Statistiques à deux variables (partie 1)

I) Introduction

A) Vocabulaire et exemple

Le problème qui se pose dans les séries statistiques à deux variables est principalement celui du lien qui existe ou non entre chacune des variables.
Le but de ce chapitre est celui-ci : savoir si il y a un éventuel lien de corrélation ou de cause à effet entre les deux variables étudiées.

Définition : Lorsque l'on étudie deux caractères (ou variables) x et y sur une même population de taille n, on associe à chaque individu de la population un couple $(x_i; y_i)$ où x_i et y_i sont les valeurs respectives des variables x et y prises par l'individu "numéro i".
On appelle série statistiques doubles $(x; y)$ l'ensemble des couples $(x_i; y_i)$ associés à chaque individu de la population. On le présente souvent sous forme de tableau :

Valeurs x_i de la variable x	x_1	x_2	x_3	...	x_n
Valeurs y_i de la variable y	y_1	y_2	y_3	...	y_n

On peut calculer la **moyenne** notée $\overline{x}$ et $\overline{y}$ de la manière usuelle :

$$\overline{x} = \frac{x_1 + ... + x_n}{n} \text{ et } \overline{y} = \frac{y_1 + ... + y_n}{n}$$

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.
Y a t-il un lien entre les variables "Notes obtenues lors de l'épreuve A", et "Notes obtenues lors de l'épreuve B" ?

Les résultats sont rassemblés dans le tableau suivant.

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

On trouve :

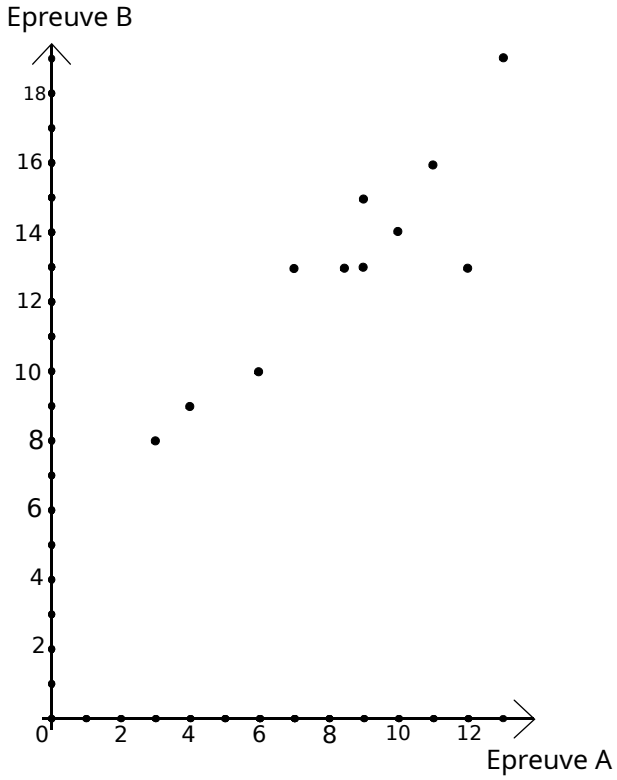
$$\overline{x} = \frac{3 + 4 + 6 + 7 + 9 + 10 + 9 + 11 + 12 + 13}{10} = 8.4$$

$$\overline{y} = \frac{8 + 9 + 10 + 13 + 15 + 14 + 13 + 16 + 13 + 19}{10} = 13$$

B) Nuage de points

Nous souhaitons maintenant réaliser un graphique qui traduise les deux séries statistiques ci-dessus.
Dans un repère orthogonal (O ; i ; j), à chaque couple $(x_i ; y_i)$ de la série statistiques double (x, y) on peut associer le point M_i de coordonnées (x_i, y_i).
L'ensemble de ces points est appelé **nuage de points** associé à la série statistique double.

Le point G de coordonnées $(\overline{x}, \overline{y})$ est appelé **point moyen** du nuage.
Dans notre exemple, nous obtenons le nuage de points suivant :



II) Ajustement

On cherche à savoir s'il existe un lien entre les deux variables étudiées. On va donc essayer de savoir s'il existe une courbe qui approche au mieux le nuage de point, c'est à dire qui passe au plus près des points de ce nuage. On dit alors que l'on a effectué un **ajustement**.

Le tracé met donc (parfois) en évidence la possibilité de "reconnaître" graphiquement la possibilité d'une **relation fonctionnelle** entre les deux grandeurs observées.
Lorsque les points du nuage semblent alignés, on appelle **droite d'ajustement** une droite qui passe au plus près des points du nuage. On dit que cette droite réalise un **ajustement affine** du nuage de points.

A) Méthode 1 : ajustement à la règle

La première méthode est de tracer au jugé une droite D passant le plus près possible des points du nuage et d'en trouver l'équation du type $y = ax + b$.
C'est une méthode "au pifométre", mais qui donne parfois des résultats convenables. Nous en ferons un exemple en TD.

B) Méthode 2 : méhode de Mayer

Cet ajustement consiste à déterminer la droite passant par **deux points moyens** du nuage de point. On notera ces points G_1 et G_2 dans notre exemple.
Dans notre exemple précédent, il y avait 10 éléments dans l'épreuve A et 10 éléments dans l'épreuve B.
On va déterminer les coordonnées du point moyen des 5 premiers éléments des épreuves A et B, puis on déterminera les coordonnées du point moyen pour les 5 derniers éléments des épreuves A et B.

Il faut penser à **trier** les données, en **classant les abscisses des points dans l'ordre croissant** !

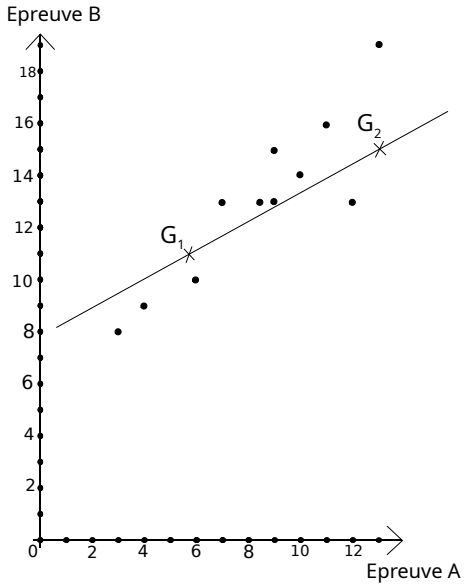
Partie 1 Épreuve A	3	4	6	7	9
Partie 1 Épreuve B	8	9	10	13	15

$$x_{G_1} = \frac{3+4+6+7+9}{5} = 5,8$$
$$y_{G_1} = \frac{8+9+10+13+15}{5} = 11$$
$$G_1(5,8;11)$$

Partie 2 Épreuve A	9	10	11	12	13
Partie 2 Épreuve B	13	14	16	13	19

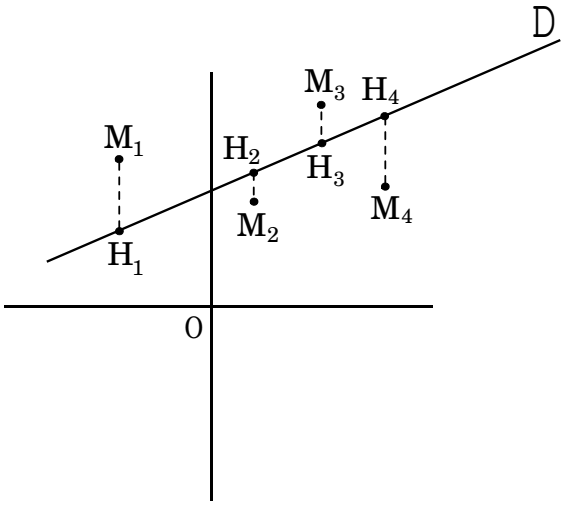
$$x_{G_2} = \frac{9+1+11+12+13=11}{5} = 11$$
$$y_{G_2} = \frac{13+14+16+13+19}{5} = 15$$
$$G_2(11;15)$$

On trace ensuite une droite passant par les points G_1 et G_2. On peut bien sûr en profiter pour calculer l'équation de cette droite, connaissant G_1 et G_2.
Dans notre exemple, nous obtenons le nuage de points suivant :



C) Méthode 3 : méthode des moindres carrés

Il s'agit d'obtenir une droite équidistante des points situés de part et d'autre d'elle-même. Pour réaliser ceci, on cherche à minimiser la somme des distances des points à la droite au carré.
On considère une série statistique à deux variables représentée par un nuage justifiant un ajustement affine (les points semblent alignés).



Définition : Dans le plan muni d'un repère orthonormal, on considère un nuage de n points de coordonnées $(x_i; y_i)$. La droite D d'équation $y = ax + b$ est appelée **droite de régression** de y en x (ou encore **droite des moindres carrés**) de la série statistique si et seulement si la quantité $M_1H1^2 + ... + M_nH_n^2$ est minimale, où $M_i(x_i; y_i)$ est un point de la série étudié (du nuage de points, donc) et $H_i(x_i, y_i)$ est un point de la droite cherchée D.

Cela revient donc à minimiser le nombre $\sum_{i=1}^{n}[y_i - (ax_i + b)]^2$

En pratique, la droite des moindres carrés a pour équation :

$$y = ax + b \text{ avec } a = \frac{cov(x,y)}{V(x)} \text{ et } b = \overline{y} - a \times \overline{x}.$$

$$\text{avec } cov(x,y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) \text{ ou encore } cov(x,y) = \frac{x_1y_1 + ... + x_ny_n}{n} - \overline{x} \times \overline{y}$$

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

Remarques :
1) $cov(x, y)$ est appelée la covariance, elle peut être positive ou négative. On note aussi $cov(x, y) = \sigma_{xy}$.
2) $V(x)$ est appelée la variance ; elle est forcément positive.
$\sigma(x)$ est l'écart type et $\sigma(x) = \sqrt{V(x)}$.
3) En pratique, on utilisera très souvent la calculatrice pour trouver les coefficients a et b qui détermine la droite des moindres carrés, pour trouver la covariance, etc. Chaque étudiant doit donc être très au clair avec l'utilisation de sa calculatrice.

On reprend notre exemple initial, et on va calculer l'équation de la droite de régression.

Nous avions trouvé $\overline{x} = 8.4$ et $\overline{y} = 13$

x_i (Épreuve A)	3	4	6	7	9	10	9	11	12	13
y_i (Épreuve B)	8	9	10	13	15	14	13	16	13	19
Écart $x_i - \overline{x}$	-5,4	-4,4	-2,4	-1,4	0,6	1,6	0,6	2,6	3,6	4,6
Écart $y_i - \overline{y}$	-5	-4	-3	0	2	1	0	3	0	6
Produit $(x_i - \overline{x})(y_i - \overline{y})$	27	17,6	7,2	0	1,2	1,6	0	7,8	0	27,6

La covariance est obtenue par

$$cov(x, y) = \frac{1}{10}\sum_{i=1}^{10}(x_i - \overline{x})(y_i - \overline{y}) = \frac{27 + 17,6 + 7,2 + 0 + 1,2 + 1,6 + 0 + 7,8 + 0 + 27,6}{10} = 9$$

La variance est obtenue par

$$V(x) = \frac{1}{10}\sum_{i=1}^{10}(x_i - \overline{x})^2 = \frac{(-5,4)^2 + (-4,4)^2 + (-2,4)^2 + (-1,4)^2 + (0,6)^2 + (1,6)^2 + (0,6)^2 + (2,6)^2 + (3,6)^2 + (4,6)^2}{10}$$

$$V(x) = \frac{1}{10} \times 100,4 = 10,04$$

On trouve $a = \dfrac{cov(x, y)}{V(x)} = \dfrac{9}{10,04} = \dfrac{225}{251} \approx 0,896$ et $b = \overline{y} - a \times \overline{x} = 13 - \dfrac{225}{251} \times 8,4 = \dfrac{1373}{251} \approx 5,47.$

L'équation de la droite de régression est donc $y = \dfrac{225}{251}x + \dfrac{1373}{251}$

En pratique, on trace la droite d'équation $y = 0,896x + 5,47$ et on vérifie que cette droite de régression est cohérente.

D) Autres méthodes

D'autres méthodes existent, lorsqu'un ajustement affine n'est pas possible (c'est à dire lorsque les points ne sont clairement pas alignés), mais ces méthodes ne seront pas abordés cette année (ajustement par une fonction exponentielle, ajustement par une fonction carrée, etc), par cause de manque de temps en ces circonstances exceptionnelles.

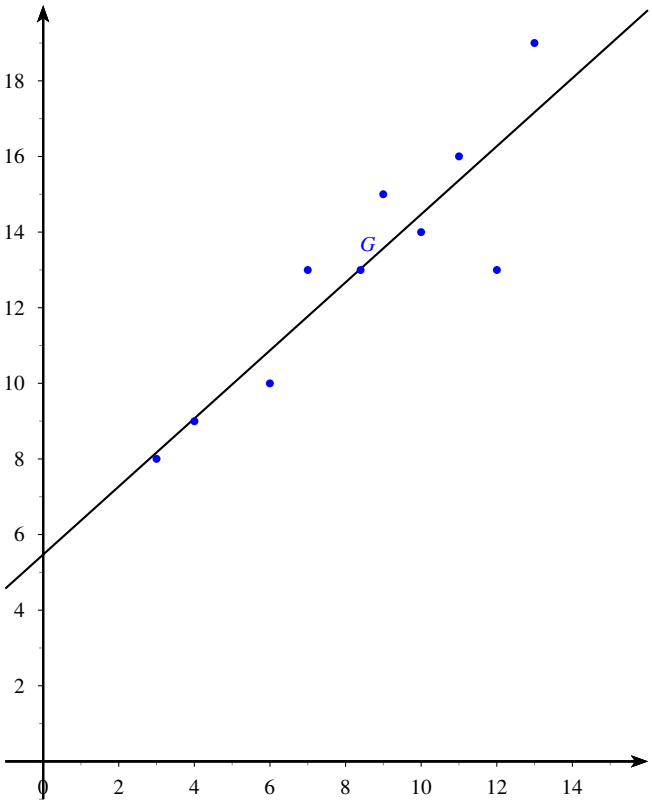
E) Coefficient de corrélation linéaire

Le **coefficient de corrélation linéaire** des variables x et y est le nombre r défini par

$$r = \frac{cov(x,y)}{\sigma(x)\sigma(y)}$$

Propriétés :
1) On a toujours : $-1 \leqslant r \leqslant 1$
2) Il y a égalité (c'est à dire r=1 ou r =-1) si et seulement si les points sont alignés (la droite de régression passe par tous les points).
3) Plus $|r|$ est proche de 1, plus l'ajustement affine est un bon modèle de corrélation entre les variables x et y.
4) Plus $|r|$ est proche de 0, moins l'ajustement n'a de sens.

Dans notre exemple, nous avions trouvé $cov(x,y) = 9$.
De plus $\sigma(x) = \sqrt{V(x)} = \sqrt{10,04} \approx 3,169$ et $\sigma(y) = \sqrt{V(y)} = \sqrt{10} \approx 3,162$.
On otbtient donc $r \approx \dfrac{9}{3,169 \times 3,162} \approx 0,898$, qui est un coefficient assez proche de 1. L'approche par la méthode moindres carrés est donc pertinente dans cet exemple.



On remarquera que la droite de régression passe par le point moyen $G(\overline{x};\overline{y})$.

On peut maintenant répondre à la question initiale : y a t-il un lien de causalité entre ces deux séries (Épreuve A et Épreuve B) ?
Dans notre exemple il y a une corrélation nette entre les deux séries (car elles semblent liées par une fonction, ici une fonction affine), il y a donc probablement une causalité entre le fait d'avoir travaillé et les résultats obtenus, la progression des notes a été linéaire.