Compare – Two Response Variables

This procedure compares the data contained in two *Response* columns. It compares the means using a t-test and the variances using an F test. It also displays the data in various ways.

The data for this analysis consist of samples from *2* populations. Let

x_{ij} = i-th measurement in sample *j*
n_j = size of sample *j*

Access

Highlight: 2 *Response* columns.

Select: *Compare* from the main menu.

Specify: You will be asked whether the data in the two columns are paired. If you answer *Yes*, the differences between the values in each row will be calculated and a one-variable analysis will be performed on the pairwise differences. For information on that type of analysis, consult *Describe – Response Variable*. In this document, we consider the case in which the data are not paired, but instead represent independent samples from two populations.

Output Page 1: A scatterplot of the data in each sample, together with summary statistics.

Output Page 2: A box-and-whisker plot for the data in each sample, together with the results of statistical tests that compare the samples.

Output Page 3: A plot of the sample means with Fisher's LSD intervals.

Output Page 4: A plot of the variability within each group.

Options

The default assumption is that the data are random samples from a normal distribution. To select a different distribution:

1. Access the *Properties* dialog box for the leftmost *Response* variable by double-clicking on its column header.

2. On the *Dist.* tab, select the assumed distribution. The default selection assumes that the data follow a normal distribution. If you select *Lognormal*, the logarithms of the data will

be assumed to follow a normal distribution. If you select *Power normal*, the data will be assumed to follow a normal distribution after raising them to the indicated *power*.

The confidence level used when constructing confidence intervals and the means plot is determined by the *Confidence* field on the *Pred*. tab of the *Column Properties* dialog box for the leftmost response column:

The alternative hypotheses used by the t-test and F-test are set using the *Comp*. tab, as is the type of intervals plotted by the *Means Plot*.

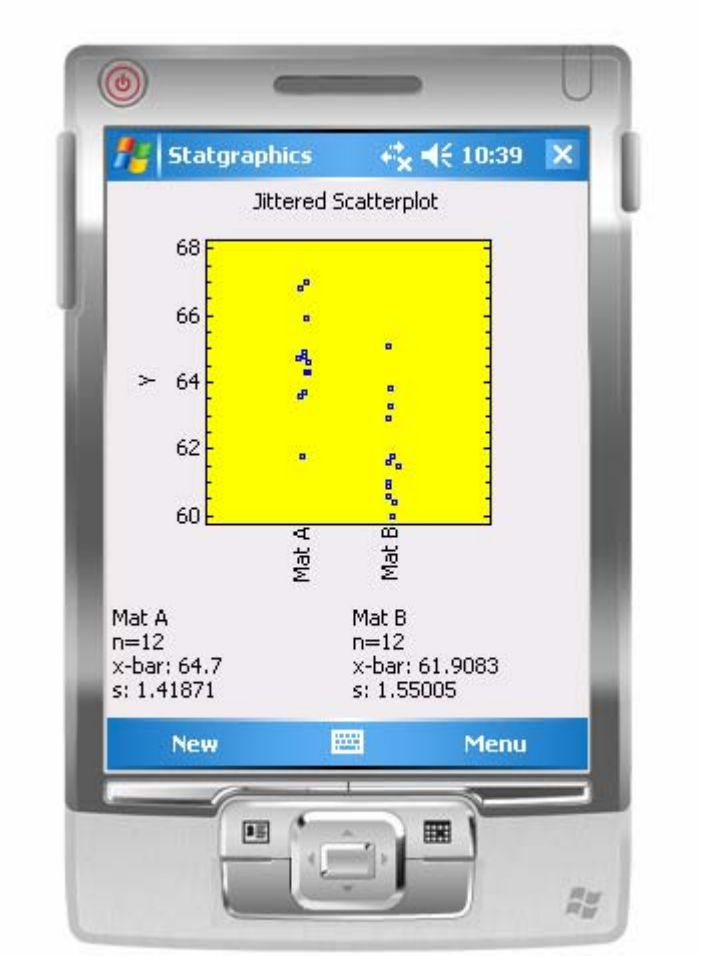
Sample Data

The file *widgets.sgm* contains data on the strength of widgets made from $m = 4$ different types of material. $n = 12$ widgets were samples from each of the four materials. In this document, we will compare the data for materials A and B, which are shown below.

Mat A	Mat B
64.7	60.4
64.8	61.8
66.8	63.3
67.0	61.6
64.9	61.0
63.7	63.8
61.8	60.9
64.3	65.1
64.3	61.5
65.9	60.0
63.6	62.9
64.6	60.6

Scatterplot

The *Scatterplot* displays the data in each column.



The points are jittered (randomly off in the horizontal direction) to help prevent overplotting. Also displayed is the number of observations in each variable n_j, the sample mean

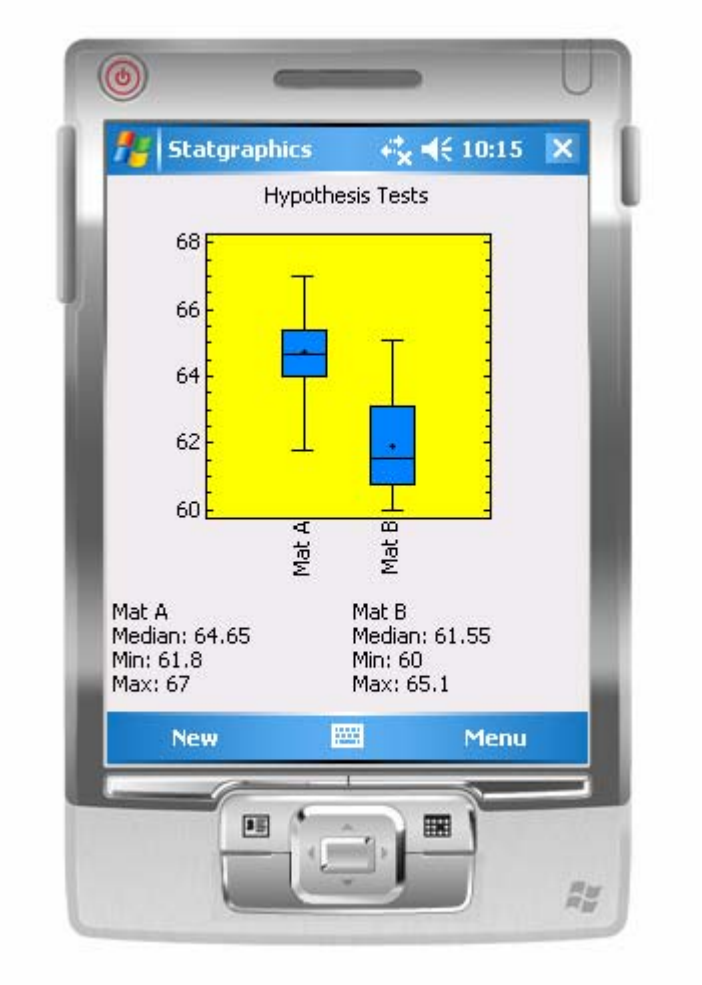
$$\bar{x}_j = \frac{\sum_{i=1}^{n_j} x_{ij}}{n_j} \tag{1}$$

and the sample standard deviation

$$s_j = \sqrt{\frac{\sum_{i=1}^{n_j} \left(x_{ij} - \bar{x}_j\right)^2}{n_j - 1}} \tag{2}$$

Boxplots

The *Boxplots* page displays a box-and-whisker plot for the data in each column.



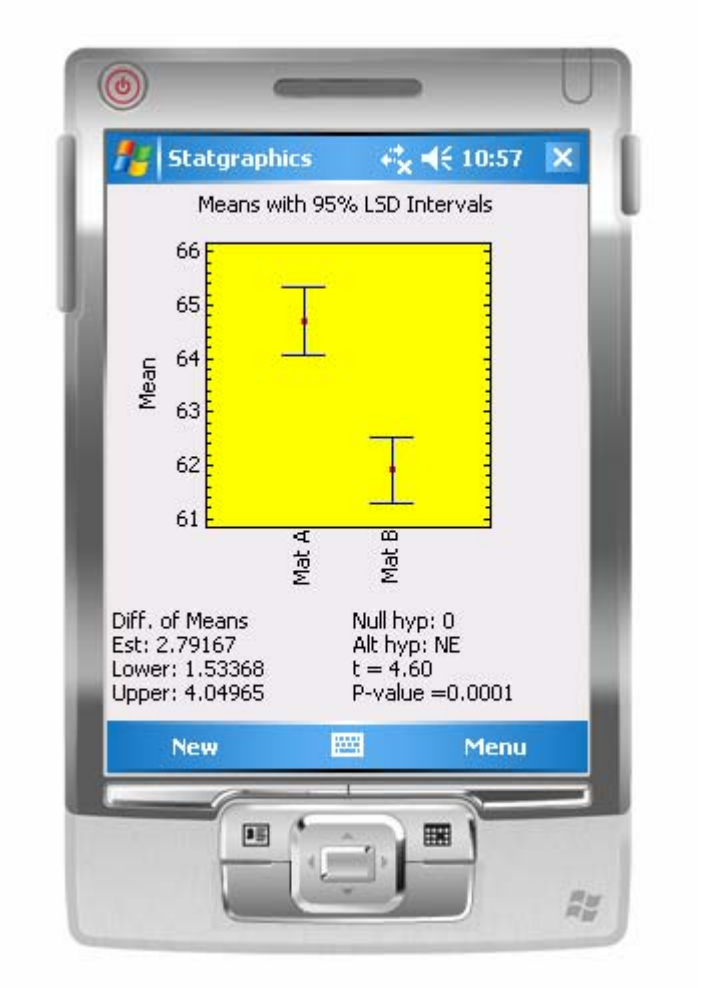
The plot is constructed in the following manner:

- A box is drawn extending from the *lower quartile* of each sample to the *upper quartile*. This is the interval covered by the middle 50% of the data values when sorted from smallest to largest.
- A horizontal line is drawn at the *median* (the middle value).
- A plus sign is placed at the location of the sample mean.
- Whiskers are drawn from the edges of the box to the largest and smallest data values.

Below the plot are summary statistics for each sample.

Means Plot

The *Means Plot* page displays the mean of each row as a point symbol, together with uncertainty intervals.



The point symbols are located at the sample means. The bars extend over uncertainty intervals, the type of which is determined by the settings on the *Comp*. tab of the leftmost *Response* column. Several different types of intervals may be constructed:

- **Standard errors (pooled s)** - displays the standard errors using the pooled within-sample standard deviation:

$$\bar{x}_j \pm \sqrt{\frac{s_p^2}{n_j}} \tag{3}$$

where

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \tag{4}$$

- **Standard errors (individual s)** - displays the standard errors using the standard deviation of each sample separately:

$$\bar{x}_j \pm \sqrt{\frac{s_j^{\,2}}{n_j}} \tag{5}$$

- **Confidence intervals (pooled s)** - displays confidence intervals for the group means using the pooled within-group standard deviation:

$$\bar{x}_j \pm t_{\alpha/2,n_1+n_2-2}\sqrt{\frac{s_p^2}{n_j}} \tag{6}$$

- **Confidence intervals (individual s)** - displays confidence intervals for the sample means using the standard deviation of each group separately:

$$\bar{x}_j \pm t_{\alpha/2,n_j-1}\sqrt{\frac{s_j^{\,2}}{n_j}} \tag{7}$$

- **LSD intervals** - designed to compare the means with the stated confidence level. The intervals are given by

$$\bar{x}_j \pm \frac{\sqrt{2}t_{\alpha/2,n_1+n_2-2}}{2}\sqrt{\frac{s_p^2}{n_j}} \tag{8}$$

  where *t* represents the value of Student's t distribution with $n_1 + n_2 - 2$ degrees of freedom leaving an area of $\alpha/2$ in the upper tail of the curve. If the sample sizes are equal, you can determine whether or not the population means are significantly different from each other by determining whether or not the intervals overlap.

- **Tukey HSD Intervals** - designed for comparing multiple pairs of means. Since this procedure compares only two pairs, the intervals are the same as the LSD intervals.

The output also displays information about the difference between the population means $\Delta = \mu_1 - \mu_2$. It displays:

(1) **Est**.: the estimated difference

$$\hat{\Delta} = \bar{x}_1 - \bar{x}_2 \tag{9}$$

(2) **Lower and upper**: confidence limits for the difference

$$\left(\bar{x}_1 - \bar{x}_2\right) \pm t_{\alpha/2,n_1+n_2-2}s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}} \tag{10}$$

(3) **t**: the calculated t statistic

$$t = \frac{(\bar{x}_1 - \bar{x}_2)}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \tag{11}$$

which is used to the hypotheses

Null hypothesis: $\mu_1 - \mu_2 = 0$

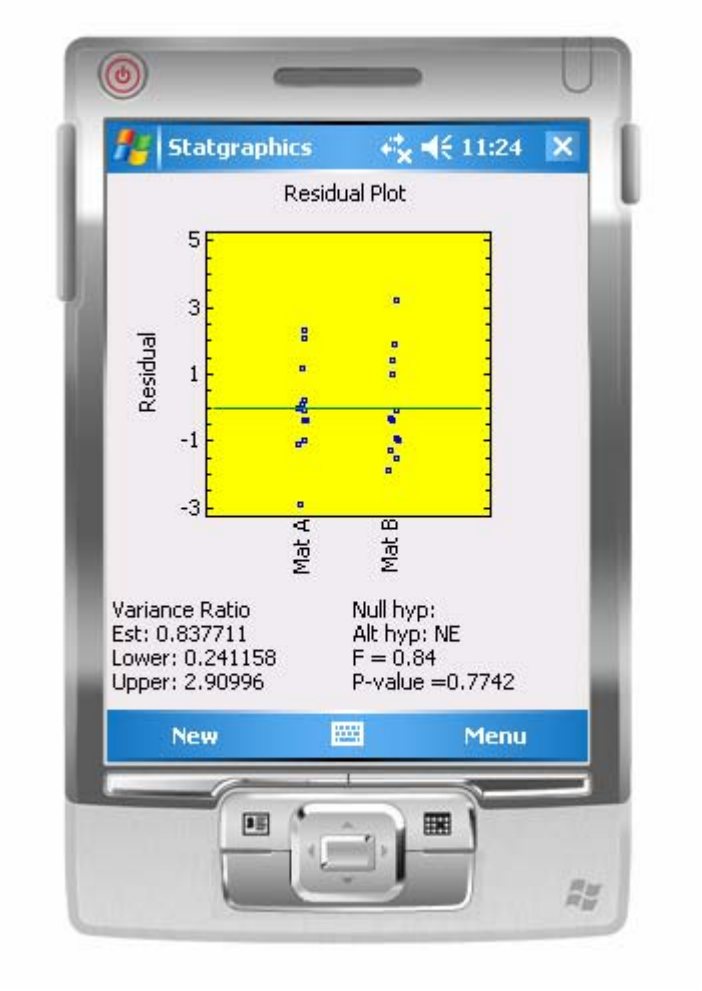
Alt. hypothesis: $\mu_1 - \mu_2 \neq 0$

(4) **P-value**: calculated P-value for the t-test computed by comparing the observed t-statistic to Student's t distribution with $n_1 + n_2$ - 2 degrees of freedom. Small P-values, as in the current example, indicate a statistically significant difference between the two population means.

NOTE: If a non-normal distribution has been specified for the leftmost *Response* column, the estimated means, difference of means, confidence limits, and hypothesis test are calculated *in the transformed metric*. An inverse transformation is applied to the results when the graph is created.

Residuals

The *Residuals* plots the deviations of each observation from its respective sample mean.



The plot may be used to help identify outliers.

Also displayed is information about the variance ratio $\rho = \sigma_1^2 / \sigma_2^2$. The output shows:

(1) **Est**.: the estimated ratio

$$\hat{\rho} = \frac{s_1^2}{s_2^2} \tag{8}$$

(2) **Lower and upper**: confidence limits for the ratio, calculated from

$$\left[\frac{s_1^2}{s_2^2} \frac{1}{F_{\alpha/2, n_1-1, n_2-1}}, \frac{s_1^2}{s_2^2} F_{\alpha/2, n_2-1, n_1-1} \right] \tag{9}$$

(3) **F**: the calculated F statistic

$$F = \frac{s_2^1}{s_2^2} \tag{10}$$

which is used to the hypotheses

Null hypothesis: $\frac{\sigma_2^1}{\sigma_2^2} = 1$

Alt. hypothesis: $\frac{\sigma_2^1}{\sigma_2^2} \neq 1$

(4) **P-value**: calculated P-value for the F-test computed by comparing the observed F-statistic to Snedecor's F distribution with n_1 -1 and n_2 - 1 degrees of freedom. Small P-values (below 0.05 is operating at the 5% significance level) indicate a statistically significant difference between the two population variances. In the example above, the P-value is well above 0.05, indicating that the population variances are not significantly different.

NOTE: If a non-normal distribution has been specified for the leftmost *Response* column, the estimated variance ratio, confidence limits, and hypothesis test are calculated *in the transformed metric*.