Correlations in SPSS (Quiz)

The development of this E-Book has been supported by the British Academy
This implementation is by National Centre for Research Methods and UK Data Service

Correlation practical questions

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

The dataset we are using is an excerpt from a cut-down dataset drawn from the Living Costs and Food Survey 2013, available from the UK Data Service: http://doi.org/10.5255/UKDA-SN-7932-2. You will be exploring the characteristics of two variables; total household expenditure and total household income. Both variables are measured in pounds per week. No conditions are required to use the data; however respondents are promised that their data will be kept confidential. As a result, high values of both variables are grouped together to prevent households being identified by their large household sizes or unusually high income. This protects respondents but it also affects the quality of the results produced in this workbook. Users who wish to use better quality data are encouraged to explore the full data from the Living Costs and Food Survey which is available through the UK Data Service (http://doi.org/10.5255/UKDA-SN-7702-1), for which users need to register and adhere to some conditions of use.

Firstly use SPSS to create a scatterplot of **income** and **expenditure** and answer the following:

- Question: What does the scatterplot say about the relationship between **income** and **expenditure**?

Next use SPSS and the Explore screen to create histograms, normality tests and QQ plots of **income** and **expenditure** and answer the following:

- Question: What do the plots and tests tell us about the normality of **income**?
- Question: What do the plots and tests tell us about the normality of **expenditure**?

Next use SPSS and the Correlate screen to answer the following:

- Question: What is the Pearson correlation coefficient between **income** and **expenditure** and is it significant?

Next use SPSS and the Correlate screen to answer the following:

- Question: What is the Spearman correlation coefficient between **income** and **expenditure** and is it significant?

Next use SPSS and the Correlate screen to answer the following:

- Question: What is the value of the Kendall tau-b correlation coefficient between **income** and **expenditure** and is it significant?

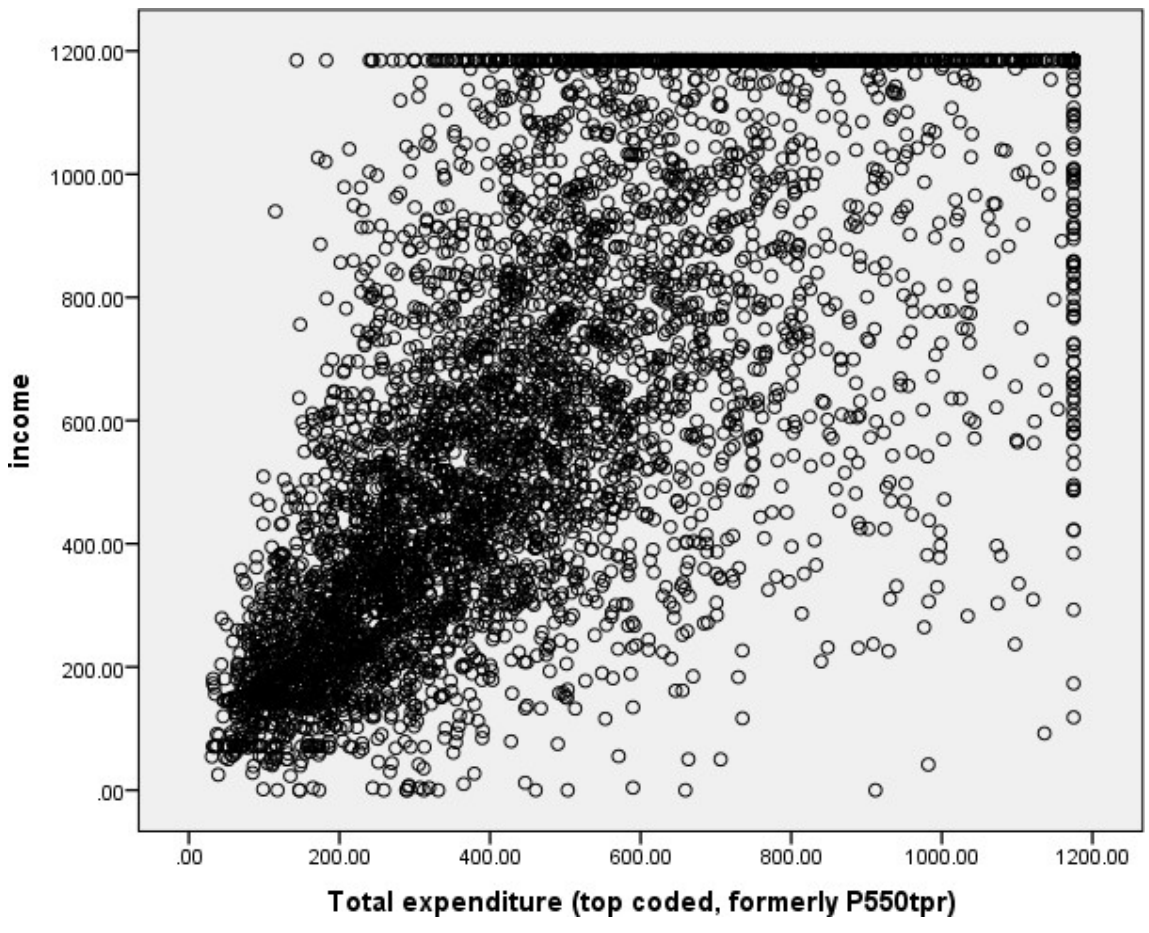
Solutions to Correlation practical questions

The SPSS instructions are as follows:

1. Select **Scatter/Dot** from the **Legacy Dialogs** available from the **Graphs** menu.
2. Select Simple Scatter and click on Define to bring up the Simple Scatterplot window.
3. Copy the **income** variable into the **Y Axis** box.
4. Copy the **Total expenditure (top coded, formerly P550tpr)[expenditure]** variable into the **X Axis** box.
5. Click on the **OK** button.

- Question: What does the scatterplot say about the relationship between **income** and **expenditure**?

Solution: The output from SPSS is as follows:



Looking at the scatterplot there appears to be a reasonably strong positive correlation between the variables with larger values of **income** associated with larger values of **expenditure** (an upward sloping relationship) as there are often in this case a considerable numbers of points in the bottom-left and top-right quarters of the plot.

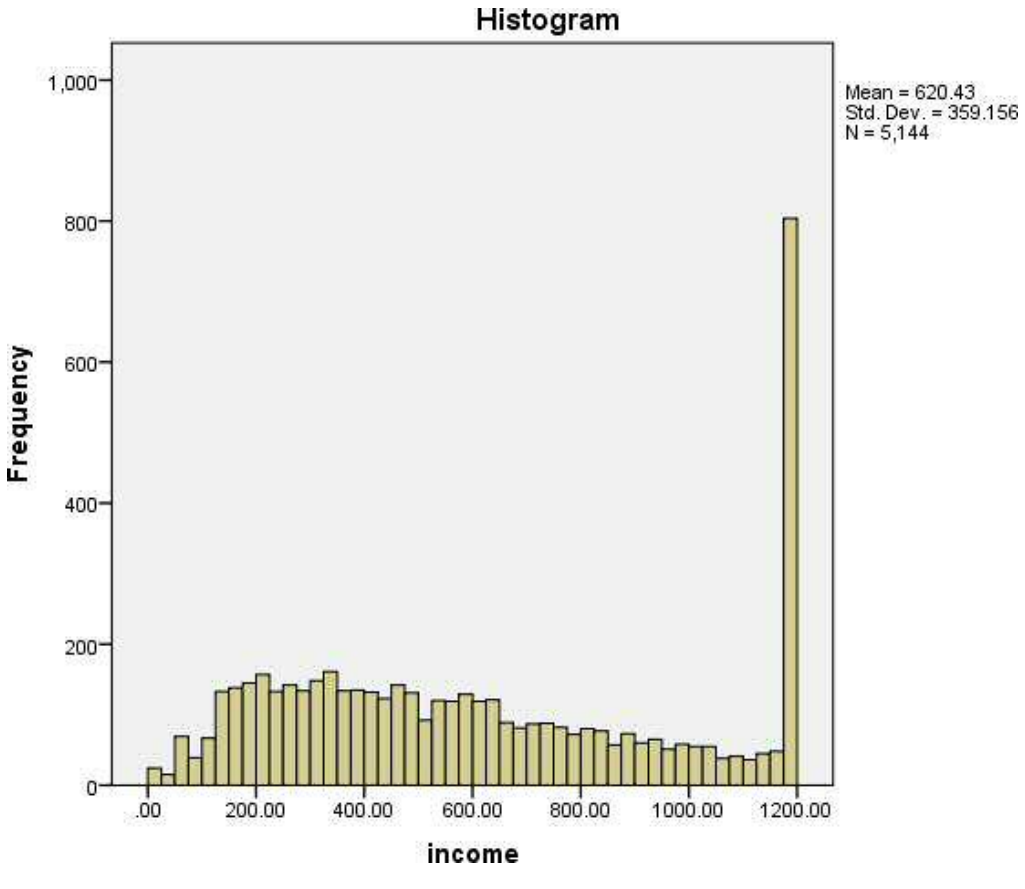
The SPSS instructions are as follows:

1. Select **Descriptive Statistics** from the **Analyze** menu.
2. Select **Explore** from the **Descriptive Statistics** sub-menu.
3. Click on the **Reset** button.
4. Copy the **income** and **Total expenditure (top coded, formerly P550tpr)[expenditure]** variables into the **Dependent List:** box.
5. Click on the **Plots...** button.
6. On the screen that appears select the **Histogram** tick box.
7. Unselect the **Stem and leaf** button.
8. Select the **Normality plots with tests** button.
9. Click on the **Continue** button.
10. Click on the **OK** button.

- Question: What do the plots and tests tell us about the normality of **income**?

Solution: The output from SPSS is as follows:

We will first look at a histogram of the variable, **income**.



Ideally for a normal distribution this histogram should look symmetric around the mean of the distribution, in this case 620.4336. This distribution appears to be significantly skewed to the right (positively skewed) and you may note that a large number of cases have been grouped into to the top income group.

Next we look at the Normality test statistics:

Tests of Normality

	Kolmogorov-Smirnov [a]		
	Statistic	df	Sig.
Income	.096	5144	.000
Total expenditure (top coded, formerly P550tpr)	.085	5144	.000

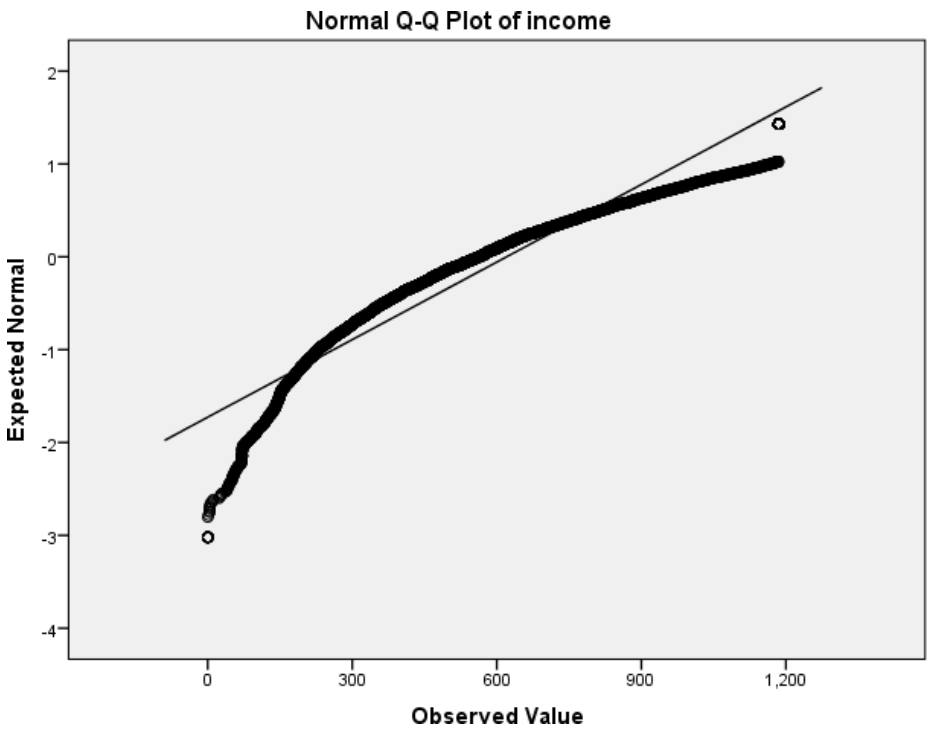
a. Lilliefors Significance Correction

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 .096 for **income**. The test has degrees of freedom which equals the number of data points, namely 5144.

For **income** 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 over interpreting such findings. Significance will be strongly affected 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.

For **income** its Quantile-Quantile plot can be seen below:

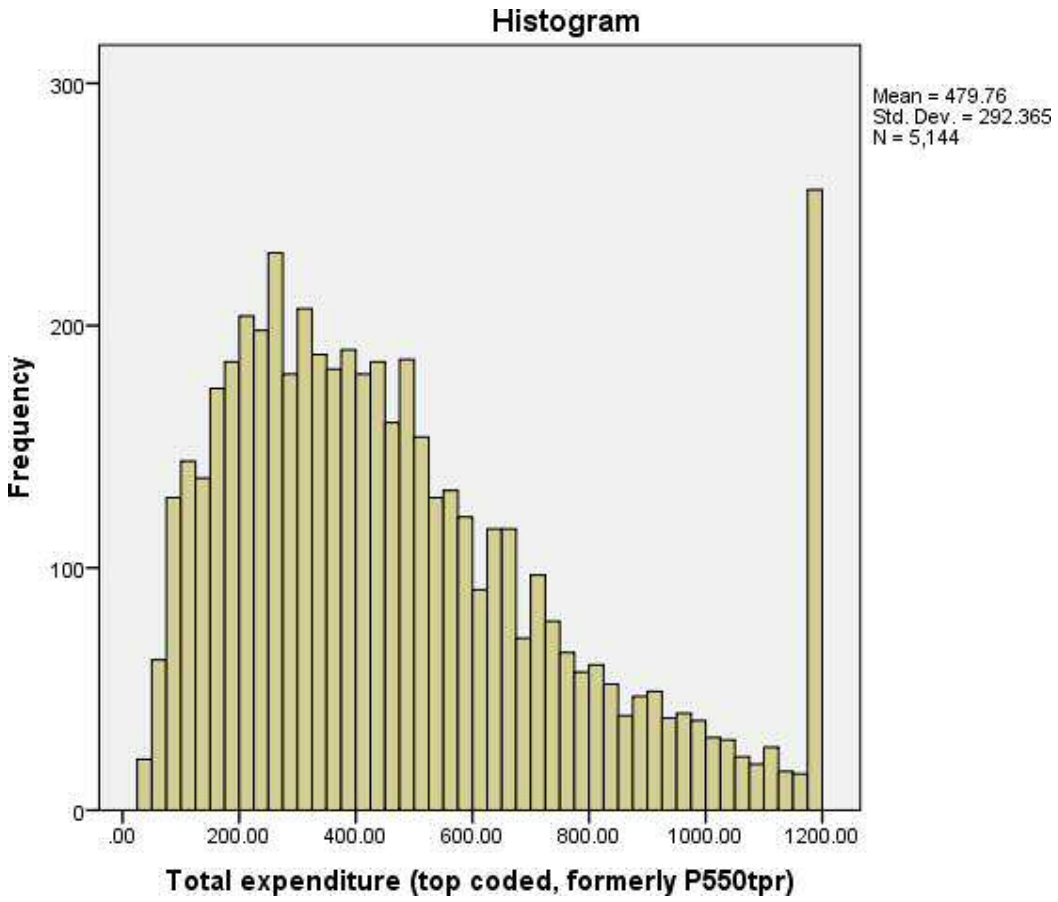


QQ plots can be used to compare the distribution of a variable with a chosen distribution (typically a normal distribution as we are doing here). The data are plotted against a theoretical normal distribution (with the same mean and variance as the sample data) in such a way that the points should form 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 **income** 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.

- Question: What do the plots and tests tell us about the normality of **expenditure**?

Solution: The output from SPSS is as follows:

We will first look at a histogram of the variable, **expenditure**.



Again for a normal distribution this histogram should look symmetric around the mean of the distribution, in this case 479.7584. This distribution appears to be significantly skewed to the right (positively skewed).

Next we look at the Normality test statistics:

Tests of Normality

	Kolmogorov-Smirnov [a]		
	Statistic	df	Sig.
Income	.096	5144	.000
Total expenditure (top coded, formerly P550tpr)	.085	5144	.000

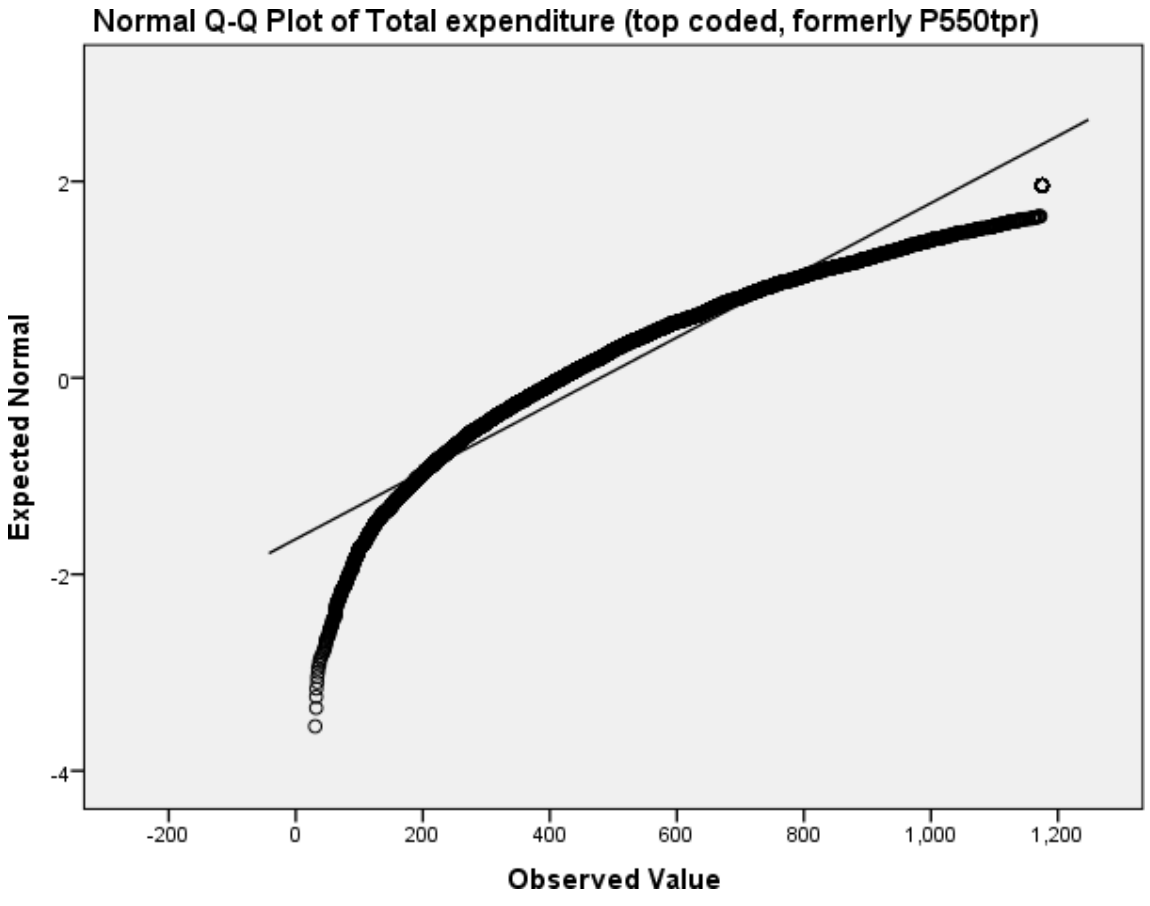
a. Lilliefors Significance Correction

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 .085 for **expenditure**. The test has degrees of freedom which equals the number of data points, namely 5144.

For **expenditure** 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 over interpreting such findings. Significance will be strongly affected 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.

For **expenditure** its Quantile-Quantile plot can be seen below:



As we found a significant effect in the Kolmogorov Smirnov test for **expenditure** 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.

The SPSS instructions are as follows:

1. Select **Bivariate...** from the **Correlate** option available from the **Analyse** menu.
2. Copy the **income** and the **Total expenditure (top coded, formerly P550tpr)[expenditure]** variables into the **Variables** box.
3. Click on the **Options** button and Select the **Means and Standard deviations** tick box.
4. Click on the **Continue** button to return to main window.
5. Click on the **OK** button.

- Question: What is the Pearson correlation coefficient between **income** and **expenditure** and is it significant?

Solution: The output from SPSS is as follows:

Descriptive Statistics

	Mean	Std. Deviation	N
Income	620.4336	359.15568	5144
Total expenditure (top coded, formerly P550tpr)	479.7584	292.36523	5144

Correlations

		income	Total expenditure (top coded, formerly P550tpr)
income	Pearson Correlation	1	.706 **
	Sig. (2-tailed)		.000
	N	5144	5144
Total expenditure (top coded, formerly P550tpr)	Pearson Correlation	.706 **	1
	Sig. (2-tailed)	.000	
	N	5144	5144

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

Here we are interested in the Pearson correlation between **income** and **expenditure** which can be found in two places in the table - either in the row for **income** and column for **expenditure** or the row for **expenditure** and column for **income**.

In this case the correlation takes value .706. This represents a large positive correlation. The correlation is given in the table, along with a significance value and a sample size which in this case is 5144. This is the number of observations in which both **income** and **expenditure** 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.

The SPSS instructions are as follows:

1. Select **Bivariate...** from the **Correlate** option available from the **Analyse** menu.
2. Check that the **income** and the **Total expenditure (top coded, formerly P550tpr)[expenditure]** variables are still in the **Variables** box.
3. Deselect the **Pearson** tick box.
4. Select the **Spearman** tick box.
5. Click on the **OK** button.

- Question: What is the Spearman correlation coefficient between **income** and **expenditure** and is it significant?

Solution: The output from SPSS is as follows:

Correlations

			income	Total expenditure (top coded, formerly P550tpr)
Spearman's rho	income	Correlation Coefficient	1.000	.750 **
		Sig. (2-tailed)	.	.000
		N	5144	5144
	Total expenditure (top coded, formerly P550tpr)	Correlation Coefficient	.750 **	1.000
		Sig. (2-tailed)	.000	.
		N	5144	5144

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

Here we are interested in the Spearman correlation between **income** and **expenditure** which can be found in two places in the table - either in the row for **income** and column for **expenditure** or the row for **expenditure** and column for **income**.

In this case the correlation takes value .750. This represents a large positive correlation. The correlation is given in the table, along with a significance value and a sample size which in this case is 5144. This is the number of observations in which both **income** and **expenditure** 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.

The SPSS instructions are as follows:

1. Select **Bivariate...** from the **Correlate** option available from the **Analyse** menu.
2. Check that the **income** and the **Total expenditure (top coded, formerly P550tpr)[expenditure]** variables are still in the **Variables** box.
3. Deselect the **Spearman** tick box.
4. Select the **Kendall tau-b** tick box.
5. Click on the **OK** button.

- Question: What is the value of the Kendall tau-b correlation coefficient between **income** and **expenditure** and is it significant?

Solution: The output from SPSS is as follows:

Correlations

			income	Total expenditure (top coded, formerly P550tpr)
Kendall's tau_b	income	Correlation Coefficient	1.000	.564 **
		Sig. (2-tailed)	.	.000
		N	5144	5144
	Total expenditure (top coded, formerly P550tpr)	Correlation Coefficient	.564 **	1.000
		Sig. (2-tailed)	.000	.
		N	5144	5144

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

Here we are interested in the Kendall Tau-b correlation between **income** and **expenditure** which can be found in two places in the table - either in the row for **income** and column for **expenditure** or the row for **expenditure** and column for **income**.

In this case the correlation takes value .564. This represents a large positive correlation. The correlation is given in the table, along with a significance value and a sample size which in this case is 5144. This is the number of observations in which both **income** and **expenditure** 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.

Copyright and citation

Jo Wathan, Vanessa Higgins, Mark Elliot, William Browne, Chris Charlton, Ana Morales Gomez and Jennifer Buckley (2019) Quantitative methods e-books: Teaching Resources, UK Data Service, NCRM, Centre for Multi-Level Modelling.