Correlations in SPSS (Practical)

The development of this E-Book has been supported by the British Academy.

Correlation practical

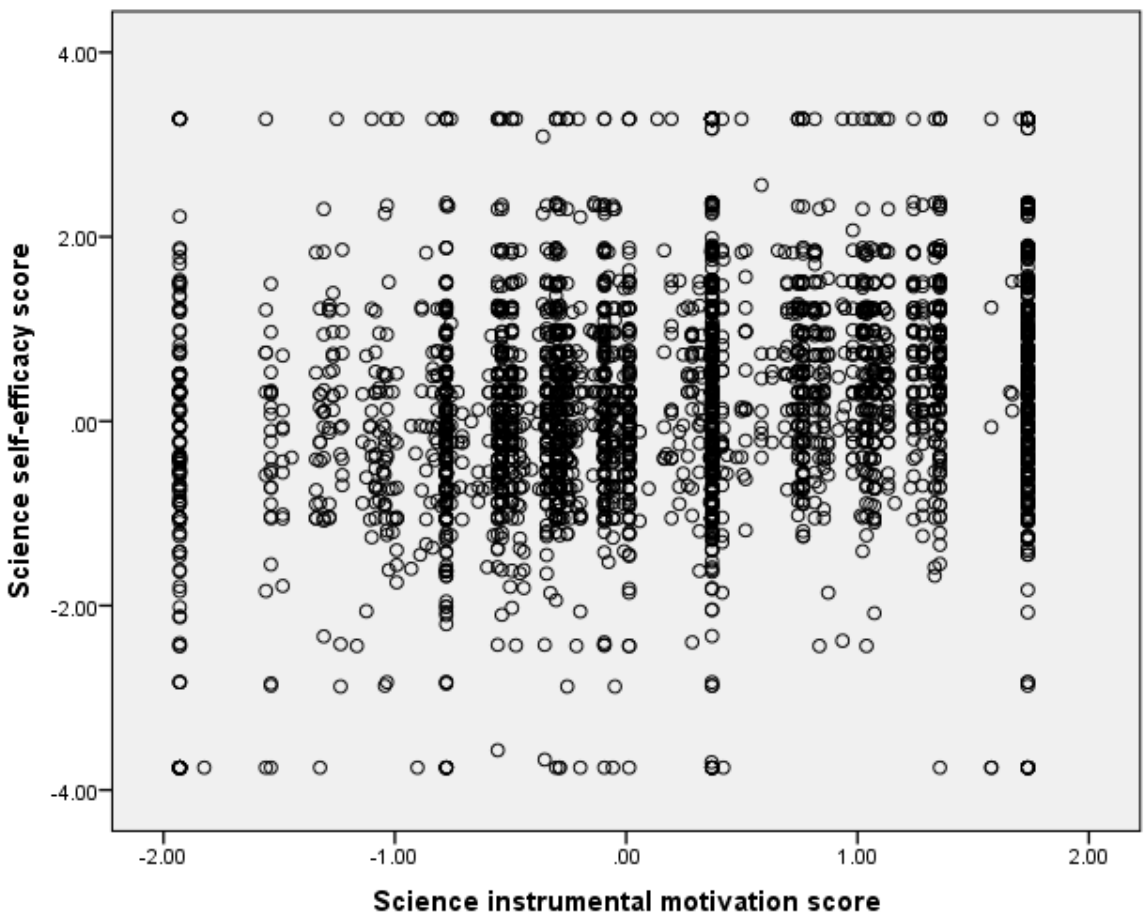
In this practical we will investigate whether there is a relationship between two variables by looking how correlated they are.

Two of the key predictors of academic achievement proposed by theories of student learning are self-efficacy and instrumental motivation. Self-efficacy refers to an individual's confidence in their ability to perform a task well, and instrumental motivation to the belief that learning will be useful for one's later career. In this practical we address the question of whether these two constructs are correlated, or specifically whether students who are more confident in their abilities in science are systematically more (or less) likely to view science learning as important for their future prospects. The PISA measure of science self-efficacy (SCIEEFF) was derived from students' responses to questions on how easy they would find it to perform eight science tasks on their own, such as "Identify the better of two explanations for the formation of acid rain". The measure of instrumental motivation (INTMOVSCI) was derived from four items in which students rated their agreement with statements like "Many things I learn in my school science subjects will help me to get a job" (see PISA datafile description for further details).

To do this we will begin by simply plotting the two variables in SPSS:

- Select **Scatter/Dot** from the **Legacy diagnostics** available from the **Graphs** menu.
- Select Simple Scatter and click on Define to bring up the Simple Scatterplot window.
- Copy the **Science self-efficacy score[SCIEEFF]** variable into the **Y Axis** box.
- Copy the **Science instrumental motivation score[INSMOVSCI]** variable into the **X Axis** box.
- Click on the **OK** button.

SPSS will then draw a scatterplot of the two variables which can be seen below:



Looking at the scatterplot there appears to be a positive correlation between the variables with larger values of **SCIEEFF** associated with larger values of **INSMOVSCI** (an upward sloping relationship) but this relationship is not that strong with possibly a few more points in the bottom-left and top-right quarters of the plot.

We want to test whether any correlation we observe in the scatterplot is significant but there are several different correlation coefficients for different situations. The first correlation coefficient that we will look at is the Pearson correlation coefficient. This correlation requires the variables to be continuous and, in smaller samples, to be normally distributed so we will firstly look at whether a normal distribution is suitable.

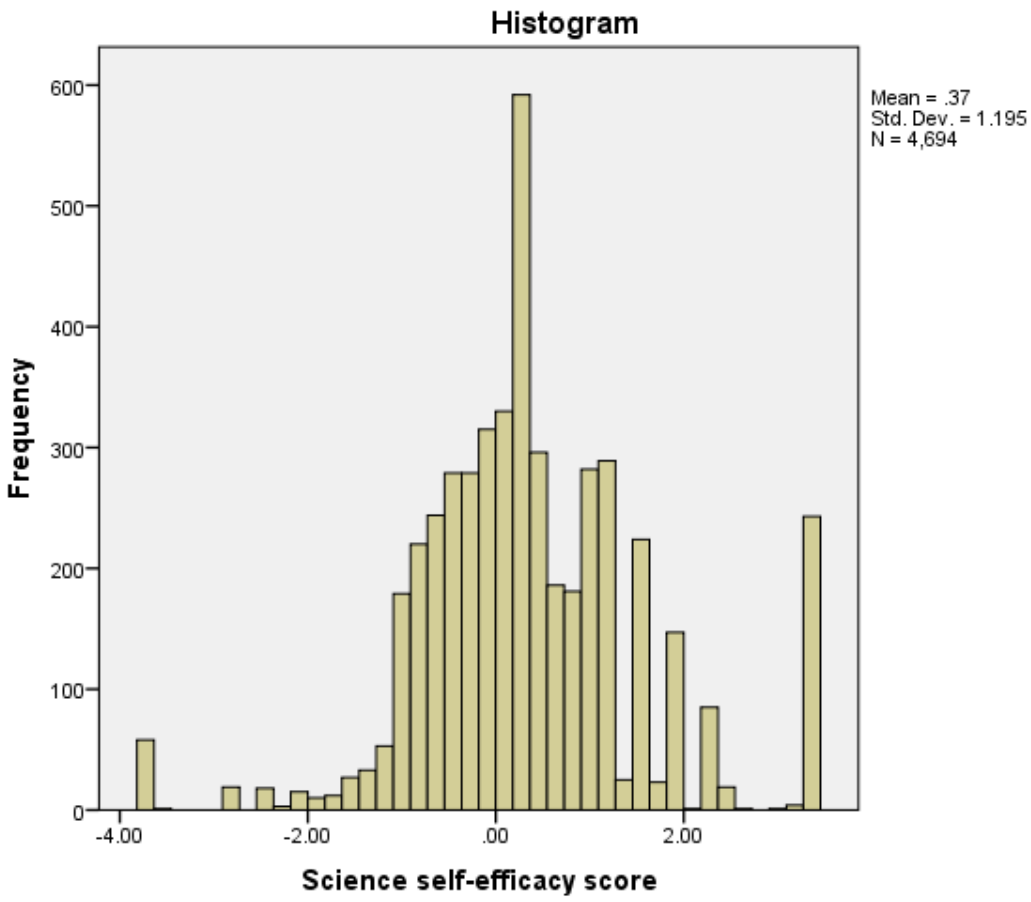
To do this we need to the following in SPSS:

- Select **Descriptive Statistics** from the **Analyze** menu.
- Select **Explore** from the **Descriptive Statistics** sub-menu.
- Click on the **Reset** button.
- Copy the **Science self-efficacy score[SCIEEFF]** and **Science instrumental motivation score[INSMOVSCI]** variables into the **Dependent List:** box.
- Click on the **Plots...** button.
- On the screen that appears select the **Histogram** tick box.
- Unselect the **Stem and leaf** button.

- Select the **Normality plots with tests** button.
- Click on the **Continue** button.
- Click on the **OK** button.

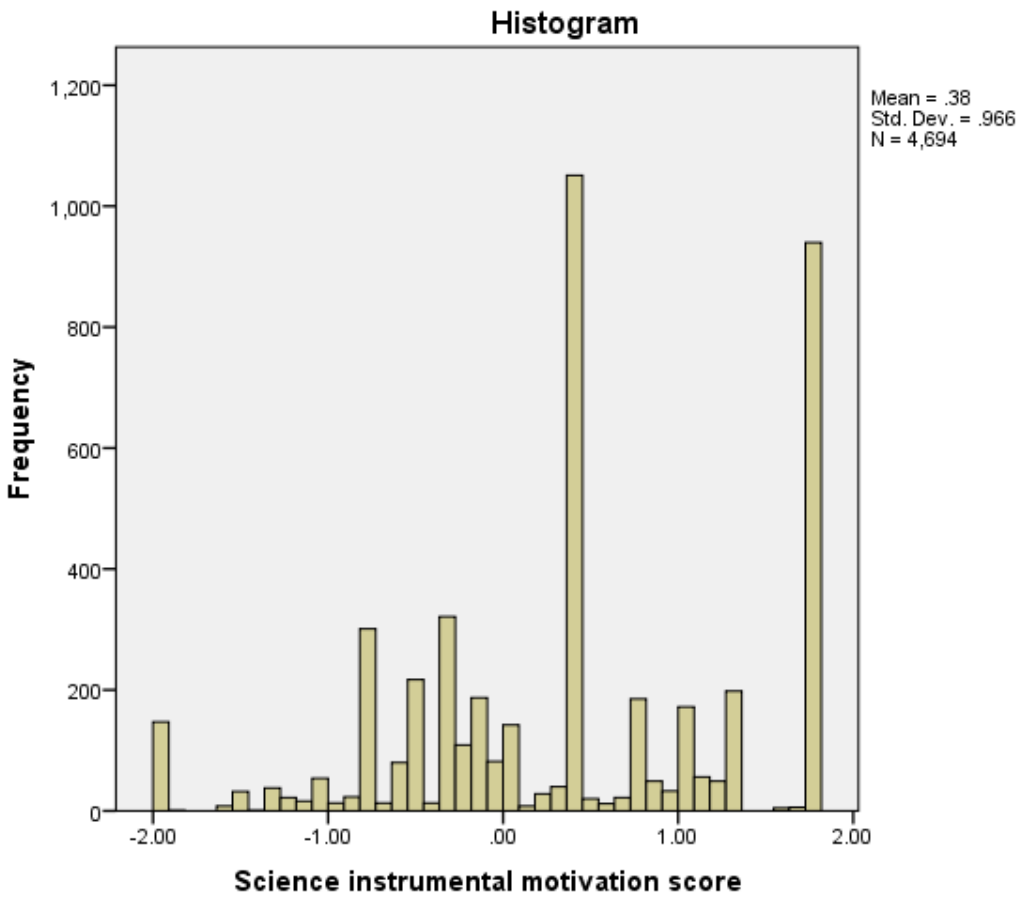
This set of instructions will create a whole list of outputs - both tables and figures - in SPSS. We will focus on two figures each for our two variables and then one table.

We will first look at a histogram of the variable, **SCIEEFF**. This can be found in amongst the set of output objects and looks as follows:



Ideally for a normal distribution this histogram should look symmetric around the mean of the distribution, in this case .3688. This distribution appears to be reasonably symmetric.

We will next look at a histogram of the variable, **INSMOVSCI**. This can also be found in amongst the set of output objects and looks as follows:



Again for a normal distribution this histogram should look symmetric around the mean of the distribution, in this case .3827. This distribution appears to be significantly skewed to the left (negatively skewed).

We will next look at statistical tests for the two variables to see if they back up our visual impressions from the histograms.

The Kolmogorov-Smirnov test is used to test the null hypothesis that a set of data comes from a normal distribution. An alternative test derived by Shapiro and Wilks is sometimes also available in SPSS but will not be described here. The available test statistics are presented in the table below that will be amongst the outputs from the Explore command:

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Science self-efficacy score	.079	4694	.000	.952	4694	.000
Science instrumental motivation score	.127	4694	.000	.940	4694	.000

a. Lilliefors Significance Correction

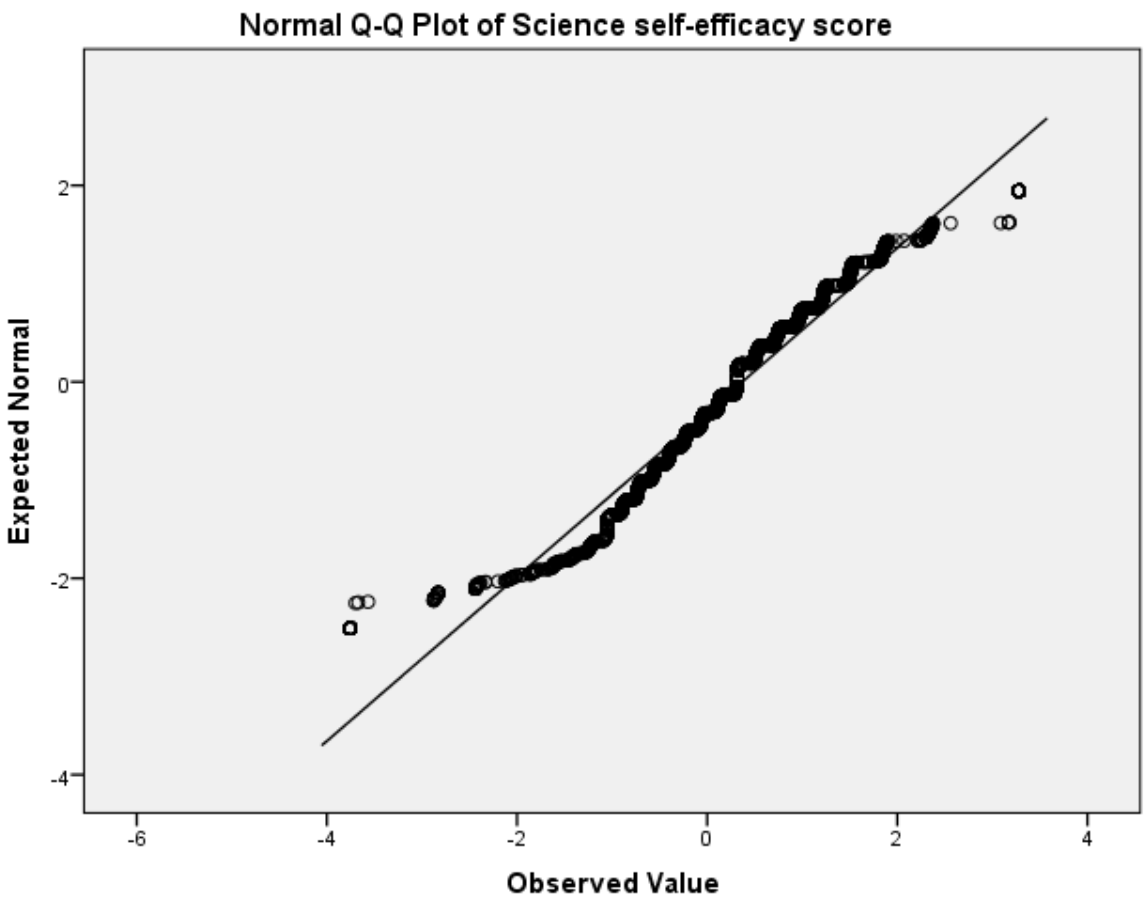
The Kolmogorov Smirnov tests produce test statistics that are used (along with a degrees of freedom parameter) to test for normality. Here we see that the Kolmogorov Smirnov statistic takes value .079 for **SCIEEFF** and value .127 for **INSMOVSCI**The test has degrees of freedom which equals the number of data points, namely 4694.

For **SCIEEFF** we see the following: The p value (quoted under Sig. for Kolmogorov Smirnov) is .000 (reported as p < .001) which is less than 0.05. We therefore have significant evidence to reject the null hypothesis that the variable follows a normal distribution.

For **INSMOVSCI** we see the following: The p value (quoted under Sig. for Kolmogorov Smirnov) is .000 (reported as p < .001) which is less than 0.05. We therefore have significant evidence to reject the null hypothesis that the variable follows a normal distribution.

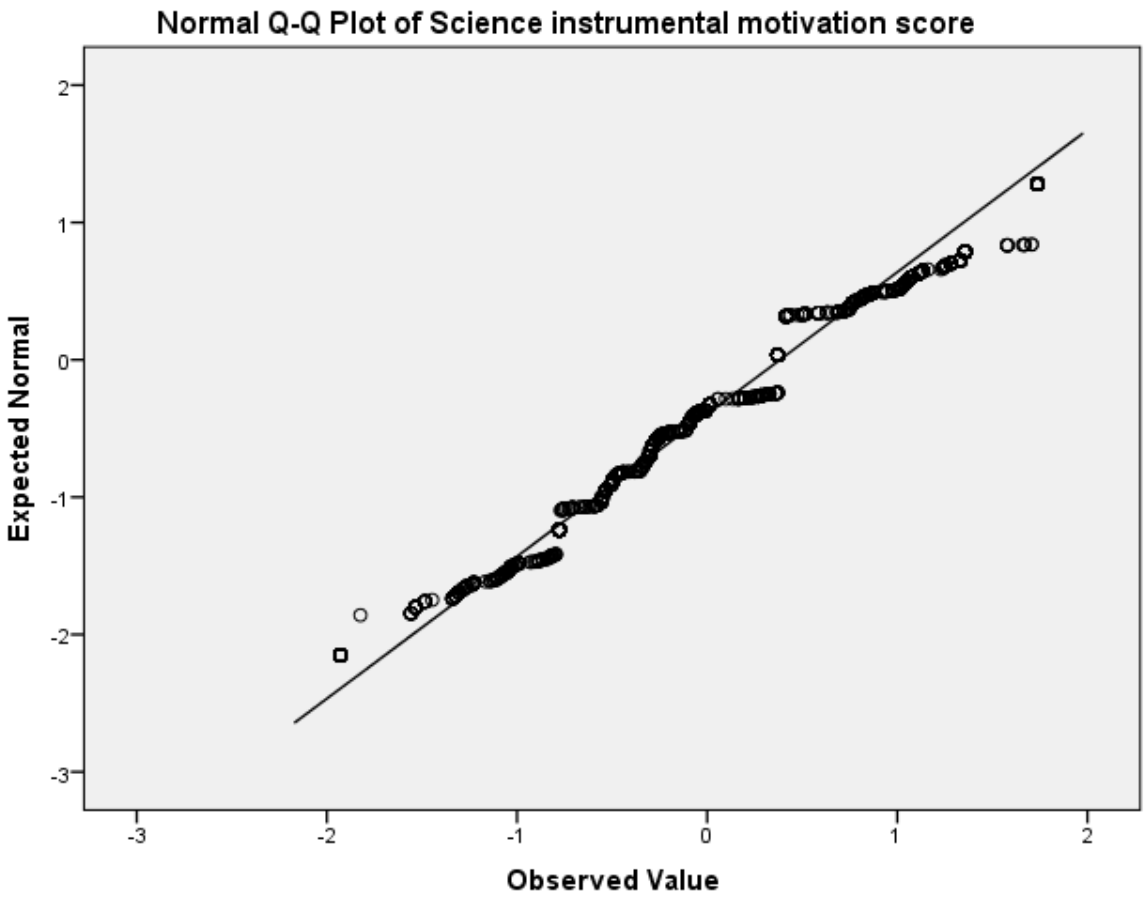
Although the Kolmogorov Smirnov test tells the researcher whether the distribution followed by a variable is statistically significantly different from a normal distribution one should take care in not overinterpreting such findings. Significance will be strongly effected by the number of observations and so only a small discrepancy from normality will be deemed significant for very large sample sizes whilst very large discrepancies will be required to reject the null hypothesis for small sample sizes.In addition, Pearson's correlation will be robust to non-normality in the data when samples are very large, as is the case here.

To complete our normality checking SPSS also produces Quantile-Quantile (or QQ) plots. We can see the one for **SCIEEFF** below:



QQ plots can be used to compare the distribution of a variable with a chosen distribution (typically a normal distribution as we are doing here). The data are plotted against a theoretical normal distribution (with the same mean and variance as the sample data) in such a way that the points should form an approximate straight line. Departures from this straight line indicate departures from normality. As we found a significant effect in the Kolmogorov Smirnov test for **SCIEEFF** we should see the points diverging from the line in the plot above with either some outlying values lying away from the line or even the shape of the points forming a non-linear pattern.

Similarly for **INSMOVSCI** its Quantile-Quantile plot can be seen below:



As we found a significant effect in the Kolmogorov Smirnov test for **INSMOVSCI** we should see the points diverging from the line in the plot above with either some outlying values lying away from the line or even the shape of the points forming a non-linear pattern.
We will now finally turn our attention to the main topic of this practical which is the calculation of the correlation between our two variables. SPSS offers several correlation coefficients and we will consider these here in turn. All three are available through the Analyse->Correlate->Bivariate option in SPSS.

- Select **Bivariate...** from the **Correlate** option available from the **Analyse** menu.
- Copy the **Science self-efficacy score[SCIEEFF]** and the **Science instrumental motivation score[INSMOVSCI]** variables into the **Variables** box.
- Click on the **Options** button and Select the **Means and Standard deviations** tick box.
- Click on the **Continue** button to return to main window.
- Click on the **OK** button.

The correlation command will produce two output tables. The first table which we show below simply gives means and standard deviations for the two variables we are comparing.

Descriptive Statistics

	Mean	Std. Deviation	N
Science self-efficacy score	.3671	1.19427	4726
Science instrumental motivation score	.3819	.96666	4791

In the next table we see the correlation matrix for the variables we are considering:

Correlations

		Science self-efficacy score	Science instrumental motivation score
Science self-efficacy score	Pearson Correlation	1	.327**
	Sig. (2-tailed)		.000
	N	4726	4694
Science instrumental motivation score	Pearson Correlation	.327**	1
	Sig. (2-tailed)	.000	
	N	4694	4791

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

The Correlate option can be used for more than two variables simultaneously and will then give all correlations hence the output table is in this matrix format. The table contains three numbers for each possible correlation (including the correlations of variables with themselves which always takes the value 1). For each correlation there is an estimate of the correlation, an accompanying p value and a sample size on which the correlation has been calculated. Here we are interested in the Pearson correlation between **SCIEEFF** and **INSMOVSCI** which can be found in two places in the table - either in the row for **SCIEEFF** and column for **INSMOVSCI** or the row for **INSMOVSCI** and column for **SCIEEFF**. Note that the SPSS table repeats exactly the same information twice, but in the write-up of results it should only be reported once!

In this case the correlation (reported as the statistic *r*) takes value .327. The widely-used rules specified by Cohen regard a correlation of *r*=.1 as small, *r*=.3 as moderate, and *r*=.5 as large. Here, then, we see a moderate positive correlation. The correlation is given in the table, along with a significance value and a sample size which in this case is 4694. This is the number of observations in which both **SCIEEFF** and **INSMOVSCI** where observed.

We can test if this correlation is significantly different from zero which will depend on (i) the magnitude of the correlation and (ii) the number of observations on which the correlation is based.

The p value (quoted under Sig. (2-tailed)) is .000 (reported as p < .001) which is less than 0.05. We therefore have significant evidence to reject the null hypothesis that the correlation is 0.

We would report the result as follows: The variables **SCIEEFF** and **INSMOVSCI** were significantly and moderately positively correlated *r* = .327, N = 4694, p < .001. Note there is no need for a table when reporting a single correlation.

The Pearson correlation coefficient is appropriate to use when both variables can be assumed to follow a normal distribution or when samples are very large.
If this is not the case then an alternative is the Spearman rank correlation. This correlation works in much the same way as the Pearson coefficient but is calculated on the ranks of the data points rather than the points themselves. To calculate the Spearman correlation we need to return to the Bivariate screen and do the following:

- Select **Bivariate...** from the **Correlate** option available from the **Analyse** menu.
- Check that the **Science self-efficacy score[SCIEEFF]** and the **Science instrumental motivation score[INSMOVSCI]** variables are still in the **Variables** box.
- Deselect the **Pearson** tick box.
- Select the **Spearman** tick box.
- Click on the **OK** button.

In the table produced we see the correlation matrix for the variables we are considering:

Correlations

			Science self-efficacy score	Science instrumental motivation score
Spearman's rho	Science self-efficacy score	Correlation Coefficient	1.000	.333**
		Sig. (2-tailed)	.	.000
		N	4726	4694
	Science instrumental motivation score	Correlation Coefficient	.333**	1.000
		Sig. (2-tailed)	.000	.
		N	4694	4791

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

For each correlation there is once again an estimate of the correlation, an accompanying p value and a sample size on which the correlation has been calculated. Here we are interested in the Spearman correlation between **SCIEEFF** and **INSMOVSCI** is repeated in two places in the table - either in the row for **SCIEEFF** and column for **INSMOVSCI** or the row for **INSMOVSCI** and column for **SCIEEFF**.

In this case the correlation (reported as the statistics rho) takes value .333. This represents a moderate positive correlation. The correlation is given in the table, along with a significance value and a sample size which in this case is 4694. This is the number of observations in which both **SCIEEFF** and **INSMOVSCI** where observed.

We can test if this correlation is significantly different from zero which will depend on (i) the magnitude of the correlation and (ii) the number of observations on which the correlation is based.

The p value (quoted under Sig. (2-tailed)) is .000 (reported as p < .001) which is less than 0.05. We therefore have significant evidence to reject the null hypothesis that the correlation is 0.

We would report the result as follows: The variables **SCIEEFF** and **INSMOVSCI** were significantly and moderatly positively correlated *r* = .333, N = 4694, p < .001.
The third possible correlation is known as Kendalls Tau-b and has desirable properties when the variables take values that are ordered categories (i.e. ordinal variables). To calculate the Kendalls Tau-b we need to return to the Bivariate screen and do the following:

- Select **Bivariate...** from the **Correlate** option available from the **Analyse** menu.
- Check that the **Science self-efficacy score[SCIEEFF]** and the **Science instrumental motivation score[INSMOVSCI]** variables are still in the **Variables** box.
- Deselect the **Spearman** tick box.
- Select the **Kendall tau-b** tick box.
- Click on the **OK** button.

In the table produced we see the correlation matrix for the variables we are considering:

Correlations

			Science self-efficacy score	Science instrumental motivation score
Kendall's tau_b	Science self-efficacy score	Correlation Coefficient	1.000	.240**
		Sig. (2-tailed)	.	.000
		N	4726	4694
	Science instrumental motivation score	Correlation Coefficient	.240**	1.000
		Sig. (2-tailed)	.000	.
		N	4694	4791

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

As in the previous correlation tables, for each pair of variables there is once again an estimate of the correlation, an accompanying p value and a sample size on which the correlation has been calculated, all repeated in two places in the table.

In this case the correlation (reported as the statistic tau) takes value .240. This correlation is small but positive. As before, the correlation coefficient is accompanied by the sample size used in the calculation and the significance value will depend on (i) the magnitude of the correlation and (ii) the number of observations on which the correlation is based.

The p value (quoted under Sig. (2-tailed)) is .000 (reported as $p < .001$) which is less than 0.05. We therefore have significant evidence to reject the null hypothesis that the correlation is 0.

We would report the result as follows: The variables **SCIEEFF** and **INSMOVSCI** were significantly and slightly positively correlated $r = .240$, $N = 4694$, $p < .001$.
This ends our practical on correlations.

In this example, Pearson's correlation is probably the most appropriate statistic to report, given the continuous nature of the variables and the very large sample size. However, the results all agree that there is a significant positive association between science self-efficacy and instrumental motivation, but not a very strong one. Students who are confident in their science ability also tend to value science in a career, but high levels on one construct certainly do not guarantee high levels on the other.