

## ***Multivariate Capability Analysis***



Revised: 10/11/2017

Summary .....	1
Data Input.....	3
Analysis Summary .....	4
Capability Plot .....	5
Capability Indices .....	7
Capability Ellipse.....	10
Correlation Matrix .....	12
Tests for Normality.....	12
Probability Plot .....	13
T-Squared Chart.....	14
Multivariate Tolerance Limits .....	16
Tolerance Region .....	18
Calculations.....	20
References.....	20

### **Summary**

The **Multivariate Capability Analysis** procedure determines the probability that items characterized by two or more variables meet established specifications limits. When variables are correlated, it is important to consider their joint behavior, since looking at each variable separately may give a misleading picture of the overall process capability.

**Sample StatFolio:** *mvcapability.sgp*

## Sample Data

The file *grit.sgd* contains measurements made on  $n = 56$  batches of “grit”, from Holmes and Mergen (1993). The data represent the percentages of large, medium, and small particles in the grit. The table below shows a partial list of the data in that file:

<i>Large</i>	<i>Medium</i>	<i>Small</i>	<i>LSL</i>	<i>Nominal</i>	<i>USL</i>
5.4	93.6	1.0		5	10
3.2	92.6	4.2		5	10
5.2	91.7	3.1			
3.5	86.9	9.6			
2.9	90.4	6.7			
4.6	92.1	3.3			
4.4	91.5	4.1			
5.0	90.3	4.7			
8.4	85.1	6.5			
4.2	89.7	6.1			
3.8	92.5	3.7			
4.3	91.8	3.9			
3.7	91.7	4.6			
3.8	90.3	5.9			
2.6	94.5	2.9			

In addition to the sample, the file also contains specification limits for the relative percentages of large and small particles, which are:

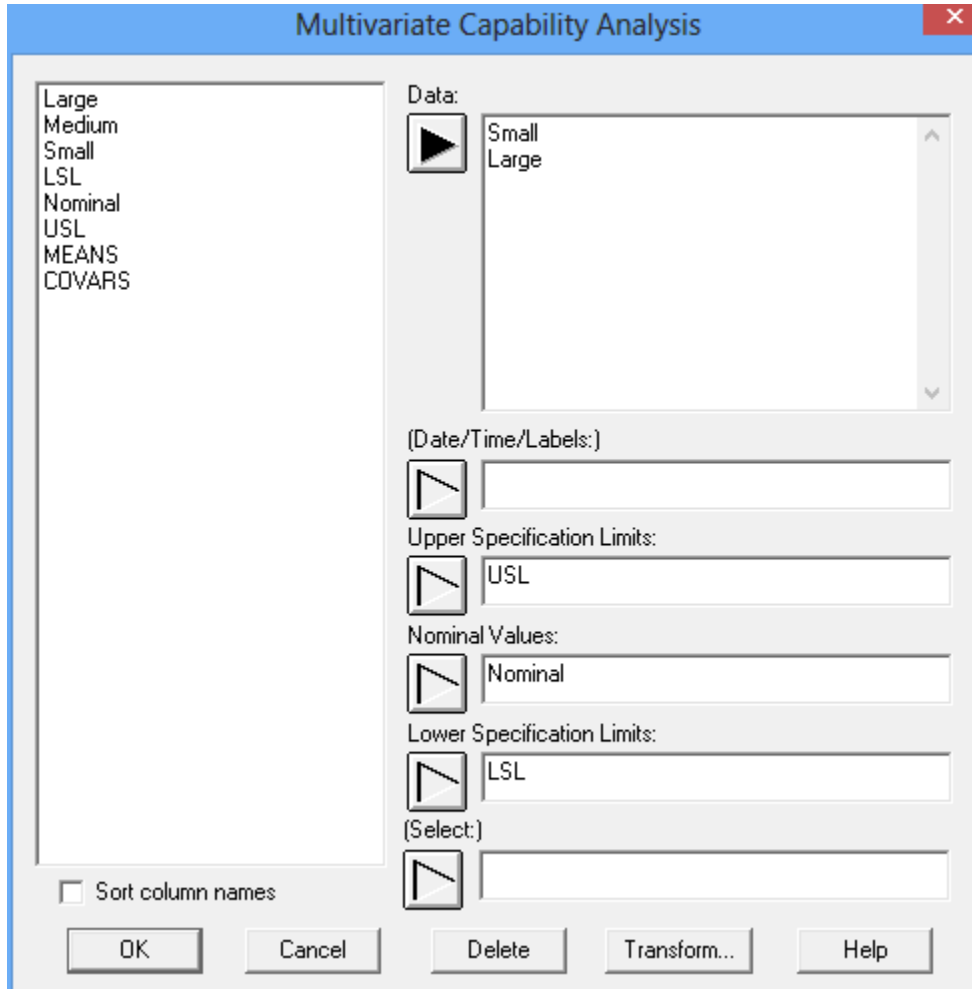
$$\text{Large} \leq 10\% \quad (1)$$

$$\text{Small} \leq 10\% \quad (2)$$

Note that the column containing the lower limits is blank, since there are no lower bounds. It is desired to estimate the percentage of batches that meet the specification limits.

## Data Input

The data to be analyzed consist of numeric columns containing the variables of interest and the specification limits.



- **Data:** numeric columns containing the  $n$  samples, one per row.
- **Date/Time/Labels:** optional identifier for each sample.
- **Upper Specification Limits:** numeric column containing the upper spec limits. If a variable does not have an upper limit, leave the corresponding cell blank.
- **Nominal Values:** numeric column containing the nominal or target values for each variable. If a variable does not have a target value, leave the corresponding cell blank.
- **Lower Specification Limits:** numeric column containing the lower spec limits. If a variable does not have a lower limit, leave the corresponding cell blank.
- **Select:** subset selection.

## Analysis Summary

The *Analysis Summary* shows summary statistics for each column and the estimated probability of being beyond the specification limits.

<u>Multivariate Capability Analysis</u>					
Data variables:					
Small					
Large					
Number of complete cases: 56					
	<i>Sample</i>	<i>Sample</i>			
<i>Variable</i>	<i>Mean</i>	<i>Std. Dev.</i>	<i>LSL</i>	<i>Nominal</i>	<i>USL</i>
Small	6.09821	2.51154		5.0	10.0
Large	5.68214	1.94171		5.0	10.0
	<i>Observed</i>	<i>Estimated</i>	<i>Estimated</i>		
<i>Variable</i>	<i>Beyond Spec.</i>	<i>Beyond Spec.</i>	<i>DPM</i>		
Small	5.35714%	6.01462%	60146.2		
Large	1.78571%	1.30827%	13082.7		
Joint	7.14286%	7.18235%	<b>71823.5</b>		

Included in the table are:

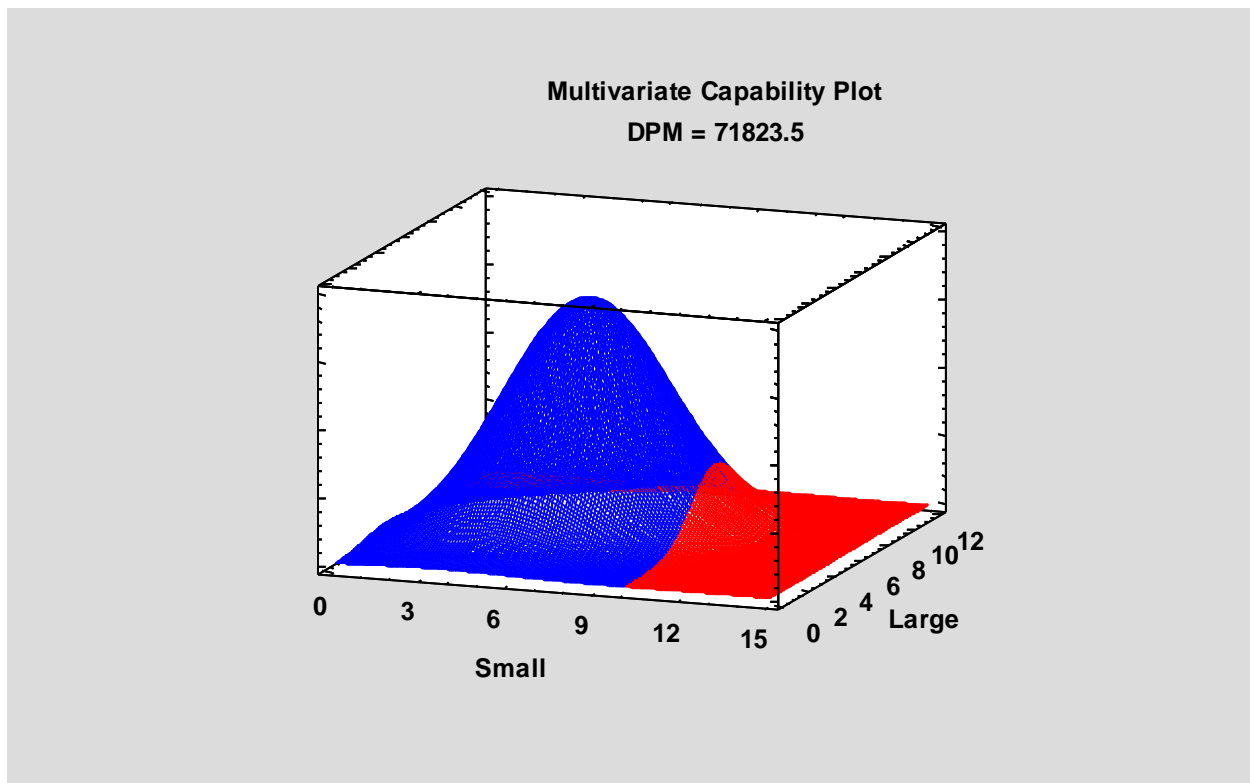
- **Number of complete cases:** the number of rows in the data file  $n$  with information on all  $p$  variables. Any row with missing data for any variable is excluded from the analysis.
- **Sample Mean and Std. Dev.:** the sample mean and sample standard deviation for each variable.
- **Observed Beyond Spec.:** the percentage of samples in which each variable was outside its individual specification limits, together the percentage of time when the samples were *jointly* outside the limits. The joint percentage includes cases in which *one or more* variables were outside their individual limits.
- **Estimated Beyond Spec.:** the percentage of similar samples in the population that are estimated to be outside the specification limits. The estimate is based on fitting a multivariate normal distribution to the data and estimating the area of that distribution that is not within the joint specification limits.
- **Estimated DPM:** the estimated “defects per million”, defined as the number of samples out of each million that are estimated to fall outside the specification limits. This is calculated from the *Estimated Beyond Spec.*

In the example, the samples of grit violated one or more specification limits 7.14% of the time. Based on the fitted multivariate normal distribution, it is estimated that 7.18% of all samples taken from the population would be jointly out of spec. It should be noted that the joint probability is *not* the sum of the individual probabilities.

NOTE: the calculation of the joint estimated percentage out of spec involves integrating a multivariate normal distribution and may be very time-consuming.

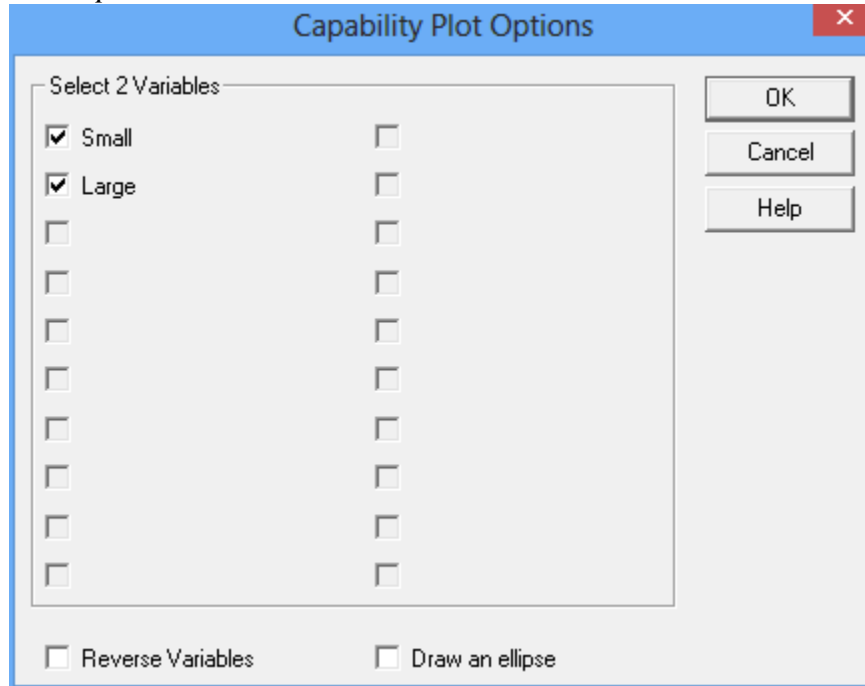
## Capability Plot

The *Capability Plot* displays the fitted multivariate normal distribution for any 2 of the variables.



The area shaded in blue corresponds to locations where all of the variables are within the specification limits. The area shaded in red corresponds to locations where one or more of the variables are out of spec.

## Pane Options



- **Select 2 Variables:** select two of the  $p$  variables. The marginal distribution of the 2 selected variables will be displayed.
- **Reverse Variables:** check this box to display the second variable along the X axis in the front of the display. Otherwise, the first variable will be plotted on the X axis.
- **Draw an ellipse:** check this box to add an ellipse covering 99.73% of the joint distribution of the 2 plotted variables.

## Capability Indices

This table displays capability indices calculated from the data.

Capability Indices	
Index	Estimate
MC <sub>pk</sub>	0.49
MC <sub>r</sub>	205.15
DPM	71823.5
Z	1.46235
SQL	2.96235

Based on 6 sigma limits. The Sigma Quality Level includes a 1.5 sigma drift in the mean.

The indices calculated are:

- **DPM** – estimated defects per million. This is the estimated number of samples taken from the population that would be out of spec on one or more of the variables.
- **MC<sub>pk</sub>** – a multivariate capability index. The index is calculated by

$$MC_{pk} = \frac{Z}{k/2} \quad (3)$$

where  $Z$  is the value of a standard normal random variable corresponding to the calculated  $DPM$ , and  $k$  is the multiple of sigma specified on the *Capability* tab of the *Preferences* dialog box, accessible from the *Edit* menu. Normally,  $k = 6$ . This index is designed to have a similar interpretation to  $C_{pk}$  in the univariate case. Normally, it is desirable for  $MC_{pk} \geq 1.33$ .

- **MC<sub>r</sub>** – the multivariate capability ratio. The index is calculated by

$$MC_r = 100 \frac{k/2}{Z} \% \quad (4)$$

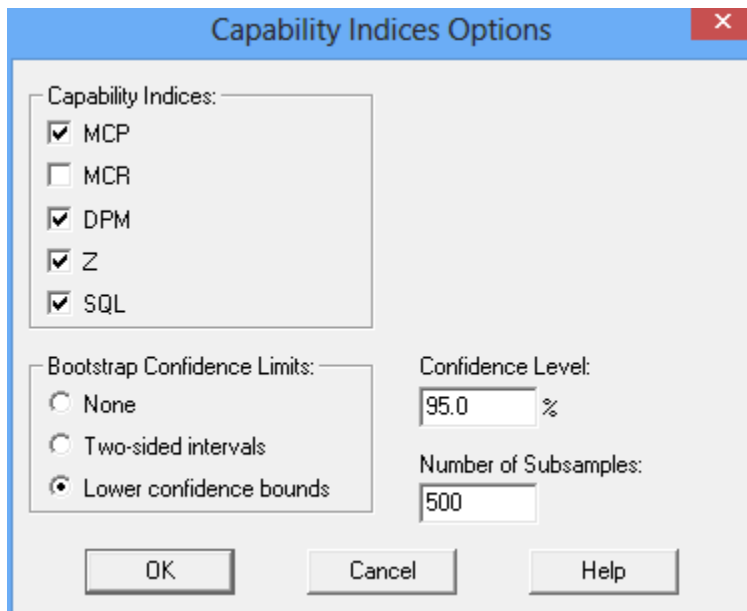
It is the inverse of  $MC_{pk}$  and expresses the percentage of the allowable variation that is being consumed by the variability in the data.

- **Z** – the value of a standard normal random variable corresponding to the calculated  $DPM$ .
- **SQL** – the estimated Sigma Quality Level. This is an index of the level of quality for the process developed as part of the Six Sigma protocol. Depending on the *1.5 Sigma Shift* setting on the *Capability* tab of the *Preferences* dialog box, the SQL equals either  $Z$  or  $(Z+1.5)$ .

The sample data have a capability index of  $MC_{pk} = 0.49$ , corresponding to a sigma quality level of approximately 3. While such levels of performance are typical, achieving world class quality would require improving process consistency.

## Pane Options

The *Pane Options* dialog box lets the analyst select the indices to be displayed and add confidence intervals for them if desired:



- **Capability Indices:** the indices to be displayed in the table.
- **Bootstrap Confidence Limits:** the type of confidence limits to be displayed, if any.
- **Confidence Level:** the confidence level used when calculating the bootstrap limits.
- **Number of Subsamples:** the number of samples created when calculating the bootstrap limits. Note: since the process can be quite time-consuming, start with a small number of subsamples until you know how long the estimation process will take.

Since there is no exact method for calculating confidence limits for the multivariate capability indices, confidence limits are estimated using a process known as bootstrapping. This process proceeds as follows:

Step 1: Create a subsample of the data by selecting  $n$  random observations from the original data, sampling *with replacement*. When sampling with replacement, the same observation may be selected multiple times.

Step 2: Calculate the capability indices using the data in the subsample.

Step 3: Repeat steps 1 and 2 many times, based on the value in the *Number of Subsamples* field, creating a distribution of each of the capability indices.



Step 4: Create the confidence limits by selecting the appropriate percentiles of the distributions accumulated for each index.

The table below shows 95% lower confidence bounds for each of the indices, using 500 subsamples.

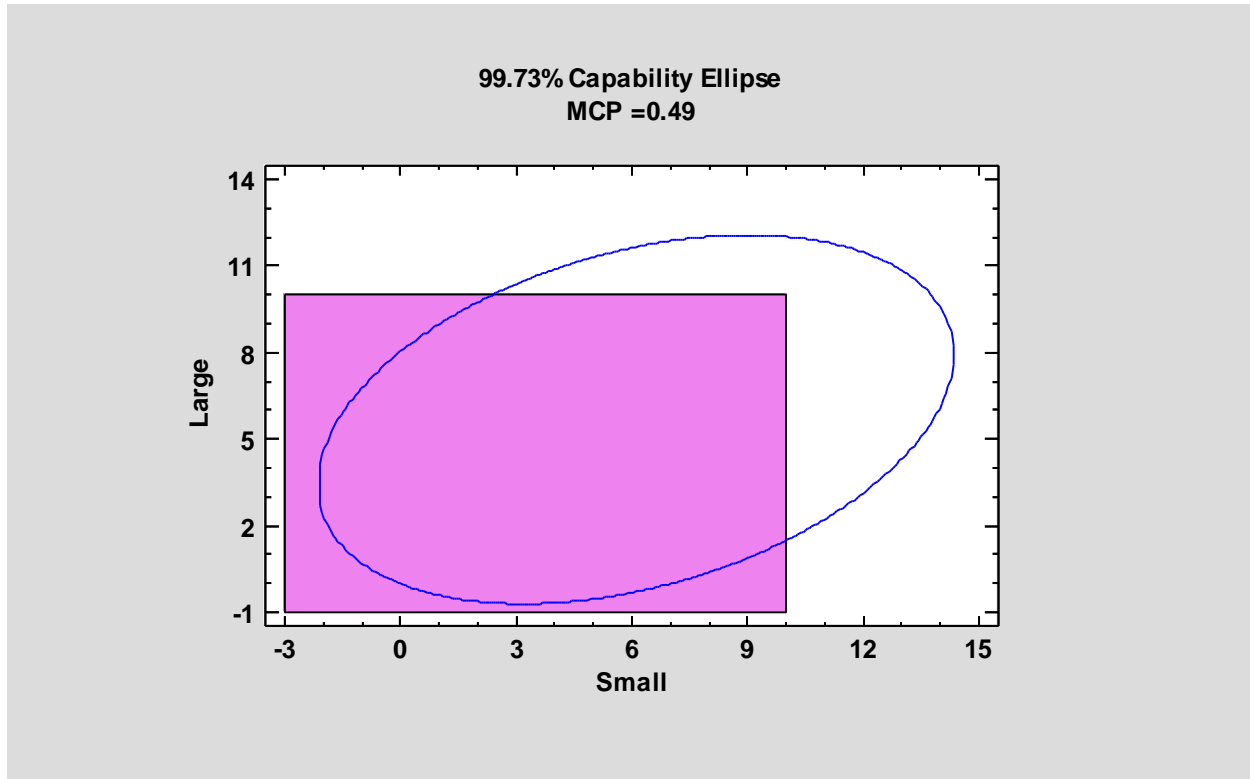
<b>95.0% Confidence Bounds - Bootstrap Method (500 subsamples)</b>	
	<i>Lower Limit</i>
MCpk	0.375221
*DPM	130155.
Z	1.12566
SQL	2.62566

\*Lower quality bound corresponds to upper limit for this index.

For the sample data, a 95% lower bound for  $MC_{pk}$  is approximately 0.375.

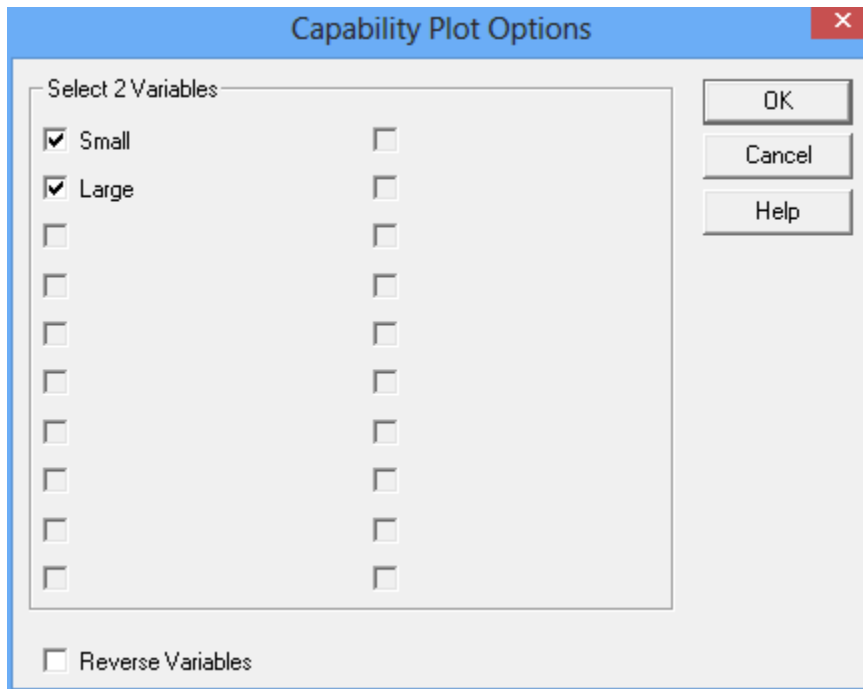
## Capability Ellipse

The *Capability Ellipse* shows a two-dimensional illustration of the estimated performance of the process with respect to 2 selected variables.



The ellipse is drawn in such a way that it includes 99.73% of the estimated multivariate normal probability for the selected variables, similar to  $\pm 3$  sigma limits on a univariate capability plot. The rectangular region indicates combinations of the 2 variables where both are within their specification limits. A “capable” process would contain an ellipse completely within the rectangular region.

## Pane Options



- **Select 2 Variables:** select two of the  $p$  variables.
- **Reverse Variables:** check this box to display the second variable along the X axis. Otherwise, the first variable will be plotted on the X axis.

## Correlation Matrix

This table displays the estimated correlation coefficients for each pair of variables:

	Small	Large
Small	1.0000	0.3538
Large	0.3538	1.0000

Correlation coefficients measure the strength of the linear relationship between a pair of variables, on a scale of  $-1$  for perfect negative correlation to  $+1$  for perfect positive correlation. The higher the correlations, the more important it is to analyze the variables jointly rather than doing a separate univariate capability analysis on each variable.

## Tests for Normality

The estimated process capability shown above is highly dependent on the assumption that the data follow a multivariate normal distribution. The *Tests for Normality* pane performs one or more tests on each variable to determine whether or not a normal distribution is a reasonable model for that variable. For each test, the hypotheses of interest are:

- Null hypothesis: data are independent samples from a normal distribution
- Alt. hypothesis: data are not independent samples from a normal distribution

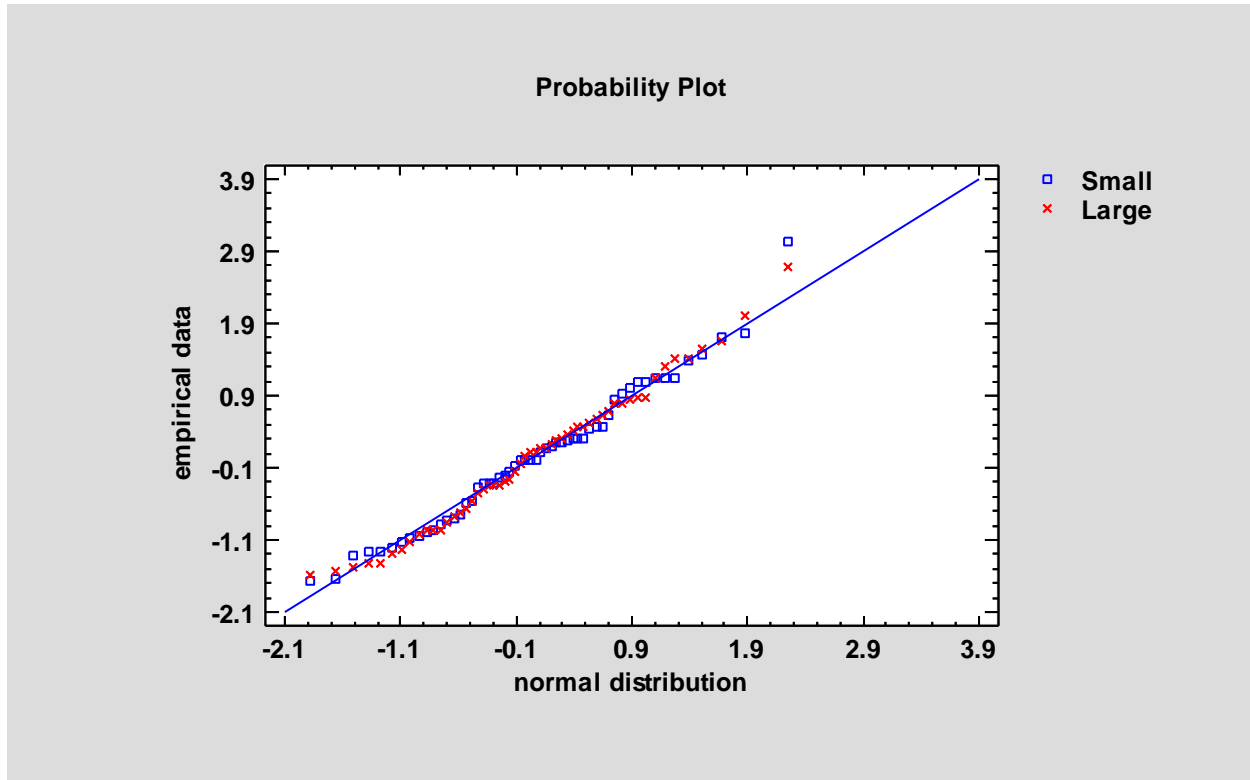
Test	Statistic	P-Value
Shapiro-Wilk W - Small	0.982	0.5805
Shapiro-Wilk W - Large	0.977	0.3662
Royston's H	1.525	0.4678

The tests displayed include the Shapiro-Wilk test, used to determine separately whether or not each variable is normally distributed, and Royston's test, which examines the joint distribution of all of the variables to determine whether they are samples from a multivariate normal distribution. Small P-values (below 0.05 if operating at the 5% significance level) lead to a rejection of the null hypothesis and thus to a rejection of the normal distribution. In the above table, the P-Values are all above 0.05, so there is not statistically significant non-normality in the data.

For a detailed description of the tests, see the documentation for *Distribution Fitting (Uncensored Data)* and for the *Multivariate Normality Test*.

## Probability Plot

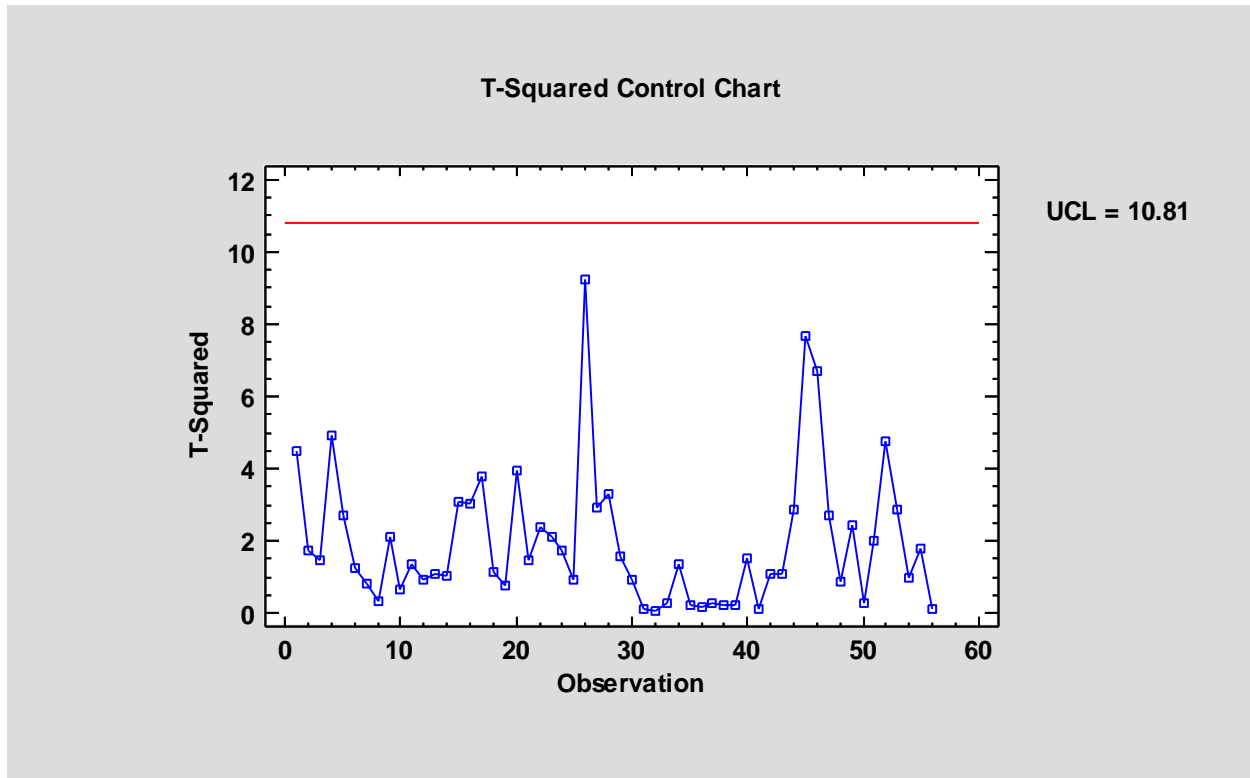
The *Probability Plot* is another method by which one can judge whether or not the currently selected distribution adequately describes the data.



The plot shows the standardized data values for each variable, sorted from smallest to largest, plotted against equivalent percentiles of the standard normal distribution. If the normal distribution is a reasonable model for the data, the points will fall approximately along a straight line. Except for one possible outlier, the above plot confirms that the normal distribution is a reasonable model for the sample data.

## T-Squared Chart

As with all capability analyses, it is important that the process be in a state of statistical control when the capability is estimated. In the case of multivariate data, a *T-Squared Chart* is useful for examining whether the data appear to have been sampled from an in-control process.

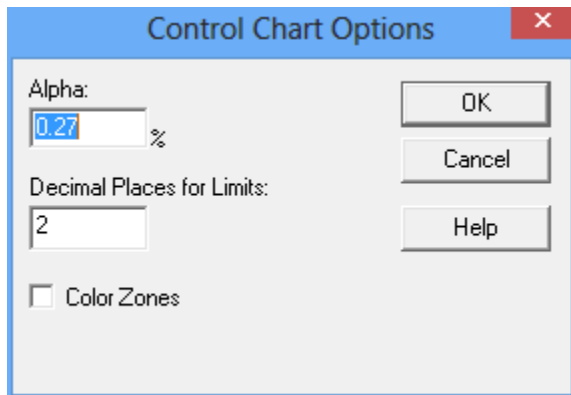


In a T-squared chart, each sample is represented by a single statistic defined by

$$T_i^2 = (x_i - \bar{x})' S^{-1} (x_i - \bar{x})$$

where  $x_i$  is the vector of observed values corresponding to the  $i$ -th sample,  $\bar{x}$  is the vector of sample means, and  $S$  is the sample covariance matrix. Included on the chart is an upper control limit, placed by default to cover 99.73% of a multivariate normal distribution with the estimated means and covariances. Any point beyond the control limit would be a signal of a potentially out-of-control situation.

## Pane Options



- **Alpha:** the false alarm probability of the control chart. For a standard 3-sigma control chart,  $\alpha = 0.27\%$ .
- **Decimal Places for Limits:** the number of decimal places used to display the control limit.
- **Color Zones:** check this box to display green and red zones.

## Multivariate Tolerance Limits

This table displays statistical tolerance limits for the multivariate data. It includes a tolerance region that bounds a selected  $p\%$  of the population with  $100(1-\alpha)\%$  confidence. It also includes joint simultaneous tolerance limits for each of the variables using a Bonferroni approach. The data are assumed to be a random sample from a multivariate normal distribution. The multivariate tolerance limits may be compared to the specifications for the multiple variables to determine whether or not most of the population is within spec.

For the sample data, the pane displays:

<b>Multivariate Tolerance Limits</b>	
Number of observations = 56	
95% Simultaneous Bonferroni Tolerance Limits for 99% of the Population	
	<i>Upper Limit</i>
Small	13.4717
Large	11.3827
Observations beyond Bonferroni limits: 1	
95% Elliptical Tolerance Region for 99% of the Population: Squared distance $\leq$ 12.9356	
Observations outside elliptical region: 0	

### *Simultaneous tolerance limits*

The output includes simultaneous tolerance limits for each of the  $m$  variables using a Bonferroni approach. This approach calculates separate tolerance limits for each of the variables using the standard K-factor

$$\bar{x} \pm Ks \tag{6}$$

where  $\bar{x}$  and  $s$  are the sample mean and sample standard deviation of the selected variable. However, instead of using a K-factor corresponding to the desired level of confidence, it uses a  $K$  with a confidence level of

$$CL = 100 (1 - \alpha/m) \% \tag{7}$$

The resulting tolerance limits bound  $p\%$  of the joint distribution of the  $m$  variables with confidence equal to or greater than  $100(1-\alpha)\%$ .

For the sample data, we can state with 95% confidence that 99% of the joint distribution of the 2 variables is such that:

$$\text{small} \leq 13.3417$$

$$\text{large} \leq 11.3827$$



The Bonferroni limits are somewhat conservative, meaning that they may contain more than the stated population percentage.

### *Elliptical tolerance region*

An exact tolerance region for the  $m$  variables is also calculated. It takes the form

$$(\mathbf{X} - \bar{\mathbf{X}})^T \mathbf{S}^{-1}(\mathbf{X} - \bar{\mathbf{X}}) \leq c \quad (8)$$

where  $\mathbf{S}$  is the  $m$  by  $m$  sample covariance matrix. This corresponds to an elliptical region in  $m$  dimensions. A multivariate observation  $\mathbf{X}_i$  is within the tolerance region if the squared generalized distance from the centroid  $\bar{\mathbf{X}}$

$$d_i^2 = (\mathbf{X}_i - \bar{\mathbf{X}})^T \mathbf{S}^{-1}(\mathbf{X}_i - \bar{\mathbf{X}}) \quad (9)$$

is no greater than  $c$ . The critical distance  $c$  depends on the number of variables  $m$ , the sample size  $n$ , the coverage percentage  $p$ , and the confidence level  $100(1-\alpha)\%$ .

Since there is no theoretical way to calculate  $c$  and the available approximations are not satisfactory for all combinations of  $m$ ,  $n$ ,  $p$  and  $\alpha$ , it is necessary to use a Monte Carlo simulation to obtain the value of  $c$ . Statgraphics does so using Algorithm 9.2 described by Krishnamoorthy and Mathew (2009). Whenever a value of  $c$  is needed, it is obtained using that algorithm with 100,000 repetitions. Tests have indicated that the value of  $c$  obtained is quite stable with that many repetitions, and the values obtained are very close to those tabulated by the authors in their textbook.

The output indicates that a 95% elliptical tolerance region for 99% of the population is given by equation (8) with  $c = 12.9356$ .

### *Distance table*

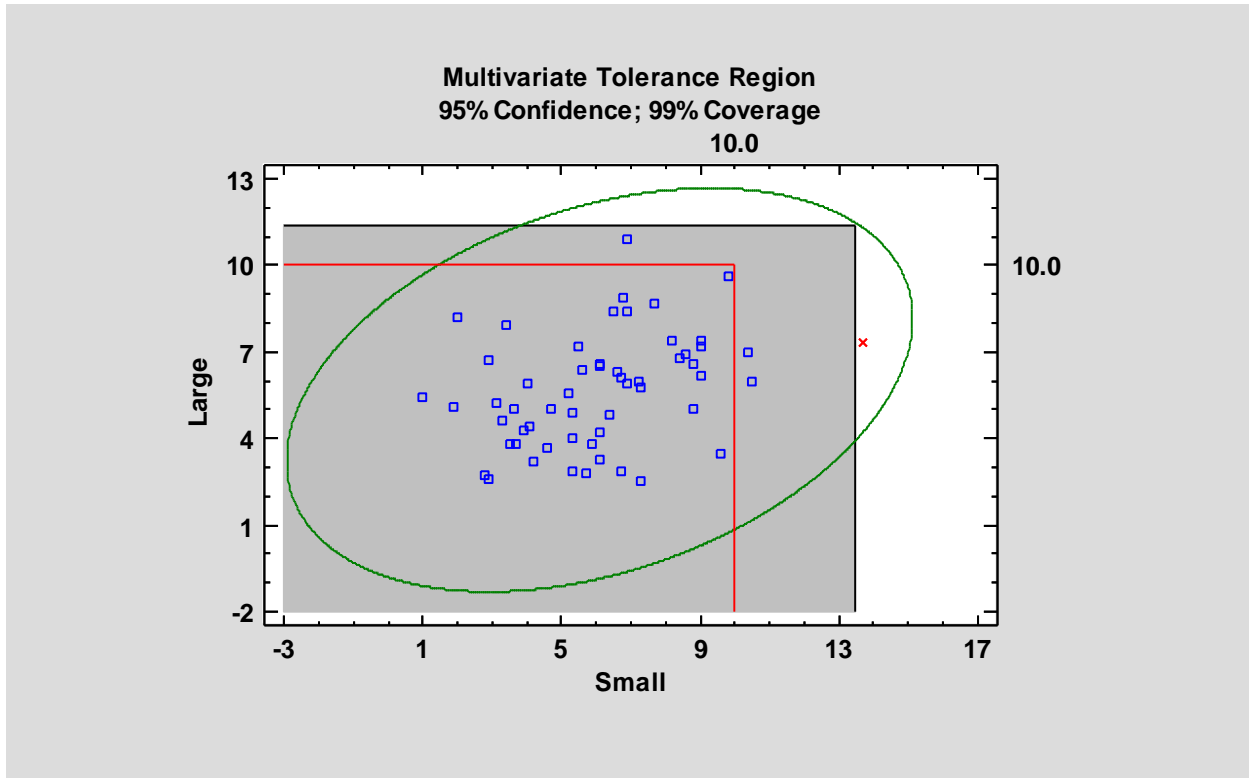
The pane also includes a table displaying each of the observations and the calculated squared distance  $d_i^2$  from equation (9), setting  $m$  equal to the total number of variables. A portion of that table is shown below:

E:Outside elliptical region B:Beyond Bonferroni limits				
Row	Small	Large	Squared distance	Beyond limits
1	1.0	5.4	4.49564	
2	4.2	3.2	1.73946	
3	3.1	5.2	1.45972	
4	9.6	3.5	4.9331	
5	6.7	2.9	2.69002	
6	3.3	4.6	1.27175	
7	4.1	4.4	0.797062	
8	4.7	5.0	0.337169	
...	...	...	...	...

The *Beyond limits* column indicates any point which is beyond the tolerance limits using each of the two methods.

## Tolerance Region

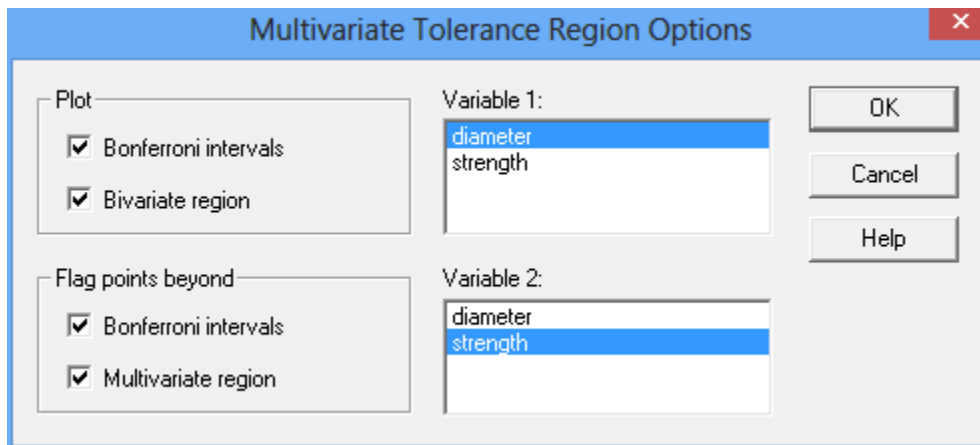
This option displays the data and calculated tolerance regions for any pair of observations:



By default, it displays both the Bonferroni limits (the shaded region) and the elliptical tolerance region. The Bonferroni limits are calculated using equation (6) with  $m$  equal to the total number of variables. The elliptical tolerance region is calculated using equation (8) with  $m = 2$  based on only the 2 variables being plotted. By default, points outside either of the tolerance limits are shown using a red X.

The red lines are drawn at the specification limits. Ideally, the tolerance regions would be entirely within the specification limits.

### Pane Options



- **Plot:** the regions to plot on the graph.
- **Flag points beyond:** the regions used to determine which points are plotted using a red X.
- **Variable 1:** the variable displayed on the horizontal axis.
- **Variable 2:** the variable displayed on the vertical axis.

## Calculations

### Sample Mean Vector

p by 1 vector with elements:

$$\bar{x}_j = \frac{1}{n} \sum_{i=1}^n x_{ij}$$

### Sample Covariance Matrix

p by p matrix with elements:

$$s_{jk} = \frac{1}{n-1} \sum_{i=1}^n (x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k)$$

### Control Limit on T-Squared Chart

$$UCL = \frac{(n-1)^2}{n} \text{Beta}_{\alpha, p/2, (n-p-1)/2}$$

## References

Holmes, D.S. and Mergen, A.E. (1993) "Improving the performance of the T<sup>2</sup> Control Chart", Quality Engineering, Vol. 5(4), pp. 619-625.

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

Krishnamoorthy, K. and Mathew, T. (2009). Statistical Tolerance Regions: Theory, Applications, and Computation. John Wiley and Sons, Hoboken, N.J.