Pearson's product moment correlation

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.

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 1.80*m* and one who is 1.70*m* 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.

To test for correlation between data at interval level, we use Pearson's product moment correlation coefficient.

Pearson's product moment correlation coefficient

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
$\Sigma t =$ 36	$\Sigma v =$ 112.7	$\Sigma tv =$ 630.3	$\Sigma t^2 =$ 204	$\Sigma v^2 =$ 1953.87

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

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

$$= 0.995 \text{ (3.S.F.)}$$

Covariance

There is another approach to finding the regression line based on the definition of *covariance*.

For *n* pairs of data $(x_1, y_1), (x_{2,} y_2) \ldots\ldots (x_n y_n)$ the covariance, S_{xy} is

$$S_{xy} = \frac{1}{n}\sum \left(x - \bar{x}\right)\left(y - \bar{y}\right)$$

Thus the variance of X can be written in terms of covariance thus

$$S_x^2 = S_{xx} = \frac{1}{n}\sum\left(x-\bar{x}\right)\left(x-\bar{x}\right) = \frac{1}{n}\sum(x-\bar{x})^2$$

and the variance of Y is

$$S_{y^2} = S_{yy} = \frac{1}{n}\sum\left(y-\bar{y}\right)\left(y-\bar{y}\right) = \frac{1}{n}\sum(y-\bar{y})^2$$

So the covariance is an extension to pairs of data of the concept of variance.

We can show that

$$S_x^2 = \frac{\sum x^2}{n} - \bar{x}^2$$
$$S_y{}^2 = \frac{\sum y^2}{n} - \bar{y}^2$$

Similarly, we can show

$$S_{xy} = \frac{\sum xy}{n} - \bar{x}\bar{y}$$

Using the covariance formulae, the product-moment correlation coefficient can be written

$$r = \frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}} = \frac{S_{xy}}{S_x S_y}$$

Example

Recalculate the product moment correlation coefficient for these data using the covariance formula

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

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

$$\overline{t} = \frac{\sum t}{n} = \frac{36}{9} = 4$$

$$\overline{v} = \frac{\sum v}{n} = \frac{112.7}{9} = 12.522...$$

$$S_{tt} = \frac{\sum t^2}{n} - \overline{t}^2$$

$$= \frac{204}{9} - 4^2$$

$$= 6.666...$$

$$S_{vv} = \frac{\sum v^2}{n} - \overline{v}^2$$

$$= \frac{1953.87}{9} - \left(12.5222...\right)^2$$

$$= 60.2906...$$

$$S_{tv} = \frac{\sum tv}{n} - \overline{t}\,\overline{v}$$

$$= \frac{630.3}{9} - 4 \times 12.5222...$$

$$= 19.9444...$$

$$r = \frac{S_{xy}}{S_x S_y}$$

$$= \frac{19.9444..}{\sqrt{6.666...} \times \sqrt{60.2906...}}$$

$$= 0.9948...$$

$$= 0.995\left(3.S.F\right)$$

Actually, there are several ways to calculate the values of S_{xx}, S_{yy} and S_{xy} and for completeness we should mention at least one other, because you meet it in other textbooks. Thus,

$$S_{xx} = S_x^2 = \frac{\sum x^2}{n} - \bar{x}^2 = \sum x^2 - \frac{\left(\sum x\right)^2}{n}$$

$$S_{yy} = S_y^2 = \frac{\sum y^2}{n} - \bar{y}^2 = \sum y^2 - \frac{\left(\sum y\right)^2}{n}$$

and $$S_{xy} = \frac{\sum xy}{n} - \bar{x}\bar{y} = \sum (xy) - \frac{\left(\sum x\right) \times \left(\sum y\right)}{n}$$

And however you calculated it, using these symbols, the product moment correlation coefficient is

$$r = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}}$$

Hypothesis testing of Pearson's Product moment correlation coefficient

The value of the correlation coefficient *suggests* whether there is a linear relationship or not. A high value of *r* indicates a strong positive relationship. However, there is the question of just how *likely* a particular value for the product moment correlation coefficient could be.

The apparent correlation between two variables *X* and *Y* could be due to chance factors alone. The question must be, what is the probability that *X* and *Y* could appear to be correlated with a coefficient of correlation *r* due to chance factors alone?

This leads to hypothesis testing of the product moment correlation coefficient.

We seek to test the hypothesis that there is a correlation, that is $r \neq 0$ against the null hypothesis that there is no correlation, $r = 0$.

Such a test would be formulated as

$H_0 \quad r = 0$

$H_1 \quad r \neq 0$

This is called a *two-tailed test* because the correlation could be either positive, $r > 0$, or negative, $r < 0$. On the other hand, we might believe in advance of collecting the data that there was either a positive or a negative correlation, in which case we would be testing, in the case of a positive correlation

$H_0 \quad r = 0$

$H_1 \quad r > 0$

or in the case of a negative correlation

$H_0 \quad r = 0$

$H_1 \quad r < 0$

In both these cases, the test is a *one-tailed test* because we are expecting either a positive correlation *or* a negative correlation, but not both.

In order to make this test we have to assume that the two variables are jointly normally distributed. However, at this level you may assume that the variables are jointly normally distributed.

Hypothesis tests are made at a *significance* level. This indicates the degree of chance that you are willing to accept and yet allow the data to pass the hypothesis test. For example, if the significance level is 5% then you are allowing that on 5% of the occasions the data would appear to be correlated *by chance.* This corresponds to a probability of 1 in 20. In other words, you are accepting that there is a correlation, even though there is a 1 in 20 chance that the same degree of correlation would appear merely due to chance alone.

In order to perform the test you require a set of tables of *critical values* for the Pearson's product moment correlation coefficient. These tables give you the critical values of the correlation coefficient for a given number of data and a given significance level. We will illustrate this by means of a complete worked example.

Example

Explain briefly your understanding of the term ‘correlation’. Describe how you used, or could have used, correlation in a project or in class work. Twelve students sat two Biology tests, one theoretical and one practical. Their marks are shown below.

Marks in theoretical test (t)	4	9	24	7	11	15	12	8	6	17	16	12
Marks in practical test (p)	6	8	27	14	8	17	12	10	8	15	22	18

(a) Draw a scatter diagram to represent these data.

(b) Find, to 4 decimal places:

(i) the value of S_{tp}.

(ii) the product-moment correlation coefficient.

(c) Use a 0.025 significance level and a suitable test to check the assertion that students who do well in theoretical Biology tests also do well in practical Biology tests.

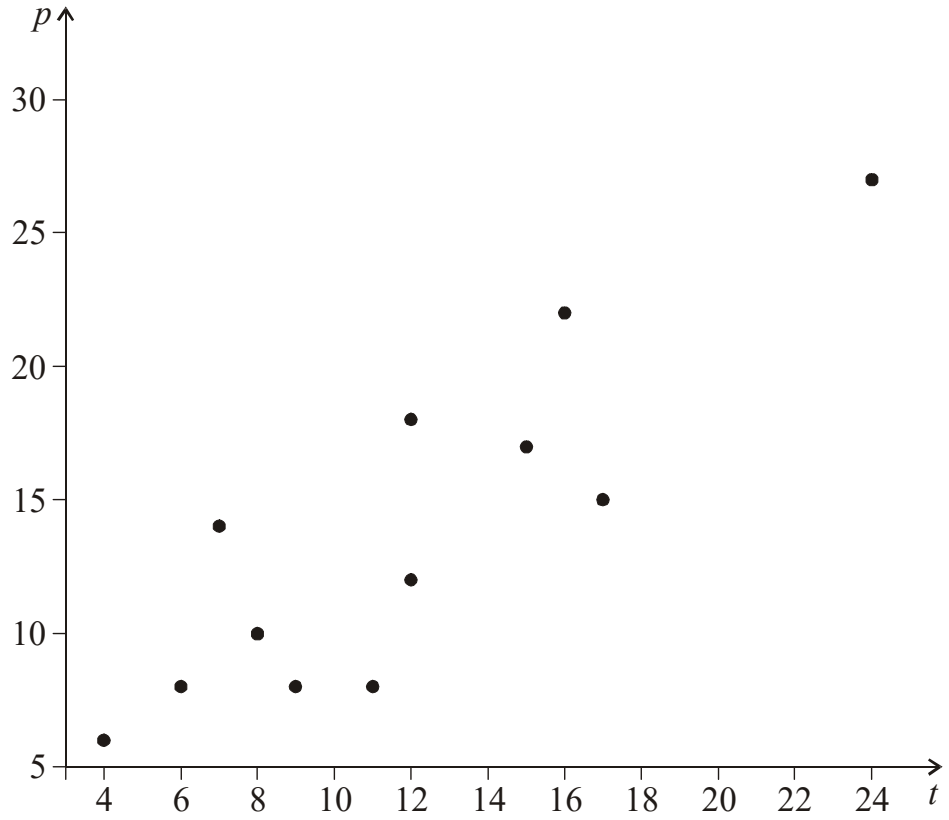
(d) Comment on whether the product-moment correlation coefficient is appropriate in this case.

Solution

Correlation in general is a measure of statistical dependence between variables. However, the term is often used in a more limited sense to indicate a linear relationship between two variables (product-moment correlation coefficient), or agreement between two sets of rank (rank correlation coefficient). Variables that show a close relationship in one of these senses are said to be highly correlated.

In a project or in class work it can be used to describe the association between different things, e.g. the marks of English and Biology tests, heights and weights of pupils, the opinion of boys and girls about the subjects they are taught, i.e. which is their favourite, second favourite and so on (rank correlation).

(a)



(b) The sample size is $n = 12$.

$$\sum t = 141, \quad \sum p165, \quad \sum t^2 = 2001, \quad \sum p^2 = 2719, \quad \sum (t \cdot p) = 2280$$

(i) $$S_{t \cdot p} = \sum (tp) - \frac{\left(\sum t\right)\left(\sum p\right)}{n} = 2280 - \frac{141 \times 165}{12} = 341.25.$$

(ii) $$S_{t \cdot t} = \sum t^2 - \frac{\left(\sum t\right)^2}{n} = 2001 - \frac{141^2}{12} = 344.25$$

$$S_{p \cdot p} = \sum p^2 - \frac{\left(\sum p\right)^2}{n} = 2719 - \frac{165^2}{12} = 450.25$$

$$\text{Therefore, } r_{test} = \frac{S_{tp}}{\sqrt{S_{tt} \times S_{pp}}} = \frac{341.25 \times 25}{\sqrt{344.25 \times 450.25}} = 0.8668$$

(c) We are testing the hypothesis that there is a positive correlation $r > 0$ against the null hypothesis that there is no correlation $r = 0$

$H_0 : \quad r = 0$
$H_1 : \quad r > 0$

The significance level is $\alpha = 0.025$ and this is one-tailed test. The sample size is $n = 12$. From tables the critical value $(\alpha = 0.025,\ n = 12,$ one-tailed test$)$ is

$r_{\text{critical}} = 0.5760$

The value of the test statistic is

$r_{test} = 0.8668 > 0.5760 = r_{critical}$

Since the test value is greater than the critical value, reject H_0 at 2.5% level of significance and accept H_1. There is evidence to support the assertion.

(d) Strictly speaking the data is not true interval level data. Therefore, the scores of the students in the tests should be ranked and a test suitable for ranked data should be used – that is, Spearman's rank correlation coefficient.

Coding

Calculations of the product moment correlation coefficient can also be simplified by coding. The final value of r is not affected if from every x {or y value} the same number is subtracted, or if each value is divided or multiplied by the same number.

For example, suppose we have the raw data

x	y
920	32
1070	40
1150	49
1230	51
1290	53

Consider the transformation

$$X = \frac{x - 1150}{10}$$

$$Y = y - 49$$

That is, we obtain new data relating to variables X and Y derived from the raw data by subtracting 1150 from x and dividing that result by 10; and by subtracting 49 from y.

Thus we obtain a simpler table of results

x	y	$X = \frac{x - 1150}{10}$	$Y = y - 49$
920	32	-23	-17
1070	40	-8	-9
1150	49	0	0
1230	51	8	2
1290	53	14	4

Using this coded data

X	Y	X^2	Y^2	XY
-23	-17	529	289	391
-8	-9	64	81	72
0	0	0	0	0
8	2	64	4	16
14	4	196	16	56
$\Sigma X = -9$	$\Sigma Y = -20$	$\Sigma X^2 = 853$	$\Sigma Y^2 = 390$	$\Sigma XY = 535$

Now we can substitute into the formula

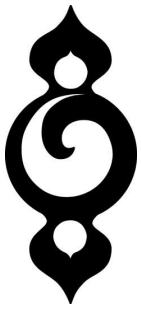
$$r = \frac{\Sigma x_i y_i - \frac{\Sigma x_i \Sigma y_i}{n}}{\sqrt{\left\{\Sigma x_i^{\,2} - \frac{\{\Sigma x_i\}^2}{n}\right\}\left\{\Sigma y_i^{\,2} - \frac{\{\Sigma y_i\}^2}{n}\right\}}}$$

$$= \frac{535 - \frac{\{-9\}\,\mathrm{x}\,\{-20\}}{5}}{\sqrt{\left\{853 - \frac{\{-9\}^2}{5}\right\}\left\{390 - \frac{\{-20\}^2}{5}\right\}}}$$

$$= \frac{499}{509.32}$$

$$= 0.980 (3.s.f.)$$

Black's Academy



Visit Black's Academy

We hope that you have found this article helpful. For more information about our database visit Black's Academy on-line at

http://www.blacksacademy.com

Contact us also at support@blacksacademy.com and tells us more about what you are studying and how you think we might be able to help you.