Solutions to the Directed Works Series n°2

Exercice n°1 : We are interested in a group of 40 employees of a certain company. The following data is presented in the form of pairs of values of the form (x_i, y_i) where x_i is the Gender of the person {Male, Female} and y_i is the last diploma {bac, licence, master} :
(M,L), (F,B), (M,L), (M,L), (F,B), (F,L), (M,B), (F,L), (M,B), (F,B),(M,L), (M,B), (F,B), (F,L) , (M,B), (M,B), (M,Ma), (F,B), (M,B), (M,B), (F,B) (M,Ma), (F,B), (M,Ma), (M,B),, (M,B), (M,L), (M,L), (M,B), (F,L), (F,B) , (M,B), (M,B), (M,B), (F,B) , (F,Ma), (M,L), (M,B), (F,B) (M,B).

1. Identify the population and characteristics being studied and their nature.
2. Fill in the contingency table :

Gender \ diploma	bac	licence	Master	total
Male	...	...	...	...
Female	...	...	...	...
total	...	...	...	...

3. Are the two variables X and Y independent ? Justify.

Solution :

1. Population : The group of 40 employees in the company.
   Variable X (Gender) : Qualitative, with two modalities Male, Female.
   Variable Y (Diploma) : Qualitative, with three modalities bac, licence, master.
2. Contingency table

Gender \ diploma	bac	licence	Master	$n_{i\bullet}$
Male	$n_{11} = 15$	$n_{12} = 7$	$n_{13} = 3$	$n_{1\bullet} = 25$
Female	$n_{21} = 10$	$n_{22} = 4$	$n_{23} = 1$	$n_{2\bullet} = 15$
$n_{\bullet j}$	$n_{\bullet 1} = 25$	$n_{\bullet 2} = 11$	$n_{\bullet 3} = 4$	$N = 40$

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

$$f_{ij} = f_{i\bullet} \times f_{\bullet j} \text{ Equivalently, for all i and j, } N \times n_{ij} = n_{i\bullet} \times n_{\bullet j}$$

But, $40 \times 15 \neq 25 \times 25$, so the variables Gender and Diploma are not independent.

Exercice n°2 : In an exam, each candidate is tested in statistics (mark X) and maths (mark Y). The results for a sample of 100 candidates are as follows :

$X\backslash Y$	$[0,4]$	$]4,8]$	$]8,12]$	$]12,16]$	$]16,20]$	total
$[0,4]$	3	4	2	0	0	...
$]4,8]$	6	9	7	4	0	...
$]8,12]$	1	8	15	12	8	...
$]12,16]$	0	1	7	7	3	...
$]16,20]$	0	0	1	0	2	...
total	...	...	...	...	...	...

1. Identify the population, its size and the type of variables being studied.
2. Determine the marginal distributions of X and Y .
3. Calculate the marginal means and variances of X and Y ;
4. Determine the conditional distribution of Y knowing that X is in the interval $]8;12]$.
5. Calculate the mean of the conditional distribution of Y given that X is in the interval $]8;12]$.

Solution :

1. Population : Candidates who took the exam.
   Size : 100 candidates.
   Variables :
   X (Statistics mark) : Quantitative continuous, grouped into intervals.
   Y (Maths mark) : Quantitative continuous, grouped into the same intervals.
2. Contingency table with totals :

$X\backslash Y$	$[0,4]$	$]4,8]$	$]8,12]$	$]12,16]$	$]16,20]$	total
$[0,4]$	3	4	2	0	0	9
$]4,8]$	6	9	7	4	0	26
$]8,12]$	1	8	15	12	8	44
$]12,16]$	0	1	7	7	3	18
$]16,20]$	0	0	1	0	2	3
total	10	22	32	23	13	100

   Marginal distributions of X :

X	$[0,4]$	$]4,8]$	$]8,12]$	$]12,16]$	$]16,20]$	total
$n_{i\bullet}$	9	26	44	18	3	100
c_i	2	6	10	14	18	/

   Marginal distributions of Y :

Y	$[0,4]$	$]4,8]$	$]8,12]$	$]12,16]$	$]16,20]$	total
$n_{\bullet j}$	10	22	32	23	13	100
c_j	2	6	10	14	18	/

3. Marginal means of X
   Mean of X : $\bar{x} = \sum_{i=1} \frac{n_{i\bullet} c_i}{N} = 9.2$
   variance of x :
   $V(x) = \sum_{i=1} \frac{n_{i\bullet} c_i^2}{N} - \bar{x}^2 = 14.08$
   Mean of y :

$\bar{y} = \sum \frac{n_{\bullet j} c_j}{N} = 10.28$ variance of y :

$V(y) = \sum_{j=1} \frac{n_{\bullet j} c_j^2}{N} - \bar{y}^2 \simeq 21.84$

4. Conditional distribution of Y given X in]8,12] :

Y_3	$[0,4]$	$]4,8]$	$]8,12]$	$]12,16]$	$]16,20]$	total
n_{3j}	1	8	15	12	8	44
c_j	2	6	10	14	18	/

5. Conditional mean of Y given X in]8,12] :

$\bar{y_3} = \sum_{j=1} \frac{n_{3j} c_j}{n_{3\bullet}} \simeq 11.64$

variance of y_3 : $V(y_3) = \sum_{j=1} \frac{n_{3j} c_j^2}{n_{3\bullet}} - \bar{y_3}^2 \simeq 21.84$

Exercice n°3 :

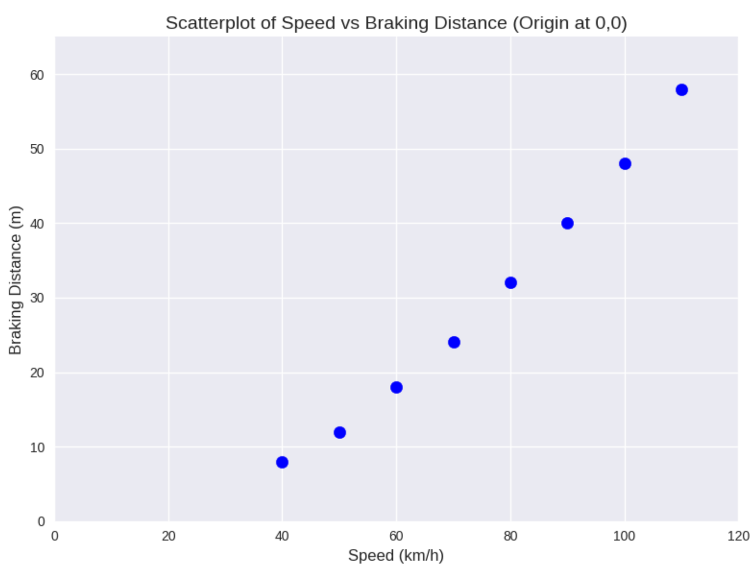
The following table gives the braking distance of a car on a dry road as a function of its speed :

x_i (Speed in kilometres/hour)	40	50	60	70	80	90	100	110
y_i (distance in metres)	8	12	18	24	32	40	48	58

1. Construct a scatterplot of this data $M_i(x_i; y_i)$.
2. Do you think the Fitting a line is justified ? Justify your answer.
3. Using Mayer's method (two-mean method), determine the equation of the straight line representing the braking distance as a function of speed .
4. Verify that the mean point is on the fitting line.
5. Using this equation, estimate the braking distance of a vehicle travelling at 120km/h.
6. Repeat the calculations using the method of least squares.

Solution :

1. Scatterplot



2. The scatter plot is elongated, so refined adjustment is justified.
3. Line by Mayer's two-mean method :

i) 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	40	50	60	70
y_i	8	12	18	24

x_i	80	90	100	110
y_i	32	40	48	58

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

$$\begin{cases} \bar{x_1} = \frac{1}{4}\sum_{i=1}^{4} x_i = \dfrac{1}{4}\sum_{i=1}^{4} x_i = \dfrac{40+50+60+70}{4} = 55 \\ \bar{y_1} = \frac{1}{4}\sum_{i=1}^{4} y_i = \dfrac{8+12+18+24}{4} = 15.5 \end{cases} \text{ and } \begin{cases} \bar{x_2} = \frac{1}{4}\sum_{i=5}^{8} x_i = 95 \\ \bar{y_2} = \frac{1}{4}\sum_{i=5}^{8} y_i = 44.5 \end{cases}$$

iii) Equation through the two mean points G_1 and G_2, therefore has the equation $y = ax + b$ where

$$a = \frac{\bar{y_2} - \bar{y_2}}{\bar{x_2} - \bar{x_2}} = \frac{44.5 - 15.5}{95 - 55} = \frac{29}{40} = 0.725 \text{ and}$$
$$b = \bar{y_2} - a\bar{x_2} = 44.5 - 0.725 \times 95 = -24.375$$

Equation of the two-mean line :

$$(D) : y = 0.725x - 24.375$$

4. Verify the global mean point $G(\bar{x}, \bar{y})$ lies on the line Global means :
$\bar{x} = \Sigma_{i=1}^{8} \dfrac{x_i}{8} = 75; \bar{y} = \Sigma_{i=1}^{8} \dfrac{y_i}{8} = 30$

Check : $0.725 \times 7524.375 = 54.37524.375 = 30 = \bar{y}$, So $G(\bar{x}, \bar{y})$ is on the Mayer line.
5. Estimate at $120km/h$, $y = 0.725 \times 120 - 24.375 = 62.625m$.
6. Least squares method : $\bar{x} = \dfrac{1}{8}\sum_{i=1}^{8} x_i = 75$

$\bar{y} = \dfrac{1}{8}\sum_{i=1}^{8} y_i = 30$

$V(x) = \dfrac{1}{8}\sum_{i=1}^{8} x_i^2 - \bar{x}^2 = 525$

$cov\,(X, Y) = \dfrac{1}{N}\sum_{i=1}^{8} x_i y_i - \bar{x}\bar{y} = 377.5$

then $a = \dfrac{377.5}{525} \simeq 0.72$ and $b = \bar{y} - a\bar{x} = 30 - 0.72 \times 75 = -24$

So , $(D) : y = 0.72x - 24.$

$y = 0.72 \times 120 - 24 = 62.4m.$

Both methods give close results ; the least-squares line is the statistically optimal linear fit for this data, while Mayer's is a quick approximation.

***Exercice n°**4 : Fit this scatter plots using 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 :

$y = \dfrac{1}{ax+b} \Leftrightarrow ax + b = \dfrac{1}{y}$. **We put** $z = \dfrac{1}{y}$

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**

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

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

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

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

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**

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

Solution :

***Exercice n°5 :**

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 :

$y = bx^a \Leftrightarrow lny = lnb + alnx$. We put $v = lny$ and $u = lnx$. We are looking for the equation of a line $v = au + B$.

This is a linear regression problem

1) 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

$\bar{u} = \dfrac{1}{10}\sum\limits_{i=1}^{10} u_i = 0,817$

$\bar{v} = \dfrac{1}{10}\sum\limits_{i=1}^{10} v_i = 1.36$

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

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

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

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

3) So the fitting is

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

***Exercice n°6 :**

Fit this scatter plots using an exponential 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 :

$y = be^{ax} \Leftrightarrow lny = lnb + ax.$

We put $z = lny$.

We are looking for the equation of a line $z = ax + B$.

This is a linear regression problem

1) 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 $a = \frac{1.125}{1.67} = 0,67$ 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

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

Exercice n°7 :

Fit this scatter plots using a power function $y = b + a\ln x$

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 :

$y = b + a\ln x.$

We put $z = lnx$.

We are looking for the equation of a line $y = az + b$.

This is a linear regression problem

1) calculate $z_i = lnx_i$;

z_i	0	0.69	1.1	1.37	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 $A = \frac{1.453}{0.485} = 3$ 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

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

Module manager. Merini.A