Comparison between more than two variances

When more than two variances have to be compared to test the following null hypothesis:

$$H_0 : \sigma_1^2 = \sigma_2^2 = \ldots = \sigma_k^2$$

i.e., to test homoscedasticity of responses obtained at different concentrations before proceeding with linear regression, the systematic use of F-test for each possible couple of variances would not provide reliable results.

In fact, although multiple pairwise comparison between variances could be made using the F-test (provided that data in each group are normally distributed), the probability of an erroneous result (declaring that two variances are significantly different when they are not) would be not negligible.

Alternative tests, some of which applicable even to data not distributed normally, are then available. The following tests:

1) Hartley's test (also called Hartley's F_{max} test)
2) Bartlett's test
3) Levene's test

are among those used more commonly to compare several variances.

Hartley's (F_{max}) test

The test was developed in 1950 by German-American statistician Herman Otto Hartley (original surname Hirschfeld) and is based on the assumptions of normality and of equal numbers of data for groups whose variances are under comparison.

It implies the calculation of the ratio, named F_{max}, between the largest and the smallest variance among those available:

$$F_{\max} = \frac{\sigma^2_{\max}}{\sigma^2_{\min}}$$

this value is then compared to a critical value taken from a table of the F_{max} distribution, which reports critical values according to the degrees of freedom (DF), corresponding to the number of data in each group subtracted of one, and to the number of groups (treatments) under comparison:

The upper and lower values in each box of the table are referred to significant levels α = 0.05 and 0.01, respectively.

As expected, the lower value is always greater than the other one (i.e., it is more difficult to reject the homogeneity of variance for a lower significance level).

By analogy with the F-test for two variances, critical values are decreased at the increase of the degrees of freedom.
On the other hand, they are increased at the increase of the number of groups under comparison.

DF (n-1)	Number of treatments (k)										
	2	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
2	39.0 199	87.5 448	142 729	202 1036	266 1362	333 1705	403 2063	475 2432	550 2813	626 3204	714 3605
3	15.4 47.5	27.8 85.0	39.2 120	50.7 151	62.0 184	72.9 21	83.5 24	93.9 28	104 31	114 33	124 36
4	9.6 23.2	15.5 37.0	20.6 49.0	25.2 59	29.5 69	33.6 79	37.5 89	41.1 97	44.6 106	48.0 113	51.4 120
5	7.2 14.9	10.8 22.0	13.7 28.0	16.3 33	18.7 38	20.8 42	22.9 46	24.7 50	26.5 54	28.2 57	29.9 60
6	5.82 11.1	8.38 15.5	10.4 19.1	12.1 22	13.7 25	15.0 27	16.3 30	17.5 32	18.6 34	19.7 36	20.7 37
7	0.99 8.89	6.94 12.1	8.44 14.5	9.70 16.5	10.8 18.4	11.8 20	12.7 22	13.5 23	14.3 24	15.1 26	15.8 27
8	4.43 7.50	6.00 9.90	7.18 11.7	8.12 13.2	9.03 14.5	9.78 15.8	10.5 16.9	11.1 17.9	11.7 18.9	12.2 19.8	12.7 21
9	4.03 6.54	5.34 8.50	6.31 9.9	7.11 11.1	7.80 12.1	8.41 13.1	8.95 13.9	9.45 14.7	9.91 15.3	10.3 16.0	10.7 16.6
10	3.72 5.85	4.85 7.40	5.67 8.6	6.34 9.6	6.92 10.4	7.42 11.1	7.87 11.8	8.28 12.4	8.66 12.9	9.01 13.4	9.34 13.9
12	3.28 4.91	4.16 6.1	4.75 6.9	5.30 7.6	5.72 8.2	6.09 8.7	6.42 9.1	6.72 9.5	7.00 9.9	7.25 10.2	7.43 10.6
15	2.86 4.07	3.54 4.9	4.01 5.5	4.37 6.0	4.68 6.4	4.95 6.7	5.19 7.1	5.40 7.3	5.59 7.5	5.77 7.8	5.95 8.0
20	2.46 3.32	2.95 3.8	3.29 4.3	3.54 4.6	3.76 4.9	3.94 5.1	4.10 5.3	4.24 5.5	4.37 5.6	4.49 5.8	4.59 5.9
30	2.07 2.63	2.40 3.0	2.61 3.3	2.78 3.4	2.91 3.6	3.02 3.7	3.12 3.8	3.21 3.9	3.29 4.0	3.36 4.1	3.39 4.2
60	1.67 1.96	1.85 2.2	1.96 2.3	2.04 2.4	2.11 2.4	2.17 2.5	2.22 2.5	2.26 2.6	2.30 2.6	2.33 2.7	2.36 2.7

Bartlett's test

The Bartlett's test, developed in 1937 by the English statistician Maurice Stevenson Bartlett, is also applied to normally distributed data but the number of data in each group under comparison (n_i, with i going from 1 to k) does not have to be necessarily the same, like in Hartley's test.

Different test procedures referred to Bartlett are available. One of the most common is described in the following.

Given k samples with sizes n_i and sample variances S_i^2, the following statistic is calculated:

$$T = \frac{(N-k)\ln(S_p^2) - \sum_{i=1}^{k}(n_i-1)\ln(S_i^2)}{1 + \frac{1}{3(k-1)}\left(\sum_{i=1}^{k}\left(\frac{1}{n_i-1}\right) - \frac{1}{N-k}\right)}$$

where:

$$N = \sum_{i=1}^{k} n_i \qquad \text{and} \qquad S_p^2 = \frac{1}{N-k}\sum_{i=1}^{k}(n_i-1)S_i^2$$

i.e., S_p^2 is the pooled estimate for the variance.

The T statistic has approximately a χ^2_{k-1} distribution, thus the homogeneity of variance is rejected, at a significance level α, if the following inequality is valid for its realization t:

$$t > \chi^2_{k-1,\,1-\alpha}$$

Bartlett's test is very sensitive to departures from normality; a different test needs thus to be used to compare variances when at least some of the groups under comparison are not normally-distributed. The Levene's test is an example.

Levene's test

The test was developed in 1960 by the American statistician and geneticist Howard Levene and is based on the following statistic:

$$W = \frac{(N-k)}{(k-1)} \cdot \frac{\sum_{i=1}^{k} N_i (Z_{i\cdot} - Z_{\cdot\cdot})^2}{\sum_{i=1}^{k} \sum_{j=1}^{N_i} (Z_{ij} - Z_{i\cdot})^2}$$

where:

k is the number of different groups of data

N_i is the number of data in the i^{th} group (each datum is counted by the j counter)

N is the total number of data

$Z_{ij} = |Y_{ij} - \bar{Y}_{i\cdot}|$ with the average being referred to the i^{th} group

$Z_{i\cdot} = \frac{1}{N_i} \sum_{j=1}^{N_i} Z_{ij}$ i.e., it is the mean of the Z_{ij} for group i

$Z_{\cdot\cdot} = \frac{1}{N} \sum_{i=1}^{k} \sum_{j=1}^{N_i} Z_{ij}$ i.e., it is the mean of all Z_{ij}

The W statistic has approximately a F distribution with k-1, N-k degrees of freedom.

The null hypothesis, stating that variances are not significantly different, is thus rejected if w, the realization of W, is higher than the value of $F_{k-1, N-k}$ at a 1-α confidence level.

Note that a test formally equivalent to the Levene's one but based on absolute values of differences between single values in a group and the corresponding median, called the Brown-Forsythe test, can be used for data following heavy-tailed distributions.

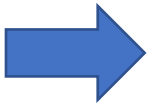
Comparison between more than two variances using Minitab 18

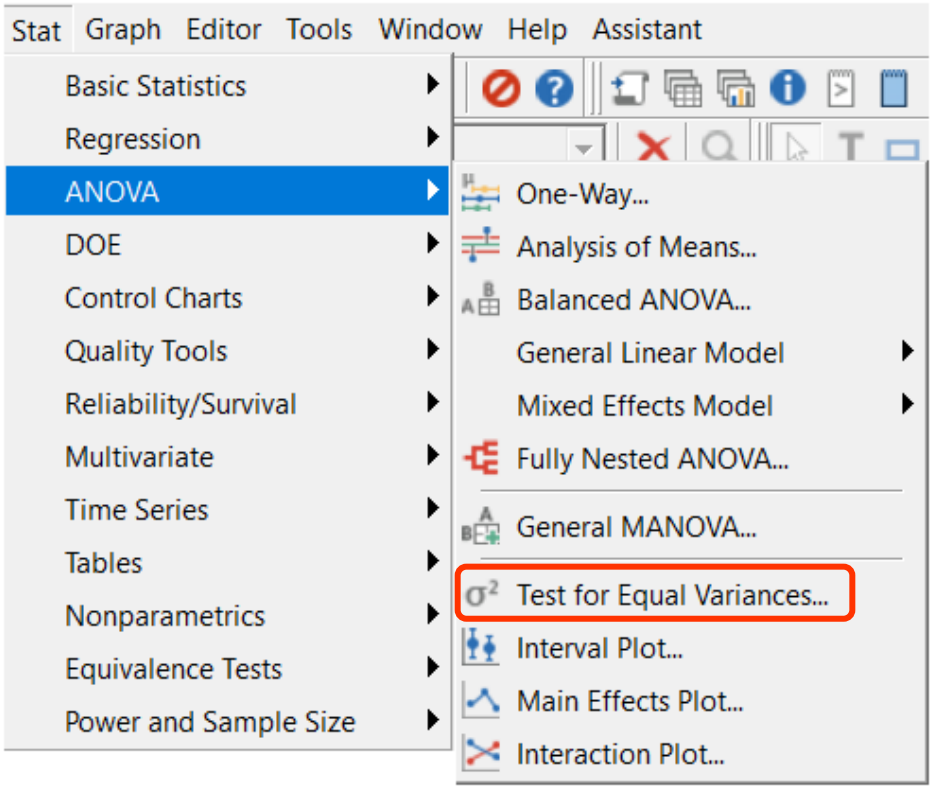
The Minitab 18 program enables a comparison between multiple variances as a part of calculations related to Analysis of Variance (ANOVA), accessible through the Stat menu.

Input data consist in replicated values obtained for different values of a variable, e.g., fluorescence intensity measurements replicated for different concentrations.

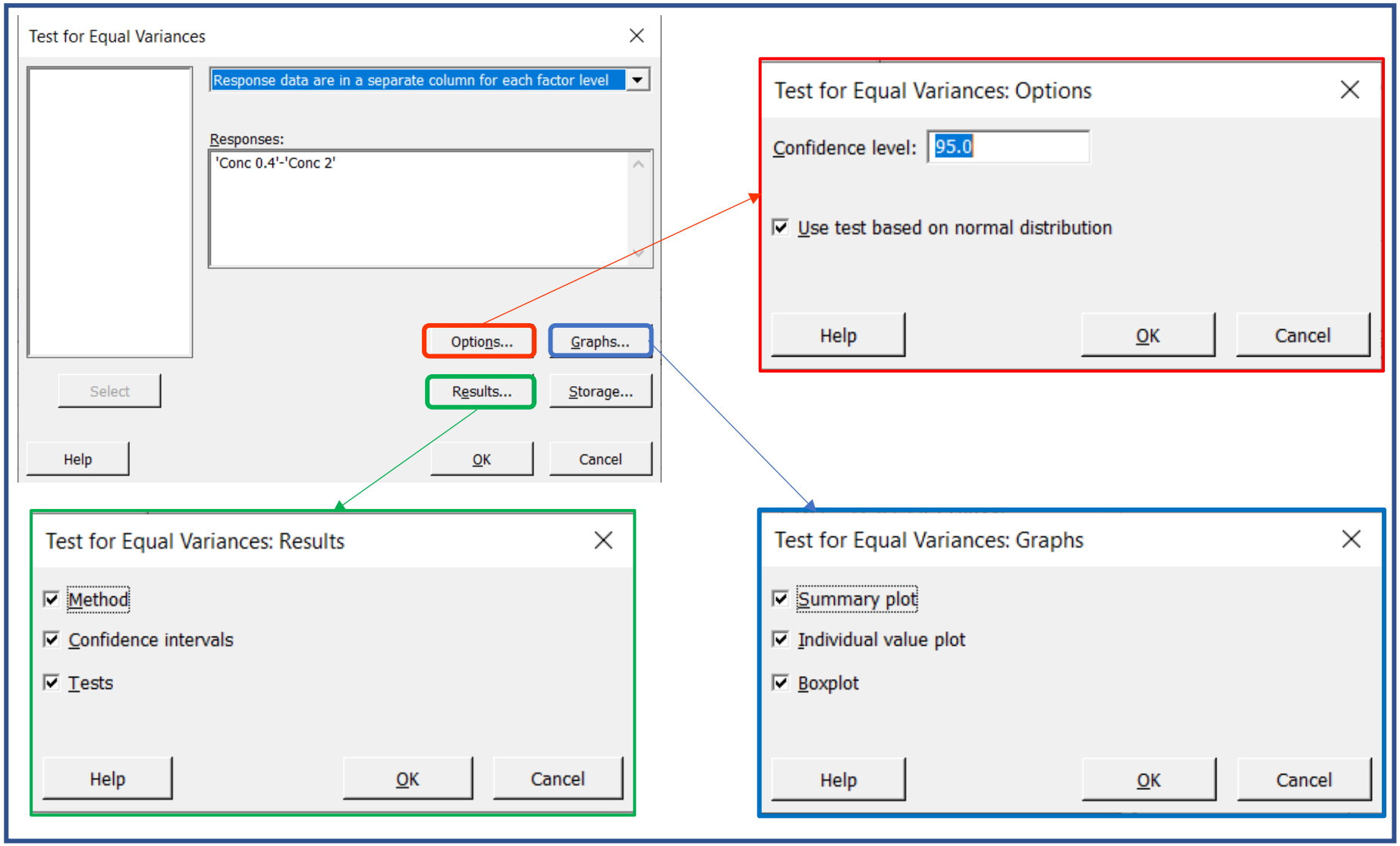
Worksheet 1 ***

↓	C1	C2	C3	C4
	Conc 0.4	Conc 0.8	Conc 1	Conc 2
1	23	39	48	120
2	18	44	56	91
3	21	41	49	102
4	20	43	52	95
5	19	42	54	131

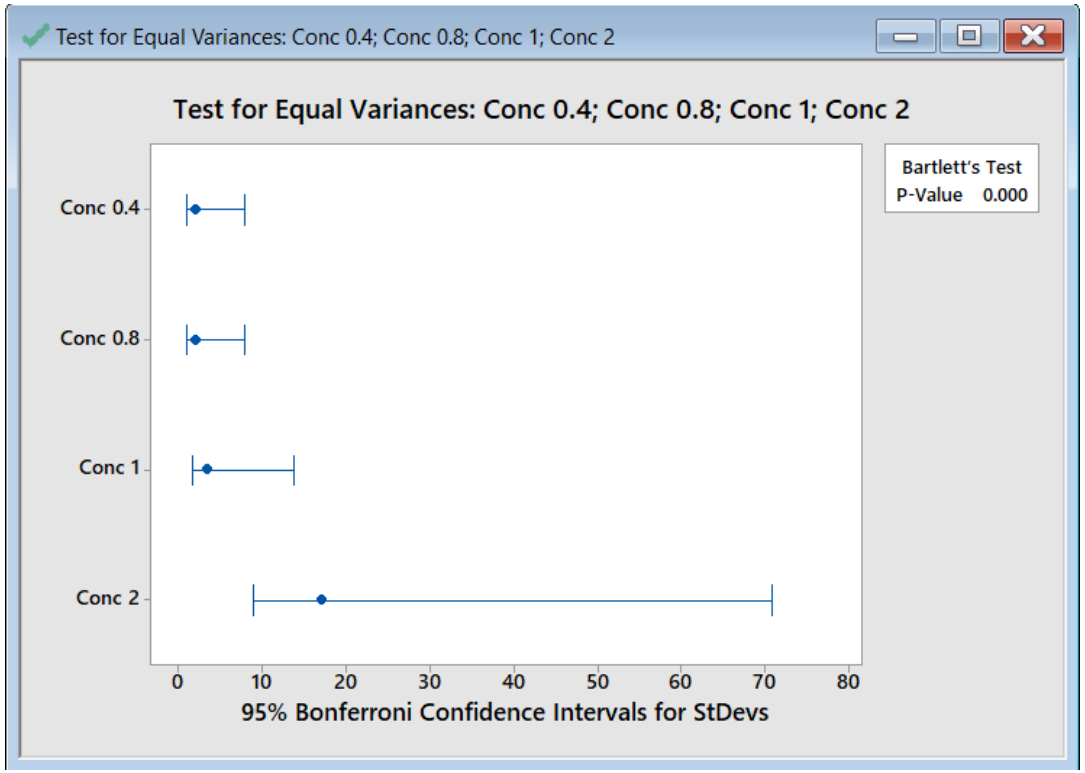




In the *Test for Equal Variances* window columns referred to Responses to be considered for the calculation of variances under comparison can be selected; then, in the Options window, the confidence level and the use of a test based on normal distribution for the comparison can be selected, if applicable. Results and graphs to be shown can be also selected:



When the normality of distributions is assumed, Minitab 18 uses Bartlett's test to make the comparison. Tables and graphs showing 95% confidence intervals for standard deviations under comparison and the realization of the Bartlett's test statistic with the corresponding P-value can be obtained as output:



Method	Test Statistic	P-Value
Bartlett	24.26	0.000

In this case the realization of Bartlett statistic is 24.26, a value much higher than the critical value, i.e., $\chi^2_{k-1,1-\alpha} = \chi^2_{3,\,0.95} = 7.81$. For this reason, P-value is much lower than 0.05 (actually, it is equal to 0, rounded to the third decimal figure), which means that there is a significant difference between some variances, obviously due to that referred to Conc 2.

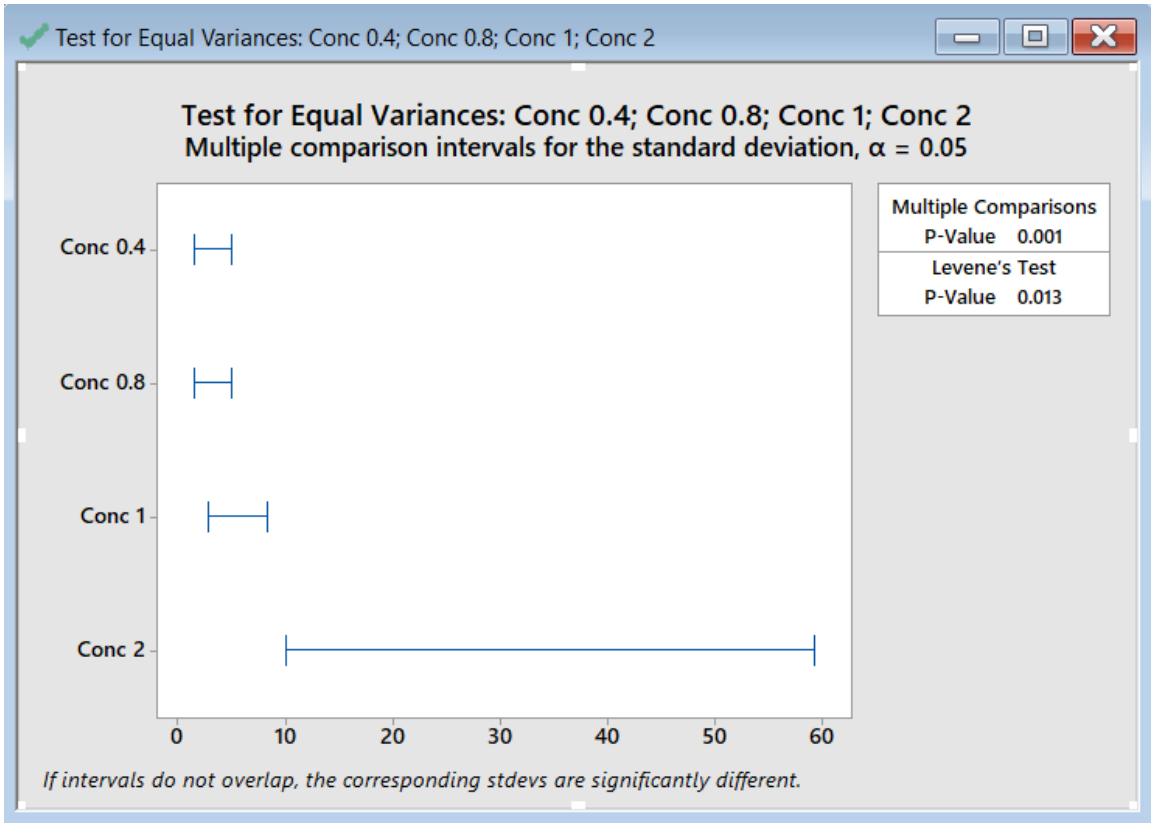
It is worth noting that confidence intervals for standard deviations are calculated by Minitab 18 according to the Bonferroni correction, named after the Italian mathematician Carlo Emilio Bonferroni, who proposed the method in 1936.

The correction proposed by Bonferroni compensates for the increase in the Type I error, i.e., the probability of incorrectly rejecting a null hypothesis (in the specific case, the absence of a significant difference between variances) when multiple comparisons are performed.

In particular, if α is needed as the overall significance level related to a multiple comparison involving m hypotheses tests, each hypothesis has to be tested at a significance level α/m, which means that single confidence intervals involved in the comparison are larger than those expected if a α significance level is adopted for each of them. Their width is clearly expected to increase with the increase of the number of comparisons to be made.

It is also important to emphasize that confidence intervals reported for standard deviations are asymmetric, since standard deviations related to normal distributions follow a χ^2 distribution.

When the normality of distributions cannot be assumed, a comparison between variances based on the Levene's Test can be made by Minitab 18 simply by not selecting the option of using a test based on normal distribution in the Options window.



Method	Test Statistic	P-Value
Multiple comparisons	—	0.001
Levene	4.95	0.013

In this case the realization of the Levene's test statistic is equal to 4.95, that is higher than the critical value $F_{k-1,\ N-k,\ (1-\alpha)} = F_{3,\ 16,\ (0.95)} = 3.239$, thus a significant difference between at least some of the four variances under comparison exist.

As shown in the previous figure, a plot of multiple comparison intervals for standard deviations at $\alpha = 0.05$ is also reported in this case. As explained in the plot caption, two standard deviations can be considered statistically different if their intervals do not overlap.

Interestingly, apart from the type of test adopted for the comparison of variances, Minitab 18 can provide a Box-and-Whisker plot for each of the four groups of data reported in the datasheet. This plot enables a visual comparison between them, eventually suggesting the presence of statistically significant differences, as in the specific case: