

## Statistiques à deux variables quantitatives

### I. Rappels sur les statistiques à une variable

#### I.1 Présentation d'une série statistique à une variable

- La **population** d'une série statistique est l'ensemble des éléments appelés **individus** sur lesquels porte l'étude statistique.
- Le caractère d'une série statistique est la propriété étudiée sur chaque individu. Il est dit :
  - **qualitatif** lorsqu'il ne prend pas que des valeurs numériques.
  - **quantitatif discret** lorsqu'il ne peut prendre qu'un nombre fini de valeurs numériques.
  - **quantitatif continu** lorsqu'il peut prendre une infinité de valeurs numériques.

- L'**effectif total** est le nombre d'éléments au sein de la population étudiée. On le note  $N$ .
- L'**effectif** d'une valeur du caractère est le nombre d'individus de la population prenant cette valeur (nombre de fois où cette valeur apparaît dans la série.). On le note  $n_i$ .
- La **fréquence** d'une valeur du caractère est le quotient de l'effectif de cette valeur par l'effectif total. On la note  $f_i$  et on a  $f_i = \frac{n_i}{N}$  (on multiplie par 100 si on veut un résultat en pourcentage).
- Le **mode** (ou **classe modale**) de la série est la valeur (ou la classe) du caractère ayant le plus grand effectif.

On note  $x_i$  une valeur prise par un caractère quantitatif : ses valeurs sont numériques.

- L'**effectif cumulé croissant** (respectivement **décroissant**) de  $x_i$  est la somme des effectifs des valeurs inférieures (respectivement supérieures) ou égales à  $x_i$ . On le note ECC.
- La **fréquence cumulée croissante** (respectivement **décroissante**) de  $x_i$  est la somme des fréquences des valeurs inférieures (respectivement supérieures) ou égales à  $x_i$ . On la note FCC.

La somme de toutes les fréquences est toujours égale à 1.

#### I.2 Paramètres d'une série statistique

|                |       |       |     |       |
|----------------|-------|-------|-----|-------|
| Variable $x_i$ | $x_1$ | $x_2$ | ... | $x_p$ |
| Effectif $n_i$ | $n_1$ | $n_2$ | ... | $n_p$ |

- La **moyenne** d'une série statistique est le nombre, noté  $\bar{x}$ , défini par :

$$\bar{x} = \frac{n_1 x_1 + n_2 x_2 + \dots + n_p x_p}{N} = f_1 x_1 + f_2 x_2 + \dots + f_n x_n$$

- Soit  $a$  et  $b$  deux réels.

Si une série statistique a pour moyenne  $\bar{x}$ , alors la série  $\{ax_i + b\}$  a pour moyenne  $a\bar{x} + b$ .

La **médiane** d'une série statistique est le nombre, noté  $Me$ , qui partage la série en deux groupes de même effectif (50% des valeurs en dessous et 50% au dessus).

Les **quartiles** d'une série partagent celle-ci en quatre groupes de même effectif :

- Le **premier quartile**  $Q_1$  est la plus petite valeur de la série telle que au moins un quart des valeurs sont inférieures ou égales à  $Q_1$ .
- Le **troisième quartile**  $Q_3$  est la plus petite valeur de la série telle que au moins trois quarts des valeurs de la série sont inférieures ou égales à  $Q_3$ .

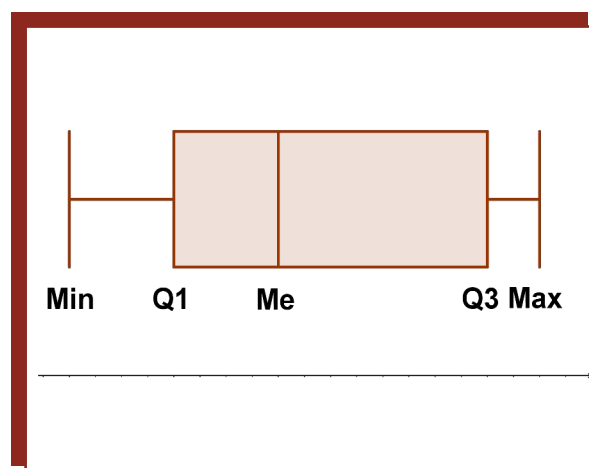
- L'**étendue** d'une série statistique est la différence entre la valeur maximale et la valeur minimale.
- L'**écart interquartile** est le nombre  $Q_3 - Q_1$  et mesure la dispersion de la série par rapport à la médiane.
- L'**intervalle interquartile**  $[Q_1; Q_3]$  contient 50% des valeurs de la série.
- L'**écart-type** d'une série statistique est le nombre réel positif, noté  $\sigma$ , défini par :

$$\sigma = \sqrt{\frac{n_1(x_1 - \bar{x})^2 + n_2(x_2 - \bar{x})^2 + \dots + n_p(x_p - \bar{x})^2}{N}}$$

### Représentation graphique d'une série

Pour résumer une série statistique sous forme de schéma, on utilisera ce que l'on appelle un diagramme en boîte (ou boîte à moustaches).

Un diagramme en boîte représente graphiquement une série en faisant apparaître les principaux indicateurs de la série. On aura ainsi environ 25% des valeurs de la série entre  $Min$  et  $Q_1$ , environ 25% entre  $Q_1$  et  $Me$ , environ 25% entre  $Me$  et  $Q_3$  (donc environ 50% des valeurs dans la boîte) et environ 25% des valeurs entre  $Q_3$  et  $Max$ .



### Méthode : Détermination de la médiane et des quartiles d'une série

Pour déterminer la médiane et les quartiles d'une série, on range tout d'abord les  $N$  valeurs de la série par ordre croissant (si ce n'est pas déjà fait).

Pour la médiane :

- Si  $N$  est un nombre impair, alors la médiane  $Me$  est la valeur située au rang  $\frac{N+1}{2}$ .
- Si  $N$  est pair, alors la médiane est la moyenne des valeurs de rangs  $\frac{N}{2}$  et  $\frac{N}{2} + 1$ .

Pour les quartiles :

- $Q_1$  est la valeur de la série de rang  $\frac{N}{4}$  (ou l'entier directement supérieur si  $\frac{N}{4} \notin \mathbb{N}$ )
- $Q_3$  est la valeur de la série de rang  $\frac{3N}{4}$  (ou l'entier directement supérieur si  $\frac{3N}{4} \notin \mathbb{N}$ )

## II. Série statistiques à deux variables

Lorsque l'on s'intéresse à un sujet, il est fréquent de recueillir plusieurs données ou caractères. Il est parfois intéressant de ne pas regarder chaque caractère indépendamment, mais d'essayer d'établir une éventuelle corrélation entre ces deux caractères. C'est le rôle des statistiques.

Lorsqu'une série statistique étudie simultanément deux caractères  $X$  et  $Y$ , on dit qu'il s'agit d'une série statistique à deux variables (ou double) que l'on note  $(X; Y)$ .

Les valeurs prises par le caractère  $X$  sont notées  $(x_1; x_2; \dots; x_n)$ .

Les valeurs prises par le caractère  $Y$  sont notées  $(y_1; y_2; \dots; y_n)$ .

On associe à chaque individu de la population un couple  $(x_i; y_i)$  où  $x_i$  et  $y_i$  sont les valeurs prises par l'individu « numéro  $i$  », avec  $i \in [1; n]$  (ou parfois entre 0 et  $n-1$ ).

On appelle alors série statistique à deux variables  $(X; Y)$  l'ensemble des couples  $(x_i; y_i)$  associés à chaque individu de la population.

## III. Point moyen et forme d'un nuage de points

### III.1 Point moyen

Dans cette partie, on représentera une série statistique à deux variables à l'aide d'un tableau de la forme :

|                                  |       |       |     |       |
|----------------------------------|-------|-------|-----|-------|
| <b>Variable <math>x_i</math></b> | $x_1$ | $x_2$ | ... | $x_n$ |
| <b>Variable <math>y_i</math></b> | $y_1$ | $y_2$ | ... | $y_n$ |

On peut associer à chaque couple  $(x_i; y_i)$ , pour  $i$  allant de 1 à  $n$ , de la série statistique à deux variables  $(X; Y)$  le point  $M_i$  de coordonnées  $(x_i; y_i)$  dans un repère orthogonal. L'ensemble des points  $M_i(x_i; y_i)$  est appelé **nuage de points** associés à la série statistique à deux variables.

Soit  $\bar{x}$  la moyenne des valeurs prises par le 1er caractère et  $\bar{y}$  celle des valeurs prises par le 2e caractère. On appelle **point moyen du nuage de points** de la série statistique le point **G** de coordonnées  $(\bar{x}; \bar{y})$  tel que  $\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n}$  et  $\bar{y} = \frac{y_1 + y_2 + \dots + y_n}{n}$ .

### III.2 Forme d'un nuage de points

La forme d'un nuage de points peut indiquer ou non une liaison entre les deux variables  $X$  et  $Y$ . L'allure du nuage de points représentant la série statistique permet d'avoir une idée de la réponse. On cherche alors s'il existe un lien entre les deux variables. Pour cela, on va essayer de trouver une courbe, si elle existe, qui « approche au mieux » le nuage, c'est-à-dire une courbe qui passe au plus près des points du nuage.

Chercher une courbe qui « approche au mieux » un nuage de point s'appelle effectuer un **ajustement**. Cette courbe d'ajustement, si elle existe, représente une fonction  $f$  qui permet quasiment d'exprimer la variable  $y$  en fonction de la variable  $x$  sous la forme  $y = f(x)$ .

### Remarque

Parmi les tendances les plus communes, on trouvera :

- les points du nuage sont dispersés, ils n'indiquent pas de tendance particulière ;
- les points du nuage sont sensiblement alignés, alors il existe une **liaison affine** entre les deux variables  $X$  et  $Y$  ;
- les points du nuage sont sensiblement sur une courbe exponentielle, alors il existe une **liaison exponentielle** entre les variables  $X$  et  $Y$ .

## IV. Ajustement affine

### IV.1 Ajustement affine

Lorsqu'un nuage de points  $M_i(x_i; y_i)$  met en évidence une liaison affine entre les deux variables étudiées, on peut la représenter par une droite d'équation  $y = ax + b$ .

Lorsque les points du nuage sont sensiblement alignés, on appelle **droite d'ajustement de  $y$  en  $x$**  de la série statistique ou **droite de régression linéaire de  $y$  en  $x$**  une droite qui passe au plus près des points du nuage. On dit que cette droite réalise un **ajustement affine** du nuage de points.

### IV.2 Méthode graphique

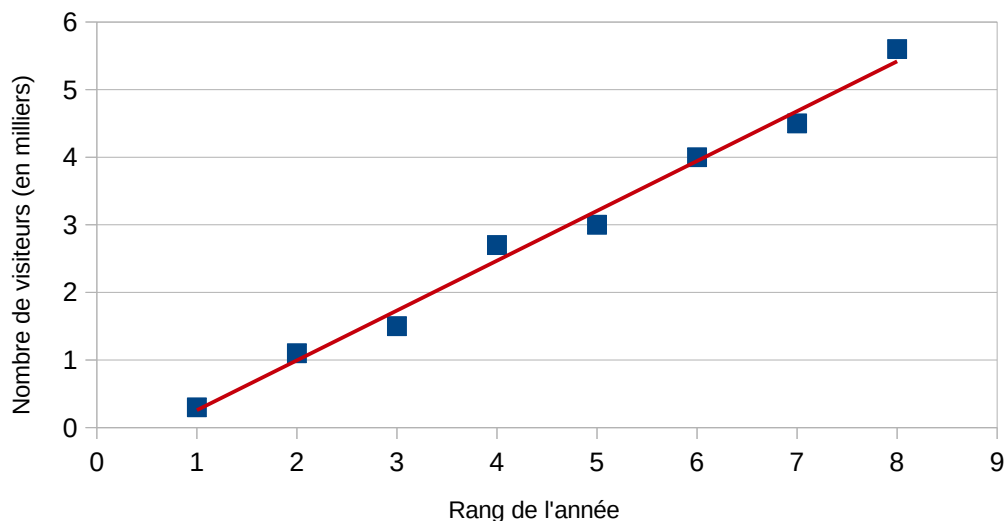
Dans le plan rapporté à un repère, on trace une droite passant au plus près de tous les points du nuage en suivant les indications fournies par l'énoncé.

### Exemple

On a relevé, pour un site internet, le nombre de visiteurs (en milliers) par an pendant les huit premières années de fonctionnement :

| Année                                      | 2011 | 2012 | 2013 | 2014 | 2015 | 2016 | 2017 | 2018 |
|--------------------------------------------|------|------|------|------|------|------|------|------|
| Rang $x_i$                                 | 1    | 2    | 3    | 4    | 5    | 6    | 7    | 8    |
| Nombre de visiteurs $y_i$<br>(en milliers) | 0,3  | 1,1  | 1,5  | 2,7  | 3    | 4    | 4,5  | 5,6  |

Dans cet exemple, les points du nuage sont sensiblement alignés. Traçons une droite d'ajustement :



### Remarque

Par cette méthode, on effectue un tracé « au jugé ». Nous n'obtiendrons donc pas à chaque fois exactement la même droite.

## IV.3 Méthode des moindres carrés et covariance

### Méthode des moindres carrés

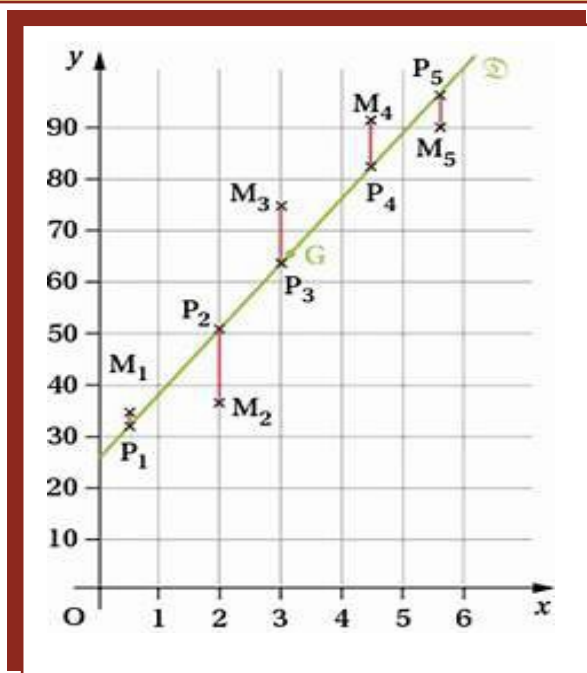
Soit un nuage de points  $M_i(x_i; y_i)_{1 \leq i \leq n}$ , une droite  $d$  d'équation  $y = ax + b$  qui ajuste le nuage et les points  $P_i(x_i; ax_i + b)$  sur la droite d'ajustement affine.

Il existe deux réels  $a$  et  $b$  tels que la somme  $M_1P_1^2 + M_2P_2^2 + \dots + M_nP_n^2$  est **minimale**.

Pour ces valeurs de  $a$  et  $b$ , la droite d'équation  $y = ax + b$  est appelée **droite des moindres carrés** associée au nuage de points  $M_i(x_i; y_i)_{1 \leq i \leq n}$ .

### Remarque

Cette méthode consiste à déterminer une droite d'ajustement  $D$ , d'équation  $y = ax + b$ , passant « en moyenne » au plus près des points du nuage, telle que la somme des carrés des écarts entre les points de même abscisse  $M_i(x_i; y_i)$  du nuage de points et  $P_i(x_i; y'_i)$  de la droite soient minimale.  $M_iP_i = y_i - (ax_i + b)$  représente l'écart entre le point  $M_i$  du nuage et la droite. Lorsque l'on fait la somme des carrés de ces écarts, on cherche à minimiser le résultat. Il s'agit donc de trouver les nombres  $a$  et  $b$  qui minimisent cette quantité.



### Remarque

Pour obtenir l'équation de la droite d'ajustement affine, on utilisera la calculatrice ou un tableur. Dans l'exemple graphique précédent, on obtiendrait la droite d'équation  $y = 0,74x - 0,48$  (en arrondissant à  $10^{-2}$  près).

La droite d'ajustement affine de  $y$  en  $x$  par la méthode des moindres carrés **passse par le point moyen** du nuage de points de la série statistique.

On appelle **covariance** des séries statistiques  $x$  et  $y$  le réel  $cov(x; y)$  défini par :

$$cov(x; y) = \frac{(x_1 - \bar{x})(y_1 - \bar{y}) + (x_2 - \bar{x})(y_2 - \bar{y}) + \dots + (x_n - \bar{x})(y_n - \bar{y})}{n}$$

### Remarques

- La covariance peut être positive ou négative.
- On note aussi la covariance  $\sigma_{xy}$ .
- Une autre formule pour calculer la covariance est :  $cov(x; y) = \sigma_{xy} = \frac{x_1 y_1 + x_2 y_2 + \dots + x_n y_n}{n} - \bar{x}\bar{y}$ .

La droite des moindres carrés a pour équation  $y = ax + b$  avec  $a = \frac{cov(x; y)}{V(x)}$  et  $b = \bar{y} - a\bar{x}$ .

### Remarques

- Le réel  $V(x)$  s'appelle la variance de la série statistique  $x$ . Ce réel est positif.
- On peut obtenir les coefficients  $a$  et  $b$  de la droite des moindres carrés directement à l'aide de la calculatrice.

On appelle **coefficient de corrélation linéaire** des séries statistiques  $x$  et  $y$  le réel défini par :

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

Le coefficient de corrélation linéaire est un réel compris entre  $-1$  et  $1$ . Il vérifie :

- 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 n'a de sens.
- Si  $|r| = 1$ , alors la droite de régression passe par tous les points du nuage.

#### IV.4 Ajustement exponentiel

Parfois le nuage de points de la série statistique  $(x; y)$  a une forme générale qui ressemble à la courbe d'une fonction exponentielle  $x \rightarrow e^{ax+b}$ .

Pour déterminer l'expression de la fonction dont la courbe représentative approche au mieux le nuage de points, on effectue un changement de variable en posant  $y' = \ln y$ .

Le nuage de points de la série statistique  $(x; y')$  présente alors une forme rectiligne que l'on peut ajuster par la droite de régression de  $y'$  en  $x$ , d'équation  $y' = ax + b$ .

On a alors  $\ln y = ax + b$  ce qui équivaut à  $y = e^{ax+b}$ .

Cette équation représente une courbe exponentielle qui ajuste au mieux le nuage.

On dit alors qu'on a réalisé un ajustement exponentiel.

#### V. Interpolations et extrapolations

Lorsque l'on a pu réaliser un ajustement (affine ou exponentiel) satisfaisant pour le nuage de points d'une série statistique à deux variables  $x$  et  $y$ , alors on peut considérer, si cela a du sens dans la situation étudiée, que les quantités  $x$  et  $y$  ne sont pas complètement indépendants l'une de l'autre et ont un lien qui peut être modélisé par une fonction  $f$  (affine ou exponentielle), qui correspond à l'ajustement trouvé, et telle que  $y = f(x)$ .

Cela permet d'effectuer des estimations de la quantité  $y$  qui pourraient correspondre à des valeurs théoriques de la quantité  $x$ , non relevées initialement dans l'étude statistique. Ces estimations s'effectuent soit par lecture graphique (à partir du tracé de la droite d'ajustement), soit par le calcul (en utilisant l'équation de la droite d'ajustement).

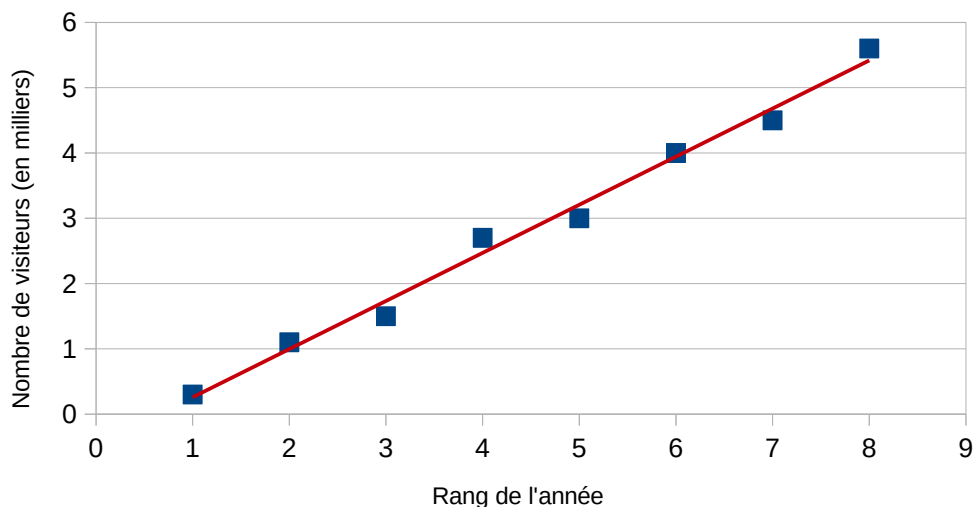
Dans le cas où les valeurs choisies restent dans le « domaine d'étude » défini par l'échantillon, on dit que l'on réalise une **interpolation**.

Réciproquement, on peut aussi estimer une valeur  $x$  qui pourrait correspondre à une valeur donnée de  $y$ . Il s'agit d'une **extrapolation**. Lorsque la variable  $x$  est une mesure du temps, la série est appelée série chronologique et il s'agit alors de se projeter dans le futur pour faire des prévisions.

#### Exemple

Reprenons l'exemple du nombre de visiteurs (en millier) par an sur un site internet pendant les huit premières années de fonctionnement :

| Année                                      | 2011 | 2012 | 2013 | 2014 | 2015 | 2016 | 2017 | 2018 |
|--------------------------------------------|------|------|------|------|------|------|------|------|
| Rang $x_i$                                 | 1    | 2    | 3    | 4    | 5    | 6    | 7    | 8    |
| Nombre de visiteurs $y_i$<br>(en milliers) | 0,3  | 1,1  | 1,5  | 2,7  | 3    | 4    | 4,5  | 5,6  |



Supposons que les statistiques de l'année numéro 5 aient été perdues. Nous pouvons estimer alors la fréquentation du site grâce à l'ajustement affine que nous avons fait.

On a  $y = 0,74x - 0,48$  avec  $x = 5$ . On obtient alors  $y = 5 \times 0,74 - 0,48 = 3,22$ . Nous estimons le nombre de visiteurs à 3 220 la cinquième année. La valeur cherchée était à l'intérieur du nuage de points, on dit qu'on a fait une **interpolation**. Il s'agit ici d'une **estimation**.

Toujours dans cet exemple, nous voulons estimer la fréquentation du site au bout de 20 ans de fonctionnement, en considérant que l'évolution reste la même.

On a  $y = 0,74x - 0,48$  avec  $x = 20$ . On obtient alors  $y = 20 \times 0,74 - 0,48 = 14,32$ . Nous estimons le nombre de visiteurs à 14 320 la vingtième année. La valeur cherchée était au-delà du nuage de points, on dit qu'on a fait une **extrapolation**. Il s'agit ici d'une **prévision**.