

# Multivariate Normality Test



Revised: 10/11/2017



Summary .....	1
Data Input.....	3
Analysis Summary .....	4
Chi-Square Plot.....	5
Analysis Options.....	7
Save Results .....	9
References.....	10

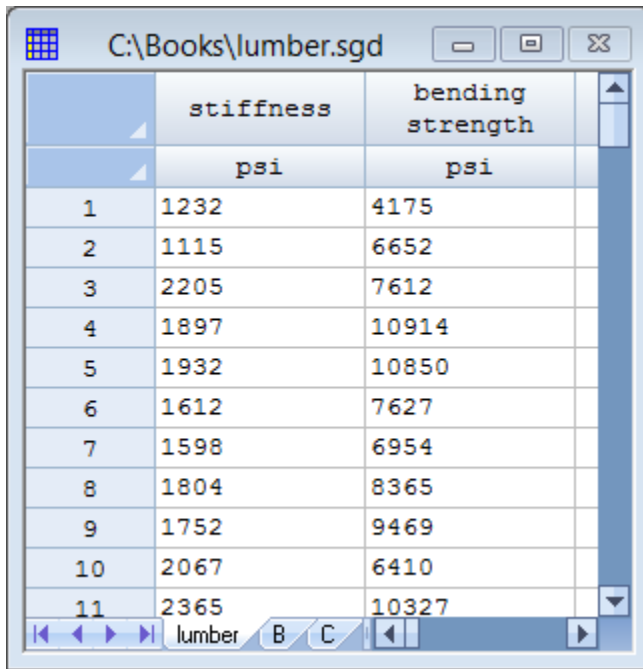
## Summary

This procedure tests whether a set of random variables could reasonably have come from a multivariate normal distribution. It includes Royston’s  $H$  test and tests based on a chi-square plot of the squared distances of each observation from the sample centroid.

**Sample StatFolio:** *mvnormal.sgp*

## Sample Data:

The file *lumber.sgd* contains measurements of the stiffness and bending strength of  $n = 30$  pieces of lumber (Johnson and Wichern, 2002). A portion of the data is shown below:

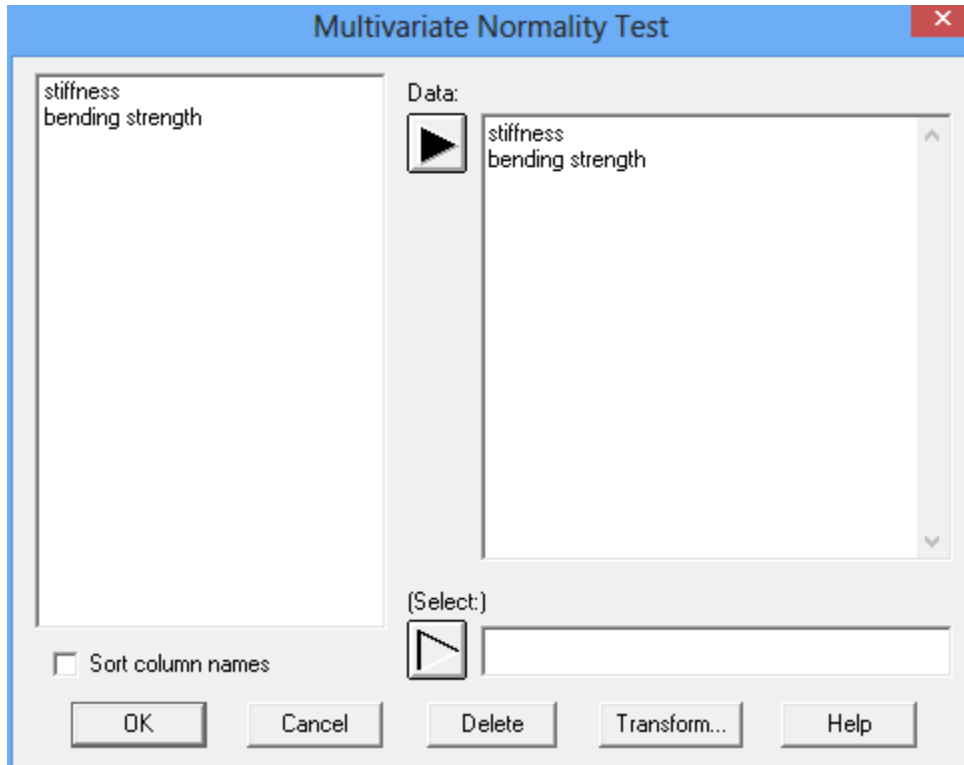


	stiffness	bending strength
	psi	psi
1	1232	4175
2	1115	6652
3	2205	7612
4	1897	10914
5	1932	10850
6	1612	7627
7	1598	6954
8	1804	8365
9	1752	9469
10	2067	6410
11	2365	10327

We wish to test the hypothesis that the samples come from a bivariate normal distribution.

## Data Input

To perform a multivariate test for normality, choose *Multivariate Normality Test* from the *Multivariate Methods* menu. The data input dialog box is shown below:



- **Data:** the names of 2 or more numeric columns containing the data.
- **Select:** subset selection.

The data for each of the  $m$  variables should be placed in a separate column. Each row corresponds to a single multivariate observation.

## Analysis Summary

The *Analysis Summary* shows sample statistics for the variables and the test results:

<u>Multivariate Normality Test</u>		
Data variables:		
stiffness (psi)		
bending strength (psi)		
	<i>Mean</i>	<i>Standard deviation</i>
stiffness	1860.5	352.214
bending strength	8354.13	1867.17
Sample Correlations		
	stiffness	bending strength
stiffness	1.0	0.549872
bending strength	0.549872	1.0
Number of observations = 30		
Goodness-of-Fit Test		
<i>Test</i>	<i>Statistic</i>	<i>P-Value</i>
Shapiro-Wilk W - stiffness	0.975	0.6798
Shapiro-Wilk W - bending strength	0.976	0.6980
Royston's H	0.325	0.8545

The first table shows the vector of sample means  $\bar{X}$  and the vector of sample standard deviations  $s$ . The second table shows the  $m$  by  $m$  sample correlation matrix  $R$ . The element in the  $j^{\text{th}}$  row and  $k^{\text{th}}$  column of the correlation matrix is calculated from

$$r_{j,k} = \frac{\sum_{i=1}^n (x_{j,i} - \bar{x}_j)(x_{k,i} - \bar{x}_k)}{\sqrt{\left[ \sum_{i=1}^n (x_{j,i} - \bar{x}_j)^2 \right] \left[ \sum_{i=1}^n (x_{k,i} - \bar{x}_k)^2 \right]}} \quad (1)$$

The lower section of the output shows the results of testing the data for normality. First, a Shapiro-Wilk  $W$  test is applied to each variable separately. P-values are calculated using the algorithm of Royston (1995). P-values below  $\alpha$  indicate rejection of the hypothesis that a particular variable follows a normal distribution at significance level  $\alpha$ . Royston's  $H$  test is then applied to the  $m$  variables simultaneously to test the hypotheses:

$H_0$ : the variables come from a multivariate normal distribution.

$H_A$ : the variables do not come from a multivariate normal distribution.

Royston's test combines the Shapiro-Wilk statistics for the separate variables and compares the result to a chi-square distribution (Royston, 1983). A small P-value for Royston's  $H$  leads to

rejection of the hypothesis of multivariate normality. For the sample data,  $P = 0.8545$  which is well above  $\alpha = 0.05$ , so the assumption of multivariate normality is not rejected.

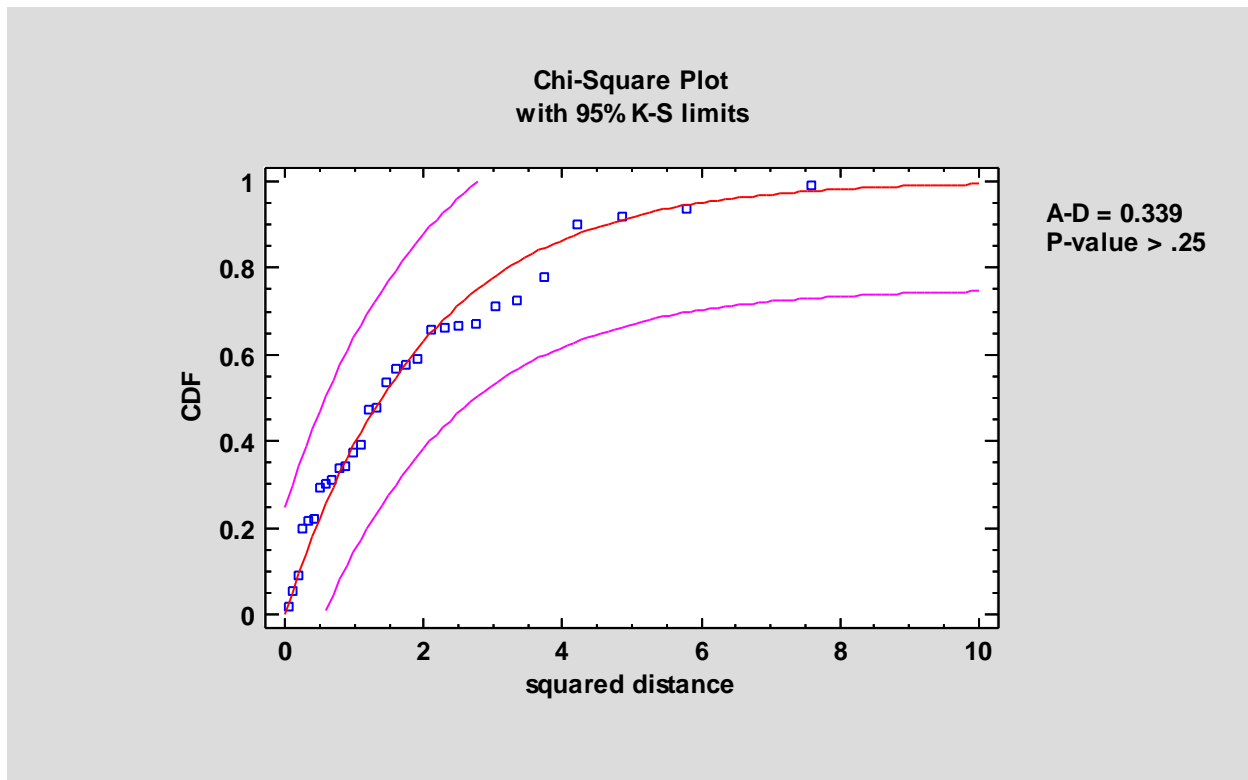
## Chi-Square Plot

The procedure also creates a chi-square plot based on the generalized distances from the observations to the sample mean vector. The squared generalized distance for observation  $i$  is defined by

$$d_i^2 = (\mathbf{X}_i - \bar{\mathbf{X}})^T \mathbf{S}^{-1} (\mathbf{X}_i - \bar{\mathbf{X}}), \quad i = 1, 2, \dots, n \quad (7.2)$$

where  $\mathbf{S}$  is the sample covariance matrix. If the data follow a multivariate normal distribution, then Johnson and Wichern (2002) suggest comparing the distances to a chi-square distribution with  $m$  degrees of freedom. Departures from that chi-square distribution would indicate that the data do not come from a multivariate normal distribution.

If you select *Chi-Square Plot* from the list of available graphs, the following plot will be displayed:



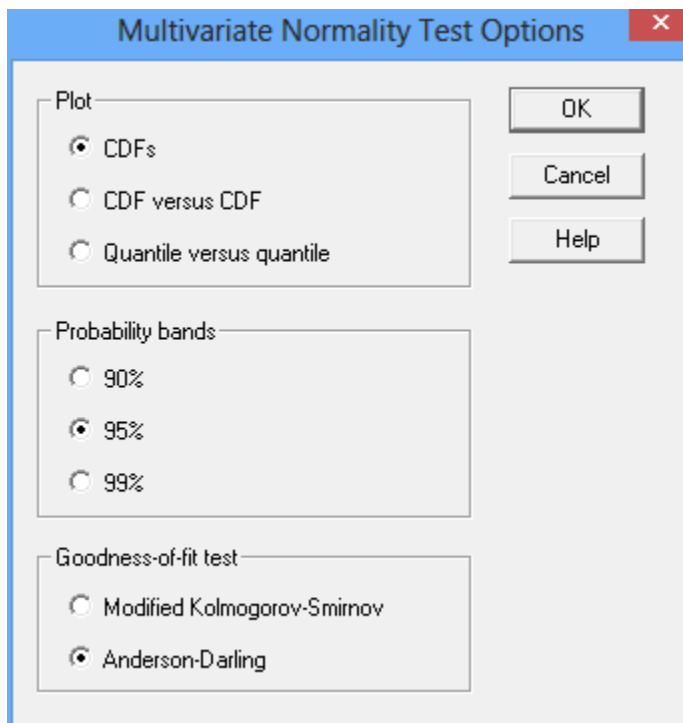
The default format plots the cumulative distribution function for the chi-square distribution with  $m$  degrees of freedom as a solid line and the empirical CDF of the squared distances as point symbols. If the hypothesis of multivariate normality is correct, the points should lie close to the solid line. Kolmogorov-Smirnov limits are also plotted on either side of the solid line, using a selected probability level such as 95%. They provide simultaneous limits for departure of the

empirical CDF from the hypothesized CDF. 95% of all samples taken from the hypothesized chi-square distribution should fall entirely within the bands.

A goodness-of-fit test is also performed comparing the squared distances to the chi-square distribution. For example, the plot above tests the goodness-of-fit using an Anderson-Darling test. Small P-values (below 0.05) would indicate that the hypothesis that the squared distances come from the hypothesized chi-square distribution could be rejected at the 5% significance level.

In the plot above, the points remain entirely within the limits and the P-value is well above 0.05. This confirms the conclusion that the data may well have come from a multivariate normal distribution.

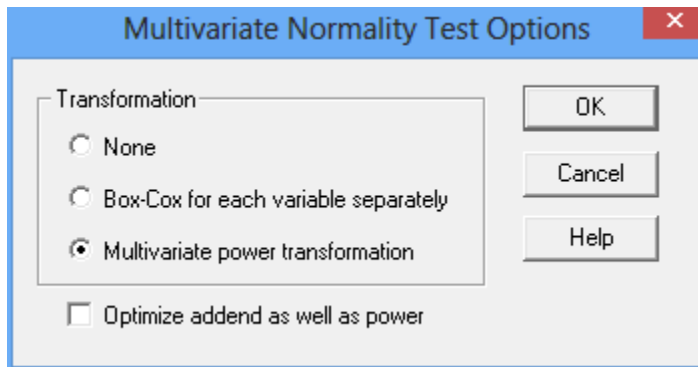
### Pane Options



- **Plot** – controls the format of the plot. *CDFs* plots the cumulative distribution function of both the data and the hypothesized chi-square distribution. *CDF versus CDF* plots the empirical CDF on the vertical axis and the hypothesized CDF on the horizontal axis. *Quantile versus quantile* creates a Q-Q plot using the empirical and theoretical quantiles.
- **Probability bands** – sets the probability level for the Kolmogorov-Smirnov limits.
- **Goodness-of-fit test** – selects the test used to compare the squared distances to the chi-square distribution.

## Analysis Options

If the original data do not follow a multivariate normal distribution, transformed values of the variables might. The *Analysis Options* dialog box allows the user to transform the variable using either a univariate or multivariate Box-Cox transformation:



### Box-Cox for each variable separately

If this option is selected, the program will find the best power transformation for each of the variables separately. The procedure automatically determines the best transformation by finding the values of  $\lambda$  and  $\Delta$  that minimize the standard deviation of the observations when transformed according to the Box-Cox transformation:

$$Y = 1 + \frac{(X + \Delta)^\lambda - 1}{\lambda g^{\lambda-1}} \quad \text{if} \quad \lambda \neq 0 \quad (2)$$

$$Y = 1 + g \ln(X + \Delta) \quad \text{if} \quad \lambda = 0 \quad (3)$$

where  $g$  is the geometric mean of the observations after adding  $\Delta$ :

$$g = \left( \prod_{i=1}^n (X_i + \Delta) \right)^{1/n} \quad (4)$$

If “Optimize addend as well as power” is not checked, the addend  $\Delta$  is set equal to 0.

### Multivariate power transformation

If this option is selected, the program find the best vector of powers  $\boldsymbol{\lambda} = \{\lambda_1, \lambda_2, \dots, \lambda_m\}$  and vector of addends  $\boldsymbol{\Delta} = \{\Delta_1, \Delta_2, \dots, \Delta_m\}$  by minimizing the profile likelihood function defined by

$$-\frac{n \log |\hat{\Sigma}(\boldsymbol{\lambda})|}{2} + \sum_{j=1}^m \{(\lambda_j - 1) \sum_{i=1}^n \log(x_{i,j} + \Delta_j)\} \quad (5)$$

where  $\hat{\Sigma}(\lambda)$  is the estimated covariance matrix of the transformed variables defined by

$$x_{i,j}^* = \begin{cases} \frac{(x_{i,j}^{\lambda_j - 1})}{\lambda_j} & \text{if } \lambda_j \neq 0 \\ \log x_{i,j} & \text{if } \lambda_j = 0 \end{cases} \quad (6)$$

For more information on this transformation, see Andrews, Gnanadesikan and Warner (1971).

For example, the table below shows the results of applying the multivariate one-parameter power transformation to the sample data:

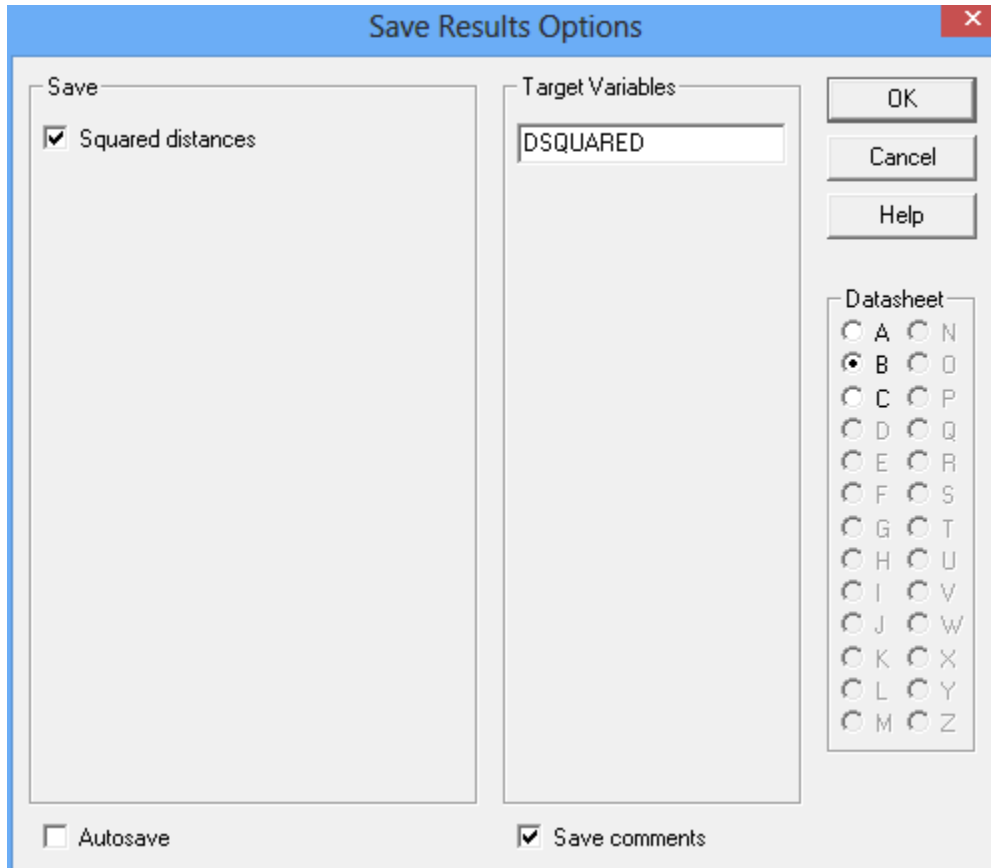
<b>Multivariate Normality Test</b>		
Data variables:		
stiffness (psi)		
bending strength (psi)		
Power transformations: estimated simultaneously		
Variable	Power	
stiffness	0.680894	
bending strength	1.18917	
	Mean	Standard deviation
stiffness	167.727	21.8233
bending strength	46366.1	12212.6
Sample Correlations		
	stiffness	bending strength
stiffness	1.0	0.556579
bending strength	0.556579	1.0
Number of observations = 30		
Normality Tests		
Test	Statistic	P-Value
Shapiro-Wilk W - stiffness	0.975	0.6959
Shapiro-Wilk W - bending strength	0.975	0.6706
Royston's H	0.338	0.8491

The optimal vector of powers was determined to be  $\hat{\lambda} = (0.680894, 1.18917)$ . The transformation has little effect on the shape of the distribution since the powers are both fairly close to 1.0. This is not unexpected since the original data showed no signs of nonnormality.



## Save Results

The squared distances may be saved to a datasheet by pressing the *Save Results* button on the analysis toolbar. This displays the following dialog box:



Check the box for *Squared distances* and press *OK*.

## References

Andrews, D.F., Gnanadesikan, R. and Warner, J.L. (1971) "Transformations of Multivariate Data". Biometrics 27, pp. 825-840.

Johnson, R.A. and Wichern, D. W. (2002). Applied Multivariate Statistical Analysis, fifth edition. Prentice Hall, Upper Saddle River, N.J.

Royston, J.P. (1982). "An Extension of Shapiro and Wilk's W Test for Normality to Large Samples". Applied Statistics 31, pp. 115-124.

Royston, J.P. (1983). "Some Techniques for Assessing Multivariate Normality Based on the Shapiro-Wilk W". Applied Statistics 32, pp. 121-133.

Royston, J.P. (1995). "Remark AS R94: A remark on Algorithm AS 181: The W test for normality". Applied Statistics 44, pp. 547-551.

Shapiro, S.S. and Wilk, M.B. (1965) "An analysis of variance test for normality (complete samples)". Biometrika 52, pp. 591-611.