Bivariate Analysis

N. Bensmain

March 8, 2025

In descriptive statistics we are interested in studying a character or a variable for a given population. We may be led to study two characteristics simultaneously, in this case we use two-variable statistics called bivariated analysis.
Bivariate analysis is a statistical method that helps you study relationships (correlation) between data sets. Many businesses, marketing, and social science questions and problems could be solved using bivariate data sets.

let X and Y be two variables defined on Ω, $card(\Omega) = n$.
$Z = (X, Y)$ is called the two dimensional variable. Values of Z are in the followig table;

ω_i	ω_1	ω_2	...	ω_n
$X(\omega_i)$	$X(\omega_1)$	$X(\omega_2)$ ...		$X(\omega_n)$
$Y(\omega_i)$	$Y(\omega_1)$	$Y(\omega_2)$	...	$Y(\omega_n)$

Let $x_1,\ x_2, ..., x_n$ values taken by X and $y_1,\ y_2, ..., y_n$ those taken by Y.

Exemple 1:

In the following table we give for each town, the average number of hours of sunshine per year as well as the average temperature.

town	T1	T2	T3	T4	T5	T6
hours of sunshine	2790	2072	2767	1729	1574	1833
Temperature	14.7	11.4	14.2	10.8	9.7	11.2

Sample (Ω): 6 towns.

1st variable : hours of sunshine per year.

2nd variable : the average temperature. **Exemple 2:**

The table below allows you to follow the evolution of the life expectancy of women in France from 1990 to 1999.

Year	1990	1991	1992	1993	1994
Life expectancy	80.9	81.1	81.4	81.8	81.9
Year	1995	1996	1997	1998	1999
Life expectancy	82	82.3	82.4	82.4	80

Sample: women in France.

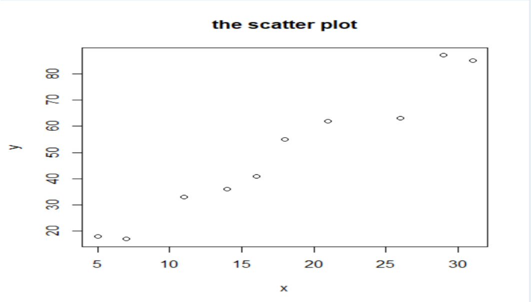
1st variable : year.

2nd variable : life expectancy.

Remark

In a statistical series with two variables if one of the variables is time (year for example), it is called a chronological series.

One way to adress the problem is to look at pairs of variables; $z_i = (x_i, y_i)$.
A scatter plot displays the bivariate data in a graphical form that maintains the pairing.



A second way is the contingency table which is a type of table in a matrix format that displays the bivariate frequency or relative

*Frequency contingency table:

n_{ij} is the frequency of (x_i, y_j).
$n_{ij} =$ $card\,\{\omega \in \Omega,$ such that $(X(\omega), Y(\omega)) = (x_i, y_j)\}$
where $1 \leq i \leq l, 1 \leq j \leq k$

$x_i \| y_j$	y_1	y_2	...	y_k	Marginal of X
x_1	n_{11}	n_{12}	...	n_{1k}	$n_{1.}$
x_2	n_{21}	n_{22}	...	n_{2k}	$n_{2.}$
...	...	...	...	...	...
x_l	n_{l1}	n_{l2}	...	n_{lk}	$n_{l.}$
Marginal of Y	$n_{.1}$	$n_{.2}$	...	$n_{.k}$	N

The marginal frequency of X is $n_{i.} = \sum_{j=1}^{k} n_{ij}$.
The marginal frequency of Y is $n_{.j} = \sum_{i=1}^{l} n_{ij}$.
$\sum_{j=1}^{k} \sum_{i=1}^{l} n_{ij} = \sum_{i=1}^{l} \sum_{j=1}^{k} n_{ij} = N$.
$card(\Omega) = N$.

*Relative frequency contingency table:

The relative frequency of (x_i, y_j) is $f_{ij} = \frac{n_{ij}}{N}$, $1 \leq i \leq l$, $1 \leq j \leq k$.
The joint distribution of $Z = (X, Y)$ is summerized in the table below

$x_i \vert y_j$	y_1	y_2	...	y_k	Marginal of X
x_1	f_{11}	f_{12}	...	f_{1k}	$f_{1.}$
x_2	f_{21}	f_{22}	...	f_{2k}	$f_{2.}$
...	...	...	...	...	...
x_l	f_{l1}	f_{l2}	...	f_{lk}	$f_{l.}$
Marginal of Y	$f_{.1}$	$f_{.2}$	...	$f_{.k}$	1

The marginal distribution of x_i is $f_{i.} = \sum_{j=1}^{k} f_{ij}$.

The marginal distribution of y_j is $f_{.j} = \sum_{i=1}^{l} f_{ij}$.

***Remark:** $\sum_{i=1}^{l} \sum_{j=1}^{k} f_{ij} = \sum_{j=1}^{k} \sum_{i=1}^{l} f_{ij} = 1$.

Example:

An experiment was carried out on 234 people to study the relationship that exists between age X in years and sleep time Y in hours, the following table was obtained;

X|Y	$[5,7[$	$[7,9[$	$[9,11[$	$[11,15]$	Marge de X
$[1,3[$	0	0	2	36	38
$[3,11[$	0	3	12	26	41
$[11,19[$	2	8	35	16	61
$[19,31[$	0	26	22	3	51
$[31,59]$	22	15	6	0	43
Marge de Y	24	52	77	81	234

The relative frequency contingency table is then

X|Y	$[5,7[$	$[7,9[$	$[9,11[$	$[11,15]$	Margin of X
$[1,3[$	0	0	2/234	36/234	38/234
$[3,11[$	0	3/234	12/234	26/234	41/234
$[11,19[$	2/234	8/234	35/234	16/234	61/234
$[19,31[$	0	26/234	22/234	3/234	51/234
$[31,59]$	22/234	15/234	6/234	0	43/234
Margin of Y	24/234	52/234	77/234	81/234	1

The conditional marginal frequency of y_j given $X = x_i$ is
$f_{j|i} = \frac{n_{ij}}{n_{i.}} = \frac{f_{ij}}{f_{i.}}$.
The marginal distribution of Y given $X = x_i$ is in the table below;

Y	y_1	...	y_k
$f_{j\|i}$	$\frac{f_{i1}}{f_{i.}}$	...	$\frac{f_{ik}}{f_{i.}}$

Example:

The conditional distribution of Y given $X \in [11, 19[$

Y	$[5, 7[$	$[7, 9[$	$[9, 11[$	$[11, 15]$
$f_{j\|X\in[11,19[}$	$2/61$	$8/61$	$35/61$	$16/61$

Interpretation: $(8/61) * 100$ is the percentage of persons aged between 11 and 19 years who have a sleep time between 7 and 9 hours.

The conditional marginal frequency of x_i given $Y = y_j$ is
$f_{i|j} = \frac{n_{ij}}{n_{.j}} = \frac{f_{ij}}{f_{.j}}$.
The marginal distribution of X given $Y = y_j$ is represnted in the table below;

X	x_1	...	x_I
$f_{i\|j}$	$\frac{f_{1j}}{f_{.j}}$	...	$\frac{f_{Ij}}{f_{.j}}$

Example:

the table below gives the coditional distribution of X given $Y \in [7, 9[$

X	$[1, 3[$	$[3, 11[$	$[11, 19[$	$[19, 31[$	$]31, 59]$
$f_{i\mid Y\in[7,9[}$	0	$3/52$	$8/52$	$26/52$	$15/52$

Interpretation; Among those with sleep duration between 7 and 9 hours $(26/52) * 100$ are aged between 19 and 31 years.

Two variables X and Y are said to be independant if

$$f_{ij} = f_{i.} * f_{.j}, \quad i \in \{1, ..., l\} \quad \text{et} \quad j \in \{1, ..., k\}.$$

We can check the independance of X and Y from the conditional distribution;

$$\forall i \in \{1, ..., l\}, \; f_{i|j} = f_{i.}, \quad \text{for all} \quad j \in \{1, ..., k\}.$$

and

$$\forall j \in \{1, ..., k\}, \; f_{j|i} = f_{.j}, \quad \text{for all} \quad i \in \{1, ..., l\}$$

Example: $f_{11} = 0$, $f_{1.} = 38/234$ and $f_{.1} = 24/234$ then there is no indépendance between age and sleep duration.

The marginal sample mean of X is $\bar{x}$; $\bar{x} = \frac{\sum_i n_{i.}x_i}{N} = \sum_i f_{i.}x_i$.

The marginal sample mean of Y is $\bar{y}$; $\bar{y} = \frac{\sum_j n_{.j}y_j}{N} = \sum_j f_{.j}y_j$.

Example:

To compute marginal sample means we need class marks of the variables X and Y.

*Marginal distribution of X

X	$[1,3[$	$[3,11[$	$[11,19[$	$[19,31[$	$]31,59]$
Class mark c_i	2	7	15	25	45
$f_{i.}$	$38/234$	$41/234$	$61/234$	$51/234$	$43/234$

Then $\sum_i f_{i.} * c_i = 2 * 38/234 + 7 * 41/234 + 15 * 61/234 + 25 * 51/234 + 45 * 43/234 = 19.179$

*Marginal distribution of Y

Y	$[5,7[$	$[7,9[$	$[9,11[$	$[11,15]$
Class mark c_j	6	8	10	13
$f_{.j}$	24/234	52/234	77/234	81/234

then;

$\sum_j f_{.j} * c_j = 6*24/234 + 8*52/234 + 10*77/234 + 13*81/234 = 10.18$

Marginal variance of X is V(X); $V(X) = \sum_i f_{i.} x_i^2 - \bar{x}^2$.
$V(X) = 2^2 * 38/234 + 7^2 * 41/234 + 15^2 * 61/234 + 25^2 * 51/234 + 45^2 * 43/234 - 19.179^2 = 576.222 - 367.834 = 208.388$.
Marginal variance of X is V(Y); $V(Y) = \sum_j f_{.j} y_j^2 - \bar{y}^2$.
Example: $V(Y) = 6^2 * 24/234 + 8^2 * 52/234 + 10^2 * 77/234 + 13^2 * 81/234 - 10.18^2 = 109.3205 - 103.6324 = 5.6881$

Conditional sample mean of X given $Y = y_j$ dented $\bar{x}_j$ is

$$\bar{x}_j = \sum_{i=1}^{I} \frac{f_{ij}}{f_{.j}} x_i.$$

Example: Conditional distribution of X given $Y \in [7, 9[$ presented in the table below:

X	$[1, 3[$	$[3, 11[$	$[11, 19[$	$[19, 31[$	$]31, 59]$
$f_{i\|Y\in[7,9[}$	0	3/52	8/52	26/52	15/52

$\bar{x}_j = 2*0+7*3/52+15*8/52+25*26/52+45*15/52 = 28.192.$

Conditional sample mean of Y given $X = x_i$ denoted $\bar{y}_i$ is

$$\bar{y}_i = \sum_{i=1}^{I} \frac{f_{ij}}{f_{i.}} y_j.$$

Example:

conditional distribution of Y given $X \in [11, 19[$

Y	$[5, 7[$	$[7, 9[$	$[9, 11[$	$[11, 15]$
$f_{j\|X \in [11,19[}$	$2/61$	$8/61$	$35/61$	$16/61$

$\bar{y}_i = 12 * 2/61 + 8 * 8/61 + 10 * 35/61 + 13 * 16/61 = 10.59.$

conditional sample variance of X given $Y = y_j$ denoted $V(X|Y = y_j)$ is

$$V(X|Y = y_j) = \sum_{i=1}^{l} f_{i|j}(x_i - \bar{x}_j)^2 = \sum_{i=1}^{l} f_{i|j} x_j^2 - \bar{x}_j^2$$

.

Example: $V(X|Y = y_j) = 2^2 * 0 + 7^2 * 3/52 + 15^2 * 8/52 + 25^2 * 26/52 + 45^2 * 15/52 - 28.192^2 = 934.077 - 794.789 = 139.2881.$

conditional sample variance of Y given $X = x_i$ denoted $V(Y|X = x_i)$,

$$V(Y|X = x_i) = \sum_{j=1}^{k} f_{j|i}(y_j - \bar{y}_i)^2 = \sum_{j=1}^{k} f_{j|i} y_j^2 - \bar{y}_i^2$$

Example:

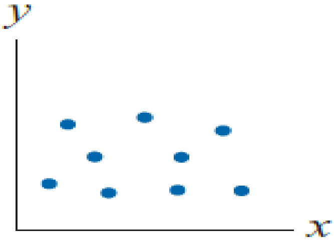
$V(Y|X = x_i) = 12^2 * 2/61 + 8^2 * 8/61 + 10^2 * 35/61 + 13^2 * 16/61 - 10.59^2 = 114.82 - 112.148 = 2.671.$

Remark:

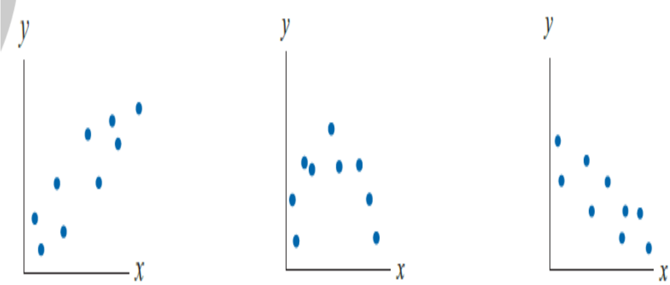
if X and Y are independant then $\bar{x}_j = \bar{x}$ and $\bar{y}_i = \bar{y}$.

When constructing a scatterplot, it is conventional to use the vertical or y-axis for the dependent variable, and the horizontal or x-axis for the independent variable.
How can The scatter plot help us to identify and describe any relationship?

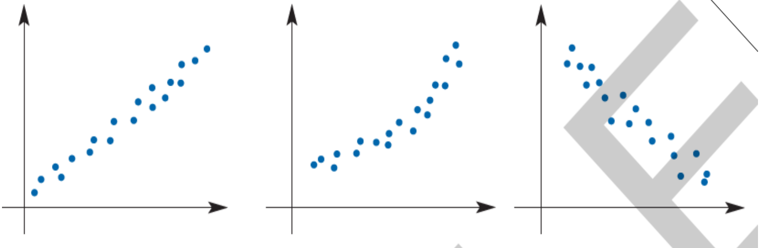
First we look to see if there is a clear pattern in the scatter plot. If the points are randomly scatered across the plot, there is no relationship.



For the three examples below, there is a clear pattern in each set of points, so we concude there is a relationship in each case.

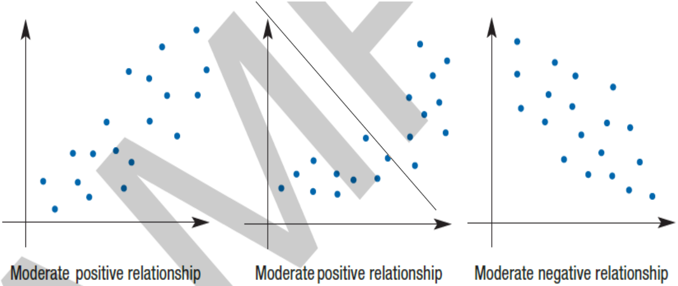


The strength of a relationonship is measured by how much scatter there is in a scatterplot.
Strong relationship:When there is a strong relationship between the variables, the points will tend to follow a single stream. A pattern is clearly seen. There is only a small amount of scatter in the plot.

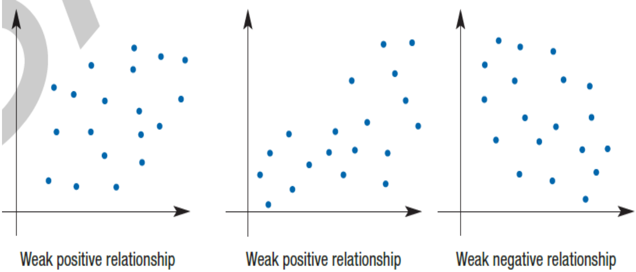


Strong positive relationship Strong positive relationship Strong negative relationship

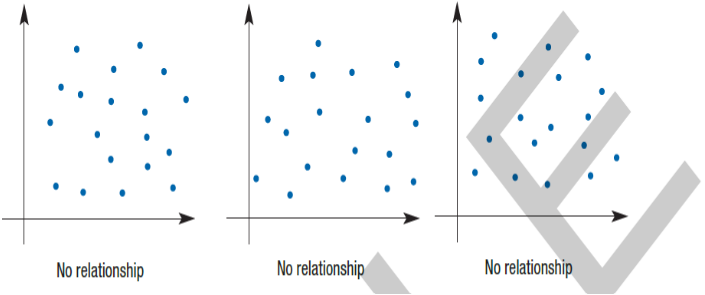
Moderate relationship: As the amount of scatter in the plot increases, the pattern becomes less clear. This indicates that the relationship is less strong. In the examples below, we might say that there is a moderate relationship between the variables.



Weak relationship: As the amount of scatter increases further the pattern bcomes even less clear. This indicates that any relationship between the variables is weak. The scatterplots below are examples of



No relationship: Finally, when all we have is scatter, as seen in the scatterplots below, no pattern can be seen. In this situation we say that there is no relationship between the variables.



Covariance of the couple (X, Y) denoted $Cov(X, Y)$ is given by the formula below ;

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

it can be written also;

$$Cov(X, Y) = \sum_{i=1}^{l}\sum_{j=1}^{k} f_{ij}x_i y_j - \bar{x}\bar{y}.$$

Properties

1) $Cov(X, Y) \in R$.
2) $Cov(X, Y) = Cov(Y, X)$.
3) $Cov(X, X) = V(X)$.
4) $V(X + Y) = V(X) + V(Y) + 2Cov(X, Y)$.

Example:

$Cov(X, Y) = 2 * (6 * 0 + 8 * 0 + 10 * 2/234 + 13 * 36/234) + 7 * (6 * 0 + 8 * 3/234 + 10 * 12/234 + 13 * 26/234) + 15 * (6 * 2/234 + 8 * 8/234 + 10 * 35/234 + 13 * 16/234) + 25 * (6 * 0 + 8 * 26/234 + 10 * 22/234 + 13 * 3/234) + 45 * (6 * 22/234 + 8 * 15/234 + 10 * 6/234 + 13 * 0) - 19.179 * 10.18 = 169.1239 - 195.2422 == -26.118.$

it is quantitative measure to help us assess the degree of association between X and Y; denoted r and given by the expression

$$r = \frac{Cov(X, Y)}{\sigma_X * \sigma_Y}$$

such that $\sigma_X = \sqrt{V(X)}$ and $\sigma_Y = \sqrt{V(Y)}$.

Properties:

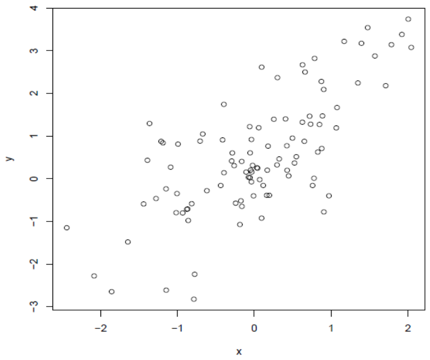
1) $r \in [-1, 1]$.
2) $|r| = 1$ means perfect linear relation ship between X and Y.
3) If X and Y are independant then $r = 0$.
4) r close to $+1$; data shows a strong positive linear correlation (large X values go with large Y).
5) r close to -1; data shows a strong negative linear correlation (large X values go with small Y).
6)r close to 0; data shows no or weak linear correlation. Other non-linear trends may be possible.

Remark:

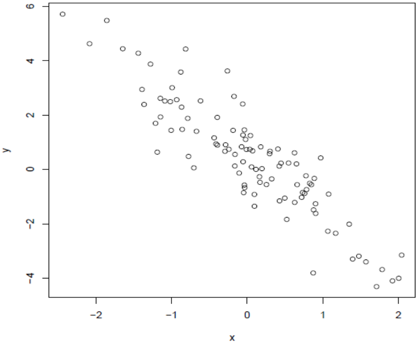
$r = 0$ doesn't imply necessarly X and Y independant.

Example: $r = \frac{-26.118}{\sqrt{(208.388)}\sqrt{(5.6881)}} = \frac{-26.118}{14.43 * 2.38} = \frac{-26.118}{34.3434} = -0.7604$.

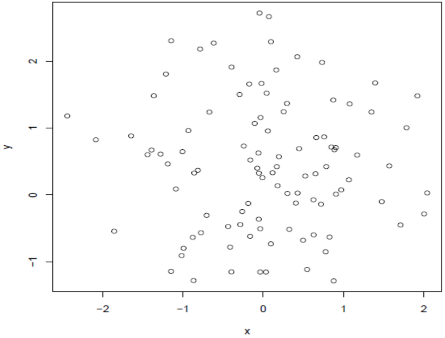
Positive slope (or upward trend)



Negative slope (or downward trend)



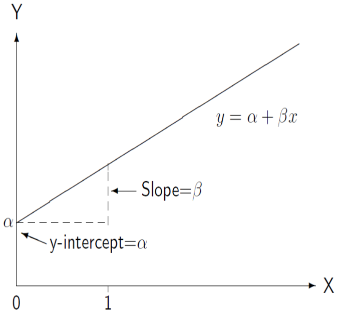
Random scatter (or no apparant pattern)



From the scatter plot, if we notice that the points are very close to a straight line with a positive slope ($r > 0$) or a negative slope ($r < 0$)

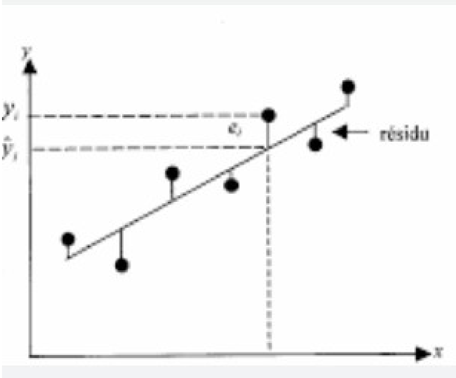
Equation of a straight line is $y = \alpha + \beta x$.
β is the slope and α is the y intercept.

Diagram



The regression line

The regression (or fitted) line is given by $\hat{y} = a + b * x$



a and b are called the least squares estimators, because they minimise the sum of squared distances between observed y's and estimated $\hat{y}$'s.
To minimize the sum S(a,b);

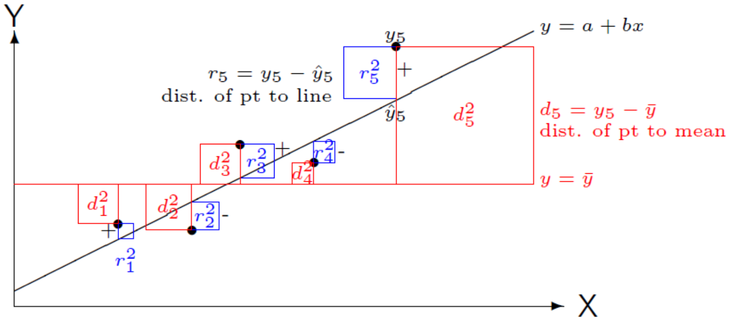
$$S(a,b) = \sum_{i=1}^{N} (y_i - b * x_i - a)^2$$

we should have

$$\begin{cases} \frac{\partial S(a,b)}{\partial a} = 0 \\ \frac{\partial S(a,b)}{\partial b} = 0 \end{cases} \Rightarrow \begin{cases} b = \frac{Cov(X,Y)}{V(X)} \\ a = \bar{y} - b\bar{x} \end{cases}$$

Remark

r^2 gives the proportion of variability explained by a linear relationship between X and Y.



Residuals

then the equation is ;

$$\hat{y} = -4.59 * x + 98.21.$$

1) The following table represents the distribution of 25 students according to the number of annual absences X and the final grade Y.

X|Y	$[0,5[$	$[5,10[$	$[10,15[$	$[15,20[$
0	1	0	3	2
1	0	2	3	4
2	3	1	α	0
3	2	1	0	0
4	1	0	1	0

1) what are the value and the interpretation of α?
2) Give the marginal distributions of X and Y.
3) Give the conditional distribution of Y given $X = 3$.
4) X and Y are they independant?justify.
5) Calculate r regression coefficient, what is your conclusion?
6) Give the regression equation of X and Y.

Solution:

1) $\alpha = 25 - (1+3+2+2+3+4+3+1+2+1+1+1) = 1.$
α is number of students who have two annual absences and their final grade are between 10 and 15.

2) The marginal distribution of X;

X	0	1	2	3	4
$f_{i.}$	$6/25$	$9/25$	$5/25$	$3/25$	$2/25$

The marginal distribution of Y;

Y	$[0,5[$	$[5,10[$	$[10,15[$	$[15,20[$
$f_{.j}$	$7/25$	$4/25$	$8/25$	$6/25$

3) The marginal distribution of Y given X=3 notée $f_{j|X=3}$;

Y	$[0, 5[$	$[5, 10[$	$[10, 15[$	$[15, 20[$
$f_{j\|X=3}$	$\frac{2/25}{3/25}$=2/3	$\frac{1/25}{3/25}$=1/3	0	0

4) X and Y aren't independant because if we consider $Y \in [10, 15[$ and $X = 3$,

$$f_{ij} = 0 \neq f_{i.}(= 3/25) * f_{.j}(= 8/25) \;\; \text{counter-example}$$

.

5) to have r, we should compute $Cov(X, Y)$;
$Cov(X, Y) = (1/25)[(2.5 * 0 + 7.5 * 2 + 12.5 * 3 + 17.5 * 4) + 2 * (2.5 * 3 + 7.5 * 1 + 12.5 * 1) + 3 * (2.5 * 2 + 7.5 * 1) + 4 * (2.5 * 1 + 12.5 * 1)] - (9/25 + 2 * 5/25 + 3 * 3/25 + 4 * 2/25) * (2.5 * 7/25 + 7.5 * 4/25 + 12.5 * 8/25 + 17.5 * 6/25) = 11 - 1.44 * 10.1 = -3.544.$
$r = \frac{Cov(X,Y)}{\sigma_X * \sigma_Y}.$
$V(X) = (9/25 + 2^2 * 5/25 + 3^2 * 3/25 + 4^2 * 2/25) - 1.44^2 = 1.4464$
$\sigma_X = \sqrt{1.4464} = 1.203.$
$V(Y) = (2.5^2 * 7/25 + 7.5^2 * 4/25 + 12.5^2 * 8/25 + 17.5^2 * 6/25 - 10.1^2) = 32.24$
$\sigma_Y = \sqrt{32.24} = 5.68.$
$r = \frac{-3.544}{1.203 * 5.68} = -0.52.$
$r < 0$ there is a negative linear correlation between X and Y.

6) Regression equation;

$$\begin{cases} b = \frac{-3.544}{1.4464} = -2.4502 \\ a = 10.1 - 1.44 * (-2.4502) = 13.6283 \end{cases}$$

then;

$$\hat{y} = -2.4502 * x + 13.6283.$$

2) We asked 10 students the number of hours spent preparing for the computer science exam and the grade was noted for each of them We had the following table:

ω_i	ω_1	ω_2	ω_3	ω_4	ω_5	ω_6	ω_7	ω_8	ω_9	ω_{10}
$X(\omega_i)$	5	1	4	10	6	8	8	9	10	3
$Y(\omega_i)$	7	6	5	15	12	10	14	19	16	7

1) Calculate marginal sample means for X and Y.
2) Calculate $Cov(X, Y)$.
3) Calculate the correlation coefficient r for (X,Y).
4) Give the regression equation of X and Y.
5) Estimate the grade of student who has spent 7 hours to prepare the exam.

Solution:

1) $\bar{x} = 64/10 = 6.4.$

$\bar{y} = 111/10 = 11.1.$

2) $Var(X) = 496/10 - 6.4^2 = 8.64.$

$Var(Y) = 1441/10 - 11.1^2 = 20.89.$

$Cov(X, Y) = 827/10 - 6.4 * 11.1 = 11.66.$

3) $r = \frac{11.66}{\sqrt{8.64} * \sqrt{20.89}} = 0.8679.$

4) $\begin{cases} b = \frac{11.66}{8.64} = 1.35 \\ a = 11.1 - b * 6.4 = 2.46 \end{cases}$

the regression equation is :$\hat{y} = 1.35 * x + 2.46.$

5)the estimate of the grade is $\hat{y} = 1.35 * 7 + 2.46 = 11.91.$