SÉRIES STATISTIQUES À DEUX VARIABLES

1) POSITION DU PROBLÈME - VOCABULAIRE

A) DÉFINITION

Définition :

On considère deux variables statistiques numériques x et y observées sur une même population de n individus.
On note $x_1 ; x_2 ; \ldots ; x_n$ les valeurs relevées pour la première variable et $y_1 ; y_2 ; \ldots ; y_n$ les valeurs relevées pour la deuxième variable.
Les couples $(x_1 ; y_1) ; (x_2 ; y_2) ; \ldots ; (x_n ; y_n)$ forment une **série statistique à deux variables**.

Pour la suite du cours, on garde les notations ci-dessus et on considère l'exemple ci-dessous :

Exemple :

Le tableau suivant donne l'évolution du nombre d'adhérents d'un club de basket de 2015 à 2020.

Année	2015	2016	2017	2018	2019	2020
Rang x_i	1	2	3	4	5	6
Nombre d'adhérents y_i	70	90	115	140	170	220

Le but est d'étudier cette série statistique à deux variables (le rang et le nombre d'adhérents) afin de prévoir l'évolution du nombre d'adhérents pour les années suivantes.

B) NUAGE DE POINTS

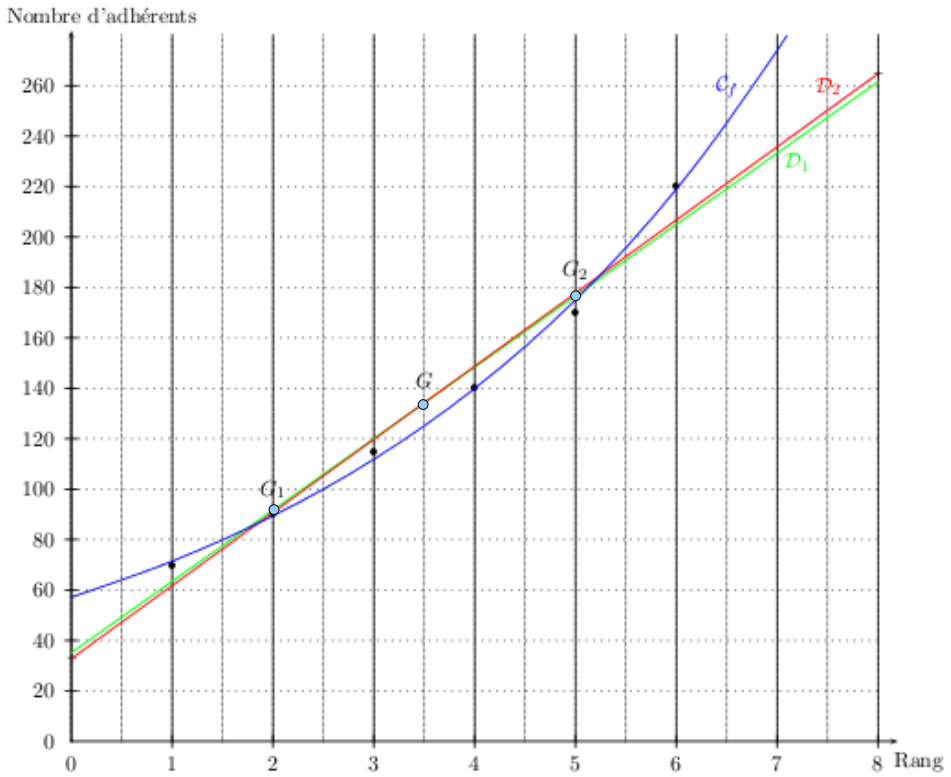
La première étape consiste à réaliser un graphique qui traduise les deux séries statistiques.

Définition :

Dans le plan rapporté à un repère orthogonal, on appelle **nuage de points** associé à cette série statistique à deux variables, l'ensemble des points $M_1(x_1 ; y_1)$; $M_2(x_2 ; y_2)$;...; $M_n(x_n ; y_n)$.

Dans notre exemple, si on place le rang en abscisses, et le nombre d'adhérents en ordonnées, on peut représenter par un point chaque valeur . On obtient ainsi une succession de points, dont les coordonnées (1; 70), (2; 90), ... (6; 220), forment un nuage de points.

Exemple - question 1 : Représenter le nuage de points associé à la série . (Graphique à compléter au fur et à mesure du cours)

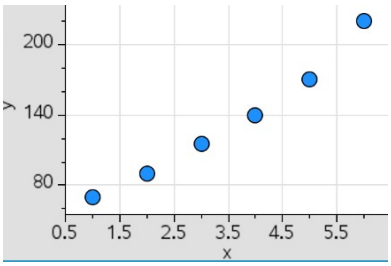


Avec une calculatrice :

Ti-nspire : https://www.youtube.com/watch?v=Nb2hmeoaMLo

Ti 83 : https://www.youtube.com/watch?v=lC1lfgYE51s

Casio : https://www.youtube.com/watch?v=xL5ixxLnING



Remarque :

Le nuage de points associé à une série statistique à deux variables donne donc immédiatement des informations de nature qualitative.
Pour en tirer des informations plus quantitatives, il nous faut poser le problème de l'ajustement.
Le tracé met en évidence la possibilité de "reconnaître" graphiquement la possibilité d'une relation fonctionnelle entre les deux grandeurs observées (ici rang et nombre d'adhérents).
Le problème de l'établissement d'une relation fonctionnelle entre les deux séries est **le problème de l'ajustement**.

C) POINT MOYEN

Définition :

On appelle **point moyen** de cette série le point G de coordonnées $(\overline{x}\,;\,\overline{y})$ où $\overline{x}$ et $\overline{y}$ sont les moyennes respectives des séries $x_1\,;\,x_2\,;\ldots\,x_n$ et $y_1\,;\,y_2\,;\ldots\,y_n$.

Exemple - question 2 : Déterminer les coordonnées des points moyens suivants :

- G_1 des années allant de 2015 à 2017 : $G_1(2\,;\,91{,}7)$
- G_2 des années allant de 2018 à 2020 : $G_2(5\,;\,176{,}7)$
- G, point moyen du nuage de points tout entier : $G(3{,}5\,;\,134{,}2)$

2) AJUSTEMENTS

A) À LA RÈGLE

On se propose, à partir des résultats obtenus, de faire des prévisions pour les années à venir.
Un moyen d'y parvenir est de tracer au juger une droite d passant le plus près possible des points du nuage et d'en trouver l'équation du type $y=ax+b$.

B) MÉTHODE DE MAYER

Cet ajustement consiste à déterminer la droite passant par deux points moyens du nuage de points.

Exemple - question 3 :

Déterminer l'équation de la droite d_1 qui passe par les points moyens G_1 et G_2 et la tracer sur le graphique précédent.

La droite d_1 n'est pas parallèle à l'axe des ordonnées, elle admet donc une équation de la forme $y=ax+b$ avec :

$$a=\frac{176{,}7-91{,}7}{5-2}\approx 28{,}3$$

On choisit $a=28{,}3$, et pour la suite du cours et les exercices si l'approximation est correcte, on choisira « y » au lieu de « $\approx$ »

De plus, elle passe par le point $G_1(2\,;\,91{,}7)$ d'où :

$$y_{G_1}=a\,x_{G_1}+b \Leftrightarrow 91{,}7=28{,}3\times 2+b \Leftrightarrow b=35{,}1$$

Conclusion : $d_1 : y=28{,}3\,x+35{,}1$.

Pour tracer d_1 , il suffit de placer G_1 et G_2 puis de tracer la droite qui les relie.

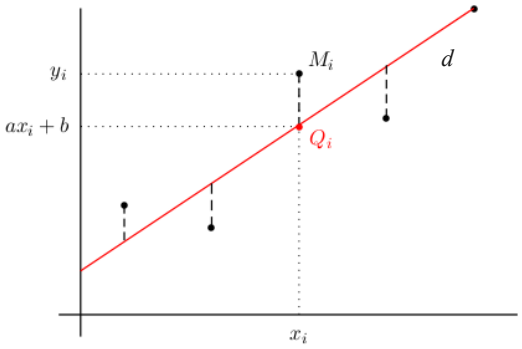
C) MÉTHODE DES MOINDRES CARRÉS

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 carrés des distances des points aux points de même abscisse de la droite.

Définition :

Dans le plan muni d'un repère orthonormal, on considère un nuage de n points de coordonnées $(x_i\,;y_i)$ justifiant un ajustement affine.
La droite d d'équation $y=ax+b$ est appelée **droite de régression** de y en x de la série statistique si et seulement si la quantité suivante est minimale :

$$\sum_{i=1}^{n}(M_i\,Q_i)^2=\sum_{i=1}^{n}(y_i-(a\,x_i+b))^2$$



Remarque :

Il serait tout aussi judicieux de s'intéresser à la droite d' qui minimise la quantité $\sum_{i=1}^{n}(x_i-(a\,y_i+b))^2$

Cette droite est appelée droite de régression de x en y.

Définition :

On appelle **covariance** notée $\text{cov}\,(x\,;y)$ ou σ_{xy} de la série statistique double de variables x et y le nombre réel :

$$\text{cov}\,(x\,;y)=\sigma_{xy}=\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})$$

Pour les calculs, on pourra aussi utiliser :

$$\sigma_{xy}=\frac{1}{n}\sum_{i=1}^{n}x_i\,y_i-\overline{x}\,\overline{y}$$

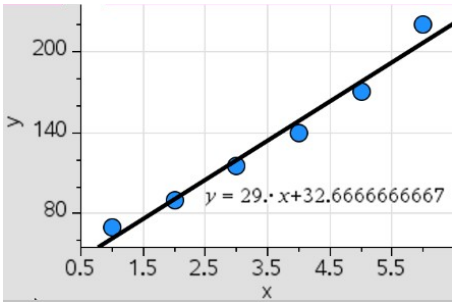
Remarque : On a $\text{cov}\,(x\,;x)=\sigma_{x^2}=\text{V}\,(x)=(\sigma\,(x))^2$

Propriété :

La droite de régression d de y en x a pour équation $y=ax+b$ où :

$$a=\frac{\sigma_{xy}}{\sigma_{x^2}}\quad\text{et } b \text{ vérifie } \overline{y}=a\,\overline{x}+b$$

Avec une calculatrice :



Propriété :

Le point moyen G du nuage appartient toujours à la droite de régression de y en x.

Exemple - question 4 : Déterminer avec la calculatrice une équation de la droite d'ajustement d_2 de y en x obtenue par la méthode des moindres carrés et la tracer sur le graphique précédent.

La calculatrice donne $d_2 : y=29\,x+32{,}7$

Pour tracer la droite d_2 , il faut choisir deux points (au moins) sur cette droite.

Par exemple :

x	0	8
y	32,7	264,7

D) AJUSTEMENT SE RAMENANT À UN AJUSTEMENT AFFINE

On remarque qu'un ajustement affine ne semble pas très approprié pour ce nuage de points à partir de 2020,
On se propose de déterminer un ajustement plus juste. (ici, un ajustement exponentiel)

Exemple - question 5 :

On pose $z = \ln y$. Compléter le tableau suivant en arrondissant les valeurs de z_i au millième.

x_i	1	2	3	4	5	6
z_i	4,248	4,500	4,745	4,942	5,136	5,394

Exemple - question 6 :

Déterminer une équation de la droite d'ajustement d_3 de z en x obtenue par la méthode des moindres carrés.

La manipulation à la calculatrice est la même que précédemment, en n'oubliant pas de changer les paramètres.

La calculatrice donne $z = 0,224x + 4,045$

Exemple - question 7 :

Dans ce cas, en déduire la relation qui lie y à x puis tracer la courbe représentative de la fonction $y = f(x)$.

On a $\begin{cases} z = 0,224\ x + 4,045 \\ z = \ln\ y \end{cases}$

On a donc : $\ln\ y = 0,224\ x + 4,045$

On obtient : $y = e^{0,224\ x + 4,045}$

Pour tracer la courbe, il suffit de placer des points, par exemple grâce au tableau de valeurs de la calculatrice.

E) COMPARAISON

Grâce aux trois derniers ajustements, on peut évaluer ce qui se passera plus tard, comparons les :

Exemple - question 8 :

En supposant que les ajustements restent valables pour les années suivantes, donner une estimation du nombre d'adhérents en 2021 suivant les trois méthodes.

Dans tous les cas, il faut calculer y lorsque x correspond à l'année 2021, c'est à dire au rang 7.

- **Méthode de Mayer :** $y = 28,3 \times 7 + 35,1 = 233,2$ soit environ 233 adhérents .
- **Ajustement affine :** $y = 29 \times 7 + 32,7 = 235,7$ soit environ 236 adhérents .
- **Ajustement exponentiel :** $y = e^{0,224 \times 7 + 4,045} \approx 274$ soit environ 274 adhérents .

Exemple - question 9 : En 2021, il y a eu 280 adhérents. Lequel des trois ajustements semble le plus pertinent ?

Le troisième ajustement semble le plus pertinent puisqu'il se rapproche le plus de la réalité.

Définitions :

On parle d'**interpolation** pour des valeurs à l'intérieur de la plage des valeurs observées et d'**extrapolation** pour des valeurs à l'extérieur de cette plage.

Bien entendu, les résultats obtenu par interpolation et par extrapolation sont à exploiter avec prudence.

3) COEFFICIENT DE CORRÉLATION LINÉAIRE

Définition :

Le **coefficient de corrélation linéaire** d'une série statistique de variables x et y est le nombre r défini par :

$$r = \frac{\sigma_{xy}}{\sigma(x) \times \sigma(y)}$$

Ce coefficient sert à mesurer la qualité d'un ajustement affine.

Interprétation graphique :

Plus le coefficient de corrélation linéaire est proche de 1 en valeur absolue, meilleur est l'ajustement affine.

Si $r \leqslant -0,75$ ou $r \geqslant 0,75$, on dit que la corrélation linéaire entre les séries x et y est forte.

Lorsque $r = \pm 1$, la droite de régression passe par tous les points du nuage, qui sont donc alignés.

Exemple - question 10 :

Déterminer le coefficient de corrélation linéaire dans le cas de l'ajustement affine (entre x et y), puis exponentiel (entre x et z). Quel est l'ajustement le plus juste ?

Grâce à la calculatrice, on trouve :

ajustement affine : $r_2 = 0,987$

ajustement exponentiel : $r_3 = 0,99$

Ce qui est conforme à ce que nous avions déduit précédemment, à savoir que l'ajustement exponentiel est plus fiable pour ce cas.

Propriété :

Le coefficient de corrélation linéaire r vérifie $-1 \leqslant r \leqslant 1$.

Remarques :

- Il ne faut pas confondre corrélation et causalité . Une forte corrélation entre deux variables ne signifie pas que l'une est la cause de l'autre ou qu'il y a un lien de cause à effet entre les deux variables.

- Si le coefficient de corrélation est proche de 0, cela signifie que le nuage de points ne peut pas être ajusté au mieux par une droite. Il faut dans ce cas essayer un autre type de courbe.