Correlation

Scatter diagrams

A scatter diagram is a graphical way of representing pairs of points. It is best illustrated by example.

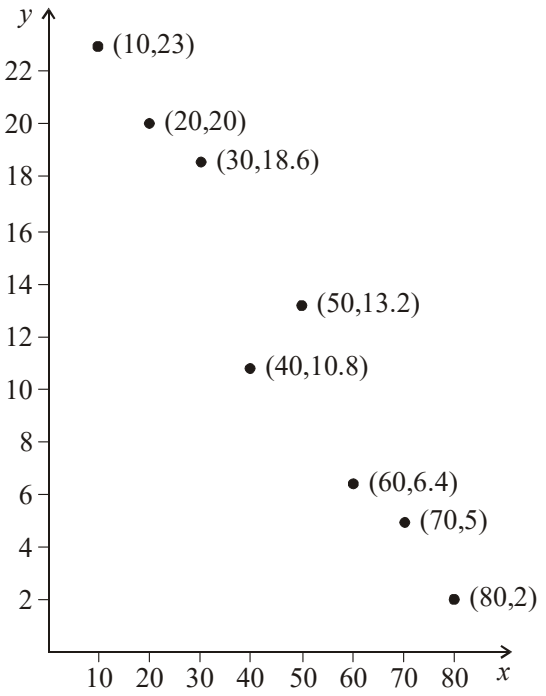
Example

The following pairs of points relating a variable x to a variable y have been observed.

x	10	20	30	40	50	60	70	80
y	23	20	18.6	10.8	13.2	6.4	5	2

By plotting a scatter diagram for the data suggest a possible relationship between x and y.

Solution



The scatter diagram above suggests a linear relationship with a negative slope.

Thus, scatter diagrams are an aid to the discovery of relationships between variables. If two variables are related in some systematic way, we say that they are correlated.

Tests of Correlation

When we are considering the question of a possible correlation between two observed physical quantities we require:

(1) a precise mathematical determination of the degree of correlation;
(2) a statement of the probability that the correlation could arise by chance.

Answers to these questions depend, however, on the type of data that the experiment or observation produces.

There can be three types of data:

Nominal Data

This is data in the form of frequencies fitting discrete, distinct categories. For example, we can count the number of boys and girls in a class. Each individual is either a boy or a girl and there is no sense in which the boys and the girls can be placed into a rank order.

Ordinal Data

Ordinal data are measures of physical quantities that can be ranked. For example, the variable X could measure the number of days individuals have been subject to a special diet; the variable Y could measure the position of those individuals in a race. Here, it is meaningful to ask how does the position of an individual, that is his rank, in terms of values of X correlate with his position, or rank, in terms of Y.

Interval Data

Data is said to be at interval level when there is a meaningful continuous scale of measurement such that equal differences between values in the scale genuinely correspond to real differences between the physical quantities that the scale measures. An example of a set of interval level data would be a collection of measurements of height. Here it is meaningful to say that the difference of height between a person who is $1.80m$ and one who is $1.70m$ tall is equal to the difference of height between a person

who is 1.90*m* and one who is 1.80*m* tall. Equal differences in the scale correspond to equal differences in the physical quantities they measure.

All interval level data can be placed in rank order; in other words, interval level data can be "reduced" to ordinal level data. Ordinal level data cannot necessarily be promoted to interval level data. Interval level data contain more information than ordinal level data.

We cannot correlate nominal level data.

Tests for ordinal level data differ from tests for nominal level data. For data at ordinal level, the appropriate test is Spearman's coefficient of rank correlation. For data at interval level, the appropriate test is Pearson's product moment correlation coefficient. Because interval data contains more information than ordinal data, the product moment correlation coefficient is said to be "more powerful" than Spearman's coefficient of rank correlation. This means that the product moment coefficient detects correlations that the rank coefficient cannot. However, in order to understand this idea we need to discuss in more detail why a statement of the probability that a given degree of correlation could arise by chance is necessary.

To understand this we must return to the ideas of population and sampling. When we make a set of observations of two physical quantities, X and Y, we are generally taking a sample of these quantities. We expect virtually all physically measurable quantities to show some variation. Consequently, there is always the possibility that observed values in the sample of the quantities of X and Y are due to chance. For example, suppose a bag contains 10 blue and 10 red balls, but when I sample 6 of these I draw only blue balls. On the strength of this sample I may conclude that all the balls in the bag are blue. I would, in fact, have made a mistake, even though my conclusion would have been reasonable! Likewise, let us imagine that in a population as a whole, there is no correlation – no systematic relation – between the two physical quantities measured by the variables X and Y. Nonetheless, my sample might, by chance alone, suggest that a correlation does exist.

Therefore, a correlation test requires a statement of the degree to which the observed correlation could arise by chance alone.

Such statements are made by comparing the calculated, "test" value for a given set of data with values corresponding to certain probabilities.

For ordinal data the appropriate test is Spearman's coefficient of rank correlation, and for interval level data the appropriate test is Pearson's product moment correlation coefficient.

Pearson's product moment correlation coefficient – testing correlation at interval level

Let X and Y be two variables at interval level. Then, the appropriate measure of correlation is Pearson's Product Moment Correlation Coefficient given by:

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

Example

The speed of a car, v metres per second, at time t after it starts to accelerate is shown in the table below for $0 \le t \le 8$.

t	0	1	2	3	4	5	6	7	8
v	0	2.8	6.7	10.2	12.8	16.2	19.6	21.5	22.9

Calculate the product moment correlation coefficient for these data.

Firstly, we must find $\sum t$, $\sum v$, $\sum tv$, $\sum t^2$, $\sum v^2$.

t	v	tv	t^2	v^2
0	0	0	0	0
1	2.8	2.8	1	7.84
2	6.7	13.4	4	44.89
3	10.2	30.6	9	104.04
4	12.8	51.2	16	163.84
5	16.2	81.0	25	262.44
6	19.6	117.6	36	384.16
7	21.5	150.5	49	462.25
8	22.9	183.2	64	524.41
Σt = 36	Σv = 112.7	Σtv = 630.3	Σt^2 = 204	Σv^2 = 1953.87

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

$$= \frac{9 \times 630.3 - 36 \times 112.7}{\sqrt{9x204 - 36^2}\sqrt{9 \times 1953.87 - 112.7^2}}$$

$$= 0.99 \text{ (2.S.F.)}$$

Spearman's Coefficient of Rank Correlation

Let X and Y be two variables at ordinal data level. Let rank X represent the order in which the values of X occur, and likewise rank Y represent the corresponding order in which values of Y occur. Each value of X is associated with a value of Y – they form pairs of values. Then, let

$$d = \text{rank}\, X - \text{rank}\, Y$$

Then, Spearman's rank correlation coefficient is:

$$r_s = 1 - \frac{6\sum d^2}{n\left(n^2 - 1\right)}$$

Where n is the number of pairs of values $\left(X_i, Y_i\right)$.

Example

Two independent examiners awarded the following marks to each of five candidates:

Candidate	A	B	C	D	E
1st Examiner	39	90	64	81	60
2nd Examiner	53	84	42	85	61

Calculate Spearman's rank correlation coefficient for these data.

Note: The data is at ordinal level because it is not meaningful to say that the difference between a score of 85 and a score of 80 is equivalent, for example, to the difference between 45 and 40.

Candidate	1st	2nd	Rank	Rank	d	d^2
A	39	53	5	4	1	1
B	90	84	1	2	-1	1
C	64	42	3	5	-2	4
D	81	85	2	1	1	1
E	60	61	4	3	1	1
						$\sum d^2 = 8$

Therefore,

$$r_s = 1 - \frac{6\sum d^2}{n\left(n^2 - 1\right)}$$

$$= 1 - \frac{6 \times 8}{5 \times 24}$$

$$= 0.6$$