Chapitre 6
Statistique à deux variables

I. Nuage de points

1) Définitions

Définition :

Sur une population donnée, on s'intéresse à deux **caractères quantitatifs** x et y.

Pour chacun des n individus de cette population, notons x_i et y_i les valeurs prises par chacun de ces caractères $(1 \leqslant i \leqslant n)$.

Les n couples $(x_i ; y_i)$ forment une **série statistique à deux variables**.

Exemple :

On mesure le poids x_i (en kg) et la tension artérielle systolique y_i (en mmHg) de 7 individus du même âge.

On a donc la série statistique à deux variables :

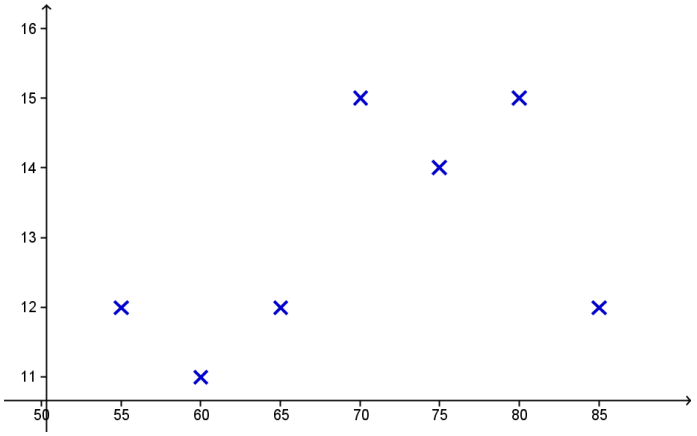
x_i	55	75	60	85	70	80	65
y_i	12	14	11	12	15	15	12

Définition :

Dans un repère orthogonal, l'ensemble des points $M_i(x_i ; y_i)$ est appelé le **nuage de points** associé à cette série statistique à deux variables.

Exemple :

Le graphique ci-contre représente le nuage de points associé à la série statistique de l'exemple précédent.



2) Point moyen

Définition :

Le point G de coordonnées $(\overline{x};\overline{y})$ avec :

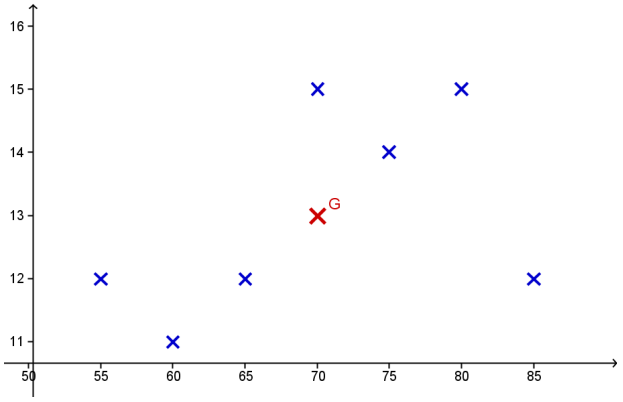
$$\overline{x}=\frac{x_1+x_2+\ldots+x_n}{n}=\frac{\sum_{i=1}^{n} x_i}{n}=\frac{1}{n}\sum_{i=1}^{n} x_i \qquad \text{et} \qquad \overline{y}=\frac{1}{n}\sum_{i=1}^{n} y_i$$

est appelé le **point moyen** du nuage de points associé à cette série statistique à deux variables.

Exemple :

Le point moyen du nuage ci-dessus est :

$$G(70;13)$$



3) Covariance

Définition :

On appelle **covariance** de x et y le nombre σ_{xy} ou $cov(x;y)$, et défini par :

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

Remarques :

- La **variance** $V(X)=\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^2$ qui sert à caractériser la dispersion d'un échantillon (et à calculer l'**écart type** $\sigma=\sqrt{V}$) vérifie donc $V(X)=\sigma_{xx}$.
- La covariance est un nombre permettant d'évaluer le sens de variation (et ainsi la dépendance) de deux variables statistiques.
- On peut aussi utiliser : $\sigma_{xy}=\frac{1}{n}\sum_{i=1}^{n} x_i y_i-\overline{x}\,\overline{y}$.

Cette formule sera intéressante pour calculer la covariance à partir de la calculatrice.

Exemple :

Dans le cas de la série statistique de l'exemple :

- Pour la série « poids » $X=\{55\ ;\ 75\ ;\ 60\ ;\ 85\ ;\ 70\ ;\ 80\ ;\ 65\}$ on a :

  $\overline{x}=70\quad ;\quad V(X)=100\quad ;\quad \sigma_x=10$

- Pour la série « tension artérielle » $Y=\{12\ ;\ 14\ ;\ 11\ ;\ 12\ ;\ 15\ ;\ 15\ ;\ 12\}$ on a :

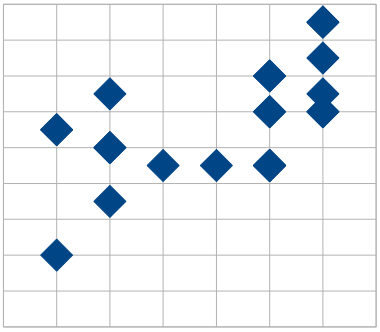
  $\overline{y}=13\quad ;\quad V(Y)=\frac{16}{7}\approx 2{,}29\quad ;\quad \sigma_y=\sqrt{\frac{16}{7}}\approx 1{,}51$

- Pour la série à deux variables X et Y : $\sigma_{xy}=\frac{50}{7}\approx 7{,}14$

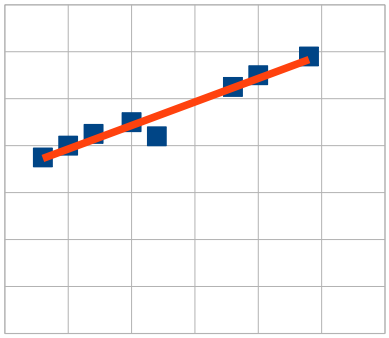
II. Ajustement

1) Introduction

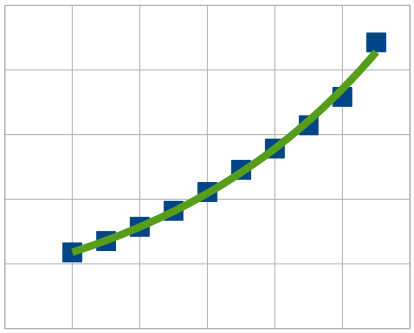
Lorsque deux variables X et Y sont liées l'une à l'autre, l'étude de la forme du nuage de points permet de modéliser cette relation.



Pas de relation



Relation affine



Relation exponentielle

Définition :

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

Remarque :

Pour les statisticiens, une formule telle que $y=f(x)$ est appelé un **modèle** (permettant d'étudier les relations entre les variables) : x est la variable explicative et y la variable expliquée.

2) Méthode des moindres carrés

Définition :

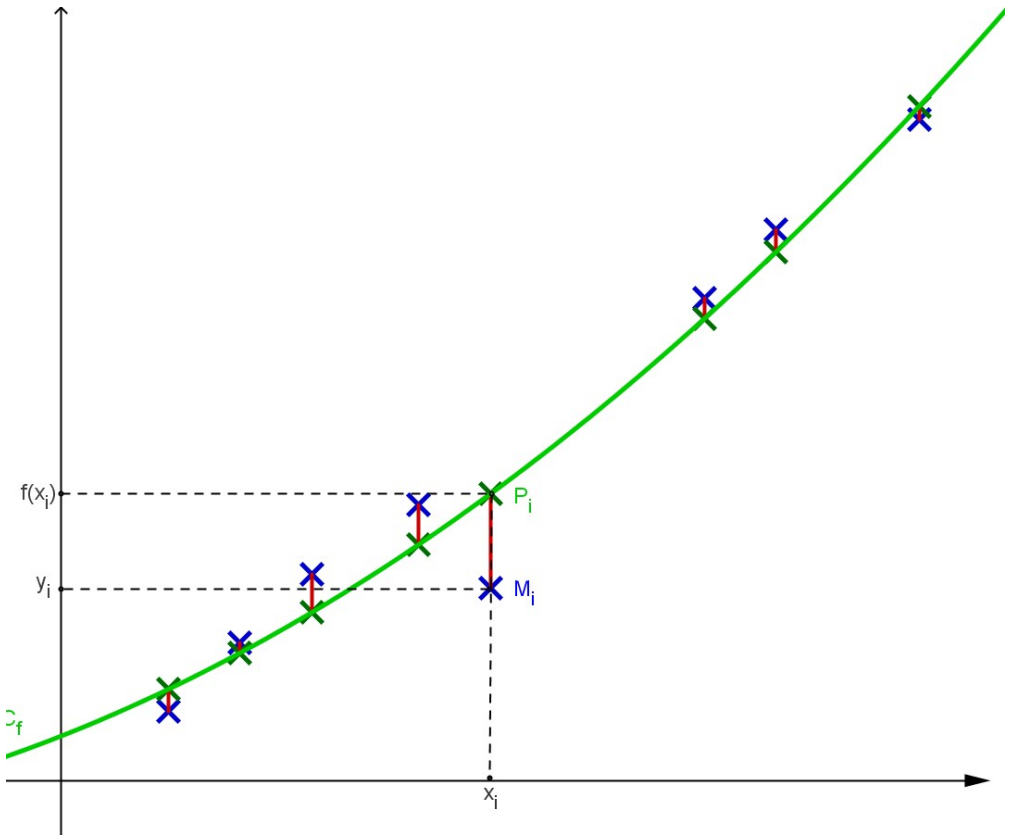
Effectuer un ajustement d'un nuage de points $(x_i\,;y_i)$ par la **méthode des moindres carrés** consiste à trouver une fonction f qui **minimise** la somme des carrés des écarts entre les valeurs y_i observées et les valeurs $f(x_i)$ données par le modèle.

La fonction f doit donc minimiser l'expression $\sum_{i=1}^{n}[y_i-f(x_i)]^2$.

Interprétation graphique :

$\mathscr{C}_f$ est la courbe représentative de la fonction f.

$M_i(x_i;y_i)$ est un point du nuage et $P_i(x_i;f(x_i))$ est un point de $\mathscr{C}_f$.



Alors $M_iP_i=|y_i-f(x_i)|$ et $(M_iP_i)^2=(y_i-f(x_i))^2$.

Donc $\mathscr{C}_f$ est la courbe qui minimise $\sum_{i=1}^{n}(M_iP_i)^2$, la somme des carrés des distances « verticales ».

Intérêt :

Pour une valeur donnée x_0 de la variable X, la fonction f permet de prévoir approximativement la valeur correspondante de Y (en calculant $f(x_0)$).

- Si x_0 appartient à l'intervalle d'observation des valeurs de X, on parle d'**interpolation**.
- Si x_0 n'appartient pas à l'intervalle d'observation des valeurs de X, on parle d'**extrapolation** (on suppose alors que le modèle est encore valable à l'extérieur de l'intervalle).

3) Ajustement affine par moindres carrés

Propriété :

Lors d'un ajustement affine de y en x par la méthode des moindres carrés, la droite d qui représente, dans un repère, la fonction f obtenue a pour coefficient directeur :

$$a=\frac{\sigma_{xy}}{V(X)}=\frac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sum_{i=1}^{n}(x_i-\overline{x})^2}$$ et passe par le point moyen $G=(\overline{x};\overline{y})$

Remarque :

- La droite d est appelée la droite d'ajustement de y en x par la méthode des moindres carrés.

  Une équation de d est :

  $$y=a(x-\overline{x})+\overline{y}$$

- On démontre que d est la seule droite pour laquelle la somme $\sum_{i=1}^{n}[y_i-(a\,x_i+b)]^2$ est minimale.

Exemple :

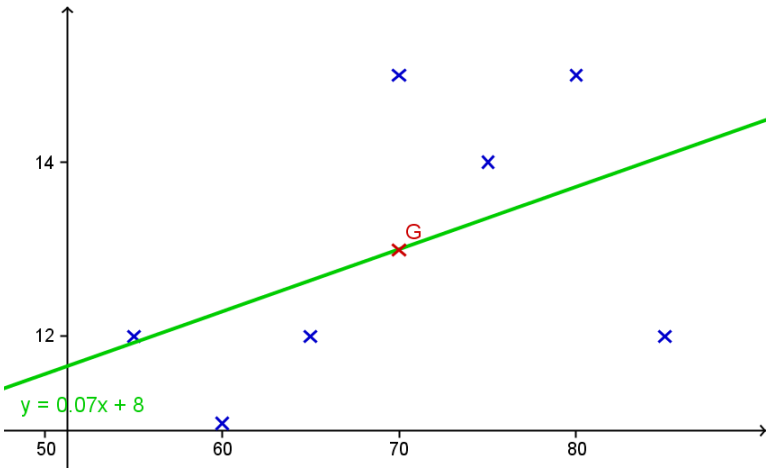
Pour la série :

x_i	55	75	60	85	70	80	65
y_i	12	14	11	12	15	15	12

On a donc :

$$a=\frac{\sigma_{xy}}{V(X)}=\frac{\frac{50}{7}}{100}=\frac{50}{700}\approx 0{,}0714$$

et nous avions $G(70;13)$.



La droite d'ajustement par la méthode des moindres carrés a pour équation :

$$y=\frac{50}{700}(x-70)+13=\frac{5}{70}x+8$$

III. Corrélation et ajustements

1) Coefficient de corrélation linéaire

Définition :

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

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

Remarques :

- $r = \dfrac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}}$
- $\sigma(x) > 0$ et $\sigma(y) > 0$, donc r a le même signe que σ_{xy}.

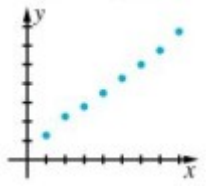
Propriété :

Pour toute série statistique à deux variables quantitatives, $-1 \leqslant r \leqslant 1$.

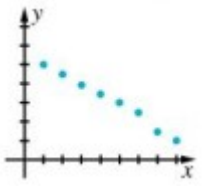
Propriétés :

- Plus $|r|$ est proche de 1, plus l'ajustement affine est un bon modèle de corrélation entre les variables x et y.
- Plus $|r|$ est proche de 0, moins l'ajustement affine a de sens.
- Si $|r| = 1$, la droite de régression passe par tous les points du nuage.

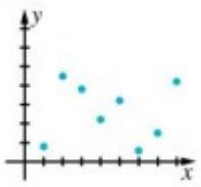
Exemple :



$r \approx 0{,}998$



$r \approx -0{,}994$



$r \approx 0{,}004$

Exemple :

Pour la série :

x_i	55	75	60	85	70	80	65
y_i	12	14	11	12	15	15	12

Le coefficient de corrélation linéaire vaut $r=\dfrac{\sigma_{xy}}{\sigma(x)\sigma(y)}=\dfrac{\frac{50}{7}}{10\times\sqrt{\frac{16}{7}}}=\dfrac{5\sqrt{7}}{28}\approx 0{,}47$.

L'ajustement affine n'est donc pas un bon modèle et il n'y a pas de corrélation forte entre x et y (le poids et la tension artérielle).

Remarque :

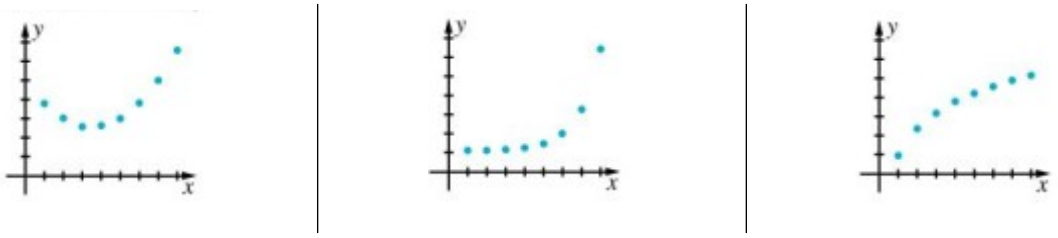
Une très forte corrélation peut exprimer un lien de cause à effet entre x et y, mais ce n'est pas toujours le cas.

Un exemple classique est celui d'une enquête réalisée en Angleterre de 1924 à 1937, révélant que le coefficient corrélation linéaire entre le nombre de permis délivrés chaque année pour l'installation d'un poste de radio et le nombre de malades mentaux dénombrés pour 10000 habitants était égal à 0,998, suggérant ainsi une relation quasiment fonctionnelle.

2) Changement de variables

Il peut arriver que le nuage de points associé à une série statistique à deux variables ait l'allure de la courbe d'une fonction autre qu'une fonction affine.

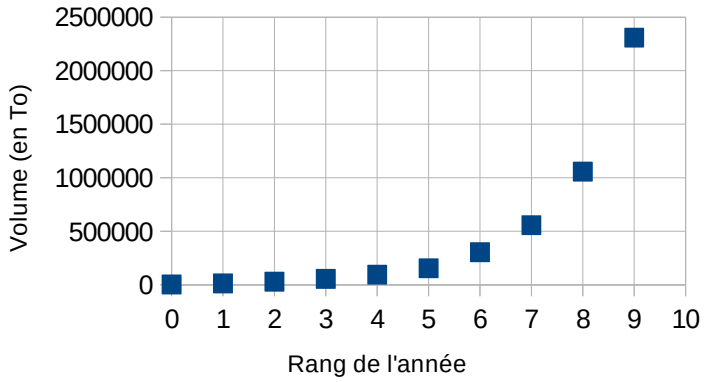
Par exemple, une fonction polynôme, une fonction exponentielle, une fonction logarithmique,...



Exemple :

Le tableau suivant donne l'évolution du volume de données mobiles consommées en France entre 2008 et 2017, en téraoctet.

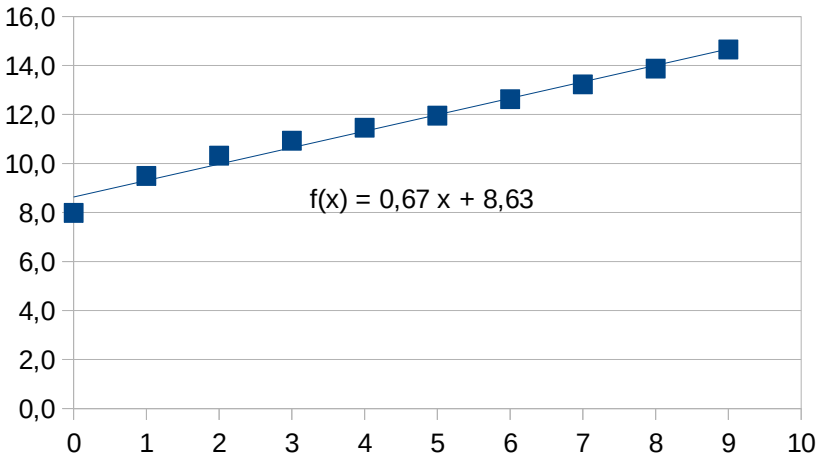
Année	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017
Rang x_i	0	1	2	3	4	5	6	7	8	9
Volume y_i	2930	13267	30331	55805	94999	155278	304471	557561	1057887	2308551



- Le nuage suggère qu'un ajustement affine n'est pas adapté, alors qu'un ajustement exponentiel de la forme $y = ke^{ax}$ pourrait convenir.
- On effectue le changement de variable $z = \ln(y)$ et on obtient le tableau de valeurs suivant, où les données sont arrondies à 0,1 près.

x_i	0	1	2	3	4	5	6	7	8	9
$z_i = \ln(y_i)$	8,0	9,5	10,3	10,9	11,5	12,0	12,6	13,2	13,9	14,7

On obtient alors l'équation de la droite de régression de z en x par la méthode des moindres carrés :



- On obtient donc $z = 0{,}67x + 8{,}63$ avec, pour coefficient de corrélation $r \approx 0{,}991$, qui est proche de 1. Cet ajustement affine est, donc, adapté.
- On revient à la variable y, en utilisant $z = \ln(y)$.

Donc $\ln(y) = 0{,}67x + 8{,}63$, c'est-à-dire $y = \mathrm{e}^{0{,}67x+8{,}63} = \mathrm{e}^{8{,}63}\,\mathrm{e}^{0{,}67x} = 5597\,\mathrm{e}^{0{,}67x}$.

On obtient donc l'ajustement $y = 5597\,\mathrm{e}^{0{,}67x}$.

- Cet ajustement peut être utilisé pour réaliser des prévisions.

Ainsi, en 2018, on peut estimer que le volume de données mobiles consommées en France était de $5597\,\mathrm{e}^{0{,}67\times 10}$ téraoctets, soit environ 4547000 téraoctets.