Standard Statistical Tests for normality

[Nematrian website page: TestsForNormality, © Nematrian 2015]

Standard statistical tests for identifying whether an observed sample is likely not to have come from a normal distribution include:

(a) Testing the extent to which the skew of the sample is non-zero, see e.g. Confidence level for skew for large sample normal distribution

(b) Testing the extent to which the (excess) kurtosis of the sample is non-zero, see e.g. Confidence level for (excess) kurtosis for large sample normal distribution

(c) The Jarque-Bera test which simultaneously tests the extent to which the skew and (excess) kurtosis of the sample are non-zero

(d) The Shapiro-Wilk test

(e) The Anderson-Darling test*

(f) The Kolmogorov-Smirnov test*

(g) The Cramer-von-Mises test*

* These tests can be used with any distributional form, i.e. they are not limited to testing merely for non-normality. Their test statistics depend on the sample data through terms that depend merely on order statistics and then only on how these map onto the hypothesised cumulative distribution function (i.e. if the sample is $x_1, x_2, \dots, x_n$ then merely through $F_1, F_2, \dots, F_n$ where $F_i = F\left(x_{(i)}\right)$ where $F(.)$ is the cumulative distribution function and $x_{(i)}$ is the i'th order statistic, i.e. the i'th smallest value in the sample). In contrast (a) to (c) are parametric, with their test statistics depending merely on specific moments of the distribution (here the skew and kurtosis and the two combined respectively). (d) depends on both order and parametric elements.

All of the above tests, as conventionally formulated, have the disadvantage that they give 'equal' weight to every observation. A possible exception is the Kolmogorov-Smirnov test, which merely refers to the single (ordered) observation that appears to exhibit the greatest deviation from where we might have expected it lie.

As explained in Kemp (2009), this generally means that they indicate mainly whether a sample appears to be deviating from normality in the middle of the distribution rather than whether it appears to be deviating from normality in its tails. Loosely speaking, this is because there are far more observations in the middle of a normal distribution than in its tail. We illustrate this with (b). Consider the proportion of observations that are in the tails of a normal distribution. Only approximately 1 in 1.7 million observations from a normal distribution should be further away from the (sample) mean than 5 standard deviations. Each one in isolation might on average contribute at least 625 times as much to the computation of kurtosis as an observation that is just one standard deviation away from the (sample) mean (since 5 x 5 x 5 x 5 = 625), but, because there are so few observations this far into the tail, they in aggregate have little impact on the overall kurtosis of the distribution.

Better, if we are interested merely in testing for deviation from normality in a *part* of a distributional form is to modify the above methodologies so that they depend just on data from the relevant part of the observed distributional form. For example, we might wish to focus on the worst 10% of outcomes. We would then estimate the mean and standard deviation of a normal distribution that would have its worst 10% of outcomes as close as possible to those actually observed, and we would then apply a modified test statistic that referred merely to the observations in the part of the distributional form in which we are interested. In general, we can view this modification as involving giving different weights w_i to the different $x_{(i)}$. To calculate critical values for such statistics (and therefore whether or not to reject the null hypothesis of normality) generally requires Monte Carlo simulation techniques, given the wide range of possible weighting schemas that could be used.

The Shapiro-Wilk test

[ShapiroWilk]

The **Shapiro-Wilk test** tests the null hypothesis that a sample, $x_1, x_2, \ldots, x_n$ comes from a Normally distributed population.

It was published in 1965, see Shapiro and Wilk (1965), and involves the following test statistic:

$$W = \frac{\left(\sum_{i=1}^{n} a_i x_{(i)}\right)^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

where $x_{(i)}$ is the i'th order statistic, i.e. the i'th smallest value in the sample and $\bar{x}, a_i, \mathbf{a}, \mathbf{m}$ are as follows, the m_i are the expected values of the order statistics of n i.i.d. random variables sampled from the standard Normal distribution and $\mathbf{V}$ is the covariance matrix of these order statistics:

$$\bar{x} = (x_1 + x_2 + \cdots + x_n)/n$$
$$\mathbf{a} = (a_1, a_2, \ldots, a_n) = \frac{\mathbf{m}^T \mathbf{V}^{-1}}{\mathbf{m}^T \mathbf{V}^{-1} \mathbf{V}^{-1} \mathbf{m}}$$
$$\mathbf{m} = (m_1, m_2, \ldots, m_n)^T$$

The Anderson-Darling test

[AndersonDarling]

The **Anderson-Darling test** tests the null hypothesis that a sample, $x_1, x_2, \ldots, x_n$ comes from a pre-specified population distribution (or a pre-specified family of such distributions).

In its basic form, the test assumes that there are no parameters to be estimated for the distribution being tested, in which case the test and its set of critical values are distribution-free.

However, it is most commonly used where a family of distributions are being tested. For example, we might be testing whether the sample comes from a Normal distribution but without specifying in advance the mean and standard deviation of that distribution. It then becomes necessary to estimate the parameters on which the particular distribution depends and this needs to be taken into account by adjusting the test statistic and/or its critical values.

The test was published in 1952, see Anderson and Darling (1952). It is based on the observation that if the data does come from the hypothesised distribution then the data can be transformed to what should be a uniform distribution. The transformed data can then be tested for uniformity with a distance test, see e.g. Shapiro (1980).

In its basic form, it involves the following test statistic, A, where we are testing the null hypothesis that the data is coming from a distribution with cumulative distribution function (cdf) F:

$$A^2 = -n - S$$

where $x_{(i)}$ is the i'th order statistic, i.e. the i'th smallest value in the sample and

$$S = \sum_{i=1}^{n} \frac{2i-1}{n} \left(\log F\left(x_{(i)}\right) + \log\left(1 - F\left(x_{(n+1-i)}\right)\right)\right)$$

Essentially the same approach can be used when testing whether data comes from a pre-specified *family* of distributions. However, the statistic must then be compared against critical values appropriate to the family in question and dependent also on the method used for parameter estimation.

A '*K-sample*' Anderson-Darling test can be used to test whether several samples appear to be coming from a single distribution, without the need to specify in advance what the distributional form might be. Sholz and Stephens (1987) indicate how this basic approach can be used to test whether a number of random samples with possibly different sample sizes are coming from the same underlying distribution, where this distribution is unspecified in advance.

The Kolmogorov-Smirnov test

[KolmogorovSmirnov]

The **Kolmogorov-Smirnov test** tests the null hypothesis that a sample, $x_1, x_2, \ldots, x_n$ comes from a pre-specified population distribution (or a pre-specified family of such distributions).

In its basic form, the test assumes that there are no parameters to be estimated for the distribution being tested, in which case the test and its set of critical values are distribution-free.

However, it is most commonly used where a family of distributions are being tested. For example, we might be testing whether the sample comes from a Normal distribution but without specifying in advance the mean and standard deviation of that distribution. It then becomes necessary to estimate the parameters on which the particular distribution depends and this needs to be taken into account by adjusting the test statistic and/or its critical values.

In its basic form, it involves the following test statistic, D_n, where we are testing the null hypothesis that the data is coming from a distribution with cumulative distribution function (cdf) F:

$$D_n = \sup_i \left| F_n\left(x_{(i)}\right) - F\left(x_{(i)}\right) \right|$$

where $x_{(i)}$ is the i'th order statistic, i.e. the i'th smallest value in the sample, $\sup S$ is the supremum (i.e. largest value) of the set S and $F_n\left(x_{(i)}\right)$ is the empirical distribution function, defined in the Wikipedia entry on this test as, in effect i/n, but perhaps more naturally defined as $(2i-1)/(2n)$, see the Cramer-von-Mises test.

Essentially the same approach can be used when testing whether data comes from a pre-specified *family* of distributions. However, the statistic must then be compared against critical values appropriate to the family in question and dependent also on the method used for parameter estimation.

The test can also be inverted to give confidence limits on $F(x)$ itself and a variant can be used to test whether two (or more) underlying one-dimensional distributions differ. Generalising the statistic to more than one dimension is also possible but complicated.

The Cramér-von Mises test

[CramervonMises]

The **Cramér-von Mises test** tests the null hypothesis that a sample, $x_1, x_2, \ldots, x_n$ comes from a pre-specified population distribution (or a pre-specified family of such distributions).

In its basic form, the test assumes that there are no parameters to be estimated for the distribution being tested, in which case the test and its set of critical values are distribution-free.

However, it is most commonly used where a family of distributions are being tested. For example, we might be testing whether the sample comes from a Normal distribution but without specifying in advance the mean and standard deviation of that distribution. It then becomes necessary to estimate the parameters on which the particular distribution depends and this needs to be taken into account by adjusting the test statistic and/or its critical values.

In its basic form, it involves the following test statistic, T, where we are testing the null hypothesis that the data is coming from a distribution with cumulative distribution function (cdf) F:

$$T = n\omega^2 = \frac{1}{12n} + \sum_{i=1}^{n} \left(\frac{2i-1}{2n} - F\left(x_{(i)}\right) \right)^2$$

where $x_{(i)}$ is the i'th order statistic, i.e. the i'th smallest value in the sample. If the empirical distribution function $F_n\left(x_{(i)}\right)$ is defined as $(2i-1)/(2n)$ then the statistic can be seen to be (up to a constant for any given n) similar to the statistic used in the Kolmogorov-Smirnov test, but using the mean squared deviation rather than the supremum of $F_n\left(x_{(i)}\right) - F\left(x_{(i)}\right)$.

Essentially the same approach can be used when testing whether data comes from a pre-specified *family* of distributions. However, the statistic must then be compared against critical values appropriate to the family in question and dependent also on the method used for parameter estimation.

Like the Kolmogorov-Smirnov test, the test can also in principle be inverted to give confidence limits on $F(x)$ itself and a variant can be used to test whether two (or more) underlying one-dimensional distributions differ. Generalising the statistic to more than one dimension is possible but complicated.