statstutor

community project

encouraging academics to share statistics support resources

stcp-marshallsamuels-1

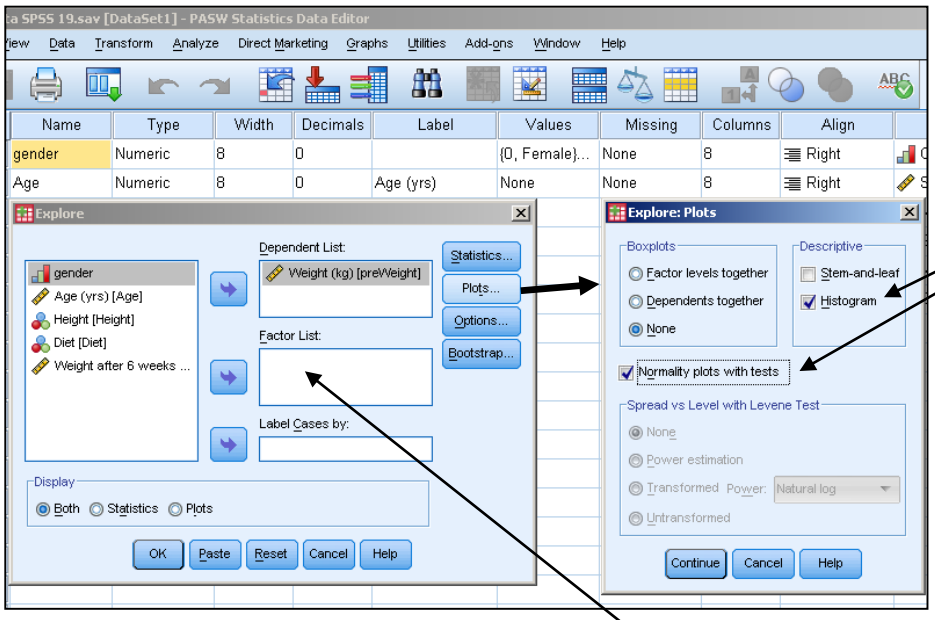
The following resources are associated:
Statistical Hypothesis Testing worksheet and Normality Checking example solutions worksheet
Workshop 7: SPSS and Workshop 8: Parametric Testing

Checking normality for parametric tests

One of the assumptions for most parametric tests to be reliable is that the data is approximately normally distributed. The normal distribution peaks in the middle and is symmetrical about the mean. Data does not need to be perfectly normally distributed for the tests to be reliable.

Checking normality in SPSS

To check if a variable is normally distributed use *Analyze → Descriptive Statistics → Explore*:



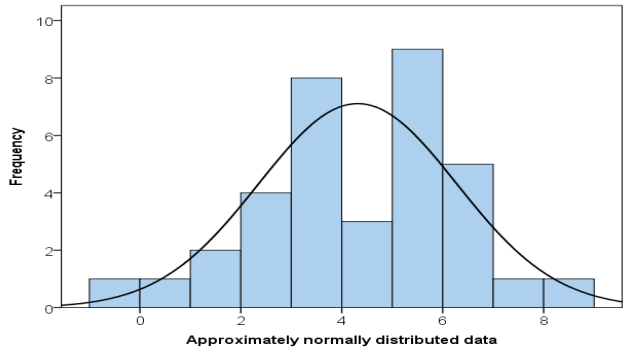
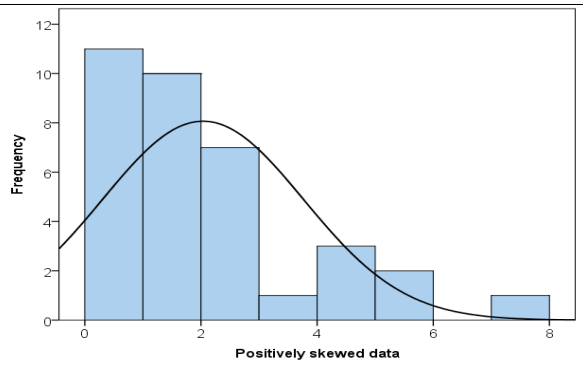
Select the Plots… button and the 'Normality plots with tests' and the 'Histogram' options

When carrying out tests comparing groups, e.g. t-tests, normality checks should be carried out separately for each group: put the appropriate grouping variable in the Factor List

Graphical methods for assessing if data is normally distributed

Plotting a **histogram** of the variable of interest will give an indication of the shape of the distribution. A **normal approximation curve** can also be added by editing the graph. Below are examples of histograms of approximately normally distributed data and heavily **skewed** data with equal sample sizes.

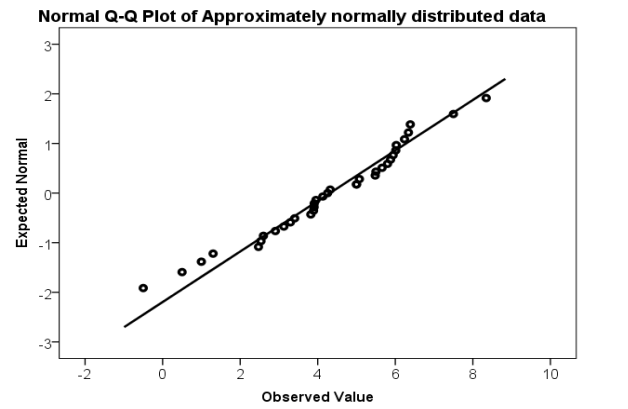
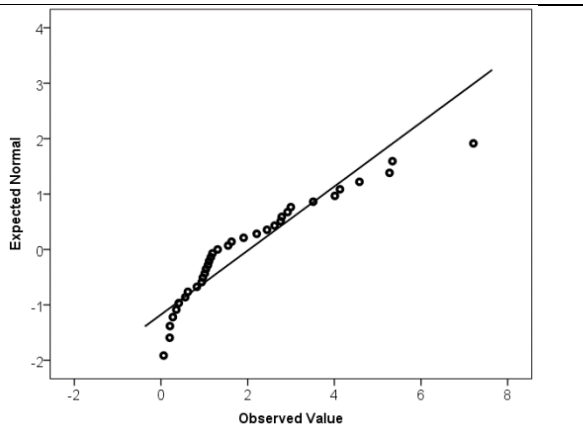
University of Sheffield / Birmingham City University
Reviewer: Cheryl Voake-Jones
University of Bath

Histogram of approximately normally distributed data	Histogram of skewed data
	

Note: Normal curves can be added to histograms by double-clicking on them and using the  button in the Chart Editor window.

It is very unlikely that a histogram of sample data will produce a perfectly smooth normal curve like the one displayed over the histogram, especially if the sample size is small. As long as the data is approximately normally distributed, with a peak in the middle and fairly symmetrical, a parametric test can be used.

The **normal Q-Q plot** is an alternative graphical method of assessing normality to the histogram and is easier to use when there are small sample sizes. The scatter should lie as close to the line as possible with no obvious pattern coming away from the line for the data to be considered normally distributed. Below are the same examples of normally distributed and skewed data.

Q-Q plot of approximately normally distributed data	Q-Q plot of skewed data
	

Note: Measures of skewness and kurtosis and their standard errors are also provided in the Explore output and can also be used in normality checking – see the workshops on SPSS and parametric testing.

Tests for assessing if data is normally distributed

There are also specific methods for testing normality but these should be used in conjunction with either a histogram or a Q-Q plot. The Kolmogorov-Smirnov test and the Shapiro-Wilk's W test whether the underlying distribution is normal. Both tests are sensitive to outliers and are influenced by sample size:

- For smaller samples, non-normality is less likely to be detected but the Shapiro-Wilk test should be preferred as it is usually relatively more sensitive
- For larger samples (i.e. more than one or two hundred), the assumption of normality might be rejected too easily but it is generally less important (see robust exceptions below). Any assessment should also include an evaluation of the normality of histograms or Q-Q plots.

Hypothesis test for a test of normality

Null hypothesis: The data is normally distributed

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Approximately normally distributed data	.095	35	.200*	.975	35	**.582**
Positively skewed data	.177	35	.007	.885	35	**.002**

For both of these examples, the sample size is 35 so the Shapiro-Wilk test should be used. For the skewed data, p = 0.002 so there is very strong evidence of non-normality and a non-parametric test should be used. For the approximately normally distributed data, p = 0.582, so the null hypothesis is retained at the 0.05 level. Therefore, normality can be assumed for this data set and, provided any other test assumptions are satisfied, an appropriate parametric test can be used.

What if the data is not normally distributed?

If the checks suggest that the data is not normally distributed, there are three options:

- **Transform the dependent variable** (repeating the normality checks on the transformed data): Common transformations include taking the log or square root of the dependent variable
- **Use a non-parametric test:** Non-parametric tests are often called distribution free tests and can be used instead of their parametric equivalent
- **Use a parametric test under robust exceptions**: These are conditions when the parametric test can still be used for data which is not normally distributed and are specific to individual parametric tests – see the appropriate test worksheet

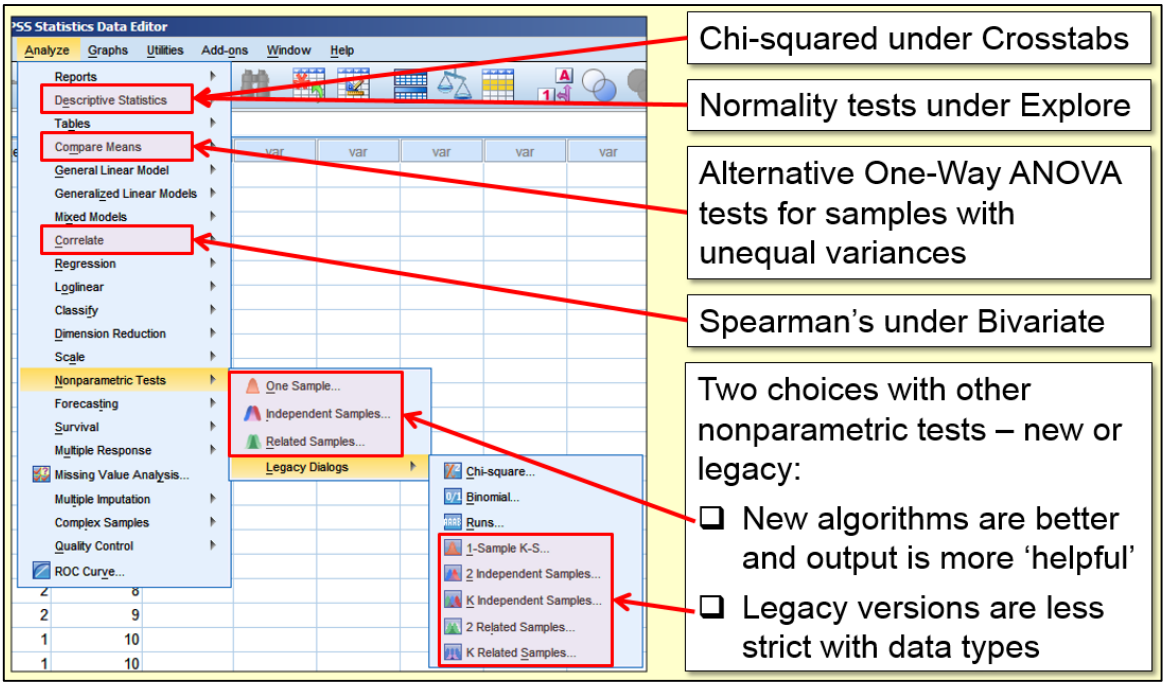
Key non-parametric tests

Parametric test	What to check for normality	Non-parametric test
Independent t-test	Dependent variable	Mann-Whitney test
Paired t-test	Paired differences	Wilcoxon signed rank test
One-way ANOVA	Residuals/dependent variable	Kruskal-Wallis test
Repeated measures ANOVA	Residuals at each time point	Friedman test
Pearson's correlation coefficient	Both variables should be normally distributed	Spearman's correlation coefficient
Simple linear regression	Residuals	N/A

Note: The residuals are the differences between the observed and expected values. They are usually available in the 'Save' options when carrying out a test. However, for ANOVA, residuals can only be obtained through the *Analyze → General Linear Model → Univariate* menu.

Although non-parametric tests require fewer assumptions and can be used on a wider range of data types, parametric tests are preferred because they are more sensitive at detecting differences between samples or an effect of the independent variable on the dependent variable. This means that to detect any given effect at a specified significance level, a larger sample size is required for the non-parametric test than the equivalent parametric test when the data is normally distributed. However, some statisticians argue that non-parametric methods are more appropriate with small sample sizes.

Where to find non-parametric tests in SPSS



Examples (see separate worksheet for solutions)

Example 1

The following standardised residuals were obtained following regression analysis:

-2.12	0.34	1.33	0.97	1.31	1.71	2.64	-0.55	0.47	-0.63
-0.68	-0.48	-0.75	0.50	1.02	-1.63	1.71	-0.46	-0.16	-1.21

Check whether they are approximately normally distributed.

Example 2

A study was carried out to compare whether exercise has an effect on the blood pressure (measure in mm Hg). The blood pressure was measured on 15 people before and after exercising. The results were as follows:

Subject	1	2	3	4	5	6	7	8	9	10
Before	85.1	108.4	79	109.1	97.3	96	102.1	91.2	89.2	100
After	86	70	78.9	69.7	59.4	55	65.6	50.2	60.5	82

Carry out the relevant normality checks and determine which test is appropriate.