statstutor

community project

encouraging academics to share statistics support resources

stcp-karadimitriou-normalR

The following resources are associated:
Statistical Hypothesis Testing and Normality Checking in R solutions, csv and script files

Checking normality for parametric tests in R

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 R

Open the 'normality checking in R data.csv' dataset which contains a column of normally distributed data (normal) and a column of skewed data (skewed)and call it normR.
You will need to change the command depending on where you have saved the file.

```
normR<-read.csv("D:\\normality checking in R data.csv",header=T,sep=",")
```

Tell R to use this dataset until further notice using `attach(normR)` which means that the variable name '`normal`' can be used instead of `normR$normal`.

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 **density curve** smoothes out the histogram and can be added to the graph.

First, produce the histogram for the normally distributed data (normal) and add a density curve.

```
hist(normal,probability=T, main="Histogram of normal
data",xlab="Approximately normally distributed data")
lines(density(normal),col=2)
```

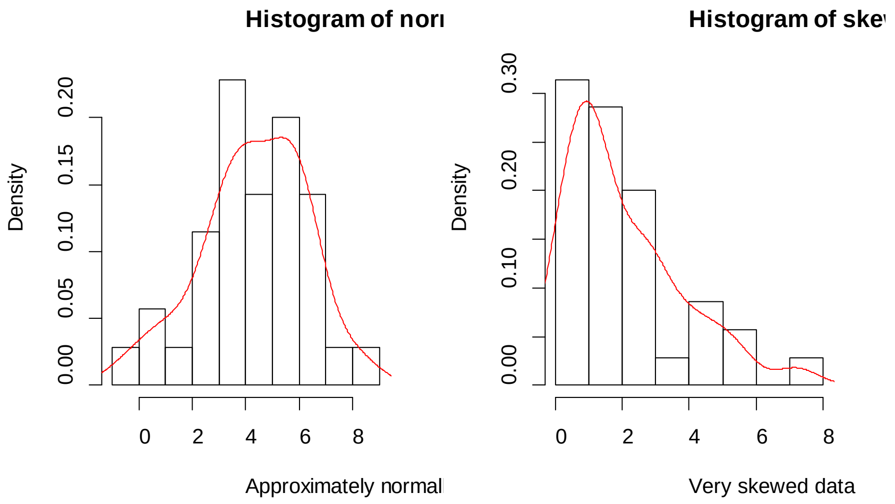
Then produce a histogram of the skewed data with a density curve

```
hist(skewed,probability=T, main="Histogram of skewed data",xlab="Very
skewed data")
lines(density(skewed),col=2)
```

It is very unlikely that a histogram of sample data will produce a perfectly smooth normal curve 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 shape of the histogram varies depending on the number of bars used so sometimes it helps to change the number of bars in the histogram by specifying the number of breaks (n) between bars using `hist(…., breaks = n)`.

Based on material provided by Ellen Marshall (University of Sheffield) and Peter Samuels (Birmingham City University)



The histogram on the left is approximately distributed as it peaks roughly in the middle and the second data set is clearly skewed so no parametric test should be carried out using the data.

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 compares the data to a perfect normal distribution. 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.

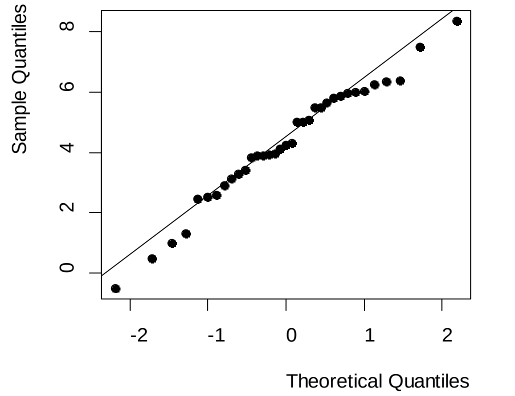
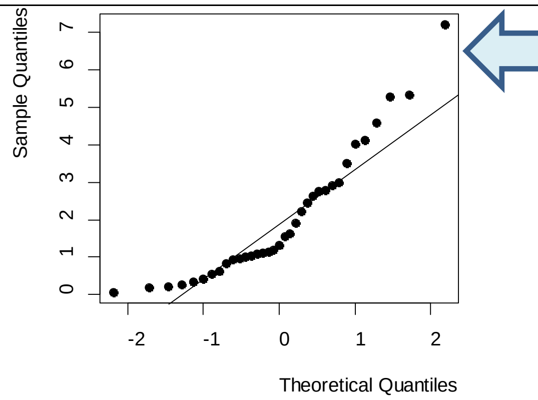
Draw the qq-plot of the normally distributed data using pch=19 to produce solid circles.

```
qqnorm(normal,main="QQ plot of normal data",pch=19)
```

Add a line where x = y to help assess how closely the scatter fits the line.

```
qqline(normal)
```

Repeat for the skewed data.

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

The scatter of skewed data tends to form curves moving away from the line at the ends

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 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 (see robust exceptions below).

Any assessment should also include an evaluation of the normality of histograms or Q-Q plots and these are more appropriate for assessing normality in larger samples.

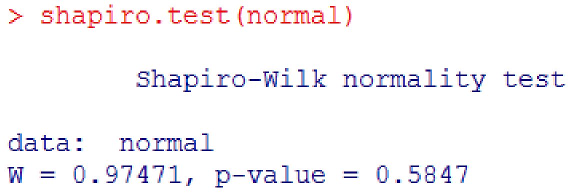
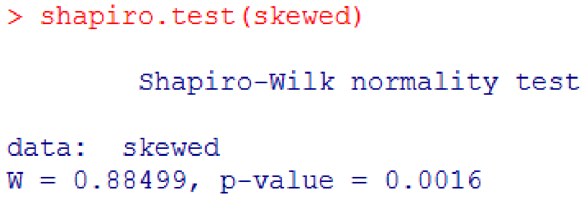
Hypothesis test for a test of normality

Null hypothesis: The data is normally distributed. If p> 0.05, normality can be assumed.

For both of these examples, the sample size is 35 so the Shapiro-Wilk test should be used.

```
shapiro.test(normal)
shapiro.test(skewed)
```

Shapiro-Wilk test of approximately normally distributed data	Shapiro-Wilk test of skewed data


`> shapiro.test(normal)`
`        Shapiro-Wilk normality test`
`data:  normal`
`W = 0.97471, p-value = 0.5847`
	
	`> shapiro.test(skewed)`
	`        Shapiro-Wilk normality test`
	`data:  skewed`
	`W = 0.88499, p-value = 0.0016`

For the skewed data, p = 0.0016 suggesting strong evidence of non-normality and a non-parametric test should be used. For the approximately normally distributed data, p = 0.5847 so the null hypothesis is retained at the 95% level of significance. 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

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.

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.

Commands for non-parametric tests in R

y = dependent variable and x = Independent variable

Test	Command in R	Comments
Mann – Whitney	*`wilcox.test(y~x)`*	y is continuous, x is binary
Wilcoxon Signed Rank Test	*`wilcox.test(y1,y2,paired=T)`*	y1 and y2 are continuous
Kruskal Wallis	*`kruskal.test(y~x)`*	y is continuous, x is categorical
Friedman Test	*`friedman.test(y~x|B)`*	y are the data, x is a grouping factor and B a blocking factor
Spearman's	*`cor.test(x,y,method='spearman')`*	one variable can also be used