Examining Continuous Variables in SPSS (Practical)

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

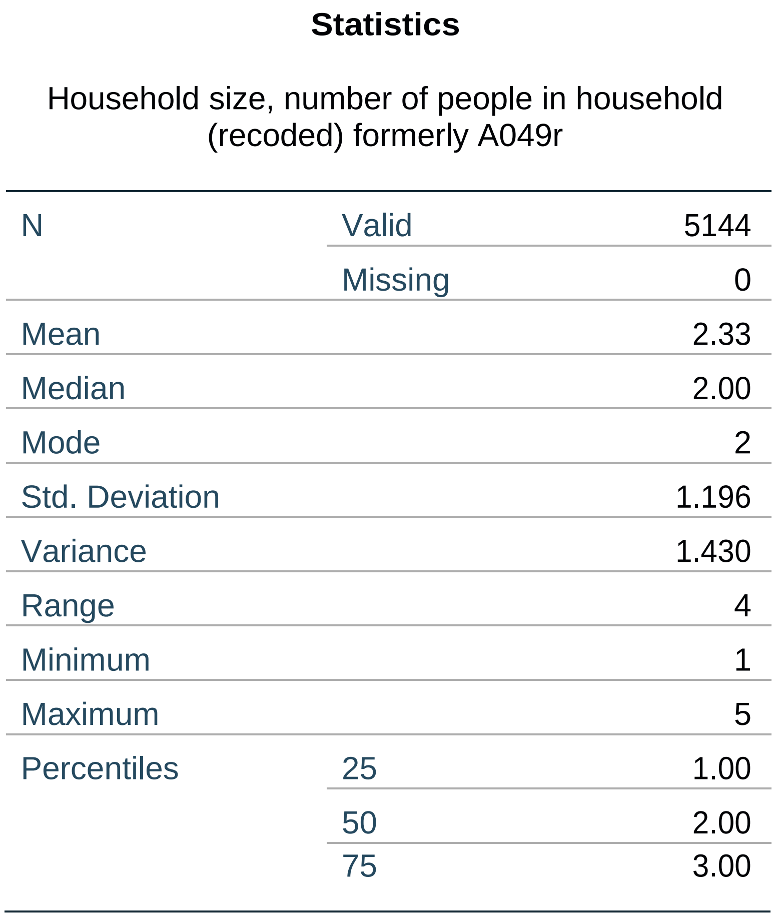
In this descriptive statistics practical we will use straightforward techniques to describe the features of continuous variables. We will look at how in SPSS we can obtain some summary statistics that describe the distribution of variables both in terms of measures of location and spread. We will also look at how we might summarise these variables graphically.

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 large household sizes or unusually high expenditures. 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.

We will begin by looking at how to use SPSS to get summary statistics for our first variable, **hhsize**.

1. Select **Frequencies** from the **Descriptive Statistics** submenu available from the **Analyze** menu.
2. Copy the **Household size, number of people in household (recoded) formerly A049r[hhsize]** variable into the **Variable(s)** box.
3. Click on the **Statistics** button to go to the statistics screen.
4. Here we need to select ALL the summary statistics that we are interested in looking at.
5. Select **Mean, Median** and **Mode** from under **Central Tendency**.
6. Select **Std. deviation, Variance, Range, Minimum** and **Maximum** from under **Dispersion**.
7. Finally Select **Quartiles** from under **Percentile Values.**
8. Click on the **Continue** button to return to the main window.
9. Click on the **OK** button to produce the tables as shown below.

The first table contains all the summary statistics that we requested for the variable as shown below:



Statistics

Household size, number of people in household (recoded) formerly A049r

N	Valid	5144
	Missing	0
Mean		2.33
Median		2.00
Mode		2
Std. Deviation		1.196
Variance		1.430
Range		4
Minimum		1
Maximum		5
Percentiles	25	1.00
	50	2.00
	75	3.00

We can see here that we have 5144 valid values for the variable **hhsize**, and no missing data.

The statistics begins with three measures of the centre of the distribution, the mean, median and the mode. For the variable **hhsize** we find that the arithmetic mean (or average) value is 2.33. The median or middle value is 2.00. This is smaller than the mean so if there is any skew to the distribution it will likely be positive. The third measure is the mode or most frequent value which takes value 2. SPSS calculates this by looking at the frequencies of each possible value so the mode is probably more useful for categorical data. You can check this by looking at the second table produced by the command (shown after this explanation) which shows the frequencies of each value.

We next look at measures of the spread of values for the variable **hhsize**. These begin with the standard deviation which takes value 1.196 and its squared value, the variance which takes value 1.430. Typically, if the data are normally distributed, approximately 95% of observations will lie within 2 standard deviations of the mean i.e. between -0.063 and 4.72. The smallest value observed is 1 and the largest value is 5 giving an overall range of length 4. We can finally see the quartiles of the distribution under the more general percentiles heading and so we see the lower (25%) quartile takes value 1.00 meaning that 25% of observations are below this value. Conversely 25% of observations are above the upper (75%) quartile which takes value 3.00. The 50% quantile is the median which we covered earlier.

The SPSS command also produces a second table shown below of all the observed values for **hhsize** and their frequencies. You may note that the top category not only includes those with household sizes of 5, but also those households with 6 or more residents.

Household size, number of people in household (recoded) formerly A049r

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	1434	27.9	27.9	27.9
	2	1926	37.4	37.4	65.3
	3	779	15.1	15.1	80.5
	4	670	13.0	13.0	93.5
	5	335	6.5	6.5	100.0
	Total	5144	100.0	100.0	

We do not have much to add about this table aside to mention that the mode, 2 can be found by finding the value with the largest frequency. The Cumulative Percent column can also be used to confirm the percentiles and you should find that if you scan down this column that the values pass 25% at the value quoted as the lower quartile and similarly passing 50% and 75% for the median and upper quartile respectively.

We will now move on to looking at a second variable, **expenditure**. This can be done in SPSS as follows:

1. Select **Frequencies** from the **Descriptive Statistics** submenu available from the **Analyze** menu.
2. Remove the **Household size, number of people in household (recoded) formerly A049r[hhsize]** variable from the **Variable(s)** box.
3. Copy the **Total expenditure (top coded, formerly P550tpr)[expenditure]** variable into the **Variable(s)** box.
4. The **Statistics** options will be remembered so do not need adding again.
5. Click on the **OK** button to produce the tables as shown below.

Once again the first table contains all the summary statistics that we requested for the variable as shown below:

Statistics

Total expenditure (top coded, formerly P550tpr)

N	Valid	5144
	Missing	0
Mean		479.7584
Median		419.9034
Mode		1175.00
Std. Deviation		292.36523
Variance		85477.425
Range		1144.48
Minimum		30.52
Maximum		1175.00
Percentiles	25	253.8735
	50	419.9034
	75	645.1281

This time we can see that we have 5144 valid values for the variable **expenditure**, and no missing data.

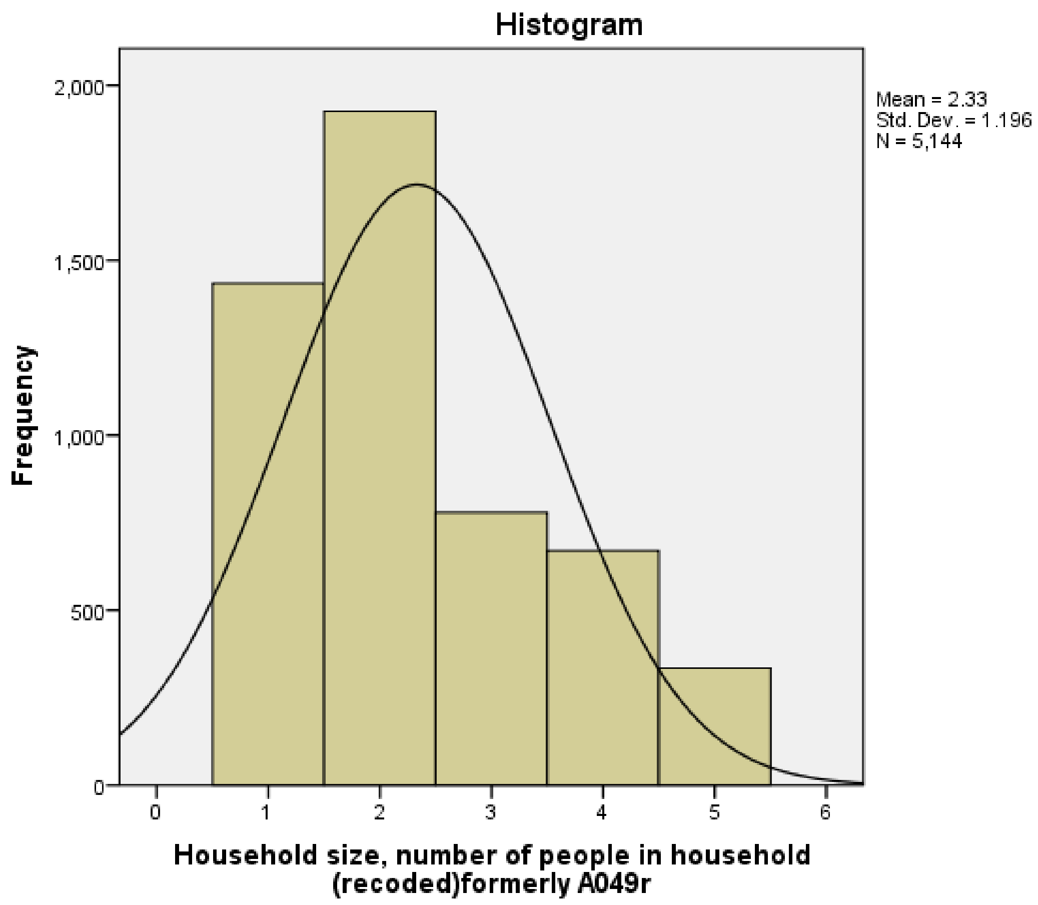
The three measures of the centre of the distribution, the mean, median and the mode appear next. For the variable **expenditure** we find that the arithmetic mean (or average) value is 479.7584. The median or middle value is 419.9034. This is smaller than the mean so if there is any skew to the distribution it will likely be positive. The third measure is the mode or most frequent value which takes value 1175.00. This value is an artefact of how this variable has been presented - can you think why this the value of this mode is so high? You can again check this is the most frequent value by looking at the second table produced by the command.

We next look at measures of the spread of values for the variable **expenditure**. These begin with the standard deviation which takes value 292.36523 and its squared value, the variance which takes value 85477.425. As noted, typically 95% of observations will lie within 2 standard deviations of the mean i.e. between -104.972 and 1064.489. The smallest value observed is 30.52 and the largest value is 1175.00 giving an overall range of length 1144.48. We can finally see the quartiles of the distribution under the more general percentiles heading and so we see the lower (25%) quartile takes value 253.8735 meaning that 25% of observations are below this value. Conversely 25% of observations are above the upper (75%) quartile which takes value 645.1281. The 50% quantile is the median which we covered earlier.

As before, the SPSS command also produces a second table of all the observed values for **expenditure** and their frequencies although this is not very useful when the data is not categorical.

We will next look at the data graphically but again using options from the Frequencies screen in SPSS for our first variable, **hhsize** as follows:

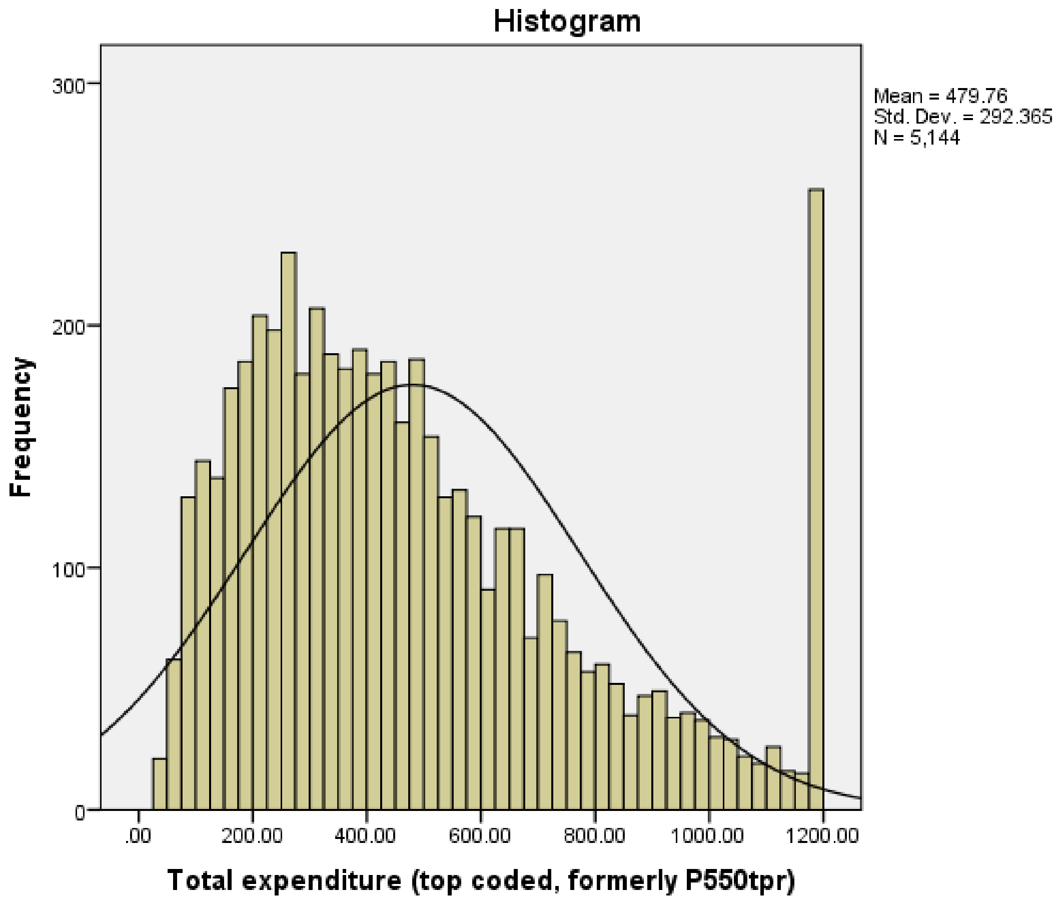
1. Select **Frequencies** from the **Descriptive Statistics** submenu available from the **Analyze** menu.
2. Remove the **Total expenditure (top coded, formerly P550tpr)[expenditure]** variable from the **Variable(s)** box.
3. Return the **Household size, number of people in household (recoded)formerly A049r[hhsize]** variable into the **Variable(s)** box.
4. Click on the **Charts...** button to bring up the chart options.
5. Click on the **Histogram** Chart type and also the **Show normal curve on histogram** tick box.
6. Click on the **Continue** button to return to the main window.
7. Click on the **OK** button to produce the graph as shown below.



The graph produced is a histogram and basically is somewhat similar to a bar graph but each bar in the histogram represents a range of values and as a result there are no gaps between bars. SPSS chooses the limits for each bar and actually plots frequencies of observations that lie between the limits. To check this you might compare the frequencies in the table with the bars in the graph. We have asked for a normal curve to be superimposed on the plot and this curve is a plot of the normal distribution that has the same mean and standard deviation as the data. Some statistical tests rely on the variable approximately following a normal distribution and if this is true then the histogram should roughly follow the curve. If the data is skewed, for example, then this will not be the case.

We will next look at plotting a similar histogram for the second variable, **expenditure** as follows:

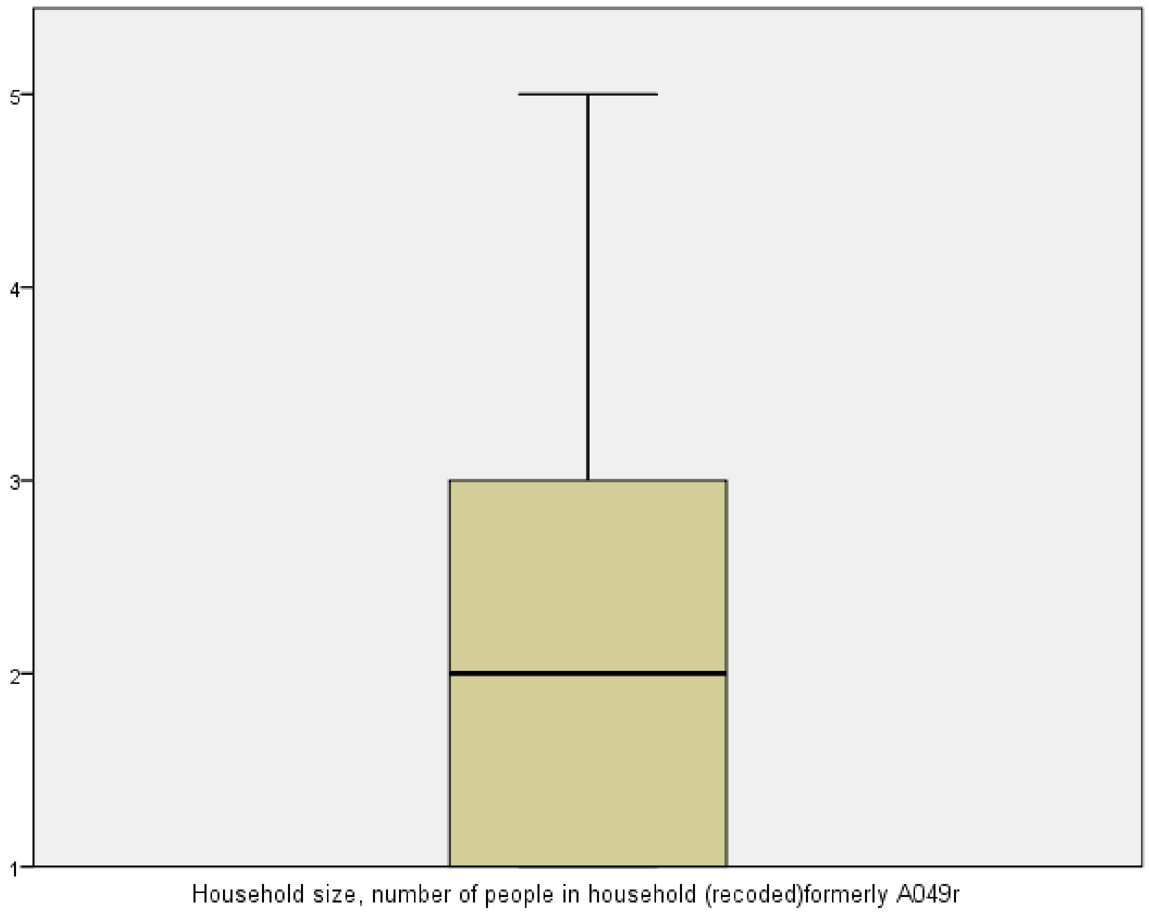
1. Select **Frequencies** from the **Descriptive Statistics** submenu available from the **Analyze** menu.
2. Remove the **Household size, number of people in household (recoded) formerly A049r[hhsize]** variable from the **Variable(s)** box.
3. Copy the **Total expenditure (top coded, formerly P550tpr)[expenditure]** variable into the **Variable(s)** box.
4. The **Charts** options will be remembered so do not need adding again.
5. Click on the **OK** button to produce the graph as shown below.



Again we can look at the shape of the histogram and check for unusual observations such as the peak at the very end of the distribution which is accounted for by values above 1175 being grouped together in this single value. Compare the graph with the plot of the normal distribution.

We will finish up this practical by looking at a second plot called a boxplot. We will do this first for variable, **hhsize**. The boxplot is not available from the Frequencies option so instead we need to do the following in SPSS:

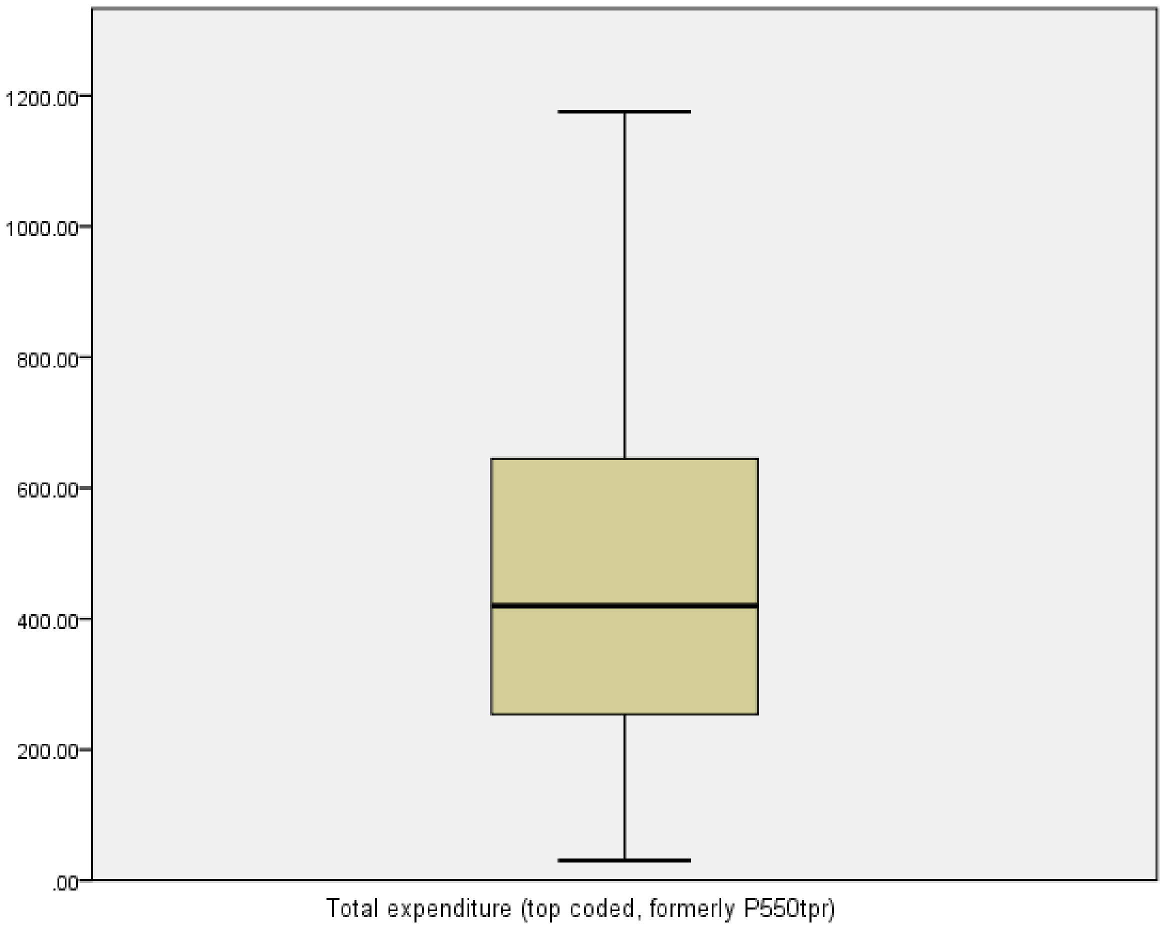
1. Select **Boxplot** from the **Legacy Dialogs** submenu available from the **Graphs** menu.
2. We want to choose **Simple** and **Summaries of separate variables** from the options here.
3. Next click on **Define** to set up the box plot.
4. Copy the **Household size, number of people in household (recoded) formerly A049r[hhsize]** variable into the **Boxes Represent:** box.
5. Ignore the rest of the options and click on the **OK** button to produce the graph as shown below.



Here we see the boxplot for the **hhsize** variable. The central box covers the interquartile range and so it has a value at the bottom of 1.00 and at the top of 3.00. The median which takes value 2.00 is represented by a vertical line in the middle of the box. The lines stretching out of the box to form T shapes are known as whiskers and will show the range unless there are any outliers (defined here as points 1.5 times the height of the box away from the box). If there are outliers the whiskers will end at 1.5 times the height of the box away from the box and the outliers will be marked as circles with a number representing which observation number they are.

We can also look at a boxplot for variable, **expenditure** as follows:

1. Select **Boxplot** from the **Legacy Dialogs** submenu available from the **Graphs** menu.
2. Keep the choices of **Simple** and **Summaries of separate variables** and click on **Define** to set up the box plot.
3. Remove the **Household size, number of people in household (recoded) formerly A049r[hhsize]** variable into the **Boxes Represent:** box.
4. Copy the **Total expenditure (top coded, formerly P550tpr)[expenditure]** variable into the **Boxes Represent:** box.
5. Ignore the rest of the options and click on the **OK** button to produce the graph as shown below.



Here we see the boxplot for the **expenditure** variable. This time the central box has a value at the bottom of 253.8735 and at the top of 645.1281. The median which takes value 419.9034 is represented by a vertical line in the middle of the box.

This ends the practical.