Multiple Sample Comparison

statgraphics 18®
centurion

Revised: 10/11/2017

Summary

The **Multiple Sample Comparison** procedure is designed to compare two or more independent samples of variable data. Tests are run to determine whether or not there are significant differences between the means, variances, and/or medians of the populations from which the

samples were taken. In addition, the data may be displayed graphically in various ways, including a multiple scatterplot, a means plot, an ANOM plot, and a medians plot.

The output of this procedure is identical to that of the *One-Way ANOVA* procedure.

Sample StatFolio: *multiple samples.sgp*

Sample Data

The file *pulse rates.sf6* contains the results of an experiment reported by Milliken and Johnson (1992) in which 78 workers were assigned at random to six groups. Each group was given a work task to perform, and pulse rates were measured after each individual had worked on his assigned task for one hour. After several individuals dropped out of the study, the final data were:

Task 1	Task 2	Task 3	Task 4	Task 5	Task 6
27	29	34	34	28	28
31	28	36	34	28	26
26	37	34	43	26	29
32	24	41	44	35	25
39	35	30	40	31	35
37	40	44	47	30	34
38	40	44	34	34	37
39	31	32	31	34	28
30	30	32	45	26	21
28	25	31	28	20	28
27	29			41	26
27	25			21	
34					

The final $n = 68$ measurements have been arranged in $q = 6$ columns, one for each group of subjects.

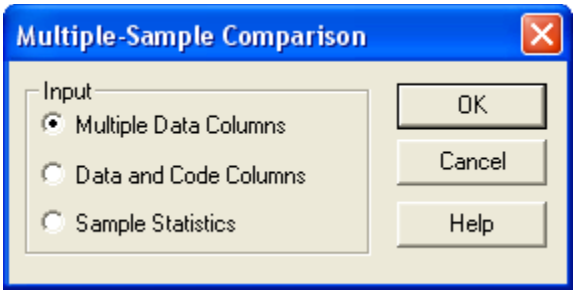
Alternatively, the data could have been arranged in a table with all of the pulse rates in a single column, together with a column identifying which task the subject was given. A portion of such a file is shown below:

Subject	*Pulse Rate*	*Task*
1	27	1
2	31	1
3	26	1
4	32	1
5	39	1
6	37	1
7	38	1
8	39	1
9	30	1
10	28	1
11	27	1
12	27	1
13	34	1
14	29	2
15	28	2
16	37	2
17	24	2
18	35	2
19	40	2
20	40	2
21	31	2
22	30	2
23	25	2
24	29	2
25	25	2
26	34	3
…	…	…

Either data structure can be analyzed by the *Multiple Sample Analysis* procedure. If the same data is to be used in other procedures such as the *General Linear Models* procedure, it should be structured in the second manner. As part of the *Save Results* option in this procedure, you can take a dataset that is structured in the multiple column format and rearrange it as data and code columns.

Data Input

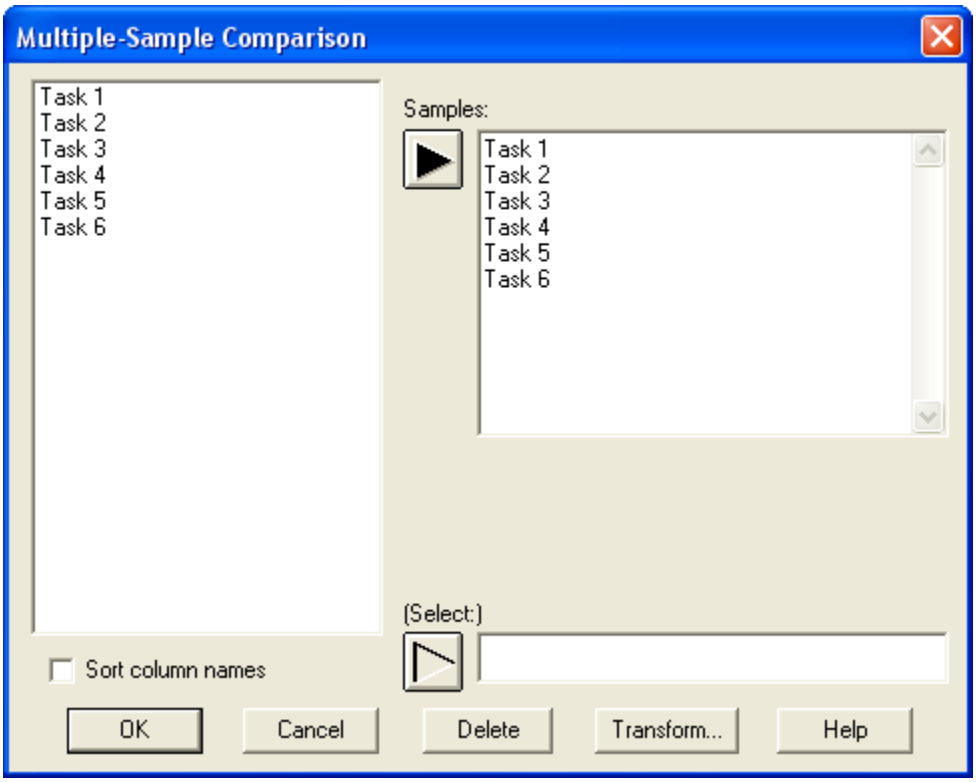
When the *Multiple Sample Comparison* procedure is selected from the main menu, the first dialog box displayed asks you to specify the format in which the data have been entered:



- **Multiple Data Columns**: indicates that each sample has been placed into a separate column.
- **Data and Code Columns**: indicates that all observations have been placed into a single column, with a second column indicating which sample each observation belongs to.
- **Sample Statistics**: indicates that the original observations are not available. However, the sample sizes, sample means, and sample standard deviations have been placed into 3 columns of the data sheet. In this case, some options will not be available.

Multiple Data Columns

If the data have been placed in separate columns for each sample, the column names must be entered on the second dialog box:

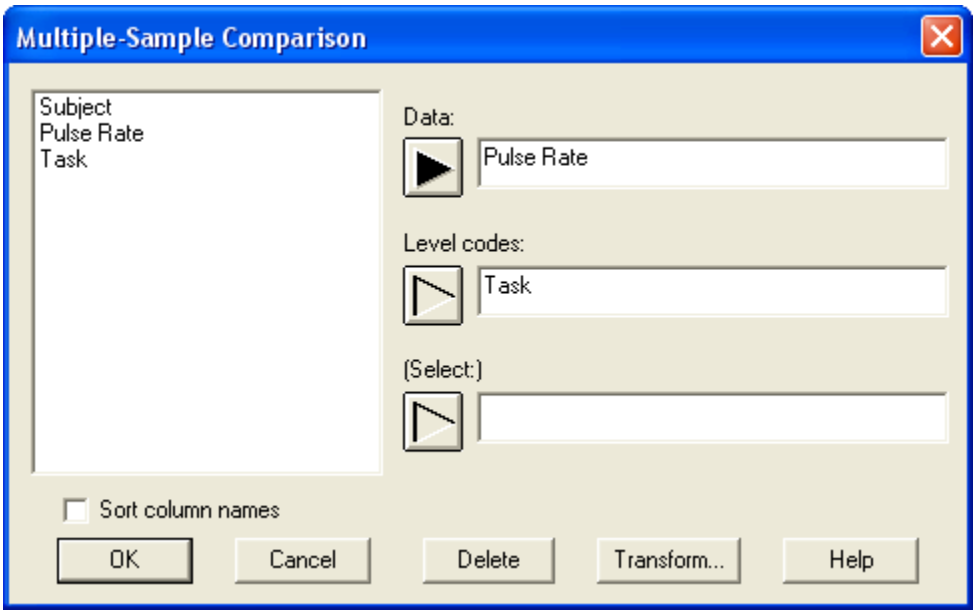


- **Samples:** two or more numeric columns containing the observations, one column for each sample.

- **Select:** subset selection.

Data and Code Columns

If the data from all samples have been placed into a single column, then enter the name of that column and the column containing the group identifiers:



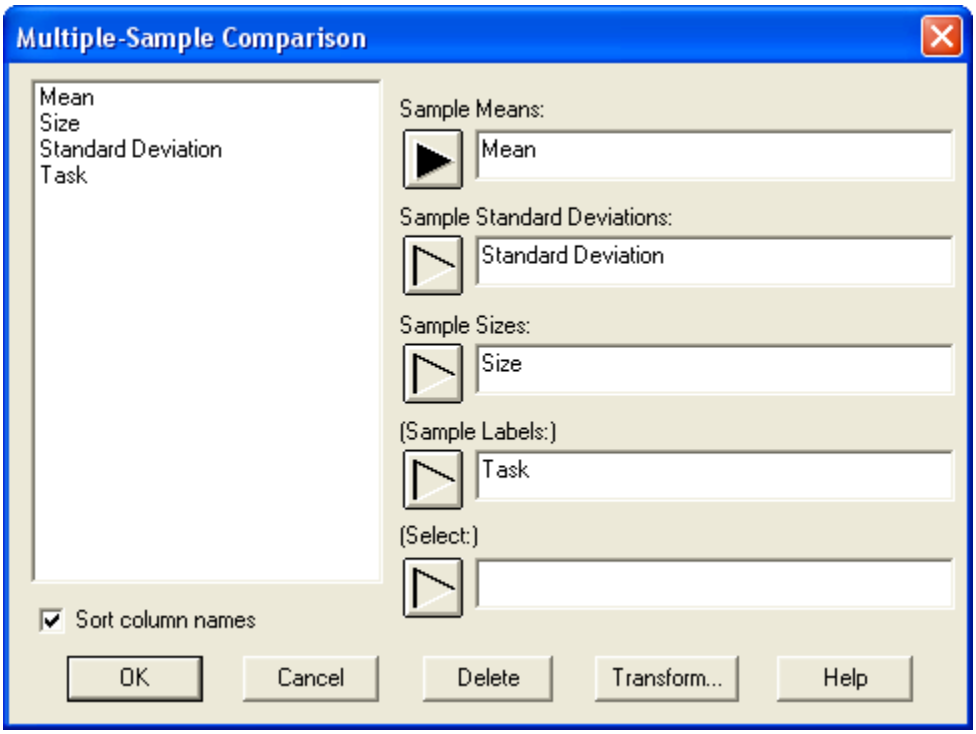
- **Data:** numeric column containing the observations from all samples.
- **Level codes:** numeric or non-numeric column containing an identifier for the sample corresponding to each data value.
- **Select:** subset selection.

Sample Statistics

If the original observations are not available but the means and standard deviations of each sample are known, enter the sample statistics into separate columns of the datasheet:

Task	Size	Mean	Standard Deviation
1	13	31.9231	4.95751
2	12	31.0833	5.66422
3	10	35.8000	5.30827
4	10	38.0000	6.59966
5	12	29.5000	6.00757
6	11	28.8182	4.75012

Then complete the second dialog box as shown below:



- **Sample Means**: numeric column containing the means of each sample.
- **Sample Standard Deviations**: numeric column containing the standard deviations of each sample.
- **Sample Sizes**: numeric column containing the sizes of each sample.
- **Sample Labels**: optional column containing labels for each sample.
- **Select**: subset selection.

Analysis Options

If the data are specified using the Data and Code columns format, this dialog box controls the order of the factor levels.

- **Levels**: drag the labels to change the order of the factors in all tables and graphs.
- **Treat levels as numeric**: if checked, graphs will plot the levels on a numeric axis rather than treating them as categorical.

Analysis Summary

The *Analysis Summary* shows the number of observations in each sample.

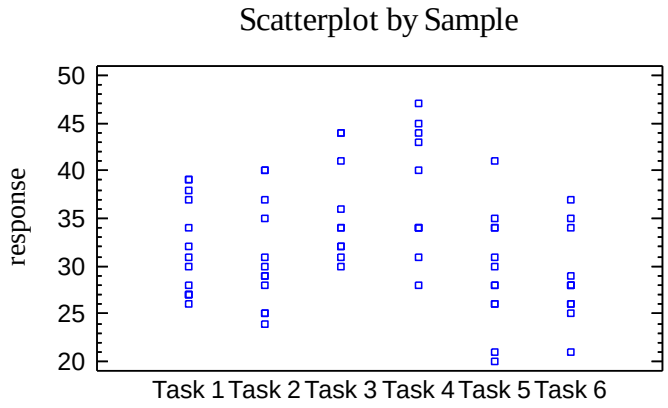
Multiple-Sample Comparison
Sample 1: Task 1
Sample 2: Task 2
Sample 3: Task 3
Sample 4: Task 4
Sample 5: Task 5
Sample 6: Task 6

Sample 1: 13 values ranging from 26.0 to 39.0
Sample 2: 12 values ranging from 24.0 to 40.0
Sample 3: 10 values ranging from 30.0 to 44.0
Sample 4: 10 values ranging from 28.0 to 47.0
Sample 5: 12 values ranging from 20.0 to 41.0
Sample 6: 11 values ranging from 21.0 to 37.0

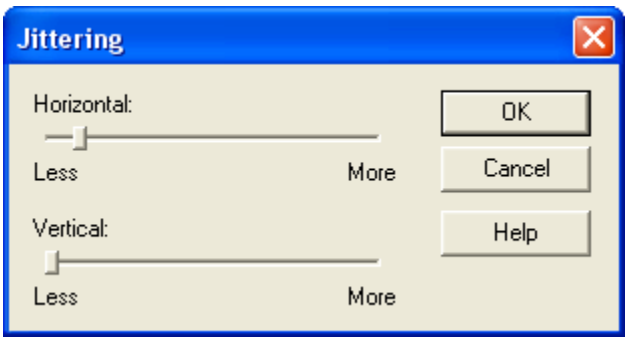
Also displayed are the largest and smallest values.

Scatterplot

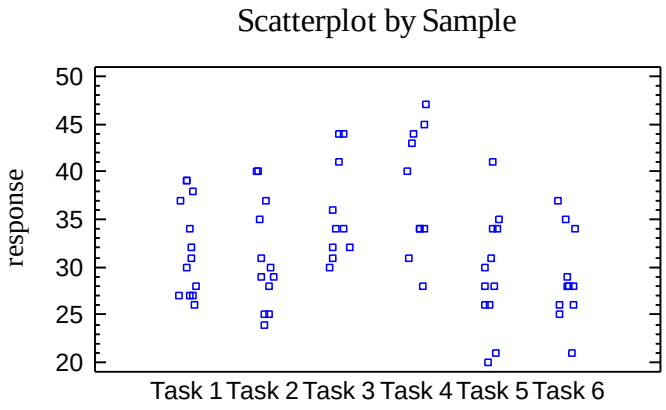
The *Scatterplot* pane plots the data within each group.



If there are many common values, you may wish to add a small amount of horizontal jitter to the plot by pressing the *Jitter* button on the analysis toolbar:



This offsets each point randomly in the horizontal direction so that identical values do not plot on top of each other:



The above plot seems to suggest that pulse rates are somewhat higher for subjects assigned to tasks 3 and 4. Note: jittering the scatterplot has no effect on any calculations.

Summary Statistics

The *Summary Statistics* pane calculates a number of different statistics that are commonly used to summarize a sample of variable data:

Summary Statistics

	Count	*Average*	*Standard deviation*	*Range*	*Stnd. skewness*	*Stnd. kurtosis*
Task 1	13	31.9231	4.95751	13.0	0.5224	-1.15872
Task 2	12	31.0833	5.66422	16.0	0.705281	-0.742333
Task 3	10	35.8	5.30827	14.0	0.978475	-0.694096
Task 4	10	38.0	6.59966	19.0	-0.0711101	-1.01365
Task 5	12	29.5	6.00757	21.0	0.21987	-0.0371778
Task 6	11	28.8182	4.75012	16.0	0.529821	-0.202849
Total	68	32.3088	6.24203	27.0	1.30662	-0.704478

Most of the statistics fall into one of three categories:

1. measures of *central tendency* – statistics that characterize the "center" of the data.
2. measure of *dispersion* – statistics that measure the spread of the data.
3. measures of *shape* – statistics that measure the shape of the data relative to a normal distribution.

The statistics included in the table by default are controlled by the settings on the *Stats* pane of the *Preferences* dialog box. Within the procedure, the selection may be changed using *Pane Options*. For a detailed description of each statistic, see the *One Variable Analysis* documentation.

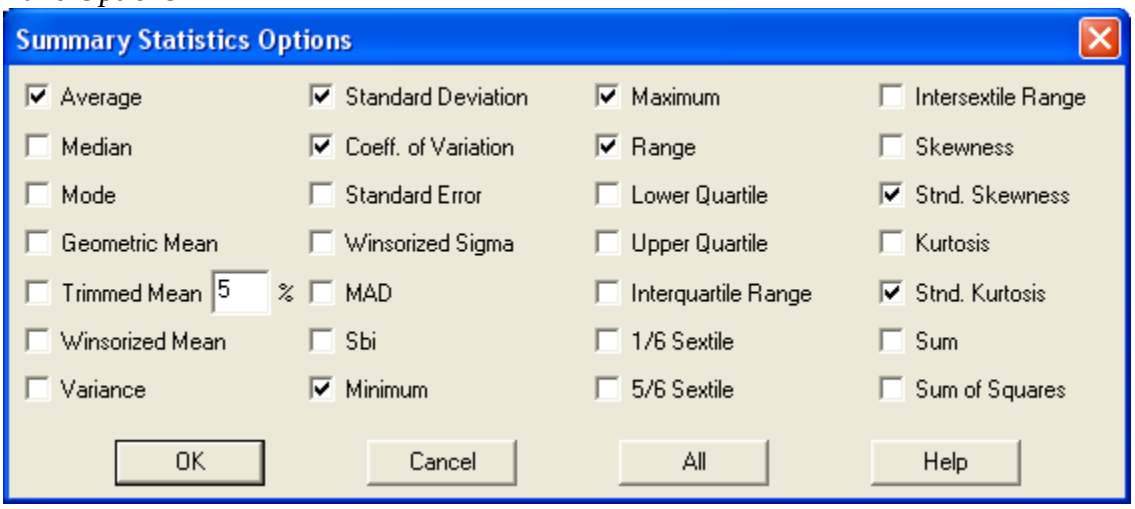
Of particular interest are:

1. *Sample means* $\bar{Y}_j$: the average pulse rate for the subjects given each of the 6 tasks.
2. *Sample standard deviations* s_j: the standard deviations of each group.

3. *Standardized skewness and kurtosis*: These statistics should be between –2 and +2 if the data come from normal distributions.

For the pulse rates, the average rate was highest for group 4, as was the standard deviation. All of the standardized skewness and kurtosis statistics are within the range expected for data from normal distributions.

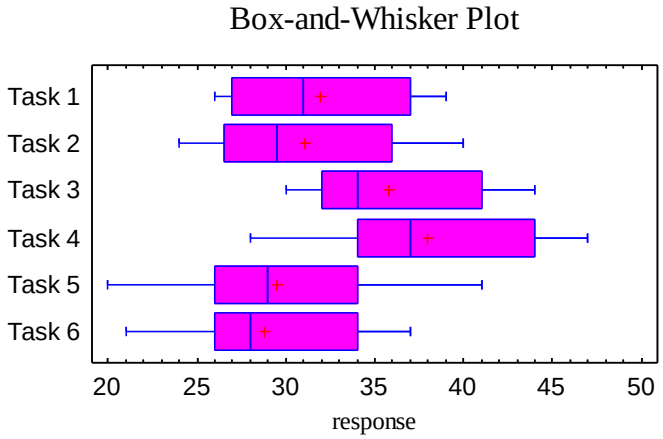
Pane Options



Select the desired statistics.

Box-and-Whisker Plot

This pane displays a box-and-whisker plot for each sample.



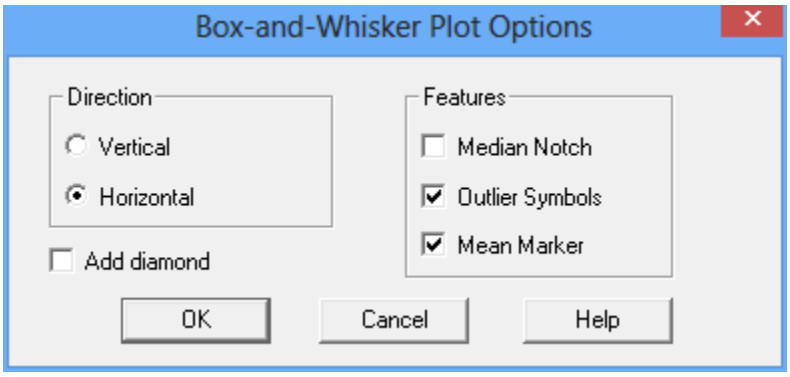
Box-and-whisker plots are constructed in the following manner:

- A box is drawn extending from the *lower quartile* of the 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 vertical line is drawn at the *median* (the middle value).
- If requested, 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, unless there are values unusually far away from the box (which Tukey calls *outside points*). Outside points, which are points more than 1.5 times the interquartile range (box width) above or below the box, are indicated by point symbols. Any points more than 3 times the interquartile range above or below the box are called *far outside points*, and are indicated by point symbols with plus signs superimposed on top of them. If outside points are present, the whiskers are drawn to the largest and smallest data values which are not outside points.

In the sample data, the variability appears to be similar within each sample, although the locations show some differences. There are no outside points.

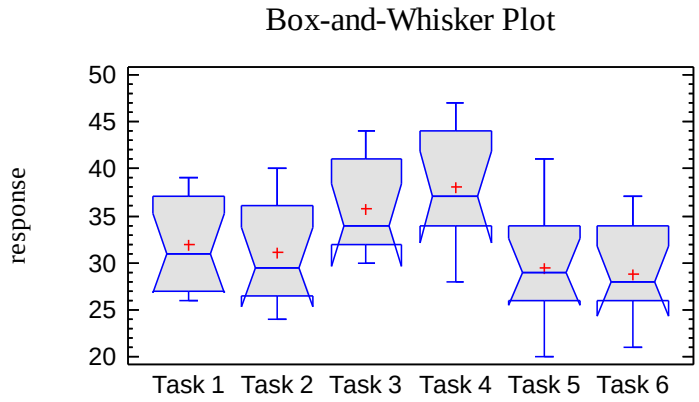
Pane Options



- **Direction**: the orientation of the plot, corresponding to the direction of the whiskers.
- **Median Notch**: If selected, a notch will be added to the plot showing the estimation error associated with each median. The notches are scaled in such a way that, for samples of equal size, if they do not overlap, the two medians are significantly different at the default system confidence level (set on the *General* tab of the *Preferences* dialog box on the *Edit* menu).
- **Outlier Symbols**: if selected, indicates the location of outside points.
- **Mean Marker**: if selected, shows the location of the sample mean as well as the median.
- **Add diamond:** if selected, a diamond will be added to the plot showing a 100(1-α)% confidence interval for the mean at the default system confidence level.

Example – Notched Box-and-Whisker Plot

The following plot adds median notches at the 95% confidence level.



Each notch covers the interval

$$\tilde{x}_j \pm \frac{z_{\alpha/2}}{2} \frac{1.25(IQR_j)}{1.35\sqrt{n_j}} \left(1 + \frac{1}{\sqrt{2}}\right) \qquad (1)$$

where $\tilde{x}_j$ is the median of the j-th sample, IQR_j is the sample interquartile range, n_j is the sample size, and $z_{\alpha/2}$ is the upper $(\alpha/2)\%$ critical value of a standard normal distribution. In cases where the sample size is small, the notch may extend beyond the box, resulting in a folding back appearance.

Since the samples vary somewhat in size, the overlap rule will not work perfectly. However, the notches for tasks 4 and 6 do not overlap, which would typically indicate a significant difference between those two medians at the 5% significance level.

ANOVA Table

In order to determine whether or not the means of the q groups are significantly different from each other, a oneway analysis of variance can be performed. The results are displayed in the *ANOVA Table*:

ANOVA Table

Source	*Sum of Squares*	*Df*	*Mean Square*	*F-Ratio*	*P-Value*
Between groups	694.439	5	138.888	4.49	0.0015
Within groups	1916.08	62	30.9045		
Total (Corr.)	2610.51	67			

The table divides the overall variability among the n measurements into two components:

1. A "within groups" component, which measures the variability among pulse rates of subjects given the same task.

2. A "between groups" component, which measures the variability among subjects given different tasks.

Of particular importance is the F-ratio, which tests the hypothesis that the mean response for all samples is the same. Formally, it tests the null hypothesis

H_0: $\mu_1 = \mu_2 = ... = \mu_q$

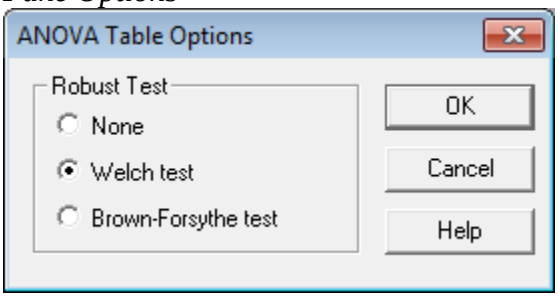
versus the alternative hypothesis

H_A: not all μ_j equal

If F is sufficiently large, the null hypothesis is rejected.

The statistical significance of the F-ratio is most easily judged by its P-value. If the P-value is less than 0.05, the null hypothesis of equal means is rejected at the 5% significance level, as in the current example. This does not imply that every mean is significantly different from every other mean. It simply implies that the means are not all the same. Determining which means are significantly different from which others requires additional tests, as discussed below.

Pane Options



Two additional robust tests for equality of the group means are also available. Unlike the F test in the ANOVA table, the Welch and Brown-Forsythe tests do not depend on the assumption that the variances within all of the groups are equal. (Note: the hypothesis of equal or *homogeneous* variances may be tested using the *Variance Check* described below.)

The robust tests add an additional table to the output pane:

Robust Test

	Test Statistic	*Df1*	*Df2*	*P-Value*
Welch test	3.78755	5.00	28.33	0.0095

The table displays the test statistic, its numerator and denominator degrees of freedom, and the resulting P-Value. The interpretation of the P-value is the same as for the F test in the ANOVA table, with small P-Values leading to the conclusion that there are statistically significant differences amongst the group means.

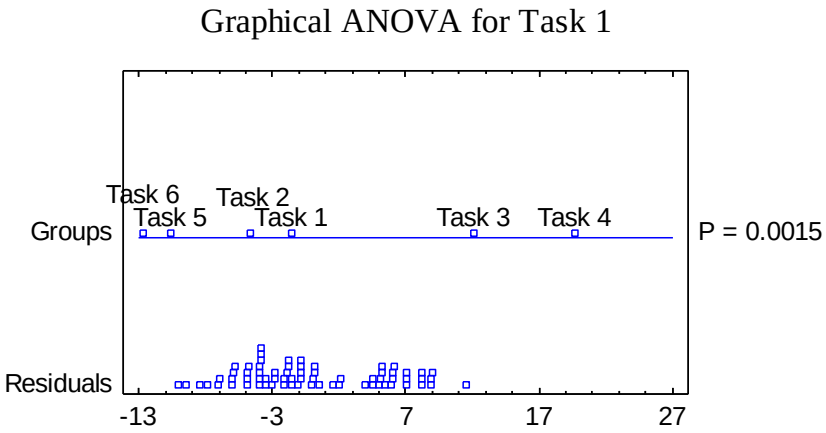
Graphical ANOVA

The *Graphical ANOVA* plot, developed by Hunter (2005), is a technique for displaying graphically the importance of the differences between the samples. It is a plot of the scaled effects, where the "effect" equals the difference between a sample mean and the estimated grand mean. Each of the effects is multiplied by a scaling factor

$$\sqrt{\frac{\nu_R n_i}{\nu_T \bar{n}}} \tag{2}$$

where ν_R is the residual degrees of freedom, ν_T is the degrees of freedom between group differences, n_i equals the number of observations in the *i-th* level of the group, and $\bar{n}$ is the average number of observations in all groups. This scales the effects so that the natural variance of the points in the diagram is comparable to that of the residuals, which are displayed at the bottom of the plot.

The plot for the sample data is shown below:



Along the right-hand side of the display is the P-Value for between group differences, taken from the ANOVA table.

By comparing the variability amongst effects in the above plot to that of the residuals, it is easy to see that the differences are of a greater magnitude than could be accounted for solely by experimental error. Depending upon the relative location of the effects, it may also be possible in some cases to visually identify which samples are significantly different from which other samples, which is done formally by the *Multiple Range Tests* described below.

Multiple Range Tests

To determine which sample means are significantly different from which others, the *Multiple Range Tests* can be performed:

Multiple Range Tests

Method: 95.0 percent LSD

	Count	*Mean*	*Homogeneous Groups*
Task 6	11	28.8182	X
Task 5	12	29.5	X
Task 2	12	31.0833	XX
Task 1	13	31.9231	XX
Task 3	10	35.8	XX
Task 4	10	38.0	X

Contrast	*Sig.*	*Difference*	*+/- Limits*
Task 1 - Task 2		0.839744	4.44862
Task 1 - Task 3		-3.87692	4.67423
Task 1 - Task 4	*	-6.07692	4.67423
Task 1 - Task 5		2.42308	4.44862
Task 1 - Task 6		3.1049	4.55256
Task 2 - Task 3		-4.71667	4.75816
Task 2 - Task 4	*	-6.91667	4.75816
Task 2 - Task 5		1.58333	4.53672
Task 2 - Task 6		2.26515	4.63869
Task 3 - Task 4		-2.2	4.96973
Task 3 - Task 5	*	6.3	4.75816
Task 3 - Task 6	*	6.98182	4.85547
Task 4 - Task 5	*	8.5	4.75816
Task 4 - Task 6	*	9.18182	4.85547
Task 5 - Task 6		0.681818	4.63869

* denotes a statistically significant difference.

The top half of the table displays each of the estimated sample means in increasing order of magnitude. It shows:

- **Count** - the number of observations n_j.
- **Mean** - the estimated sample mean $\overline{Y}_j$.
- **Homogeneous groups** - a graphical illustration of which means are significantly different from which others, based on the contrasts displayed in the second half of the table. Each column of X's indicates a group of means within which there are no statistically significant differences. For example, the first column in the above table contains an X for tasks 1, 2, 5, and 6, indicating that there are no significant differences amongst those four means. Likewise, tasks 1, 2, and 3 show no significant differences, nor do tasks 3 and 4. Any two tasks that do not have an X in the same column are significantly different from each other, such as tasks 4 and 6.
- **Difference** - the difference between the two sample means

$$\hat{\Delta}_{j_1 j_2} = \overline{Y}_{j_1} - \overline{Y}_{j_2} \qquad (3)$$

- **Limits** - an interval estimate of that difference, using the currently selected multiple comparisons procedure:

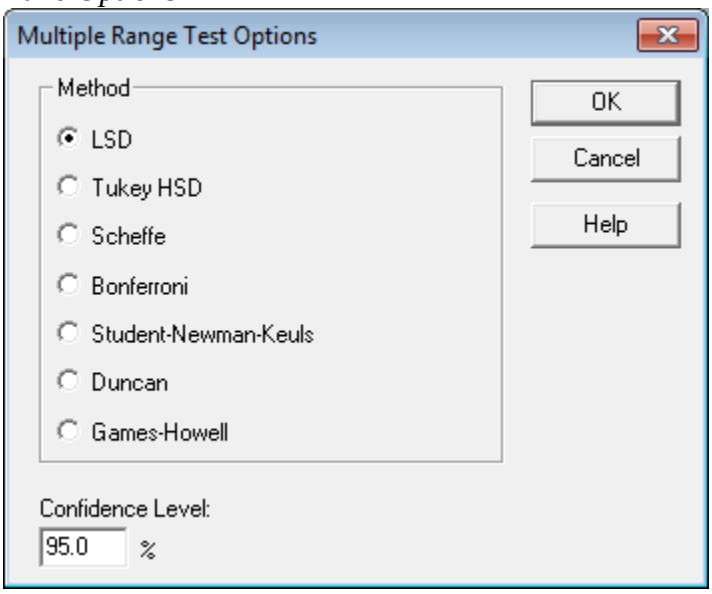
$$\hat{\Delta}_{j_1 j_2} \pm M\sqrt{MS_{within}\left(\frac{1}{n_{j_1}}+\frac{1}{n_{j_2}}\right)} \qquad (4)$$

  where M is a constant that depends upon the procedure selected.

- **Sig.** - An asterisk is placed next to any difference that is statistically significantly different from 0 at the currently selected significance level, i.e., any interval that does not contain 0.

For the pulse rate data, Task 4 has a significantly higher mean pulse rate than every task except Task 3. In addition, Task 3 is significantly higher than Tasks 5 and 6.

Pane Options



- **Method**: the method used to make the multiple comparisons.

- **Confidence Level**: the level of confidence used by the selected multiple comparison procedure.

The available methods are:

- **LSD** - forms a confidence interval for each pair of means at the selected confidence level using:

$$M = t_{\alpha/2,n-q} \qquad (5)$$

where *t* represents the value of Student's t distribution with *n* - *q* degrees of freedom leaving an area of $\alpha/2$ in the upper tail of the curve. This procedure is due to Fisher and is called the *Least Significant Difference* procedure, since the magnitude of the limits indicates the smallest difference between any two means that can be declared to represent a statistically significant difference. It should only be used when the F-test in the ANOVA table indicates significant differences amongst the sample means. The probability of making a Type I error α applies to each pair of means separately. If making more than one comparison, the overall probability of calling at least one pair of means significantly different when they are not may be considerably larger than α.

- **Tukey HSD** - widens the intervals to allows for multiple comparisons amongst all pairs of means, using:

$$M = T_{\alpha/2,q,n-q} \tag{6}$$

which uses Tukey's T instead of Student's t. Tukey's T is equal to $\left(1/\sqrt{2}\right)$ times the Studentized range distribution, which is tabulated in books such as Neter et al. (1996). Tukey called his procedure the *Honestly Significant Difference* procedure since it controls the experiment-wide error rate at α. If all of the means are equal, the probability of declaring *any* of the pairs to be significantly different in the entire experiment equals α. Tukey's procedure is more conservative than Fisher's LSD procedure, since it makes it harder to declare any particular pair of means to be significantly different.

- **Scheffe** - designed to permit the estimation of all possible contrasts amongst the sample means (not just pairwise comparisons). It uses a multiple related to the F distribution:

$$M = \sqrt{(q-1)F_{\alpha,q-1,n-q}} \tag{7}$$

In the current instance, this procedure is likely to be very conservative, since only pairs are being estimated.

- **Bonferroni** - designed to permit the estimation of any preselected number of contrasts. In this case, it uses a multiple equal to

$$M = t_{\alpha/(q(q-1)),n-q} \tag{8}$$

since *q(q-1)/2* pairwise differences are being estimated. These limits are usually wider than Tukey's limits when all pairwise comparisons are being made.

- **Student-Newman-Keuls** - Unlike the previous methods, this method does not create intervals for the pairwise differences. Instead, it sorts the means in increasing order and then begins to separate them into groups according to values of the Studentized range distribution. Eventually, the means are separated into homogeneous groups within which there are no significant differences.

- **Duncan** - similar to the Student-Newman-Keuls procedure, except that it uses a different critical value of the Studentized range distribution when defining the homogeneous groups. A detailed discussion of the Duncan and Student-Newman-Keuls procedures is given by Milliken and Johnson (1992).
- **Games-Howell** - similar to the Tukey procedure, except that it does not assume equal variances within each group. This is a good choice if the *Variance Check* indicates significant differences amongst the group variances. The limits for the Games-Howell procedure are given by

$$\hat{\Delta}_{j_1 j_2} \pm T_{\alpha/2,q,v} \sqrt{\left(\frac{s_{j_1}^2}{n_{j_1}} + \frac{s_{j_2}^2}{n_{j_2}} \right)} \tag{9}$$

where

$$v = \frac{\left(\frac{s_{j_1}^2}{n_{j_1}} + \frac{s_{j_2}^2}{n_{j_2}} \right)^2}{\frac{\left(\frac{s_{j_1}^2}{n_{j_1}} \right)^2}{n_{j_1} - 1} + \frac{\left(\frac{s_{j_2}^2}{n_{j_2}} \right)^2}{n_{j_2} - 1}} \tag{10}$$

The choice between the LSD procedure and a multiple comparisons procedure such as Tukey's HSD should depend on the relative cost of making a Type I error (calling a pair of means different when they're really not) versus the cost of making a Type II error (not calling a pair of means different when they really are). In early stages of an investigation, one may not want to be as conservative as when final verifications are being made.

Table of Means

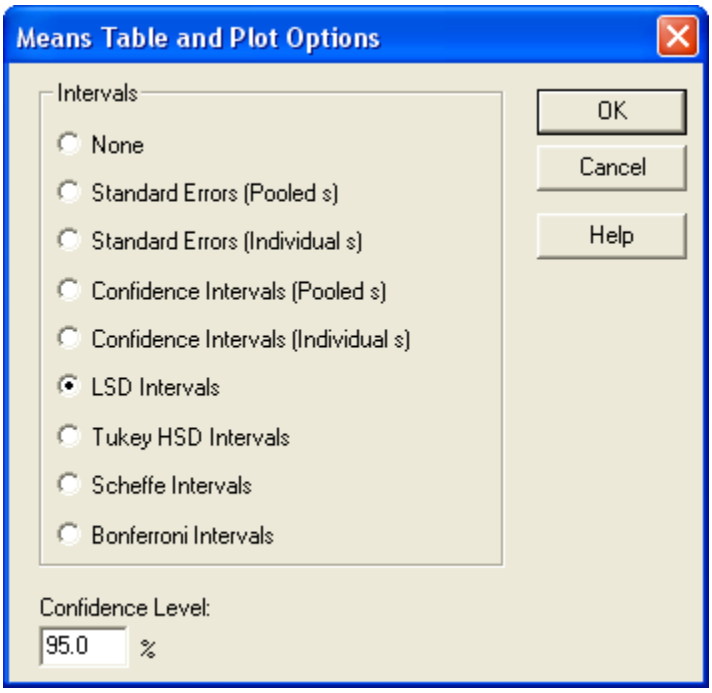
This table displays each sample mean together with an uncertainty interval:

Table of Means with 95.0 percent LSD intervals

			Stnd. error		
	Count	*Mean*	*(pooled s)*	*Lower limit*	*Upper limit*
Task 1	13	31.9231	1.54184	29.7437	34.1024
Task 2	12	31.0833	1.6048	28.815	33.3517
Task 3	10	35.8	1.75797	33.3151	38.2849
Task 4	10	38.0	1.75797	35.5151	40.4849
Task 5	12	29.5	1.6048	27.2316	31.7684
Task 6	11	28.8182	1.67616	26.449	31.1874
Total	68	32.3088			

The type of interval displayed depends on *Pane Options*.

Pane Options



- **Intervals**: the method used to construct the intervals.
- **Confidence Level:** the level of confidence associated with each interval.

The type of intervals that may be selected are:

- **None** - no intervals are displayed.
- **Standard errors (pooled s)** - displays the standard errors using the pooled within-sample standard deviation:

$$\bar{Y}_j \pm \sqrt{\frac{MS_{within}}{n_j}} \tag{11}$$

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

$$\bar{Y}_j \pm \sqrt{\frac{s_j^{\,2}}{n_j}} \tag{12}$$

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

$$\bar{Y}_j \pm t_{\alpha/2,n-q}\sqrt{\frac{MS_{within}}{n_j}} \tag{13}$$

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

$$\bar{Y}_j \pm t_{\alpha/2,n_j-1}\sqrt{\frac{s_j^{\,2}}{n_j}} \tag{14}$$

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

$$\bar{Y}_j \pm \frac{\sqrt{2}M}{2}\sqrt{\frac{MS_{within}}{n_j}} \tag{15}$$

where M is defined as in the *Multiple Range Tests*. This formula also applies to the three selections below.

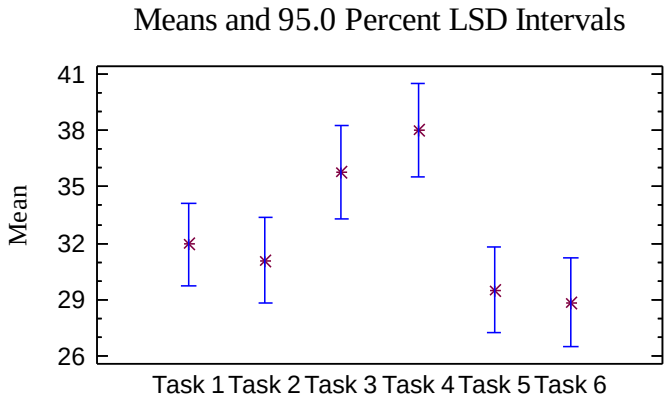
- **Tukey HSD Intervals** - designed for comparing all pairs of means. The stated confidence level applies to the entire family of pairwise comparisons.

- **Scheffe Intervals** - designed for comparing all contrasts. Not usually relevant here.

- **Bonferroni Intervals** - designed for comparing a selected number of contrasts. Tukey's intervals are usually tighter.

Means Plot

The sample means may be plotted together with the uncertainty intervals:



The types of intervals that may be used are the same as for the *Means Table* above.

Provided all of the sample sizes are the same (or close), the analyst can determine which means are significantly different from which others using the LSD, Tukey, Scheffe, or Bonferroni procedure simply by looking at whether or not a pair of intervals overlap in the vertical direction. A pair of intervals that do not overlap indicates a statistically significant difference between the means at the selected confidence level. In this case, note that the interval for Task 4 overlaps only the interval for Task 3, indicating that it is significantly different from all of the other tasks.

Variance Check

One of the assumptions underlying the analysis of variance is that the variances of the populations from which the samples come are the same. The *Variance Check* pane performs any of several tests to verify this assumption:

Variance Check

	Test	*P-Value*
Levene's	0.687841	0.634445

Comparison	*Sigma1*	*Sigma2*	*F-Ratio*	*P-Value*
Task 1 / Task 2	4.95751	5.66422	0.766034	0.6522
Task 1 / Task 3	4.95751	5.30827	0.872209	0.8067
Task 1 / Task 4	4.95751	6.59966	0.564266	0.3513
Task 1 / Task 5	4.95751	6.00757	0.680973	0.5187
Task 1 / Task 6	4.95751	4.75012	1.08923	0.9051
Task 2 / Task 3	5.66422	5.30827	1.1386	0.8594
Task 2 / Task 4	5.66422	6.59966	0.736607	0.6227
Task 2 / Task 5	5.66422	6.00757	0.888959	0.8487
Task 2 / Task 6	5.66422	4.75012	1.4219	0.5868
Task 3 / Task 4	5.30827	6.59966	0.646939	0.5267
Task 3 / Task 5	5.30827	6.00757	0.780744	0.7216
Task 3 / Task 6	5.30827	4.75012	1.24881	0.7301
Task 4 / Task 5	6.59966	6.00757	1.20683	0.7561
Task 4 / Task 6	6.59966	4.75012	1.93034	0.3200
Task 5 / Task 6	6.00757	4.75012	1.59952	0.4676

The hypotheses to be tested are:

Null Hypothesis: all σ_j are equal

Alt. Hypothesis: not all σ_j are equal

The four tests are:

1. *Cochran's test:* compares the maximum within-sample variance to the average within-sample variance. A P-value less than 0.05 indicates a significant difference amongst the within-sample standard deviations at the 5% significance level. The test is appropriate only if all group sizes are equal.

2. *Bartlett's test:* compares a weighted average of the within-sample variances to their geometric mean. A P-value less than 0.05 indicates a significant difference amongst the within-sample standard deviations at the 5% significance level. The test is appropriate for both equal and unequal group sizes.

3. *Hartley's test:* computes the ratio of the largest sample variance to the smallest sample variance. This statistic must be compared to a table of critical values, such as the one contained in Neter et al. (1996). For 6 samples and 62 degrees of freedom for experimental error, *H* would have to exceed approximately 2.1 to be statistically significant at the 5% significance level. Note: this test is only appropriate if the number of observations within each treatment level is the same.

4. *Levene's test:* performs a one-way analysis of variance on the variables

$$Z_{ij} = \left| y_{ij} - \bar{y}_j \right| \tag{16}$$

The tabulated statistic is the F statistic from the ANOVA table.

For the pulse rate data, there is no reason to reject the assumption that the standard deviations are the same for all groups, since the P-values is greater than 0.05. Any apparent differences among the sample standard deviations are not statistically significant at the 5% significance level.

The table also displays the results of a set of two-sample F-tests that compare the standard deviations for each pair of samples. Any pair with a small P-Value would be a pair whose standard deviations were significantly different. In the current example, there are no significant pairs. NOTE: It is recommended that the F-tests be ignored if the initial overall test does not show significant differences amongst the sigmas.

Residual Plots

As with all statistical models, it is good practice to examine the residuals. In a oneway analysis of variance, the residuals are defined by:

$$e_{ij} = y_{ij} - \bar{y}_j \tag{17}$$

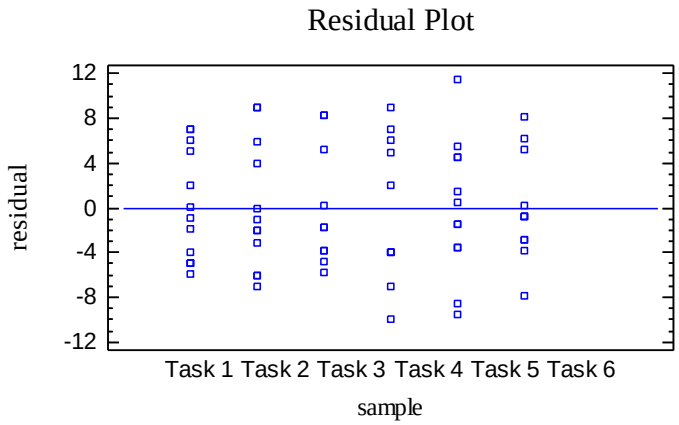
i.e., the residuals are the differences between the observed data values and their respective group means.

The *Multiple Sample Comparison* procedure creates 3 residual plots:

1. versus sample indicator.
2. versus predicted value.
3. versus observation number.

Residuals versus Samples

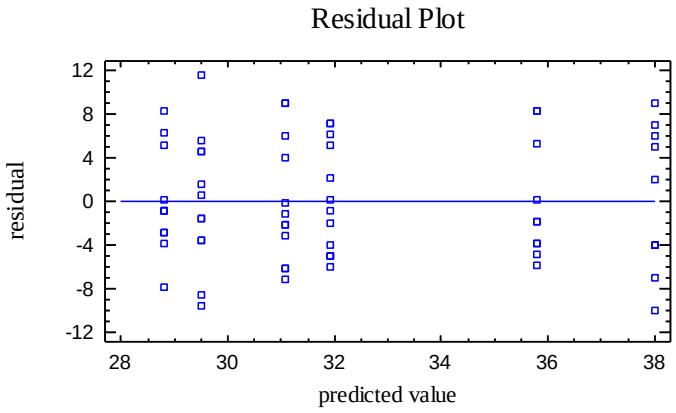
This plot is helpful in visualizing any differences in variability amongst the samples.



The average residual in each group equals 0.

Residuals versus Predicted

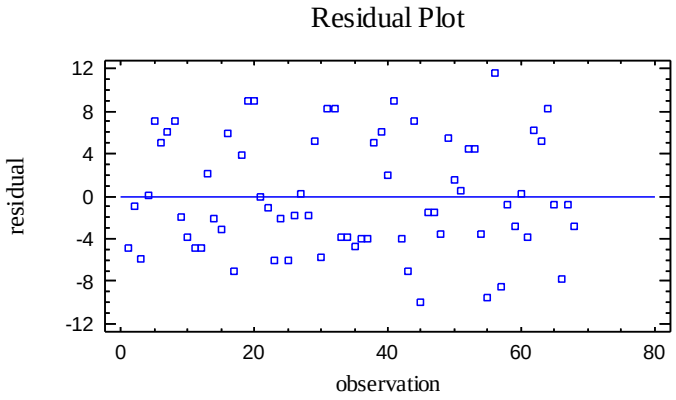
This plot is helpful in detecting any heteroscedasticity in the data.



Heteroscedasticity occurs when the variability of the data changes as the mean changes, and might necessitate transforming the data before performing the ANOVA. It is usually evidenced by a funnel-shaped pattern in the residual plot.

Residuals versus Observation

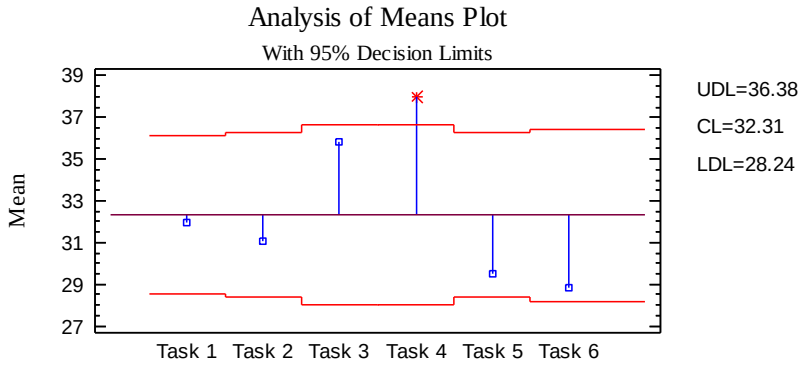
This plot shows the residuals versus row number in the datasheet:



If the data are arranged in chronological order, any pattern in the data might indicate an outside influence. No such pattern is evident in the above plot.

Analysis of Means (ANOM) Plot

If the number of samples is between 3 and 20, a somewhat different approach to the comparison of sample means is presented in the *Analysis of Means* or *ANOM Plot*:



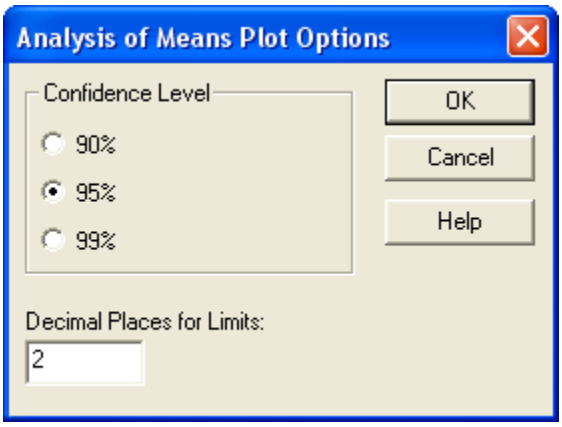
This plot constructs a chart similar to a standard control chart, where each sample mean is plotted together with a centerline and upper and lower decision limits. The centerline is located at the grand average of all of the observations $\bar{Y}$. The decision limits are located at

$$\bar{Y} \pm h_{n-q,1-\alpha}\sqrt{\frac{MS_{within}}{n_j}\left(\frac{q-1}{q}\right)} \qquad (18)$$

where *h* is a critical value obtained from a table of the multivariate t distribution. The chart tests the null hypothesis that all of the sample means are equal to the grand mean. Any means that fall outside the decision limits indicate that the corresponding sample differs significantly from that overall mean.

The advantage of the ANOM plot is that it shows at a glance which means are significantly different than the average of all the samples. It also does so using a type of chart with which many engineers and operators are quite familiar. It is easy to see from the above chart that Task 4 has a significantly higher pulse rate than average, while all of the other task means are within the decision limits. The procedure is exact if all sample sizes are equal and approximate if they don't differ too much.

Pane Options



- **Confidence Level:** level used to position the decision limits.
- **Decimal Places for Limits:** number of decimal places shown when displaying the decision limits.

Kruskal-Wallis and Friedman Tests

An alternative to the standard analysis of variance that compares group *medians* instead of means is the *Kruskal-Wallis Test*. This test is much less sensitive to the presence of outliers than a standard oneway ANOVA and should be used whenever the assumption of normality within samples is not reasonable. It tests the hypotheses:

Null Hypothesis: all group medians are equal

Alt. Hypothesis: not all group medians are equal

The test is conducted by:

1. Sorting all of the *n* data values from smallest to largest and ranking them, assigning a rank of 1 to the smallest and *n* to the largest. If any observations are exactly equal, then the tied observations are given a rank equal to the average of the positions at which the tie occurs.

2. Computing the average ranks of the observations within each group $\overline{R}_j$.

3. Calculating a test statistic to compare the differences amongst the average ranks.

4. Calculating a P-value to test the hypotheses.

The top section of the output is shown below:

Kruskal-Wallis Test

	Sample Size	*Average Rank*
Task 1	13	33.3846

Task 2	12	30.5833
Task 3	10	46.4
Task 4	10	50.35
Task 5	12	26.7083
Task 6	11	23.3636

Test statistic = 15.9995 P-Value = 0.00684551

Small P-Values (less than 0.05 if operating at the 5% significance level) indicate that there are significant differences amongst the group medians, as in the example above.

The bottom section of the output shows a comparison of the average rank for each pair of groups:

95.0 percent Bonferroni intervals

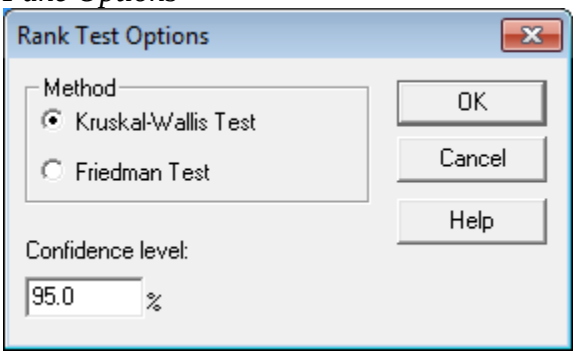
Contrast	*Sig.*	*Difference*	*+/- Limits*
Task 1 - Task 2		2.80128	23.2346
Task 1 - Task 3		-13.0154	24.4129
Task 1 - Task 4		-16.9654	24.4129
Task 1 - Task 5		6.67628	23.2346
Task 1 - Task 6		10.021	23.7774
Task 2 - Task 3		-15.8167	24.8512
Task 2 - Task 4		-19.7667	24.8512
Task 2 - Task 5		3.875	23.6947
Task 2 - Task 6		7.2197	24.2272
Task 3 - Task 4		-3.95	25.9562
Task 3 - Task 5		19.6917	24.8512
Task 3 - Task 6		23.0364	25.3595
Task 4 - Task 5		23.6417	24.8512
Task 4 - Task 6	*	26.9864	25.3595
Task 5 - Task 6		3.3447	24.2272

* denotes a statistically significant difference.

The table shows the difference between the two average ranks, together with the width of a Bonferroni interval for the difference. Any pair of groups for which the difference exceeds the "+/- Limits" value is statistically significant at the stated confidence level.

In the sample data, the only difference that is statistically significant is the difference between Task 4 and Task 6. It should be noted, however, that the Bonferroni procedure can be quite conservative, meaning that some real differences may not show up as statistically significant.

Pane Options



- **Method**: procedure to use to compare the medians. *Kruskal-Wallis* is appropriate when comparing *q* independent samples. *Friedman* is appropriate when analyzing a blocked experiment, i.e., when the data is each row correspond to the same experimental unit or block.

- **Confidence Level:** level used to construct the Bonferroni intervals.

The Friedman test is appropriate for a *randomized block design*, in which each row of the datasheet represents a particular condition or "block". In the current example, this would apply if the same 13 subjects performed each of the 6 tasks, rather than different subjects for each task.

The output from the Friedman test is interpreted in the same manner as the Kruskal-Wallis test output.

Mood's Median Test

Mood's Median Test is another method of determining whether or not the medians of all *q* groups are equal. It is less sensitive to outliers than the Kruskal-Wallace test, but is also less powerful when the data come from distributions such as the normal. The output is shown below.

Mood's Median Test
Total n = 68
Grand median = 31.0

Sample	*Sample Size*	*n<=*	*n>*	*Median*	*95% lower CL*	*95% upper CL*
Task 1	13	7	6	31.0	27.0	38.6056
Task 2	12	8	4	29.5	25.0	39.6809
Task 3	10	2	8	34.0	30.3244	44.0
Task 4	10	2	8	37.0	28.9733	46.3511
Task 5	12	8	4	29.0	21.5318	34.8936
Task 6	11	8	3	28.0	23.8509	35.5745

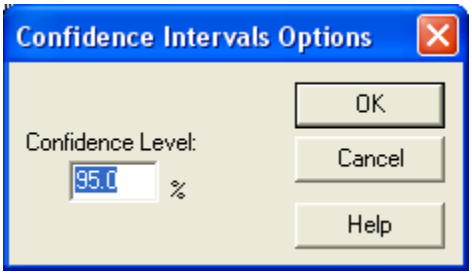
Test statistic = 12.168 P-Value = 0.0325567

Displayed at the top of the table are the total number of observations *n* and the overall median. For each sample, the table shows

1. *Sample Size*: the number of observations in the sample n_j.
2. *n*<=: of the observations in the sample, how many are less than or equal to the overall median.
3. *n*>: of the observations in the sample, how many are greater than the overall median.
4. *Median*: the sample median.
5. *CL*: the lower and upper confidence limits for the median of the population from which the sample came.

Displayed at the bottom of the screen is a test statistic and P-Value. Treating the *n*<= and the *n*> columns as columns of a two-way contingency table, a chi-squared test statistic is calculated. Small P-Values (less than 0.05 if operating at the 5% significance level) lead to the conclusion that the medians are not all equal, as in the current example.

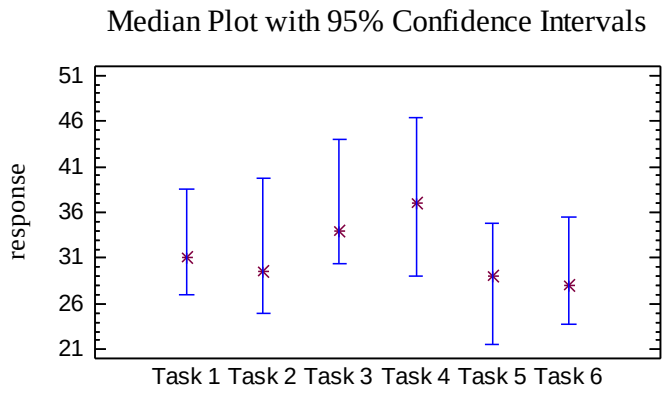
Pane Options



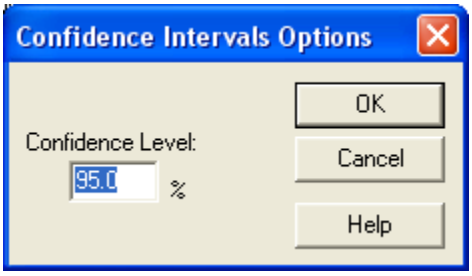
- **Confidence Level:** level used for the confidence limits.

Medians Plot

The *Medians Plot* displays the confidence intervals for the medians displayed by the *Mood's Median Test* pane.



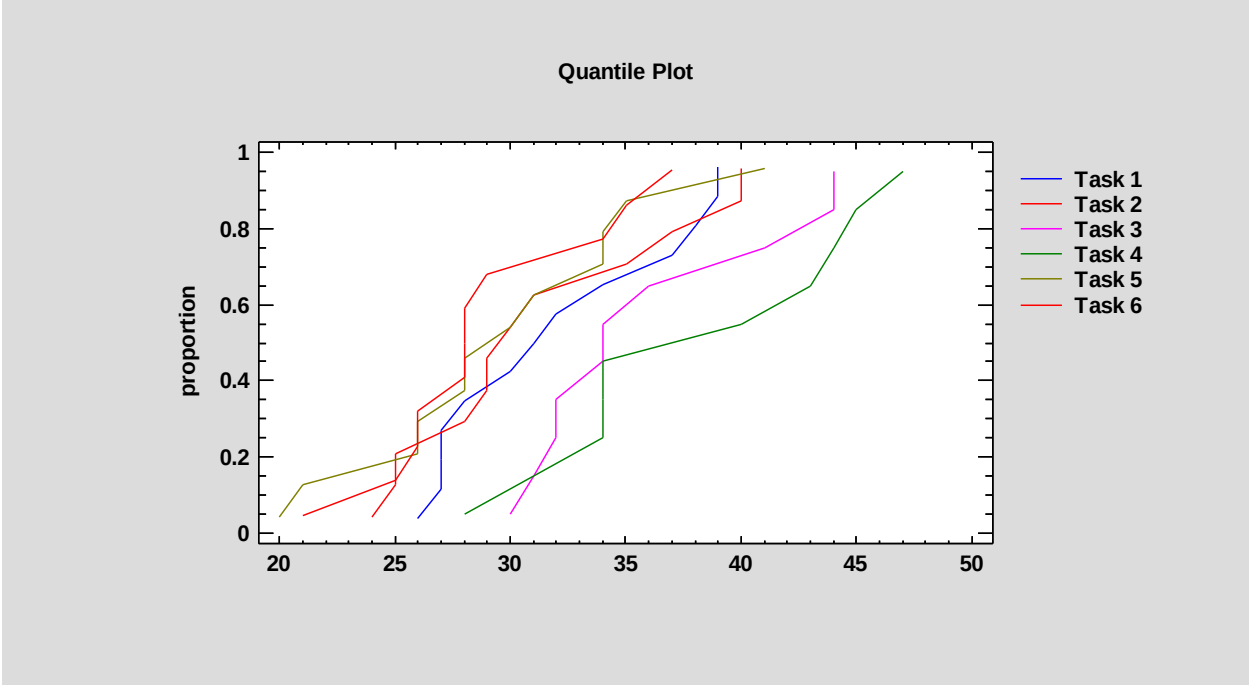
Pane Options



- **Confidence Level:** level used for the confidence limits.

Quantile Plot

The *Quantile Plot* plots the quantiles (percentiles) of the data for each group.



In this plot, the data are sorted from smallest to largest and plotted at the coordinates

$$\left(x_{(j)}, \frac{j-0.5}{n} \right) \tag{19}$$

Save Results

The following results can be saved to the datasheet:

1. *Counts* – the *q* sample sizes n_j.
2. *Means* – the *q* sample means.
3. *Medians* – the *q* sample medians.
4. *Standard Deviations* – the *q* sample standard deviations.
5. *Standard Errors* – the standard errors of each sample mean, $\sqrt{MS_{within} / n_j}$.
6. *Labels* – a label for each sample.
7. *Residuals* – the *n* residuals.
8. *Ranges* – the *q* sample ranges.
9. *Data Column* – the *n* observations arranged into a single column.
10. *Code Column* – *n* codes identifying the sample corresponding to each observation in the *Data Column*.

Calculations

Analysis of Variance

Source	Sum of Squares	D.F.	Mean Square	F-Ratio
Between groups	$SS_{between} = \sum_{j=1}^{q} n_j \left(\overline{Y}_j - \overline{\overline{Y}}\right)^2$	$df_{between} = q-1$	$MS_{between} = \frac{SS_{between}}{df_{between}}$	$F = \frac{MS_{between}}{MS_{within}}$
Within groups	$SS_{within} = \sum_{j=1}^{q} \sum_{i=1}^{n_j} \left(Y_{ij} - \overline{Y}_j\right)^2$	$df_{within} = \sum_{j=1}^{q} \left(n_j - 1\right)$	$MS_{within} = \frac{SS_{within}}{df_{within}}$	
Total	$SS_{total} = \sum_{j=1}^{q} \sum_{i=1}^{n_j} \left(Y_{ij} - \overline{\overline{Y}}\right)^2$	n-1		

Cochran's Test

The statistic displayed is calculated by

$$A = \frac{\max\left(s_j^2\right)}{\sum_{j=1}^{q} s_j^2} \tag{20}$$

To test for statistical significance,

$$C = \left(q-1\right)\left(\frac{A}{1-A}\right) \tag{21}$$

is compared to an F distribution with *(n/q - 1)* and *(n/q - 1)(q - 1)* degrees of freedom.

Bartlett's Test

The statistic displayed is calculated by

$$B = \frac{1}{C}\left[\left(dfe\right)\ln\left(MSE\right) - \sum_{j=1}^{q}\left(n_j - 1\right)\ln\left(s_j^2\right)\right] \tag{22}$$

where

$$C = 1 + \frac{1}{3(q-1)}\left[\left(\sum_{j=1}^{q}\left(n_j - 1\right)^{-1}\right) - \frac{1}{dfe}\right] \quad (23)$$

$$MSE = \frac{1}{dfe}\sum_{j=1}^{q}\left(n_j - 1\right)s_j^2 \quad (24)$$

$$dfe = \sum_{j=1}^{q}\left(n_j - 1\right) \quad (25)$$

B is compared to the chi-squared distribution with *(q-1)* degrees of freedom.

Hartley's Test

$$H = \frac{\max\left(s_j^2\right)}{\min\left(s_j^2\right)} \quad (26)$$

Median Confidence Limits

The limits displayed are a nonlinear interpolation of the confidence intervals at the nearest confidence levels above and below the level requested. After sorting the observations, the interval extending from the *d*-th smallest observation in the sample to the *d*-th largest observation forms a confidence interval for the median with confidence level $1 - 2\ P_B(d\text{-}1)$, where P_B represents the cumulative binomial distribution with $p = 0.5$ and $n = n_j$.

Welch Test

The statistic displayed is calculated by

$$F = \frac{\sum_{j=1}^{q} W_j (\bar{Y}_j - \bar{Y}^*)^2 / (q-1)}{1 + 2(q-2)\Lambda / \left(q^2 - 1\right)} \quad (27)$$

where

$$W_j = \frac{n_j}{s_j^2} \quad (28)$$

$$\bar{Y}^* = \frac{\sum_{j=1}^{q} W_j \bar{Y}_j}{\sum_{j=1}^{q} W_j} \tag{29}$$

$$\Lambda = \sum_{j=1}^{q} \frac{\left(1 - W_j / W_.\right)^2}{n_j - 1} \tag{30}$$

$$W_. = \sum_{j=1}^{q} W_j \tag{31}$$

The test statistic has an approximate F distribution with (q-1) and (q^2-1)/3Λ degrees of freedom.

Brown-Forsythe Test

The statistic displayed is calculated by

$$F = \frac{\sum_{j=1}^{q} n_j (\bar{Y}_j - \bar{Y})^2}{\sum_{j=1}^{q} s_j^2 \left(1 - n_j / n\right)} \tag{32}$$

The test statistic has an approximate F distribution with v_1= (q-1) and v_2 degrees of freedom where

$$v_2 = \frac{1}{\sum_{j=1}^{q} c_j^2 / \left(n_j - 1\right)} \tag{33}$$

$$c_j = \frac{s_j^2 \left(1 - n_j / n\right)}{\sum_{i=1}^{q} s_i^2 \left(1 - n_i / n\right)} \tag{34}$$

Bonferroni Intervals for Kruskal-Wallis and Friedman Tests

The width of the intervals is calculated from:

Kruskal-Wallis Test: $$\pm B \left[\frac{n(n+1)}{12} \left(\frac{1}{n_i} + \frac{1}{n_j} \right) \right]^{1/2} \tag{35}$$

Friedman Test: $$\pm B\left[\frac{q(q+1)}{6n}\right]^{1/2} \quad (36)$$

where

$$B = z(1-\alpha/2g) \quad (37)$$

$$g = \frac{q(q-1)}{2} \quad (38)$$