Unit 9

Bivariate Statistics

My goals

By the end of this unit, I will be able to:

- find measures of central tendency in two quantitative variables.
- find measures of variability in two quantitative variables.
- determine the linear regression line of a given series.
- calculate a linear correlation coefficient of a given double series and interpret it.

Introduction

Descriptive statistics is a set of brief descriptive coefficients that summarises a given data set, which can either be a representation of the entire population or sample. Data may be **qualitative** such as sex, color and so on or **quantitative** represented by numerical quantity such as height, mass, time and so on.

The measures used to describe the data are measures of central tendency and measures of variability or dispersion. Until now, we know how to determine the measures of central tendency in one variable. In this unit, we will use those measures in two quantitative variables known as **double series**.

In statistics, double series include technique of analyzing data in two variables, when focus on the relationship between a dependent variable-y and an independent variable-x. The **linear regression** method will be used

in this unit. The estimation target is a function of the independent variable called the **regression function** which will be a function of a straight line.

Descriptive statistics provide useful summary of security returns when performing empirical and analytical analysis, as they provide historical account of return behavior. Although past information is useful in any analysis, one should always consider the expectations of future events.

Some variables are **discrete**, others are **continuous**. If the variable can take only certain values, for example, the number of apples on a tree, then the variable is discrete. If however, the variable can take any decimal value (in some range), for example, the heights of the children in a school, then the variables are continuous. In this unit, we will consider discrete variables.

9.1. Covariance

Activity 9.1

Complete the following table

i	x_i	y_i	$x_i-\bar{x}$	$y_i-\bar{y}$	$(x_i-\bar{x})(y_i-\bar{y})$
1	3	6			
2	5	9			
3	7	12			
4	3	10			
5	2	7			
6	6	8			
	$\sum_{i=1}^{6} x_i = \ldots$	$\sum_{i=1}^{6} y_i = \ldots$			$\sum_{i=1}^{6}(x_i-\bar{x})(y_i-\bar{y}) = \ldots$
	$\bar{x} = \ldots\ldots$	$\bar{y} = \ldots\ldots$			

What can you get from the following expressions:

1. $\sum_{i=1}^{k}\left(x_i-\bar{x}\right)\left(x_i-\bar{x}\right)$ 2. $\sum_{i=1}^{k}\left(x_i-\bar{x}\right)\left(y_i-\bar{y}\right)$

In case of two variables, say x and y, there is another important result called **covariance of** x **and** y, denoted $\operatorname{cov}(x,y)$.

The **covariance of variables** x **and** y is a measure of how these two variables change together. If the greater values of one variable mainly correspond with the greater values of the other variable, and the same holds for the smaller values, i.e. the variables tend to show similar behavior, the covariance is positive. In the opposite case, when the greater values of one variable mainly correspond to the smaller values of the other, i.e. the variables tend to show opposite behavior, the covariance is negative. If covariance is zero, the variables are said to be **uncorrelated,** meaning that there is no linear relationship between them.

Therefore, the sign of covariance shows the tendency in the linear relationship between the variables. The magnitude of covariance is not easy to interpret.

Covariance of variables x and y, where the summation of frequencies $\sum_{i=1}^{k} f_i = n$ are equal for both variables, is defined to be

$$\operatorname{cov}(x,y)=\frac{1}{n}\sum_{i=1}^{k} f_i\left(x_i-\bar{x}\right)\left(y_i-\bar{y}\right)$$

Developing this formula, we have

$$\begin{aligned}\operatorname{cov}(x,y) &= \frac{1}{n}\sum_{i=1}^{k} f_i\left(x_i y_i - x_i\bar{y} - \bar{x}y_i + \bar{x}\,\bar{y}\right)\\ &= \frac{1}{n}\sum_{i=1}^{k} f_i x_i y_i - \frac{1}{n}\sum_{i=1}^{k} f_i x_i\,\bar{y} - \frac{1}{n}\sum_{i=1}^{k} f_i \bar{x}\,y_i + \frac{1}{n}\sum_{i=1}^{k} f_i \bar{x}\,\bar{y}\\ &= \frac{1}{n}\sum_{i=1}^{k} f_i x_i y_i - \frac{1}{n}\bar{y}\sum_{i=1}^{k} f_i x_i - \frac{1}{n}\bar{x}\sum_{i=1}^{k} f_i\, y_i + \bar{x}\,\bar{y}\frac{1}{n}\sum_{i=1}^{k} f_i \qquad \left[\frac{1}{n}\sum_{i=1}^{k} f_i = \frac{1}{n}\times n = 1\right]\\ &= \frac{1}{n}\sum_{i=1}^{k} f_i x_i y_i - \bar{x}\,\bar{y} - \bar{x}\,\bar{y} + \bar{x}\,\bar{y}\\ &= \frac{1}{n}\sum_{i=1}^{k} f_i x_i y_i - \bar{x}\,\bar{y}\end{aligned}$$

Thus, the covariance is also given by

$$\operatorname{cov}(x,y) = \frac{1}{n}\sum_{i=1}^{k} f_i x_i y_i - \bar{x}\,\bar{y}$$

Example 9.1

Find the covariance of x and y in the following data sets

x	3	5	6	8	9	11
y	2	3	4	6	5	8

Solution

We have

x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$
	2	-4	-2.6	10.4
	3	-2	-1.6	3.2
	4	-1	-0.6	0.6
	6	1	1.4	1.4
	5	2	0.4	0.8
11	8	4	3.4	13.6
$\sum_{i=1}^{6} x_i = 42$	$\sum_{i=1}^{6} y_i = 28$			$\sum_{i=1}^{6}(x_i - \bar{x})(y_i - \bar{y}) = 30$
$\bar{x} = 7$	$\bar{y} = 4.6$			

Thus,

$$\operatorname{cov}(x,y) = \frac{1}{6}\sum_{i=1}^{6} f_i\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)$$
$$= \frac{1}{6}(30)$$
$$= 5$$

Example 9.2

Find the covariance of the following distribution

y \ x	0	2	4
1	2	1	3
2	1	4	2
3	2	5	0

Solution

Convert the double entry into a simple table and compute the arithmetic means

x_i	y_i	f_i	$x_i f_i$	$y_i f_i$	$x_i y_i f_i$
0	1	2	0	2	0
0	2	1	0	2	0
0	3	2	0	6	0
2	1	1	2	2	2
2	2	4	8	8	16
2	3	5	10	15	30
4	1	3	12	3	12
4	2	2	8	4	16
4	3	0	0	0	0
		$\sum_{i=1}^{9} f_i = 20$	$\sum_{i=1}^{9} x_i f_i = 40$	$\sum_{i=1}^{9} y_i f_i = 41$	$\sum_{i=1}^{9} x_i y_i f_i = 76$

$$\bar{x} = \frac{40}{20} = 2\,,\quad \bar{y} = \frac{41}{20} = 2.05$$

$$\operatorname{cov}(x,y)=\frac{76}{20}-2\times 2.05=-0.3$$

Alternative method

$$\operatorname{cov}(x,y)=\frac{1}{n}\sum_{i=1}^{k} f_i x_i y_i-\bar{x}\,\bar{y} \qquad \bar{x}=\frac{1}{n}\sum_{i=1}^{k} x_i f_i,\quad \bar{y}=\frac{1}{n}\sum_{i=1}^{k} y_i f_i$$

y	0	2	4	Total
1	2	1	3	6
2	1	4	2	7
3	2	5	0	7
Total	5	10	5	20

$$\bar{x}=\frac{1}{20}(0\times 5+2\times 10+4\times 5)$$
$$=\frac{40}{20}=2$$
$$\bar{y}=\frac{1}{20}(1\times 6+2\times 7+3\times 7)$$
$$=\frac{41}{20}=2.05$$
$$\operatorname{cov}(x,y)=\frac{1}{20}\begin{pmatrix}0\times 1\times 2+0\times 2\times 1+0\times 3\times 2+2\times 1\times 1+2\times 2\times 4\\+2\times 3\times 5+4\times 1\times 3+4\times 2\times 2+4\times 3\times 0\end{pmatrix}-2\times 2.05$$
$$=\frac{1}{20}(0+0+0+2+16+30+12+16+0)-4.1$$
$$=\frac{76}{20}-4.1$$
$$=-0.3$$

Exercise 9.1

1. The scores of 12 students in their mathematics and physics classes are

Mathematics	2	3	4	4	5	6	6	7	7	8	10	10
Physics	1	3	2	4	4	4	6	4	6	7	9	10

Find the covariance of the distribution.

2. The values of two variables x and y are distributed according to the following table

y \ x	100	50	25
14	1	1	0
18	2	3	0
22	0	1	2

Calculate the covariance

9.2. Regression lines

We use the regression line to **predict** a value of y for any given value of x and vice versa. The "best" line would make the best predictions: the observed y-values should stray as little as possible from the line. This straight line is the regression line from which we can adjust its algebraic expressions and it is written as $y = ax + b$.

Activity 9.2

The regression line y on x has the form $y = ax + b$. We need the distance from this line to each point of the given data to be small, so that the sum of the square of such distances be very small. That is $D = \sum_{i=1}^{k} \left[y_i - (ax_i + b) \right]^2$ or $D = \sum_{i=1}^{k} (y_i - ax_i - b)^2$ (1) is minimum.

1. Differentiate relation (1) with respect to b. In this case, y, x and a will be considered as constants.
2. Equate relation obtained in 1) to zero, divide each side by n and give the value of b.

3. Take the value of b obtained in 2) and put it in relation obtained in 1). Differentiate the obtained relation with respect to a, equate it to zero and divide both sides by n to find the value of a.
4. Using the relations: The variance for variable x is $\sigma_x^2 = \frac{1}{n}\sum_{i=1}^{k}\left(x_i - \bar{x}\right)^2$ and the variance for variable y is $\sigma_y^2 = \frac{1}{n}\sum_{i=1}^{k}\left(y_i - \bar{y}\right)^2$ and the covariance of these two variables is $\text{cov}(x,y) = \frac{1}{n}\sum_{i=1}^{k}\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)$, give the simplified expression equal to a.
5. Put the value of b obtained in 2) and the value of a obtained in 4) in relation $y = ax + b$ and give the expression of regression line y on x.

From Activity 9.2, the regression line y on x is written as

$$y = \frac{\text{cov}(x,y)}{\sigma_x^2}x + \left(\bar{y} - \frac{\text{cov}(x,y)}{\sigma_x^2}\bar{x}\right)$$

We may write

$$L_{y\,x} \equiv y - \bar{y} = \frac{\text{cov}(x,y)}{\sigma_x^2}\left(x - \bar{x}\right)$$

Note that the regression line x on y is $x = cy + d$ given by

$$x - \bar{x} = \frac{\text{cov}(x,y)}{\sigma_y^2}\left(y - \bar{y}\right)$$

This line is written as

$$L_{x/y} \equiv x - \bar{x} = \frac{\text{cov}(x,y)}{\sigma_y^2}\left(y - \bar{y}\right)$$

Short cut method of finding regression line

To abbreviate the calculations, the two regression lines can be determined as follows:

a) Relation y-x is $L_{y/x} \equiv y = ax + b$ and the values of a and b are found by solving the simultaneous equations:

$$\begin{cases} \sum_{i=1}^{k} f_i y_i = a\sum_{i=1}^{k} f_i x_i + b\,n \\ \sum_{i=1}^{k} f_i x_i y_i = a\sum_{i=1}^{k} f_i x_i^2 + b\sum_{i=1}^{k} f_i x_i \end{cases}$$

These equations are called the normal equations for y on x.

a) Relation x-y is $L_{x/y} \equiv x = cy + d$ and the values of c and d are found by solving the simultaneous equations:

$$\begin{cases} \sum_{i=1}^{k} f_i x_i = c\sum_{i=1}^{k} f_i y_i + d\,n \\ \sum_{i=1}^{k} f_i x_i y_i = c\sum_{i=1}^{k} f_i y_i^2 + d\sum_{i=1}^{k} f_i y_i \end{cases}$$

These equations are called the normal equations for x on y.

Example 9.3

Find the regression line of y on x for the following data and estimate the value of y for $x = 4, x = 7, x = 16$ and the value of x for $y = 7, y = 9, y = 16$.

x	3	5	6	8	9	11
y	2	3	4	6	5	8

Solution

x	y	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})^2$	$(y-\bar{y})^2$	$(x-\bar{x})(y-\bar{y})$
3	2	-4	-2.6	16	6.76	10.4
5	3	-2	-1.6	4	2.56	3.2
6	4	-1	-0.6	1	0.36	0.6
8	6	1	1.4	1	1.96	1.4
9	5	2	0.4	4	0.16	0.8
11	8	4	3.4	16	11.56	13.6
$\sum_{i=1}^{6} x_i = 42$	$\sum_{i=1}^{6} y_i = 28$			$\sum_{i=1}^{6}(x_i-\bar{x})^2 = 42$	$\sum_{i=1}^{6}(y_i-\bar{y})^2 = 23.36$	$\sum_{i=1}^{6}(x_i-\bar{x})(y_i-\bar{y}) = 30$

$$\bar{x} = \frac{42}{6} = 7\,,\ \bar{y} = \frac{28}{6} = 4.7$$

$$\operatorname{cov}(x, y) = \frac{1}{n}\sum_{i=1}^{k}\left(x - \bar{x}\right)\left(y - \bar{y}\right) = \frac{30}{6} = 5$$

$$\sigma_x^2 = \frac{42}{6} = 7\,,\ \sigma_y^2 = \frac{23.36}{6} = 3.89$$

$$L_{y/x} \equiv y - \bar{y} = \frac{\operatorname{cov}(x, y)}{\sigma_x^2}\left(x - \bar{x}\right)$$

$$L_{y/x} \equiv y - 4.7 = \frac{5}{7}(x - 7)$$

Finally, the line of y on x is

$$L_{y/x} \equiv y = \frac{5}{7}x - 0.3$$

And

$$L_{x/y} \equiv x - \bar{x} = \frac{\operatorname{cov}(x, y)}{\sigma_y^2}\left(y - \bar{y}\right)$$

$$L_{x/y} \equiv x - 7 = \frac{5}{3.89}(y - 4.7)$$

Finally, the line of x on y is

$$L_{x/y} \equiv y = 1.3x + 1$$

Alternative method

x	y	x^2	y^2	xy
3	2	9	4	6
5	3	25	9	15
6	4	36	16	24
8	6	64	36	48
9	5	81	25	45
11	8	121	64	88
$\sum_{i=1}^{6} x_i = 42$	$\sum_{i=1}^{6} y_i = 28$	$\sum_{i=1}^{6} x_i^2 = 336$	$\sum_{i=1}^{6} y_i^2 = 154$	$\sum_{i=1}^{6} x_i y_i = 226$

$$L_{y/x} \equiv y = ax + b$$

$$\begin{cases} \sum_{i=1}^{k} f_i y_i = a\sum_{i=1}^{k} f_i x_i + b\,n \\ \sum_{i=1}^{k} f_i x_i y_i = a\sum_{i=1}^{k} f_i x_i^2 + b\sum_{i=1}^{k} f_i x_i \end{cases}$$

$$\begin{cases} 28 = 42a + 6b \\ 226 = 336a + 42b \end{cases} \Leftrightarrow \begin{cases} a = \dfrac{5}{7} \\ b = -0.3 \end{cases}$$

Thus, the line of y on x is

$$L_{y/x} \equiv y = \frac{5}{7}x - 0.3$$

If

$$x = 4 \Rightarrow y = 2.5$$
$$x = 7 \Rightarrow y = 4.7$$
$$x = 16 \Rightarrow y = 11.1$$

$$L_{x/y} \equiv x = cy + d$$

$$\begin{cases} \sum_{i=1}^{k} f_i x_i = c\sum_{i=1}^{k} f_i y_i + d\,n \\ \sum_{i=1}^{k} f_i x_i y_i = c\sum_{i=1}^{k} f_i y_i^2 + d\sum_{i=1}^{k} f_i y_i \end{cases}$$

$$\begin{cases} 42 = 28c + 6d \\ 226 = 154c + 28d \end{cases} \Leftrightarrow \begin{cases} c = 1.3 \\ d = 1 \end{cases}$$

Thus, the line of x on y is

$$L_{x/y} \equiv x = 1.3y + 1$$

If

$$y = 7 \Rightarrow x = 10.1$$
$$y = 9 \Rightarrow x = 12.7$$
$$y = 16 \Rightarrow x = 21.8$$

Exercise 9.2

1. Consider the following table:

x	y
60	3.1
61	3.6
62	3.8
63	4
65	4.1

a) Find the regression line of y on x

b) Calculate the approximate y value for the variable $x = 64$.

2. The values of two variables x and y are distributed according to the following table.

y \ x	100	50	25
14	1	1	0
18	2	3	0
22	0	1	2

Find the regression lines.

9.3. Coefficient of correlation

Pearson's coefficient of correlation (or product moment coefficient of correlation)

Activity 9.3

Consider the following table:

x	y
3	6
5	9
7	12
3	10
2	7
6	8

1. Find the standard deviations σ_x, σ_y
2. Find covariance $\text{cov}(x, y)$
3. Calculate the ratio $\dfrac{\text{cov}(x, y)}{\sigma_x \sigma_y}$.

The **Pearson's coefficient of correlation (**or **product moment coefficient of correlation** or simply **coefficient of correlation),** denoted by r, is a measure of the strength of linear relationship between two variables.

The coefficient of correlation between two variables x and y is given by

$$r = \frac{\text{cov}(x, y)}{\sigma_x \sigma_y}$$

Where,

$\text{cov}(x, y)$ is covariance of x and y

σ_x is the standard deviation for x

σ_y is the standard deviation for y

Properties of the coefficient of correlation

a) The coefficient of correlation does not change the measurement scale. That is, if the height is expressed in metres or feet, the coefficient of correlation does not change.

b) The sign of the coefficient of correlation is the same as the covariance.

c) The square of the coefficient of correlation is equal to the product of angular coefficients (slopes) of two regression lines.

In fact, $r = \frac{\text{cov}(x, y)}{\sigma_x \sigma_y}$. Squaring both sides gives

$$r^2 = \left[\frac{\text{cov}(x, y)}{\sigma_x \sigma_y}\right]^2$$

$$= \frac{\text{cov}^2(x, y)}{\sigma_x^2 \sigma_y^2} = \frac{\text{cov}(x, y)}{\sigma_x^2} \times \frac{\text{cov}(x, y)}{\sigma_y^2}$$

d) If the coefficient of correlation is known, it can be used to find the angular coefficients of two regression lines.

We know that the angular coefficient of the regression line y on x is $\frac{\text{cov}(x,y)}{\sigma_x^2}$. From this, we have;

$$\frac{\text{cov}(x,y)}{\sigma_x^2} = \frac{\text{cov}(x,y)}{\sigma_x\sigma_x} \times \frac{\sigma_y}{\sigma_y}$$

$$= \frac{\text{cov}(x,y)}{\sigma_x\sigma_y} \times \frac{\sigma_y}{\sigma_x} = r\frac{\sigma_y}{\sigma_x}$$

We know that the angular coefficient of the regression line x on y is $\frac{\text{cov}(x,y)}{\sigma_y^2}$. From this, we have;

$$\frac{\text{cov}(x,y)}{\sigma_y^2} = \frac{\text{cov}(x,y)}{\sigma_y\sigma_y} \times \frac{\sigma_x}{\sigma_x}$$

$$= \frac{\text{cov}(x,y)}{\sigma_x\sigma_y} \times \frac{\sigma_x}{\sigma_y} = r\frac{\sigma_x}{\sigma_y}$$

Thus, the angular coefficient of the regression line y on x is given by $r\frac{\sigma_y}{\sigma_x}$ and the angular coefficient of the regression line y on x is given by $r\frac{\sigma_x}{\sigma_y}$.

e) Cauchy Inequality: $\text{cov}^2(x,y) \le \sigma_x^2\sigma_y^2$

In fact, $r = \frac{\text{cov}(x,y)}{\sigma_x\sigma_y} \Leftrightarrow \text{cov}(x,y) = r\sigma_x\sigma_y$.

Squaring both sides gives $\text{cov}^2(x,y) = r^2\sigma_x^2\sigma_y^2$

Or $\text{cov}^2(x,y) \le \sigma_x^2\sigma_y^2$

f) The coefficient of correlation takes value ranging between -1 and +1. That is, $-1 \le r \le 1$

In fact, from Cauchy Inequality, we have,

$\text{cov}^2(x,y) \le \sigma_x^2\sigma_y^2$

$$\Leftrightarrow \frac{\text{cov}^2(x,y)}{\sigma_x^2\sigma_y^2} \le 1 \Leftrightarrow \left[\frac{\text{cov}(x,y)}{\sigma_x\sigma_y}\right]^2 \le 1 \Leftrightarrow r^2 \le 1$$

Taking square roots both sides

$\Leftrightarrow \sqrt{r^2} \le 1$

$\Leftrightarrow |r| \le 1$ since $\sqrt{x^2} = |x|$

$|r| \le 1$ is equivalent to $-1 \le r \le 1$.

Thus, $-1 \le r \le 1$

g) If the linear coefficient of correlation takes values closer to **−1**, the **correlation** is **strong and negative**, and will become stronger the closer r approaches -1.

h) If the linear coefficient of correlation takes values close to **1,** the **correlation** is **strong and positive**, and will become stronger the closer r approaches 1.

i) If the linear coefficient of correlation takes values close to **0**, the **correlation** is **weak**.

j) If r = 1 or r = −1, there is **perfect correlation** and the line on the scatter plot is increasing or decreasing respectively.

k) If r = 0, there is **no linear correlation**.

Example 9.4

Considering Example 9.3, we have seen that

$\text{cov}(x,y) = 5$

$\sigma_x^2 = \frac{42}{6} = 7$, $\sigma_y^2 = \frac{23.36}{6} = 3.89$

Then, the Pearson's coefficient of correlation is

$$r = \frac{\text{cov}(x,y)}{\sigma_x\sigma_y} \qquad r = \frac{5}{\sqrt{7}\sqrt{3.89}} = \frac{5}{\sqrt{27.23}} = 0.96$$

Then, there is a very strong positive linear relationship between two variables.

We have also seen that the two regression lines are

$$L_{y/x} \equiv y = \frac{5}{7}x - 0.3$$

$$L_{x/y} \equiv x = 1.3y + 1$$

Their slopes are $\alpha = \frac{5}{7}$ and $\beta = 1.3$

We see that $r^2 = (0.96)^2 = 0.92$. On the other hand,

$\alpha.\beta = \frac{5}{7} \times 1.3 = 0.92$.

Thus, $r^2 = \alpha.\beta$

Example 9.5

A test is made over 200 families on number of children x and number of beds y per family. Results are collected in the table below:

x \ y	0	1	2	3	4	5	6	7	8	9	10
1	0	2	7	5	2	0	0	0	0	0	0
2	2	2	10	8	15	1	0	0	0	0	0
3	1	3	5	6	8	6	1	0	0	0	0
4	0	2	8	2	6	12	10	8	0	0	0
5	0	1	0	2	5	6	10	5	7	3	3
6	0	0	0	2	4	5	5	2	3	3	2

a) What is the average number for children and beds per a family?

b) Find the regression line of y on x.

c) Can we confirm that there is a high linear correlation between the number of children and number of beds per family?

Solution

a) Average number of children per family:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{k} f_i x_i, \quad \bar{y} = \frac{1}{n}\sum_{i=1}^{k} f_i y_i$$

x \ y	0	1	2	3	4	5	6	7	8	9	10	Total
1	0	2	7	5	2	0	0	0	0	0	0	16
2	2	2	10	8	15	1	0	0	0	0	0	38
3	1	3	5	6	8	6	1	0	0	0	0	30
4	0	2	8	2	6	12	10	8	0	0	0	48
5	0	1	0	2	5	6	10	5	7	3	3	42
6	0	0	0	2	4	5	5	2	3	3	2	26
Total	3	10	30	25	40	30	26	15	10	6	5	200

$$\bar{x} = \frac{1}{200}(3\times0+10\times1+30\times2+25\times3+40\times4+30\times5+26\times6+15\times7+10\times8+6\times9+5\times10)$$

$$= \frac{900}{200} = 4.5$$

Or there are about 5 children per family.

Average number of beds per family:

$$\bar{y} = \frac{1}{200}(16\times1+38\times2+30\times3+48\times4+42\times5+26\times6)$$

$$= \frac{740}{200} = 3.7$$

Or there are about 4 beds per family.

b) The equation of regression line of y on x is given by equation

$$y-\bar{y} = \frac{\operatorname{cov}(x,y)}{\sigma_x^2}(x-\bar{x})$$

where $\bar{y} = 3.7$ and $\bar{x} = 4.5$

$$\sigma_x^2 = \frac{1}{n}\sum_{i=1}^{k} x_i^2 f_i - \left(\bar{x}\right)^2$$

$$= \frac{1}{200}\begin{pmatrix} 3\times 0^2 + 10\times 1^2 + 30\times 2^2 + 25\times 3^2 + 40\times 4^2 + 30\times 5^2 \\ +26\times 6^2 + 15\times 7^2 + 10\times 8^2 + 6\times 9^2 + 5\times 10^2 \end{pmatrix} - (4.5)^2$$

$$= \frac{5042}{200} - 20.25$$

$$= 4.96$$

$$\operatorname{cov}(x,y) = \frac{1}{200}\sum_{i=1}^{66} f_i x_i y_i - \bar{x}\,\bar{y}$$

$$= \frac{1}{200}\begin{pmatrix} 0+2+14+15+8+0+4+40+48+120+10+0 \\ +9+30+54+96+90+18+0+8+64+24+96 \\ +240+240+224+0+5+0+30+100+150 \\ +300+175+280+135+150+0+36+96+150 \\ +180+84+144+162+120 \end{pmatrix} - 4.5\times 3.7$$

$$= \frac{3751}{200} - 16.65$$

$$= 18.7555 - 16.65$$

$$= 2.105$$

The required equation of regression of y on x is

$$y - 3.7 = \frac{2.105}{4.96}(x - 4.5)$$

Or

$$y = 0.4x + 1.8$$

c) Coefficient of correlation is given by $\dfrac{\operatorname{cov}(x,y)}{\sigma_x \sigma_y}$

$$\sigma_y^2 = \frac{1}{200}\sum_{i=1}^{6} f_i y_i^2 - \left(\bar{y}\right)^2$$

$$= \frac{1}{200}(16 + 38\times 4 + 30\times 9 + 48\times 16 + 42\times 25 + 26\times 36) - (3.7)^2$$

$$= 15.96 - 13.69$$

$$= 2.27$$

Therefore, the coefficient of correlation is

$$r = \frac{2.105}{\sqrt{4.96}\sqrt{2.27}} \approx 0.63$$

There is a high linear correlation.

Notice

Spearman's coefficient of rank correlation

A Spearman coefficient of rank correlation or Spearman's rho is a measure of statistical dependence between two variables. It assesses how well the relationship between two variables can be described using a monotonic function.

The **Spearman's coefficient of rank correlation** is denoted and defined by

$$\rho = 1 - \frac{6\sum_{i=1}^{k} d_i^2}{n\left(n^2 - 1\right)}$$

Where, d refers to the difference of ranks between paired items in two series and n is the number of observations..

It is much easier to calculate the Spearman's coefficient of rank correlation than to calculate the Pearson's coefficient of correlation as there is far less working involved. However, in general, the Pearson's coefficient of correlation is a more accurate measure of correlation.

Method of ranking

Example 9.6

Suppose that we have the marks, x, of seven students in this order:

12, 18, 10, 13, 15, 16, 9

We assign the rank 1, 2, 3, 4, 5, 6, 7 such that the smallest value of x will be ranked 1.

That is

x	12	18	10	13	15	16	9
$Rank\left(x\right)$	3	7	2	4	5	6	1

If we have two or more equal values, we proceed as follows:
Consider the following series

x	66	65	66	67	66	64	68	68

To assign the rank to this series, we do the following:
$x = 64$ will take rank 1, since it is the smallest value of x
$x = 65$ will be ranked 2.
$x = 66$ appears 3 times, since the previous value was ranked 2, here, 66 would be ranked 3, another 66 would be ranked 4 and another 5 but since there are three 66's, we need to find the average of those ranks which is $\frac{3+4+5}{3} = 4$ so that each 66 will be ranked 4.
$x = 67$ will be ranked 6 since we are on the 6th position
$x = 68$ appears 2 times, since the previous value was ranked 6, here, 68 would be ranked 7, and another 66 would be ranked 8 but since there are two 68's, we need to find the average of those ranks which is $\frac{7+8}{2} = 7.5$ so that each 68 will be ranked 7.5

Thus we have the following:

x	66	65	66	67	66	64	68	68
$Rank(x)$	4	2	4	6	4	1	7.5	7.5

Example 9.7

Compute the Spearman's coefficient of rank correlation for the data given in Example 9.3

Solution

x	y	$Rank(x)$	$Rank(y)$	$Rank(x) - Rank(y) = d$	d^2
3	2	1	1	0	0
5	3	2	2	0	0
6	4	3	3	0	0
8	6	4	5	-1	1
9	5	5	4	1	2
11	8	6	6	0	0
					$\sum_{i=1}^{6} d_i^2 = 3$

Then the Spearman's coefficient of correlation is

$$\rho = 1 - \frac{6\sum_{i=1}^{6} d_i^2}{n\left(n^2-1\right)}$$

$$= 1 - \frac{6\times 3}{6\left(36-1\right)}$$

$$= 1 - \frac{18}{210}$$

$$= 0.91$$

Example 9.8

Calculate the Spearman's coefficient of rank correlation for the series.

x	12	8	16	12	7	10	12	16	12	9
y	6	5	7	7	4	6	8	13	10	10

Solution

x	y	$Rank\left(x\right)$	$Rank\left(y\right)$	$Rank\left(x\right)-Rank\left(y\right)=d$	d^2
12	6	6.5	3.5	3	9
8	5	2	2	0	0
16	7	9.5	5.5	4	16
12	7	6.5	5.5	1	1
7	4	1	1	0	0
10	6	4	3.5	0.5	0.25
12	8	6.5	7	0.5	0.25
16	13	9.5	10	0.5	0.25
12	10	6.5	8.5	2	4
9	10	3	8.5	5.5	30.25
					$\sum_{i=1}^{10} d_i^2 = 61$

Then

$$\rho = 1 - \frac{6 \times 61}{10(100-1)} \Leftrightarrow \rho = 1 - \frac{366}{990} \Leftrightarrow \rho = \frac{990 - 366}{990}$$

Or

$$\rho = 0.63$$

Exercise 9.3

1. The scores of 12 students in their mathematics and physics classes are:

Mathematics	2	3	4	4	5	6	6	7	7	8	10	10
Physics	1	3	2	4	4	4	6	4	6	7	9	10

Find the coefficient of correlation distribution and interpret it.

2. The values of the two variables x and y are distributed according to the following table:

y \ x	0	2	4
1	2	1	3
2	1	4	2
3	2	5	0

Calculate the coefficient of correlation.

3. The marks of eight candidates in English and Mathematics are:

Candidate	1	2	3	4	5	6	7	8
English	50	58	35	86	76	43	40	60
Mathematics	65	72	54	82	32	74	40	53

Rank the results and hence find Spearman's rank coefficient of correlation between the two sets of marks. Comment on the value obtained.

4. Find Spearman's rank coefficient of correlation for the following data and interpret the value:

x	1	2.5	6	7	4.5	3	6.5
y	0.5	1	3.5	6.5	3	2.5	5.5

9.4. Applications

Activity 9.4

Discuss how statistics, especially bivariate statistics, can be used in our daily life.

Bivariate statistics can help in prediction of a value for one variable if we know the value of the other.

Example 9.9

One measure of personal fitness is the time taken for an individual's pulse rate to return to normal after strenuous exercise; the greater the fitness, the shorter the time. Following a short program of strenuous exercise, Norman recorded his pulse rates P at time t minutes after he had stopped exercising. Norman's results are given in the table below:

t	0.5	1.0	1.5	2.0	3.0	4.0	5.0
P	125	113	102	94	81	83	71

Estimate Norman's pulse rate 2.5 minutes after stopping the exercise program.

Solution

t	P	t^2	P^2	tP
0.5	125	0.25	15625	62.5
1	113	1	12769	113
1.5	102	2.25	10404	153
2	94	4	8836	188
3	81	9	6561	243
4	83	16	6889	332
5	71	25	5041	355
$\sum_{i=1}^{7} t_i = 17$	$\sum_{i=1}^{7} P_i = 669$	$\sum_{i=1}^{7} t_i^2 = 57.5$	$\sum_{i=1}^{7} P_i^2 = 66125$	$\sum_{i=1}^{7} t_i P_i = 1446.5$

We need the line $P = at + b$

Use the formula

$$\begin{cases} \sum_{i=1}^{7} P_i = a\sum_{i=1}^{7} t_i + bn \\ \sum_{i=1}^{7} t_i P_i = a\sum_{i=1}^{7} t_i^2 + b\sum_{i=1}^{7} t_i \end{cases}$$

We have

$$\begin{cases} 669 = 17a + 7b \\ 1446.5 = 57.5a + 17b \end{cases}$$

Solving, we have

$$\begin{cases} a = -11 \\ b = 122.3 \end{cases}$$

Then, $P = -11t + 122.3$

So,

Norman's pulse rate 2.5 minutes after stopping the exercise program is estimated to be $P = -11(2.5) + 122.3$ or 94.8.

Unit Summary

1. The **covariance of variables** x **and** y is a measure of how these two variables change together. It is defined to be $\text{cov}(x, y) = \frac{1}{n}\sum_{i=1}^{k} f_i(x_i - \bar{x})(y_i - \bar{y})$ or $\text{cov}(x, y) = \frac{1}{n}\sum_{i=1}^{k} f_i x_i y_i - \bar{x}\,\bar{y}$
2. The regression line y on x is $L_{y/x} \equiv y - \bar{y} = \frac{\text{cov}(x, y)}{\sigma_x^2}(x - \bar{x})$
3. The regression line x on y is $L_{x/y} \equiv x - \bar{x} = \frac{\text{cov}(x, y)}{\sigma_y^2}(y - \bar{y})$
4. The coefficient of correlation between two variables x and y is given by $r = \frac{\text{cov}(x, y)}{\sigma_x \sigma_y}$

5. The Spearman's coefficient of rank correlation is given by

$$\rho = 1 - \frac{6\sum_{i=1}^{k} d_i^2}{n\left(n^2 - 1\right)}$$

Where, d refers to the difference of ranks between paired items in two series.

End of Unit Assessment

1. For each set of data, find;
   a) equation of the regression line of y on x
   b) equation of the regression line of x on y

Data set 1

x	3	7	9	11	14	14	15	21	22	23	26
y	5	12	5	12	10	17	23	16	10	10	25

Data set 2

x	1	5	5	5	6	7.5	7.5	7.5	10	11	12.5	14	14.5
y	85	82	85	89	78	66	77	81	70	74	65	69	63

2. The following is a summary of the results of given two variables:

$$\sum_{i=1}^{k} f_i x_i = 500,\ \sum_{i=1}^{k} f_i y_i = 300,\ \sum_{i=1}^{k} f_i x_i^2 = 27818,\ \sum_{i=1}^{k} f_i x_i y_i = 16837,\ \sum_{i=1}^{k} f_i y_i^2 = 10462$$

Find the equation of regression line of y on x.

Estimate the value of y for $x = 60$.

3. Compute the coefficient of correlation for the following series:

x	80	45	55	56	58	60	65	68	70	75	85
y	81	56	50	48	60	62	64	65	70	74	90

4. The following results were obtained from lineups in Mathematics and Physics examinations:

	Mathematics (x)	Physics (y)
Mean	475	39.5
Standard deviation	16.8	10.8

$r = 0.95$

Find both equations of the regression lines. Also estimate the value of y for $x = 30$.

5. The following results were obtained from records of age (x) and systolic blood pressure (y) of a group of 10 men:

	(x)	(y)
Mean	53	142
Variance	130	165

$\sum_{i=1}^{k} f_i\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right) = 1220$

Find both equations of the regression lines. Also estimate the blood pressure of a man whose age is 45.

6. For a given set of data:

$\sum_{i=1}^{k} f_i x_i = 15,\ \sum_{i=1}^{k} f_i y_i = 43,\ \sum_{i=1}^{k} f_i x_i^2 = 55,\ \sum_{i=1}^{k} f_i x_i y_i = 145,\ \sum_{i=1}^{k} f_i y_i^2 = 397,\ \sum_{i=1}^{k} f_i = 5$

Find the equations of the regression lines y on x, and x on y.

7. For a set of 20 pairs of observation s of the variables x and y, it is known that $\sum_{i=1}^{k} f_i x_i = 250,\ \sum_{i=1}^{k} f_i y_i = 140$, and that the regression line of y on x passes through (15,10). Find the equation of that regression line and use it to estimate y when $x = 10$.

8. The gradient of the regression line x on y is -0.2 and the line passes through $(0,3)$. If the equation of the line is $x = c + dy$, find the value of c and d and sketch the line on a diagram.

9. The heights h, in cm, and weights w, in kg, of 10 people are measured. It is found that

$$\sum_{i=1}^{k} f_i h_i = 1710,\ \sum_{i=1}^{k} f_i w_i = 760,\ \sum_{i=1}^{k} f_i h_i^2 = 293162,\ \sum_{i=1}^{k} f_i h_i w_i = 130628,\ \sum_{i=1}^{k} f_i w_i^2 = 59390$$

Calculate the coefficient of correlation between the value of h and w.

What is the equation of the regression line of w on h?

10. The regression equations are $7x - 16y + 9 = 0$ and $5y - 4x - 3 = 0$. Find $\bar{x}$, $\bar{y}$ and r.

11. If two regression coefficients are 0.8 and 0.2, what would be the value of coefficient of correlation?

12. For a given set of data:

$$\sum_{i=1}^{k} f_i x_i = 680,\ \sum_{i=1}^{k} f_i y_i = 996,\ \sum_{i=1}^{k} f_i x_i^2 = 20154,\ \sum_{i=1}^{k} f_i x_i y_i = 24844,\ \sum_{i=1}^{k} f_i y_i^2 = 34670,\ \sum_{i=1}^{k} f_i = 30$$

Find the coefficient of correlation.

13. For a set of data, the equations of the regression lines are

$y = 0.648x + 2.64$ and $x = 0.917y - 1.91$

Find the coefficient of correlation.

14. For a set of data, the equations of the regression lines are

$y = -0.219x + 20.8$ and $x = -0.785y + 16.2$

Find the coefficient of correlation.

15. For a set of data, the equations of the regression lines are $y = 1.3x + 0.4$ and $x = 0.7y - 0.1$

Find;

a) the coefficient of correlation. b) $\bar{x}$ and $\bar{y}$.

16. In a partially destroyed laboratory record of an analysis of correlation data, the following results only are legible: Variance of x is 9

Equations of regression lines: $8x-10y+66=0$ and $40x-18y-214=0$

What were:

a) the mean values of x and y.

b) the standard deviation of y, and

c) the coefficient of correlation between x and y.

17. The following equations of regression lines and variance are obtained from a correlation table:

$20x-9y-107=0$, $4x-5y+33=0$, variance of x is 9.

Find

a) the mean value of x and y.

b) the standard deviation of y.

18. The table below shows the marks awarded to six students in a competition:

Student	A	B	C	D	E	F
Judge 1	6.8	7.3	8.1	9.8	7.1	9.2
Judge 2	7.8	9.4	7.9	9.6	8.9	6.9

Calculate a coefficient of rank correlation.

19. At the end of a season, a league of eight hockey clubs produced the following table showing the position of each club in the league and the average attendances (in hundreds) at home matches.

Club	Position	Average attendance
A	1	27
B	2	29
C	3	9
D	4	16
E	5	24
F	6	15
G	7	12
H	8	22

a) Calculate the Spearman's coefficient of rank correlation between position in the league and average attendance.

b) Comment on your results.

20. A company is to replace its fleet of cars. Eight possible models are considered and the transport manager is asked to rank them, from 1 to 8, in order of preference. A saleswoman is asked to use each type of car for a week and grade them according to their suitability for the job (*A*-very suitable to *E*-unsuitable).

The price is also recorded:

Model	Transport manager's ranking	Saleswoman's grade	Price (£10s)
S	5	B	611
T	1	B+	811
U	7	D-	591
V	2	C	792
W	8	B+	520
X	6	D	573
Y	4	C+	683
Z	3	A-	716

a) Calculate the Spearman's coefficient of rank correlation between:

(i) price and transport manager's rankings,

(ii) price and saleswoman's grades.

b) Based on the result of a, state, giving a reason, whether it would be necessary to use all the three different methods of assessing the cars.

c) A new employee is asked to collect further data and to do some calculations. He produces the following results:

The coefficient of correlation between

(i) price and boot capacity is 1.2,

(ii) maximum speed and fuel consumption in miles per gallons is -0.7,

(iii) price and engine capacity is -0.9.

For each of his results, say giving a reason, whether you think it is reasonable.

d) Suggest two sets of circumstances where Spearman's coefficient of rank correlation would be preferred to the Pearson's coefficient of correlation as a measure of association.

21. The scores obtained by a group of students in tests that measure verbal ability (x) and abstract reasoning (y) are represented in the following table:

y \ x	20	30	40	50
(25-35)	6	4	0	0
(35-45)	3	6	1	0
(45-55)	0	2	5	3
(55-65)	0	1	2	7

a) Is there a correlation between the two variables?

b) According to the data, if one of these students obtained a score of 70 points in abstract reasoning, what would be the estimated score in verbal ability?

Unit 10 Conditional Probability and Bayes Theorem

My goals

By the end of this unit, I will be able to:

- use tree diagram to find probability of events.
- find probability of independent events.
- find probability of one event given that the other event has occurred.
- use and apply Bayes theorem.

Introduction

Probability is a common sense for scholars and people in modern days. It is the chance that something will happen- how likely it is that some event will happen. No engineer or scientist can conduct research and development works without knowing the probability theory. Some academic fields based on the probability theory are statistics, communication theory, computer performance evaluation, signal and image processing, game theory etc. Some applications of the probability theory are character recognition, speech recognition, opinion survey, missile control, seismic analysis...

The theory of game of chance formed the foundations of probability theory, contained at the same time the principle for combinations of elements of a finite set, and thus establishes the traditional connection between combinatorial analysis and probability theory.