Correlation and the Pearson's r correlation coefficient

CHARLES STURT UNIVERSITY

Learning Skills

Introduction:

A correlation is a measure of the strength of the linear relationship between two measurable variables. The Pearson correlation coefficient, represented as r, gives the strength and direction of this relationship. The closer r is to 1 or to -1 then the stronger the linear relationship between the two variables.

This sheet will teach you to:

- Calculate a range of required sums
- Introduce two alternate Pearson correlation coefficient formulae
- Calculate the Pearson correlation coefficient for a set of data

1. Background information

To demonstrate the concept of a correlation two simple sets of data, representing the scores for variable X and Y will be used.

X = 2, 6, 9, 10

Y = 5, 6, 4, 12

Both the x and y scores are increasing (with one exception in the y set). This means variable X and Y are positively correlated. The correlation coefficient r when calculated will lie between 0 and 1.

If the x scores were increasing and the y scores decreasing then there would be a negative relationship and the r would lie between 0 and -1.

2. Calculating the required sums

To calculate Pearson's correlation coefficient five sums need to be found: the sums of X , Y , X^2 , Y^2 and XY.

The sums are best worked out by using a table format and totalling each column. The symbol Σ is used to represent each sum or total.

- The X and Y scores are given
- X squared is found by squaring each X score
- Y squared is found by squaring each Y score
- XY is found by multiplying each X score by the pairing Y score

X	Y	X^2	Y^2	XY
2	5	2x2=4	5x5=25	2x5=10
6	6	6x6=36	6x6=36	6x6=36
9	4	9x9=81	4x4=16	9x4=36
10	11	10x10=100	11x11=121	10x11=110
$\sum X$ =27	$\sum Y$ = 26	$\sum X^2$ =221	$\sum Y^2$ =198	$\sum XY$ =192

The sums required are:

$\sum X = 27$ so $(\sum X)^2 = 27^2 = 729$

$\sum Y = 26$ so $(\sum Y)^2 = 26^2 = 676$

$\sum X^2 = 221$ $\sum Y^2 = 198$ and $\sum XY = 192$

The number of pairs (N) is also needed : $N = 4$

3. Introducing the alternate formulae

Alternate 1

$$r = \frac{SP}{\sqrt{(SS_x \times SS_y)}}$$

where

- $SP = \sum XY - \frac{\sum X \times \sum Y}{N}$
- $SS_x = \sum X^2 - \frac{(\sum X)^2}{N}$
- $SS_y = \sum Y^2 - \frac{(\sum Y)^2}{N}$

Alternate 2

$$r = \frac{N \times \sum XY - \sum X \times \sum Y}{\sqrt{[N \times \sum X^2 - (\sum X)^2] \times [N \times \sum Y^2 - (\sum Y)^2]}}$$

4. Calculating the Pearson correlation coefficient using the first formula

$$SP = \sum XY - \frac{\sum X \times \sum Y}{N} = 192 - \frac{27 \times 26}{4} = 192 - 175.5 = 16.5$$

$$SS_x = \sum X^2 - \frac{(\sum X)^2}{N} = 221 - \frac{729}{4} = 221 - 182.25 = 38.75$$

$$SS_y = \sum Y^2 - \frac{(\sum Y)^2}{N} = 198 - \frac{676}{4} = 198 - 169 = 29$$

so

$$r = \frac{SP}{\sqrt{(SS_x \times SS_y)}} = \frac{16.5}{\sqrt{(38.75 \times 29)}} = \frac{16.5}{\sqrt{1123.75}} = \frac{16.5}{33.5224} = 0.49$$

5. Calculating the Pearson correlation coefficient using the second formula

$$r = \frac{N \times \sum XY - \sum X \times \sum Y}{\sqrt{[N \times \sum X^2 - (\sum X)^2] \times [N \times \sum Y^2 - (\sum Y)^2]}}$$

$$r = \frac{4 \times 192 - 27 \times 26}{\sqrt{[4 \times 221 - 729] \times [4 \times 198 - 676]}}$$

$$r = \frac{768 - 702}{\sqrt{[884 - 729] \times [792 - 676]}}$$

$$r = \frac{66}{\sqrt{155 \times 116}} = \frac{66}{\sqrt{17980}} = \frac{66}{134.09} = 0.49$$

Work out the top and bottom lines separately

Work out the bottom line bit by bit:

Work out each bracket by multiplying before subtracting

After multiplying, check that the first number in both brackets is larger than the second. If this is not the case you may need to check your column values and totals.

You can see that either formula leads to the same answer.

6. For more information

Visit our Learning Skills website at http://www.csu.edu.au/division/studserv/maths/index.htm