Cours de Probabilités & Statistique (L1 MI)

A. Menni

1[er] mars 2016

Chapitre2 : Statistique Descriptive Bivariée

---

1 Introduction.

La mesure de deux variables, désignées par X et Y, sur des unités statistiques donne lieu à une série statistique bivariée. L'analyse statistique d'une telle série s'effectue selon une démarche ayant de nombreux points communs avec celle suivie dans le cas univarié. Cependant, l'objectif poursuivi ici est double : il consiste à explorer, organiser et décrire les données afin de :

1. analyser les valeurs observées pour X d'une part et pour Y d'autre part.
2. analyser le lien éventuel entre les valeurs prises par X et celles prises par Y.

La série statistiques bivariée (ou double) est alors une suite de n couples de valeurs $\{(x_i, y_i)\ ;\ i = 1, 2, \ldots, n\}$ prises par les deux variables statistiques X et Y sur chacun des n individus.

2 Tableau de contingence

Ce tableau est intéressant à construire lorsqu'on observe à plusieurs reprises les mêmes couples de valeurs. Cette situation se présente surtout lorsque n est élevé, que les variables sont qualitatives ou quantitatives et que le nombre de valeurs distinctes de chaque variable est faible.

Le tableau de contingence (des effectifs) s'obtient alors en associant à chaque couple de valeurs distinctes observées l'effectif correspondant au nombre de fois qu'il est apparu. Nous désignerons par $x_1,\ x_2, \ldots, x_p$ et $y_1,\ y_2, \ldots, y_q$ les valeurs ou modalités distinctes des variables X et Y respectivement, ou encore les centres de classe en cas de groupement des données.

Remarque

Nous limitons l'étude dans ce cours au cas où les variables X et Y sont quantitatives.

X \ Y	y_1	y_2	$\dots$	y_j	$\dots$	y_q	Distribution marginale de X
x_1	n_{11}	n_{12}	$\dots$	n_{1j}	$\dots$	n_{1q}	$n_{1\bullet}$
x_2	n_{21}	n_{22}	$\dots$	n_{2j}	$\dots$	n_{2q}	$n_{2\bullet}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
x_i	n_{i1}	n_{i2}	$\dots$	n_{ij}	$\dots$	n_{iq}	$n_{i\bullet}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
x_p	n_{p1}	n_{p2}	$\dots$	n_{pj}	$\dots$	n_{pq}	$n_{p\bullet}$
Distribution marginale de Y	$n_{\bullet 1}$	$n_{\bullet 2}$	$\dots$	$n_{\bullet j}$	$\dots$	$n_{\bullet q}$	$n_{\bullet\bullet} = n$

Exemple

On a mesuré la taille X et le poids Y de 20 individus. Ci-après les résultats obtenus :

x_i	y_i	x_i	y_i
155	60	180	75
162	61	175	76
157	64	173	78
170	67	175	80
164	68	179	85
162	69	175	90
169	70	180	96
170	70	185	96
178	72	187	98
173	73	187	99

Tableau 1 – Tailles et poids de 20 individus

Les classes d'une même variable étant de même amplitude. Compléter le tableau de contingence ci-dessous :

X \ Y		[60 - 68[	[68 - 76[	[76 - 84[	[84 - 92[	[92 - 100[	$n_{i\bullet}$ → marginal
	x_i \ y_j	64	72	80	88	96	
[155 - 163[	159	3 n_{11}	1 n_{12}				4 $n_{1\bullet}$
[163 - 171[	167	1 n_{21}	3 n_{22}				4 $n_{2\bullet}$
[171 - 179[	175		2 n_{32}	3 n_{34}	1 n_{35}		6 $n_{3\bullet}$
[179 - 187]	183		1 n_{42}		1 n_{45}	4 n_{46}	6 $n_{4\bullet}$
$n_{\bullet j}$	X	4 $n_{\bullet 1}$	7 $n_{\bullet 2}$	3 $n_{\bullet 3}$	2 $n_{\bullet 4}$	4 $n_{\bullet 5}$	

Tableau 2 – Tableau de contingence Tailles-poids $n = 20$

- $n_{32} = 2$ nombre d'individus ayant une taille comprise entre 171 et 179 cm et un poids comprise entre 68 et 76 Kg

- $n_{3\bullet}$ (3ème effectif marginal) $= 6 \Rightarrow$ nombre d'individus pour lesquels la taille est comprise entre 171 et 179 cm

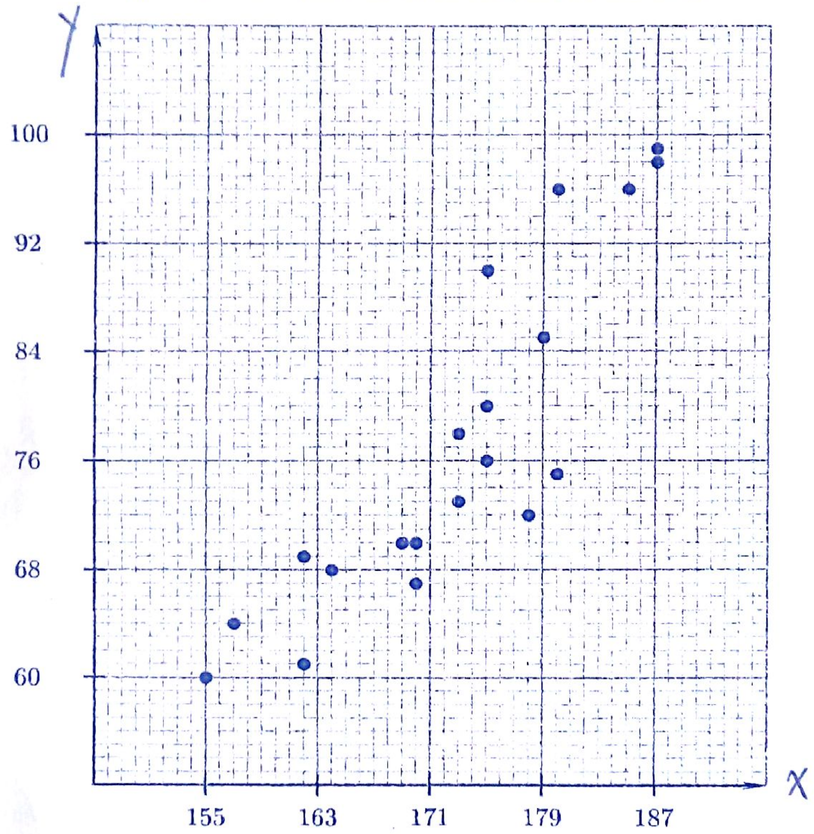
- $n_{\bullet 2} = 7$ nombre d'individus pour lesquels le poids et compris entre 68 et 76 Kg

3 Nuage de points

Si les deux variables X et Y sont quantitatives, une manière simple de visualiser les données consiste à représenter chaque individu par un point dans le plan $\mathbb{R}^2$, de coordonnées (x_i, y_i) ; $i = 1, 2, \ldots, n$.
Cette représentation graphique des données porte le nom de nuage de points. Puisque ce graphique nous indique la façon dont les points se dispersent dans le plan, on lui donne aussi le nom de *graphique de dispersion*.

Exemple

Tracer le nuage de points représentant les données du tableau1.



4 Les distributions

4.1 La distribution conjointe

Elle est définie par l'ensemble des triplets $\{(x_i, y_j, n_{ij}); i = 1, 2, \ldots, p \text{ et } j = 1, 2, \ldots, p\}$.
L'effectif conjoint n_{ij} donné dans la cellule (i, j) est le nombre d'individus présentant simultanément les modalités x_i de X et y_j de Y.

La somme des effectifs n_{ij} est bien sûr égale à la taille de la série statistique bivariée de départ :

$$\sum_{i=1}^{p}\sum_{j=1}^{q} n_{ij} = n$$

On parle également de distribution conjointe des fréquences relatives $f_{ij} = \frac{n_{ij}}{n}$ dont la somme est évidemment égale 1 :

$$\sum_{i=1}^{p}\sum_{j=1}^{q} f_{ij} = 1$$

4.2 Les distributions marginales

L'étude d'une série bivariée comporte en premier lieu l'analyse des séries marginales univariées obtenues en ne considérant qu'une seule variable à la fois :

la série marginale en X : $x_1,\ x_2,\ \dots,\ x_p$
la série marginale en Y : $y_1,\ y_2,\ \dots,\ y_q$

La distribution marginale de X est définie par l'ensembles des couples $\{(x_i, n_{i\bullet})\ ;\ i = 1, 2, \dots, p\}$ où l'on associe aux modalités x_i $(i = 1, 2, \dots, p)$ les effectifs marginaux définis par

$$n_{i\bullet} = \sum_{j=1}^{q} n_{ij}$$

Il va de soi que la distribution marginale des fréquences relatives de X est définie par l'ensemble $\{(x_i, f_{i\bullet})\ ;\ i = 1, 2, \dots, p\}$ où

$$f_{i\bullet} = \frac{n_{i\bullet}}{n}$$

De façon similaire, on définit la distribution marginale de la variable Y par l'ensemble des couples $\{(y_j, n_{\bullet j})\ ;\ j = 1, 2, \dots, q\}$.

$$n_{\bullet j} = \sum_{i=1}^{p} n_{ij}$$

La distribution marginale des fréquences relatives de Y est alors définie par l'ensemble $\{(y_j, f_{\bullet j})\ ;\ j = 1, 2, \dots, q\}$ où

$$f_{\bullet j} = \frac{n_{\bullet j}}{n}$$

4.3 Les distributions conditionnelles

Une distribution conditionnelle consiste à fixer a priori la valeur d'une des deux variables et à décrire la distribution des valeurs de l'autre variable.
En effet, la $j^{ème}$ colonne du tableau statistique de contingence décrit la sous-population des individus possédant la modalité y_j (ou appartenant à la $j^{ème}$ classe dans le cas d'un caractère continu) suivant le caractère X. La fréquence conditionnelle de la modalité x_i sachant y_j est donnée par :

$$f_{i/j} = \frac{f_{ij}}{f_{\bullet j}} = \frac{n_{ij}}{n_{\bullet j}} \quad ; \quad i = 1, 2, \dots, p \quad et \quad j \ fixé\ dans\ \{1, 2, \dots, q\}$$

la valeur encadrée, veut dire que parmi les individus dont le poids est compris entre 60 et 68 Kg, 25% ont une taille dans l'intervalle [163, 171[

De manière analogue est décrite la distribution conditionnelle de Y sachant x_i :

$$f_{j/i} = \frac{f_{ij}}{f_{i\bullet}} = \frac{n_{ij}}{n_{i\bullet}} \quad ; \quad j = 1, 2, \ldots, q \quad et \quad i \; fixé \; dans \; \{1, 2, \ldots, p\}$$

On constate que $f_{ij} = f_{i\bullet} f_{j/i} = f_{\bullet j} f_{i/j}$ et montre facilement que $\sum_{i=1}^{p} f_{i/j} = \sum_{j=1}^{q} f_{j/i} = 1$.

Exemples

1. Distribution de X sachant que $Y \in [60 - 68[$.

Classes de X	[155 - 163[	[163 - 171[	[171 - 179[	[179 - 187]
$f_{X/Y} \in [60\text{-}68[$	$3/4$	$\boxed{1/4}$	0	0

on doit calculer $f_{i/j=1} = \frac{n_{i1}}{n_{\bullet 1}}$ pour $i = 1 \ldots 4$

2. Distribution de Y sachant que $X \in [171 - 179[$.

Classes de Y	[60 - 68[	[68 - 76[	[76 - 84[	[84 - 92[	[92 - 100[
$f_{Y/X} \in [171\text{-}179[$	0	$\boxed{2/6}$	$3/6$	$1/6$	0

on doit calculer $f_{j/i=3} = \frac{n_{3j}}{n_{3\bullet}}$ pour tout $j \in \{1 \ldots 5\}$

5 Indépendance entre deux variables statistiques

Les deux variables X et Y sont dites indépendantes ssi

$$f_{ij} = f_{i\bullet} f_{\bullet j} \;\; \forall (i, j) \in \{1, 2, \ldots, p\} \times \{1, 2, \ldots, q\}$$

On peut également vérifier l'indépendance entre X et Y à partir des distributions conditionnelles : il y a indépendance entre X et Y ssi

$$pour \;\; tout \;\; i \;\; on \;\; a \;\; f_{i/j} = f_{i\bullet} \;\; \forall j \in \{1, 2, \ldots, q\}$$

ou bien

$$pour \;\; tout \;\; j \;\; on \;\; a \;\; f_{j/i} = f_{\bullet j} \;\; \forall i \in \{1, 2, \ldots, p\}$$

Exemple de variables indépendantes. $\perp\!\!\!\perp$

X \ Y	1	2	3	$n_{i\bullet}$
[0 , 10[	8	4	4	16
[10 , 20]	4	2	2	8
$n_{\bullet j}$	12	6	6	$n = 24$

$X \perp\!\!\!\perp Y \iff \forall (i,j), \; f_{ij} = f_{i\bullet} \times f_{\bullet j}$

$\frac{8}{24} = \frac{16}{24} \times \frac{12}{24} =$

$\frac{4}{24} \stackrel{?}{=} \frac{16}{24} \times \frac{6}{24}$

$\frac{4}{24} \stackrel{?}{=} \frac{8}{24} \times \frac{12}{24}$

$\frac{2}{24} \stackrel{?}{=} \frac{8}{24} \times \frac{6}{24}$

Donc $X \perp\!\!\!\perp Y$

X = moyenne generale annuelle

Y = nombre de frère et soeurs

la valeur encadrées veut dire que parmi les individus dont la taille comprise entre 171 et 179 cm, environ 33,33% ont un poids qui est comprise entre 60 et 76 kg

6 Paramètres marginaux

6.1 Les moyennes arithmétiques

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{p} n_{i\bullet}x_i$$

$$\bar{Y} = \frac{1}{n}\sum_{j=1}^{q} n_{\bullet j}y_j$$

6.2 Les variances

$$V(X) = \sigma_X^2 = \frac{1}{n}\sum_{i=1}^{p} n_{i\bullet}(x_i - \bar{X})^2 = \bar{X^2} - (\bar{X})^2$$

$$V(Y) = \sigma_Y^2 = \frac{1}{n}\sum_{j=1}^{q} n_{\bullet j}(y_j - \bar{Y})^2 = \bar{Y^2} - (\bar{Y})^2$$

Exemple

X \ Y		[60 - 68[	[68 - 76[	[76 - 84[	[84 -92[	[92 - 100[	$n_{i\bullet}$	$n_{i\bullet}x_i$	$n_{i\bullet}x_i^2$
	x_i \ y_j	64	72	80	88	96			
[155 - 163[	159	2							
[163 - 171[	167								
[171 - 179[	175								
[179 - 187]	183								
$n_{\bullet j}$									
$n_{\bullet j}y_j$									
$n_{\bullet j}y_j^2$									

6.3 Les moyennes arithmétiques conditionnelles

Moyenne de X sachant que $Y = y$: $\bar{X}_{/Y=y} = \sum_{i=1}^{p} f_{i/j}x_i$

Moyenne de Y sachant que $X = x$: $\bar{Y}_{/X=x} = \sum_{j=1}^{q} f_{j/i}y_j$

6.4 Les variances conditionnelles

Variance de X sachant que $Y = y$:

$$V(X/Y=y) = \sigma^2_{X/Y=y} = \sum_{i=1}^{p} f_{i/j}(x_i - \bar{X}_{/Y=y})^2 = \sum_{i=1}^{p} f_{i/j}x_i^2 - (\bar{X}_{/Y=y})^2$$

Variance de Y sachant que $X = x$:

$$V(Y/X=x) = \sigma^2_{Y/X=x} = \sum_{j=1}^{q} f_{j/i}(y_j - \bar{Y}_{/X=x})^2 = \sum_{j=1}^{q} f_{j/i}y_j^2 - (\bar{Y}_{/X=x})^2$$

Remarque

Si X et Y sont indépendantes alors $\bar{X}_{/Y=y} = \bar{X}$ et $\bar{Y}_{/X=x} = \bar{Y}$.

Exemple

Calculer la moyenne et la variance de X sachant que $Y \in [60-68[$

Classes	[155 - 163[	[163 - 171[	[171 - 179[	[179 -187]
x_i	159	167	175	183
$f_{X/Y\in[60-68[}$				
$x_i f_{X/Y\in[60-68[}$				
$x_i^2 f_{X/Y\in[60-68[}$				

7 Covariance et coefficient de corrélation

7.1 La covariance

La covariance de la série statistique $\{(x_i, y_i) \ ; \ i = 1, 2, \ldots, n\}$ est désignée par S_{XY} et est définie par l'expression

$$S_{XY} = \frac{1}{n}\sum_{i=1}^{p}\sum_{j=1}^{q} n_{ij}(x_i - \bar{X})(y_j - \bar{Y}) = \sum_{i=1}^{p}\sum_{j=1}^{q} f_{ij}(x_i - \bar{X})(y_j - \bar{Y})$$

On montre que

$$S_{XY} = \frac{1}{n}\sum_{i=1}^{p}\sum_{j=1}^{q} n_{ij}x_i y_j - \bar{X}\bar{Y} = \sum_{i=1}^{p}\sum_{j=1}^{q} f_{ij}x_i y_j - \bar{X}\bar{Y}$$

Propriétés

1. $S_{XY} \in \mathbb{R}$
2. La covariance est symétrique : $S_{XY} = S_{YX}$
3. $S_{XX} = V(X)$
4. $V(X+Y) = V(X) + V(Y) + 2S_{XY}$
5. $S^2_{XY} \leq V(X)V(Y)$

7.2 Le coefficient de corrélation linéaire

C'est un indice rendant compte numériquement de la manière dont les deux variables X et Y varient simultanément (il mesure le degré de liaison entre X et Y). Il est noté par ρ_{XY} et défini par l'expression

$$\rho_{XY} = \frac{S_{XY}}{\sigma_X \sigma_Y}$$

Propriétés

1. $\rho_{XY} = \rho_{YX}$
2. $\rho_{XY} \in [-1, 1]$
3. $|\rho_{XY}| = 1 \Leftrightarrow$ il existe une liaison parfaitement linéaire entre X et Y
4. X et Y sont indépendantes $\Rightarrow \rho_{XY} = 0$

Attention !

$\rho_{XY} = 0 \nRightarrow X$ et Y sont indépendantes.

Exemple

X \ Y		[60 - 68[	[68 - 76[	[76 - 84[	[84 -92[	[92 - 100[	$n_{i\bullet}$
	x_i \ y_j	64	72	80	88	96	
[155 - 163[	159						
[163 - 171[	167						
[171 - 179[	175						
[179 - 187]	183						
$n_{\bullet j}$							

8 La régression linéaire simple

8.1 Droite de régression de Y en X

Le coefficient de corrélation linéaire mesure la dépendance linéaire entre les deux variables Y et X. Si la valeur de ce coefficient est assez importante, on peut exprimer la variable Y comme fonction affine de X. Théoriquement on écrit :

$$(D_{Y/X}) : Y = aX + b$$

Les paramètres a et b sont déterminés de telle sorte que la droite de régression passe au plus près possible de tous les points dans le diagramme de dispersion. La méthode utilisée pour l'estimation de ces deux paramètres est appelée " méthode des Moindres Carrés Ordinaires " (notée en abrégé MCO). Elle consiste à trouver les paramètres a et b qui minimisent le carré des distances entre chaque points (x_i, y_i) et sa projection sur la droite parallèlement a l'axe des ordonnées $(\overrightarrow{OY})$. Evidemment, le paramètre a définit la pente de la droite alors que b donne la valeur de l'ordonnée à l'origine de la droite de régression (appelée également droite d'ajustement). Comme la projection des points du nuage se fait parallèlement à l'axe des ordonnées, la distance entre un point (x_i, y_i) et se projection sur la droite est donnée par

$$y_i - \hat{y}_i = y_i - ax_i - b$$

La somme des carrés de ces distances qu'il faut donc minimiser est

$$S(a,b) = \sum_{i=1}^{n} (y_i - ax_i - b)^2$$

$S(a,b)$ atteint son minimum lorsque ses dérivées partielles par rapport à a et b s'annullent :

$$\frac{\partial S(a,b)}{\partial a} = \frac{\partial S(a,b)}{\partial b} = 0 \Rightarrow \hat{a} = \frac{S_{XY}}{V(X)} \quad et \quad \hat{b} = \bar{y} - \hat{a}\bar{x}$$

La forme du paramètre estimé $\hat{b}$ permet de constater que la droite d'ajustement passe par le point moyen (appelé aussi "centre de gravité") de coordonnées $\bar{x}$ et $\bar{y}$.
L'équation de cette droite devient donc $\hat{y} = \hat{a}x + \hat{b}$.

8.2 Droite de régression de X en Y

En inversant les rôles des deux variables X et Y, et en effectuant le même calcul que précédemment, on définit la droite de régression de X en Y d'équation : $\hat{x} = \hat{a}'y + \hat{b}'$ avec

$$\hat{a}' = \frac{S_{XY}}{V(Y)} \quad et \quad \hat{b}' = \bar{x} - \hat{a}'\bar{y}$$

Remarques

— Les deux droites de régression sont en général différentes, mais on ne peut dire laquelle des deux donne un meilleur ajustement.

— Le carré du coefficient de corrélation linéaire (appelé coefficient de détermination et noté R^2) est le produit des deux pentes :

$$R^2 = \rho^2_{XY} = \hat{a}\hat{a}'$$

- Pour valider (ou rejeter) l'ajustement, on compare la valeur du coefficient de détermination R^2 avec 1. Plus il est proche de 1 mieux c'est.
- Si $R^2 = 1$ alors les points dans le graphe de dispersion sont parfaitement alignés.
- Si $R^2 = 0$ alors les variables X et Y sont non corrélés.
- Si $\rho_{XY} > 0$ alors les variables X et Y varient dans le même sens.
- Si $\rho_{XY} < 0$ alors les variables X et Y varient dans deux sens différents.

8.3 Prévision

- La droite de régression de Y en X permet de prédire une valeur inconnue de Y lorsqu'une nouvelle valeur x_{n+1} de X, n'ayant pas participé à la construction de $D_{Y/X}$, est donnée :

$$\hat{y}_{n+1} = \hat{a}x_{n+1} + \hat{b}$$

- La droite de régression de X en Y permet de prédire une valeur inconnue de X lorsqu'une nouvelle valeur y_{n+1} de Y, n'ayant pas participé à la construction de $D_{X/Y}$, est donnée :

$$\hat{x}_{n+1} = \hat{a}'y_{n+1} + \hat{b}'$$

9 Ajustement non linéaire

Dans certains cas, l'ajustement à une fonction affine n'est pas adéquat et le coefficient de corrélation linéaire prend une valeur non significative. Nous présentons dans ce cours deux cas où un ajustement non linéaire est envisagé, mais à l'aide d'un simple changement de variable, la méthode des MCO nous permet d'effectuer cet ajustement.

9.1 Ajustement par une fonction puissance

Lorsqu'on constate, sur le graphique de dispersion, que les deux variables X et Y sont éventuellement liées l'une à l'autre par une relation de la forme $Y = \beta X^\alpha$, le changement de variable suivant nous permet l'utilisation de la méthode des MCO pour l'estimation des deux paramètres β et α :

$$Z = \ln Y,\ W = \ln X,\ a = \alpha \ et\ b = \ln \beta$$

La droite d'ajustement (affine) est ainsi obtenu sur les nouvelles variables W et Z :

$$Z = \hat{a}W + \hat{b}$$

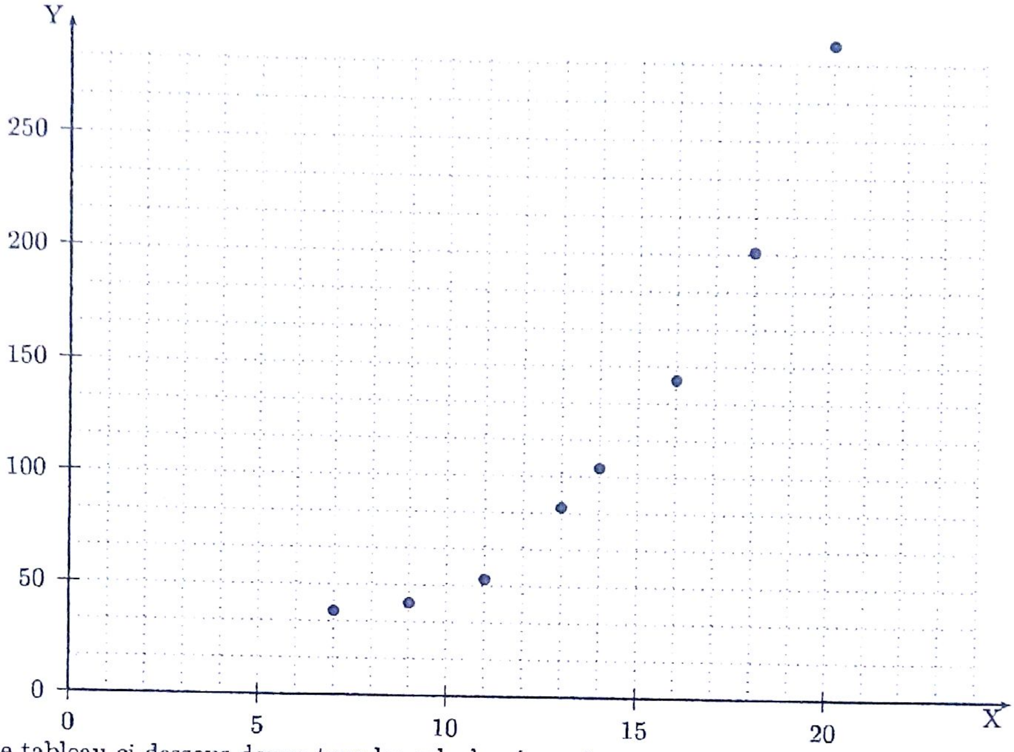
Et la fonction, à laquelle les données originales $\{(x_i, y_i);\ i = 1, 2, \ldots, n\}$ sont ajustées, est estimée par : $Y = e^{\hat{b}}X^{\hat{a}}$ où $\hat{a}$ et $\hat{b}$ sont les estimations des MCO de a et b respectivement.

Exemple

Un responsable logistique a effectué 8 observationx mesurant le temps de préparation d'une commande en minutes et le nombre de colis à préparer :

nombre de colis	7	9	11	13	14	16	18	20
temps de préparation(en minutes)	38	42	53	86	104	144	201	292

En examinant le nuage de points, nous nous proposons d'effectuer un ajustement linéaire ainsi qu'un ajustement à une fonction puissance pour voir lequel des deux ajustements est meilleur.



Le tableau ci-dessous donne tous les calculs nécessaires.

	nombre de colis	temps de préparation	W=ln X	Z=ln Y	X²	Y²	W²	Z²	XY	WZ
	7	38	1,9459	3,6376	49	1444	3,7866	13,2320	266	7,0784
	9	42	2,1972	3,7377	81	1764	4,8278	13,9702	378	8,2125
	11	53	2,3979	3,9703	121	2809	5,7499	15,7632	583	9,5203
	13	86	2,5649	4,4543	169	7396	6,5790	19,8412	1118	11,4252
	14	104	2,6391	4,6444	196	10816	6,9646	21,5704	1456	12,2568
	16	144	2,7726	4,9698	256	20736	7,6872	24,6990	2304	13,7792
	18	201	2,8904	5,3033	324	40401	8,3542	28,1250	3618	15,3285
	20	292	2,9957	5,6768	400	85264	8,9744	32,2255	5840	17,0060
moyenne	13,5	120	2,5505	4,5493	199,5000	21328,7500	6,6155	21,1783	1945,3750	11,8259

La droite de régression de Y en X est estimée par $D_{Y/X}$: $Y = 18.862X - 134.641$, avec un coefficient de corrélation linéaire $\rho_1 = 0.941$ et donc un coefficient de détermination $R^2 = 0.886$

La droite de régression de Z en W est estimée par $D_{Z/W}$: $Z = 2.018X - 0.596$, avec un coefficient de corrélation linéaire $\rho_2 = 0.966$ et un coefficient de détermination $R^2 = 0.933$

La courbe d'ajustement à une fonction puissance est donc meilleure et elle est obtenue en posant $\alpha = 2.018$ *et* $\beta = e^{-0.596} = 0.551$
Autrement dit, $Y = 0.551X^{2.018}$

9.2 Ajustement par une fonction exponentielle

Si le graphique de dispersion nous donne l'impression que les deux variables X et Y sont liées l'une à l'une par une relation non linéaire de la forme : $Y = \beta e^{\alpha X}$, le changement de variable suivant nous permet l'utilisation de la méthode des MCO pour l'estimation de β et α :

$$Z = \ln Y, \;\; a = \alpha \;\; \text{et} \;\; b = \ln \beta$$

La droite d'ajustement pour le nouveau couple de variables (X, Z) est donnée par :

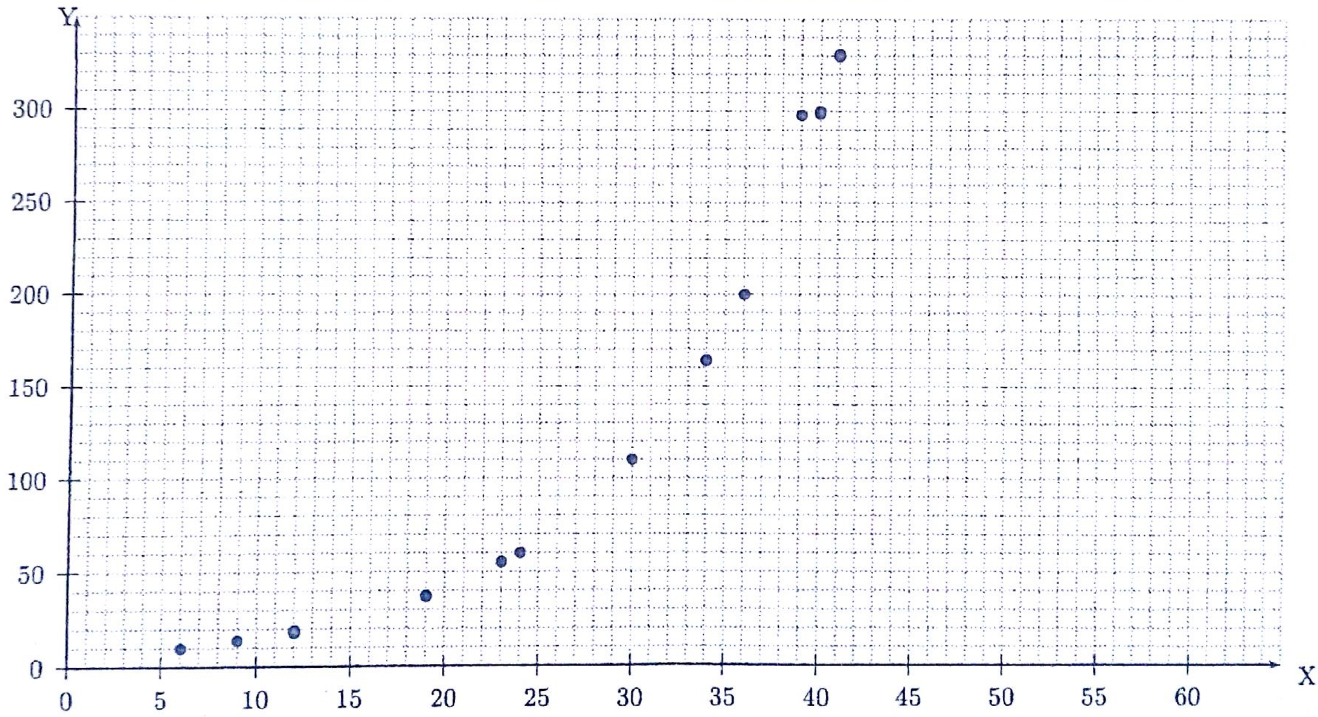
$$Z = \hat{a}X + \hat{b}$$

Et la fonction, à laquelle les données originales $\{(x_i, y_i);\ i = 1, 2, \ldots, n\}$ sont ajustées, est estimée par : $Y = \hat{a}X + e^{\hat{b}}$ où $\hat{a}$ et $\hat{b}$ sont toujours les estimations des MCO de a et b respectivement.

Exemple

Le tableau ci-dessous donne l'évolution du montant des ventes d'une certaine marque de téléphones portables (en millions de dinars) entre 1999 et 2013 en fonction des frais publicitaires encourus.

année	99	00	01	02	03	04	05	06	07	08	09	10	11	12	13
X	6	9	12	12	19	23	24	30	34	36	39	40	40	41	41
Y	10	14	18	19	37	55	60	110	164	200	298	299	300	330	331



Le tableau ci-dessous donne tous les calculs nécessaires.

	Frais. Pub (X)	Montant. Ventes (Y)	Z=lnY	X²	Z²	XZ
	6	10	2,3026	36	5,3019	13,8155
	9	14	2,6391	81	6,9646	23,7515
	12	18	2,8904	144	8,3542	34,6845
	12	19	2,9444	144	8,6697	35,3333
	19	37	3,6109	361	13,0387	68,6074
	23	55	4,0073	529	16,0587	92,1687
	24	60	4,0943	576	16,7637	98,2643
	30	110	4,7005	900	22,0945	141,0144
	34	164	5,0999	1156	26,0086	173,3955
	36	200	5,2983	1296	28,0722	190,7394
	39	298	5,6971	1521	32,4569	222,1866
	40	299	5,7004	1600	32,4951	228,0177
	40	300	5,7038	1600	32,5331	228,1513
	41	330	5,7991	1681	33,6295	237,7628
	41	331	5,8021	1681	33,6646	237,8869
moyennes	27,0667	149,6667	4,4193	887,0667	21,0737	135,0520

L'équation de la courbe de régression exponentielle de Y en X est alors donnée par :

$$Y = 5.5434e^{0.1X}$$

9.3 Ajustement par une fonction Logarithmique

Si maintenant le nuage de points (x_i, y_i) semble suivre le graphique d'une fonction logarithmique du type $Y = \alpha \ln(X) + \beta$, le changement de variable suivant nous permet d'estimer α et β à l'aide des MCO :

$$Z = Y \text{ et } W = \ln(X)$$

Exemple

Considérons la série statistique double suivante :

X	1,2	2,5	3,9	4,8	4,8	5,1	5,2	5,7	6,8	7,7	8,8	8.9	9,8	10,6	11,5	12,0	12,7	13,4	13,4	14,2
Y	0	14,1	22,7	26,4	26,7	27,7	28,1	29,8	33,3	35,6	38,3	38,5	40,4	41,7	43,5	44,2	45,3	46,3	46.3	46,9

Le tableau ci-dessous donne tous les calculs nécessaires.

																					moyenne
X	1,2	2,5	3,9	4,8	4,8	5,1	5,2	5,7	6,8	7,7	8,8	8,9	9,8	10,6	11,5	12	12,7	13,4	13,4	14,2	8,15
Y	0	14,1	22,7	26,4	26,7	27,7	28,1	29,8	33,3	35,6	38,3	38,5	40,4	41,7	43,5	44,2	45,3	46,3	46,3	46,9	33,79
W = lnX	0,18	0,92	1,36	1,57	1,57	1,63	1,65	1,74	1,92	2,04	2,17	2,19	2,28	2,36	2,44	2,48	2,54	2,60	2,60	2,65	1,94
W^2	0,03	0,84	1,85	2,46	2,46	2,65	2,72	3,03	3,67	4,17	4,73	4,78	5,21	5,57	5,97	6,17	6,46	6,74	6,74	7,04	4,16
Y^2	0	198,81	515,29	696,96	712,89	767,29	789,61	888,04	1108,89	1267,36	1466,89	1482,25	1632,16	1738,89	1892,25	1953,64	2052,09	2143,69	2143,69	2199,61	1282,515
WY	0,00	12,92	30,89	41,41	41,88	45,13	46,33	51,87	63,83	72,67	83,29	84,16	92,21	98,45	106,24	109,83	115,13	120,16	120,16	124,44	71,05