UNIT 2
Correlation Analysis

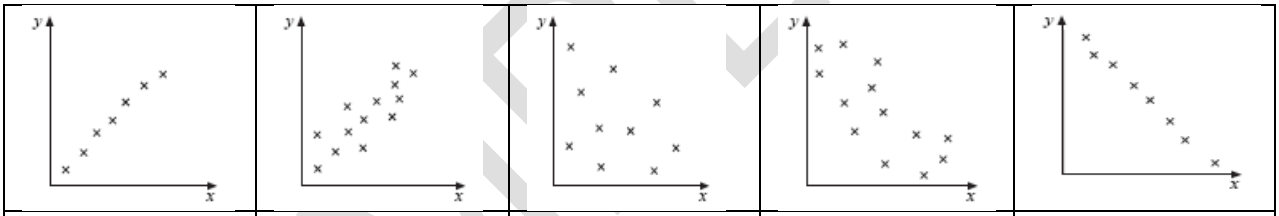
2.1. Definition

Correlation is a statistical measure that indicates the extent to which two or more variables fluctuate together. A positive correlation indicates the extent to which those variables increase or decrease in parallel; a negative correlation indicates the extent to which one variable increases as the other decreases.

When the fluctuation of one variable reliably predicts a similar fluctuation in another variable, there's often a tendency to think that means that the change in one causes the change in the other. However, correlation does not imply causation. There may be an unknown factor that influences both variables similarly.

Correlation is a statistical technique that can show whether and how strongly pairs of variables are related. Although this correlation is fairly obvious your data may contain unsuspected correlations. You may also suspect there are correlations, but don't know which are the strongest. An intelligent correlation analysis can lead to a greater understanding of your data.

- Correlation is **Positive** or direct when the values **increase** together, and
- Correlation is **Negative** when one value **decreases** as the other increases, and so called inverse or contrary correlation.



a	b	c	d	e

If the points plotted were all on a straight line we would have perfect correlation, but it could be positive or negative as shown in the diagrams above,

a. Strong positive correlation between x and y. The points lie close to a straight line with y increasing as x increases.
b. Weak, positive correlation between x and y. The trend shown is that y increases as x increases but the points are not close to a straight line
c. No correlation between x and y; the points are distributed randomly on the graph.
d. Weak, negative correlation between x and y. The trend shown is that y decreases as x increases but the points do not lie close to a straight line
e. Strong, negative correlation. The points lie close to a straight line, with y decreasing as x increases

Correlation can have a value:

3. 1 is a perfect positive correlation
4. 0 is no correlation (the values don't seem linked at all)
5. -1 is a perfect negative correlation

The value shows how good the correlation is (not how steep the line is), and if it is positive or negative. Usually, in statistics, there are three types of correlations: Pearson correlation, Kendall rank correlation and Spearman correlation.

2.2. Assumption of Correlation

Employing of correlation rely on some underlying assumptions. The variables are assumed to be independent, assume that they have been randomly selected from the population; the two variables are normal distribution; association of data is homoscedastic (homogeneous), homoscedastic data have the same standard deviation in different groups where data are heteroscedastic have different standard deviations in different groups and assumes that the relationship between the two variables is linear. The correlation coefficient is not satisfactory and difficult to interpret the associations between the variables in case if data have outliers.

An inspection of a scatterplot can give an impression of whether two variables are related and the direction of their relationship. But it alone is not sufficient to determine whether there is an association between two variables. The relationship depicted in the scatterplot needs to be described qualitatively. Descriptive statistics that express the degree of relation between two variables are called correlation coefficients. A commonly employed correlation coefficient are Pearson correlation, Kendall rank correlation and Spearman correlation.

Correlation used to examine the presence of a linear relationship between two variables providing certain assumptions about the data are satisfied. The results of the analysis, however, need to be interpreted with care, particularly when looking for a causal relationship.

2.3. Bivariate Correlation

Bivariate correlation is a measure of the relationship between the two variables; it measures the strength and direction of their relationship, the strength can range from absolute value 1 to 0. The stronger the relationship, the closer the value is to 1. Direction of The relationship can be positive (direct) or negative (inverse or contrary); correlation generally describes the effect that two or more phenomena occur together and therefore they are linked For example, the positive relationship of .71 can represent positive correlation between the statistics degrees and the science degrees. The student who has high degree in statistics has also high degree in science and vice versa.

The Pearson correlation coefficient is given by the following equation:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Where $\bar{x}$ is the mean of variable x values, and $\bar{y}$ is the mean of variable y values.

Example – Correlation of statistics and science tests

A study is conducted involving 10 students to investigate the association between statistics and science tests. The question arises here; is there a relationship between the degrees gained by the 10 students in statistics and science tests?

Table (2.1) Student degree in Statistic and science

Students	1	2	3	4	5	6	7	8	9	10
Statistics	20	23	8	29	14	12	11	20	17	18
Science	20	25	11	24	23	16	12	21	22	26

Notes: the marks out of 30

Suppose that (x) denotes for statistics degrees and (y) for science degree

Calculating the mean $(\bar{x}\ ,\ \bar{y})$;

$$\bar{x}=\frac{\sum x}{n}=\frac{173}{10}=17.3\ ,\ \bar{y}=\frac{\sum y}{n}=\frac{200}{10}=20$$

Where the mean of statistics degrees $\bar{x}$ = 17.3 and the mean of science degrees $\bar{y}$ = 20

Table (2.2) Calculating the equation parameters

Statistics	Science					
x	y	$x-\bar{x}$	$(x-\bar{x})^2$	$y-\bar{y}$	$(y-\bar{y})^2$	$(x-\bar{x})(y-\bar{y})$
20	20	2.7	7.29	0	0	0
23	25	5.7	32.49	5	25	28
8	11	-9.3	86.49	-9	81	83
29	24	11.7	136.89	4	16	46
14	23	-3.3	10.89	3	9	-9.9
12	16	-5.3	28.09	-4	16	21.2
11	12	-6.3	39.69	-8	64	50.4
21	21	3.7	13.69	1	1	3.7
17	22	-0.3	0.09	2	4	-0.6
18	26	0.7	0.49	6	36	4.2
173	200	0	356.1	0	252	228

$$\sum(x-\bar{x})^2=356.1\ ,\ \sum(y-\bar{y})^2=252\ ,$$
$$\sum(x-\bar{x})(y-\bar{y})=228$$

Calculating the Pearson correlation coefficient;

$$r=\frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2}\sqrt{\sum(y-\bar{y})^2}}=\frac{228}{\sqrt{356.1}\sqrt{252}}$$
$$=\frac{228}{(18.8706)(15.8745)}=\frac{228}{299.5614}=0.761$$

Other solution

Also; the Pearson correlation coefficient is given by the following equation:

$$r = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sqrt{\left(\sum x^2 - \frac{(\sum x)^2}{n}\right)\left(\sum y^2 - \frac{(\sum y)^2}{n}\right)}}$$

Table (2.3) Calculating the equation parameters

x	y	xy	x^2	y^2
20	20	400	400	400
23	25	575	529	625
8	11	88	64	121
29	24	696	841	576
14	23	322	196	529
12	16	192	144	256
11	12	132	121	144
21	21	441	441	441
17	22	374	289	484
18	26	468	324	676
173	200	3688	3349	4252

Required calculation
$\sum x = 173$, $\sum y = 200$ $\sum xy = 3688$ $\sum x^2 = 3349$ $\sum y^2 = 4252$

Calculating the Pearson correlation coefficient by substitute in the aforementioned equation;

$$\boldsymbol{r = \frac{3688 - \frac{(173)(200)}{10}}{\sqrt{\left(3349 - \frac{(173)^2}{n10}\right)\left(4252 - \frac{(200)^2}{10n}\right)}} = \frac{228}{\sqrt{(356.1)(252)}} = \frac{228}{299.5614} = 0.761}$$

Pearson Correlation coefficient r = 0.761 exactly the same output of the first equation.

The calculation shows a strong positive correlation (0.761) between the student's statistics and science degrees. This means that as degrees of statistics increases the degrees of science increase also. Generally the student who has a high degree in statistics has high degree in science and vice versa.

2.4. Partial Correlation

The Partial Correlations procedure computes partial correlation coefficients that describe the linear relationship between two variables while controlling for the effects of one or more

additional variables. Correlations are measures of linear association. Two variables can be perfectly related, but if the relationship is not linear, a correlation coefficient is not an appropriate statistic for measuring their association**.**

Partial correlation is the correlation between two variables after removing the effect of one or more additional variables. Suppose we want to find the correlation between y and x controlling by W . This is called the partial correlation and its symbol is $r_{YX.W}$. This command is specifically for the case of one additional variable. In this case, the partial correlation can be computed based on standard correlations between the three variables as follows:

$$r_{YX.W} = \frac{r_{XY} - r_{XW}r_{YW}}{\sqrt{(1 - r_{XW}^2)(1 - r_{YW}^2)}}$$

$r_{YX.W}$ Stands for the correlation between x and y controlling by W .

As with the standard correlation coefficient, a value of +1 indicates a perfect positive linear relationship, a value of -1 indicates a perfect negative linear relationship, and a value of 0 indicates no linear relationship. *For more information see unit 4 of this book.*

2.5. Correlation Coefficients Pearson, Kendall and Spearman

Correlation is a Bivariate analysis that measures the strengths of association between two variables. In statistics, the value of the correlation coefficient varies between +1 and -1. When the value of the correlation coefficient lies around ± 1, then it is said to be a perfect degree of association between the two variables. As the correlation coefficient value goes towards 0, the relationship between the two variables will be weaker. Usually, in statistics, we measure three types of correlations: Pearson correlation, Kendall rank correlation and Spearman correlation.

Pearson r correlation: Pearson correlation is widely used in statistics to measure the degree of the relationship between linear related variables. For example, in the stock market, if we want to measure how two commodities are related to each other, Pearson correlation is used to measure the degree of relationship between the two commodities. The following formula is used to calculate the Pearson correlation coefficient r: See Example

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Kendall's Tau rank correlation: Kendall rank correlation is a non-parametric test that measures the strength of dependence between two variables. If we consider two samples, x and y, where each sample size is n, we know that the total number of pairings with $x\ y$ is n (n-1)/2.

The following formula is used to calculate the value of Kendall rank correlation:

$$\tau = \frac{n_c - n_d}{\frac{1}{2}n(n-1)}$$

Where:

τ = Kendall rank correlation coefficient

n_c = number of concordant (Ordered in the same way).

n_d = Number of discordant (Ordered differently).

Kendall's Tau Basic Concepts

Definition 1: Let $x_1, \ldots, x_n$ be a sample for random variable x and let $y_1, \ldots, y_n$ be a sample for random variable y of the same size n. There are $C(n, 2)$ possible ways of selecting distinct pairs (x_i, y_i) and (x_j, y_j). For any such assignment of pairs, define each pair as concordant, discordant or neither as follows:

- Concordant © if ($x_i > x_j$ and $y_i > y_j$) or ($x_i < x_j$ and $y_i < y_j$)
- Discordant (D) if ($x_i > x_j$ and $y_i < y_j$) or ($x_i < x_j$ and $y_i > y_j$)
- Neither if $x_i = x_j$ or $y_i = y_j$ (i.e. ties are not counted).

Observation: To facilitate the calculation of $C - D$ it is best to first put all the x data elements in ascending order. If x and y are perfectly positively correlated, then all the values of y would be in ascending order too, and so if there are no ties then $C = C(n, 2)$ and $\tau = 1$.

Otherwise, there will be some inversions. For each i, count the number of $j > i$ for which $x_j < x_i$. This sum is D. If x and y are perfectly negatively correlated, then all the values of y would be in descending order, and so if there are no ties then $D = C(n, 2)$ and $\tau = -1$.

An example of calculating Kendall's Tau correlation

To calculate a ***Kendall's Tau correlation*** coefficient on same data without any ties we use the following data:

Students	1	2	3	4	5	6	7	8	9	10
Statistics	20	23	8	29	14	12	11	20	17	18
Science	20	25	11	24	23	16	12	21	22	26

Table (2.4) Set rank to the data

data				Arranged Rank	
statistics (degree)	science (degree)	Rank (statistics)	Rank (science)	Rank (science)	Rank (statistics)
20	20	4	7	1	5
23	25	2	2	2	2
8	11	10	10	3	1
29	24	1	3	4	7
14	23	7	4	5	6
12	16	8	8	6	3
11	12	9	9	7	4
21	21	3	6	8	8
17	22	6	5	9	9
18	26	5	1	10	10

Continued Table (2.4) Calculating the Number of Concordant C and Discordant (D)

D	C											
		1	--									
1		2	D	--								
2		3	D	D	--							
	3	4	C	C	C	--						
1	3	5	D	C	C	C	--					
3	2	6	D	D	C	C	D	--				
3	3	7	C	D	D	C	C	D	--			
	7	8	C	C	C	C	C	C	C	--		
	8	9	C	C	C	C	C	C	C	C	--	
	9	10	C	C	C	C	C	C	C	C	C	--
			1	2	3	4	5	6	7	8	9	10
10	35	Total of (D) and (C)										

Then substitute into the main equation

$$\tau = \frac{n_c - n_d}{\frac{1}{2}\, n(n-1)} \qquad \tau = \frac{35 - 10}{\frac{1}{2} * 10(10-1)}$$

$$\tau = \frac{25}{45} = 0.556$$

Kendall's Tau coefficient τ = 0.556; this indicates a moderate positive relationship between the ranks individuals obtained in the statistics and science exam. This means the higher you ranked in statistics, the higher you ranked in science also, and vice versa.

Calculating Kendall's Tau manually can be very tedious without a computer and is rarely done without a computer. Large dataset make it almost impossible to do manually by hand. . For more information see unit4 in this book

Spearman rank correlation: Spearman rank correlation is a non-parametric test that is used to measure the degree of association between two variables. It was developed by Spearman, thus it is called the Spearman rank correlation. Spearman rank correlation test does not assume any assumptions about the distribution of the data and is the appropriate correlation analysis when the variables are measured on a scale that is at least ordinal.

The following formula is used to calculate the Spearman rank correlation coefficient:

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$

Where:

ρ = Spearman rank correlation coefficient
di= the difference between the ranks of corresponding values Xi and Yi
n= number of value in each data set.

The Spearman correlation coefficient,ρ, can take values from +1 to -1. A ρ of +1 indicates a perfect association of ranks, a ρ of zero indicates no association between ranks and a ρ of -1 indicates a perfect negative association of ranks. The closer ρ to zero, the weaker the association between the ranks.

An example of calculating Spearman's correlation

To calculate a Spearman rank-order correlation coefficient on data without any ties use the following data:

Students	1	2	3	4	5	6	7	8	9	10
Statistics	20	23	8	29	14	12	11	20	17	18
Science	20	25	11	24	23	16	12	21	22	26

Table (2.5) Calculating the Parameters of Spearman rank Equation:

statistics (degree)	science (degree)	Rank (statistics)	Rank (science)	| d |	d^2
20	20	4	7	3	9
23	25	2	2	0	0
8	11	10	10	0	0
29	24	1	3	2	4
14	23	7	4	3	9
12	16	8	8	0	0
11	12	9	9	0	0
21	21	3	6	3	9
17	22	6	5	1	1
18	26	5	1	4	16

Where d = absolute difference between ranks and d^2 = difference squared.
Then calculate the following:

$$\sum d_i^2 = 9 + 0 + 0 + 4 + 9 + 0 + 0 + 9 + 1 + 16 = 48$$

Then substitute into the main equation as follows:

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)} \quad ; \quad \rho = 1 - \frac{6*48}{10(10^2-1)}$$

$$\rho = 1 - \frac{288}{990} \quad ; \quad \rho = 1 - 0.2909$$

$$\rho = 0.71$$

Hence, we have a ρ= 0.71; this indicates a strong positive relationship between the ranks individuals obtained in the statistics and science exam. This means the higher you ranked in statistics, the higher you ranked in science also, and vice versa.

So; the **Pearson r correlation** coefficient = 0.761 and ***Spearman's correlation*** = 0.71 for the same data which means that correlation coefficients for both techniques are approximately equal. For more information see unit4 in this book

2.6 Exercises

Study is conducted involving 14 infants to investigate the association between gestational age at birth, measured in weeks, and birth weight, measured in grams.

Table (2.6) Gestational age and their Weight at birth

Infant No.	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Gestational age	34.7	36	29.3	40.1	35.7	42.4	40.3	37.3	40.9	38.3	38.5	41.4	39.7	39.7
Birth Weight	1895	2030	1440	2835	3090	3827	3260	2690	3285	2920	3430	3657	3685	3345

Applying the proper method; Estimate the association between gestational age and infant birth weight.

(Guide values $r = 0.882, \rho = 0.779, \tau = 0.641$)