Mohamed Boudiaf University - M'sila
Faculty of Technology
Department of Civil Engineering- Department of Electrical Engineering
Module: Probability-Statistics

Chapter 2: Bivariate statistical series

Merini Abdelaziz*

10 novembre 2024

Table des matières

*

1 Introduction

In this chapter, you will learn about the statistical methods used to analyse data recorded about two related variables, such asa person's weight and height. Such data is called **bivariate data** (two-variable data).
When we analyse bivariate data, we are interested in how the two variables relate to each other.
We try to answer questions such as :
'Is there a relationship between these two variables ? and 'Does knowing the value of one of the variables tell us anything about the value of the second variable ?

Example 1. **1)** *Relationship between the average number of hours of study and the marks obtained in exams.*

2) *The relationship between the height and weight of newborn babies in a maternity hospital*

2 Data tables (scatter plots - contingency table)

Definition 1. *A bivariate statistical series X and Y is any list of pairs of the type :*

$$(x_1, y_1), (x_2, y_2), ..., (x_n, y_n), \text{ where } n \text{ is the number of pairs.}$$

Remark 1. *We can represent the bivariate statistical series in one of the following two tables :*

1

Observation	*1*	*2*	...	*n*
Variable X	x_1	x_2	...	x_n
Variable Y	y_1	y_2	...	y_n

Or

Variable X	x_1	x_2	...	x_n
Variable Y	y_1	y_2	...	y_n

Example 2. *The data below gives the marks (y_i) out of 100 obtained by the students in an exam and the time (x_i) they spent studying for it.*

student number i	1	2	3	4	5	6	7	8	9	10
Time (hours), x_i	4	36	23	19	1	11	18	13	18	8
Marks y_i	41	87	67	62	23	52	61	43	64	52

2.1 Scatter plots

Bivariate data are usually represented graphically on scatterplots.

Definition 2. *A **scatter plot** is a graph that shows whether there is a relationship between two variables x and y.*
Each data value (x_i, y_i) on a scatter plot is shown by a point on a Cartesian plane.

Example 3. *Referring to example 2,*

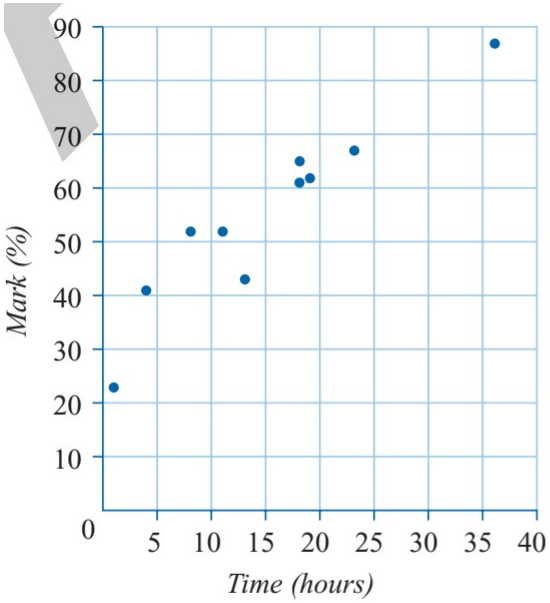


FIGURE 1 – The scatter plot shows relationship between the marks and time they spent studying for it

2.1.1 The mean point

Definition 3. *The mean point of a scatter plots is the point G with coordinates* $(\bar{x}, \bar{y})$*, where* $\bar{x} = \dfrac{\Sigma x_i}{n}$ *and* $\bar{y} = \dfrac{\Sigma y_i}{n}$

Example 4. *Referring to example 2,* $\bar{x} = \dfrac{\Sigma x_i}{n} = 15$ *and* $\bar{y} = \dfrac{\Sigma y_i}{n} = 55.6$

2.2 Contingency table

Let the statistical variable Z be given by the pair (X, Y). Let $x_1, x_2, ...x_k$ and $y_1, y_2, ...y_l$ be the values taken by X and Y respectively. In this case, we define the values of Z as follows : for i ranging from 1 to k and for j ranging from 1 to l, $z_{ij} := (x_i; y_j)$: The statistical variable Z takes k × l values.

Definition 4. *During this study, we use the following During this study, we use the following or contingency table (double-entry table)*

$X \diagdown Y$	y_1	y_2	...	y_l	Marginal X
x_1	n_{11} or f_{11}	n_{12} ou f_{12}	...	n_{1l} ou f_{1l}	$n_{1\bullet}$ or $f_{1\bullet}$
x_2	n_{21} or f_{21}	n_{22} or f_{22}	...	n_{2l} ou f_{2l}	$n_{2\bullet}$ ou $f_{2\bullet}$
$\vdots$	...	...	...	...	$\vdots$
x_k	n_{k1} or f_{k1}	n_{k2} or f_{k2}	...	n_{kl} ou f_{kl}	$n_{k\bullet}$ or $f_{k\bullet}$
Marginale Y	$n_{\bullet 1}$ or $f_{\bullet 1}$	$n_{\bullet 2}$ ou $f_{\bullet 2}$	...	$n_{\bullet l}$ or $f_{\bullet l}$	N or 1

with

n_{ij} : frequency of the pair (x_i, y_j).

$n_{i\bullet}$: marginal frequency of x_i is given by

$$n_{i\bullet} = \sum_{j=1}^{l} n_{ij} = n_{i1} + n_{i2} + ... + n_{il} \ = \text{total de la ligne i}$$

$n_{\bullet j}$:marginal frequencye of y_j is given by

$$n_{\bullet j} = \sum_{i=1}^{k} n_{ij} = n_{1j} + n_{2j} + ... + n_{kj} \ = \text{total de la colonne j}$$

f_{ij} :the relative frequency of the pair (x_i, y_j), is given by $f_{ij} = \dfrac{n_{ij}}{N}$

$f_{i\bullet}$: the relative frequency of x_i is given by $f_{i\bullet} = \dfrac{n_{i\bullet}}{N}$.

$f_{\bullet j}$:the relative frequency of y_j is given by $f_{\bullet j} = \dfrac{n_{\bullet j}}{N}$

Remark 2.

$$\begin{aligned}
\sum_{i=1}^{k}\sum_{j=1}^{l} n_{ij} &= \sum_{i=1}^{k}(n_{i1} + n_{i2} + ... + n_{il}) = \sum_{i=1}^{k} n_{i1} + \sum_{i=1}^{k} n_{i2} + ... + \sum_{i=1}^{k} n_{il} \\
&= (n_{11} + n_{21} + ... + n_{kl}) + (n_{12} + n_{22} + ... + n_{2k}) + ... + (n_{1l} + n_{2l} + ... + n_{kl}) \\
&= N
\end{aligned}$$

Remark 3. *We have the following property,*

$$\sum_{i=1}^{k}\sum_{j=1}^{l} f_{ij} = 1$$

Example 5. *A survey of 500 families was carried out. It looked at the relationship between the number of children in families (Y) and their annual expenditure on school supplies in Dinars (X). The survey produced the following results :*

$X \diagdown Y$	1	2	3	*marginal* $n_{i\bullet}$
$[0; 20000[$	80	54	13	147
$[20000; 40000[$	56	97	43	186
$[40000; 70000[$	10	71	76	157
Marginal $n_{\bullet j}$	146	222	132	500

3 Marginal and conditional distributions

3.1 Marginal distributions

On the margin of the contingency table, we can extract data solely with respect to X and solely with respect to Y (see the previously established contingency table).
The k pairs $(x_i, n_{i\bullet})$ form the marginal distribution of the variable X :

X	x_1	x_2	...	x_k	Σ
$n_{i\bullet}$	$n_{1\bullet}$	$n_{2\bullet}$	...	$n_{k\bullet}$	N

The l pairs $(y_j, n_{\bullet j})$ form the marginal distribution of the variable Y :

y	y_1	y_2	...	y_l	total
$n_{\bullet j}$	$n_{\bullet 1}$	$n_{\bullet 2}$	...	$n_{\bullet l}$	N

Example 6. *Referring to example 5, the marginal distribution of the number of children Y :*

Y	1	2	3	Σ
$n_{\bullet j}$	146	222	132	500

The marginal distribution of X

X	$n_{i\bullet}$
$[0; 20000[$	147
$[20000; 40000[$	186
$[40000; 70000[$	157
Σ	500

3.1.1 independent

Definition 5. *Two statistical variables X and Y are said to be independent if and only if, for all i and j*

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

If there exists (i, j) such that $f_{ij} \neq f_{i\bullet} \times f_{\bullet j}$, we say that the two variables are not independent.
Equivalently, for all iand j, $N \times n_{ij} = n_{i\bullet} \times n_{\bullet j}$ In this case, if X and Y are independent

$$N \times n_{ij} = n_{i\bullet} \times n_{\bullet j}$$

3.2 Characteristics of marginal distributions

3.2.1 Means

In the case of a two-dimensional statistical variable X and Y , the means are given respectively by :

$$\bar{x} = \frac{1}{N} \sum_{i=1}^{k} n_{i\bullet} x_i = \sum_{i=1}^{k} f_{i\bullet} x_i$$

$$\bar{y} = \frac{1}{N} \sum_{i=1}^{l} n_{\bullet j} y_j = \sum_{i=1}^{l} f_{\bullet j} y_j$$

Remark 4. *In the continuous case, x_i and y_j represent respectively the center of the classes for X and Y , that is,*

$$x_i = \frac{L_{i+1} + L_i}{2} \text{ et } y_j = \frac{L_{j+1} + L_j}{2}.$$

3.2.2 variances and standard deviations

We now define the variance of X and the variance of Y as follows,

$$V(X) = \left(\frac{1}{N} \sum_{i=1}^{k} n_{i\bullet} x_i^2 \right) - \bar{x}^2 = \sum_{i=1}^{k} f_{i\bullet} x_i^2 - \bar{x}^2.$$

$$V(Y) = \left(\frac{1}{N} \sum_{i=1}^{l} n_{\bullet j} y_j^2 \right) - \bar{y}^2 = \sum_{i=1}^{l} f_{\bullet j} y_j^2 - \bar{y}^2.$$

The standard deviations of X and Y are given, respectively, by

$$\sigma(X) = \sqrt{V(X)}.$$

$$\sigma(Y) = \sqrt{V(Y)}.$$

Example 7. *Referring to example 5,*

$$\bar{y} = \frac{1}{N} \sum_{i=1}^{l} n_{\bullet j} y_j = \frac{1}{500} (146 \times 1 + 222 \times 2 + 132 \times 3) = 1,\, 972 \approx 2.$$

$$V(X) = \left(\frac{1}{N} \sum_{i=1}^{l} n_{\bullet j} y_j^2 \right) - \bar{y}^2 = \frac{1}{500} \left(146 \times 1^2 + 222 \times 2^2 + 132 \times 3^2\right) - 1,\, 972^2 = 0.555$$

3.3 Conditional distributions

The distribution of the variable Y knowing $X = x_i$ is called the conditional distribution of Y for $X = x_i$:

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

: endtabular

In this case we calculate the conditional relative frequency $f_{j/i}$ (f_j knowing that i), for $j = 1, ..., l$, by :

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

The conditional mean $\bar{y}_i$ is given by :

$$\bar{y}_i = \frac{1}{n_{ibullet}} \sum_{j=1}^{l} n_{ij} y_j = \sum_{j=1}^{l} f_{ij} y_j$$

The conditional variance $V(Y_i)$ is given by :

$$V(Y_i) = \frac{1}{n_{i\bullet}} \sum_{j=1}^{l} n_{ij} y_j^2 - \bar{y}_i^2 = \sum_{j=1}^{l} f_{ij} y_j^2 - \bar{y}_i^2$$

The conditional standard deviation $\sigma(Y_i)$

$$\sigma(Y_i) = \sqrt{V(Y_i)}.$$

The distribution of the variable X knowing that $Y = y_j$, is called ‘ee conditional distribution of X for $Y = y_j$:

$X/Y = y_j$	x_1	...	x_i	...	x_k	Total
frequency	n_{1j}	...	n_{ij}	...	n_{kj}	$n_{\bullet j}$

In this case, we calculate the conditional sequence f_i (f_i knowing that j), for $i = 1, ..., k$, by :

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

The conditional mean $\bar{x}_j$ is given by :

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

The conditional variance $V(X_j)$ is given by :

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

The conditional standard deviation $\sigma(X_i)$

$$\sigma(X_i) = \sqrt{V(X_j)}.$$

Example 8. *In example 5, we have :*
The conditional distribution of Y *for* $X = x_1 \in [0; 20000[$

Y	1	2	3	*total*
$frequency$	80	54	13	147

The conditional mean $\bar{y}_1$ *is given by :*

$$\bar{y}_1 = \frac{1}{147}(80 \times 1 + 54 \times 2 + 13 \times 3) \approx 1,54$$

The conditional variance $V(Y_1)$ *is given by :*

$$V(Y_1) = \frac{1}{147}\left(80 \times 1^2 + 54 \times 2^2 + 13 \times 3^2\right) - 1.54^2 \approx 0.44$$

The conditional standard deviation $\sigma(Y_1)$

$$\sigma(Y_1) = \sqrt{0.44} \approx 0.66.$$

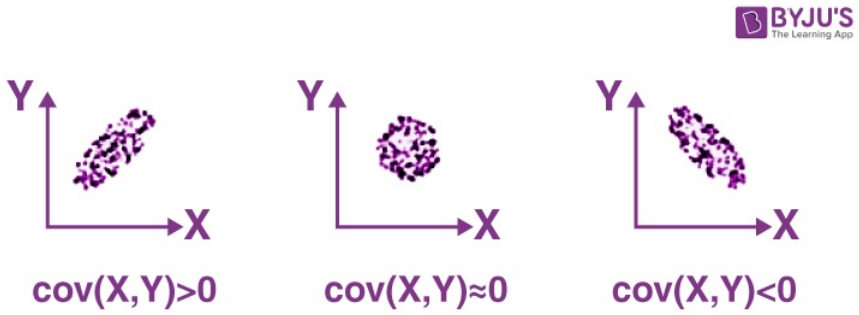
3.4 Covariance

Definition 6.

Covariance is a measure of relationship between two variables.
Let X and Y be two variables then covariance between them is given by

$$Cov(X, Y) = \sigma_{XY} = \frac{1}{N}\Sigma_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})$$

3.4.1 Interprétation



Covariance indicates whether the variables x and y vary in the same direction or in two opposite directions.
We have the following cases :
- If $Cov(x, y) > 0$ then the two variables are not independent, they are positively related and vary in the same direction.

- If $Cov(x, y) < 0$ then the two variables are not independent, they are negatively related and vary in two opposite directions.
- If $Cov(x, y) = 0$ then variations in one of the two variables do not cause variations in the other.

3.4.2 Practical calculation formula

$$cov\,(X, Y) = \sigma_{_{XY}} = \frac{1}{N}\sum_{i=1}^{N} x_i y_i - \bar{x}\bar{y}$$

Démonstration.

$$\begin{aligned}
Cov\,(X, Y) &= \frac{1}{N}\Sigma_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y}) \\
&= \frac{1}{N}\sum_{i=1}^{N}(x_i y_i - \bar{x}y_i - x_i\bar{y} + \bar{x}\bar{y}) \\
&= \frac{1}{N}\sum_{i=1}^{N} x_i y_i - \frac{1}{N}\sum_{i=1}^{N}\bar{x}y_i - \frac{1}{N}\sum_{i=1}^{n} x_i\bar{y} + \frac{1}{N}\sum_{i=1}^{N}\bar{x}\bar{y} \\
&= \frac{1}{N}\sum_{i=1}^{n} x_i y_i - \bar{x}\frac{1}{N}\sum_{i=1}^{n} y_i - \bar{y}\frac{1}{N}\sum_{i=1}^{N} x_i + \frac{1}{N}\sum_{i=1}^{N}\bar{x}\bar{y} \\
&= \frac{1}{N}\sum_{i=1}^{N} x_i y_i - \bar{x}\bar{y} - \bar{y}\bar{x} + \bar{x}\bar{y} \\
&= \frac{1}{N}\sum_{i=1}^{N} x_i y_i - \bar{x}\bar{y}
\end{aligned}$$

□

3.4.3 Properties of the covariance

If X and Y be two continuous variables and a, b, c and d be any constants, then :

1. $Cov(X, X) = Var(X)$.
2. $Cov(X, Y) = Cov(Y, X)$.
3. $Cov(\alpha X, Y) = \alpha Cov(X, Y)$, when α is constant.
4. $Cov(X + \alpha, Y) = Cov(X, Y)$,when α is constant.
5. If x and y are independent $\Rightarrow Cov(X, Y) = 0$..

Démonstration. 1.

$$\begin{aligned} Cov\,(X,X) &= \frac{1}{n}\Sigma_{i=1}^{N}(x_i-\bar{x})(x_i-\bar{x}) \\ &= \frac{1}{N}\Sigma_{i=1}^{N}(x_i-\bar{x})^2 \\ &= Var(X) \end{aligned}$$

2.

$$\begin{aligned} Cov\,(X,Y) &= \frac{1}{N}\Sigma_{i=1}^{N}(x_i-\bar{x})(y_i-\bar{y}) \\ &= \frac{1}{N}\sum_{i=1}^{N}(y_i-\bar{y})(x_i-\bar{x}) \\ &= Cov(Y,X) \end{aligned}$$

3.

$$\begin{aligned} Cov(\alpha X,Y) &= \frac{1}{N}\sum_{i=1}^{N}(\alpha x_i)y_i-(\alpha\bar{x}\bar{y}) \\ &= \alpha\frac{1}{n}\sum_{i=1}^{N}x_iy_i-\bar{x}\bar{y} \\ &= \alpha Cov(X,Y) \end{aligned}$$

4.

$$\begin{aligned} Cov(X+\alpha,Y) &= \frac{1}{N}\sum_{i=1}^{N}(x_i+\alpha)y_i-(\bar{x}+\alpha)\bar{y} \\ &= \frac{1}{N}\sum_{i=1}^{N}(x_iy_i+\alpha y_i)-(\bar{x}\bar{y}+\alpha\bar{y}) \\ &= \frac{1}{N}\sum_{i=1}^{N}x_iy_i+\alpha\frac{1}{n}\sum_{i=1}^{N}y_i-\bar{x}\bar{y}-\alpha\bar{y} \\ &= \frac{1}{N}\sum_{i=1}^{N}x_iy_i+\alpha\bar{y}-\bar{x}\bar{y}-\alpha\bar{y} \\ &= \frac{1}{N}\sum_{i=1}^{N}x_iy_i-\bar{x}\bar{y} \\ &= Cov(X,Y) \end{aligned}$$

□

4 Linear correlation coefficient-Regression line and Mayer line

Correlation is the study of links between variables in a series of observations. This study is important because it allows us to measure the degree of dependence or otherwise between variables.

4.1 Linear correlation coefficient

Definition 7. *The linear correlation coefficient (or Pearson's correlation coefficient) between X and Y is*

$$\begin{aligned} r_{xy} &= \frac{Cov(X,Y)}{\sqrt{V(x)}\sqrt{V(y)}} \\ &= \frac{Cov(X,Y)}{\sigma_x \sigma_y} \end{aligned}$$

4.1.1 Properties of the Linear correlation coefficient

Following is a gallery of scatterplots with the corresponding value of r for each.

1) The correlation coefficient of two statistical variables is always between -1 and 1 :

$$-1 \leq r_{xy} \leq 1$$

2) if $r_{xy} = 1$, indicates a perfect positive relationship between x and y.
3)if $0.75 \leq r_{xy} < 1$ indicates Strong positive linear association between the variables.
4) if $0.5 \leq r_{xy} < 0.75$ indicates moderate positive linear association between the variables.
5) if $0.25 \leq r_{xy} < 0.5$ indicates Weak positive linear association between the variables.
7) if $-0.25 < r_{xy} < 0.25$ indicates No linear association association between the variables.
8) if $-0.5 \leq r_{xy} \leq -0.25$ indicates Weak negative linear association between the variables.
9) if $-0.75 \leq r_{xy} \leq -0.5$ indicates Moderate negative linear association between the variables.
10) if $-1 \leq r_{xy} \leq -0.75$ indicates Strong negative linear association between the variables.
11) if $r_{xy} = -1$ indicates a perfect negative relationship between x and y.
12) For two independent variable correlation coefficient is zero.
13) It is always unit free.

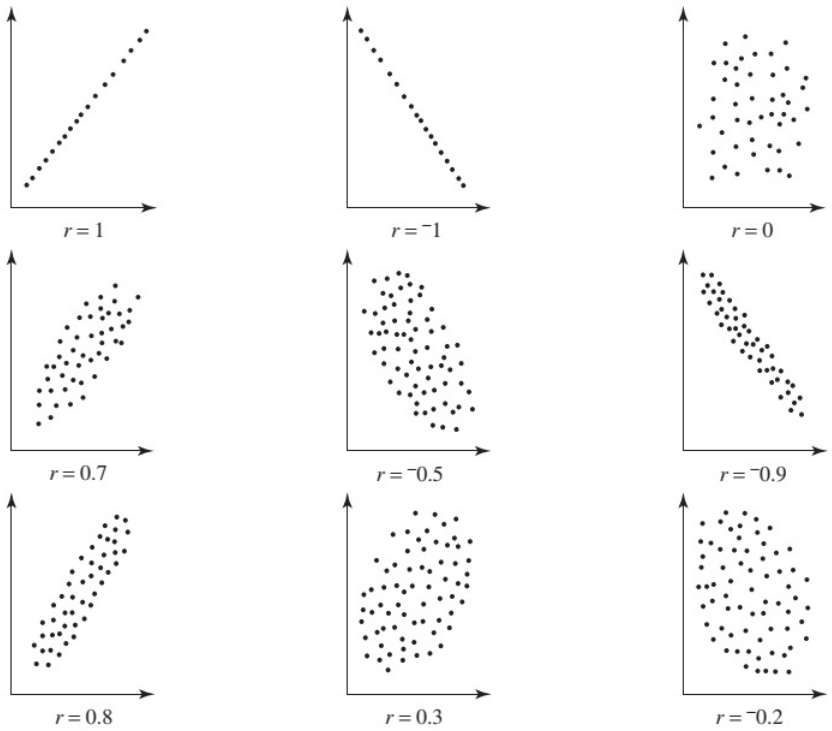


FIGURE 2 – The Scatter diagrams for different values of r

4.1.2 Regression lines

The method of least squares Equation of regression lines Using the least squares method, we assume that : the regression line is given by :

$$(D)_{y/x} : y = ax + b$$

where $a = \dfrac{cov(x,y)}{V(x)}$ and $b = \bar{y} - a\bar{x}$

Or

$$D(X/Y) : x = a'y + b'$$

where $a' = \dfrac{cov(x,y)}{V(y)}$ and $b' = \bar{x} - a'\bar{y}$

Example 9. *Referring to example 2. Fit a line to the scatterplot using the method of least squares and write its equation* $(D)_{y/x} : y = ax + b$.

Time (hours), x_i	4	36	23	19	1	11	18	13	18	8
Marks y_i	41	87	67	62	23	52	61	43	64	52

Solution

$$\bar{x} = \frac{1}{10}\sum_{i=1}^{10} x_i = \frac{4+36+23+19+1+11+18+13+18+8}{10} = 15.1$$

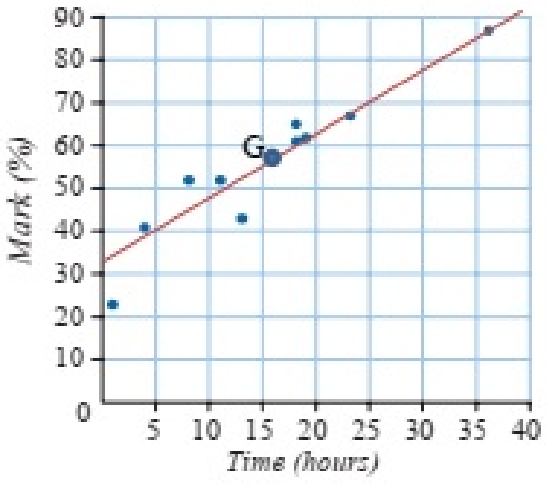
$$\bar{y} = \frac{1}{10}\sum_{i=1}^{10} y_i = \frac{23+41+52+52+43+61+65+62+67+87}{10} = 55.3$$

$$V(x) = \frac{1}{10}\sum_{i=1}^{10} x_i^2 - \bar{x}^2 = 92.49$$

$$cov\,(X,Y) = \frac{1}{N}\sum_{i=1}^{10} x_i y_i - \bar{x}\bar{y} = 148.47$$

$$then\ a = \frac{148.47}{92.49} \simeq 1.6\ \ and\ b = \bar{y} - a\bar{x} = 55.3 - 1.6 \times 15.1 = 31.1$$

$$So\ ,\ (D) : y = 31.1 + 1.6x$$



4.1.3 Fitting a line using the two-mean method (Mayer line)

1) Rewrite the data pairs in order, according to the x values.

2) Divide the ordered table into two new tables : one for the lower half of data values, the other for the top half of data values.

3) Find the mean values of x and y for each new table $G_1(\bar{x_1}, \bar{y_1})$ and $G_2(\bar{x_2}, \bar{y_2})$.

4) The line (D) passes through points G_1 and G_2 and therefore has the equation $y = ax + b$ where

$$a = \frac{\bar{y}_2 - \bar{y}_2}{\bar{x}_2 - \bar{x}_2} \text{ and } b = \bar{y_2} - a\bar{x_2}$$

Example 10. *Referring to example 2. Fit a line to the scatterplot using the two-mean method and write its equation .*

Time (hours), x_i	4	36	23	19	1	11	18	13	18	8
Marks y_i	41	87	67	62	23	52	61	43	64	52

solution

1) *Rewrite the data pairs in order, according to the x values.*

Time (hours), x_i	1	4	8	11	13	18	18	19	23	36
Marks y_i	23	41	52	52	43	61	65	62	67	87

2) *Divide the ordered table into two new tables : one for the lower half of data values, the other for the top half of data values.*

x_i	1	4	8	11	*13*
y_i	23	41	52	52	43

x_i	18	18	19	23	36
y_i	61	65	62	67	87

3) *Find the mean values* $G_1(\bar{x_1}, \bar{y_1})$ *and* $G_2(\bar{x_2}, \bar{y_2})$

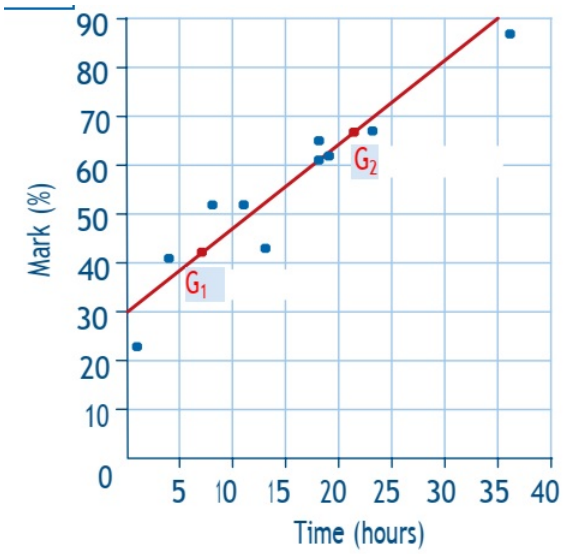
$$\begin{cases} \bar{x_1} = \frac{1}{5}\sum_{i=1}^{5} x_i x_i = \frac{1}{5}\sum_{i=1}^{5} x_i = \frac{1+4+8+11+13}{5} = 7.4 \\ \bar{y_1} = \frac{1}{5}\sum_{i=1}^{5} x_i y_i = \frac{23+41+52+52+43}{5} = 42.2 \end{cases} \text{ and } \begin{cases} \bar{x_2} = \frac{1}{5}\sum_{i=6}^{10} x_i x_i = 22.8 \\ \bar{y_2} = \frac{1}{5}\sum_{i=6}^{10} x_i y_i = 68.4 \end{cases}$$

4) *The line (D) passes through points* G_1 *and* G_2 *and therefore has the equation* $y = ax + b$ *where*

$$a = \frac{\bar{y}_2 - \bar{y}_2}{\bar{x}_2 - \bar{x}_2} = \frac{68.4 - 42.2}{22.8 - 7.4} = \frac{26.2}{15.4} = 1.7 \text{ and}$$
$$b = \bar{y_2} - a\bar{x_2} = 68.4 - 1.7 \times 22.8 = 29.6$$

Equation of the two-mean line :

$$(D) : y = 29.6 + 1.7x$$



5 Non-linear fitting

In some cases, the fit to a linear function is not adequate : a fit of the data to a non-linear function must be considered. The cases we will consider are those where a simple transformation can be made to an affine fit.

5.1 Fitting using a hyperbola

The points $(x_i; y_i)$ are not aligned, but rather close to a certain hyperbola of of the form

$$y = \frac{1}{ax + b}$$

5.1.1 Method

1 calculate $z_i = \frac{1}{y_i}$;

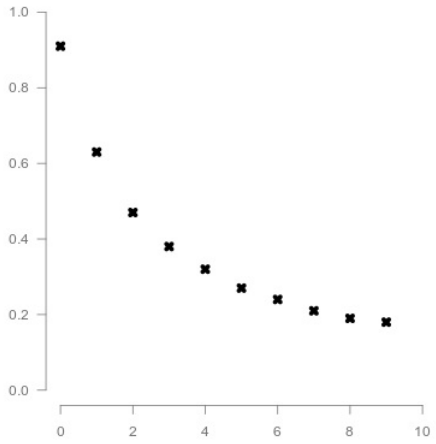
2 determine the equation of the regression line from z to x using the method of least squares ;

3 from the equation obtained $z = ax + b$, immediately deduce the equation of the hyperbola

$$y = \frac{1}{ax + b}$$

Example 11. *Fit this point cloud with a hyperbola* $y = \dfrac{1}{ax + b}$

x_i	0	1	2	3	4	5	6	7	8	9
y_i	0.91	0.63	0.47	0.38	0.32	0.27	0.24	0.21	0.19	0.18



solution

1) *calculate* $z_i = \dfrac{1}{y_i}$

x_i	0	1	2	3	4	5	6	7	8	9
z_i	1.1	1.6	2.1	2.6	3.1	3.6	4.1	4.6	5.1	5.6

2) *determine the equation of the regression line from z to x using the method of least squares*

$\bar{x} = \dfrac{1}{10}\sum_{i=1}^{10} x_i = 4.5$

$\bar{z} = \dfrac{1}{10}\sum_{i=1}^{10} z_i = 3.35$

$V(x) = \dfrac{1}{10}\sum_{i=1}^{10} x_i^2 - \bar{x}^2 = 8,25$

$cov\,(X, Z) = \dfrac{1}{N}\sum_{i=1}^{10} x_i z_i - \bar{x}\bar{z} = 4.125$

then $a = \dfrac{4.125}{8,25} = 0.5$ *and* $b = \bar{z} - a\bar{x} = 1.1$
So , $(D) : z = 1.1 + 0.5x$

3) *So the fitting is*

$$y = \frac{1}{1.1 + 0.5x}$$

5.2 Fitting the power function

The points $(x_i; y_i)$ are not aligned, a power function curve such as

$$y = bx^a$$

Note that $ln(y) = aln(x) + ln(b)$.

5.2.1 Method

1 calculate $u_i =$ and $v_i = lny_i$;

2 determine the equation of the regression line from v to u using the least squares method least squares ;

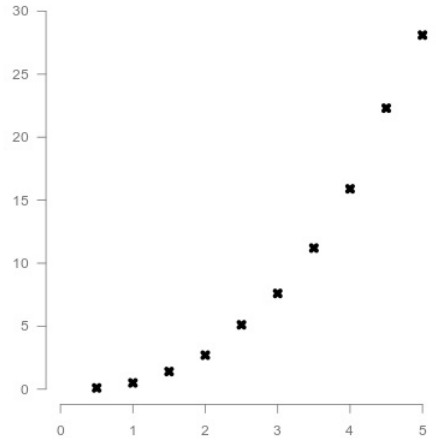
3 From the equation $v = Au + B$, we can deduce the equation of the power function

$$y = bx^a$$

since $a = A$ and $b = e^B$

Example 12. *Fit this scatter plots using a power function* $y = bx^a$

x_i	0.5	1	1.5	2	2.5	3	3.5	4	4.5	5
y_i	0.1	0.5	1.4	2.7	5.1	7.6	11.2	15.9	22.3	28.1



solution

1) calculate calculate $u_i = lnx$ and $v_i = lny_i$;

u_i	-0.69	0	0.41	0.69	0.91	1.1	1.25	1.39	1.5	1.61
v_i	-2.3	-0.69	0.34	1	1.63	2.03	2.42	2.77	3.1	3.34

2) determine the equation of the regression line from v to u using the method of least squares

$\bar{u} = \frac{1}{10}\sum_{i=1}^{10} u_i = 0,817$

$\bar{v} = \frac{1}{10}\sum_{i=1}^{10} v_i = 1.36$

$V(u) = \frac{1}{10}\sum_{i=1}^{10} x_i^2 - \bar{x}^2 = 0,482$

$cov\,(u,v) = \frac{1}{10}\sum_{i=1}^{10} u_i v_i - \bar{u}\bar{v} = 1.19$

then $A = \frac{1.19}{0.482} = 2.47$ and $B = \bar{v} - A\bar{u} = 1.36 - 2.47 \times 0.817 = -0.66$

So , $(D) : v = -0.66 + 2.47u$ Since $lnb = -0.66 sob = e^{-0.66} = 0.52$

3) So the fitting is

$$y = 0.52x^{2.47}$$

5.3 Fitting by an exponential function

The points $(x_i; y_i)$ are near a curve of an exponential function of the form

$$y = be^{ax}$$

Note that $ln(y) = ax + ln(b)$.

5.3.1 Method

1 calculate $z_i = lny_i$;

2 determine the equation of the regression line from z to x using the least squares method least squares ;

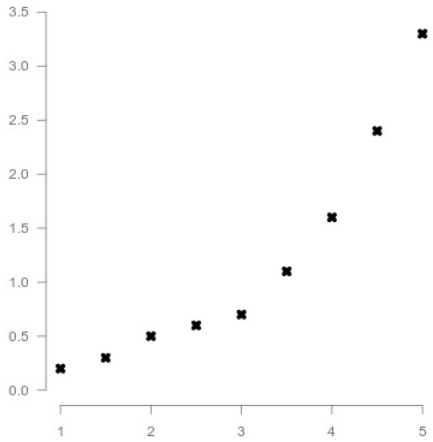
3 From the equation $z = ax + B$, we can deduce the equation of the power function

$$y = be^{ax}$$

since $b = e^B$

Example 13. *Fit this scatter plots using a power function* $y = be^{ax}$

x_i	1	1.5	2	2.5	3	3.5	4	4.5	5
y_i	0.2	0.3	0.5	0.6	0.7	1.1	1.6	2.4	3.3



solution

1) calculate calculate $z_i = lny_i$;

x_i	1	1.5	2	2.5	3	3.5	4	4.5	5
z_i	-1.61	-1.2	-0.69	-0.51	-0.36	0.1	0.47	0.88	1.19

2) determine the equation of the regression line from z to x using the method of least squares

$\bar{x} = \frac{1}{9}\sum_{i=1}^{9} x_i = 3$

$\bar{z} = \frac{1}{9}\sum_{i=1}^{9} z_i = -0.19$

$v(x) = \frac{1}{9}\sum_{i=1}^{9} x_i^2 - \bar{x}^2 = 1.67$

$cov\,(x,z) = \frac{1}{9}\sum_{i=1}^{9} x_i z_i - \bar{x}\bar{z} = 1.125$

then $a = \dfrac{1.125}{1.67} = 0,67$ and $B = \bar{z} - A\bar{x} = -0.19 - 0.67 \times 3 = -2.2$

So , $(D) : z = -2.2 + 0.67x$ Since $lnb = -2.2$, so $:b = e^{-2.2} \simeq 0.11$

3) So the fitting is

$$y = 0.11e^{0.67x}$$

5.4 Fitting using a logarithmic function

The points $(x_i; y_i)$ are near a logarithmic curve of the form

$$y = b + alnx$$

5.4.1 Method

1 calculate $z_i = lnx_i$;

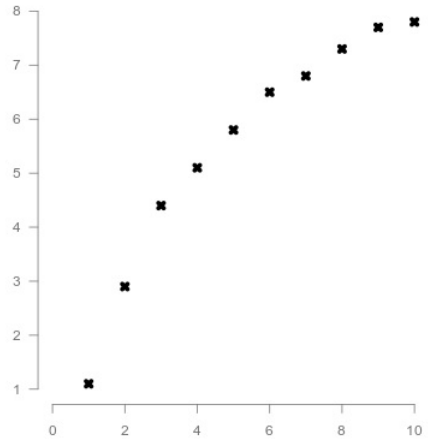
2 determine the equation of the regression line from y to z using the least squares method least squares ;

3 From the equation $y = az + b$, we can deduce the equation of the power function

$$y = b + alnx$$

Example 14. *Fit this scatter plots using a power function* $y = b + alnx$

x_i	1	2	3	4	5	6	7	8	9	10
y_i	1.1	2.9	4.4	5.1	5.8	6.5	6.8	7.3	7.7	7.8



solution

1) calculate calculate $z_i = lnx_i$;

z_i	0	0.69	1.1	1.87	1.61	1.79	1.95	2.08	2.2	2.3
y_i	1.1	2.9	4.4	5.1	5.8	6.5	6.8	7.3	7.7	7.8

2) determine the equation of the regression line from y to z using the method of least squares

$$\bar{z} = \frac{1}{10}\sum_{i=1}^{10} z_i = 1.51$$

$$\bar{y} = \frac{1}{10}\sum_{i=1}^{10} y_i = 5.54$$

$$v(z) = \frac{1}{10}\sum_{i=1}^{10} z_i^2 - \bar{z}^2 = 0.485$$

$$cov\,(y,z) = \frac{1}{10}\sum_{i=1}^{10} y_i z_i - \bar{y}\bar{z} = 1.453$$

then $A = \dfrac{1.453}{0.485} = 3$ and $B = \bar{z} - A\bar{x} = 5.54 - 3 \times 1.51 = 1.01$

So , $(D) : y = 1.01 + 3z$ Since

3) So the fitting is

$$(D) : y = 1.01 + 3lnx$$