Correlations in SPSS (Practical)

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

Note: Weights have not been applied to the analyses. You can find out more about weighting survey data on the UK Data Service website.

Correlation practical

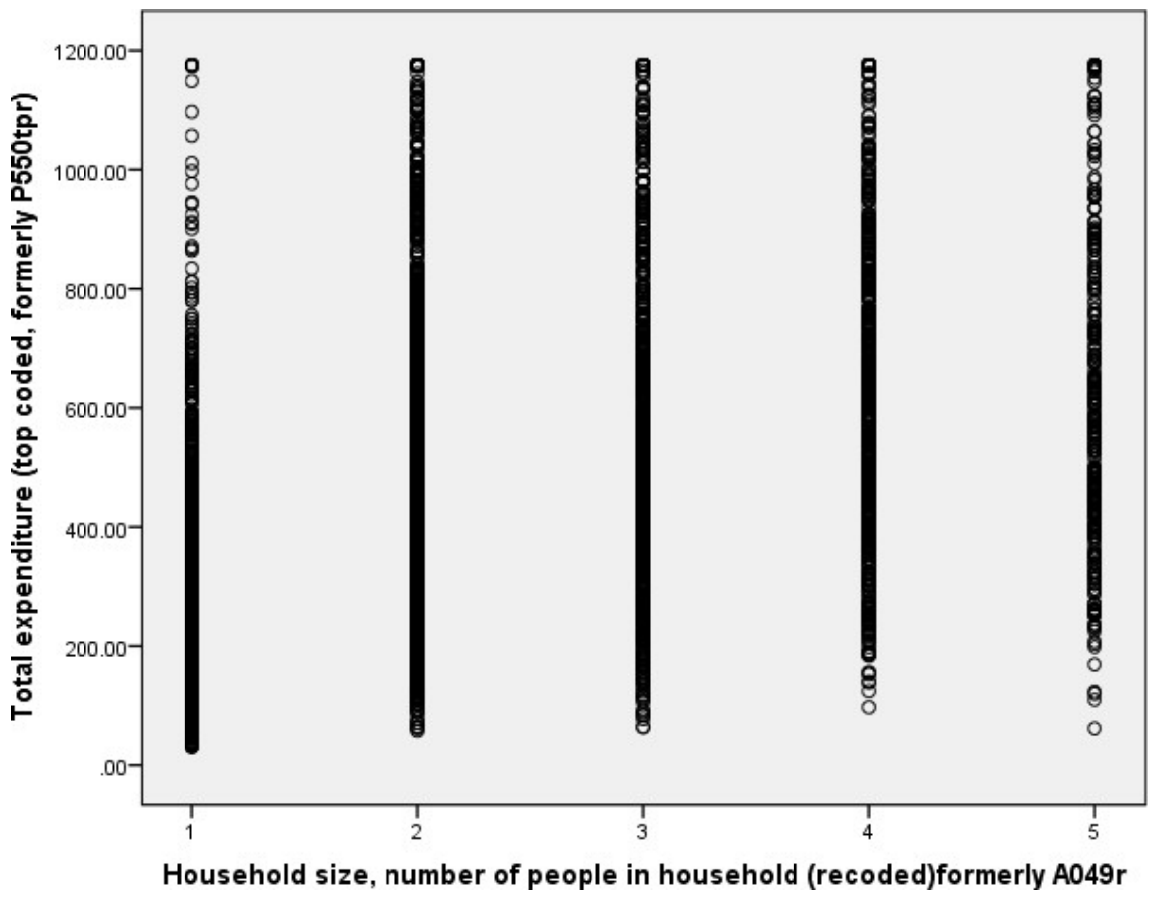
In this practical we will investigate whether there is a relationship between two variables by looking how correlated they are.

The dataset we are using is an excerpt from a cut-down dataset drawn from the Living Costs and Food Survey, available from the UK Data Service: http://doi.org/10.5255/UKDA-SN-7932-2, and we will be exploring the characteristics of two variables; household size and total household expenditure (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 are grouped together to prevent households being identified by their large household sizes or unusually high expenditure. 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.

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

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 **Total expenditure (top coded, formerly P550tpr)[expenditure]** variable into the **Y Axis** box.
4. Copy the **Household size, number of people in household (recoded) formerly A049r[hhsize]** variable into the **X Axis** box.
5. Click on the **OK** button.

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



Because there are only 5 possible values of household size and there are a lot of observations, it is quite difficult to see clearly what is going on. However the greatest concentration of values for expenditure do seem to increase somewhat as household size increases, therefore looking at the scatterplot there appears to be a positive correlation between the variables with larger values of **expenditure** associated with larger values of **hhsize** (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.

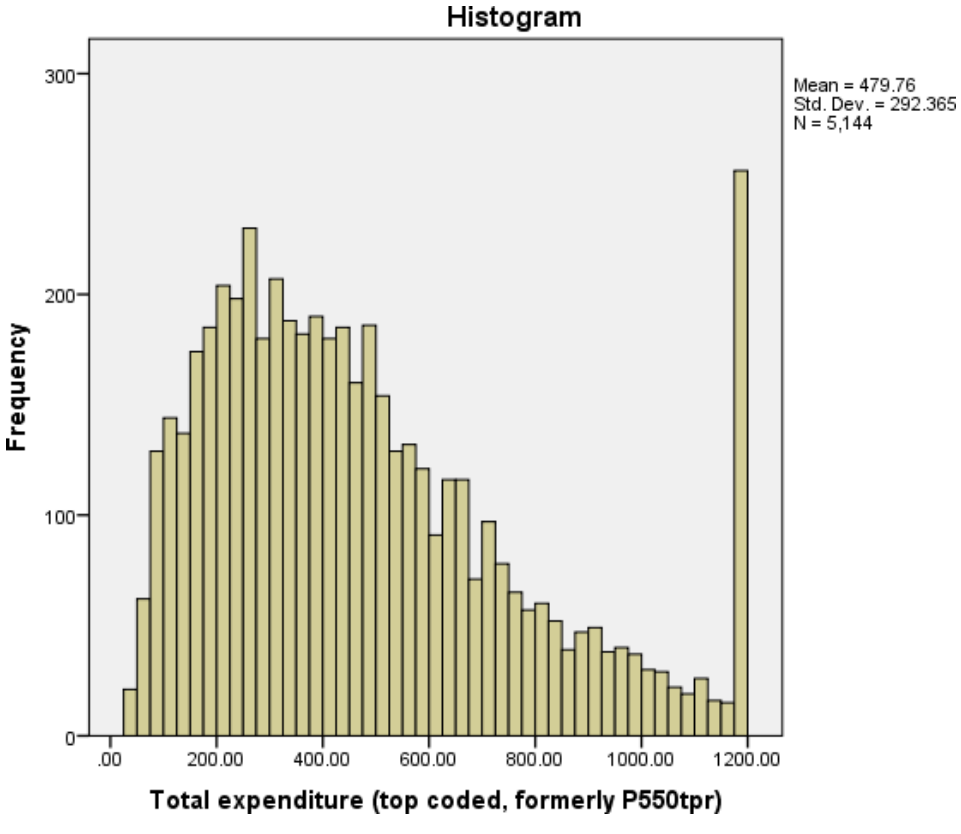
We want to test whether any correlation we observe in the scatterplot is significant but there are several different correlation coefficients for different situations. The first correlation coefficient that we will look at is the Pearson correlation coefficient. This correlation requires the variables 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:

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 **Total expenditure (top coded, formerly P550tpr)[expenditure]** and **Household size, number of people in household (recoded) formerly A049r[hhsize]** 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.

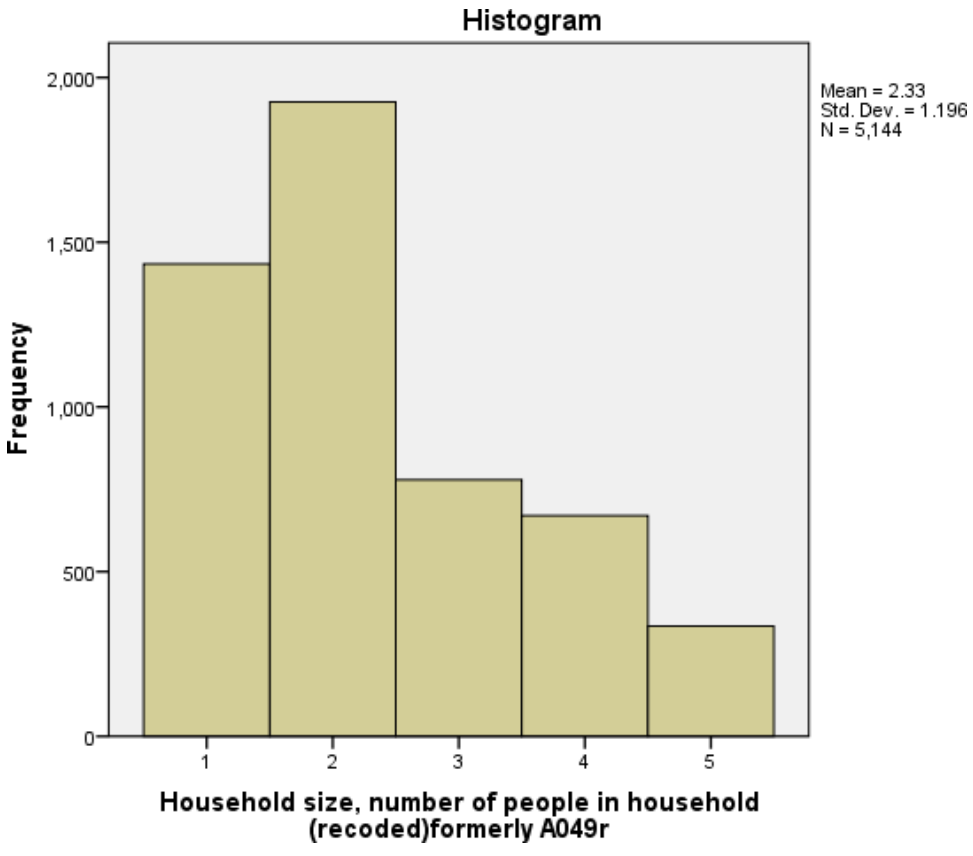
This set of instructions will create a whole list of outputs - both tables and figures - in SPSS. We will focus on two figures each for our two variables and then one table.

We will first look at a histogram of the variable, **expenditure**. 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 479.7584. This distribution appears to be significantly skewed to the right (positively skewed) and there is one maximum value that stands out as containing a disproportionate number of cases. This is an artefact of high values of expenditure being grouped which reduces the overall accuracy of the data and results in an underestimate of overall expenditure.

We will next look at a histogram of the variable, **hhsize**. 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 2.33. This distribution appears to be significantly skewed to the right (positively 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]		
	Statistic	df	Sig.
Total expenditure (top coded, formerly P550tpr)	.085	5144	.000
Household size, number of people in household (recoded)formerly A049r	.261	5144	.000

a. Lilliefors Significance Correction

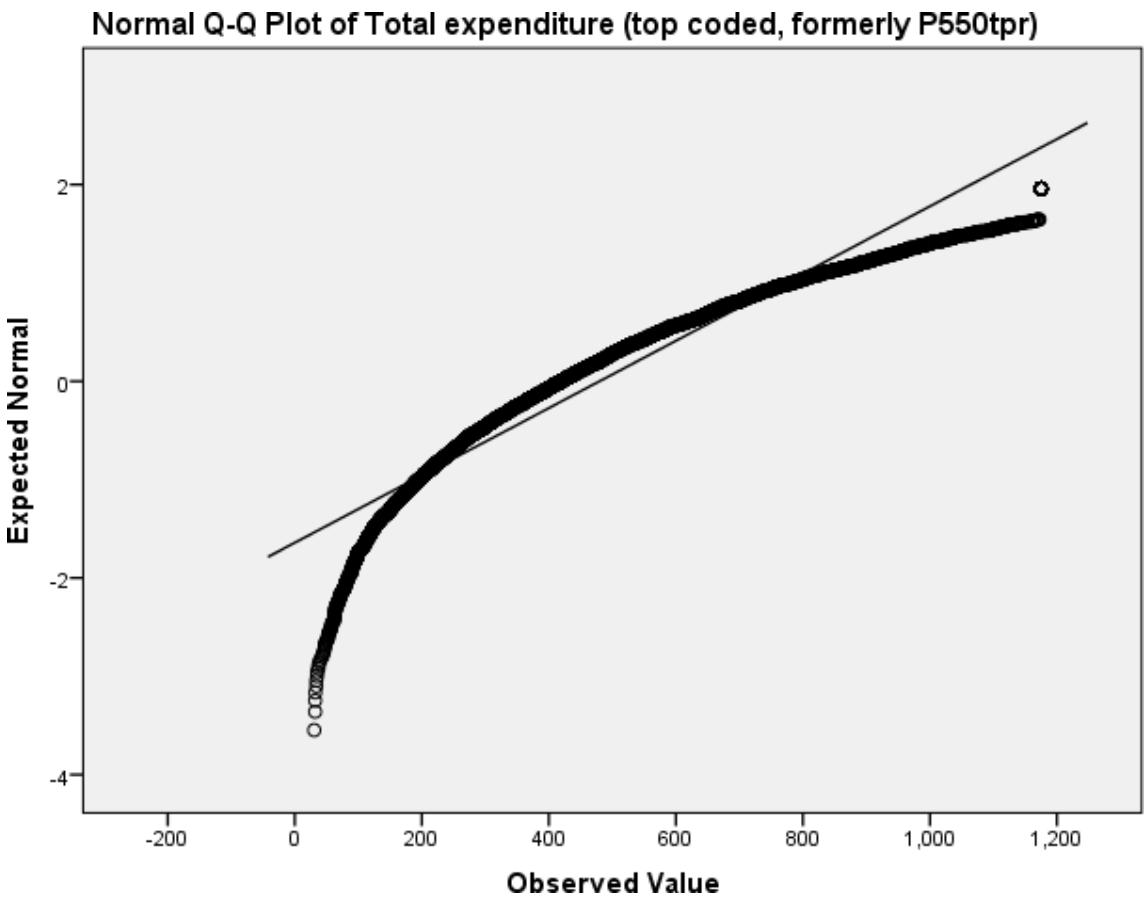
The Kolmogorov Smirnov tests produce test statistics that are used (along with a degrees of freedom parameter) to test for normality. Here we see that the Kolmogorov Smirnov statistic takes value .085 for **expenditure** and value .261 for **hhsize.** The test has degrees of freedom which equals the 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.

For **hhsize** 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.

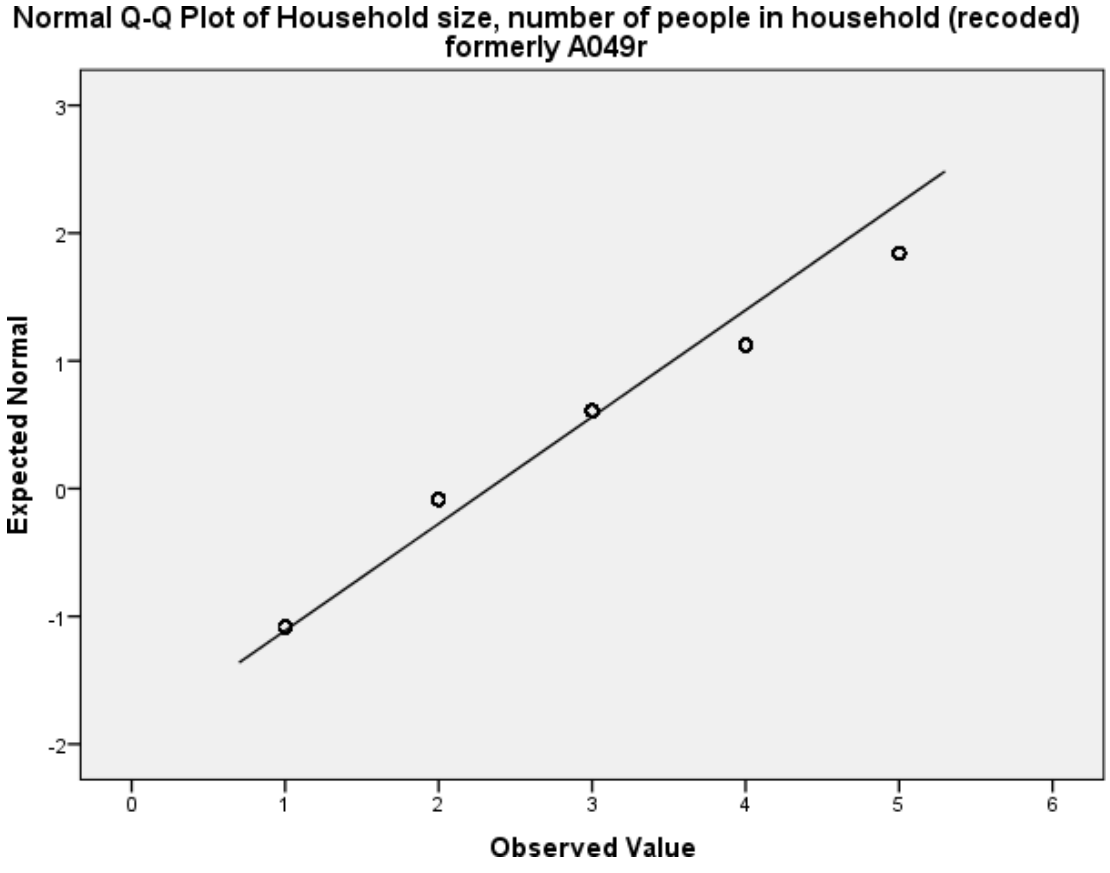
Although the Kolmogorov Smirnov test tells the researcher whether the distribution followed by a variable is statistically significantly different from a normal distribution, one should take care in not over interpreting such findings. Significance will be strongly affected by the number of observations 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.

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

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



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

1. Select **Bivariate...** from the **Correlate** option available from the **Analyse** menu.
2. Copy the **Total expenditure (top coded, formerly P550tpr)[expenditure]** and the **Household size, number of people in household (recoded)formerly A049r[hhsize]** 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.

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
Total expenditure (top coded, formerly P550tpr)	479.7584	292.36523	5144
Household size, number of people in household (recoded)formerly A049r	2.33	1.196	5144

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

Correlations

		Total expenditure (top coded, formerly P550tpr)	Household size, number of people in household (recoded) formerly A049r
Total expenditure (top coded, formerly P550tpr)	Pearson Correlation	1	.449 **
	Sig. (2-tailed)		.000
	N	5144	5144
Household size, number of people in household (recoded) formerly A049r	Pearson Correlation **	.449	1
	Sig. (2-tailed)	.000	
	N	5144	5144

**. 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 **expenditure** and **hhsize** which can be found in two places in the table - either in the row for **expenditure** and column for **hhsize** or the row for **hhsize** and column for **expenditure**. 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 .449. 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 5144. This is the number of observations in which both **expenditure** and **hhsize** 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 **expenditure** and **hhsize** were significantly and moderately positively correlated *r* = .449, N = 5144, 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:

1. Select **Bivariate...** from the **Correlate** option available from the **Analyse** menu.
2. Check that the **Total expenditure (top coded, formerly P550tpr)[expenditure]** and the **Household size, number of people in household (recoded) formerly A049r[hhsize]** 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.

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

Correlations

			Total expenditure (top coded, formerly P550tpr)	Household size, number of people in household (recoded)formerly A049r
Spearman's rho	Total expenditure (top coded, formerly P550tpr)	Correlation Coefficient	1.000	.519 **
		Sig. (2-tailed)	.	.000
		N	5144	5144
	Household size, number of people in household (recoded)formerly A049r	Correlation Coefficient	.519 **	1.000
		Sig. (2-tailed)	.000	.
		N	5144	5144

**. 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 **expenditure** and **hhsize** is repeated in two places in the table - either in the row for **expenditure** and column for **hhsize** or the row for **hhsize** and column for **expenditure**.

In this case the correlation (reported as the statistics rho) takes value .519. 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 **expenditure** and **hhsize** 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 **expenditure** and **hhsize** were significantly and positively correlated *r* = .519, N = 5144, 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:

1. Select **Bivariate...** from the **Correlate** option available from the **Analyse** menu.
2. Check that the **Total expenditure (top coded, formerly P550tpr)[expenditure]** and the **Household size, number of people in household (recoded) formerly A049r[hhsize]** 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.

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

Correlations

			Total expenditure (top coded, formerly P550tpr)	Household size, number of people in household (recoded)formerly A049r
Kendall's tau_b	Total expenditure (top coded, formerly P550tpr)	Correlation Coefficient	1.00	.397 **
		Sig. (2-tailed)	.	.000
		N	5144	5144
	Household size, number of people in household (recoded)formerly A049r	Correlation Coefficient	.397 **	1.000
		Sig. (2-tailed)	.000	.
		N	5144	5144

**. 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 .397. This represents a moderate positive correlation. 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 **expenditure** and **hhsize** were significantly and moderately positively correlated $r = .397$, N = 5144, $p < .001$.

This ends our practical on correlations.