CORRELATION

Soient X et Y deux variables quantitatives concernant un même échantillon. Le problème est ici de savoir comment évaluer l'existence d'un lien entre elles ou, au contraire, leur indépendance.

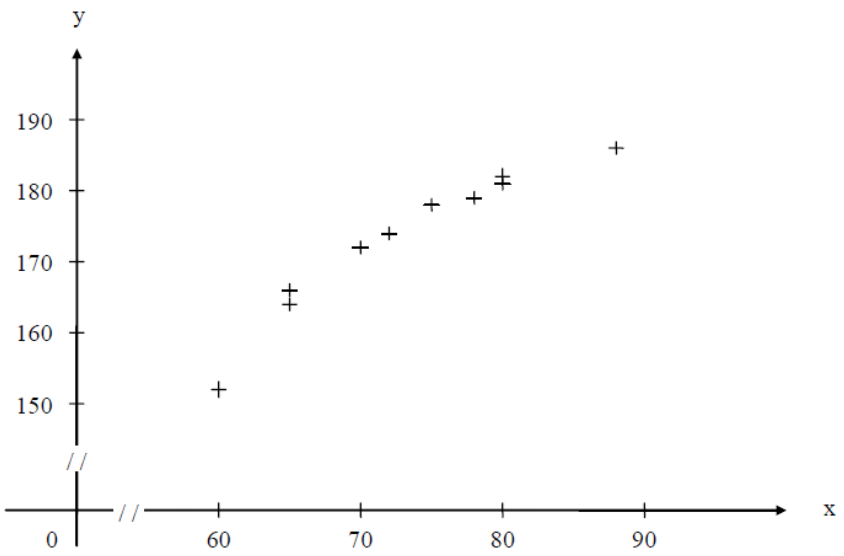
On considère les valeurs prises par ces variables dans un échantillon de n individus.
Les valeurs relevées forment n couples $(x_i\ ;\ y_i)$ de réels (pour $i \in [\![1;n]\!]$).
On obtient avec l'ensemble de ces couples une série statistique à deux variables (ou série statistique double).
Cette série peut être représentée graphiquement dans un repère orthogonal par l'ensemble des points M_i de coordonnées $(x_i\ ;\ y_i)$, appelé le **nuage de points** représentatif de la série statistique double.

Exemple : La direction d'un quotidien compare chaque mois, sur une durée de 10 mois :
- son investissement publicitaire, exprimé en k€ : variable X
- le nombre d'exemplaires vendus, exprimé en centaines de milliers d'unités : variable Y

Les résultats sont rassemblés dans le tableau suivant :

Mois	1	2	3	4	5	6	7	8	9	10
X : Invest. pub	60	65	65	70	75	72	80	88	78	80
Y : Ventes	153	164	166	172	178	174	181	186	179	182

Nuage de points :



Le nuage de points fournit une information qualitative quant au lien entre X et Y que l'on cherche à étudier.
Si ses points sont disséminés anarchiquement, il semble que les deux variables en présence soient indépendantes (pas de corrélation).
Si sa forme tend à se conformer à une courbe, il suggère la possibilité de **lier fonctionnellement *Y* et *X***, c'est-à-dire de déterminer une fonction f telle que $f(x_i)$ soit une "bonne" approximation de y_i pour tout i de $[\![1;n]\!]$. On dit que la courbe représentative d'une telle fonction réalise un **ajustement** du nuage de points. Il est d'autant "meilleur" qu'il passe "au plus près" des points du nuage.
Si la fonction f est affine (du type $f(X) = aX + b$), on parle d'**ajustement affine**.

1. Coefficient de corrélation linéaire

La propension d'une série statistique double (X, Y) à subir un ajustement affine s'exprime par son **coefficient de corrélation linéaire** :

$$\boldsymbol{r = \frac{s_{XY}}{s_X\, s_Y}}$$

où : s_{XY} désigne la covariance de la série statistique (X, Y) : $s_{XY} = \frac{1}{n}\sum_{i=1}^{i=n}(x_i - \bar{X})(y_i - \bar{Y})$

s_X désigne l'écart type de la série statistique X : $s_X = \sqrt{\frac{1}{n}\sum_{i=1}^{i=n}(x_i - \bar{X})^2}$

s_Y désigne l'écart type de la série statistique Y : $s_Y = \sqrt{\frac{1}{n}\sum_{i=1}^{i=n}(y_i - \bar{Y})^2}$

On a toujours : $\boldsymbol{-1 \leq r \leq 1}$

Plus $|r|$ est proche de 1, meilleure est la corrélation linéaire entre X et Y, et d'autant plus significatif sera un ajustement affine entre X et Y.
Néanmoins, une forte corrélation linéaire ne démontre en rien l'existence d'un lien causal entre X et Y : elle ne fait au mieux qu'en suggérer l'existence.

Le **signe de r** exprime le sens de variation : Y est une fonction croissante de X lorsque $r > 0$, et Y est une fonction décroissante de X lorsque $r < 0$.

2. Ajustement affine par la méthode des moindres carrés

La recherche d'un ajustement par une droite du nuage de points associé à l'échantillon observé se justifie s'il existe une forte corrélation linéaire entre X et Y.
Le problème est de trouver une droite du plan dont l'écart par rapport aux points du nuage est minimal, tout en sachant que plusieurs définitions de la mesure de cet écart sont possibles.

Droite de régression de Y en X : $D_{Y/X}$

La droite de régression de Y en X est la droite qui minimise la somme des écarts verticaux entre les points $M_i(x_i\ ;\ y_i)$ du nuage et les points de mêmes abscisses $M_i'(x_i\ ;\ y_i')$ appartenant à cette droite.

Une équation de $D_{Y/X}$ est : $Y - \bar{Y} = a(X - \bar{X})$

Son coefficient directeur est : $a = \frac{s_{XY}}{{s_X}^2}$

Droite de régression de X en Y : $D_{X/Y}$

La droite de régression de X en Y est la droite qui minimise la somme des écarts horizontaux entre les points $M_i(x_i\ ;\ y_i)$ du nuage et les points de mêmes ordonnées $M_i'(x_i'\ ;\ y_i)$ appartenant à cette droite.

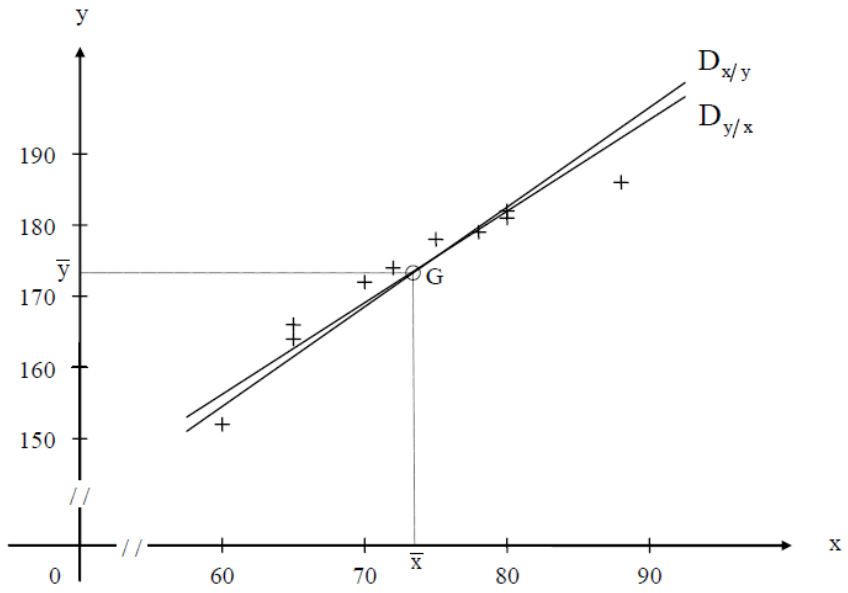
Une équation de $D_{X/Y}$ est : $X - \bar{X} = a'(Y - \bar{Y})$

Son coefficient directeur est : $\frac{1}{a'}$ avec $a' = \frac{s_{XY}}{{s_Y}^2}$

Ces deux droites de régression passent par le point $G(\bar{X}\ ;\ \bar{Y})$, appelé point moyen.

On remarque de plus que $\boldsymbol{r^2 = aa'}$
Ainsi, plus r est proche de 1, plus les coefficients directeurs a et $1/a'$ sont proches l'un de l'autre, (donc plus les droites $D_{Y/X}$ et $D_{X/Y}$ sont proches l'une de l'autre), et meilleur est l'ajustement du nuage de points que constitue chacune d'entre elles.

Exemple (suite) : Coefficient de corrélation linéaire : $r \approx 0{,}962$
Equation de la droite de régression $D_{Y/X}$: $y = 1{,}128\,x + 90{,}799$
Equation de la droite de régression $D_{X/Y}$: $y = 1{,}218\,x + 84{,}207$



$$\bar{X} = \frac{1}{10}\sum_{i=1}^{i=10} x_i = 73{,}3$$

$$\bar{Y} = \frac{1}{10}\sum_{i=1}^{i=10} y_i = 173{,}5$$

Chacun de ces deux ajustements affines permet de faire des **extrapolations**, en permettant de calculer une valeur de Y à partir d'une valeur de X qui ne fait pas partie des valeurs observées.
La **fiabilité d'une telle extrapolation** dépend de la qualité de la corrélation linéaire entre X et Y, et diminue fortement au fur et à mesure que l'on s'éloigne l'intervalle correspondant à l'étendue de X.
Rappelons enfin que la relation affine trouvée ne démontre en rien l'existence d'un lien causal entre X et Y : elle ne fait au mieux qu'en suggérer l'existence.

3. Ajustements non affines

On peut être amené à chercher un ajustement du nuage de points à l'aide d'une fonction non affine, par exemple une fonction puissance, logarithmique, exponentielle…
Pour étudier la pertinence du choix de cette fonction, on peut effectuer un **changement de variable** par le biais d'une fonction u (non affine) appliquée à la variable X.
On définit ainsi une nouvelle variable Z telle que $z_i = u(x_i)$ pour tout i de $[\![1; n]\!]$, puis on étudie la corrélation linéaire entre Y et Z.
Si celle-ci est forte, il est justifié de déterminer la droite de régression de Y en Z.
On en déduit un ajustement non affine de Y en X sous la forme : $Y = a\,u(X) + b$

Un changement de variable peut éventuellement être effectué sur Y, ou à la fois sur X et sur Y.
Par la même méthode que précédemment, on aboutit dans ce dernier cas à une relation du type $v(Y) = a\,u(X) + b$ et on en déduit : $Y = v^{-1}[a\,u(X) + b]$
(à condition que v soit bijective sur l'intervalle dans lequel on souhaite l'utiliser).

EXERCICES

EXERCICE 1

Cet exercice est un prolongement de l'exemple du cours :
La direction d'un quotidien compare chaque mois, sur une durée de 10 mois :

- son investissement publicitaire, exprimé en k€ : variable X
- le nombre d'exemplaires vendus, exprimé en centaines de milliers d'unités : variable Y

Les résultats sont rassemblés dans le tableau suivant :

Mois	1	2	3	4	5	6	7	8	9	10
X : Invest. pub	60	65	65	70	75	72	80	88	78	80
Y : Ventes	153	164	166	172	178	174	181	186	179	182

1. a) Calculer le coefficient de corrélation linéaire de la série double (X, Y) et les équations des deux droites de régression $D_{Y/X}$ et $D_{X/Y}$.
   b) Utiliser ces équations pour prévoir le nombre d'exemplaires vendus avec un investissement publicitaire mensuel de 100 k€.

2. On considère les variables $Z = \ln(X)$ et $T = \ln(Y)$.
   a) Calculer les coefficients de corrélation linéaire $r(X,T)$, $r(Z,Y)$, $r(Z,T)$.
   b) En déduire l'ajustement le plus pertinent exprimant Y en fonction de X, et prévoir ainsi le nombre d'exemplaires vendus avec un investissement publicitaire mensuel de 100 k€.

EXERCICE 2

Des tests effectués sur une voiture ont permis de dresser le tableau suivant, qui fournit ses distances de freinage sur route sèche (en m) correspondant à sa vitesse (en km/h) :

Vitesse	20	30	50	70	90	110	130
Distance de freinage	2	5	14	28	46	68	95

1) Représenter le nuage de point associé à cette série double, puis calculer son coefficient de corrélation linéaire. Quelle conclusion en tirer ?
2) Répondre aux mêmes questions avec la série double constituée des distances de freinage et des carrés des vitesses, puis comparer avec les résultats du 1).
3) Idem avec la série double constituée des distances de freinage et des exponentielles de base e des vitesses.
4) Déterminer au moyen d'une régression non linéaire une relation qui semble pertinente entre vitesse et distance de freinage.
5) Quelle distance de freinage peut-on attendre pour une vitesse de 150 km/h ?

EXERCICE 3

Sur une chaîne de production, le temps nécessaire par opération diminue lorsque la répétition de l'opération augmente. On se propose d'étudier cet effet d'apprentissage sur le temps moyen de fabrication d'un moteur.

Nombre de moteurs fabriqués par série	Nombre cumulé C de moteurs fabriqués	Temps moyen T par moteur fabriqué (h) sur l'ensemble des séries
10	10	18
10	20	10,4
15	35	7,5
15	50	6,3
30	80	5,1
40	120	4,2
40	160	3,8

1. Un ajustement affine de la série double (C, T) est-il justifié ?

2. On considère les variables $X = \ln(C)$ et $Y = \ln(T)$.
   a) Calculer les coefficients de corrélation linéaire $r(C, Y)$, $r(X, T)$, $r(X, Y)$.
   b) En déduire l'ajustement le plus pertinent exprimant T en fonction de C, et prévoir ainsi le temps moyen de fabrication d'un moteur lorsqu'on aura fabriqué 200 moteurs.