Multiple Comparison Methods for Means (MCMs)

When the overall null hypothesis in ANOVA:

$$H_0 : \mu_1 = \mu_2 = \ldots\ldots = \mu_i = \ldots = \mu_h$$

is rejected, at least two means could differ significantly.

Multiple Comparison Methods for Means are designed to investigate differences existing between specific pairs of means in a group.

The t-test is a typical example of a method investigating the difference existing between two means. When it is applied to several couples of means using always the same level of significance, α, the latter is called per-comparison level of significance or per-comparison (Type I) error rate (PCER).

As demonstrated before, the problem with the application of a per-comparison approach to multiple comparisons is the possible inflation of the overall probability of Type I error or (equivalently) the possible deflation of the overall confidence level (1-α).

As shown previously, if every pair of h means had to be tested, a total of C = h(h-1)/2 t-tests, each at a significance level α, would be required.

The probability of finding at least one erroneous difference would then be:

$$\alpha_C = 1 - (1 - \alpha)^C$$

As an example, for h = 3 it would result: C = 3 and $\alpha_3 = 1 - (1 - 0.05)^3 = 0.143$.
If h = 10, then C = 45 and $\alpha_{45} = 1 - (1 - 0.05)^{45} = 0.901$!

Thus, α_C approaches unity already for 10 means under comparison.

In other words, insisting on performing many pairwise comparisons, each at a per-comparison level of significance α, would almost certainly lead to conclude that two of the treatments are different even though they are not.

In statistical nomenclature a family is a collection of inferences for which it is meaningful to consider an overall measure of error.

In the specific case, the collection of all possible pairwise comparisons is a family (containing C elements) and the probability of encountering at least one Type I error, α_c, represents the overall measure of errors.

α_c is an example of family-wise error rate (FWER), generally indicated as α_{FW}.

Once a FWER is specified, the researcher must be careful to choose a multiple comparison method able to guarantee that error rate under all possible configurations of the population means.

Multiple Comparison Methods are statistical procedures usually designed to take into account and control the inflation of the overall probability of Type I error, i.e., designed to maintain a specified α_{FW} level regardless of how many pairs of means are compared.

Some of the most common MCMs for means, namely:

✓ **Fisher-Hayter** test
✓ **Tukey** test (equal group sizes)
✓ **Tukey-Kramer** test (unequal groups sizes)
✓ **Bonferroni** test
✓ **Duncan multiple range** test
✓ **Dunnett** test (for the comparison of several means with a control mean)

will be described in the following slides, after describing one of the first tests devised for the multiple comparison of means, the Fisher's Least Significant Difference (LSD) test, that is important from an historical point of view, although it does not account for the inflation of Type I error probability.

Fisher Least Significant Difference (LSD) test

The LSD test, developed by Ronald Fisher in 1935, begins by testing the overall null hypothesis (i.e., means are not statistically different) by ANOVA and, if it is rejected, moves to the next step.
On succeeding steps, the null hypothesis is tested for each couple of means, i.e., a number of *h(h-1)/2* t-tests at a *per-comparison level of significance* α is performed to see which pair of means can be considered different.

The main idea of the LSD test is to compute the least significant difference between two means and to declare significant any difference larger than the LSD.

The rationale behind the LSD technique value comes from the observation that, when the null hypothesis is true (i.e., the two means are not significantly different), the value of the t statistics evaluating the difference between, for instance, means related to Groups 1 and 2 of observations (n_1 and n_2 data, respectively) is equal to:

$$t = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{MS_{Error}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

This statistic follows a t distribution with N-h degrees of freedom, with MS_{error} being the mean square calculated during the ANOVA procedure.

The ratio *t* is thus declared significant, at a given significance level *α*, if it is larger than the critical value, denoted as $t_{(N\text{-}h),\ (1\text{-}\alpha/2)}$.

Rewriting this ratio shows that a difference between the means of Group *1* and *2* will be significant if:

$$|\bar{X}_1 - \bar{X}_2| \geq LSD = t_{N-h,\,1-\alpha/2}\sqrt{MS_{Error}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

When there is an equal number of observation per group (*n*), the above equation can be simplified as:

$$LSD = t_{N-h,\,1-\alpha/2}\sqrt{MS_{Error}\left(\frac{2}{n}\right)}$$

In order to evaluate the difference between the means of Groups *1* and *2*, the absolute value of the difference between the means is then calculated and compared to the value of LSD.

The procedure is then repeated for all the h(h-1)/2 comparisons.

Modified LSD (MLSD) or Fisher-Hayter procedure

By definition, the LSD test does not correct for multiple comparisons, thus inflating Type I error (i.e., finding a difference when it does not actually exist).

As a consequence, a revised version of the **LSD** test was proposed by Hayter (and then it is known as the Fisher-Hayter procedure), in which a modified **LSD** (**MLSD**) is computed using the *Studentized* range distribution *q*:

$$MLSD = q_{\alpha(h-1,\upsilon)}\sqrt{\frac{MS_{Error}}{2}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$$

where $q_{\alpha(h\text{-}1,\nu)}$ is the α-level critical value of the Studentized range distribution depending on indexes *h-1* and ν = *N-h*.

The expression of MLSD can be simplified if $n_1 = n_2$:

$$MLSD = q_{\alpha(h-1,\upsilon)}\sqrt{\frac{MS_{Error}}{n}}$$

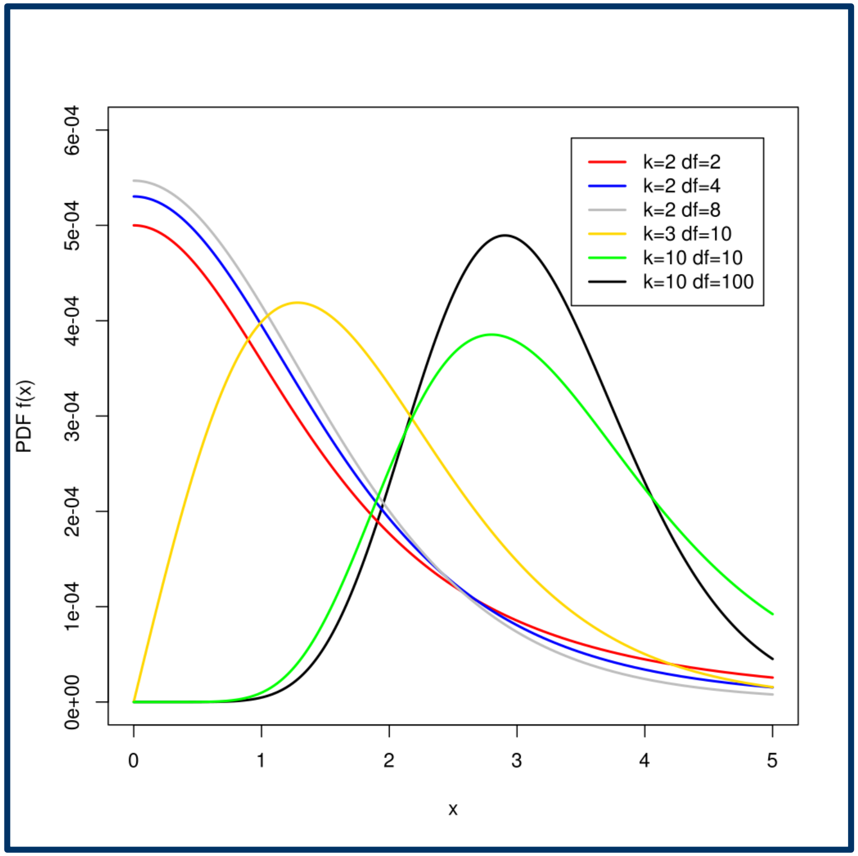
The *Studentized* range distribution is a continuous probability distribution that arises when estimating the range of data extracted from a normally distributed population in situations where the sample size is small, and population standard deviation is unknown.

Suppose that we take a sample of size n from each of k populations with the same normal distribution N(μ,σ) and that $\bar{y}_{\mathrm{min}}$ and $\bar{y}_{\mathrm{max}}$ represent the smallest and the largest of the sample means, respectively, whereas S^2 is the pooled sample variance from these samples.

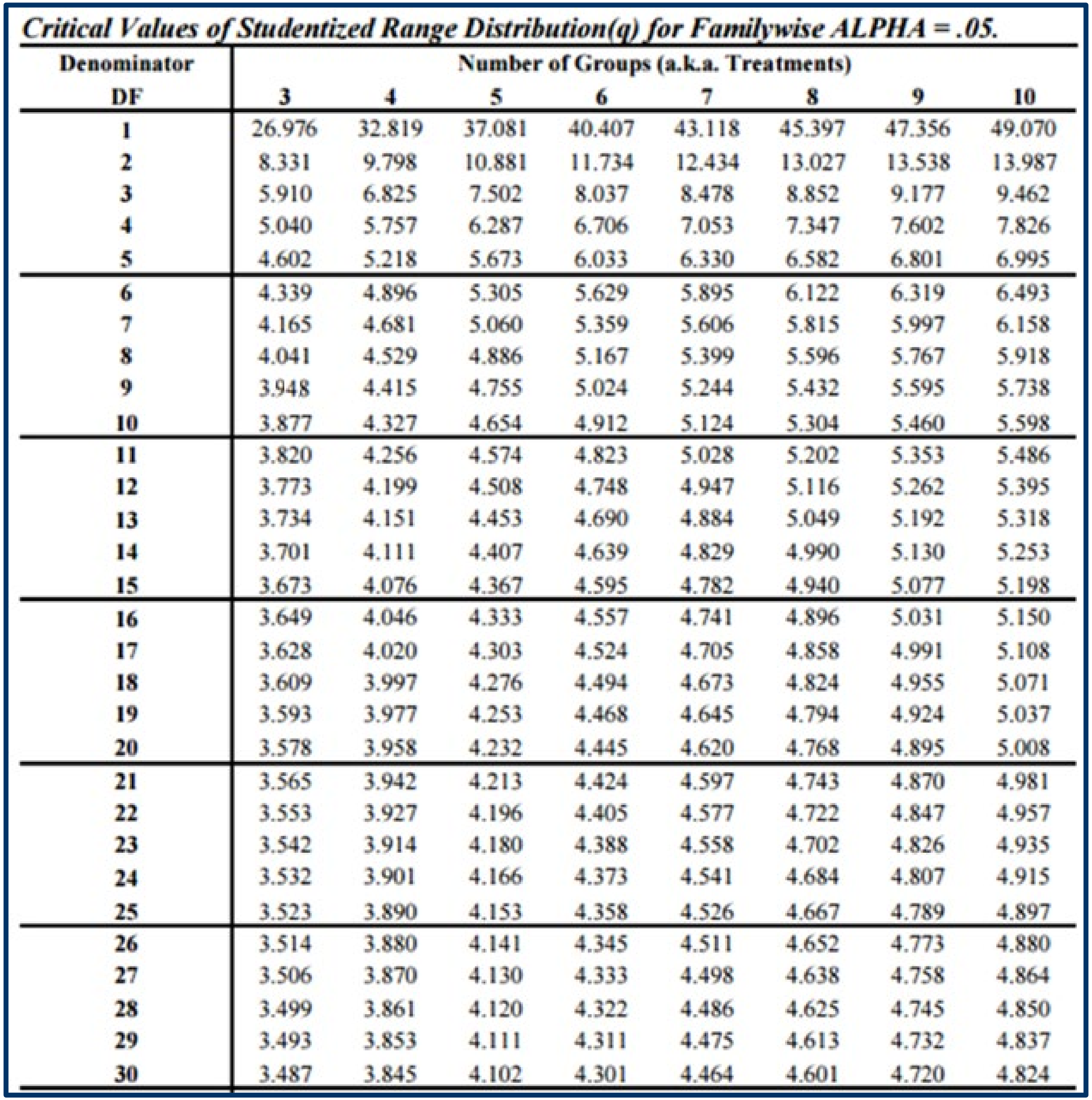
Under these conditions, the following random variable:

$$q = \frac{\bar{y}_{\max} - \bar{y}_{\min}}{S\sqrt{2/n}}$$

has a Studentized range distribution, which depends on the number k and on the degrees of freedom, i.e., the difference between the total number of data and the number of populations k:



Tables of critical values for the Studentized Range distribution are available:

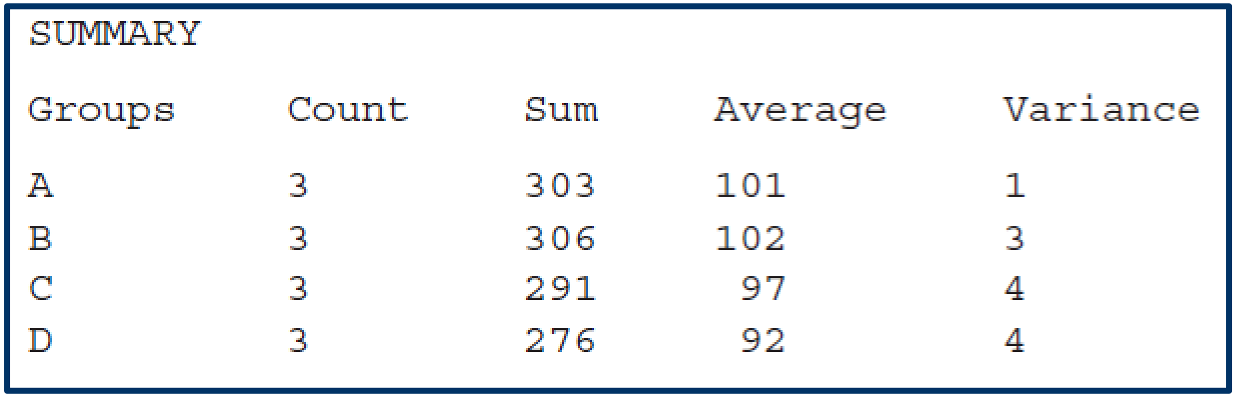


Critical Values of Studentized Range Distribution(q) for Familywise ALPHA = .05.

Denominator	Number of Groups (a.k.a. Treatments)							
DF	3	4	5	6	7	8	9	10
1	26.976	32.819	37.081	40.407	43.118	45.397	47.356	49.070
2	8.331	9.798	10.881	11.734	12.434	13.027	13.538	13.987
3	5.910	6.825	7.502	8.037	8.478	8.852	9.177	9.462
4	5.040	5.757	6.287	6.706	7.053	7.347	7.602	7.826
5	4.602	5.218	5.673	6.033	6.330	6.582	6.801	6.995
6	4.339	4.896	5.305	5.629	5.895	6.122	6.319	6.493
7	4.165	4.681	5.060	5.359	5.606	5.815	5.997	6.158
8	4.041	4.529	4.886	5.167	5.399	5.596	5.767	5.918
9	3.948	4.415	4.755	5.024	5.244	5.432	5.595	5.738
10	3.877	4.327	4.654	4.912	5.124	5.304	5.460	5.598
11	3.820	4.256	4.574	4.823	5.028	5.202	5.353	5.486
12	3.773	4.199	4.508	4.748	4.947	5.116	5.262	5.395
13	3.734	4.151	4.453	4.690	4.884	5.049	5.192	5.318
14	3.701	4.111	4.407	4.639	4.829	4.990	5.130	5.253
15	3.673	4.076	4.367	4.595	4.782	4.940	5.077	5.198
16	3.649	4.046	4.333	4.557	4.741	4.896	5.031	5.150
17	3.628	4.020	4.303	4.524	4.705	4.858	4.991	5.108
18	3.609	3.997	4.276	4.494	4.673	4.824	4.955	5.071
19	3.593	3.977	4.253	4.468	4.645	4.794	4.924	5.037
20	3.578	3.958	4.232	4.445	4.620	4.768	4.895	5.008
21	3.565	3.942	4.213	4.424	4.597	4.743	4.870	4.981
22	3.553	3.927	4.196	4.405	4.577	4.722	4.847	4.957
23	3.542	3.914	4.180	4.388	4.558	4.702	4.826	4.935
24	3.532	3.901	4.166	4.373	4.541	4.684	4.807	4.915
25	3.523	3.890	4.153	4.358	4.526	4.667	4.789	4.897
26	3.514	3.880	4.141	4.345	4.511	4.652	4.773	4.880
27	3.506	3.870	4.130	4.333	4.498	4.638	4.758	4.864
28	3.499	3.861	4.120	4.322	4.486	4.625	4.745	4.850
29	3.493	3.853	4.111	4.311	4.475	4.613	4.732	4.837
30	3.487	3.845	4.102	4.301	4.464	4.601	4.720	4.824

Exercise on Fisher and Fisher-Hayter Least Significant Difference (LSD/MLSD) tests

Let us reconsider the following dataset, for which MS_{Error} = 3 and ANOVA showed the presence of a significant variability between groups:



SUMMARY

Groups	Count	Sum	Average	Variance
A	3	303	101	1
B	3	306	102	3
C	3	291	97	4
D	3	276	92	4

Mean values can be arranged in increasing order:

$$\bar{x}_D = 92 \qquad \bar{x}_C = 97 \qquad \bar{x}_A = 101 \qquad \bar{x}_B = 102$$

In this case LSD = $[(3) \times (2/3)]^{1/2} \times t_{8,0.975} = 3.26$ ($\alpha = 0.05$).

Comparing this value with the differences between the means suggests that Groups D and C give results differing significantly from each other (5 > 3.26) and from the results obtained for Groups A and B (differences equal to 4, 5, 9 and 10, according to the case). Groups A and B (difference equal to 1) do not differ significantly from each other.

As shown by the table of critical values of the Studentized range distribution q, reported before, the critical value to use for a MLSD test on the same data would be 4.041.

Indeed, in the specific case n = 12 and h = 4, thus DF = n - h = 8 and the distribution with k = h -1 = 3 and DF = 8 has to be considered.

Since 4.041 > 4, the means of Groups C and A cannot be considered significantly different from each other if the MLSD test is used instead of the LSD one.

Generally speaking, the MLSD procedure is more conservative than the LSD one, i.e., it provides a lower number of significant differences for the same data.

Tukey test

The Tukey test, also called Tukey Honestly Significant Difference (HSD) test is a single step MCM whose Family-Wise Error Rate (FWER) for a family of C = h(h-1)/2 comparisons is exactly α. It was introduced by John Tukey in 1949.

The Tukey test is optimal in the sense that, among all procedures that give equal-width confidence intervals for all pairwise differences with a familywise confidence level at least (1- α), it provides the shortest intervals.

In other words, if the family consists of all pairwise comparisons and the Tukey test can be used, it will have shorter confidence intervals than any of the other single-step MCMs.

Tukey's test declares two means μ_i and μ_j significantly different if the absolute value of the difference between their estimates exceeds a critical value related to the Studentized range distribution with h, N-h degrees of freedom:

$$\left|\overline{X}_i - \overline{X}_j\right| \geq T_\alpha = q_{\alpha(h,\upsilon)}\sqrt{\frac{MS_{Error}}{n}} \qquad \text{with } \nu = N\text{-}h$$

Given a FWER of α, the Tukey confidence interval for $(\mu_i - \mu_j)$ is thus given by:

$$\left(\overline{X}_i - \overline{X}_j\right) \pm q_{\alpha(h,\upsilon)} \sqrt{\frac{MS_{Error}}{n}} \qquad i \neq j$$

Notably, all Tukey confidence intervals will have the same width since the latter depends on the total number of means h and on the common sample size, n, so it is not affected by the number of comparisons in the family.

The limitation of the Tukey's test is that it requires a balanced design (i.e., $n_1 = n_2 = ... = n_h = n$).

For unbalanced designs a simple modification, corresponding to the Tukey-Kramer test, is required.

Tukey-Kramer test

The Tukey-Kramer test arose from a modification to the Tukey test introduced in 1956 by the American statistician Clyde Kramer.

It is based on the confidence interval for $(\mu_i - \mu_j)$ expressed as:

$$\left(\bar{X}_i - \bar{X}_j\right) \pm q_{\alpha(h,\upsilon)} \sqrt{\frac{MS_{Error}}{2}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)} \qquad \text{with } \nu = N\text{-}h$$

Notably, if the modified **LSD**:

$$MLSD = q_{\alpha(h-1,\upsilon)} \sqrt{\frac{MS_{Error}}{2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

is considered and the variation of q values with the number of groups is evaluated from the table reported before, it can be seen that $q_{\alpha(h,\nu)} > q_{\alpha(h\text{-}1,\nu)}$. This means that the Tukey-Kramer test might lead to a lower number of significant differences between means than the modified Fisher-Hayter test.

Bonferroni test

The Bonferroni test was developed by the Italian mathematician Carlo Emilio Bonferroni in 1935 and consists in performing a t-test for each pair of means but accounting for the number, c, of pairwise comparisons.

This approach compensates for the increase in Type I error occurring when multiple pairwise comparisons between means are performed.

Indeed, the maximum familywise error rate, FWER, is α for any configuration of the populations means.

Given a FWER = α, the Bonferroni *confidence interval* for $(\mu_i - \mu_j)$ is given by:

$$\left(\overline{X}_i - \overline{X}_j\right) \pm t_{\alpha^*,\upsilon}\sqrt{MS_{Error}\left(\frac{1}{n_i}+\frac{1}{n_j}\right)} \qquad \text{with } \nu = N\text{-}h$$

where $\alpha^* = (\alpha/c)$.

Duncan multiple range test

The Duncan multiple range test was developed by David B. Duncan in 1955 to increase the protection against Type II error.

To apply Duncan multiple range test for ***equal sample sizes*** (n) the averages of the h treatments are arranged in ascending order, and the standard error of each average is determined as:

$$S_{\overline{X}_i} = \sqrt{\frac{MS_{Error}}{n}}$$

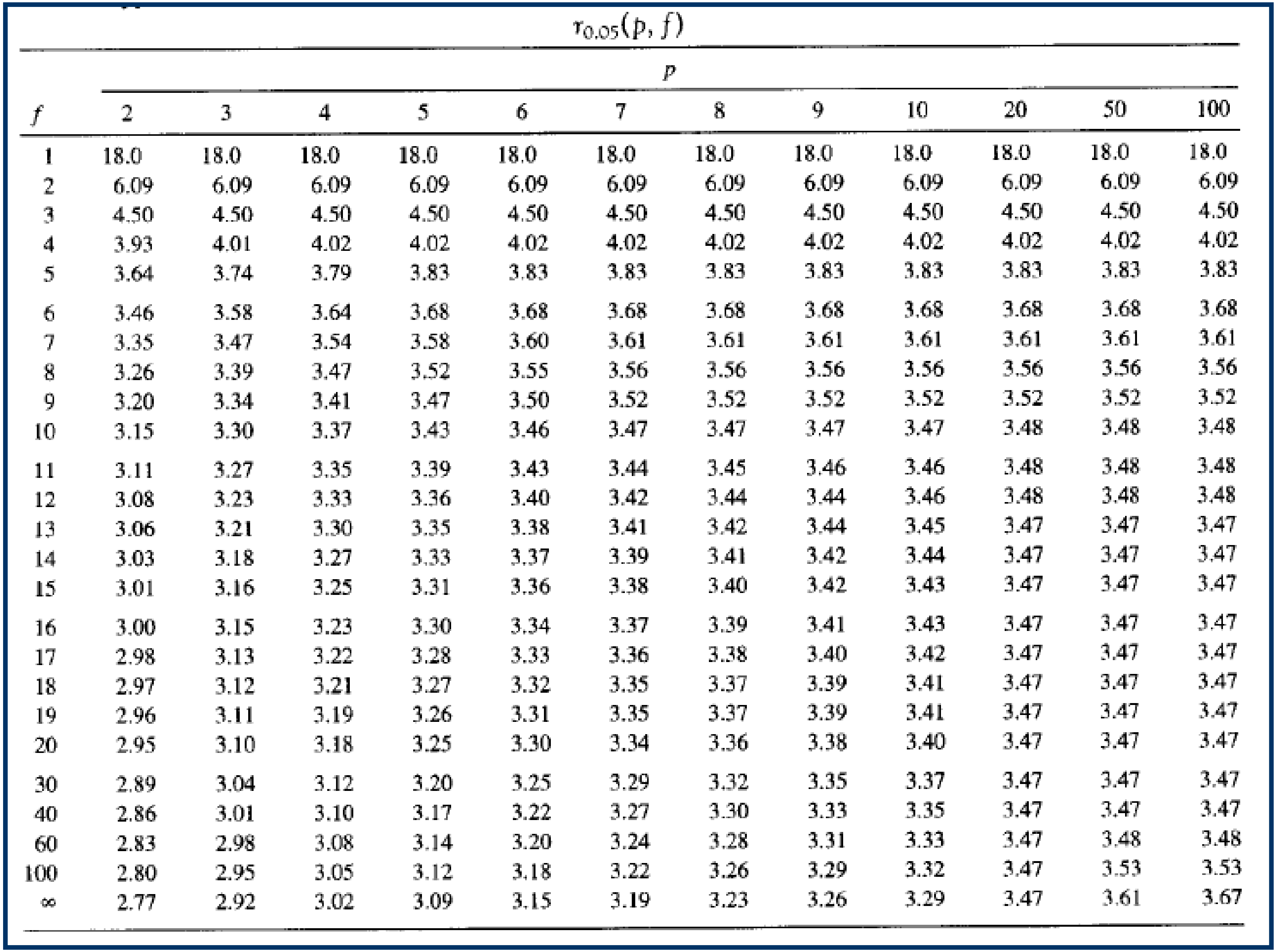
Duncan's table of significant range coefficients, shown in the next slide, is then considered to obtain values $r_{\alpha}(p,f)$ for p = 2,3,….,h where α is the significance level and f = N-h is the number of degrees of freedom.

The coefficients are subsequently converted into a set of h-1 significance ranges R_p for p = 2,3,…, h, by calculating:

$$R_p = r_{\alpha}(p, f) S_{\overline{X}_i} \qquad \text{for} \quad p = 2,3,\ldots\ldots,h$$

The differences observed between means are then tested, beginning with the difference between the largest and the smallest mean, which is compared with the R_p value obtained for p = h (indicated as R_h).

Table of significant ranges for Duncan's Multiple Range Test



$r_{0.05}(p, f)$

	p											
f	2	3	4	5	6	7	8	9	10	20	50	100
1	18.0	18.0	18.0	18.0	18.0	18.0	18.0	18.0	18.0	18.0	18.0	18.0
2	6.09	6.09	6.09	6.09	6.09	6.09	6.09	6.09	6.09	6.09	6.09	6.09
3	4.50	4.50	4.50	4.50	4.50	4.50	4.50	4.50	4.50	4.50	4.50	4.50
4	3.93	4.01	4.02	4.02	4.02	4.02	4.02	4.02	4.02	4.02	4.02	4.02
5	3.64	3.74	3.79	3.83	3.83	3.83	3.83	3.83	3.83	3.83	3.83	3.83
6	3.46	3.58	3.64	3.68	3.68	3.68	3.68	3.68	3.68	3.68	3.68	3.68
7	3.35	3.47	3.54	3.58	3.60	3.61	3.61	3.61	3.61	3.61	3.61	3.61
8	3.26	3.39	3.47	3.52	3.55	3.56	3.56	3.56	3.56	3.56	3.56	3.56
9	3.20	3.34	3.41	3.47	3.50	3.52	3.52	3.52	3.52	3.52	3.52	3.52
10	3.15	3.30	3.37	3.43	3.46	3.47	3.47	3.47	3.47	3.48	3.48	3.48
11	3.11	3.27	3.35	3.39	3.43	3.44	3.45	3.46	3.46	3.48	3.48	3.48
12	3.08	3.23	3.33	3.36	3.40	3.42	3.44	3.44	3.46	3.48	3.48	3.48
13	3.06	3.21	3.30	3.35	3.38	3.41	3.42	3.44	3.45	3.47	3.47	3.47
14	3.03	3.18	3.27	3.33	3.37	3.39	3.41	3.42	3.44	3.47	3.47	3.47
15	3.01	3.16	3.25	3.31	3.36	3.38	3.40	3.42	3.43	3.47	3.47	3.47
16	3.00	3.15	3.23	3.30	3.34	3.37	3.39	3.41	3.43	3.47	3.47	3.47
17	2.98	3.13	3.22	3.28	3.33	3.36	3.38	3.40	3.42	3.47	3.47	3.47
18	2.97	3.12	3.21	3.27	3.32	3.35	3.37	3.39	3.41	3.47	3.47	3.47
19	2.96	3.11	3.19	3.26	3.31	3.35	3.37	3.39	3.41	3.47	3.47	3.47
20	2.95	3.10	3.18	3.25	3.30	3.34	3.36	3.38	3.40	3.47	3.47	3.47
30	2.89	3.04	3.12	3.20	3.25	3.29	3.32	3.35	3.37	3.47	3.47	3.47
40	2.86	3.01	3.10	3.17	3.22	3.27	3.30	3.33	3.35	3.47	3.47	3.47
60	2.83	2.98	3.08	3.14	3.20	3.24	3.28	3.31	3.33	3.47	3.48	3.48
100	2.80	2.95	3.05	3.12	3.18	3.22	3.26	3.29	3.32	3.47	3.53	3.53
∞	2.77	2.92	3.02	3.09	3.15	3.19	3.23	3.26	3.29	3.47	3.61	3.67

The procedure is subsequently repeated for the difference between the largest and the second smallest mean, which is compared to the R_{h-1} value.

The test proceeds with the same logic, until all means have been compared with the largest mean.

If an observed difference is greater than the corresponding least significant range the two means under evaluation can be considered significantly different.

Dunnett test for comparisons with a control

In many experiments one of the treatments (whose number is indicated as *a*) is a control, and the analyst is interested in comparing each of the other *a-1* treatment means with the one of control.

In 1964 the Canadian statistician Charles Dunnett developed a procedure to make such comparisons.
Let us suppose that treatment *a* is the control, thus the hypotheses under test are:

$$H_0: \mu_i = \mu_a$$
$$H_1: \mu_i \neq \mu_a$$

for i = 1, 2, …., a-1.

The observed differences in the sample means are computed for each i value:

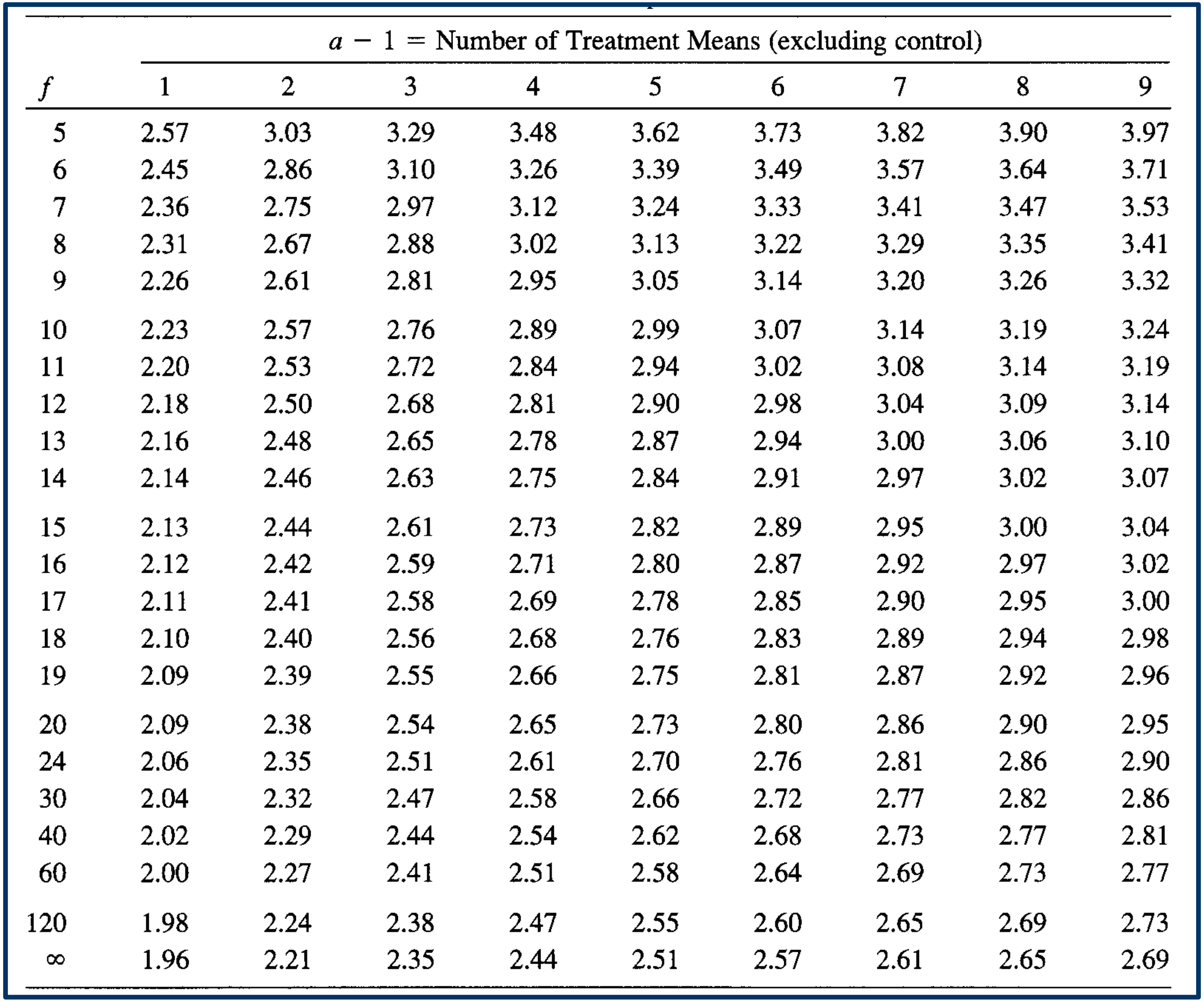
$$|\bar{y}_{i.} - \bar{y}_{a.}| \qquad i = 1, 2, \ldots, a-1$$

The null hypothesis H_0 is rejected, with a type I error rate α, if:

$$|\bar{y}_{i.} - \bar{y}_{a.}| > d_\alpha(a-1, f)\sqrt{MS_E\left(\frac{1}{n_i} + \frac{1}{n_a}\right)}$$

where the d_α (a-1, f) value is provided by the table shown in the next slide:

Table of critical values for Dunnett’s Test for α = 0.05 (two-sided comparisons)



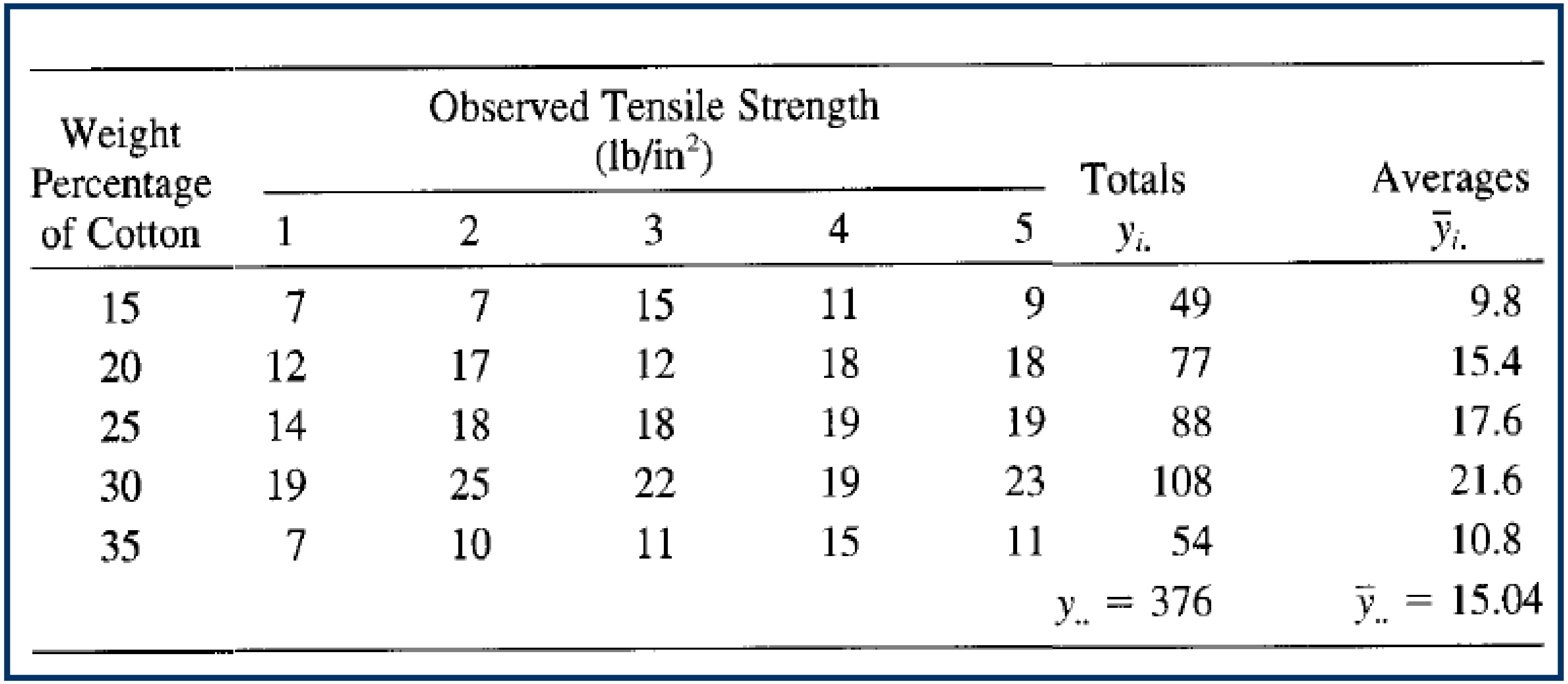
	$a - 1$ = Number of Treatment Means (excluding control)								
f	1	2	3	4	5	6	7	8	9
5	2.57	3.03	3.29	3.48	3.62	3.73	3.82	3.90	3.97
6	2.45	2.86	3.10	3.26	3.39	3.49	3.57	3.64	3.71
7	2.36	2.75	2.97	3.12	3.24	3.33	3.41	3.47	3.53
8	2.31	2.67	2.88	3.02	3.13	3.22	3.29	3.35	3.41
9	2.26	2.61	2.81	2.95	3.05	3.14	3.20	3.26	3.32
10	2.23	2.57	2.76	2.89	2.99	3.07	3.14	3.19	3.24
11	2.20	2.53	2.72	2.84	2.94	3.02	3.08	3.14	3.19
12	2.18	2.50	2.68	2.81	2.90	2.98	3.04	3.09	3.14
13	2.16	2.48	2.65	2.78	2.87	2.94	3.00	3.06	3.10
14	2.14	2.46	2.63	2.75	2.84	2.91	2.97	3.02	3.07
15	2.13	2.44	2.61	2.73	2.82	2.89	2.95	3.00	3.04
16	2.12	2.42	2.59	2.71	2.80	2.87	2.92	2.97	3.02
17	2.11	2.41	2.58	2.69	2.78	2.85	2.90	2.95	3.00
18	2.10	2.40	2.56	2.68	2.76	2.83	2.89	2.94	2.98
19	2.09	2.39	2.55	2.66	2.75	2.81	2.87	2.92	2.96
20	2.09	2.38	2.54	2.65	2.73	2.80	2.86	2.90	2.95
24	2.06	2.35	2.51	2.61	2.70	2.76	2.81	2.86	2.90
30	2.04	2.32	2.47	2.58	2.66	2.72	2.77	2.82	2.86
40	2.02	2.29	2.44	2.54	2.62	2.68	2.73	2.77	2.81
60	2.00	2.27	2.41	2.51	2.58	2.64	2.69	2.73	2.77
120	1.98	2.24	2.38	2.47	2.55	2.60	2.65	2.69	2.73
∞	1.96	2.21	2.35	2.44	2.51	2.57	2.61	2.65	2.69

Application of different tests for multiple comparison of means

1) Effect of the percent of cotton on the tensile strength of a synthetic fiber

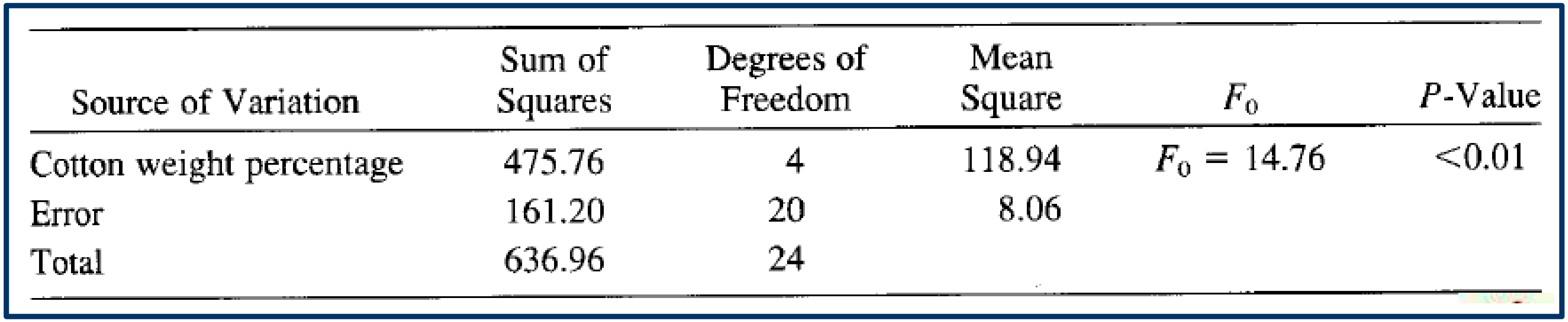
An engineer decides to test for tensile strength specimens of a synthetic fiber containing one of five different levels of cotton weight percent: 15, 20, 25, 30 and 35%.
In this case h = 5 and n = 5 and the 25 measurements of tensile strength are performed in random order.

The resulting data are:



Weight Percentage of Cotton	Observed Tensile Strength (lb/in^2)					Totals	Averages
	1	2	3	4	5	$y_{i.}$	$\bar{y}_{i.}$
15	7	7	15	11	9	49	9.8
20	12	17	12	18	18	77	15.4
25	14	18	18	19	19	88	17.6
30	19	25	22	19	23	108	21.6
35	7	10	11	15	11	54	10.8
						$y_{..} = 376$	$\bar{y}_{..} = 15.04$

The results of ANOVA are the following:



Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0	P-Value
Cotton weight percentage	475.76	4	118.94	$F_0 = 14.76$	<0.01
Error	161.20	20	8.06		
Total	636.96	24			

The F_0 value is high enough to lead to a p-value lower than 0.01, which means that some significant differences exist between the different groups of samples, i.e., the cotton percentage has some influence on the observed tensile strength.

Different tests can then be performed to understand which means differ from each other.

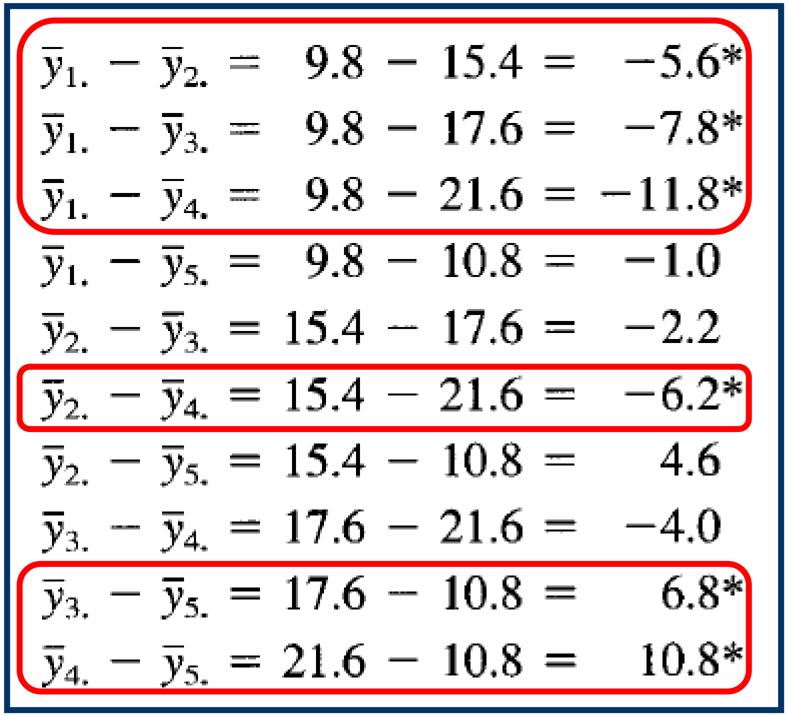
1a) Tukey test

The Tukey’s test can be used since the numbers of data in each group are the same.
For α = 0.05 the critical value for the test is:

$$T_{0.05} = q_{0.05}(5,\ 20)\sqrt{\frac{MS_E}{n}} = 4.23\sqrt{\frac{8.06}{5}} = 5.37$$

Thus, significant differences would be inferred for any pair of groups means differing in absolute value by more than 5.37.

As shown in the table, six pairs of means, indicated by an asterisk, are found to be significant different:



$$\begin{array}{rclcl}
\bar{y}_{1.} - \bar{y}_{2.} & = & 9.8 - 15.4 & = & -5.6^{*} \\
\bar{y}_{1.} - \bar{y}_{3.} & = & 9.8 - 17.6 & = & -7.8^{*} \\
\bar{y}_{1.} - \bar{y}_{4.} & = & 9.8 - 21.6 & = & -11.8^{*} \\
\bar{y}_{1.} - \bar{y}_{5.} & = & 9.8 - 10.8 & = & -1.0 \\
\bar{y}_{2.} - \bar{y}_{3.} & = & 15.4 - 17.6 & = & -2.2 \\
\bar{y}_{2.} - \bar{y}_{4.} & = & 15.4 - 21.6 & = & -6.2^{*} \\
\bar{y}_{2.} - \bar{y}_{5.} & = & 15.4 - 10.8 & = & 4.6 \\
\bar{y}_{3.} - \bar{y}_{4.} & = & 17.6 - 21.6 & = & -4.0 \\
\bar{y}_{3.} - \bar{y}_{5.} & = & 17.6 - 10.8 & = & 6.8^{*} \\
\bar{y}_{4.} - \bar{y}_{5.} & = & 21.6 - 10.8 & = & 10.8^{*}
\end{array}$$

1b) Fisher LSD test

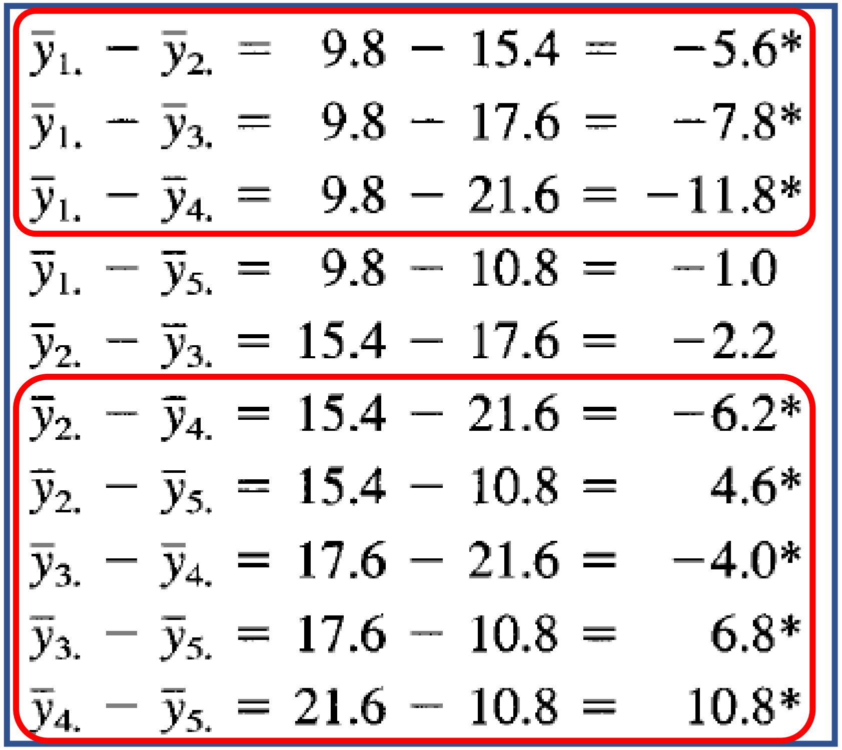
For α = 0.05 the critical value for the Fisher LSD test is:

$$\text{LSD} = t_{.025,20}\sqrt{\frac{2MS_E}{n}} = 2.086\sqrt{\frac{2(8.06)}{5}} = 3.75$$

Thus, significant differences would be inferred for any pair of groups means differing in absolute value by more than 3.75.

As apparent from the table, eight pairs of means, indicated by an asterisk, are now found to be significant different:

Calculations confirm that Tukey's test is more conservative than the Fisher LSD one, i.e., it leads to evidence a lower number of significant differences between means.



$$\begin{aligned}
\bar{y}_{1.} - \bar{y}_{2.} &= 9.8 - 15.4 = -5.6^{*}\\
\bar{y}_{1.} - \bar{y}_{3.} &= 9.8 - 17.6 = -7.8^{*}\\
\bar{y}_{1.} - \bar{y}_{4.} &= 9.8 - 21.6 = -11.8^{*}\\
\bar{y}_{1.} - \bar{y}_{5.} &= 9.8 - 10.8 = -1.0\\
\bar{y}_{2.} - \bar{y}_{3.} &= 15.4 - 17.6 = -2.2\\
\bar{y}_{2.} - \bar{y}_{4.} &= 15.4 - 21.6 = -6.2^{*}\\
\bar{y}_{2.} - \bar{y}_{5.} &= 15.4 - 10.8 = 4.6^{*}\\
\bar{y}_{3.} - \bar{y}_{4.} &= 17.6 - 21.6 = -4.0^{*}\\
\bar{y}_{3.} - \bar{y}_{5.} &= 17.6 - 10.8 = 6.8^{*}\\
\bar{y}_{4.} - \bar{y}_{5.} &= 21.6 - 10.8 = 10.8^{*}
\end{aligned}$$

1c) Duncan multiple range test

In this case the group means must be first arranged in ascending order:

$$\begin{aligned} \bar{y}_{1.} &= 9.8 \\ \bar{y}_{5.} &= 10.8 \\ \bar{y}_{2.} &= 15.4 \\ \bar{y}_{3.} &= 17.6 \\ \bar{y}_{4.} &= 21.6 \end{aligned}$$

The following values need to be used:

$MS_E = 8.06$, N = 25, h = 5, n = 5

The standard error of each average is then:

$$S_{\bar{Y}_i} = \sqrt{\frac{MS_{Error}}{n}} = 1.27$$

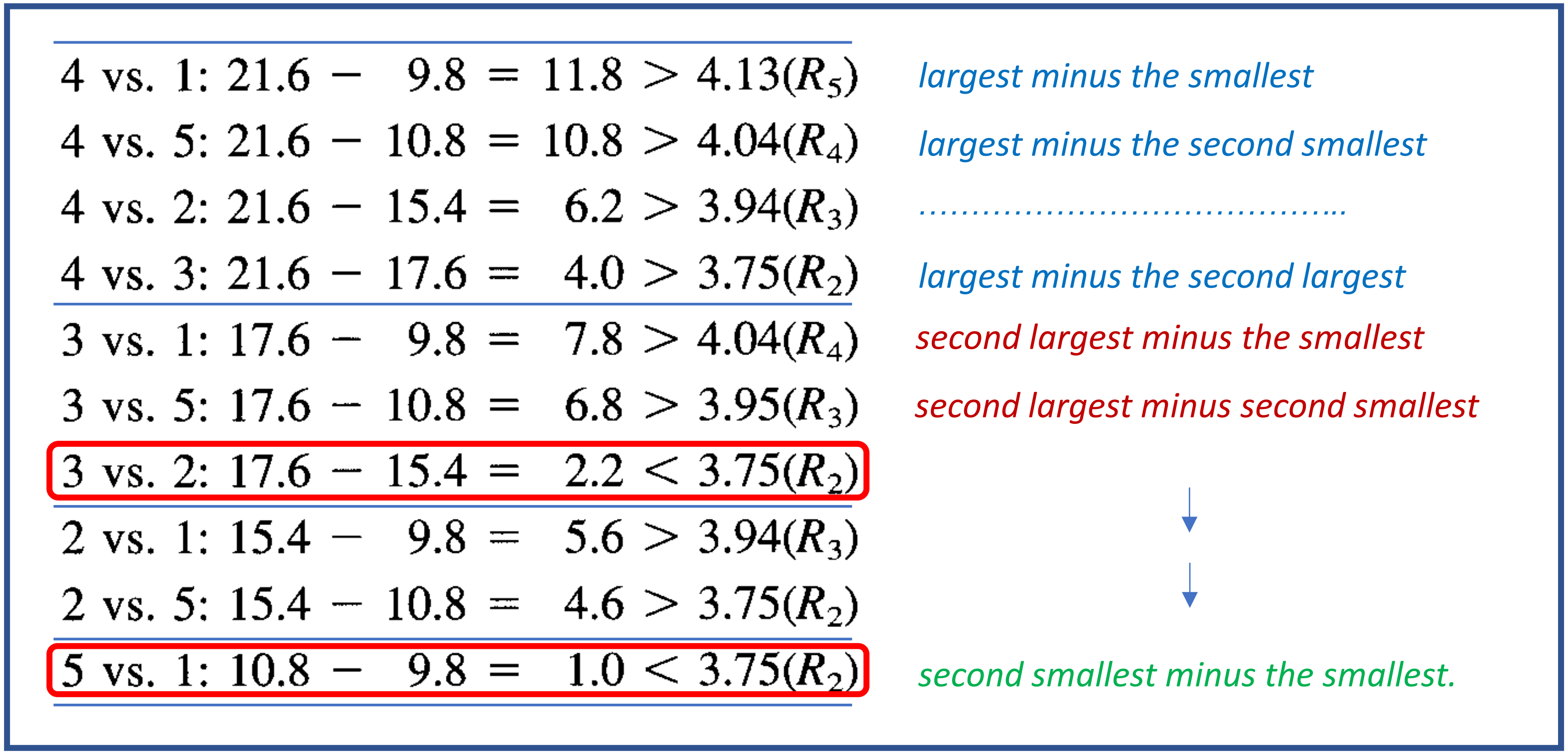
In this case f = N-h = 20, thus the following values are extracted from Duncan's table:

$r_{0.05}(2, 20) = 2.95$, $r_{0.05}(3, 20) = 3.10$, $r_{0.05}(4, 20) = 3.18$ and $r_{0.05}(5, 20) = 3.25$

The least significant ranges are:

$$\begin{aligned} R_2 &= r_{0.05}(2, 20)S_{\bar{y}_{i.}} = (2.95)(1.27) = 3.75 \\ R_3 &= r_{0.05}(3, 20)S_{\bar{y}_{i.}} = (3.10)(1.27) = 3.94 \\ R_4 &= r_{0.05}(4, 20)S_{\bar{y}_{i.}} = (3.18)(1.27) = 4.04 \\ R_5 &= r_{0.05}(5, 20)S_{\bar{y}_{i.}} = (3.25)(1.27) = 4.13 \end{aligned}$$

The tests are subsequently performed in the following order:
the largest minus the smallest, the largest minus the second smallest and then up to the largest minus the second largest; then the second largest minus the smallest, the second largest minus the second smallest, and so on, finishing with the second smallest minus the smallest.



$4 \text{ vs. } 1: 21.6 - 9.8 = 11.8 > 4.13(R_5)$	*largest minus the smallest*
$4 \text{ vs. } 5: 21.6 - 10.8 = 10.8 > 4.04(R_4)$	*largest minus the second smallest*
$4 \text{ vs. } 2: 21.6 - 15.4 = 6.2 > 3.94(R_3)$	……………………………………
$4 \text{ vs. } 3: 21.6 - 17.6 = 4.0 > 3.75(R_2)$	*largest minus the second largest*
$3 \text{ vs. } 1: 17.6 - 9.8 = 7.8 > 4.04(R_4)$	*second largest minus the smallest*
$3 \text{ vs. } 5: 17.6 - 10.8 = 6.8 > 3.95(R_3)$	*second largest minus second smallest*
$\boxed{3 \text{ vs. } 2: 17.6 - 15.4 = 2.2 < 3.75(R_2)}$	
$2 \text{ vs. } 1: 15.4 - 9.8 = 5.6 > 3.94(R_3)$	↓
$2 \text{ vs. } 5: 15.4 - 10.8 = 4.6 > 3.75(R_2)$	↓
$\boxed{5 \text{ vs. } 1: 10.8 - 9.8 = 1.0 < 3.75(R_2)}$	*second smallest minus the smallest.*

As a result, there are significant differences between all pairs of means except 3 vs 2 and 5 vs 1.

Consequently, the Duncan's multiple range test and the LSD test produce identical conclusions.

1d) Dunnett test

Since the Dunnett's test was developed to compare group means with the one obtained for a control, the group represented by the synthetic fiber containing 35% of cotton will be supposed to represent the control.

According to the Dunnett's test nomenclature, *a* = 5 and f = 20, thus the critical difference, for α = 0.05, can be calculated as follows (note that in this case each group includes the same number of data):

$$d_\alpha(a-1,f)\sqrt{MS_E\left(\frac{1}{n_i}+\frac{1}{n_a}\right)} = d_{0.05}(4,\,20)\sqrt{\frac{2MS_E}{n}} = 2.65\sqrt{\frac{2(8.06)}{5}} = 4.76$$

Thus, any group mean that differs from the control by more than 4.76 can be declared significantly different from the control.

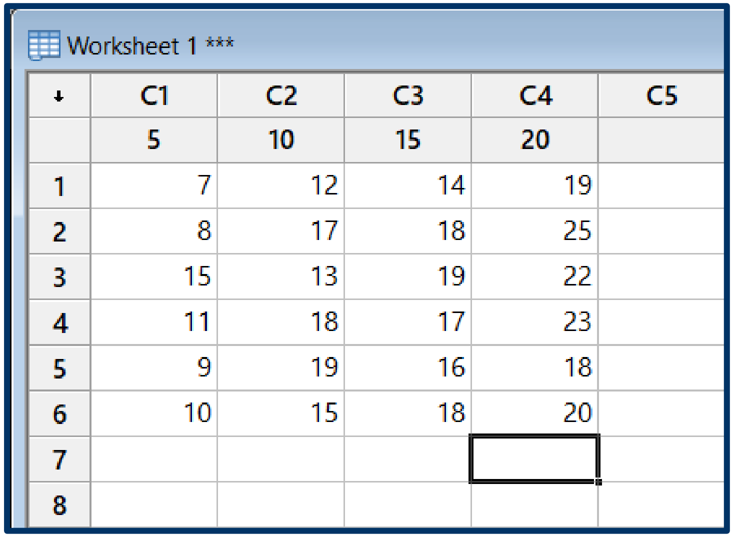
The observed differences are:

$$\begin{array}{l}
1 \text{ vs. } 5\text{: } \bar{y}_{1.} - \bar{y}_{5.} = 9.8 - 10.8 = -1.0 \\
2 \text{ vs. } 5\text{: } \bar{y}_{2.} - \bar{y}_{5.} = 15.4 - 10.8 = 4.6 \\
3 \text{ vs. } 5\text{: } \bar{y}_{3.} - \bar{y}_{5.} = 17.6 - 10.8 = 6.8* \\
4 \text{ vs. } 5\text{: } \bar{y}_{4.} - \bar{y}_{5.} = 21.6 - 10.8 = 10.8*
\end{array}$$

Consequently, only means of groups 3 and 4 differ from that of the control group.

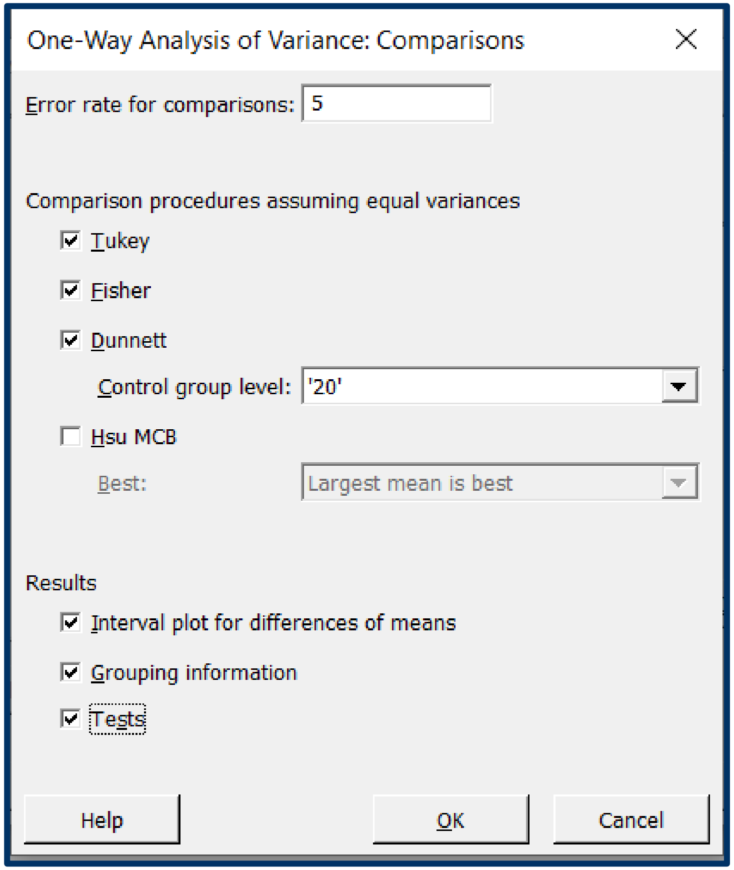
Use of Minitab 18 for multiple comparison of means

Let us reconsider the data set relevant to the tensile strength of paper as a function of hardwood percentages, with four percentages considered (5, 10, 15 and 20%) and six replicated measurements made randomly for each percentage:



Worksheet 1 ***

↓	C1	C2	C3	C4	C5
	5	10	15	20	
1	7	12	14	19	
2	8	17	18	25	
3	15	13	19	22	
4	11	18	17	23	
5	9	19	16	18	
6	10	15	18	20	
7					
8					



One-Way Analysis of Variance: Comparisons

Error rate for comparisons: 5

Comparison procedures assuming equal variances

☑ Tukey

☑ Fisher

☑ Dunnett

Control group level: '20'

☐ Hsu MCB

Best: Largest mean is best

Results

☑ Interval plot for differences of means

☑ Grouping information

☑ Tests

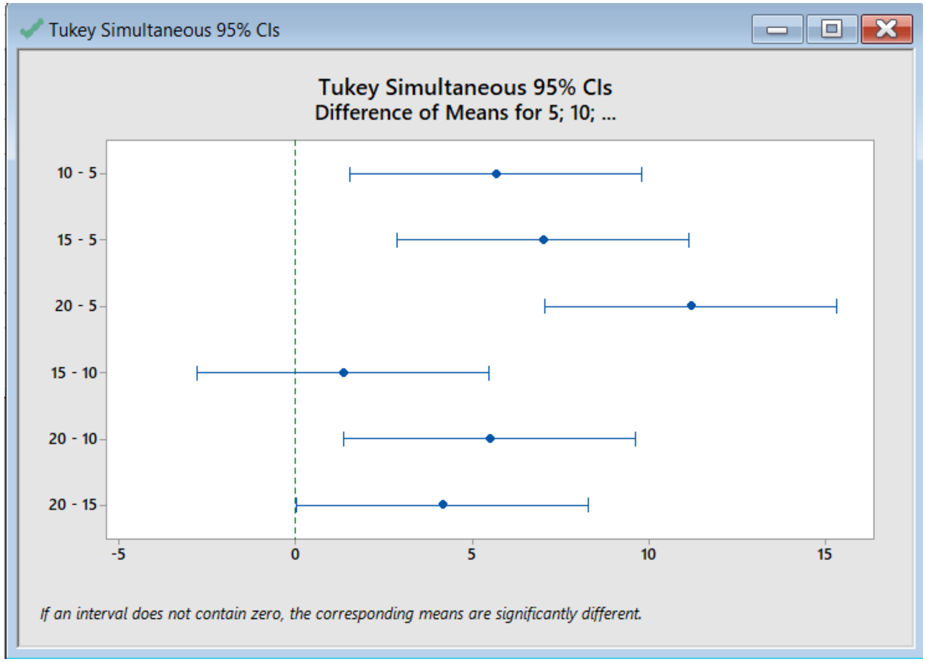
The option “Comparisons” in the One-Way ANOVA menu of the Minitab 18 software covers four of the multiple comparisons of means described so far: Tukey (or Tukey-Kramer), Fisher and Dunnett.

In the case of Dunnett’s test a control group has to be specified (in the example the group related to a 20% hardwood percentage has been indicated).

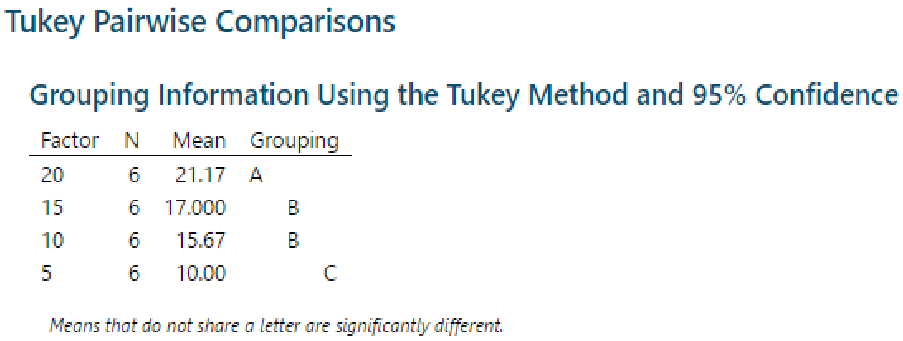
In the case of the Tukey test Minitab shows confidence intervals for all the possible differences of means.

When a specific interval does not include the 0 value a significant difference is inferred between the corresponding means.

In the "Session" window of the Minitab software a table is reported after the Tukey test, with capital letters indicating which means can be grouped together, since a not significant difference has been found between them (in the specific case means of tensile strength for paper samples containing 10 and 15% of hardwood).



If an interval does not contain zero, the corresponding means are significantly different.



Tukey Pairwise Comparisons

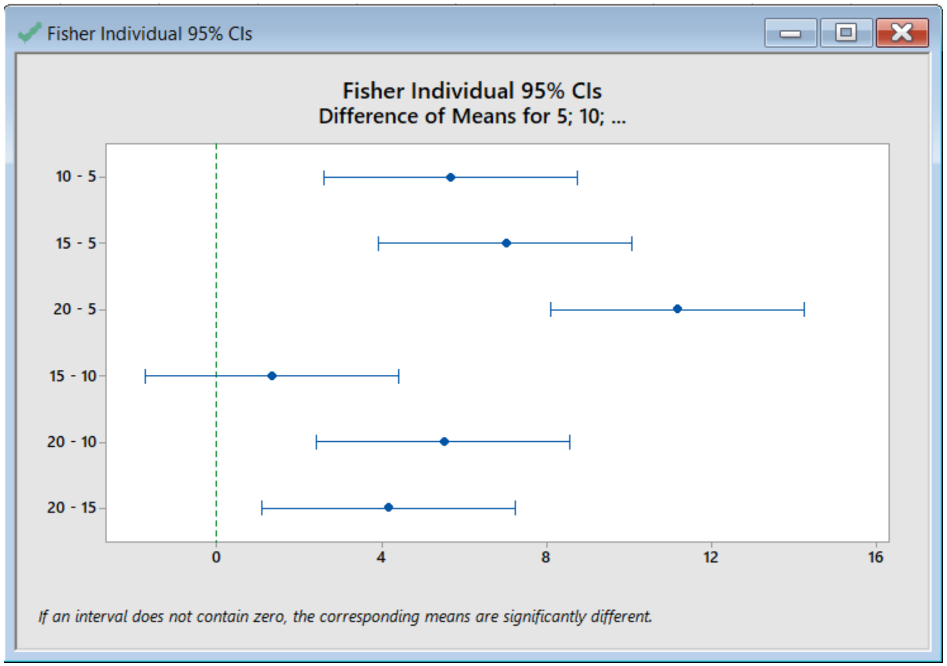
Grouping Information Using the Tukey Method and 95% Confidence

Factor	N	Mean	Grouping
20	6	21.17	A
15	6	17.000	B
10	6	15.67	B
5	6	10.00	C

Means that do not share a letter are significantly different.

A similar approach is adopted for the Fisher LSD test.

Notably, the Fisher confidence intervals are narrower than those of the Tukey’s test, thus the observation of significant differences is more likely.



If an interval does not contain zero, the corresponding means are significantly different.

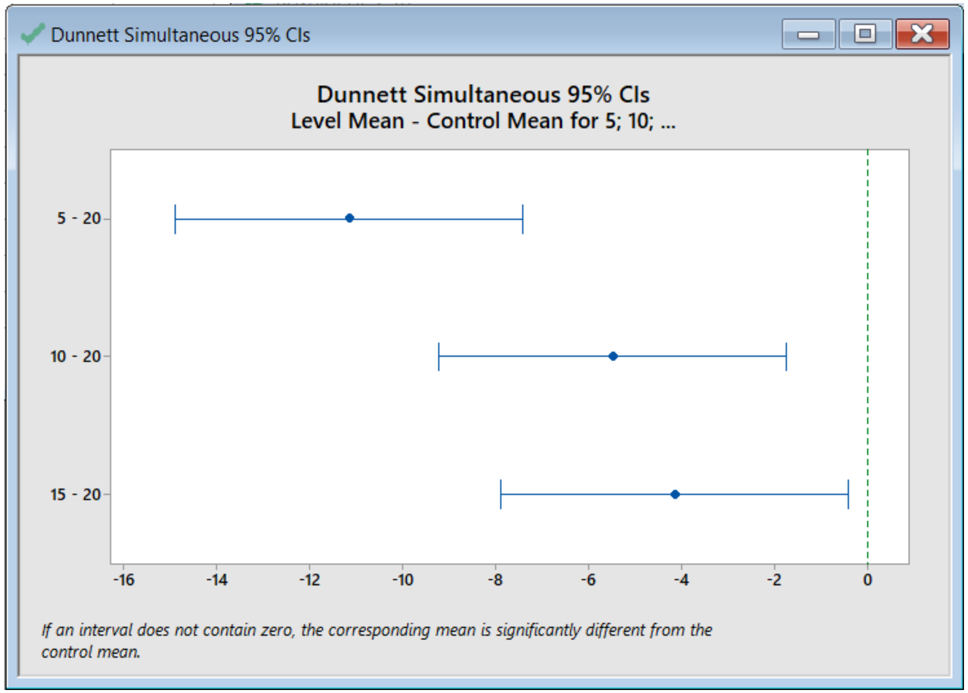
Fisher Pairwise Comparisons

Grouping Information Using the Fisher LSD Method and 95% Confidence

Factor	N	Mean	Grouping		
20	6	21.17	A		
15	6	17.000		B	
10	6	15.67		B	
5	6	10.00			C

When the Dunnett’s test is performed the confidence intervals for differences between group means and the control mean are reported.

If intervals do not include the 0 value a significant difference with the control mean can be inferred.

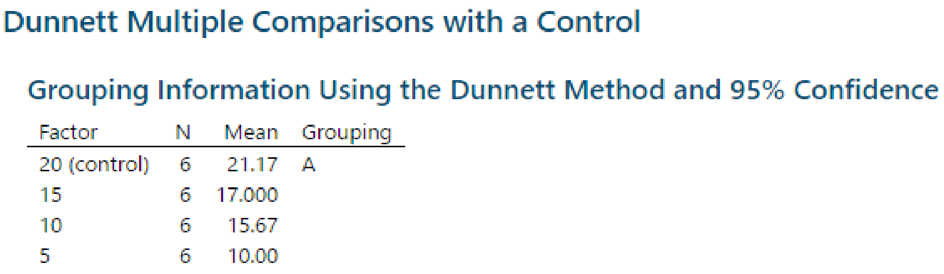


Dunnett Simultaneous 95% CIs
Level Mean - Control Mean for 5; 10; ...

If an interval does not contain zero, the corresponding mean is significantly different from the control mean.

In the summary table group means eventually not differing from the control one would be classified with the same letter as the control.

This was not the case in the specific example.



Dunnett Multiple Comparisons with a Control

Grouping Information Using the Dunnett Method and 95% Confidence

Factor	N	Mean	Grouping
20 (control)	6	21.17	A
15	6	17.000	
10	6	15.67	
5	6	10.00	

Multiple comparison between means using non-parametric methods

As other statistical procedures, even the multiple comparison between means can be performed using non-parametric (also called distribution-free) methods, i.e., methods that make no assumption about the distribution from which data are taken.

Such tests are useful when the assumption of normality cannot be proved, or it has been demonstrated to be not true.

An example of them is the Kruskal-Wallis test, developed in 1952 by American statisticians William Kruskal and Wilson Allen Wallis, whose goal is assessing if at least two of the medians related to several groups of data differ significantly, thus it is often considered the nonparametric equivalent of ANOVA, although it can be performed also on groups including different numbers of data.

As in other non-parametric tests, the ranks of data must be calculated.
In particular, a single ranking is made for data arising from all groups together, i.e., the ranking is made from 1 to N by ignoring group membership.

Moreover, if present, tied values are assigned ranks corresponding to the average of the ranks they would have received if they had not been tied.

The general statistic for the Kruskal-Wallis test is:

$$H = (N-1)\frac{\sum_{i=1}^{g} n_i(\bar{r}_{i\cdot} - \bar{r})^2}{\sum_{i=1}^{g}\sum_{j=1}^{n_i}(r_{ij} - \bar{r})^2}$$

where:

N is the total number of observations

g is the number of groups

n_i is the number of observations in group i

r_{ij} is the rank (among all observations) of observation j from group i

$\bar{r}_{i\cdot} = \frac{\sum_{j=1}^{n_i} r_{ij}}{n_i}$ is the average rank of all observations in group i

$\bar{r} = \frac{1}{2}(N+1)$ is the average of all the r_{ij} values

Starting from the general definition of the H statistic:

$$H = (N-1)\frac{\sum_{i=1}^{g} n_i(\bar{r}_{i\cdot} - \bar{r})^2}{\sum_{i=1}^{g}\sum_{j=1}^{n_i}(r_{ij} - \bar{r})^2}$$

it can be seen that when the average rankings referred to different groups are similar, which implies, indirectly, that observations in the different groups are comparable, the realization of the H statistic is lower than when the average rankings are not similar.

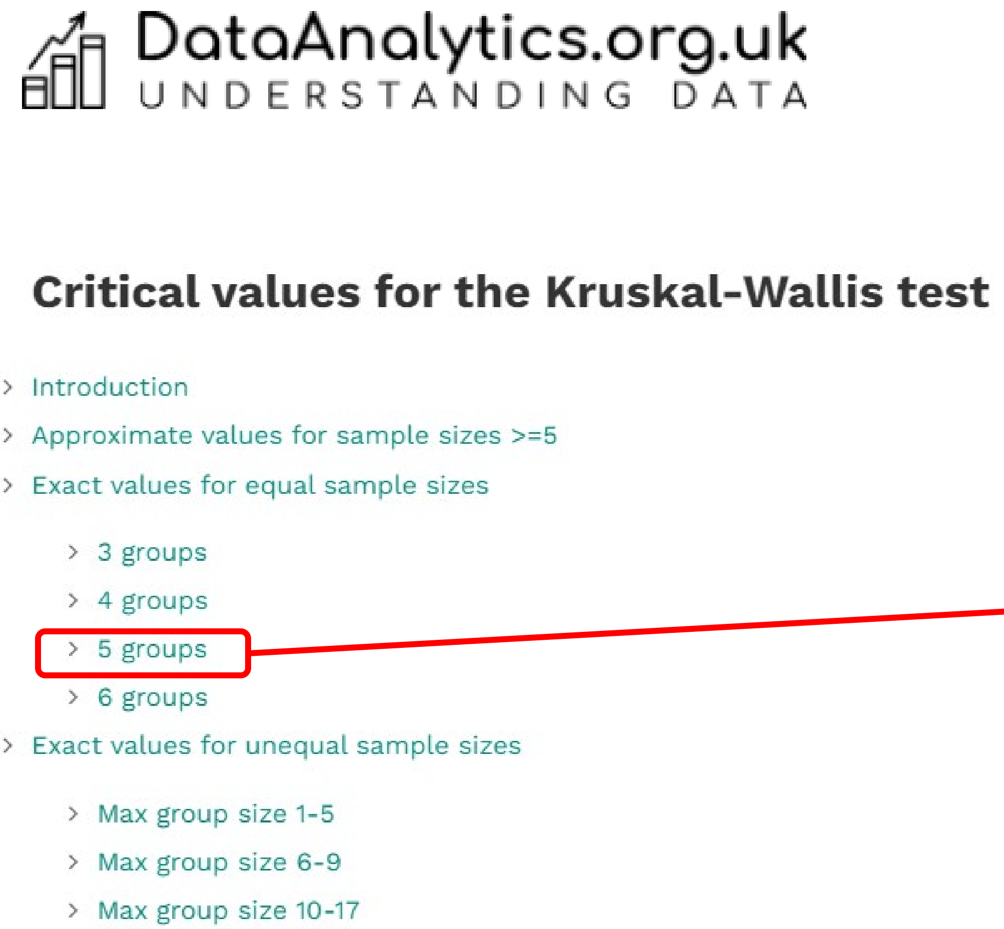
This explains why the test indicates the presence of a significant difference between the groups' medians if the value assumed by H is higher than a critical value.

The latter can be obtained from a χ^2 distribution with g-1 degrees of freedom when N > 15 and each n_i is not lower than 5.

Special tables should be used for smaller numbers of measurements.

A source of critical values for the Kruskal-Wallis test can be found on the Internet site:

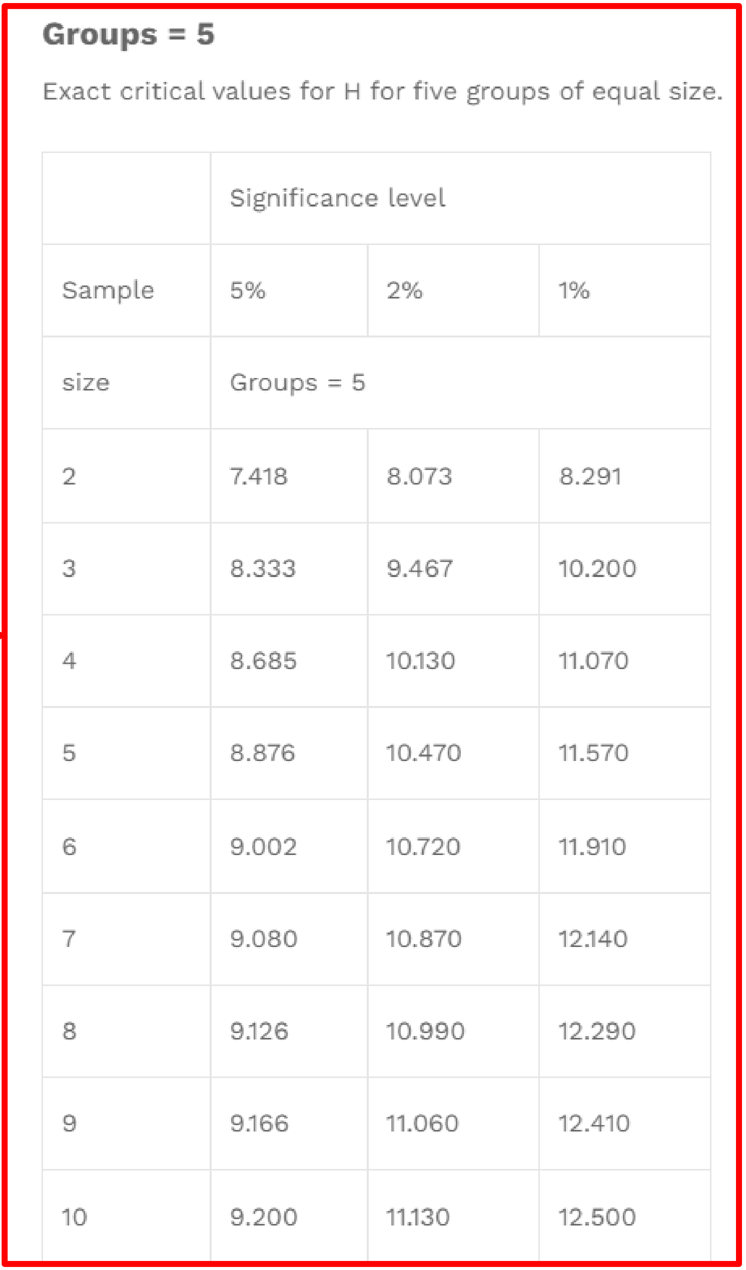
https://www.dataanalytics.org.uk/critical-values-for-the-kruskal-wallis-test/



DataAnalytics.org.uk
UNDERSTANDING DATA

Critical values for the Kruskal-Wallis test

- Introduction
- Approximate values for sample sizes >=5
- Exact values for equal sample sizes
  - 3 groups
  - 4 groups
  - 5 groups
  - 6 groups
- Exact values for unequal sample sizes
  - Max group size 1-5
  - Max group size 6-9
  - Max group size 10-17



Groups = 5

Exact critical values for H for five groups of equal size.

	Significance level		
Sample	5%	2%	1%
size	Groups = 5		
2	7.418	8.073	8.291
3	8.333	9.467	10.200
4	8.685	10.130	11.070
5	8.876	10.470	11.570
6	9.002	10.720	11.910
7	9.080	10.870	12.140
8	9.126	10.990	12.290
9	9.166	11.060	12.410
10	9.200	11.130	12.500

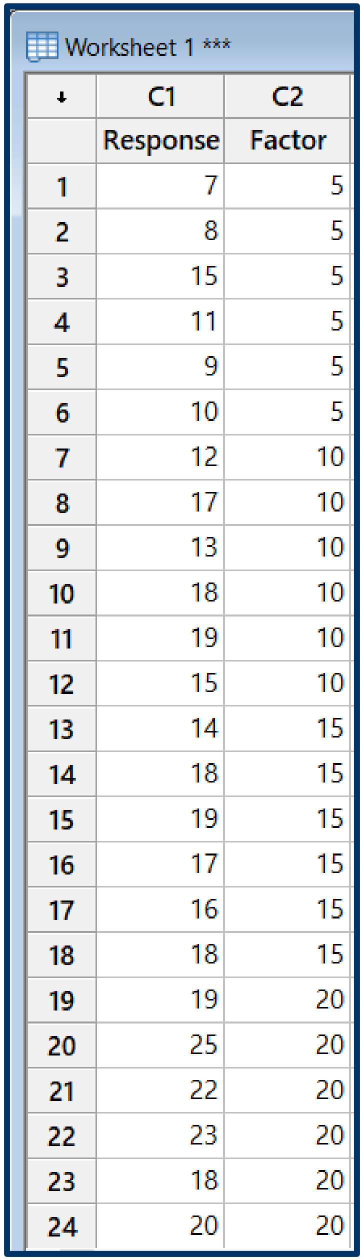
Use of Minitab 18 to perform the Kruskal-Wallis test

A different approach has to be adopted to input data in the Minitab 18 worksheet before proceeding to a Kruskal-Wallis test.

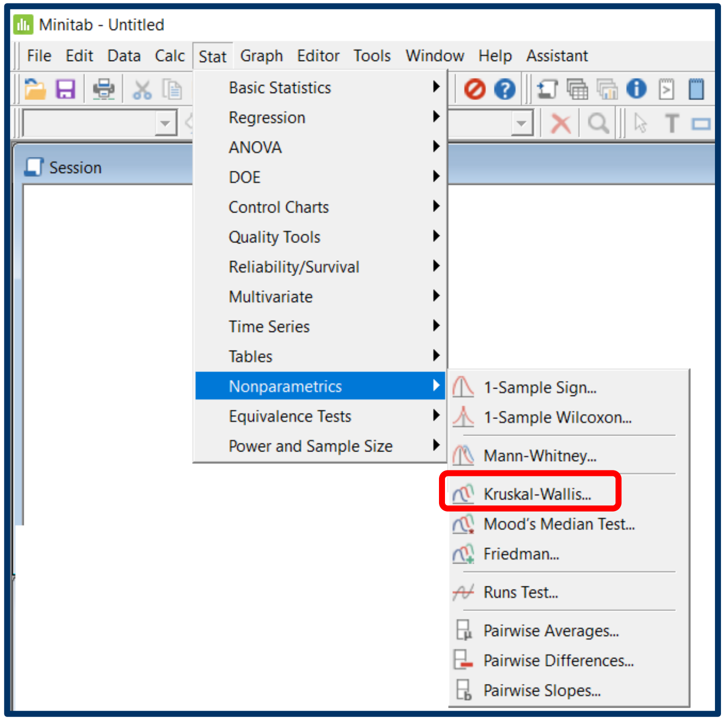
Specifically, one column is used to introduce all response values and another to introduce the values of the factor. In the specific example the factor is the percentage of hardwood in paper and the response is the tensile strength.

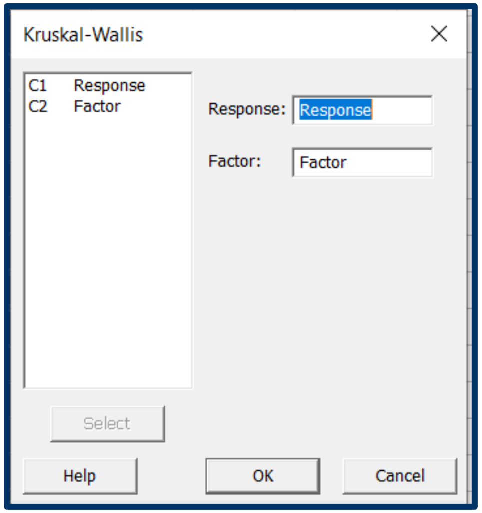
The Kruskal-Wallis test is accessed through the *Nonparametrics* section in the Stat menu.

Columns corresponding to the Response and to the Factor are then selected.



↓	C1	C2
	Response	Factor
1	7	5
2	8	5
3	15	5
4	11	5
5	9	5
6	10	5
7	12	10
8	17	10
9	13	10
10	18	10
11	19	10
12	15	10
13	14	15
14	18	15
15	19	15
16	17	15
17	16	15
18	18	15
19	19	20
20	25	20
21	22	20
22	23	20
23	18	20
24	20	20





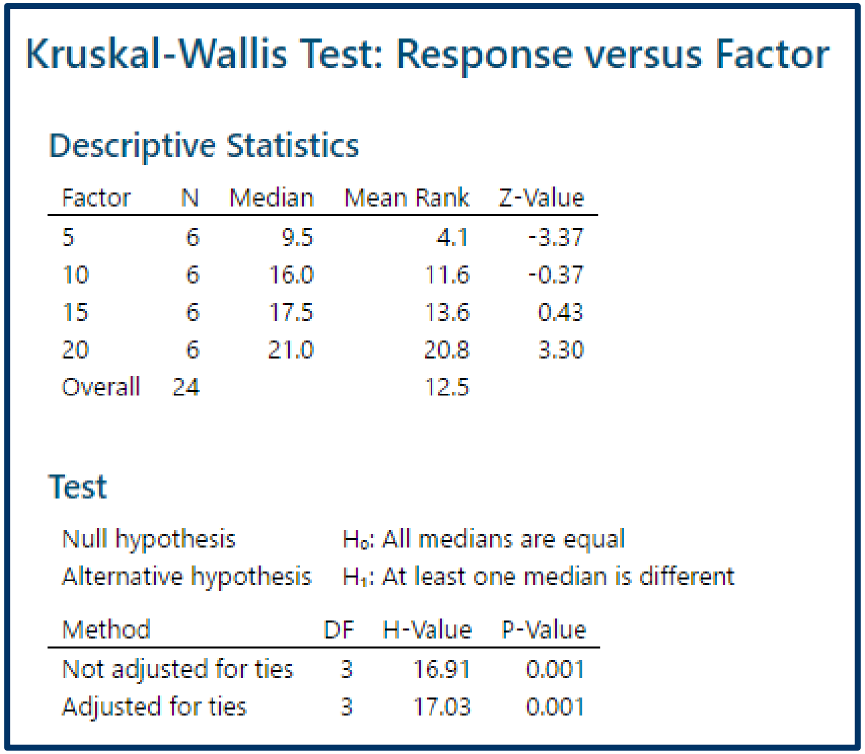
Response: Response

Factor: Factor

In the Session window the test results are summarized as a table, in which median and mean rank values are reported for each group.

In the specific example mean ranks are quite different from each other.

As a result, the H-value is quite high and the p-value, 0.001, very low, thus suggesting that a significant difference exists between some of the groups.



Kruskal-Wallis Test: Response versus Factor

Descriptive Statistics

Factor	N	Median	Mean Rank	Z-Value
5	6	9.5	4.1	-3.37
10	6	16.0	11.6	-0.37
15	6	17.5	13.6	0.43
20	6	21.0	20.8	3.30
Overall	24		12.5	

Test

Null hypothesis H_0: All medians are equal
Alternative hypothesis H_1: At least one median is different

Method	DF	H-Value	P-Value
Not adjusted for ties	3	16.91	0.001
Adjusted for ties	3	17.03	0.001

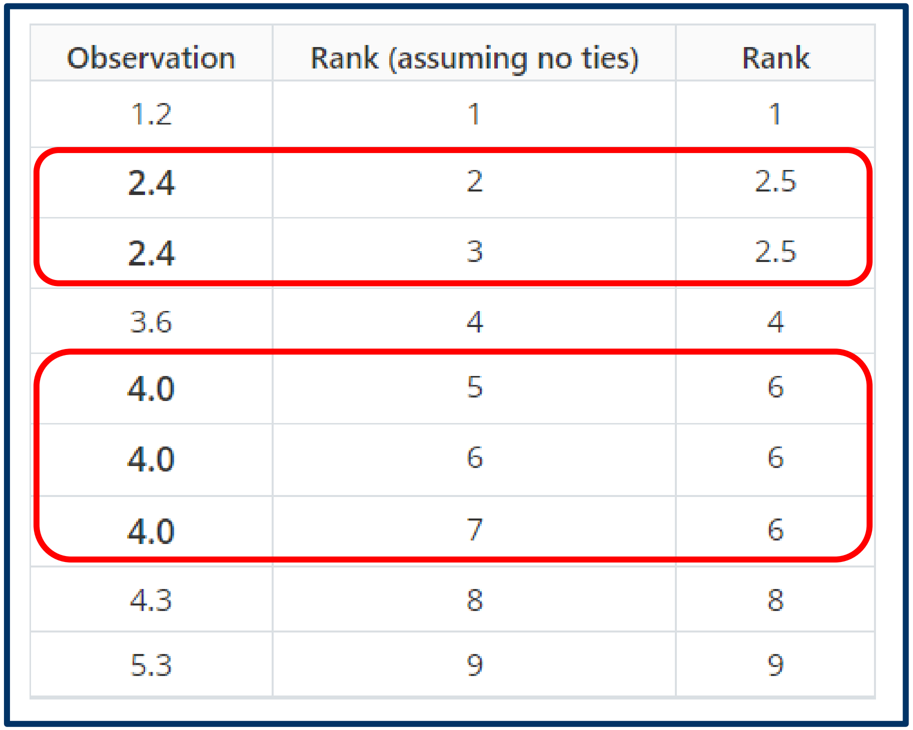
Interestingly, the software calculates also an H-value adjusted for ties, if they occur, using the following formula:

$$H(adj) = \frac{H}{1 - \frac{\Sigma(t_i^3 - t_i)}{N^3 - N}}$$

In this formula the sum refers to the number of groups of ties occurring in the data series and t_i represents the number of ties occurring in each group.

As an example, two groups of ties can be observed in the dataset shown on the right. The first group includes two data (thus $t_1 = 2$), whereas the second group includes three data ($t_2 = 3$).

Notably, the correction usually makes little difference in the value of H unless there are many ties.



Observation	Rank (assuming no ties)	Rank
1.2	1	1
2.4	2	2.5
2.4	3	2.5
3.6	4	4
4.0	5	6
4.0	6	6
4.0	7	6
4.3	8	8
5.3	9	9

Minitab also calculates a Z-value for each group, using the following formula:

$$z_j = \frac{\overline{R}_j - \overline{R}}{\sqrt{\frac{(N+1)(\frac{N}{n_j} - 1)}{12}}}$$

The z-value indicates how the average rank for each group compares to the average rank of all observations.

Dunn’s test

If the Kruskal-Wallis test indicates that not all medians are statistically comparable, the Dunn’s Test, proposed by the American mathematician and statistician Olivia Jean Dunn in 1964, can be employed to determine which medians are significantly different.

Dunn’s Test performs pairwise comparisons between independent groups and indicates which groups are significantly different at some level of significance α.

The formula to calculate the z statistic for the difference between two groups in the Dunn’s test is:

$z_i = y_i / \sigma_i$

where:

i is one of the 1 to *m* comparisons

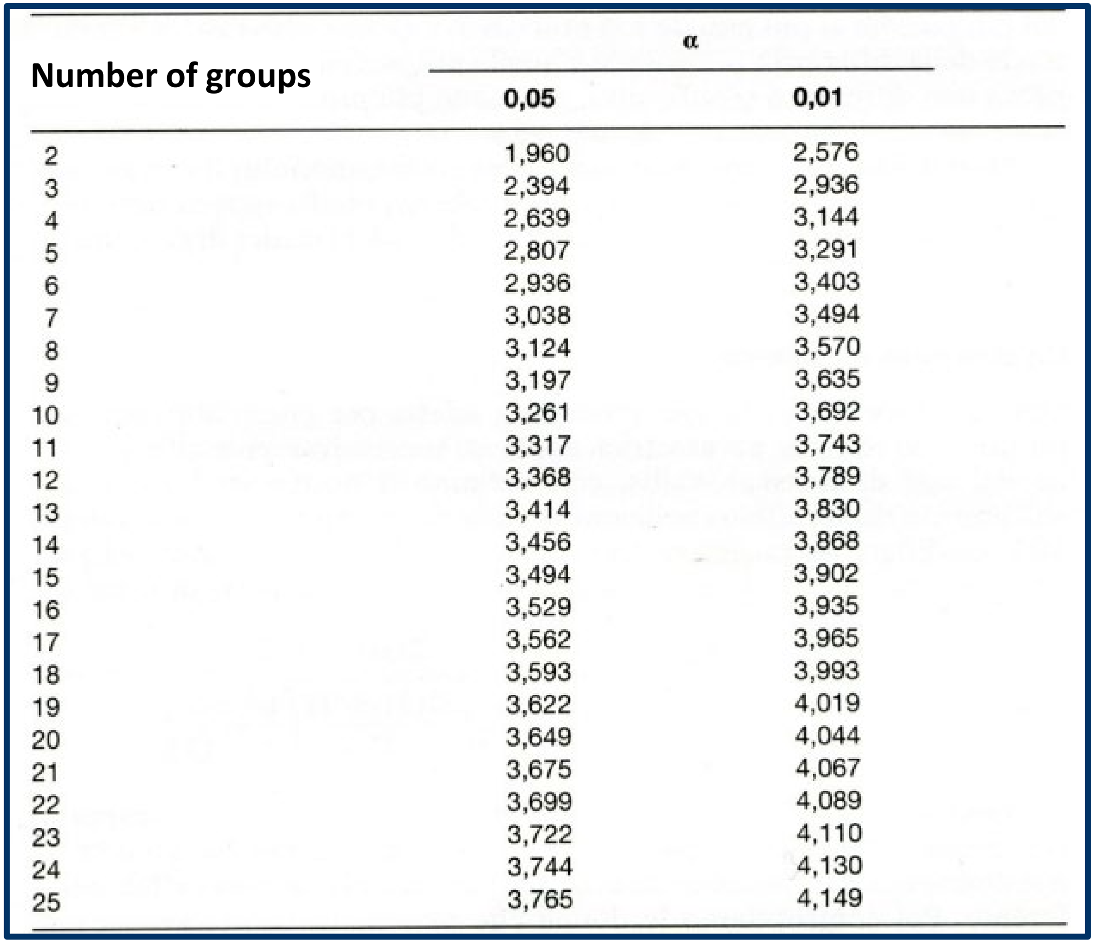
$y_i = \overline{r}_A - \overline{r}_B$ is the difference between the average ranks of groups A and B under test

and:

$$\sigma_i = \sqrt{\left\{\frac{N(N+1)}{12} - \frac{\sum_{s=1}^{r} \tau_s^3 - \tau_s}{12(N-1)}\right\}\left(\frac{1}{n_A} + \frac{1}{n_B}\right)}$$

where τ_s represents the number of ties in the sth group of ties (if no ties are present, the sum shown in the formula is equal to 0).

Each z_i value (often referred to also as Q value) has to be compared with a critical value depending on the significativity value α, as shown in the table on the right:



Number of groups	α	
	0,05	0,01
2	1,960	2,576
3	2,394	2,936
4	2,639	3,144
5	2,807	3,291
6	2,936	3,403
7	3,038	3,494
8	3,124	3,570
9	3,197	3,635
10	3,261	3,692
11	3,317	3,743
12	3,368	3,789
13	3,414	3,830
14	3,456	3,868
15	3,494	3,902
16	3,529	3,935
17	3,562	3,965
18	3,593	3,993
19	3,622	4,019
20	3,649	4,044
21	3,675	4,067
22	3,699	4,089
23	3,722	4,110
24	3,744	4,130
25	3,765	4,149

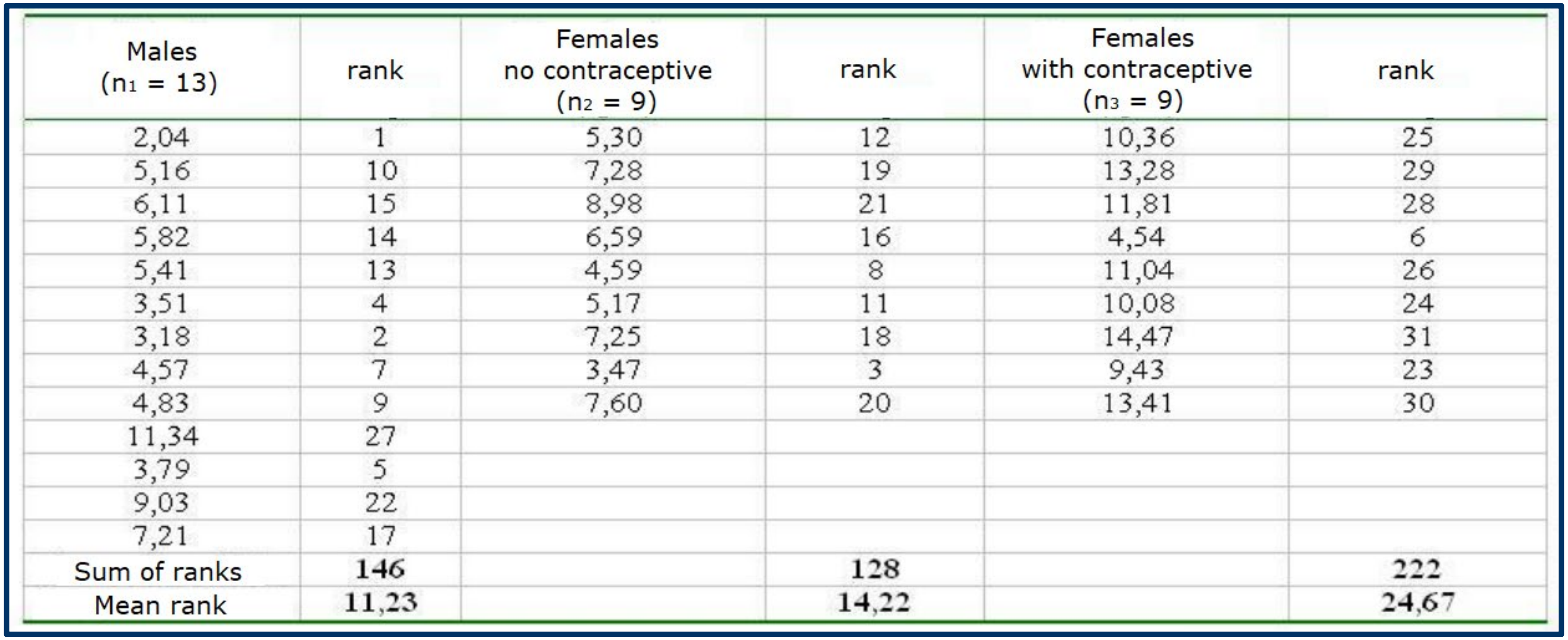
Notably, the critical values reported in the table arise from the standard normal distribution after making the Bonferroni correction, i.e., after dividing α by the number of pairwise comparisons performed with the Dunn's test.

For example, if 5 groups are considered, a total of 5 * 4 / 2 = 10 comparisons needs be made, thus a $\alpha/10$ value has to be used for each comparison. If $\alpha = 0.05$, then $\alpha/10 = 0.005$ and the critical value to be used is $z_{1-0.005/2} = z_{1-0.0025} = z_{0.9975} = 2.807$.

An example of Dunn test

The half life of caffeine, expressed in hours, was measured in the blood of individuals from three groups, namely, 13 men, 9 women not using contraceptives and 9 women using contraceptives, after the oral assumption of a tablet containing 250 mg of caffeine.

The following results were obtained:



Males (n_1 = 13)	rank	Females no contraceptive (n_2 = 9)	rank	Females with contraceptive (n_3 = 9)	rank
2,04	1	5,30	12	10,36	25
5,16	10	7,28	19	13,28	29
6,11	15	8,98	21	11,81	28
5,82	14	6,59	16	4,54	6
5,41	13	4,59	8	11,04	26
3,51	4	5,17	11	10,08	24
3,18	2	7,25	18	14,47	31
4,57	7	3,47	3	9,43	23
4,83	9	7,60	20	13,41	30
11,34	27				
3,79	5				
9,03	22				
7,21	17				
Sum of ranks	**146**		**128**		**222**
Mean rank	**11,23**		**14,22**		**24,67**

In this case the H value calculated for the Kruskal-Wallis test, 12.07, is higher than the critical value related to a χ^2 distribution with 2 degrees of freedom and α = 0.01, thus a significant difference exists between at least two of the three groups.

Since no ties were observed between data, the realizations of the Dunn statistic (Q) were calculated using the simplified equation for σ_i :

$$\sigma_i = \sqrt{\left\{\frac{N(N+1)}{12}\right\}\left(\frac{1}{n_A}+\frac{1}{n_B}\right)}$$

Consequently, the following values were obtained:

$$Q = \frac{\bar{R}_{contrac} - \bar{R}_{\text{males}}}{\sqrt{\frac{N(N+1)}{12}\left(\frac{1}{n_{contrac}}+\frac{1}{n_{\text{males}}}\right)}} = \frac{24{,}67 - 11{,}23}{\sqrt{\frac{31(31+1)}{12}\left(\frac{1}{9}+\frac{1}{13}\right)}} = 3{,}409$$

$$Q = \frac{\bar{R}_{contrac} - \bar{R}_{no\_contrac}}{\sqrt{\frac{N(N+1)}{12}\left(\frac{1}{n_{contrac}}+\frac{1}{n_{no\_contrac}}\right)}} = \frac{24{,}67 - 14{,}22}{\sqrt{\frac{31(31+1)}{12}\left(\frac{1}{9}+\frac{1}{9}\right)}} = 2{,}438$$

$$Q = \frac{\bar{R}_{no\_contrac} - \bar{R}_{\text{males}}}{\sqrt{\frac{N(N+1)}{12}\left(\frac{1}{n_{no\_contrac}}+\frac{1}{n_{\text{males}}}\right)}} = \frac{14{,}22 - 11{,}23}{\sqrt{\frac{31(31+1)}{12}\left(\frac{1}{9}+\frac{1}{13}\right)}} = 0{,}7583$$

Since the critical value for a = 0.05 and three groups is 2.394, the test shows that the half-like in blood of caffeine is longer in women assuming contraceptives than in women not assuming them and in men.