Statistical Analysis 2: Pearson Correlation

Research question type: Relationship between 2 variables

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

Common Applications: Exploring the relationship (linear) between 2 variables; eg, as variable A increases, does variable B increase or decrease? The relationship is measured by a quantity called **correlation**

Example 1:

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

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

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

Hypotheses:

The 'null hypothesis' might be:

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

And an 'alternative hypothesis' might be:

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

Data can be found in W:\EC\STUDENT\ MATHS SUPPORT CENTRE STATS WORKSHEETS\calcium.sav

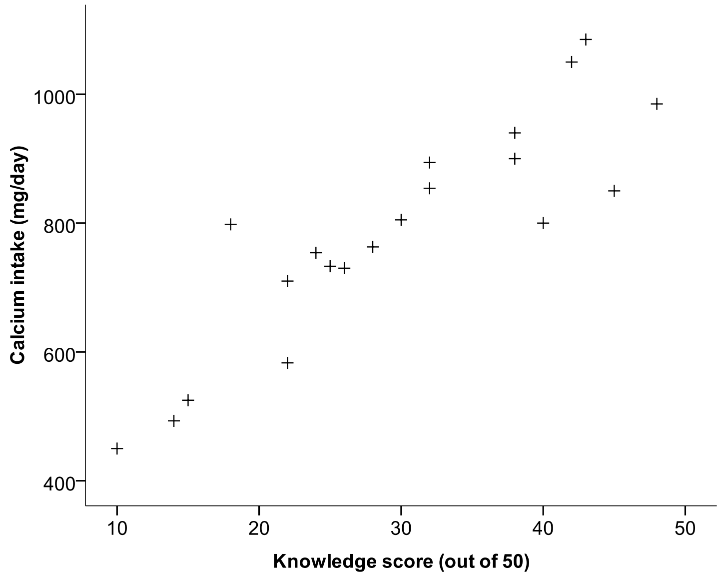
Steps in SPSS (PASW):

Step 1: Draw a **scatter plot** of the data to see any underlying trend in the relationship:

A scatter plot can be drawn in MS Excel or in SPSS, as right, using the **Graphs> Chart Builder** options

- choose Scatter/Dot
- drag the Simple Scatter plot into the plotting region
- drag the required variables into the two axes boxes
- click OK

[Note that the chart has been edited in the Chart Editor].



In this example there is perhaps an underlying assumption that 'calcium intake' quantity is in response to the amount of 'knowledge'.

It can be perceived from the scatter plot that the points are reasonably closely scattered about an underlying straight line (as opposed to a curve or nothing), so we say there is a **strong linear relationship** between the two variables. The scatter plot implies that as the knowledge score increases so the calcium intake increases. This shows a **positive linear relationship**. **Pearson's coefficient of linear correlation** is a **measure** of this strength.

Pearson's correlation coefficient can be positive or negative; the above example illustrates positive correlation – one variable increases as the other increases. An example of negative correlation would be the amount spent on gas and daily temperature, where the value of one variable increases as the other decreases.

Pearson's correlation coefficient has a value between -1 (**perfect negative correlation**) and 1 (**perfect positive correlation**).

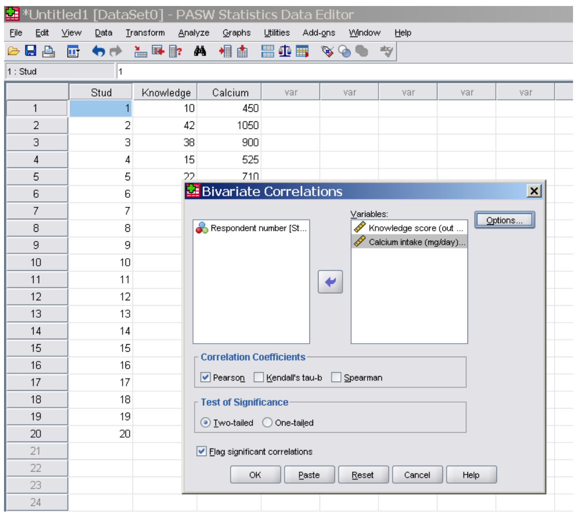
If no underlying straight line can be perceived, there is **no point** going on to the next calculation.

Step 2: Calculating the correlation coefficient

With the data in the Data Editor, choose

Analyze > Correlate > Bivariate…

- Select the 2 variables to be correlated – in this case calcium intake and knowledge score – into the Variable list
- Ensure the Pearson Correlation Coefficients box is ticked
- Click OK



Output should look something like:

Correlations

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

Pearson's correlation coefficient, r

***p*-value**

number of pairs of readings

NB The information is given twice.

Results:
From the Correlations table, it can be seen that the correlation coefficient (r) equals 0.882, indicating a strong relationship, as surmised earlier. $p < 0.001$ [**NEVER write p = 0.000]** and indicates that the coefficient is significantly different from 0.

Conclusion:
We can conclude that for sports science students there is evidence that knowledge about calcium is related to calcium intake. In particular, it seems that the more a sports science student knows about calcium, the greater their calcium intake is ($r = 0.88$, $p <0.001$).

Note:
We CANNOT readily assume that knowledge about calcium CAUSES an increase in calcium intake

Comments:

- Validity of Pearson correlation calculations are based on several assumptions:
    - data is at continuous (scale/interval/ratio) level
    - data values are independent of each other; ie, only one pair of readings per participant is used
    - a linear relationship is assumed when calculating Pearson's coefficient of correlation
    - observations are random samples from normal or symmetric distributions

- Other coefficients can be calculated for data at ordinal level of measurement:
    - Kendall's τ ('tau') measures the degree to which a relationship is always positive or always negative
    - Spearman's coefficient of rank correlation, ρ ('rho') behaves in a similar way to Kendall's τ, but has less direct interpretation

- A relationship between two variables does not necessarily imply causation. Could a third variable be involved?

- As sample size increases, so the value of r at which a significant result occurs, decreases. So it is important to look at the size of r, rather than the p-value. A value of r below 0.5 is 'weak'

- Conclusions are only valid within the range of data collected.

Example 2:

A correlation coefficient of 0.79 ($p < 0.001$) was calculated for 18 data pairs plotted in the scatter graph in figure A, right.

A Pearson correlation coefficient of 0.53 ($p = 0.005$) was calculated for the 27 data pairs plotted in the scatter graph in figure B below.

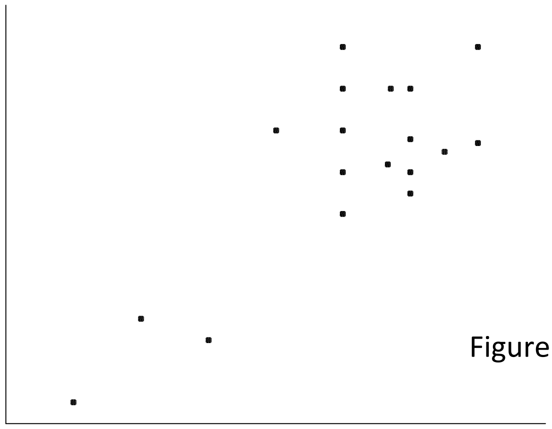
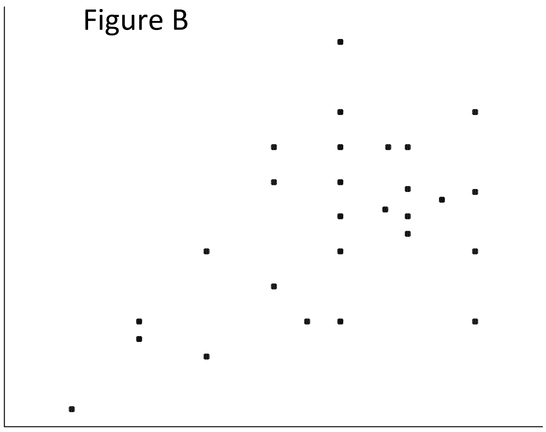


Figure A

Figure B



Comment on the pattern of dots and these results. Would you have calculated correlation coefficients for A and B?

See below for some suggestions.

Example 3:

Data were collected from a group of students to investigate the relationship between their shoe size (European) and their forearm length (cm). Using the data provided in W:\EC\STUDENT\ MATHS SUPPORT CENTRE STATS WORKSHEETS\shoe.sav explore this relationship.
Note that there are some missing values coded 888, and some anomalous data readings.

Example 4:

In the above data set would it be sensible to calculate a Pearson correlation coefficient for age and shoe size?

Suggested Answers

Example 2: No – neither chart shows an underlying straight line!
A: cover-up the three points in the bottom left – what do you see?
B: the points 'fan-out' as values increase – ie showing greater variability for larger values

Example 3: Assuming the data is at the appropriate level, a scatterplot shows an underlying straight line, although the points are widely spread out. Using all the data as it's given, $r=0.338$, $p<0.001$. Do all the readings make sense? May some students have given their forearm length in inches rather than cm? May some of them have 'guessed' their forearm length?

Example 4: Not for adults, but perhaps for growing children?