Examining Continuous Variables in SPSS (Practical)

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

Descriptives statistics for all variables practical

In this descriptive statistics practical we will expand our investigation of variables to include 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.

In this example we explore the distribution of two measures of 15 year-olds academic achievement from PISA – the overall science proficiency score SCISCORE and the overall mathematics proficiency score MATHSCORE. PISA test scores are designed to have a mean of 500 and a standard deviation of 100 across all OECD countries combined. PISA further explain that 30 test scores points correspond to progress expected in approximately one year of schooling, and that a score of 410 is regarded as the baseline level of proficiency needed for participation in social, economic and civic life in adulthood. With these benchmarks in mind, we can use descriptive statistics to get a sense of the distribution of performance of students in England.

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

- Select **Frequencies** from the **Descriptive Statistics** submenu available from the **Analyze** menu.
- Copy the **Science test score[SCISCORE]** variable into the **Variable(s)** box.
- Click on the **Statistics** button to go to the statistics screen.
- Here we need to select ALL the summary statistics that we are interested in looking at.
- Select **Mean, Median** and **Mode** from under **Central Tendency**.
- Select **Std. deviation, Variance, Range, Minimum** and **Maximum** from under **Dispersion**.
- Finally Select **Quartiles** from under **Percentile Values.**
- Click on the **Continue** button to return to the main window.
- 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

Science test score

N	Valid	5194
	Missing	0
Mean		523.5944
Median		527.6530
Mode		350.78[a]
Std. Deviation		102.57804
Variance		10522.254
Range		677.91
Minimum		175.22
Maximum		853.13
Percentiles	25	449.8808
	50	527.6530
	75	596.8897

a. Multiple modes exist. The smallest value is shown

We can see here that we have 5194 valid values for the variable **SCISCORE**, 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 **SCISCORE** we find that the arithmetic mean (or average) value is 523.5944. The median or middle value is 527.6530. This is larger than the mean so if there is any skew to the distribution it will likely be negative. The third measure is the mode or most frequent value which takes value 350.78. 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 which shows the frequencies of each value.

We next look at measures of the spread of values for the variable **SCISCORE**. These begin with the standard deviation which takes value 102.57804 and its squared value, the variance which takes value 10522.254. Typically, if the data are normally distributed, appoximately 95% of observations will lie within 2 standard deviations of the mean i.e. between 318.438 and 728.75. The smallest value observed is 175.22 and the largest value is 853.13 giving an overall range of length 677.91. 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 449.8808 meaning that 25% of observations are below this value. Conversely 25% of observations are above the upper (75%) quartile which takes value 596.8898. The 50% quantile is the median which we covered earlier.

The SPSS command also produces a second table of all the observed values for **SCISCORE** and their frequencies although this is not very useful when the data is not categorical.
We will now move on to looking at a second variable, **MATHSCORE**.

This can be done in SPSS as follows:

- Select **Frequencies** from the **Descriptive Statistics** submenu available from the **Analyze** menu.
- Remove the **Science test score[SCISCORE]** variable from the **Variable(s)** box.
- Copy the **Math test score[MATHSCORE]** variable into the **Variable(s)** box.
- The **Statistics** options will be remembered so do not need adding again.
- 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

Math test score

N	Valid	5194
	Missing	0
Mean		501.8590
Median		505.6065
Mode		559.14
Std. Deviation		97.10044
Variance		9428.495
Range		680.52
Minimum		109.66
Maximum		790.17
Percentiles	25	434.5785
	50	505.6065
	75	572.1085

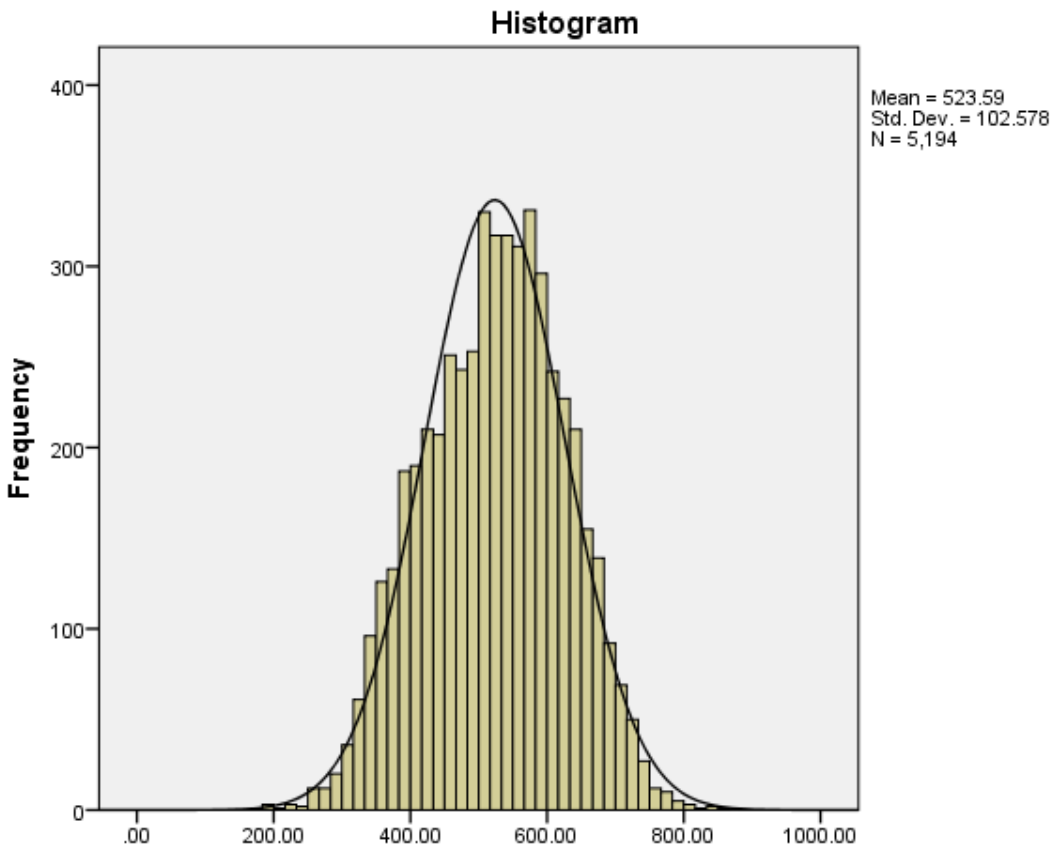
This time we can see that we have 5194 valid values for the variable **MATHSCORE**, and no missing data.

The three measures of the centre of the distribution, the mean, median and the mode appear next. For the variable **MATHSCORE** we find that the arithmetic mean (or average) value is 501.8590. The median or middle value is 505.6065. This is larger than the mean so if there is any skew to the distribution it will likely be negative. The third measure is the mode or most frequent value which takes value 559.14. 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 **MATHSCORE**. These begin with the standard deviation which takes value 97.10044 and its squared value, the variance which takes value 9428.495. As noted, typically 95% of observations will lie within 2 standard deviations of the mean i.e. between 307.658 and 696.06. The smallest value observed is 109.66 and the largest value is 790.17 giving an overall range of length 680.52. 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 434.5785 meaning that 25% of observations are below this value. Conversely 25% of observations are above the upper (75%) quartile which takes value 572.1085. The 50% quantile is the median which we covered earlier.

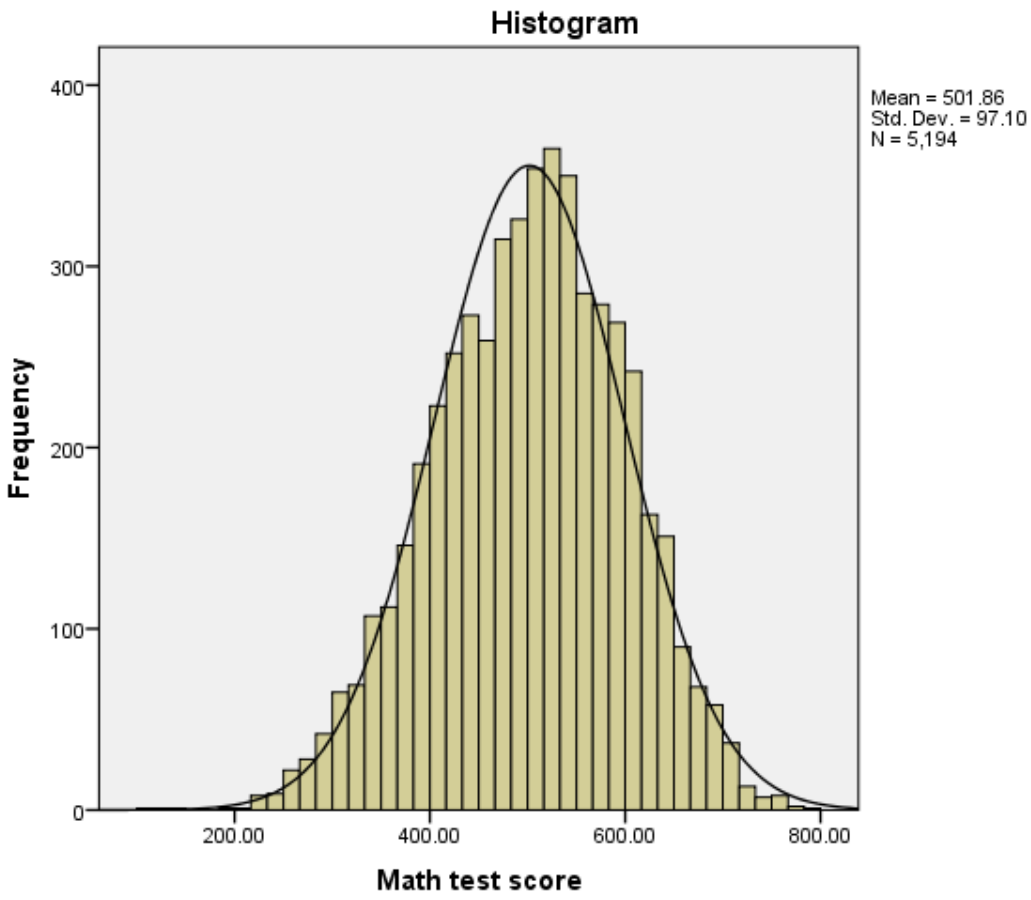
As before, the SPSS command also produces a second table of all the observed values for **MATHSCORE** and their frequencies although this is not very useful when the data is not categorical.
We will next look at the data graphically but again using options from the Frequencies screen in SPSS for our first variable, **SCISCORE** as follows:

- Select **Frequencies** from the **Descriptive Statistics** submenu available from the **Analyze** menu.
- Remove the **Math test score[MATHSCORE]** variable from the **Variable(s)** box.
- Return the **Science test score[SCISCORE]** variable into the **Variable(s)** box.
- Click on the **Charts...** button to bring up the chart options.
- Click on the **Histogram** Chart type and also the **Show normal curve on histogram** tick box.
- Click on the **Continue** button to return to the main window.
- Click on the **OK** button to produce the graph as shown below.



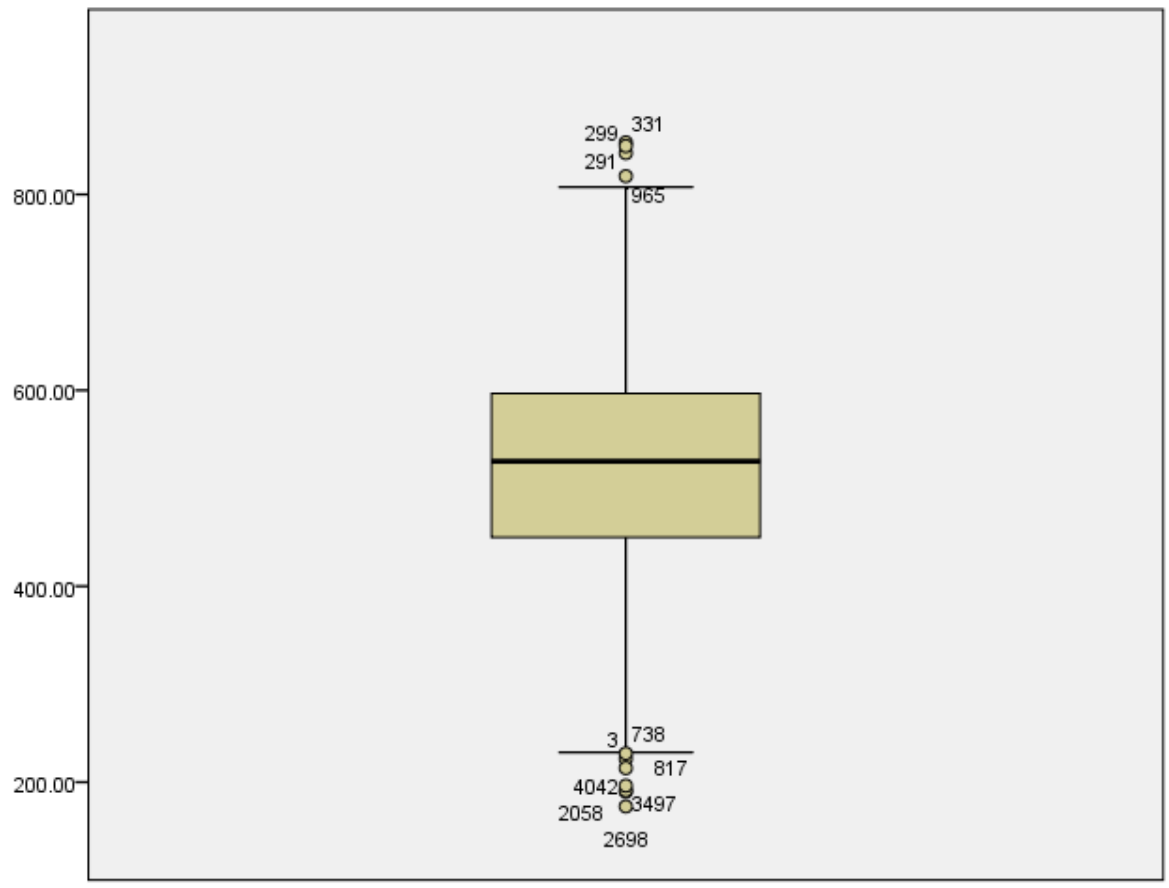
The graph produced is a histogram and basically is somewhat similar to a bar graph but each bar in the histogram represents a range of values and as a result there are no gaps between bars. SPSS chooses the limits for each bar and actually plots frequencies of observations that lie between the limits. 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, **MATHSCORE** as follows:

- Select **Frequencies** from the **Descriptive Statistics** submenu available from the **Analyze** menu.
- Remove the **Science test score[SCISCORE]** variable from the **Variable(s)** box.
- Copy the **Math test score[MATHSCORE]** variable into the **Variable(s)** box.
- The **Charts** options will be remembered so do not need adding again.
- Click on the **OK** button to produce the graph as shown below.



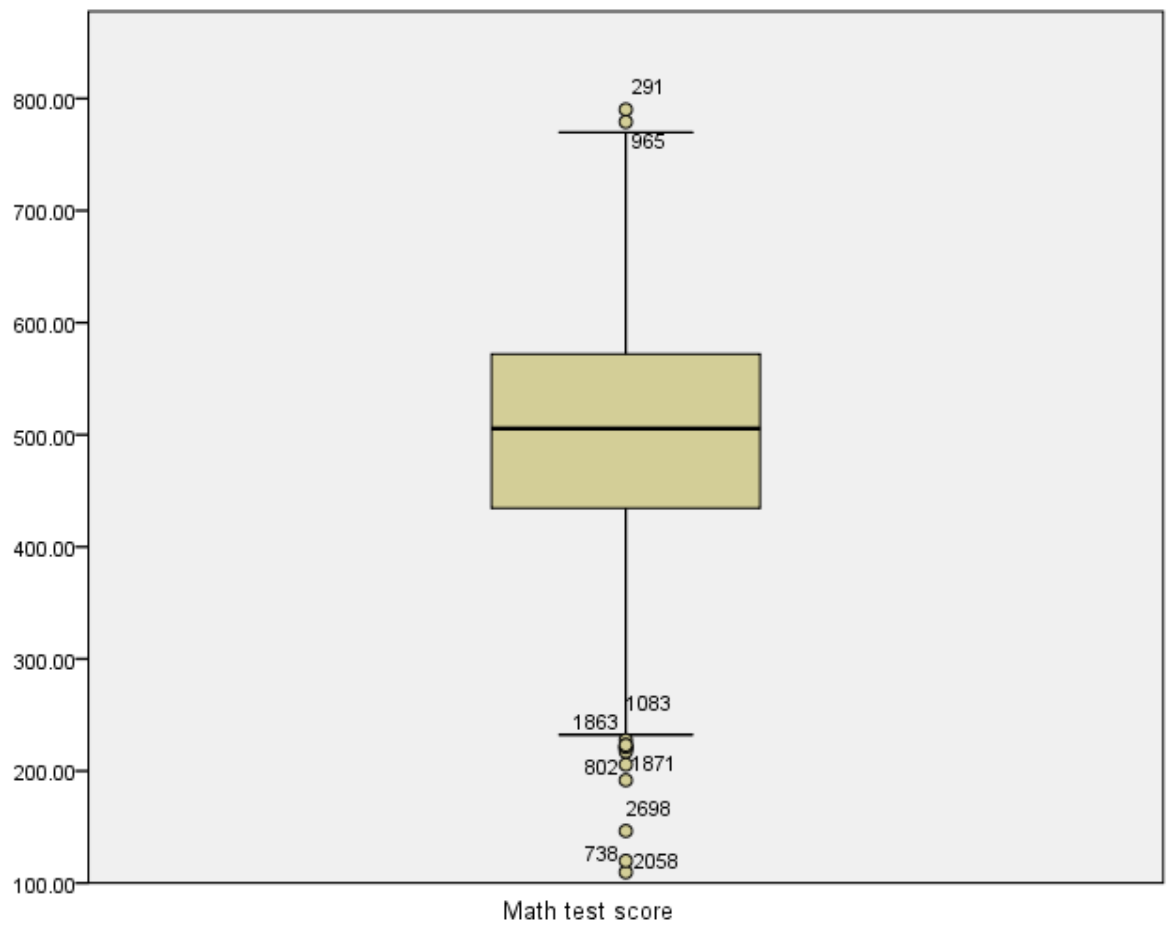
Again we can look at the shape of the histogram and check for unusual observations and compare the graph with the plot of the normal distribution.
We will finish up this practical by looking at a second plot called a boxplot. We will do this first for variable, **SCISCORE**. The boxplot is not available from the Frequencies option so instead we need to do the following in SPSS:

- Select **Boxplot** from the **Legacy Dialogs** submenu available from the **Graphs** menu.
- We want to choose **Simple** and **Summaries of separate variables** from the options here.
- Next click on **Define** to set up the box plot.
- Copy the **Science test score[SCISCORE]** variable into the **Boxes Represent:** box.
- 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 **SCISCORE** variable. The central box covers the interquartile range and so it has a value at the bottom of 449.8808 and at the top of 596.8898. The median which takes value 527.6530 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, **MATHSCORE** as follows:

- Select **Boxplot** from the **Legacy Dialogs** submenu available from the **Graphs** menu.
- Keep the choices of **Simple** and **Summaries of separate variables** and click on **Define** to set up the box plot.
- Remove the **Science test score[SCISCORE]** variable into the **Boxes Represent:** box.
- Copy the **Math test score[MATHSCORE]** variable into the **Boxes Represent:** box.
- 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 **MATHSCORE** variable. This time the central box has a value at the bottom of 434.5785 and at the top of 572.1085. The median which takes value 505.6065 is represented by a vertical line in the middle of the box.
We have seen that, among 15 year-olds in England, mean science performance is considerably above the OECD average of 500, while mean maths performance is about equal to the OECD average. Science performance is more variable than maths performance.The 25 th percentiles of both scores are well above the threshold of 410, showing the proportion of students in England not able to meet the baseline proficiency will be considerably less than a quarter.

This ends the practical.