Statistiques

Régression linéaire

http://www.i2m.univ-amu.fr/perso/jean-philippe.preaux/

1. Introduction : corrélation entre deux variables

Lorsque nous nous intéressons à des questions du type :

- quelle est la taille moyenne des garçons âgés d'une vingtaine d'années ?
- le poids moyen des pains produits dans une boulangerie est-il supérieur à 800 grammes ?

nous étudions le comportement statistique d'une seule variable : taille, poids des pains, etc...

Il existe cependant toute une gamme de problèmes statistiques où l'on s'intéresse à la relation entre plusieurs variables.

Exemples :

- les individus les plus grands sont-ils les plus lourds ?
- le revenu d'une famille a-t-il une influence sur les résultats scolaires des enfants ?
- y a-t-il une relation entre le tabagisme et les cancers du poumon ?
- le rendement en céréales dépend-il de la quantité d'engrais utilisée ?
- la productivité d'une entreprise est-elle liée au salaire des ouvriers ou employés ?

Dans ces questions, nous désirons savoir si le comportement d'une variable est influencé par la valeur d'une autre variable :

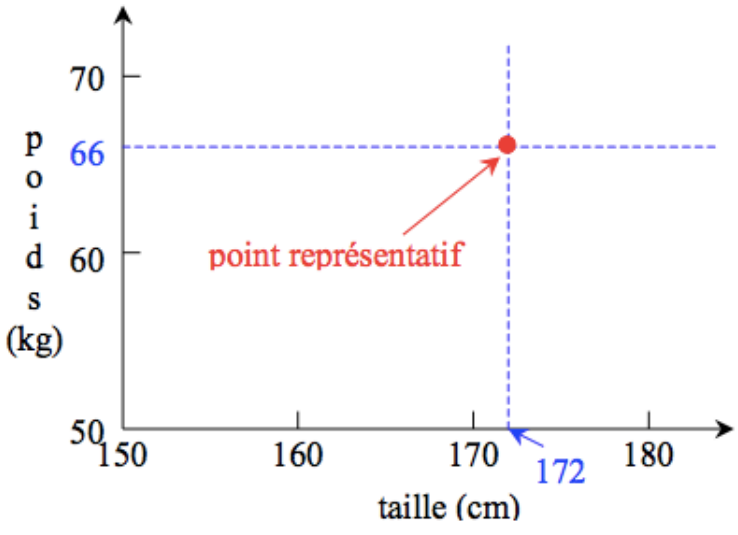
La relation peut être causale ou non Pour étudier les relations ou corrélations entre deux variables statistiques, on peut les porter sur un graphique.

Exemple. : relation entre la taille et le poids des individus.

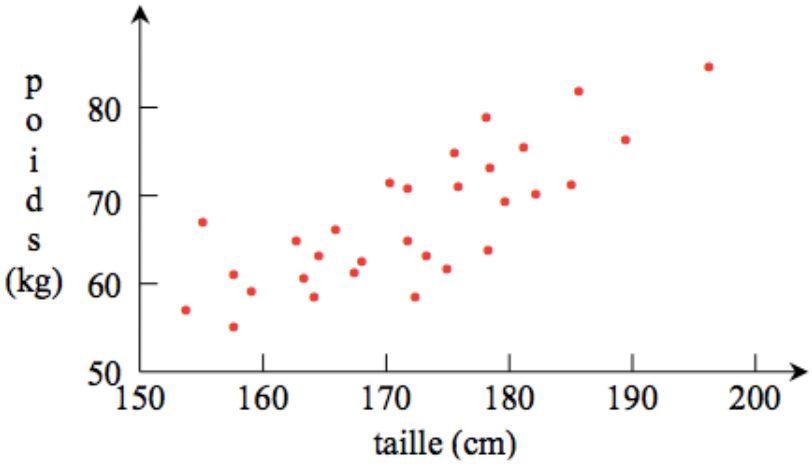
Pour chaque individu de l'échantillon, on porte sur un graphique :

- sa taille en abscisse,
- son poids en ordonnée.

Chaque individu est donc, dans ce graphique, représenté par un point (point représentatif) soit un individu mesurant 172 cm et pesant 66 kg :



L'ensemble des individus est représenté par un nuage de points ; il y a autant de points que d'individus :

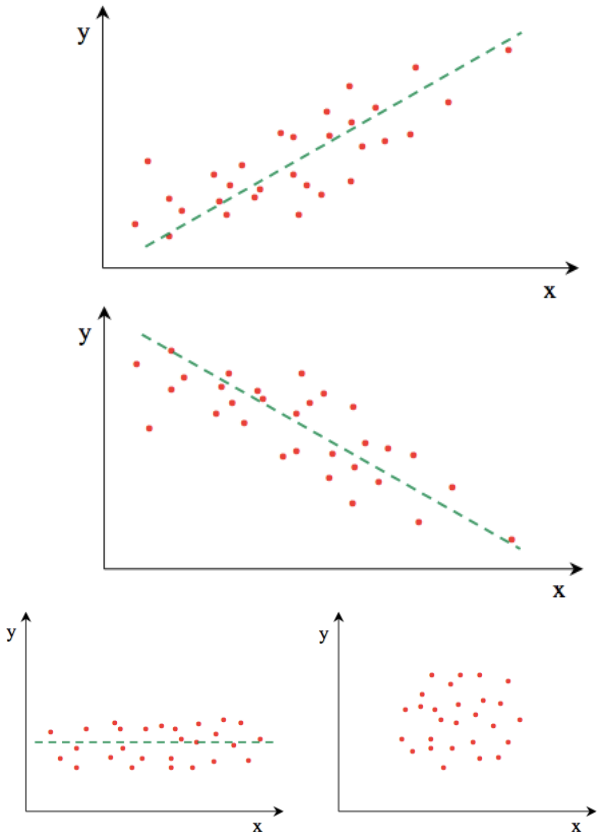


On peut (par la pensée ou réellement) tracer une droite qui passe "au plus proche " des points du nuage :

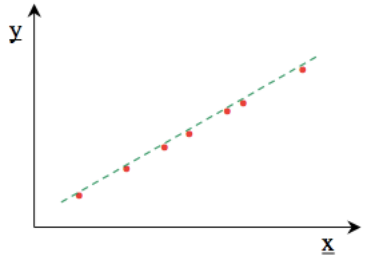
• Si cette droite a une pente positive, on dira qu'il y a corrélation positive entre les deux variables :
Dans notre population, lorsque la taille augmente, le poids aurait tendance à augmenter.

• Si elle a une pente négative, c'est une corrélation négative :
Dans notre population, lorsque la taille augmente, le poids aurait tendance à diminuer.

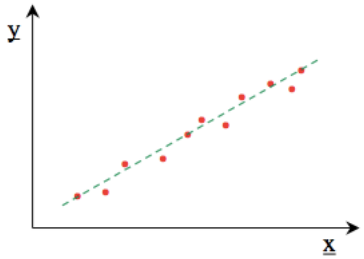
• Si elle est "horizontale", ou si on ne peut pas décider, c'est qu'il y a absence de corrélation :
Dans notre population, une augmentation de taille n'aurait pas tendance à entraîner une modification du poids.



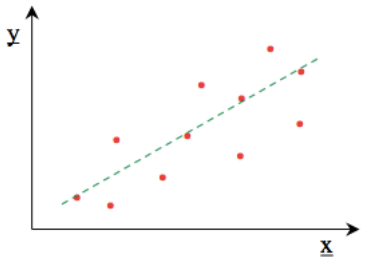
La qualité de la corrélation entre deux variables peut se mesurer par la dispersion des points autour de la relation moyenne.



Corrélation parfaite



Bonne corrélation



Corrélation moyenne

Le plan $\mathscr{P}$ est rapporté à un repère orthonormé. Étant données deux séries numériques X et Y de même effectif n :

$$X = (x_1, x_2, \dots, x_n) \quad ; \quad Y = (y_1, y_2, \dots, y_n)$$

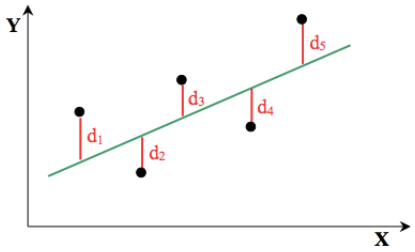
On appelle nuage de points d'abscisses X et d'ordonnées Y, le sous-ensemble de $\mathscr{P}$ constitué des points M_i d'abscisse x_i et d'ordonnée y_i :

$$\left\{M_i(x_i, y_i) \in \mathscr{P} \mid i \in [\![1, n]\!]\right\}$$

2. Méthode des moindres carrés ; droite de régression linéaire

2.1. Distance d'une droite à un nuage de points.

On donne un sens à la droite "qui approche au mieux le nuage de points" par la méthode dite des moindres carrés.



Soit Δ une droite d'équation :

$$y = ax + b$$

Pour deux séries statistiques $X = (x_1, \dots, x_n)$ et $Y = (y_1, \dots, y_n)$ de même effectif, on considère pour tout $k \in [\![1, n]\!]$:

$$d_k = y_k - (ax_k + b)$$

qui est la distance algébrique entre le point $M_k(x_y, y_k)$ et le point de même abscisse sur la droite Δ, et on définit :

$$S_{a,b} = \sum_{k=1}^{n} (d_k)^2$$

que l'on appelle la distance au sens des moindres carrés de la droite Δ au nuage de points d'abscisse X et d'ordonnée Y.

Cette quantité vérifie des propriétés propres aux distances :

Sous les mêmes hypothèses :

$S_{a,b} \geqslant 0$ *et* $S_{a,b} = 0 \iff$ *tous les points $M_i(x_i, y_i)$ du nuage sont sur la droite Δ*

Démonstration. C'est une somme de carrés, donc positive, et une somme de nombres positifs est nulle si et seulement si chacun de ses termes est nul. ■

Les résultats suivants de cette parties seront admis pour l'instant, et démontrés ultérieurement (cf. Chapitre "Fonctions de deux variables réelles").

2.2. La droite de régression linéaire.

Le résultat fondamental est qu'au sens de cette distance, il existe une unique droite qui approche au mieux le nuage de points au sens des moindres carrés :

Si on suppose en outre que $X = (x_1, x_2, \dots, x_n)$ est une série de valeurs non constantes, i.e.

$$\exists (i,j) \in [\![1,n]\!], \quad \text{tel que } i \neq j \text{ et } x_i \neq x_j$$

alors l'ensemble :

$$\left\{ S_{a,b} \mid (a,b) \in \mathbb{R}^2 \right\}$$

admet un minimum qui est atteint si et seulement si :

$$a = \frac{\sum_{k=1}^{n} (x_i - \overline{X})(y_i - \overline{Y})}{\sum_{k=1}^{n} (x_i - \overline{X})^2} \quad ; \quad b = \overline{Y} - a\overline{X}$$

où :

$$\overline{X} = \frac{1}{n}\sum_{k=1}^{n} x_k \; ; \overline{Y} = \frac{1}{n}\sum_{k=1}^{n} y_k \text{ désignent les moyennes de } X \text{ et } Y$$

Ces formules s'écrivent de manière plus concises en utilisant la variance et la covariance.

Soient $X = (x_1, x_2, \dots, x_n)$ et $Y = (y_1, y_2, \dots, y_n)$ deux séries numériques de même effectif.
En notant :

$$\overline{X} = \frac{1}{n}\sum_{k=1}^{n} x_i \quad ; \quad \overline{Y} = \frac{1}{n}\sum_{k=1}^{n} y_i$$

les moyennes de X et Y ;

- *la variance de X est :*

$$\boxed{V(X) = \frac{1}{n}\sum_{i=1}^{n} \left(x_i - \overline{X}\right)^2}$$

- *la covariance de X et Y est :*

$$\boxed{\mathrm{cov}(X,Y) = \frac{1}{n}\sum_{i=1}^{n} \left(x_i - \overline{X}\right)\left(y_i - \overline{Y}\right)}.$$

Remarque. Il vient immédiatement :

$$\mathrm{cov}(X,Y) = \mathrm{cov}(Y,X) \quad ; \quad V(X) = \mathrm{cov}(X,X)$$

$$V(X) = 0 \iff X \text{ est constante}$$

et :

$$a = \frac{\sum_{k=1}^{n}(x_i - \overline{X})(y_i - \overline{Y})}{\sum_{k=1}^{n}(x_i - \overline{X})^2} = \frac{\text{cov}(X,Y)}{V(X)}$$

Soient deux séries statistiques $X = (x_1, \ldots, x_n)$ et $Y = (y_1, \ldots, y_n)$ de même effectif, avec X non constante. Pour le nuage de points :

$$\left\{M_i(x_i, y_i) \in \mathscr{P} \mid i \in [\![1, n]\!]\right\}$$

sa droite de régression linéaire est la droite d'équation $y = ax + b$ avec :

$$a = \frac{\text{cov}(X,Y)}{V(X)} \qquad \text{et} \quad b = \overline{Y} - a\overline{X}.$$

C'est la droite qui approche le mieux le nuage de points au sens des moindres carrés.

2.3. **Coefficient de corrélation.**

Le signe de la pente a donne le sens de corrélation, mais pas sa qualité.

$a > 0$ corrélation positive

$a < 0$ corrélation négative

$a = 0$ pas de corrélation

La qualité de la corrélation peut être mesurée par un coefficient de corrélation ρ :

Sous les mêmes hypothèses, le coefficient de corrélation est :

$$\boxed{\rho = \frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y}}$$

où :

$$\sigma_X = \sqrt{V(X)} \qquad ; \qquad \sigma_Y = \sqrt{V(Y)}$$

sont les écarts-types de X et de Y.

Le coefficient de corrélation vérifie :

Le coefficient de corrélation est compris entre -1 et $+1$.

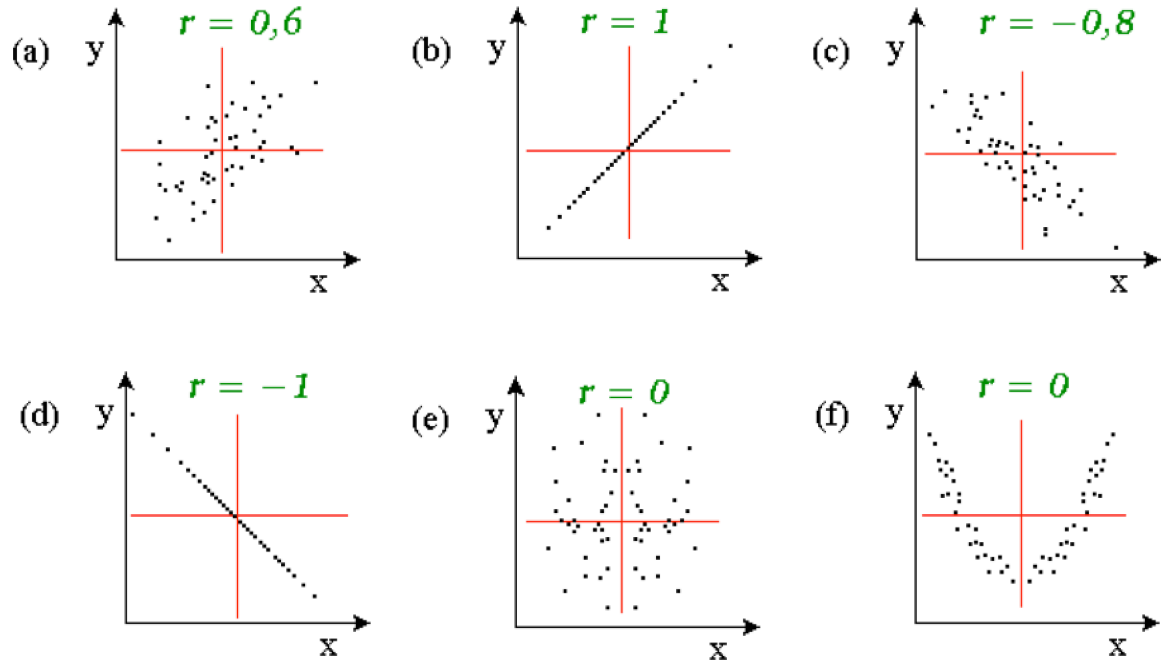
Plus il s'éloigne de zéro, meilleure est la corrélation :

- *Si $\rho = +1$: corrélation positive parfaite (la droite de régression linéaire passe par tous les points du nuage, et a une pente $a > 0$.*
- *Si $\rho = -1$: corrélation négative parfaite (la droite de régression linéaire passe par tous les points du nuage, et a une pente $a < 0$.*
- *Si $\rho = 0$: absence totale de corrélation.*

Le coefficient de corrélation peut aussi s'exprimer :

$$\rho = \frac{\sum_{i=1}^{n}(x_i - \overline{X})(y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{X})^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{Y})^2}}$$

• Quelques exemples de corrélation (le coefficient de corrélation ρ est indiqué dans chaque cas).



2.4. **Exemples.**

Supposons un échantillon aléatoire de 4 firmes pharmaceutiques présentant les dépenses de recherche X et les profits Y suivants (en milliers de dollars) :

X	Y
40	50
40	60
30	40
50	50

Déterminons la droite de régression et le coefficient de corrélation.

On commence par calculer les moyennes $\overline{X}$ et $\overline{Y}$:

$$\overline{X} = \frac{1}{4} \times (40 + 40 + 30 + 50) = \frac{160}{4} = 40$$
$$\overline{Y} = \frac{1}{4} \times (50 + 60 + 40 + 50) = \frac{200}{4} = 50$$

Complétons le tableau suivant :

X	Y	$X - \overline{X}$	$Y - \overline{Y}$	$(X - \overline{X})^2$	$(Y - \overline{Y})^2$	$(X - \overline{X})(Y - \overline{Y})$
40	50	0	0	0	0	0
40	60	0	+10	0	100	0
30	40	−10	−10	100	100	100
50	50	+10	0	100	0	0

donc :

$$\sum_i (x_i - \overline{X})(y_i - \overline{Y}) = 100 \quad ; \quad \sum_i (x_i - \overline{X})^2 = 200 \quad ; \quad \sum_i (y_i - \overline{Y})^2 = 200$$

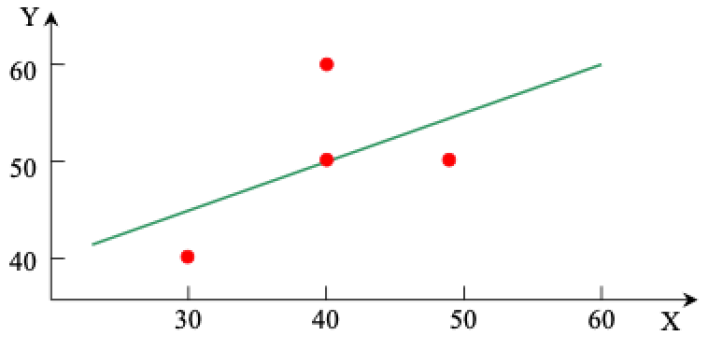
On en déduit les coefficients a et b de la droite de régression linéaire $y = ax + b$:

$$a = \frac{\sum_i (x_i - \overline{X})(y_i - \overline{Y})}{\sum_i (x_i - \overline{X})^2} = \frac{1}{2} \quad ; \quad b = \overline{Y} - a.\overline{X} = 30$$

et le coefficient de corrélation :

$$\rho = \frac{\sum_i (x_i - \overline{X})(y_i - \overline{Y})}{\sqrt{\sum_i (x_i - \overline{X})^2}\sqrt{\sum_i (y_i - \overline{Y})^2}} = \frac{100}{200} = \frac{1}{2}$$

elle est positive et de qualité médiocre.



Remarques.

• Le coefficient de corrélation nous donne des informations sur l'existence d'une relation linéaire (sous forme d'une droite) entre les deux grandeurs considérées.

Un coefficient de corrélation nul ne signifie pas l'absence de toute relation entre les deux grandeurs. Il peut exister une relation non linéaire entre elles.

On peut être amené à corréler $f(X)$ et Y pour une fonction f bien choisie.
(Exemple : pour le nuage (f) ci-dessus on pourrait essayer : $f(x) = x^2$, ou encore x^4, etc. . .)

• Il ne faut pas confondre corrélation et relation causale. Une bonne corrélation entre deux grandeurs peut révéler une relation de cause à effet entre elles, mais pas nécessairement.

Exemples :

– Si on compare la durée de vie des individus à la quantité de médicaments pour le cœur qu'ils ont absorbée, on observera probablement une corrélation négative. Il serait imprudent de conclure que la prise de médicaments pour le cœur abrège la vie des individus. . . (en fait, dans ce cas, la corrélation est l'indice d'une cause commune : la maladie de cœur).

– Le soleil tire son énergie de réactions nucléaires transformant l'hydrogène en hélium. Notre société tire une bonne part de son énergie de la combustion du pétrole. Si on compare, année après année, la quantité d'hélium contenue dans le soleil au prix moyen du pétrole, on obtiendra une bonne corrélation positive, sans qu'il y ait la moindre relation de cause à effet, ni aucune cause commune.

– Depuis une dizaine d'années, la taille d'un élève de la classe, né en 2002, est très

bien corrélée avec la puissance de calcul des ordinateurs personnels. Cette excellente corrélation ne révèle bien évidemment aucune relation de cause à effet, ni cause commune.

L'existence d'une corrélation, aussi bonne soit elle, n'est jamais la preuve d'une relation de cause à effet.