

Chapitre :

Corrélation et causalité



Ce chapitre a pour objet les statistiques à deux variables.

On fait des statistiques à deux variables quand on recueille, pour chaque individu d'un échantillon étudié, les valeurs de deux grandeurs, X et Y .

Autrement dit, pour chaque individu, on recueille un couple de valeurs (x_i, y_i) .

L'ensemble de ces couples est appelé série statistique double, que l'on résume par la notation $(X; Y)$.

On peut s'intéresser à une éventuelle corrélation entre les grandeurs X et Y , autrement dit à la possibilité d'exprimer la valeur Y en fonction de la valeur X , c'est à dire obtenir une fonction f telle que $Y = f(X)$.

Au préalable, on pourra réviser les méthodes pour obtenir l'équation réduite d'une droite tracée dans un repère (ou dont on connaît les coordonnées de deux points). Vous pouvez trouver [ici](#) et [là](#) deux liens vers des vidéos sur le sujet.

⊗ **Activité** : Corrélation

I. Nuages de points et ajustement

Définition Le **nuage de point** associé à la série statistique double $(x_1; y_1), \dots, (x_n; y_n)$ est l'ensemble des points de coordonnées $(x_i; y_i)$.

Définition Soit $(X; Y)$ une série statistiques double. Si l'on note $\bar{x}$ (resp. $\bar{y}$) la moyenne des x_i (resp. y_i), Alors le **point moyen** est le point G de coordonnées $(\bar{x}; \bar{y})$.

Rappel On a donc $\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n}$ et $\bar{y} = \frac{y_1 + y_2 + \dots + y_n}{n}$.

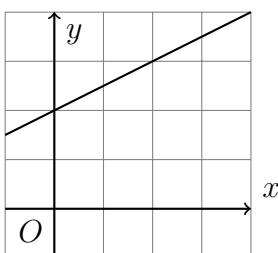
Les nuages de points peuvent avoir toutes sortes d'allure. Dans certains cas, on peut « voir » ce qui semble être une relation fonctionnelle entre X et Y , sous la forme $Y = f(X)$.

Cette relation n'existe généralement pas de façon exacte, on cherche donc seulement une relation approchée et satisfaisante.

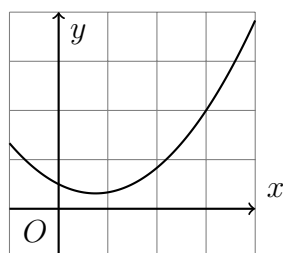
Effectuer un **ajustement** de la série consiste à trouver une fonction f telle que la courbe d'équation $y = f(x)$ passe « près » des points du nuage.

Quand un ajustement semble possible, on dit que les variables X et Y sont **corrélées**.

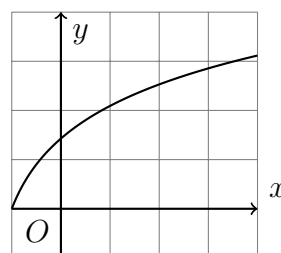
La fonction f de l'ajustement peut être de n'importe quelle forme ; les plus courantes sont :



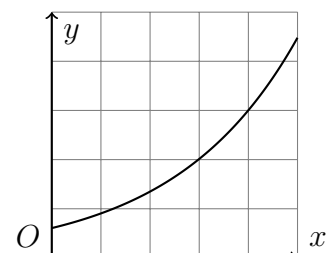
affine



quadratique
(polynomiale de degré 2)



logarithmique



exponentiel

Nous verrons dans la section suivante comment réaliser des ajustements affines. Les autres utilisent généralement ce premier type d'ajustement.

► **Exercices** : Nuages de points et point moyen

Un ajustement a pour but de permettre d'**interpoler** ou d'**extrapoler** des valeurs :

- **interpoler** c'est estimer une valeur intermédiaire (par exemple une valeur de y pour une valeur de x située entre les valeurs x_i de la série) ;
- **extrapoler** c'est estimer une valeur au-delà du nuage du point (par exemple une valeur de y pour une valeur de x située après les valeurs x_i).



Corrélation n'est pas causalité !

Un lien « mathématique » entre deux grandeurs ne démontre pas de lien de cause à effet entre les deux grandeurs.

Il est possible de trouver une corrélation statistiquement significative et fiable pour deux variables qui ne sont en fait liées par aucun lien de cause à effet. En fait, ce type de corrélation est fréquent ! Souvent, elles sont dues au fait que les deux variables sont associées à une variable causale différente, qui intervient en parallèle des données que nous mesurons.

Exemple (activité physique et cancer de la peau)

Imaginez que vous avez des données médicales. Vous observez une corrélation positive statistiquement significative entre activité physique et cancer de la peau, ce qui veut dire que les personnes qui ont plus d'activité physique tendent à être les personnes qui font un cancer de la peau. Cette corrélation semble forte et fiable et apparaît dans plusieurs populations de patients. Sans chercher plus loin, vous pourriez conclure que l'activité physique peut, d'une manière ou d'une autre, provoquer le cancer ! D'après ces résultats, vous pourriez même développer une hypothèse plausible : peut-être que le stress dû à l'activité physique prive l'organisme de sa capacité à se protéger des risques de cancer.

Mais imaginez qu'en réalité, cette corrélation existe dans votre ensemble de données parce que les personnes qui vivent dans des régions très ensoleillées toute l'année sont significativement plus actives au quotidien que celles qui vivent dans des régions moins bien loties. Cela apparaît dans leurs données sous la forme d'une activité physique supérieure. Dans le même temps, une exposition accrue au soleil implique un plus grand nombre de cancers de la peau. Les deux variables, activité physique et cancer de la peau, sont affectées par une troisième variable causale, l'exposition au soleil, mais elles ne sont pas liées par un lien de cause à effet.

II. Ajustement affine

⊗ **Activité** : Ajustement

On parle d'**ajustement affine** lorsque la fonction de corrélation est une fonction affine, autrement dit quand les points du nuage de point sont presque alignés.

Il y a plusieurs manières de choisir une droite « proche » des points, notamment :

- On peut choisir par exemple la droite passant par les **points extrêmes** (à gauche et à droite).
- On peut aussi partager le nuage de points en deux, déterminer pour chaque partie son point moyen, et choisir la droite qui passe par les deux points moyens. C'est ce que l'on appelle la **droite de Mayer**.

Une des méthodes les plus utilisées (en tout cas connues), est celle des **moindres carrés**.

C'est la droite d'équation $y = ax + b$ telle que la somme des $(y_i - (ax_i + b))^2$ (les carrés des écarts verticaux des points à la droite, voir ci-dessous les segments rouges) est la plus petite possible (d'où les moindres carrés).

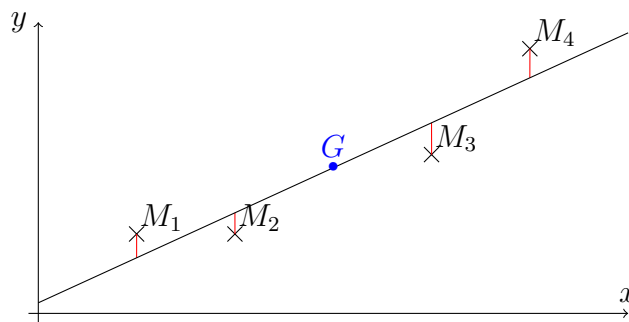
La figure ci-contre illustre, mais n'est pas exacte (la droite n'est pas nécessairement celle des moindres carrés).

La droite des moindres carrés a la particularité de passer par le point moyen G du nuage de points.

Techniquement, on a $a = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$

et $b = \bar{y} - a\bar{x}$

Où le symbole $\sum$ désigne la somme.



Il existe une autre manière de formuler : $a = \frac{\text{cov}(X; Y)}{V(X)}$, où :

- $\text{cov}(X; Y) = \frac{1}{n} \sum (x_i - \bar{x})(y_i - \bar{y})$ est la **covariance** de X et Y ;
- $V(X) = \frac{1}{n} \sum (x_i - \bar{x})^2$ est la **variance** de X .

Définition Le **coefficient de corrélation** est le nombre r défini par :

$$r = \frac{\text{cov}(X; Y)}{\sqrt{V(X) \times V(Y)}} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2} \times \sqrt{\sum (y_i - \bar{y})^2}}$$

On a toujours $-1 \leq r \leq 1$.

C'est un indicateur de qualité de l'ajustement :

- Si $|r| \simeq 1$, alors les points du nuage sont quasiment alignés, l'ajustement affine par les moindres carrés est adapté.
- Si $r \simeq 0$, alors les points sont très dispersés autour de la droite, l'ajustement affine par les moindres carrés n'est pas adapté.

Les valeurs de a , b et r , ainsi que la covariance et les variances, peuvent être obtenues grâce à une calculatrice.

Pour les exercices, sauf demande (implicite) contraire, nous utiliserons directement la calculatrice.

► **Exercices** : Droite des moindres carrés ; Divers ajustements affines

III. Autres ajustements

Bien sûr, d'autres ajustements qu'affines sont possibles. Pour certains, on peut passer par des changements de variables qui permettent de transformer un nuage de points en un nuage de points presque alignés, ce qui permet un ajustement affine, puis on revient en arrière pour obtenir l'ajustement du nuage initial.

► **Exercices** : Autres ajustements