statstutor

community project

encouraging academics to share statistics support resources

stcp-marshall-normalityX

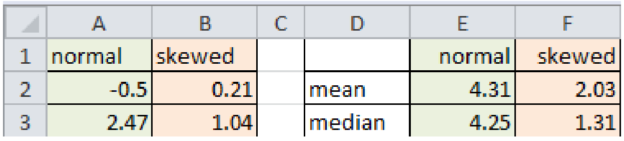
The following resources are associated:
Excel data file 'NormX' and 'Additional toolpaks in Excel'

Checking normality in Excel

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. There are several methods for checking normality which include graphical methods, tests for normality and assessing skewness figures. It is not necessary to use all the methods; just select one or two.

Excel's options are limited for methods for checking normality. A histogram can be constructed using the standard 'Data analysis toolpak' add in package. You could use the 'Real-statistics' add in package, http://www.real-statistics.com/tests-normality-and-symmetry/ or an online calculator tool such as the webpage https://www.ai-therapy.com/psychology-statistics/distributions/normal which contain more options. This sheet discusses the methods for checking normality and how to obtain the output using the three methods.

Data for this sheet: Open the Excel file 'NormX' which contains two columns of data. The two columns of data are not related and they will be checked for normality separately. Comparing the means (`=AVERAGE(A:A)`) and medians (`=MEDIAN(A:A)`) can help assess normality as for skewed data they are very different. Here they are similar for the 'normal' data and quite different for the 'skewed' data.

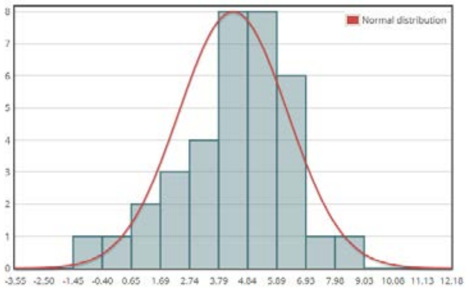


	A	B	C	D	E	F
1	normal	skewed			normal	skewed
2	-0.5	0.21		mean	4.31	2.03
3	2.47	1.04		median	4.25	1.31

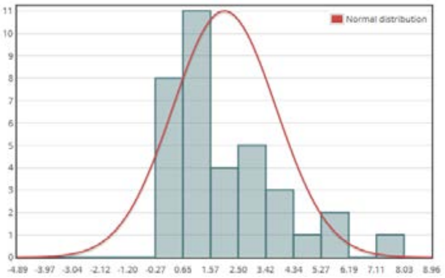
Graphical methods for assessing if data is normally distributed

The most commonly used method is the histogram. Plotting a **histogram** of the variable of interest will give an indication of the shape of the distribution and is the most commonly used. 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.

Histogram of approximately normally distributed data



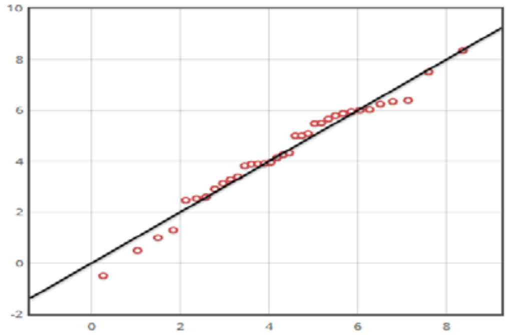
Histogram of skewed data



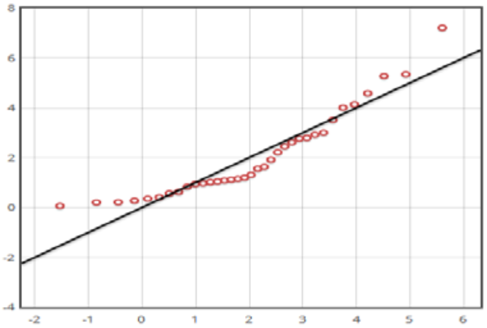
It is very unlikely that a histogram of sample data will produce a perfectly smooth normal curve like the one displayed over the histogram, especially if the sample size is small. As long as the data is approximately normally distributed, with a peak in the middle and fairly symmetrical, the assumption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

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



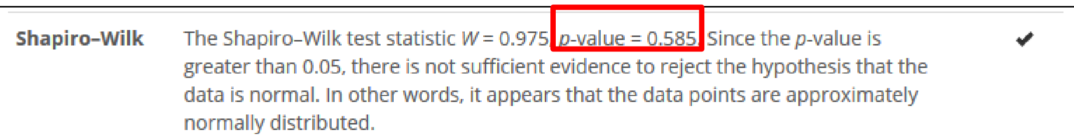
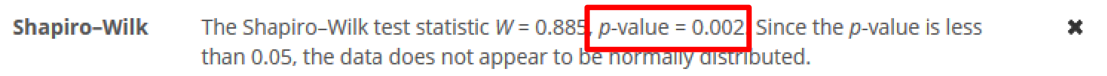
Tests for assessing if data is normally distributed

The Kolmogorov-Smirnov test and the Shapiro-Wilk's W test are two specific methods for testing normality of data but these should be used in conjunction with either a histogram or a Q-Q plot as 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 generally more sensitive
- For larger samples (i.e. more than one hundred), the normality tests are overly conservative and the assumption of normality might be rejected too easily.

Null hypothesis for test of normality: The data is normally distributed.

If the p-value is under 0.05, the null is rejected and there is significant evidence of non-normal data.

normal	 **Shapiro–Wilk** The Shapiro–Wilk test statistic $W = 0.975$ *p*-value = 0.585 Since the *p*-value is greater than 0.05, there is not sufficient evidence to reject the hypothesis that the data is normal. In other words, it appears that the data points are approximately normally distributed. ✔
skewed	 **Shapiro–Wilk** The Shapiro–Wilk test statistic $W = 0.885$ *p*-value = 0.002 Since the *p*-value is less than 0.05, the data does not appear to be normally distributed. ✖

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 suggesting strong evidence of non-normality. For the approximately normally distributed data, p = 0.585, so normality can be assumed 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 two 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.

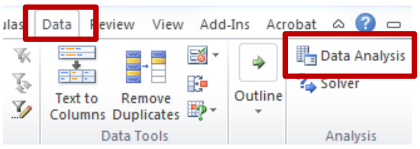
Key non-parametric tests

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

Note: The residuals are the differences between the observed and expected values.

Excel will not perform non-parametric tests even with the data analysis toolpak add in. Both AI-therapy and the 'Real-Statistics' add in will though.

Using the data analysis toolpak



Although this is a standard Excel add in package, you will need to add it via *File > Options > Add-Ins* (see '*Additional toolpaks in Excel*' sheet for details). The '*Data Analysis'* button then appears on the **Data** menu.

Before creating the histogram decide on the cut-offs where the bars of the histogram meet (called bins). It is better to specify these yourself as Excel chooses badly! Calculate the minimum and maximum and create 5 -10 equally spaced categories between the minimum and a value above the maximum.

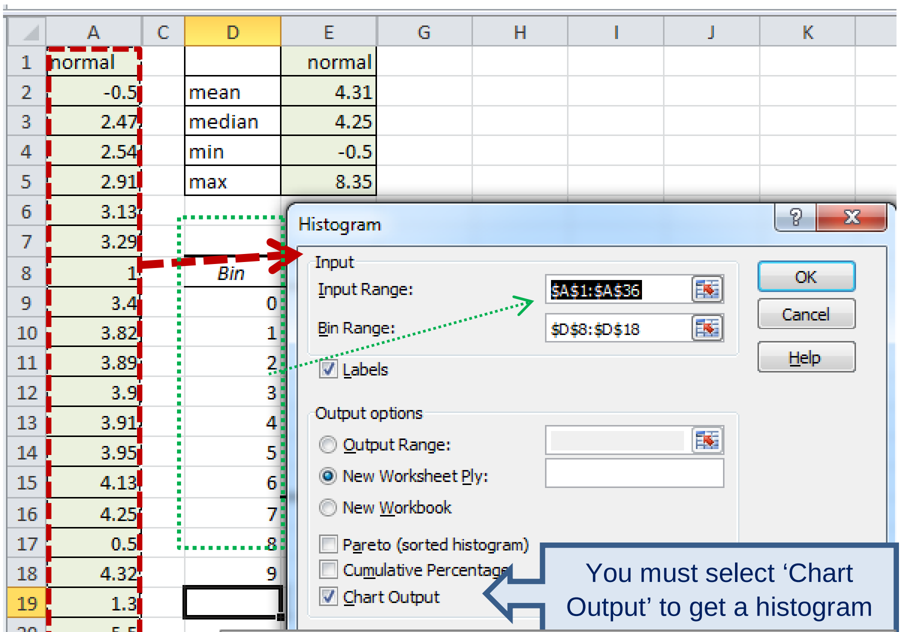
Go to *Data > Data Analysis* and select the 'Histogram' option from the first menu which opens the Histogram dialog box.

Input range: Select the data you wish to plot in a histogram.

Bin Range: Select the upper cut off's you defined. **Labels**: Tick this if the range includes the data name.

Output range: Where you want the output to go.

Select the '*Chart Output*' option and click **OK**.



	A	C	D	E
1	normal			normal
2	-0.5		mean	4.31
3	2.47		median	4.25
4	2.54		min	-0.5
5	2.91		max	8.35
6	3.13			
7	3.29			
8	1		Bin	
9	3.4		0	
10	3.82		1	
11	3.89		2	
12	3.9		3	
13	3.91		4	
14	3.95		5	
15	4.13		6	
16	4.25		7	
17	0.5		8	
18	4.32		9	
19	1.3			
20	5.5			

Histogram

Input

Input Range: $A$1:$A$36

Bin Range: $D$8:$D$18

☑ Labels

Output options

○ Output Range:

● New Worksheet Ply:

○ New Workbook

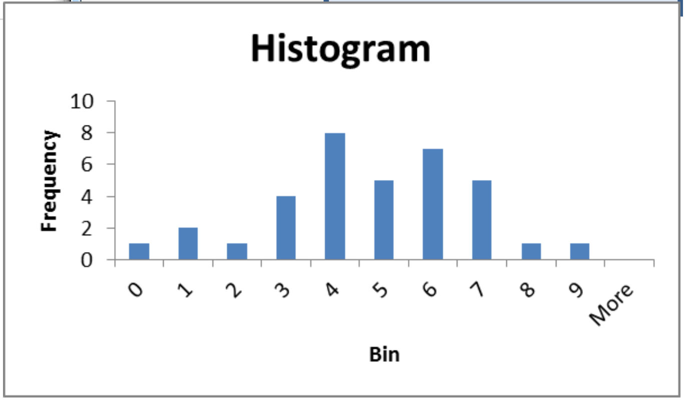
□ Pareto (sorted histogram)

□ Cumulative Percentage

☑ Chart Output

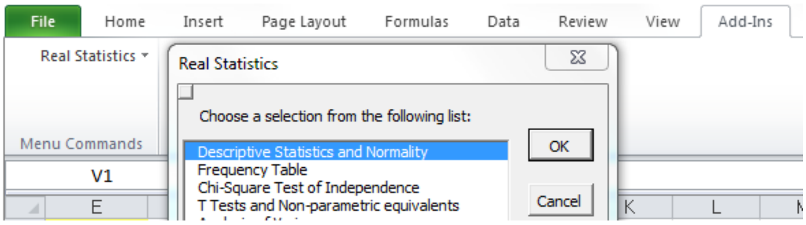
You must select 'Chart Output' to get a histogram

The frequencies for each of your specified 'bins' appear along with this chart. Excel does not appear to know what a histogram is as the bars should meet and the 'bin' labels should separate the bars! However for the purposes of assessing normality, the data can be seen to be approximately normal. If it's hard to tell, changing the number of bins can help so choose different bin cut-offs and follow the procedure again.



Using the real-statistics add in package

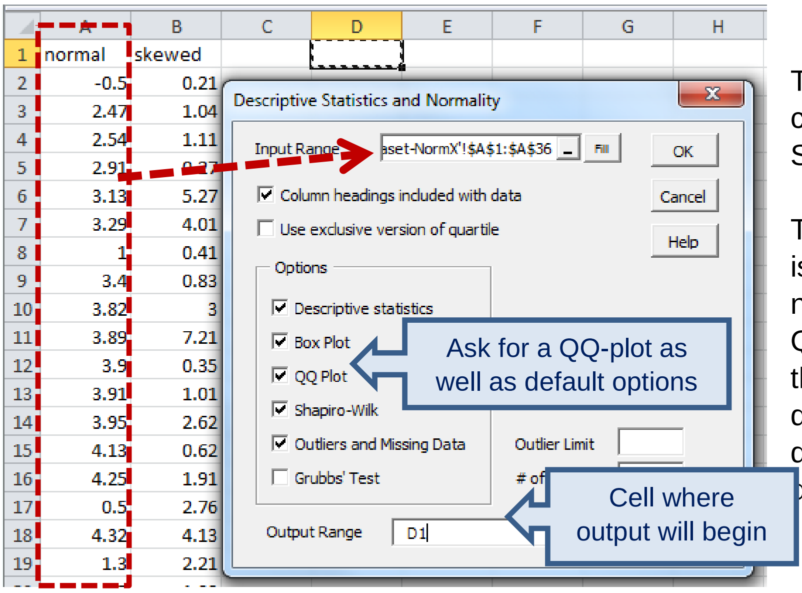
You must download this package (see ‘Additional toolpaks in Excel’ sheet for details). When downloaded, you can access the menu from the ‘Add-Ins’ tab.



Choose a selection from the following list:

- Descriptive Statistics and Normality
- Frequency Table
- Chi-Square Test of Independence
- T Tests and Non-parametric equivalents

Select ‘*Descriptive Statistics and Normality’* and click **OK** to open the following dialog box.



	A	B
1	normal	skewed
2	-0.5	0.21
3	2.47	1.04
4	2.54	1.11
5	2.91	0.27
6	3.13	5.27
7	3.29	4.01
8	1	0.41
9	3.4	0.83
10	3.82	3
11	3.89	7.21
12	3.9	0.35
13	3.91	1.01
14	3.95	2.62
15	4.13	0.62
16	4.25	1.91
17	0.5	2.76
18	4.32	4.13
19	1.3	2.21

Descriptive Statistics and Normality

Input Range: aset-NormX'!$A$1:$A$36

☑ Column headings included with data

☐ Use exclusive version of quartile

Options

☑ Descriptive statistics

☑ Box Plot

☑ QQ Plot

☑ Shapiro-Wilk

☑ Outliers and Missing Data

☐ Grubbs' Test

Outlier Limit

Output Range: D1

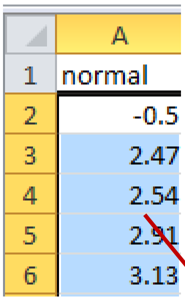
Ask for a QQ-plot as well as default options

Cell where output will begin

There’s a lot of output but concentrate on the QQ-plot and the Shapiro-Wilk to assess normality.

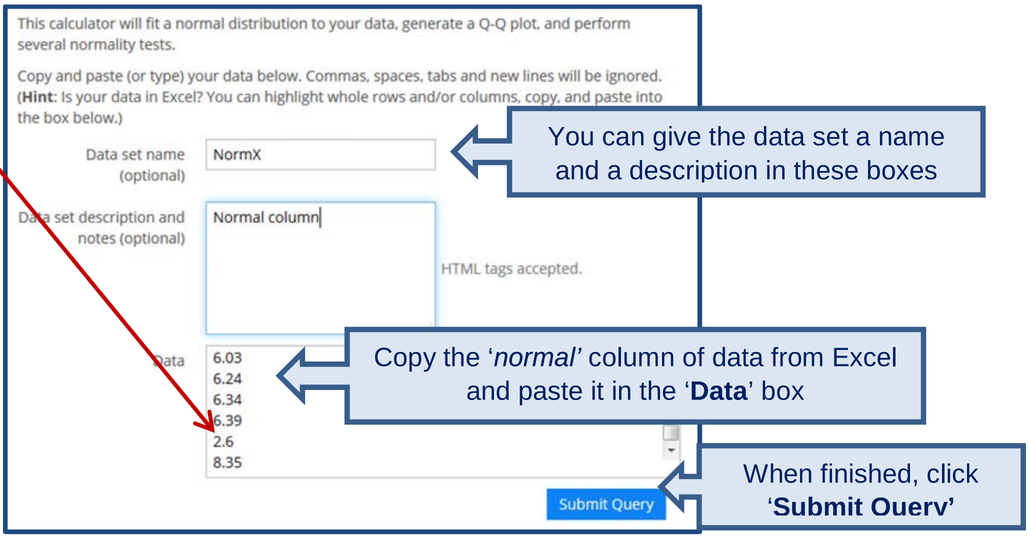
The Shapiro-Wilk p-value of 0.5847 is much bigger than 0.05 so normality can be assumed. In the QQ-plot the data points are close to the line of the perfect normal distribution also suggesting that the data are approximately normally distributed.

Using AI-therapy to check normality



	A
1	normal
2	-0.5
3	2.47
4	2.54
5	2.91
6	3.13

Copy the data from the ‘*normal’* column in the Excel file and add it to the ‘**Data**’ section of the webpage https://www.ai-therapy.com/psychology-statistics/distributions/normal.



This calculator will fit a normal distribution to your data, generate a Q-Q plot, and perform several normality tests.

Copy and paste (or type) your data below. Commas, spaces, tabs and new lines will be ignored. (**Hint:** Is your data in Excel? You can highlight whole rows and/or columns, copy, and paste into the box below.)

Data set name (optional): NormX

Data set description and notes (optional): Normal column

HTML tags accepted.

Data:
6.03
6.24
6.34
6.39
2.6
8.35

You can give the data set a name and a description in these boxes

Copy the ‘*normal’* column of data from Excel and paste it in the ‘**Data**’ box

When finished, click ‘**Submit Query**’

If your data is skewed and a non-parametric test is needed, comparisons of two sets of data can be accessed at https://www.ai-therapy.com/psychology-statistics/hypothesis-testing/two-samples and for comparisons of more than two sets at: https://www.ai-therapy.com/psychology-statistics/hypothesis-testing/anova