

## community project

encouraging academics to share statistics support resources

All stcp resources are released under a Creative Commons licence

stcp-gilchristsamuels-3

The following resources are associated:

Simple Linear Regression worksheet

### Pearson Correlation

**Research question type:** Relationship between two variables

**What kind of variables:** Continuous (scale/interval/ratio)

**Common Applications:** Exploring the (linear) relationship between two variables; e.g. as variable X increases does variable Y increase or decrease? The relationship is measured by correlation.

### Definition

Pearson correlation measures the existence (given by a p-value) and strength (given by the coefficient  $r$  between -1 and +1) of a linear relationship between two variables. It should only be used when its underlying assumptions are satisfied. If the outcome is significant we conclude that a correlation exists.

According to Cohen (1988) an absolute value of  $r$  of **0.1** is classified as **small**, an absolute value of **0.3** is classified as **medium** and of **0.5** is classified as **large**.

### Example: Dietetics

A dietetics student wanted to look at the relationship between calcium intake and knowledge about calcium in sports science students. Table 1 below shows the data she collected.

**Research question:** Is there a relationship between calcium intake and knowledge about calcium in sports science students?

### Hypotheses:

$H_0$ : There is **no correlation** between calcium intake and knowledge about calcium in sports science students (equivalent to saying  $r = 0$ )

$H_1$ : There **is a correlation** between calcium intake and knowledge about calcium in sports science students (equivalent to saying  $r \neq 0$ ),



**Table 1: Dietetics study data**

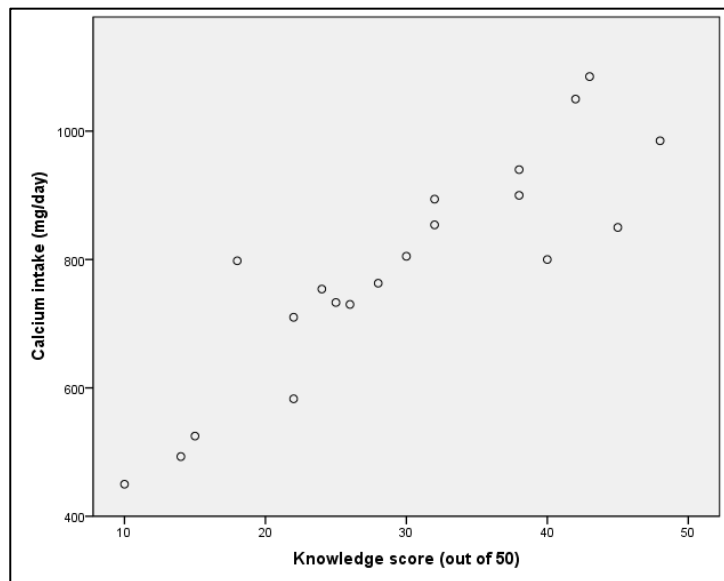
Respondent number	Knowledge score (Out of 50)	Calcium intake (mg/day)	Respondent number	Knowledge score (Out of 50)	Calcium intake (mg/day)
1	10	450	11	38	940
2	42	1050	12	25	733
3	38	900	13	48	985
4	15	525	14	28	763
5	22	710	15	22	583
6	32	854	16	45	850
7	40	800	17	18	798
8	14	493	18	24	754
9	26	730	19	30	805
10	32	894	20	43	1085

### Steps in SPSS

#### Step 1: Draw a scatterplot of the data to see any underlying trend in the relationship

Scatterplots can be drawn either in Excel or in SPSS. The scatterplot below was created using the Graphs - Chart Builder option in SPSS:

- Choose Scatter/Dot
- Drag the first Scatter plot into the plotting region
- Drag the two variables onto the axes as shown
- Click OK



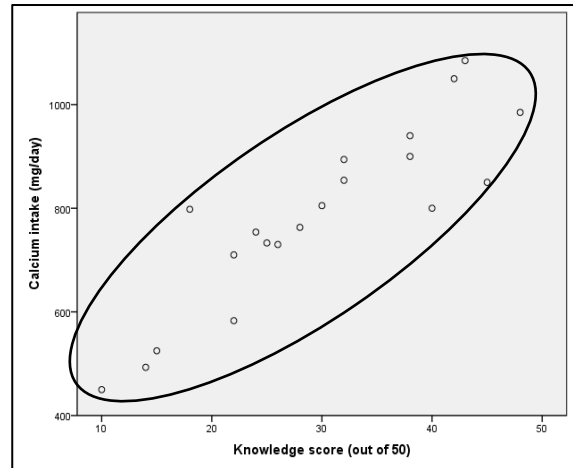
In this example there is perhaps an underlying assumption that calcium intake quantity is in response to the amount of knowledge about calcium. Therefore the intake variable has been placed on the vertical (Y) axis and the knowledge variable has been placed on the horizontal (X) axis.

**Firstly** we observe that the circles on the scatterplot are **reasonably closely scattered about an underlying straight line** (as opposed to a curve or a random scattering), so we say there is a **linear relationship** between the two variables. The scatterplot implies that as the knowledge score increases the calcium intake increases so we would expect the Pearson correlation coefficient to be **positive**. The exact size of the coefficient is a measure of the strength of the correlation (with 1 being a perfect positive correlation).



An example of negative correlation is the amount spent on heating and daily temperature: as the temperature increases the amount spent on heating decreases (although we would again need to inspect a scatter plot to determine whether this relationship is linear).

**Secondly**, we observe from the scatterplot that the values of the calcium variable do not vary more from this linear model as the values of the knowledge variable change. This means that most of the points lie within an **ellipse** or **cigar** shape orientated in the direction of the linear model (see diagram on right).

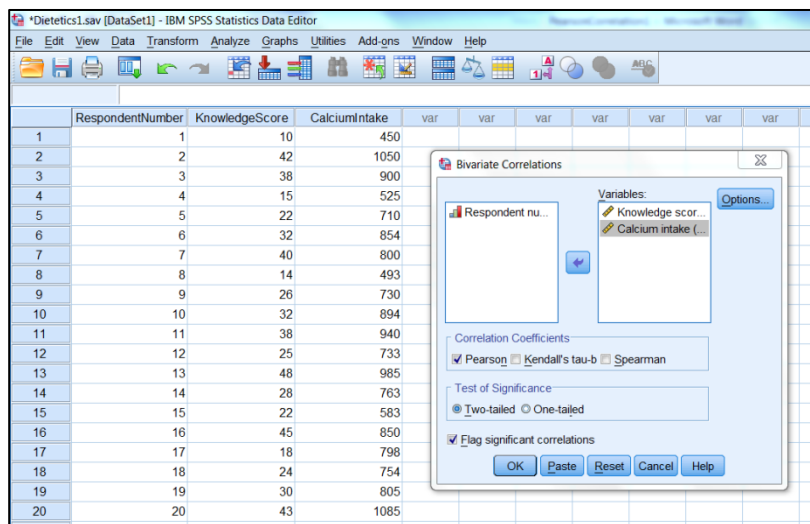


If no underlying straight line and cigar-shaped scattering can be perceived then there is **no point** going on to the next calculation.

**Step 2: Calculate the correlation coefficient**

With the data in the Data Editor, choose Analyze – Correlate – Bivariate...

- Select the variables to be correlated – in this case Calcium intake and Knowledge score – into the Variable list
- Ensure the Pearson Correlation Coefficients box is ticked (the default option)
- Click OK



The output should then look like this:

		Knowledge score (out of 50)	Calcium intake (mg/day)
Knowledge score (out of 50)	Pearson Correlation	1	.882 <sup>**</sup>
	Sig. (2-tailed)		.000
	N	20	20
Calcium intake (mg/day)	Pearson Correlation	.882 <sup>**</sup>	1
	Sig. (2-tailed)	.000	
	N	20	20

\*\* . Correlation is significant at the 0.01 level (2-tailed).

Pearson's correlation coefficient, *r*

p-value

Number of pairs of readings

**Note:** The information is given twice.



## Results

From the table above we observe that the correlation coefficient ( $r$ ) is 0.882, which is interpreted as a large coefficient. The p-value is less than 0.001 (SPSS outputs 0.000 as it has been rounded to 3 decimal places, but you should never write this) indicating a highly significant result so we reject  $H_0$  in favour of  $H_1$ .

## Interpretation

For sports science students there is very strong evidence that knowledge about calcium is linearly related to calcium intake. In particular we conclude that the more a sports science student knows about calcium, the greater their calcium intake is ( $r = 0.88$ ,  $p < 0.001$ ). The Pearson correlation coefficient is large.

### Note:

1. The p-value for a Pearson correlation test and the Pearson correlation coefficient are not the same thing. The larger the sample size, the lower the value of  $r$  at which a significant result occurs. For small samples it is possible to have a high correlation coefficient which is not significant and for large samples it is possible to have a small correlation coefficient which is significant. Thus it is important to look at the value of  $r$  **as well as the p-value**.
2. We **cannot** conclude that knowledge about calcium **causes** an increase in calcium intake. Perhaps a third (mediating) variable is involved? **Causality can only be established by a randomised control trial.**

## Comments

- Conclusions are only valid within the range of data collected.
- Pearson correlation also assumes the data values are **independent**. If the assumptions of Pearson correlation are not met or the data is ordinal other coefficients can be calculated:
  - **Kendall's  $\tau$**  ('tau') measures the degree to which a relationship is always positive or always negative
  - **Spearman's coefficient of rank correlation,  $\rho$**  ('rho'), behaves in a similar way to Kendall's  $\tau$  but has a less direct interpretation

## Reference

Cohen, J. (1988) *Statistical Power Analysis for the Behavioral Sciences*, 2<sup>nd</sup> ed. Hillsdale, NJ: Erlbaum.

