des années suivantes :
$\overline{x} = 1999.5$, $\overline{y} = 6.189$, $V(x) = 5.25$, $C_{xy} = 0.749$, $a = 0.1428$, $b = -279,365$ pour année/ln(production) ; $a = 4.710$, $b = 1.153$ pour année/prod, où $y = a * b^x$ (avec $r = 0.968$, donc très bonne corrélation).

3.2 Ajustement par une fonction puissance

Si les points $M_i(x_i, y_i)$ sont "proches" de la courbe $y = \lambda x^a$, alors les points $P_i(\ln x_i, \ln y_i)$ sont proches de la droite $y = \ln \lambda + ax$ et réciproquement. Pour ce faire, on construit le nuage sur du papier gradué de façon logarithmique en abscisse et en ordonnée (cela nous donne la droite cherchée).

4 Compléments

4.1 Remarques

1. Les preuves importantes de l'exposé sont l'inégalité de C-S, et théorême existence-unicité de la droite de régression.
2. Si on présente la preuve (de C-S) comme ds cette exposé, ne pas mettre C-S dans les prérequis. Sinon faire la preuve en utilisant ce théorème (gagne du temps).
3. On a pas forcément $y_1 \leqslant y_2 \leqslant ... \leqslant y_n$ (cf. exemple).

4. Il existe d'autres méthodes d'ajustement affine d'une série statistique double. On peut par exemple distinguer deux sous-ensembles de la série statistique, et considérer la droite joignant les points moyen de chaque sous-nuage (méthode de Meyer). Dans ce cas, la droite obtenue dépend du découpage du nuage en deux sous nuages (la calculatrice propose un ajustement en trois parties, avec la fonction *med − med* : on obtient trois points M_1, M_2 et M_3 qui sont les médianes des valeurs de x et y, puis on trace la droite passant par le point moyen de ses trois points, parallèlement à M_1M_3).
5. Le théorême de Cauchy-Schwartz nous sert à montrer que $r_{XY} \in [-1, 1]$.
6. Si $r_{XY} \geqslant 0$: croissant (cf. exemple) ; si $r_{XY} \leqslant 0$: décroissant (ex : vitesse de la voiture, distance de freinage).
7. Le coefficient de corrélation n'est plus au programme en T-ES (et ce car vient du théorême de Cauchy-Schwartz).
8. Historiquement, on nomme la droite cherchée "droite de régression" car **Galton** étudie en 1877 la taille des enfants de sujets de grande taille par rapports à la taille des parents ; cet écart à tendance à régresser (diminuer) au court du temps vers la grandeur moyenne de la population. La droite décrivant cette relation fût ainsi dite "de régression".
9. Stats : corrélation entre deux caractères statistiques. Proba : 2 variables aléatoires sont-elles indépendantes ? (on va des stats vers les proba.) $C_{XY} = E(XY) - E(X)E(Y)$, et X, Y indépendantes sisi $C_{XY} = 0$, ie $E(XY) = E(X)E(Y)$.

4.2 preuves

preuve (conséquence) : on étudie ici des séries statistiques, la moyenne est notée $\overline{x}$ et la fréquence f_{x_i} (ce qui correspond en terme probalistique à respectivement l'espérance $E(X)$ et à la probabilitée $P(X = x_i)$). On considère que les éléments (x_i, y_i) ont tous le même poids, ie une fréquence de $\frac{1}{n}$, ie sous forme "probalistique" que $P(X = x_i) = \frac{1}{n} = f_{x_i}$, donc (avec des guillements dans les notations)

$E(X)=\sum_{i=1}^{n} x_i P(X = x_i)=\sum_{i=1}^{n} x_i f_{x_i}=\frac{1}{n}\sum_{i=1}^{n} x_i=\overline{x}$. De plus $E(X^2)=\sum_{i=1}^{n} x_i^2 f_{x_i}=\frac{1}{n}\sum_{i=1}^{n} x_i^2$

De même $V(X) = E[(X - E(X))^2]=\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$. Or $V(X) = E(X^2) - E(X)^2$, donc

$V(X)=\frac{1}{n}\sum_{i=1}^{n} x_i^2$-$(\frac{1}{n}\sum_{i=1}^{n} x_i)^2=\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \overline{x}^2$ □

preuve (Propriétés covariance) :

(1) simple calcul : $C_{XY} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})=\frac{1}{n}\sum_{i=1}^{n}(x_i y_i + \overline{xy} - \overline{x}y_i - x_i\overline{y})$

$=\frac{1}{n}\sum_{i=1}^{n} x_i y_i + \frac{n}{n}\overline{xy}$ - $\frac{\overline{y}}{n}\sum_{i=1}^{n} x_i$ - $\frac{\overline{x}}{n}\sum_{i=1}^{n} y_i=\frac{1}{n}\sum_{i=1}^{n} x_i y_i+\overline{xy}$-$\overline{y}\overline{x}$-$\overline{x}\overline{y}=\frac{1}{n}\sum_{i=1}^{n} x_i y_i$-$\overline{xy}$

(2) simple calcul : $cov(aX + b, cY + d)=\frac{1}{n}\sum_{i=1}^{n}((ax_i + b)(cy_i + d) - (a\overline{x} + b)(c\overline{y} + d))$=... □

preuve (coeff. de corrélation) :

(i) $|C_{XY}| \leqslant \sigma_X\sigma_Y$ donc $-1 \leqslant \frac{C_{XY}}{\sigma_X\sigma_Y} \leqslant 1$

(ii) $|C_{XY}| = \sigma_X\sigma_Y \Leftrightarrow$ les points sont alignés (d'après propriétés covariance).

(iii) $r_{aX+b,cy+d}=\frac{C_{aX+b,cY+d}}{\sigma_{aX+b}\sigma_{cX+d}}=\frac{acC_{XY}}{a\sigma_X\sigma_Y}=\frac{C_{XY}}{\sigma_X\sigma_Y}=r_{XY}$ □

4.3 Autre méthode pour le minimum de φ

$\varphi(a,b) = \sum_{i=1}^{n}(y_i - ax_i - b)^2$=$\varphi(a,b) = nb^2 - 2nb(\overline{y} - a\overline{x})+\sum_{i=1}^{n}(y_i - ax_i)^2$, donc : $\varphi(a,b)$=$\sum_{i=1}^{n}(y_i - ax_i - b)^2$

C'est un polynôme du second degré en b, donc φ admet un minimum. On a : $\frac{\delta\varphi}{\delta b}(a,b)$=$2n.b + 2(\overline{y} - a\overline{x})$

Donc $\frac{\delta\varphi}{\delta b}(a,b) = 0 \Leftrightarrow b = \overline{y} - a\overline{x}$.

Donc $\varphi(a,b) = \sum_{i=1}^{n}(y_i - ax_i - b)^2$=$\sum_{i=1}^{n}(y_i - ax_i - \overline{y} + a\overline{x})^2$=$\sum_{i=1}^{n}(-a(x_i - \overline{x}) + (y_i - \overline{y}))^2$

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

Donc $\frac{\delta\varphi}{\delta a}(a,b) = 0 \Leftrightarrow -2a.n.\sum_{i=1}^{n}(x_i - \overline{x})^2$+$2.n.C_{XY} \Leftrightarrow a = \frac{C_{XY}}{V(X)}$

4.4 Lien de causalité

Il est possible d'avoir des séries statistiques x et y où les points du nuage sont "presque" alignés, sans qu'il y ait un lien entre x et y (ex : x : nombre de hamburguers mangés à Moscou au cours de l'année 1986 ; y : nombre de joueurs de pétanques qui s'inscrivent au grand tournoi de Montélimard en 1950). Il est très simple de construire deux séries ainsi : il suffit de prendre des séries qui dépendent du temps, du type :

X tq. $X = \alpha.t + \beta$, Y tq. $Y = \alpha'.t + \beta'$, donc $t = \frac{X - \beta}{\alpha}$, donc $Y = \frac{\alpha'}{\alpha}X + (\beta - \frac{\alpha'.\beta}{\alpha})$, donc $Y = A.X + B$ (pour un certain couple $(A,B) \in \mathbb{R}^2$). Donc les deux séries statistiques sont bien liés par une relation de type affine (sans qu'il y ait forcément une relation de cause à effet entre X et Y).

4.5 Introduction de la droite des moindres carrés

Il est "de bon ton" d'introduire en T-ES la droite des moindres carrés via un tableur, pour bien faire sentir au élève ce qu'elle représente (et que "ça marche bien") : dans une colonne, on fait varier a, dans une autre on fait varier b, dans une troisième on étudie le réel $\varphi(a,b)$.

4.6 Changement de repère affine

Que se passe t-il si on fait un changement de repère ? Le droite des moindres carrés est t-elle changée ? Réponse : la pente a ne change pas, b varie, la droite passe toujours par $G(\overline{x},\overline{y})$ (qui lui varie)

Preuve : soit $X = x + C, Y = y + C'$ $a = \frac{C_{XY}}{V(X)}$. Or $V(X + C) = V(X)$ (revenir à la définition de $V(X)$), et $C_{X+C,Y+C'} = C_{XY}$. $b = \overline{y} - a.\overline{x}$, or $\overline{y}$ et $\overline{x}$ varient, donc b varie. Pour tracer la nouvelle droite : trouver $G(\overline{x},\overline{y})$, puis avec a connu, déterminer b.

4.7 Méthode de Meyer

On distingue deux sous-ensembles de la série statistique, et on considére la droite joignant les points moyens de chaque sous-nuage. Dans ce cas, la droite obtenue dépend du découpage du nuage en deux sous nuages.

Il est intéressant de noter que parfois, cette méthode est plus judicieuse que la méthode d'ajustement des moindres carrés (par une droite ou une fonction). Ex : prod. linéaire des papier dans une usine, avec un "trou" au milieu du nuage, causée par une grève ou une panne de machine. Au contraire, la méthode de Meyer peut parfois être très mauvaise. Ex : nuage en forme de haricot.