Statistique à 2 variables

©Pascal Brachet (CC BY NC SA)

https://www.xm1math.net

1. Rappels sur les séries statistiques à 1 variable

<table>
<tr><td>Rappels</td><td>Exemple</td></tr>
<tr><td>Étant donné une série statistique définie par :
<table>
<tr><td>Valeurs du caractère</td><td>x_1</td><td>x_2</td><td>$\cdots$</td><td>x_p</td></tr>
<tr><td>Effectif</td><td>n_1</td><td>n_2</td><td>$\cdots$</td><td>n_p</td></tr>
</table></td><td>
<table>
<tr><td>x_i</td><td>1</td><td>2</td><td>3</td><td>4</td><td>5</td></tr>
<tr><td>n_i</td><td>2</td><td>3</td><td>1</td><td>3</td><td>1</td></tr>
</table></td></tr>
<tr><td>L'effectif total est : $N = n_1 + n_2 + \cdots + n_p$</td><td>$N = 2 + 3 + \cdots + 1 = 10$</td></tr>
<tr><td>La moyenne est définie par :
$\overline{x} = \dfrac{n_1 x_1 + n_2 x_2 + \cdots + n_p x_p}{N}$</td><td>$\overline{x} = \dfrac{2 \times 1 + 3 \times 2 + \cdots + 1 \times 5}{10} = 2,8$</td></tr>
<tr><td>La variance est définie par :
$V = \dfrac{n_1 (x_1 - \overline{x})^2 + n_2 (x_2 - \overline{x})^2 + \cdots + n_p (x_p - -\overline{x})^2}{N}$</td><td>$V = \dfrac{2 \times (1 - 2,8)^2 + \cdots + 1 \times (5 - 2,8)^2}{10}$
$= 1,76$</td></tr>
<tr><td>L'écart-type est $\sigma = \sqrt{V}$</td><td>$\sigma = \sqrt{1,76} \approx 1,327$</td></tr>
</table>

2. Séries statistiques à 2 variables

a) Situation

Sur une même population, on étudie deux caractères. Pour chacun des individus, on note x_1, x_2, $\cdots$, x_n les valeurs du 1er caractère et y_1, y_2, $\cdots$, y_n les valeurs du 2e caractère. Les données sont présentées sous la forme suivantes :

Caractère x_i	x_1	x_2	$\cdots$	x_n
Caractère y_i	y_1	y_2	$\cdots$	y_n

b) Nuage de points et point moyen

Définition

• Dans un repère orthogonal, on appelle **nuage de points** associé à une série statistique à 2 variables, l'ensemble des points d'abscisse x_1, x_2, $\cdots$, x_n et d'ordonnée y_1, y_2, $\cdots$, y_n.
• Le **point moyen** du nuage de points est le point G d'abscisse $\overline{x}$ (moyenne des x_i) et d'ordonnée $\overline{y}$ (moyenne des y_i).

2. Séries statistiques à 2 variables

Exemple(s)

Avec la série :

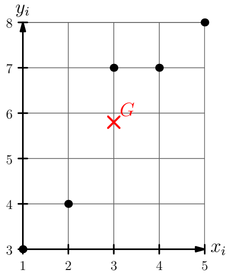
x_i	1	2	3	4	5
y_i	3	4	7	7	8

On a :

$\overline{x} = \dfrac{1+2+3+4+5}{5} = 3$

$\overline{y} = \dfrac{3+4+7+7+8}{5} = 5,8$

Le point moyen G a pour abscisse 3 et pour ordonnée $5,8$



b) Covariance d'une série statistique à 2 variables

Définition

La covariance des x_i et y_i est le réel noté $\text{cov}(x;y)$ défini par :

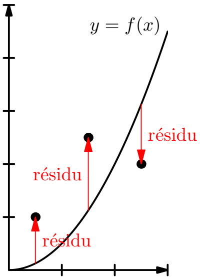
$$\text{cov}(x;y) = \frac{(x_1 - \overline{x})(y_1 - \overline{y}) + (x_2 - \overline{x})(y_2 - \overline{y}) + \cdots + (x_n - \overline{x})(y_n - \overline{y})}{n}$$

xm1math.net

3. Ajustement d'un nuage de points

a) Introduction - Méthode des moindres carrés

- Effectuer un **ajustement** de y en x d'un nuage de points, c'est chercher une fonction f dont la courbe est la « plus proche possible » des points du nuage.
- On peut alors de servir de l'ajustement pour **estimer** une valeur de y pour un x non donné dans le nuage :
  - si le x est à l'extérieur du nuage, l'estimation est appelée **extrapolation** ;
  - si le x est à l'intérieur du nuage, l'estimation est appelée **interpolation** ;
- Pour chaque point du nuage d'abscisse x_i et d'ordonnée y_i, on appelle **résidu** la différence $y_i - f(x_i)$.
- La méthode d'ajustement par **les moindres carrés** consiste à trouver une fonction f telle que la somme des carrés des résidus soit la plus petite possible.



b) Cas des ajustements affines

Lorsqu'on cherche un ajustement avec une fonction f affine dont la courbe est une droite, on dit qu'on effectue un **ajustement affine** de y en x. Ce type d'ajustement est adapté aux cas où les points du nuage semblent « à peu près alignés ».

3. Ajustement d'un nuage de points

Définition-Propriété

- *L'utilisation de la méthode des moindres carrés pour déterminer un ajustement affine donne une droite, appelée **droite des moindres carrés ou droite de régression** de y en x, telle que :*
  - *cette droite passe par le point moyen G du nuage de points ;*
  - *cette droite admet comme équation $y = ax + b$ avec $a = \dfrac{cov(x;y)}{V(x)}$ et $b = \overline{y} - a\overline{x}$ (a et b sont fournis directement par la calculatrice)*
- *La « qualité » de l'ajustement peut-être vérifié par le calcul du **coefficient de corrélation linéaire** $r = \dfrac{cov(x;y)}{\sigma(x)\sigma(y)}$. Plus ce coefficient est proche de 1 (pour des données croissantes) ou de -1 (pour des données décroissantes), plus les points du nuage sont proches de la droite des moindres carrés.*

Obtention de la droite des moindres carrés avec une CASIO

`MENU` → `STAT`

• Entrée des données : rentrer les valeurs x_i dans la liste 1 et les valeurs y_i dans la liste 2.

• Affichage des résultats : `CALC` → `SET`

Pour `2Var XList`, choisir `List 1`
Pour `2Var YList`, choisir `List 2`
Pour `2VarFreq`, taper `1`

Choisir `REG`, puis `X`

On peut lire a et b dans la liste des résultats (ainsi que r)

3. Ajustement d'un nuage de points

Obtention de la droite des moindres carrés avec une TI

- Entrée des données : `stats` `EDIT` `1:Edite` ; rentrer les valeurs x_i dans L1 et les valeurs y_i dans L2.
- Affichage des résultats :

Pour a et b : `stats` `CALC` `4:RegLin(ax+b)`

`Xlist : L1` et `Ylist : L2`, puis `Calculs`

Pour r : `var` `5:Statistiques` `EQ` `r`

Exemple(s)

Avec la série :

x_i	1	2	3	4	5
y_i	8	9	12	12	14

- Point moyen : $\bar{x} = \dfrac{1+2+3+4+5}{5} = 3$; $\bar{y} = \dfrac{8+9+12+12+14}{5} = 11$

  Donc, on a $G\,(3;11)$
- Droite des moindres carrés : La calculatrice donne $a = 1,5$ et $b = 6,5$. Une équation de la droite des moindres carrés est donc : $y = 1,5x + 6,5$
- Estimation de la valeur de y pour $x = 7$: $y = 1,5 \times 7 + 6,5 = 17$

3. Ajustement d'un nuage de points

c) Ajustement se ramenant par changement de variable à un ajustement affine

Remarque(s)

Quand les points du nuage ne semblent pas « regroupés autour d'une même droite » mais plutôt autour d'une courbe, on peut être amené à effectuer des ajustements affines entre x_i et $\ln(y_i)$, ou entre $\ln(x_i)$ et $\ln(y_i)$, ou entre x_i^2 et y_i ...

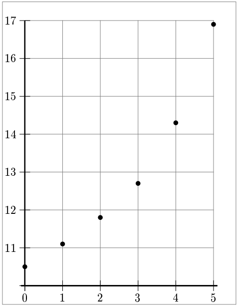
Exemple(s)

Le tableau ci-dessous indique la population d'une ville en milliers d'habitants selon l'année :

Année	2015	2016	2017	2018	2019	2020
Rang de l'année x_i	0	1	2	3	4	5
Nb d'habitants y_i	10,5	11,1	11,8	12,7	14,3	16,4

Le nuage de points associé suggère qu'un ajustement affine de y en x ne semble pas très adapté : la progression du nombre d'habitants semble plus être exponentielle qu'affine.

On pense donc à effectuer un ajustement affine non pas entre x_i et y_i, mais entre x_i et $\ln(y_i)$.



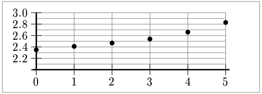
3. Ajustement d'un nuage de points

Exemple(s)

1. On pose $z = \ln y$. La série de z en x devient :

Rang de l'année x_i	0	1	2	3	4	5
$z_i = \ln(y_i)$	2,35	2,41	2,47	2,54	2,66	2,88

Et le nuage de points de z en x devient :



2. Avec la calculatrice, on obtient $z = 0,092\,x + 2,313$ comme droite de régression de z en x par la méthode des moindres carrés.

3. Reste à en déduire le nombre de milliers d'habitants y sous la forme $y = A\,\mathrm{e}^{Bx}$:
$z = 0,092\,x + 2,313$
$\Leftrightarrow \ln y = 0,092\,x + 2,313$
$\Leftrightarrow y = \mathrm{e}^{0,092\,x+2,313}$
$\Leftrightarrow y = \mathrm{e}^{2,313} \times \mathrm{e}^{0,092\,x}$
$\Leftrightarrow y = 10,105\,\mathrm{e}^{0,092\,x}$

4. On peut dés lors obtenir, par exemple, une estimation du nombre de milliers d'habitants en 2021 en prenant $x = 6$: pour $x = 6$, on a $y = 10,105\,\mathrm{e}^{0,092\times 6} \approx 17,55$

Fin du chapitre