Tutorials and worked examples for simulation,
curve fitting, statistical analysis, and plotting.
http://www.simfit.org.uk

Of the many tests for normality, the Shapiro-Wilks test is usually recommended to check if a sample is consistent with a normal distribution, and it leads naturally to the normal scores plot which is a convenient technique to examine normality graphically.

To be precise, the user has a sample (i.e. vector X) of n observations

$$X = (x_1, x_2, \ldots, x_n)$$

and wishes to test if these numbers are consistent with a normal distribution, where the parameters have been previously estimated with great precision from an independent very large sample, or are known due to further information. Preferably the data should cover a wide range and n should not be too small, say $n > 20$?

There are many statistical methods provided by SimFIT program **normal** for doing this as now summarized.

- **Kolmogorov-Smirnov**
  This only has advantages in the case when both parameters are known, and not estimated from the sample.
- **One sample t**
  This always uses the sample variance, and also is best when the true mean is known in advance.
- **Chi-square**
  This is also a rather poor test, especially if the expected values are estimated from the sample.
- **Shapiro-Wilks**
  This is now generally thought to be the best all purpose test where parameters are estimated from the sample, but it does require intensive computation which can limit the maximum value of n, e.g. $n \leq 5000$ in SimFIT.

In addition there are the following graphical methods.

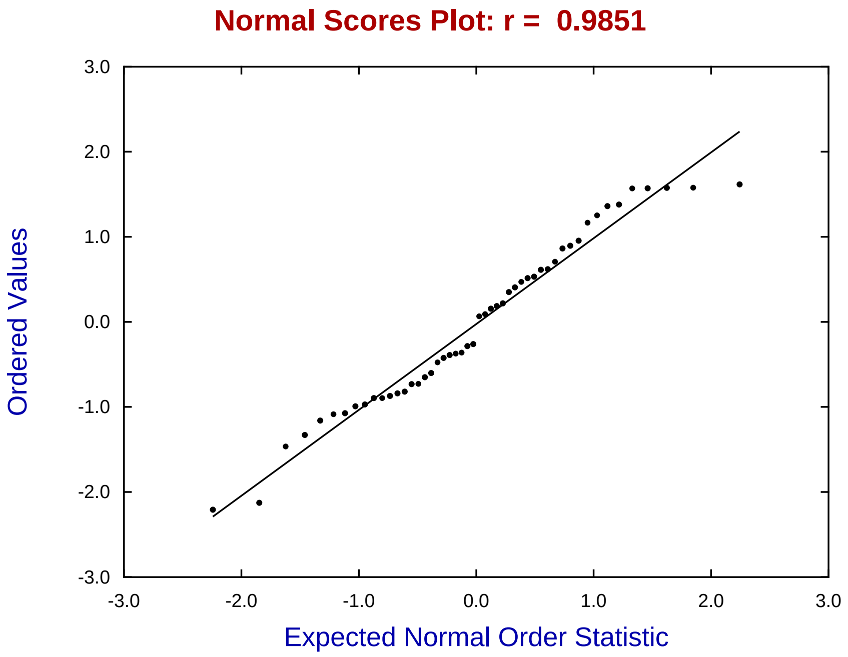
- **Histogram**
  This can easily be done, but SimFIT can also be used to rationalize the situation by analyzing the data after transformation to $U(0, 1)$, as it is somewhat easier to detect deviations of a histogram from the case where all cells have equal expected frequency from one where a bell-shaped curve is anticipated.
- **Cumulative distribution**
  This is much better than a histogram, as histogram shape depends on the number of bins whereas the sample cumulative distribution is of fixed shape.
- **Normal scores**
  The n values that can be calculated to divide a standard normal cumulative distribution into $n + 1$ sections, each of area $1/(n + 1)$, are referred to as normal scores. A plot of sample scores against normal scores should be close to a straight line, and is widely recognized as the best plot for detecting departure from normality. A variant is the half normal plot, where negative values are changed in sign, and this is usually preferred for testing that residuals from regression do not differ too widely from normality.

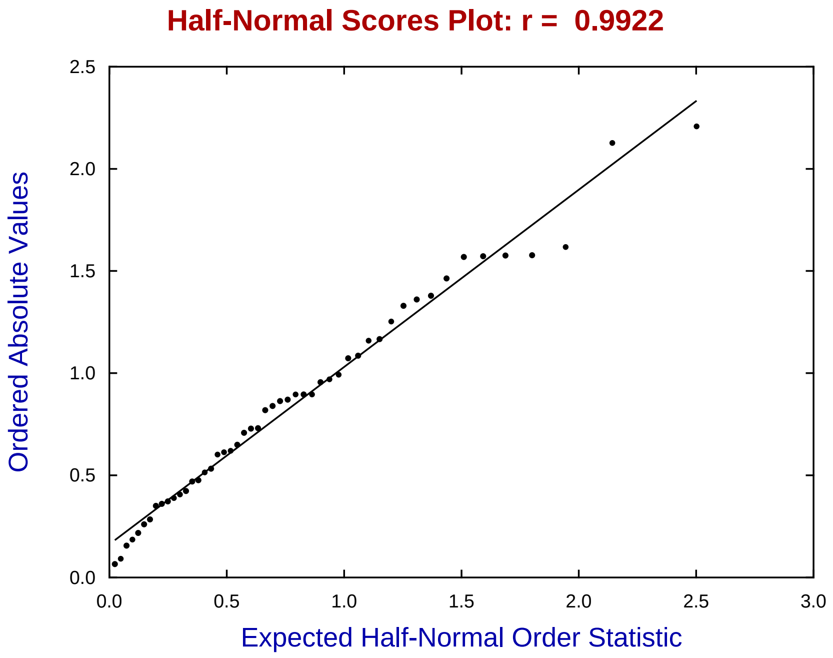
In the case of mixtures of normal distributions there are dedicated statistical tests, but SimFIT program **qnfit** also provides the ability to fit histograms or cumulative distributions if the sample size is very large, and the distributions well separated. The advantage here is that the SimFIT graphical deconvolution technique can be used to display how the distribution is made up from sums of normal distributions.

Here is the conclusion from program **normal** and the test data `normal.tf1` provided, which establishes that a normal distribution cannot be rejected, as the Shapiro-Wilks test statistic is $W = 0.9627$ with $p = 0.1153$, i.e. close to $W = 1$, which indicates strong correlation between the sample scores and normal scores.

Normal distribution test

Data: Test file `normal.tf1` with 50 pseudo-random numbers	
Shapiro-Wilks statistic W	0.9627
Significance level for W	0.1153
Conclusion: *Tentatively accept normality*	


Normal Scores Plot: r = 0.9851
Ordered Values
Expected Normal Order Statistic


Half-Normal Scores Plot: r = 0.9922
Ordered Absolute Values
Expected Half-Normal Order Statistic