SPSS - Exploring Normality (Practical)

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

Checking for Normality practical

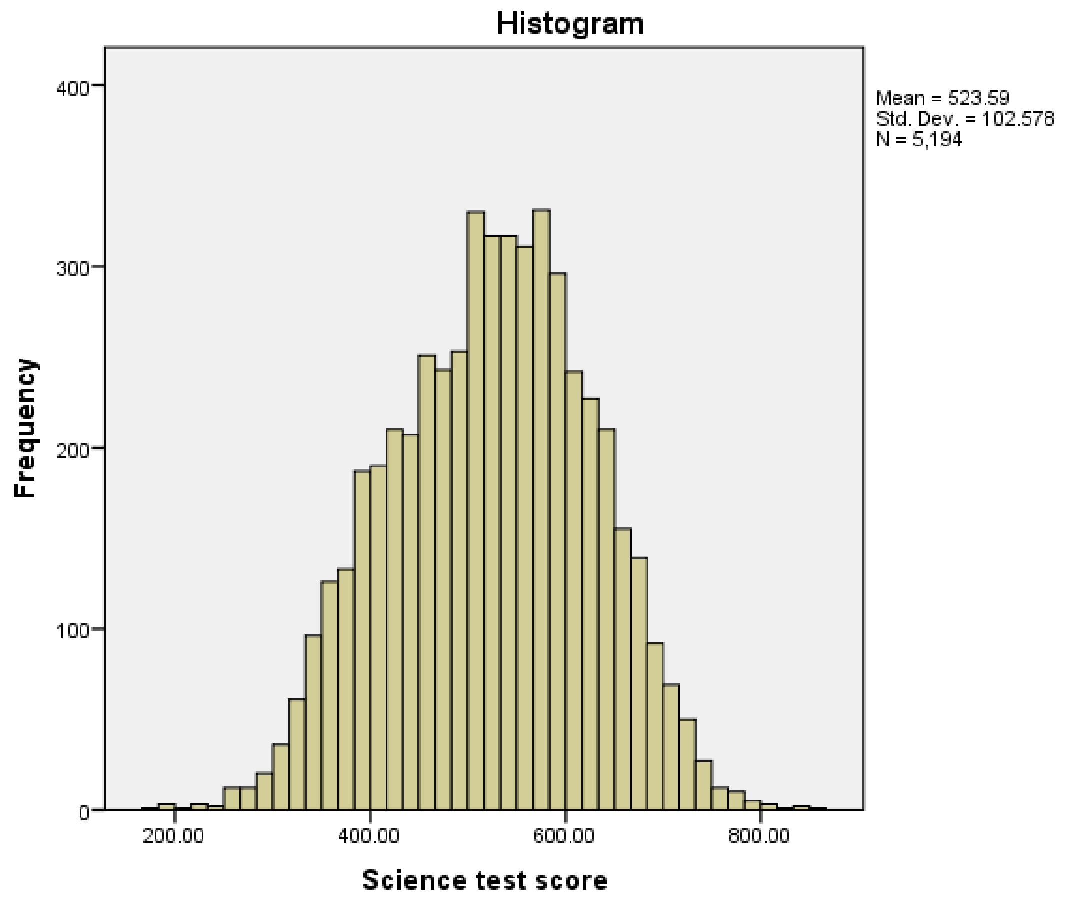
In this practical we look at how we can use SPSS to investigate whether a variable can be assumed to be normally distributed. This is an important decision as most of the parameteric statistical tests that we consider rely on the assumption that variables are normally distributed, unless sample sizes are very large. We will look at this both graphically and through a statistical test known as the Kolmogorov-Smirnov test.

Here we explore whether the PISA science test score (SCISCORE) appears normally distributed in the sample as a whole.

We start by giving instructions on how to get the required graphs and the test statistics in SPSS which are accessed via the **Explore** option as detailed here:

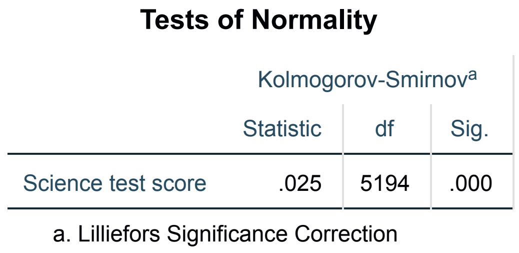
- Select **Descriptive Statistics** from the **Analyze** menu.
- Select **Explore** from the **Descriptive Statistics** sub-menu.
- Click on the **Reset** button.
- Copy the **Science test score[SCISCORE]** variables into the **Dependent List:** box.
- Click on the **Plots...** button.
- On the screen 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.

We will first look at a histogram of the variable, **SCISCORE**. 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 523.5944. This distribution appears to be skewed to the left (negatively skewed).
We will next look at a statistical test to see if this backs up our visual impressions from the histogram.

The Kolmogorov-Smirnov test is used to test the null hypothesis that a set of data comes from a Normal distribution.



Tests of Normality

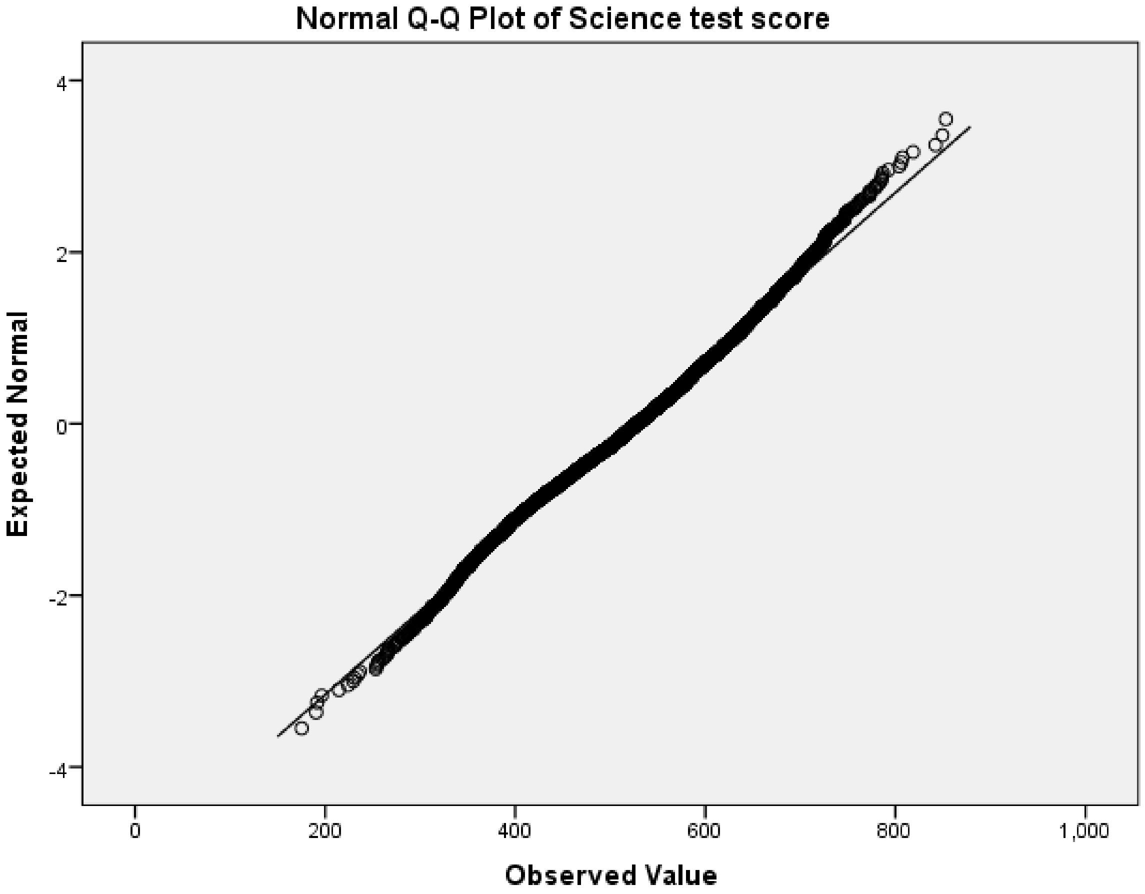
	Kolmogorov-Smirnov[a]		
	Statistic	df	Sig.
Science test score	.025	5194	.000

a. Lilliefors Significance Correction

The Kolmogorov Smirnov test produces 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 .025. This has degrees of freedom which equals the number of data points, namely 5194.

Here we see the p-value provided by SPSS (quoted under Sig. for Kolmogorov-Smirnov) is .000 (reported as $p < .001$). We therefore have significant evidence to reject the null hypothesis that the variable follows a normal distribution.

Although both these statistics tell 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 affected by the number of observations and so only a small discrepancy from normality will be deemed significant for very large sample sizes as may be the case here whilst very large discrepancies will be required to reject the null hypothesis for small sample sizes. It should also be remembered that many parametric statistics are robust to non-normality when sample sizes are very large (employing the Central Limit Theorem), so the implications of non-normality are primarily of interest is designs with smaller sample sizes.
To complete our practical on checking for normality SPSS also produces a Quantile-Quantile (or QQ) plot that 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 we should see 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.
In this example the PISA science test scores do seem to follow a bell-shaped distribution that is broadly symmetric, although scores are clustered more closely around the mean (with too few extremely low or high scores observed) for the distribution to be strictly normal. Given the very large sample size in this application, however, these small departures from normality would not usually prevent us from employing parametric statistical methods.