Ecole Nationale Polytechnique de Constantine

COURS DE STATISTIQUE DESCRIPTIVE

Zaher MOHDEB

E-mail: z.mohdeb@gmail.com, zaher.mohdeb@umc.edu.dz

Chapitre 2 Série statistique à deux caractères

1- Distributions statistiques à deux variables

1.1- Distribution conjointe

Soient X et Y deux variables statistiques qui peuvent être qualitatives ou quantitatives, et qui peuvent ne pas être de même nature.
Les k modalités de X sont désignées par $x_1, \ldots, x_i, \ldots, x_k$;
les l modalités de Y sont désignées par $y_1, \ldots, y_j, \ldots, y_l$.
La i^{e} modalité d'une variable désigne le centre de la i^{e} classe dans le cas d'une variable quantitative continue.
Supposons que les deux variables X et Y étudiées sont des variables discrètes et que les caractères sont des caractères quantitatifs.
La répartition des n observations, ou distribution conjointe, suivant les modalités de X et Y se présente sous forme d'un tableau à double entrée, appelée tableau de contingence (cf. tableau 1 suivant).

Distributions statistiques à deux variables

X \ Y	y_1	$\dots$	y_j	$\dots$	y_l	Total
x_1	n_{11}	$\dots$	n_{1j}	$\dots$	n_{1l}	$n_{1\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_i	n_{i1}	$\dots$	n_{ij}	$\dots$	n_{il}	$n_{i\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_k	n_{k1}	$\dots$	n_{kj}	$\dots$	n_{kl}	$n_{k\bullet}$
Total	$n_{\bullet 1}$	$\dots$	$n_{\bullet j}$	$\dots$	$n_{\bullet l}$	n

TABLE: (Tableau 1) : Tableau de contingence de la distribution conjointe de deux variables X et Y.

Distributions statistiques à deux variables

On définit les fréquences absolues suivantes :
- Les fréquences marginales :

$$n_{i\bullet} = \sum_{j=1}^{l} n_{ij} \quad \text{et} \quad n_{\bullet j} = \sum_{i=1}^{k} n_{ij} \, ,$$

- La fréquence marginale $n_{i\bullet}$ est donc le nombre d'individus possédant la modalité x_i du caractère X quelle que soit la modalité de Y ; par exemple tous les individus ayant le même poids quelle que soit leur taille.
De même, l'effectif ou la fréquence marginale $n_{\bullet j}$ est le nombre d'individus de la modalité y_j de Y, quelle que soit la modalité de X.
On a évidemment : $\sum_{i=1}^{k} n_{i\bullet} = \sum_{j=1}^{l} n_{\bullet j} = n$.

Distributions statistiques à deux variables

- La fréquence conditionnelle $f_{j/i}$ est la distribution de la variable Y quand on a fixé la modalité x_i pour la variable X ; on s'intéresse, par exemple, à la répartition des tailles des individus ayant tous le même poids. Elle est définie par :

$$f_{j/i} = \frac{n_{ij}}{n_{i\bullet}} .$$

- On définit de la même façon la fréquence conditionnelle $f_{i/j}$ par :

$$f_{i/j} = \frac{n_{ij}}{n_{\bullet j}} .$$

On s'intéresse, par exemple, à la répartition des poids des individus ayant tous la même taille.

Distributions statistiques à deux variables

- Les fréquences relatives f_{ij}, $f_{i\bullet}$ et $f_{\bullet j}$ sont obtenues en divisant les effectifs n_{ij} et les fréquences marginales $n_{i\bullet}$ et $n_{\bullet j}$ par l'effectif total n :

$$f_{ij} = \frac{n_{ij}}{n}, \qquad f_{i\bullet} = \frac{n_{i\bullet}}{n} \qquad \text{et} \qquad f_{\bullet j} = \frac{n_{\bullet j}}{n}.$$

- Les distributions X et Y sont statistiquement indépendantes si :

$$f_{ij} = f_{i\bullet} f_{\bullet j}$$

pour toutes les valeurs des indices i et j.

1.2- Distribution marginale

Les k couples $(x_i, n_{i\bullet})$ forment la distribution marginale de la variable X. Les l couples $(y_j, n_{\bullet j})$ forment la distribution marginale de la variable Y. Les distributions marginales peuvent aussi être données sous forme de fréquences :

$$f_{i\bullet} = \frac{n_{i\bullet}}{n} \quad \text{et} \quad f_{\bullet j} = \frac{n_{\bullet j}}{n}\,.$$

Disposant d'une distribution conjointe, on peut déduire les distributions marginales qui permettent d'étudier séparément chaque variable en représentant graphiquement sa distribution et s'il s'agit d'une variable quantitative, en calculant ses caractéristiques de tendance centrale, de dispersion, de forme ...

Distributions statistiques à deux variables

Il est clair que la moyenne de X et celle de Y sont données respectivement par

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{k} n_{i\bullet}\, x_i \quad \text{et} \quad \overline{y} = \frac{1}{n}\sum_{j=1}^{l} n_{\bullet j}\, y_j$$

et les variances respectives sont données par

$$\mathrm{Var}(X) = \frac{1}{n}\sum_{i=1}^{k} n_{i\bullet}(x_i - \overline{x})^2 = \frac{1}{n}\sum_{i=1}^{k} n_{i\bullet}\, x_i^2 - \overline{x}^2$$

et

$$\mathrm{Var}(Y) = \frac{1}{n}\sum_{j=1}^{l} n_{\bullet j}(y_j - \overline{y})^2 = \frac{1}{n}\sum_{j=1}^{l} n_{\bullet j}\, y_j^2 - \overline{y}^2$$

1.3- Distribution conditionnelle

La distribution de la variable Y, la variable X étant égale à x_i, est appelée distribution conditionnelle de Y pour $X = x_i$, (ou distribution conditionnelle de Y relatif à $X = x_i$).

$Y/X = x_i$	y_1	$\ldots$	y_j	$\ldots$	y_l	Total
Effectif	n_{i1}	$\ldots$	n_{ij}	$\ldots$	n_{il}	$n_{i\bullet}$

TABLE: Distribution conditionnelle des effectifs de Y relatif à $X = x_i$.

Distributions statistiques à deux variables

Cette distribution des $n_{i\bullet}$ observations, satisfaisant à la condition $X = x_i$, est présentée sous la forme de fréquences conditionnelles :

$$f_{j/i} = \frac{n_{ij}}{n_{i\bullet}} \qquad \text{avec} \qquad \sum_{j=1}^{l} f_{j/i} = 1$$

$Y/X = x_i$	y_1	$\ldots$	y_j	$\ldots$	y_l	Total
Fréquence	$f_{1/i}$	$\ldots$	$f_{j/i}$	$\ldots$	$f_{l/i}$	1

TABLE: Distribution des fréquences conditionnelles de Y relatif à $X = x_i$.

Distributions statistiques à deux variables

La fréquence $f_{j/i}$ se lit « f indice j si i », c'est-à-dire fréquence de y_j si $X = x_i$.
Il y a k distributions conditionnelles de Y pour $(i = 1, \ldots, k)$.
Lorsque la variable Y est quantitative, on peut calculer pour chaque valeur x_i sa moyenne conditionnelle $\overline{y}_i$, appelée moyenne conditionnelle de Y relatif à $x = x_i$ et sa variance conditionnelle S_i^2, appelée variance conditionnelle de Y relatif à $x = x_i$:

$$\overline{y}_i = \frac{1}{n_{i\bullet}} \sum_{j=1}^{l} n_{ij} y_j = \sum_{j=1}^{l} f_{j/i} y_j$$

et

$$S_i^2 = \frac{1}{n_{i\bullet}} \sum_{j=1}^{l} n_{ij} (y_j - \overline{y}_i)^2 = \sum_{j=1}^{l} f_{j/i} (y_j - \overline{y}_i)^2 .$$

Distributions statistiques à deux variables

Les k modalités de X induisant une partition des observations en k sous-groupes, la moyenne $\overline{y}$ peut s'exprimer comme somme pondérée des k moyennes $\overline{y}_i$. En effet, on a :

$$\begin{aligned}
\overline{y} &= \frac{1}{n}\sum_{j=1}^{l} n_{\bullet j} y_j = \frac{1}{n}\sum_{j=1}^{l}\left(\sum_{i=1}^{k} n_{ij}\right) y_j \\
&= \frac{1}{n}\sum_{i=1}^{k} n_{i\bullet}\left(\frac{1}{n_{i\bullet}}\sum_{j=1}^{l} n_{ij} y_j\right) \\
&= \frac{1}{n}\sum_{i=1}^{k} n_{i\bullet}\overline{y}_i = \sum_{i=1}^{k} f_{i\bullet}\overline{y}_i
\end{aligned}$$

Distributions statistiques à deux variables

Symétriquement, on a l distributions conditionnelles de X et on définit les fréquences conditionnelles "f indice i si j" :

$$f_{i/j} = \frac{n_{ij}}{n_{\bullet j}} \qquad \text{avec} \qquad \sum_{i=1}^{k} f_{i/j} = 1$$

$X/Y = y_i$	x_1	$\ldots$	x_i	$\ldots$	x_k	Total
Fréquence	$f_{1/j}$	$\ldots$	$f_{i/j}$	$\ldots$	$f_{k/j}$	1

TABLE: Distribution des fréquences conditionnelles de X relatif à $Y = y_j$.

Distributions statistiques à deux variables

Lorsque la variable X est quantitative, on peut calculer pour chaque valeur y_j sa moyenne conditionnelle $\overline{x}_j$ et sa variance conditionnelle S_j^2 :

$$\overline{x}_j = \frac{1}{n_{\bullet j}} \sum_{i=1}^{k} n_{ij} x_i = \sum_{i=1}^{l} f_{i/j} x_i$$

et

$$S_j^2 = \frac{1}{n_{\bullet j}} \sum_{i=1}^{k} n_{ij} (x_i - \overline{x}_j)^2 = \sum_{i=1}^{k} f_{i/j} (x_i - \overline{x}_j)^2$$

et on a la relation suivante entre la moyenne $\overline{x}$ et les l moyennes conditionnelles $\overline{x}_j$:

$$\begin{aligned}
\overline{x} &= \frac{1}{n}\sum_{i=1}^{k} n_{i\bullet}x_i = \frac{1}{n}\sum_{i=1}^{k}\left(\sum_{j=1}^{l} n_{ij}\right)x_i \\
&= \frac{1}{n}\sum_{j=1}^{l} n_{\bullet j}\left(\frac{1}{n_{\bullet j}}\sum_{i=1}^{k} n_{ij}x_i\right) \\
&= \frac{1}{n}\sum_{j=1}^{l} n_{\bullet j}\overline{x}_j = \sum_{j=1}^{l} f_{\bullet j}\overline{x}_j
\end{aligned}$$

2- Covariance entre deux variables statistiques

Définition

La covariance *généralise à deux variables la notion de variance. Sa formule de définition est la suivante :*

- Cas de données individuelles, par exemple : les deux variables statistiques représentées dans le tableau suivant

X	x_1	x_2	$\ldots$	x_n
Y	y_1	y_2	$\ldots$	y_n

$$Cov(X,Y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \overline{x}\,\overline{y}$$

Covariance entre deux variables statistiques

- Cas de deux variables statistiques représentées dans le Tableau 1 de contingence (page 3).

$$\mathrm{Cov}(X,Y) = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{l} n_{ij}(x_i - \overline{x})(y_j - \overline{y}) = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{l} n_{ij}x_i y_j - \overline{x}\,\overline{y}$$

La covariance est donc la moyenne des produits des écarts aux moyennes (dans chaque produit, chacun des deux écarts est relatif à l'une des deux variables considérées).
Intuitivement, la covariance caractérise les variations simultanées de deux variables statistiques : elle sera positive lorsque les écarts entre les variables et leurs moyennes ont tendance à être de même signe, négative dans le cas contraire.

Covariance entre deux variables statistiques

Propriétés de la covariance

1) $\mathrm{Cov}(X, Y) = \mathrm{Cov}(Y, X)$

2) $\mathrm{Cov}(X, X) = \mathrm{Var}(X)$

3) $\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\ \mathrm{Cov}(X, Y)$

4) $\forall a, b, c, \alpha, \beta \in \mathbb{R},\ \mathrm{Cov}(aX + \alpha, bY + \beta) = ab\ \ \mathrm{Cov}(X, Y)$

$\Longrightarrow \mathrm{Var}(aX + bY + c) = a^2\ \mathrm{Var}(X) + b^2\mathrm{Var}(Y) + 2ab\,\mathrm{Cov}(X, Y)$

5) $|\mathrm{Cov}(X, Y)| \leq \sqrt{\mathrm{Var}(X)\,\mathrm{Var}(Y)}$.

(Cette inégalité découle de l'inégalité de Cauchy-Schwarz).

Coefficients de corrélation linéaire

Comme la variance, la covariance n'a pas de signification concrète. Dans le cas de la variance, on doit passer à l'écart-type pour avoir un indicateur interprétable ; dans celui de la covariance, il faudra passer au coefficient de corrélation linéaire.

Définition

On appelle coefficient de corrélation linéaire entre deux variables statistiques X et Y, le rapport de leur covariance par le produit de leurs écarts-types :

$$r(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)\,Var(Y)}}\,.$$

Propriétés du coefficients de corrélation linéaire

$\forall a, b, \alpha, \beta \in \mathbb{R}$, on a :

$$\begin{aligned} r(aX+\alpha, bY+\beta) &= \frac{\mathrm{Cov}(aX+\alpha, bY+\beta)}{\sqrt{\mathrm{Var}(aX+\alpha)\,\mathrm{Var}(bY+\beta)}} \\ &= \frac{ab\,\mathrm{Cov}(X,Y)}{|ab|\sqrt{\mathrm{Var}(X)\,\mathrm{Var}(Y)}} \\ &= \begin{cases} +r(X,Y) & \text{si } a \text{ et } b \text{ de même signe} \\ -r(X,Y) & \text{si } a \text{ et } b \text{ de signe opposé}. \end{cases} \end{aligned}$$

Ce coefficient, invariant par changement d'origine et d'échelle, est un nombre sans dimension qui, d'après la propriété 5 de la covariance, varie entre -1 et $+1$.
On montrera que s'il est égal à ± 1, les n points sont alignés.

3- Ajustement d'un nuage de points

3.1- Ajustement à l'aide d'une droite

Soient (X, Y) un couple de variables statistiques et supposons que les observations de (X, Y) des n individus de l'échantillon, ont conduit aux valeurs suivantes : $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$.

Une représentation graphique du nuage de points $\{(x_i, y_i), i = 1, \ldots, n\}$ dans un plan (on porte en abscisse, les valeurs x_i de X et en ordonnée, les valeurs y_i de la variable Y) donne une première indication sur la nature de la liaison pouvant exister entre ces variables.

Le calcul de la covariance et du coefficient de corrélation linéaire

$$r(X,Y) = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^2}}$$

$$= \frac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\overline{x})^2}\sqrt{\sum_{i=1}^{n}(x_i-\overline{x})^2}}$$

entre ces deux variables statistiques X et Y nous renseigne sur la relation entre ces variables et sur sa nature.

Ajustement d'un nuage de points

Le nombre $r(X, Y)$ est sans dimension et vérifie

$$-1 \leq r(X, Y) \leq +1 \quad \text{et} \quad r(X, Y) = r(Y, X).$$

Par ailleurs $r(X, Y)$ donne une indication sur l'intensité de la relation linéaire entre les variables X et Y.

- Si $r(X, Y) = 0$, il n'y a pas de corrélation linéaire entre les variables statistiques X et Y.
- Les valeurs extrêmes traduisent la corrélation parfaite, positive si $r(X, Y) = +1$, négative si $r(X, Y) = -1$.

Ces cas extrêmes sont rares ; cependant, si $r(X, Y)$ est voisin de $+1$ ou de -1, les points de coordonnées (x_i, y_i) sont sensiblement alignés. Dans ce cas un il y a une relation du type $Y = aX + b$.

Ajustement d'un nuage de points

Il s'agit, ainsi, de trouver une droite D d'équation

$$Y = aX + b$$

qui ajuste au "mieux" le nuage de points $\{(x_i, y_i),\ i = 1, \ldots, n\}$ qui ne sont pas tout à fait colinéaires.

Nous supposerons que la variable X est notre variable contrôlée et qui n'est pas sujette à erreurs que nous appellerons **variable explicative**. C'est en fonction des valeurs de X que nous essayerons d'expliquer les variations de Y appelée **variable expliquée**.

Ajustement d'un nuage de points

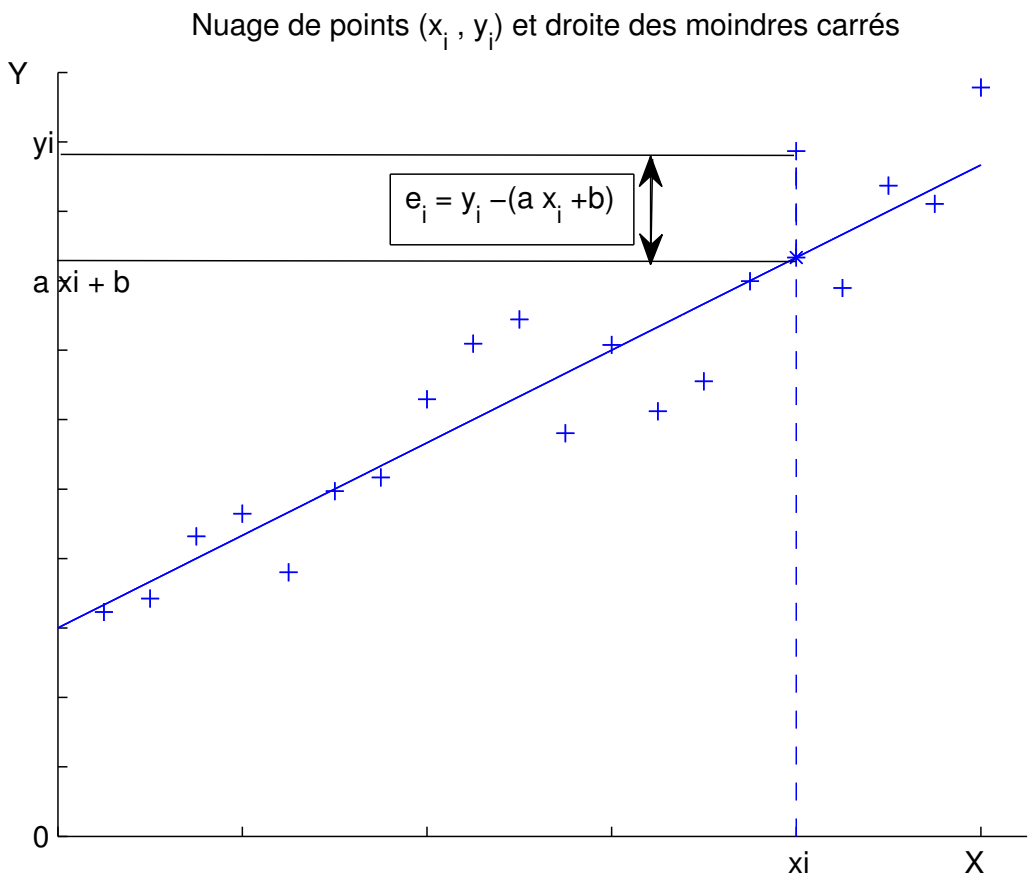

FIGURE: (Figure 1) Nuage de points (x_i, y_i) et droite des moindres carrés.

Ajustement d'un nuage de points

La méthode utilisée pour obtenir une droite qui s'ajuste le mieux possible au nuage de points, est **la méthode des moindres carrés** qui consiste à rendre minimale la somme des carrés des écarts (ou valeurs résiduelles) $e_i = y_i - (ax_i + b)$ des valeurs observées y_i à la droite. Autrement dit, il s'agit de trouver les coefficients a et b vérifiant la propriété :

$$\text{Rendre} \quad \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \Big[y_i - (ax_i + b) \Big]^2 \quad \text{minimum.}$$

Dans l'expression de la fonction à minimiser, les inconnues sont a et b. Les valeurs x_i et y_i étant des nombres connus, la quantitié $\sum_{i=1}^{n} e_i^2$ à minimiser est donc une fonction de a et b que nous allons noter par $F(a, b)$.

Ajustement d'un nuage de points

Le minimum de $F(a,b) = \sum_{i=1}^{n} \Big[y_i - (ax_i + b) \Big]^2$ est obtenu pour :

$$\begin{cases} \dfrac{\partial F(a,b)}{\partial b} = 0 \implies \displaystyle\sum_{i=1}^{n} (y_i - ax_i - b) = 0 \\ \dfrac{\partial F(a,b)}{\partial a} = 0 \implies \displaystyle\sum_{i=1}^{n} x_i (y_i - ax_i - b) = 0 \,. \end{cases}$$

La première équation a pour solution

$$\overline{y} = a\,\overline{x} + b\,, \quad \text{où} \quad \overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad \text{et} \quad \overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i\,,$$

et la deuxième, compte tenu de la première solution

$$a = \frac{\dfrac{1}{n}\displaystyle\sum_{i=1}^{n} x_i\,y_i - \overline{x}\,\overline{y}}{\dfrac{1}{n}\displaystyle\sum_{i=1}^{n} x_i^2 - \overline{x}^2} = \frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(X)}\,.$$

Ajustement d'un nuage de points

Ainsi, les valeurs

$$\begin{cases} a = \dfrac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(X)} = \dfrac{\frac{1}{n}\sum_{i=1}^{n} x_i\, y_i - \overline{x}\,\overline{y}}{\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \overline{x}^2} \\ b = \overline{y} - a\,\overline{x} \end{cases} \tag{1}$$

sont les coefficients de la droite cherchée obtenus en minimisant la somme des carrés des écarts des observations y_i par rapport à cette droite. La droite ainsi obtenue est appelée **la droite d'ajustement par la méthode des moindres carrés**.

Ajustement d'un nuage de points

La droite D d'ajustement ainsi obtenue a pour équation

$$y - \overline{y} = \frac{Cov(X,Y)}{Var(X)}(x - \overline{x}). \tag{2}$$

Notons que cette droite D passe par le point $(\overline{x}, \overline{y})$.
Par ailleurs, puisque $r(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)\,Var(Y)}}$ et en remplaçant $Cov(X,Y)$ par $r(X,Y)\sqrt{Var(X)\,Var(Y)}$ dans l'équation (2), on a une autre forme de l'équation de la droite D donnée par

$$y - \overline{y} = r(X,Y)\frac{\sqrt{Var(Y)}}{\sqrt{Var(X)}}(x - \overline{x}). \tag{3}$$

Ajustement d'un nuage de points

Droite d'ajustement D' de la variable X en fonction de la variable Y

Notons que sur la base de l'observation d'un nuage de points $\{(x_i, y_i),\ i = 1, \ldots, n\}$, nous pouvons également déterminer la droite d'ajustement D' de la variable X en fonction de la variable Y :

$$X = c\,Y + d\,.$$

Dans ce cas Y est la variable explicative et X est la variable expliquée. Les paramètres c et d de la doite sont déterminés avec le même principe de la méthode précédente et qui consiste à **minimiser la quantité** :

$$\sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} \Big[x_i - (c\,y_i + d)\Big]^2.$$

Ajustement d'un nuage de points

D'après la relation (1) et par symétrie des variables X et Y on obtient les paramètres c et d de la droite D' de X en Y

$$\begin{cases} c = \dfrac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(Y)} = \dfrac{\frac{1}{n}\sum_{i=1}^{n} x_i\, y_i - \overline{x}\,\overline{y}}{\frac{1}{n}\sum_{i=1}^{n} y_i^2 - \overline{y}^2} \\ d = \overline{x} - c\,\overline{y} \end{cases} \tag{4}$$

Ajustement d'un nuage de points

La droite D' d'ajustement ainsi obtenue a pour équation

$$x - \overline{x} = \frac{Cov(X,Y)}{Var(Y)}(y - \overline{y})\,. \tag{5}$$

Par ailleurs, en remplaçant $Cov(X,Y)$ par $r(X,Y)\sqrt{Var(X)\,Var(Y)}$ dans l'équation (5), on obtient une autre forme de l'équation de la droite D' donnée par

$$x - \overline{x} = r(X,Y)\frac{\sqrt{Var(X)}}{\sqrt{Var(Y)}}(y - \overline{y})$$

$$\iff \quad y - \overline{y} = \frac{1}{r(X,Y)}\frac{\sqrt{Var(Y)}}{\sqrt{Var(X)}}(x - \overline{x})\,. \tag{6}$$

Remarque

Notons que :

1) les deux droites D et D' passent par le point $(\overline{x}, \overline{y})$,

2) si $r(X, Y) = \pm 1$, d'après les équations (3) et (6), les droites D et D' sont confondues et que les n points (x_i, y_i), $i = 1, \ldots, n$ sont alignés.

3.2- Ajustement à l'aide d'une fonction exponentielle

Parfois, on remarque que les valeurs de la variable Y augmentent de façon exponentielle en fonction des valeurs de la variable X, c'est le cas par exemple des observations suivantes représentant le nombre Y d'articles vendus d'un nouveau produit durant les 10 premiers mois X (Tableau 5) :

X (Mois)	1	2	3	4	5	6	7	8	9	10
Y	2	5	8	17	32	64	128	257	512	1024

TABLE: (Tableau 5) Les ventes Y enregistrées d'un nouveau produit durant les 10 premiers mois X.

Ajustement d'un nuage de points

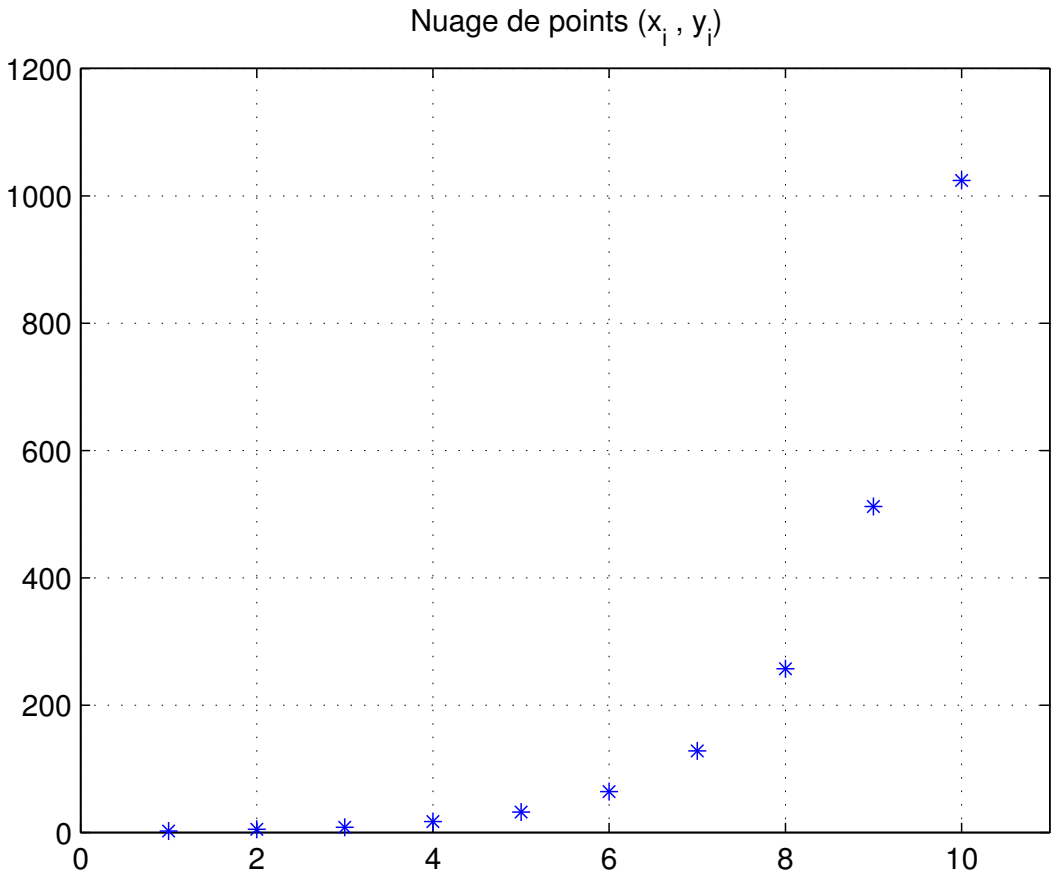

FIGURE: (Figure 2) Nuage de points (x_i, y_i) des données du Tableau 5.

Ajustement d'un nuage de points

Nous remarquons sur cet exemple que l'évolution de la variable Y est très rapide et que le nuage de points $\{(x_i, y_i),\ i = 1, \ldots, n\}$ ne peut être ajusté par une droite. La forme du nuage de points nous fait penser à un ajustement par une fonction exponentielle de la forme :

$$Y = B\, A^X\,. \tag{7}$$

Il s'agit donc de déterminer les paramètres A et B pour connaître l'expression de la fonction à ajuster au nuage de points.

Ajustement d'un nuage de points

Les coefficients A et B peuvent être déterminés en utilisant la droite d'ajustement par la méthode des moindres carrés par une transformation logarithmique de la relation (7) :

$$Log(Y) = X\,Log(A) + Log(B)\,.$$

Le nuage de points des variables $(X, Log(Y))$ est donc aligné quand le nuage de points des variables (X, Y) s'ajuste à une fonction exponentielle.
En effet, la relation précédente est l'équation d'une droite de la forme

$$Z = a\,X + b$$

où les nouvelles variables sont

$$\begin{cases} X &= X \\ Z &= Log(Y) \end{cases}$$

avec les coefficients

$$\begin{cases} a &= Log(A) \\ b &= Log(B)\,. \end{cases}$$

Ajustement d'un nuage de points

En utilisant la méthode d'ajustement de la droite des moindres carrés et d'après la relation (1), on a

$$\begin{cases} a = Log(A) = \dfrac{\mathrm{Cov}(X,Z)}{\mathrm{Var}(X)} = \dfrac{\frac{1}{n}\sum_{i=1}^{n} x_i\, z_i - \overline{x}\,\overline{z}}{\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \overline{x}^2} \\ b = Log(B) = \overline{z} - a\,\overline{x}\,, \end{cases} \tag{8}$$

où

$$z_i = Log(y_i) \qquad \overline{z} = \frac{1}{n}\sum_{i=1}^{n} z_i = \frac{1}{n}\sum_{i=1}^{n} Log(y_i)\,.$$

Une fois les valeurs de $a = Log(A)$ et $b = Log(b)$ déterminées, on en déduit celles de A et B par transformation exponentielle, c'est-à-dire

$$\begin{cases} A = e^a \\ B = e^b\,. \end{cases}$$

Ajustement d'un nuage de points

A titre d'application, reprenons l'exemple des données du Tableau 5 représentant les ventes Y enregistrées d'un nouveau produit durant les 10 premiers mois X.

La représentation graphique du nuage de points (Figure 2) montre clairement que l'ajustement le plus approprié est celui d'une fonction exponentielle de la forme

$$Y = B\,A^X\,.$$

Pour les calculs, on peut s'appuyer sur le Tableau 6 récapitulatif suivant pour déterminer les paramètres A et B de la fonction $Y = B\,A^X$.

Tableau 6 récapitulatif des ventes Y d'un nouveau produit

x_i	y_i	$z_i = Log(y_i)$	x_i^2	$x_i z_i$
1	2	0.6931	1	0.6931
2	5	1.6094	4	3.2189
3	8	2.0794	9	6.2383
4	17	2.8332	16	11.3329
5	32	3.4657	25	17.3287
6	64	4.1589	36	24.9533
7	128	4.8520	49	33.9642
8	257	5.5491	64	44.3926
9	512	6.2383	81	56.1449
10	1024	6.9315	100	69.3147
$\sum_{i=1}^{n} x_i = 55$	$\sum_{i=1}^{n} y_i = 2049$	$\sum_{i=1}^{n} z_i = 38.4108$	$\sum_{i=1}^{n} x_i^2 = 385$	$\sum_{i=1}^{n} x_i z_i = 267.5816$
$\frac{1}{n}\sum_{i=1}^{n} x_i = 5.5$	$\frac{1}{n}\sum_{i=1}^{n} y_i = 204.9$	$\frac{1}{n}\sum_{i=1}^{n} z_i = 3.84108$	$\frac{1}{n}\sum_{i=1}^{n} x_i^2 = 38.5$	$\frac{1}{n}\sum_{i=1}^{n} x_i z_i = 26.75816$

Ainsi les formules (8) nous donnent

$$\begin{cases} a = Log(A) = 0.6827 \\ b = Log(B) = 0.0862\,. \end{cases}$$

$$\Longrightarrow \quad \begin{cases} A = e^{a} = 1.9792 \\ B = e^{b} = 1.0900\,. \end{cases}$$

La courbe ajustée est donc :

$$Y = (1.0900)\,(1.9792^{X})\,.$$

3.3- Ajustement à l'aide d'une fonction puissance

Soit $\{(x_i, y_i)\,,\ i = 1, \ldots, n\}$ un nuage de points d'un couple (X, Y) de variables statistiques. Une fonction puissance s'écrit :

$$Y = B\,X^A\,, \quad \text{avec} \quad A > 0$$

soit, en prenant le logarithme népérien de cette expression, on obtient :

$$Log(Y) = A\,Log(X) + Log(B)\,.$$

On pose :

$$b = Log(B), \quad V = Log(X) \quad \text{et} \quad Z = Log(Y),$$

on obtient une équation d'une droite de la forme $Z = A\,V + b$.
Le nuage de points des variables $(V, Z) = (Log(X), Log(Y))$ est donc aligné quand le nuage de points des variables (X, Y) s'ajuste à une fonction puissance.

Ajustement d'un nuage de points

L'application de la méthode d'ajustement de la droite des moindres carrés fournit comme paramètres de la droite $Z = A\,V + b$:

$$\begin{cases} A & = & \dfrac{\mathrm{Cov}(V,Z)}{\mathrm{Var}(V)} = \dfrac{\dfrac{1}{n}\displaystyle\sum_{i=1}^{n} v_i\, z_i - \overline{v}\,\overline{z}}{\dfrac{1}{n}\displaystyle\sum_{i=1}^{n} v_i^2 - \overline{v}^2} \\ b = Log(B) & = & \overline{z} - A\,\overline{v}\,, \end{cases} \tag{9}$$

où

$$v_i = Log(x_i), \quad \overline{v} = \frac{1}{n}\sum_{i=1}^{n} v_i = \frac{1}{n}\sum_{i=1}^{n} Log(x_i),$$

$$z_i = Log(y_i) \;\text{ et }\; \overline{z} = \frac{1}{n}\sum_{i=1}^{n} z_i = \frac{1}{n}\sum_{i=1}^{n} Log(y_i)\,.$$