Statistiques à deux variables

Table des matières

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 texte en bleu concerne les calculatrices (TI et Casio)

I Position du problème. Vocabulaire

Par soucis de clarté, ce cours est élaboré à partir de l'exemple suivant :

Exemple
Le tableau suivant donne l'évolution du nombre d'adhérents d'un club de rugby de 2001 à 2006.

Année	2001	2002	2003	2004	2005	2006
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.

I.1 Nuage de points

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

Définition 1
Soit X et Y deux variables statistiques numériques observées sur n individus.
Dans un repère orthogonal $(O;\vec{i};\vec{j})$, l'ensemble des n points de coordonnées (x_i, y_i) forme le nuage de points associé à cette série statistique.

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 sont $(1;70)$, $(2;90)$, ... $(6;220)$, forment un nuage de points.

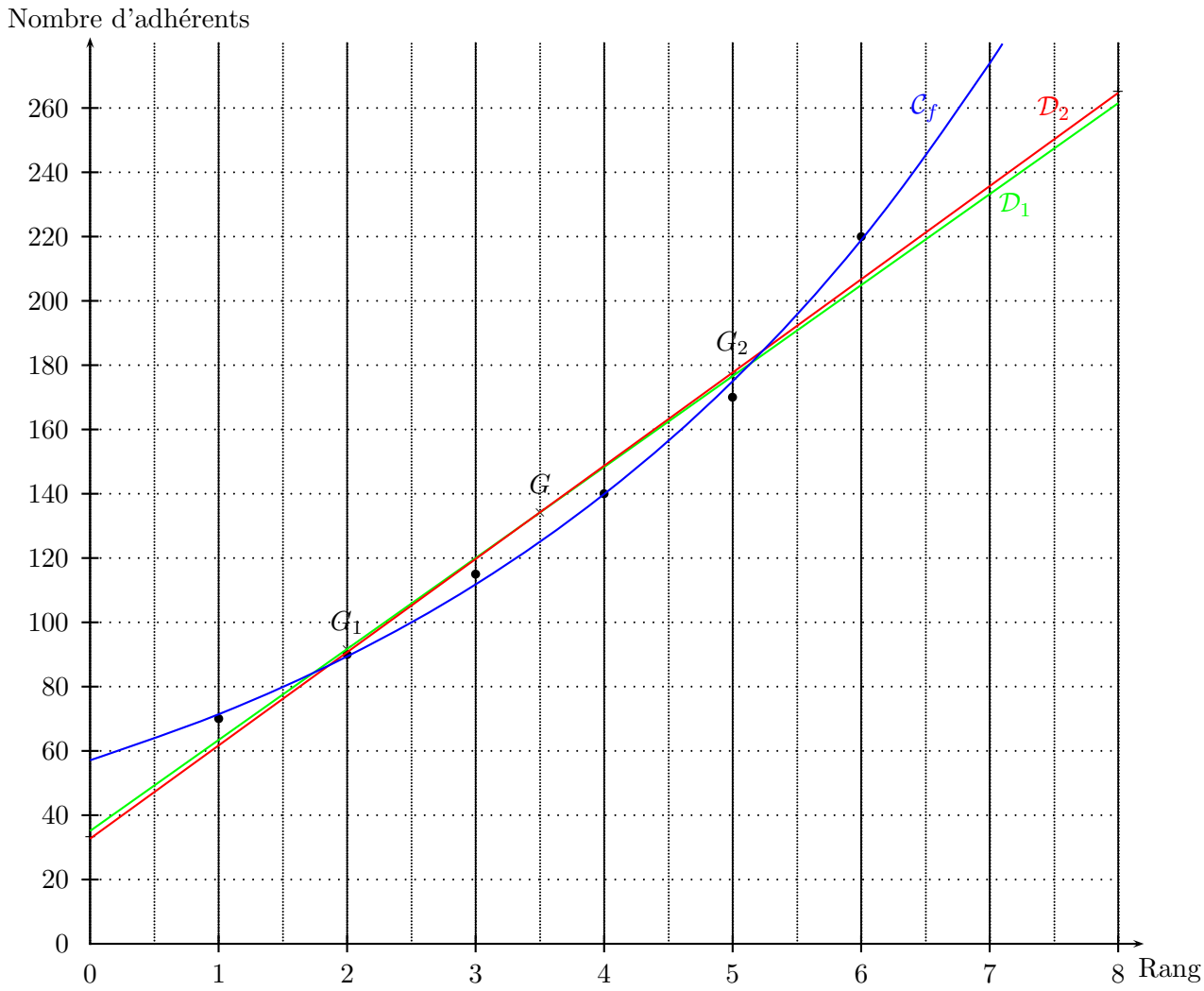
Question 1
Dans le plan muni d'un repère orthogonal d'unités graphiques : 2 cm pour une année sur l'axe des abscisses et 1 cm pour 20 adhérents sur l'axe des ordonnées, représenter le nuage de points associé à la série $(x_i;y_i)$.

T.I.

- Touche STAT
- Menu EDIT
- Entrer les valeurs x_i dans L_1
- Entrer les valeurs y_i dans L_2
- Règler les valeurs du repère avec la touche WINDOWS
- Appuyer sur la touche TRACE

Casio

- Menu STAT
- Entrer les valeurs x_i dans $List1$
- Entrer les valeurs y_i dans $List2$
- Choisir GRPH
- Règler les paramètres avec SET
- Choisir GPH1



I.2 Le problème de l'ajustement

Le nuage de points associé à une série statistique à deux variables donne donc immédiatement des informations de nature qualitatives.
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'adhérent).
Le problème de l'établissement d'une relation fonctionnelle entre les deux séries est le problème de l'ajustement.

I.3 Point moyen

Définition 2
Soit une série statistique à deux variables, X et Y, dont les valeurs sont des couples $(x_i; y_i)$.
On appelle point moyen de la série le point G de coordonnées

- $x_G = \dfrac{x_1 + x_2 + \cdots + x_n}{n}$.
- $y_G = \dfrac{y_1 + y_2 + \cdots + y_n}{n}$.

Question 2
Déterminer les coordonnées des points moyens suivants :

- G_1 des années allant de 2001 à 2003,
- G_2 des années allant de 2004 à 2006,
- G, point moyen du nuage de points tout entier.

Calcul des coordonnées de G_1 : $\left\{ \begin{array}{rcccl} x_{G_1} & = & \frac{1+2+3}{3} & = & 2 \\ y_{G_1} & = & \frac{70+90+115}{3} & = & 91,7 \end{array} \right.$ donc, $\boxed{G_1(\ 2\ ;\ 91,7\ )}$.

Calcul des coordonnées de G_2 : $\left\{ \begin{array}{rcccl} x_{G_2} & = & \frac{4+5+6}{3} & = & 5 \\ y_{G_2} & = & \frac{140+170+220}{3} & = & 176,7 \end{array} \right.$ donc, $\boxed{G_2(\ 5\ ;\ 176,7\ )}$.

Calcul des coordonnées de G : $\left\{ \begin{array}{rcccl} x_{G} & = & \frac{1+2+3+4+5+6}{6} & = & 3,5 \\ y_{G} & = & \frac{70+90+115+140+170+220}{3} & = & 134,2 \end{array} \right.$ donc, $\boxed{G(\ 3,5\ ;\ 134,2\ )}$.

II Ajustements

II.1 Ajustement à la règle

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

II.2 Méthode de Mayer

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

Question 3
Déterminer l'équation de la droite $\mathcal{D}_1$ qui passe par les points moyens G_1 et G_2 et la tracer sur le graphique précédent.

La droite $\mathcal{D}_1$ n'est pas parallèle à l'axe des ordonnées, elle a donc pour équation $y = ax + b$ avec :
$a = \frac{y_{G_2} - y_{G_1}}{x_{G_2} - x_{G_2}} = \frac{176,7-91,7}{5-2} = 28,3.$

De plus, elle passe par le point $G_1(\ 2\ ;\ 91,7\ )$ d'où :
$y_{G_1} = ax_{G_1} + b \quad \Rightarrow \quad 91,7 = 28,3 \times 2 + b \quad \Rightarrow \quad b = 35,1.$

Conclusion : $\boxed{\mathcal{D}_1 : y = 28,3x + 35,1}$.

Pour tracer $\mathcal{D}_1$, il suffit de placer G_1 et G_2 puis de tracer la droite qui les relie.

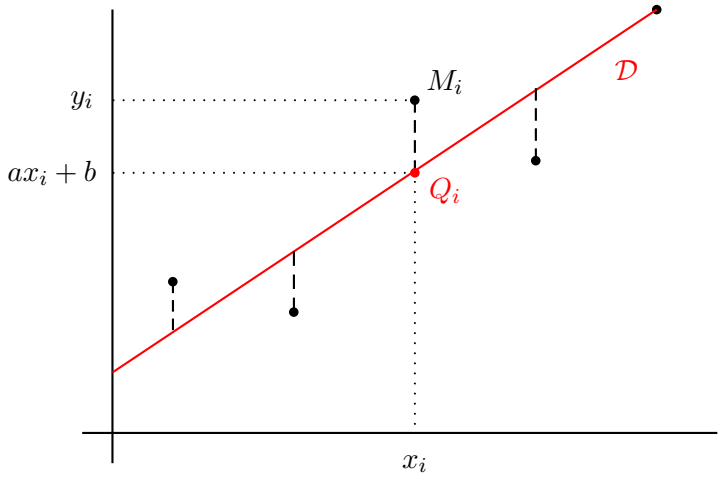
II.3 Méthode des moindres carrés

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.

Définition 3
Dans le plan muni d'un repère orthonormal, on considère un nuage de n points de coordonnées $(x_i; y_i)$. La droite $\mathcal{D}$ d'équation $y = ax + b$ est appelée droite de régression de y en x de la série statistique ssi la quantité suivante est minimale :

$$\sum_{i=1}^{n} (M_iQ_i)^2 = \sum_{i=1}^{n}[y_i - (ax_i + b)]^2$$



Remarque 1
Il serait tout aussi judicieux de s'intéresser à la droite $\mathcal{D}'$ qui minimise la quantité $\sum_{i=1}^{n}[x_i - (ay_i + b)]^2$.

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

Définition 4
On appelle covariance de la série statistique double de variables x et y le nombre réel

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

Pour les calculs, on pourra aussi utiliser :

$$\sigma_{xy} = \frac{1}{n}\sum_{i=1}^{n} x_iy_i - \bar{x}\bar{y}.$$

Remarque 2
On a : $cov(x, x) = \sigma_{x^2} = V(x) = [\sigma(x)]^2$.

Propriété 1
La droite de régression D de y en x a pour équation $y = ax + b$ où

$$\begin{cases} a = \dfrac{\sigma_{xy}}{[\sigma(x)]^2} \\ b \quad \text{vérifie} \quad \bar{\text{y}} = \text{a}\bar{\text{x}} + \text{b}. \end{cases}$$

Remarque 3
Les réels a et b sont donnés par la calculatrice.

T.I.

- Touche STAT
- Menu CALC
- Item LinReg
- LinReg L_1, L_2

Casio

- Menu STAT
- Item CALC
- Règler les paramètres avec set
- Item REG
- Choisir X

Propriété 2
Le point moyen G du nuage appartient toujours à la droite de régression de y en x.

Question 4
Déterminer une équation de la droite d'ajustement $\mathcal{D}_2$ de y en x obtenue par la méthode des moindres carrés et la tracer sur le graphique précédent.

La calculatrice donne $\mathcal{D}_2 : y = ax + b$ avec $a = 29$ et $b = 32,7$.

Conclusion : $\boxed{\mathcal{D}_2 : y = 29x + 32,7}$

Pour tracer la droite $\mathcal{D}_2$, il faut choisir deux points (au moins) sur cette droite.

Par exemple :

x	0	8
y	32,7	264,7

, les placer dans le repère puis tracer la droite.

II.4 Ajustement exponentiel

On remarque qu'un ajustement affine ne semble pas très approprié pour ce nuage de points à partir de 2006, on se propose de déterminer un ajustement plus juste.

Question 5
On pose $z = \ln y$. Recopier et 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					

Il suffit de calculer $\ln y_i$ pour chaque caleur de i :

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

On peut déterminer les éléments de ce tableau grâce à la calculatrice :

T.I.

- Touche STAT
- Menu EDIT
- Se placer dans L_3
- Entrer la formule "$= \ln L_2$"

Casio

- Touche STAT
- Menu EDIT
- Se placer dans $List3$
- Entrer la formule "$= \ln List2$"

Question 6
Déterminer une équation de la droite d'ajustement $\mathcal{D}_3$ de z en x obtenue par la méthode des moindres carrés.

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

La calculatrice donne $\mathcal{D}_3 : z = ax + b$ avec $a = 0,224$ et $b = 4,045$.

Conclusion : $\boxed{\mathcal{D}_3 : z = 0,224x + 4,045}$.

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 $\begin{cases} z &= 0,224x + 4,045 \\ z &= \ln y \end{cases}$ donc : $\ln y = 0,224x + 4,045$

On compose par la fonction exponentielle : $\begin{aligned} e^{\ln y} &= \mathrm{e}^{0,224x+4,045} \\ &= (\mathrm{e}^{0,224})^x \times e^{4,045} \\ &= (1,251)^x \times 57,111 \end{aligned}$

Conclusion : $\boxed{y = 57,111 \times 1,251^x}$.

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

II.5 Comparaison

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

Question 8
En supposant que les ajustements restent valables pour les années suivantes, donner une estimation du nombre d'adhérents en 2007 suivant les trois méthodes.

Dans tous les cas, il faut calculer y lorsque x correspond à l'année 2007, c'est à dire au rang 7.

- Méthode de Mayer : $y = 28,3 \times 7 + 35,1 = 233,2$ soit environ $\boxed{233 \text{ adhérents}}$.
- Ajustement affine : $y = 29 \times 7 + 32,7 = 235,7$ soit environ $\boxed{236 \text{ adhérents}}$.
- Ajustement exponentiel : $y = 57,112 \times 1,024^7 = 273,9$ soit environ $\boxed{274 \text{ adhérents}}$.

Question 9
En 2007, il y a eu 280 adhérents. Lequel des trois ajustements semble le plus pertinent ?

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

III Coefficient de corrélation linéaire

Définition 5
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 régression linéaire est proche de 1 en valeur absolue, meilleur est l'ajustement linéaire. Lorque $r = \pm 1$, la droite de régression passe par tous les points du nuage, qui sont donc alignés.

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'ajustemet le plus juste ?

Grâce à la calculatrice, on trouve successivement $r_2 = 0,987$ puis $r_3 = 0,999$.
Ce qui est conforme à ce que nous avions déduit précédemment, à savoir que l'ajustement exponentiel est plus fiable pour ce cas.

Propriété 3
Le coefficient de corrélation linéaire r vérifie $-1 \leq r \leq 1$.