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) ; c'est l'objet du cadre « Consolider les bases » page 258. Vous pouvez trouver ici et là deux liens vers des vidéos sur le sujet.

⊛ **Activité :** Situation 1 page 258

I. Nuages de points

Définition Le **nuage de point** associé à la série statistique double $(x_1;y_1)$, …, (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 $\overline{x}$ (resp. $\overline{y}$) la moyenne des x_i (resp. y_i), Alors le **point moyen** est le point G de coordonnées $(\overline{x};\overline{y})$.

Rappel On a donc $\overline{x} = \dfrac{x_1 + x_2 + \cdots + x_n}{n}$ et $\overline{y} = \dfrac{y_1 + y_2 + \cdots + y_n}{n}$.

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 la courbe d'équation $y = f(x)$ passe « près » des points du nuage.
Voir le manuel page 260 pour quelques représentations possibles.
Quand un ajustement semble possible, on dit que les variables X et Y sont **corrélées**. Cependant, ce lien « mathématique » entre les deux grandeurs ne démontre pas de lien de cause à effet entre les deux grandeurs.
Voir la Situation 3 page 259 pour observer cela.

▶ **Exercices :** 36-39p270
Comme nous l'avons vu dans la situation 1, 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).

Nous devons cependant introduire un ajustement particulier avant faire des exercices là-dessus.

II. Ajustement affine

⊛ **Activité :** Situation 2 page 259

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.

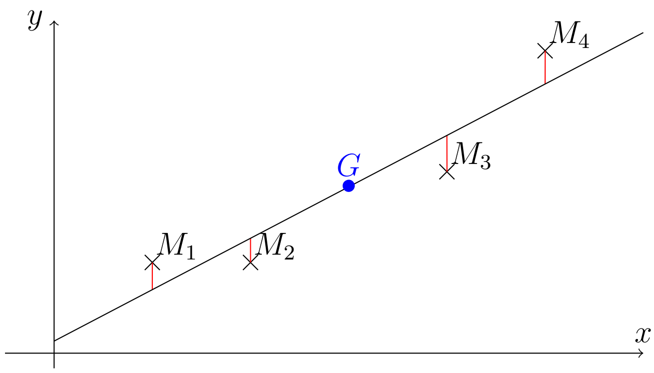
Il y a plusieurs manières de choisir une droite « proche » des points, notamment :

- On peut choisir par exemple la droite passant par les points extrêmes (comme nous l'avons fait dans la situation 1).
- 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-dessous 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 $a = \dfrac{\sum(x_i - \overline{x})(y_i - \overline{y})}{\sum(x_i - \overline{x})^2}$ et $b = \overline{y} - a\overline{x}$

Où le symbole $\sum$ désigne la somme.

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

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

<u>Définition</u> Le **coefficient de corrélation** est le nombre r défini par :

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

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

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. Voir page 263.
Pour les exercices, sauf demande (implicite) contraire, nous utiliserons directement la calculatrice.

▶ **Exercices :** 3p263

▶ **Exercices :** 4,5p265 (interpolation, extrapolation), 40p270 (diverses droites)

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.
Voir le manuel page 264.

▶ **Exercices :** 56p274 (corrigé dans le manuel), 57p274

★ **Approfondissement :** choix d'un exercice par groupe parmi les 59 à 66 pages 276 à 279.