

## Chapter 194

# Normality Tests

---

## Introduction

This procedure provides seven tests of data normality. If the variable is normally distributed, you can use parametric statistics that are based on this assumption.

If a variable fails a normality test, it is critical to look at the histogram and the normal probability plot to see if an outlier or a small subset of outliers has caused the non-normality. If there are no outliers, you might try a transformation (such as, the log or square root) to make the data normal. If a transformation is not a viable alternative, nonparametric methods that do not require normality may be used.

Always remember that a reasonably large sample size is required to detect departures from normality. Only extreme types of non-normality can be detected with samples less than fifty observations.

There is a common misconception that a histogram is always a valid graphical tool for assessing normality. Since there are many subjective choices that must be made in constructing a histogram, and since histograms generally need large sample sizes to display an accurate picture of normality, preference should be given to other graphical displays such as the box plot, the density trace, and the normal probability plot.

Normality tests generally have small statistical power (probability of detecting non-normal data) unless the sample sizes are at least over 100.

---

## Technical Details

This section provides details of the seven normality tests that are available.

### Shapiro-Wilk $W$ Test

This test for normality has been found to be the most powerful test in most situations. It is the ratio of two estimates of the variance of a normal distribution based on a random sample of  $n$  observations. The numerator is proportional to the square of the best linear estimator of the standard deviation. The denominator is the sum of squares of the observations about the sample mean. The test statistic  $W$  may be written as the square of the Pearson correlation coefficient between the ordered observations and a set of weights which are used to calculate the numerator. Since these weights are asymptotically proportional to the corresponding expected normal order statistics,  $W$  is roughly a measure of the straightness of the normal quantile-quantile plot. Hence, the closer  $W$  is to one, the more normal the sample is.

The probability values for  $W$  are valid for sample sizes greater than 3. The test was developed by Shapiro and Wilk (1965) for sample sizes up to 20. **NCSS** uses the approximations suggested by Royston (1992) and Royston (1995) which allow unlimited sample sizes. Note that Royston only checked the results for sample sizes up to 5000, but indicated that he saw no reason larger sample sizes should not work.  $W$  may not be as powerful as other tests when ties occur in your data. The test is not calculated when a frequency variable is specified.

### Anderson-Darling Test

This test, developed by Anderson and Darling (1954), is a popular among those tests that are based on EDF statistics. In some situations, it has been found to be as powerful as the Shapiro-Wilk test.

Note that this test is not calculated when a frequency variable is specified.

## Normality Tests

### Martinez-Iglewicz Test

This test for normality, developed by Martinez and Iglewicz (1981), is based on the median and a robust estimator of dispersion. The authors have shown that this test is very powerful for heavy-tailed symmetric distributions as well as a variety of other situations. A value of the test statistic that is close to one indicates that the distribution is normal. This test is recommended for exploratory data analysis by Hoaglin (1983). The formula for this test is:

$$I = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{(n-1)s_{bi}^2}$$

where  $s_{bi}^2$  is a biweight estimator of scale.

### Kolmogorov-Smirnov Test

This test for normality is based on the maximum difference between the observed distribution and expected cumulative-normal distribution. Since it uses the sample mean and standard deviation to calculate the expected normal distribution, the Lilliefors' adjustment is used. The smaller the maximum difference the more likely that the distribution is normal.

This test has been shown to be less powerful than the other tests in most situations. It is included because of its historical popularity.

### D'Agostino Skewness Test

D'Agostino (1990) describes a normality test based on the skewness coefficient,  $\sqrt{b_1}$ . Recall that because the normal distribution is symmetrical,  $\sqrt{b_1}$  is equal to zero for normal data. Hence, a test can be developed to determine if the value of  $\sqrt{b_1}$  is significantly different from zero. If it is, the data are obviously non-normal. The statistic,  $z_s$ , is, under the null hypothesis of normality, approximately normally distributed. The computation of this statistic, which is restricted to sample sizes  $n > 8$ , is

$$z_s = d \ln \left( \frac{T}{a} + \sqrt{\left( \frac{T}{a} \right)^2 + 1} \right)$$

where

$$b_1 = \frac{m_3^2}{m_2^3}$$

$$T = \sqrt{b_1 \left( \frac{(n+1)(n+3)}{6(n-2)} \right)}$$

$$C = \frac{3(n^2 + 27n - 70)(n+1)(n+3)}{(n-2)(n+5)(n+7)(n+9)}$$

$$W^2 = -1 + \sqrt{2(C-1)}$$

$$a = \sqrt{\frac{2}{W^2 - 1}}$$

$$d = \frac{1}{\sqrt{\ln(W)}}$$

## Normality Tests

### D'Agostino Kurtosis Test

D'Agostino (1990) describes a normality test based on the kurtosis coefficient,  $b_2$ . Recall that for the normal distribution, the theoretical value of  $b_2$  is 3. Hence, a test can be developed to determine if the value of  $b_2$  is significantly different from 3. If it is, the data are obviously non-normal. The statistic,  $z_k$ , is, under the null hypothesis of normality, approximately normally distributed for sample sizes  $n > 20$ . The calculation of this test proceeds as follows:

$$z_k = \frac{\left(1 - \frac{2}{9A}\right) - \left(\frac{1 - \frac{2}{A}}{1 + G\sqrt{\frac{2}{A-4}}}\right)^{1/3}}{\sqrt{\frac{2}{9A}}}$$

where

$$b_2 = \frac{m_4}{m_2^2}$$

$$G = \frac{b_2 - \left(\frac{3n-3}{n+1}\right)}{\sqrt{\frac{24n(n-2)(n-3)}{(n+1)^2(n+3)(n+5)}}}$$

$$E = \frac{6(n^2 - 5n + 2)}{(n+7)(n+9)} \sqrt{\frac{6(n+3)(n+5)}{n(n-2)(n-3)}}$$

$$A = 6 + \frac{8}{E} \left( \frac{2}{E} + \sqrt{1 + \frac{4}{E^2}} \right)$$

### D'Agostino Omnibus

D'Agostino (1990) describes a normality test that combines the tests for skewness and kurtosis. The statistic,  $K^2$ , is approximately distributed as a chi-square with two degrees of freedom. After calculated  $z_s$  and  $z_k$ , calculate  $K^2$  as follows:

$$K^2 = z_s^2 + z_k^2$$

---

## Data Structure

The data are contained in a single variable.

### Height dataset (subset)

| Height |
|--------|
| 64     |
| 63     |
| 67     |
| .      |
| .      |
| .      |

---

## Procedure Options

This section describes the options available in this procedure. To find out more about using a procedure, turn to the Procedures chapter.

Following is a list of the procedure's options.

---

### Variables Tab

The options on this panel specify which variables to use.

---

#### Data

##### Data Variable(s)

Specify a list of one or more variables upon which the normality tests are to be generated. You can double-click the field or single click the button on the right of the field to bring up the Variable Selection window.

##### Frequency Variable

This optional variable specifies the number of observations that each row represents. When omitted, each row represents a single observation. If your data is the result of a previous summarization, you may want certain rows to represent several observations. Note that negative values are treated as a zero weight and are omitted.

---

#### Break Variables

##### Break Variables

You can select up to five categorical-break variables. When one or more of these are specified, a separate set of reports is generated for each unique set of values for these variables.

---

#### Box-Cox Power Transformations

##### Exponent

Occasionally, you might want to obtain a statistical report on the square root or logarithm of your variable. This option lets you specify an on-the-fly transformation of the variable. The form of this transformation is  $X = Y^A$ , where Y is the original value, A is the selected exponent, and X is the value that is summarized.

##### Additive Constant

Occasionally, you might want add a constant to each value so that no zero or negative values occur. This option lets you specify an on-the-fly transformation of the variable. The form of this transformation is  $X = Y+B$ , where Y is the original value, B is the Additive Constant, and X is the value that results.

Note that if you apply both the *Exponent* and the *Additive Constant*, the form of the transformation is  $X = (Y+B)^A$ .

---

### Reports Tab

The options on this panel control the format of the report.

---

#### Select Reports

##### Summary Section, Counts Section, and Normality Tests

Each of these options indicates whether to display the corresponding report.

## Normality Tests

---

### Report Options

#### Precision

Specify the precision of numbers in the report. A single-precision number will show seven-place accuracy, while a double-precision number will show thirteen-place accuracy. Note that the reports were formatted for single precision. If you select double precision, some numbers may run into others. Also note that all calculations are performed in double precision regardless of which option you select here. This is for reporting purposes only.

#### Value Labels

This option applies to the *Break Variable(s)*. It lets you select whether to display data values, value labels, or both. Use this option if you want the output to automatically attach labels to the values (like 1=Yes, 2=No, etc.). See the section on specifying *Value Labels* elsewhere in this manual.

#### Variable Names

This option lets you select whether to display only variable names, variable labels, or both.

---

### Report Options - Decimal Places

#### Values, Means, Probabilities

Specify the number of decimal places when displaying this item. Select 'General' to display all possible decimal places.

---

### Plots Tab

These options specify the histogram and probability plot.

---

### Select Plots

#### Histogram and Probability Plot

Specify whether to display the indicated plots. Click the plot format button to change the plot settings.

## Example 1 – Running Normality Tests

This section presents a detailed example of how to run normality tests on the *SepalLength* variable in the Fisher dataset. To run this example, take the following steps.

You may follow along here by making the appropriate entries or load the completed template **Example 1** by clicking on Open Example Template from the File menu of the Normality Tests window.

### 1 Open the Fisher dataset.

- From the File menu of the NCSS Data window, select **Open Example Data**.
- Click on the file **Fisher.NCSS**.
- Click **Open**.

### 2 Open the Normal Tests window.

- Using the Analysis menu or the Procedure Navigator, find and select the **Normality Tests** procedure.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

### 3 Specify the SepalLength variable.

- On the Normality Tests window, select the **Variables tab**.
- Double-click in the **Data Variables** text box. This will bring up the variable selection window.
- Select **SepalLength** from the list of variables and then click **Ok**. The word “SepalLength” will appear in the Data Variables box. Remember that you could have entered a “1” here signifying the first (left-most) variable on the dataset.

### 4 Run the procedure.

- From the Run menu, select **Run Procedure**. Alternatively, just click the green Run button.

The following reports and charts will be displayed in the Output window.

## Summary Section

| Summary Section of SepalLength |          |                    |                |         |         |       |
|--------------------------------|----------|--------------------|----------------|---------|---------|-------|
| Count                          | Mean     | Standard Deviation | Standard Error | Minimum | Maximum | Range |
| 150                            | 58.43333 | 8.280662           | 0.6761132      | 43      | 79      | 36    |

### Count

This is the number of non-missing values. If no frequency variable was specified, this is the number of non-missing rows.

### Mean

This is the average of the data values.

### Standard Deviation

This is the standard deviation of the data values

### Standard Error

This is the standard error of the mean.

### Minimum

The smallest value in this variable.

## Normality Tests

### Maximum

The largest value in this variable.

### Range

The difference between the largest and smallest values for a variable.

## Count Section

### Counts Section of SepalLength

| Rows | Sum of Frequencies | Missing Values | Distinct Values |
|------|--------------------|----------------|-----------------|
| 150  | 150                | 0              | 35              |

### Rows

This is the total number of rows available in this variable.

### Sum of Frequencies

This is the number of non-missing values. If no frequency variable was specified, this is the number of non-missing rows.

### Missing Values

The number of missing (empty) rows.

### Distinct Values

This is the number of unique values in this variable. This value is useful for finding data entry errors and for determining if a variable is continuous or discrete.

## Normality Test Section

### Normality Test Section of SepalLength

| Test Name           | Test Value | Prob Level | 10% Critical Value | 5% Critical Value | Decision (Alpha = 5%)  |
|---------------------|------------|------------|--------------------|-------------------|------------------------|
| Shapiro-Wilk W      | 0.976      | 0.0102     |                    |                   | Reject normality       |
| Anderson-Darling    | 0.894      | 0.0225     |                    |                   | Reject normality       |
| Martinez-Iglewicz   | 0.909      |            | 1.036              | 1.056             | Can't reject normality |
| Kolmogorov-Smirnov  | 0.089      |            | 0.066              | 0.072             | Reject normality       |
| D'Agostino Skewness | 1.596      | 0.1104     | 1.645              | 1.96              | Can't reject normality |
| D'Agostino Kurtosis | -1.785     | 0.0742     | 1.645              | 1.96              | Can't reject normality |
| D'Agostino Omnibus  | 5.736      | 0.0568     | 4.605              | 5.991             | Can't reject normality |

### Shapiro-Wilk W Test

This test for normality, developed by Shapiro and Wilk (1965), has been found to be the most powerful test in most situations. The test is not calculated when a frequency variable is specified.

### Anderson-Darling Test

This test, developed by Anderson and Darling (1954), is the most popular normality test that is based on EDF statistics. In some situations, it has been found to be as powerful as the Shapiro-Wilk test. The test is not calculated when a frequency variable is specified.

## Normality Tests

### Martinez-Iglewicz Test

This test for normality, developed by Martinez and Iglewicz (1981), is based on the median and a robust estimator of dispersion.

### Martinez-Iglewicz (10% Critical and 5% Critical)

The 10% and 5% critical values are given here. If the value of the test statistic is greater than this value, reject normality at that level of significance.

### Kolmogorov-Smirnov

This test for normality is based on the maximum difference between the observed distribution and expected cumulative-normal distribution. This test has been shown to be less powerful than the other tests in most situations. It is included because of its historical popularity.

### Kolmogorov-Smirnov (10% Critical and 5% Critical)

The 10% and 5% critical values are given here. If the value of the test statistic is greater than this value, reject normality at that level of significance. The critical values are the Lilliefors' adjusted values as given by Dallal (1986). If the test value is greater than the reject critical value, normality is rejected at that level of significance.

### D'Agostino Skewness

D'Agostino (1990) describes a normality test based on the skewness coefficient.

### Skewness Test (Prob Level)

This is the two-tail, significance level for this test. Reject the null hypothesis of normality if this value is less than a pre-determined value, say 0.05.

### Skewness Test Decision (5%)

This reports the outcome of this test at the 5% significance level.

### D'Agostino Kurtosis

D'Agostino (1990) describes a normality test based on the kurtosis coefficient.

### Prob Level of Kurtosis Test

This is the two-tail significance level for this test. Reject the null hypothesis of normality if this value is less than a pre-determined value, say 0.05.

### Decision of Kurtosis Test

This reports the outcome of this test at the 5% significance level.

### D'Agostino Omnibus

D'Agostino (1990) describes a normality test that combines the tests for skewness and kurtosis.

### Prob Level D'Agostino Omnibus

This is the significance level for this test. Reject the null hypothesis of normality if this value is less than a pre-determined value, say 0.05.

### Decision of D'Agostino Omnibus Test

This reports the outcome of this test at the 5% significance level.

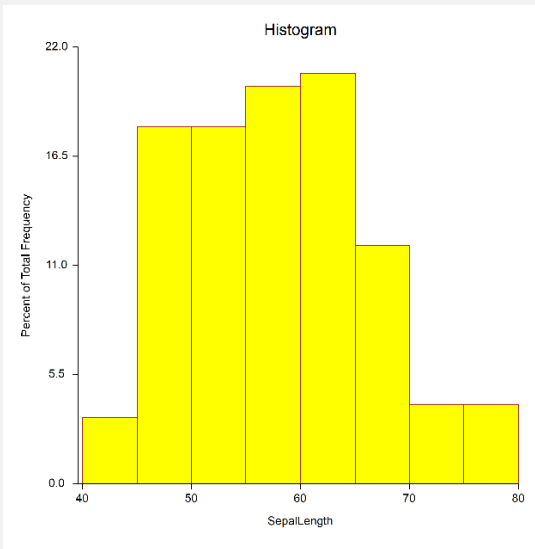
---

## Histogram Plot

The following plot shows a histogram of the data.



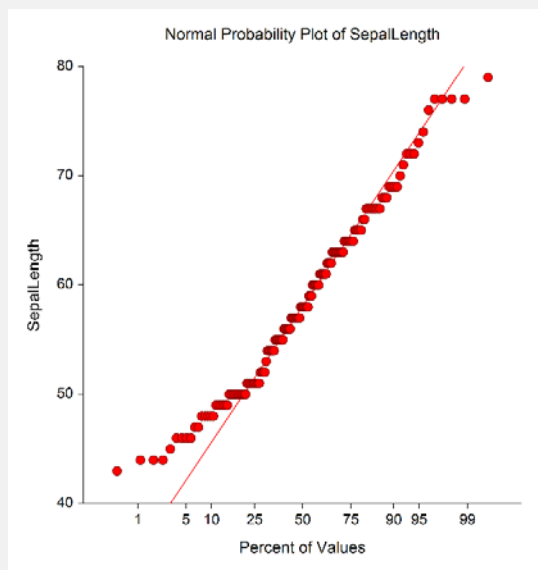
## Normality Tests



## Histogram

The histogram is a traditional way of displaying the shape of a group of data. It is constructed from a frequency distribution, where choices on the number of bins and bin width have been made. These choices can drastically affect the shape of the histogram. The ideal shape to look for in the case of normality is a bell-shaped distribution.

## Normal Probability Plot



This is a plot of the inverse of the standard normal cumulative versus the ordered observations. If the underlying distribution of the data is normal, the points will fall along a straight line. Deviations from this line correspond to various types of non-normality. Stragglers at either end of the normal probability plot indicate outliers. Curvature at both ends of the plot indicates long or short distribution tails. Convex, or concave, curvature indicates a lack of symmetry. Gaps, plateaus, or segmentation in the plot indicate certain phenomenon that need closer scrutiny.