## 4.4. Further Analysis within ANOVA

*1) Estimation of the effects*

Fixed effects model:

$\alpha_i = \mu_i - \mu$ is estimated by $a_i = (\bar{x}_i - \bar{\bar{x}})$ if $H_0 \colon \mu_1 = \mu_2 = \cdots = \mu_k$ is rejected.

Random effects model:

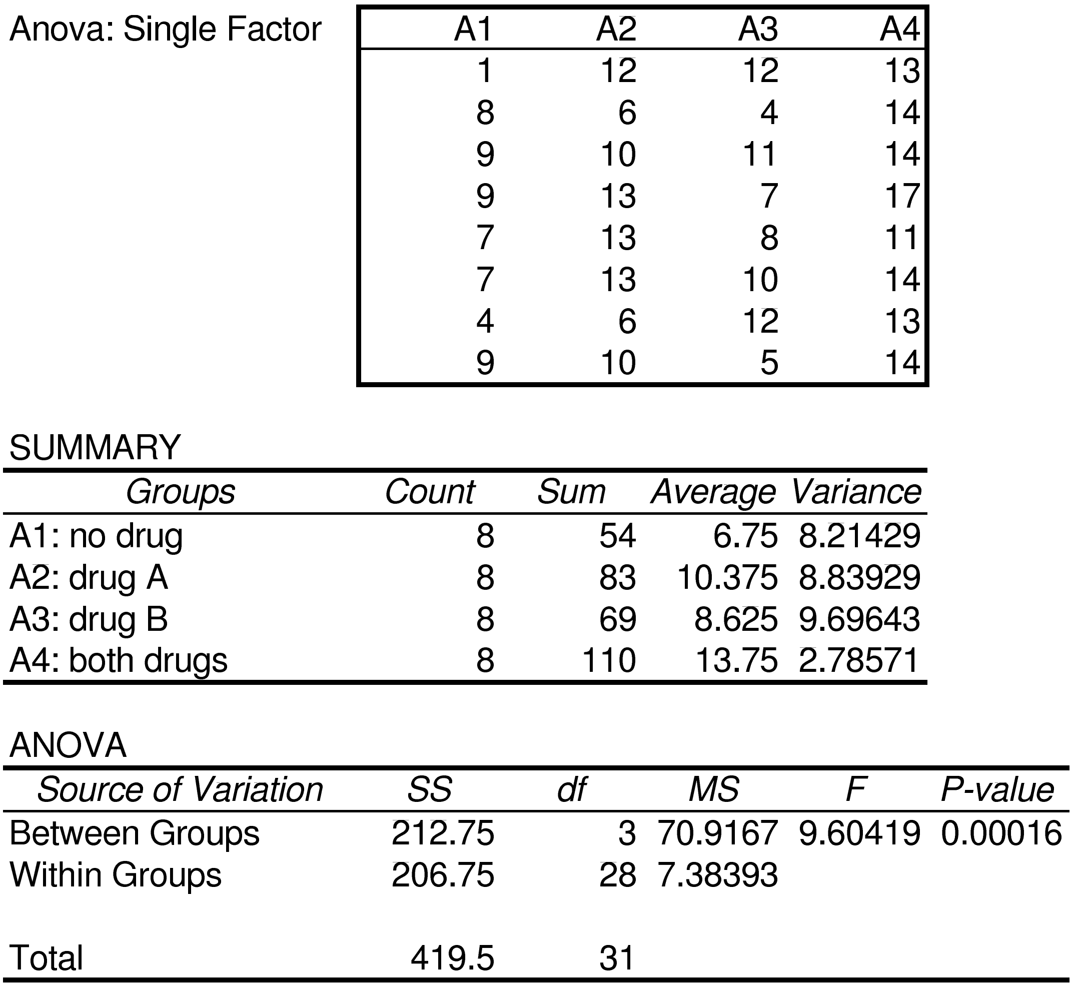
If $H_0 \colon \sigma_A^2 = 0$ is rejected, then we estimate the variability $\sigma_A^2$ among the population means by

$$s_A^2 = \frac{\text{MSB} - \text{MSW}}{n_0} \text{ with } n_0 = \frac{N^2 - \sum_{i=1}^{k} n_i^2}{N(k-1)},$$

where $N = n_1 + n_2 + \cdots + n_k$, with $n_0 = n$ in the special case of equal sampel sizes $n_i = n$ in all groups.

## 2) Planned Comparisons: Contrasts

The hypothetical data below shows errors in a test made by subjects under the influence of two drugs in 4 groups of 8 subjects each. Group A1 is a control group, not given any drug. The other groups are experimental groups, group A2 gets drug A, group A3 gets drug B, and group A4 gets both drugs.



Anova: Single Factor

| A1 | A2 | A3 | A4 |
|---|---|---|---|
| 1 | 12 | 12 | 13 |
| 8 | 6 | 4 | 14 |
| 9 | 10 | 11 | 14 |
| 9 | 13 | 7 | 17 |
| 7 | 13 | 8 | 11 |
| 7 | 13 | 10 | 14 |
| 4 | 6 | 12 | 13 |
| 9 | 10 | 5 | 14 |

SUMMARY

| *Groups* | *Count* | *Sum* | *Average* | *Variance* |
|---|---|---|---|---|
| A1: no drug | 8 | 54 | 6.75 | 8.21429 |
| A2: drug A | 8 | 83 | 10.375 | 8.83929 |
| A3: drug B | 8 | 69 | 8.625 | 9.69643 |
| A4: both drugs | 8 | 110 | 13.75 | 2.78571 |

ANOVA

| *Source of Variation* | *SS* | *df* | *MS* | *F* | *P-value* |
|---|---|---|---|---|---|
| Between Groups | 212.75 | 3 | 70.9167 | 9.60419 | 0.00016 |
| Within Groups | 206.75 | 28 | 7.38393 | | |
| | | | | | |
| Total | 419.5 | 31 | | | |

Suppose we are really interested in answering specific questions such as:

1. On the average do drugs have any effect on learning at all?
2. Do subjects make more errors if given both drugs than if given only one?
3. Do the two drugs differ in the number of errors they produce?

All these questions can be formulated as null hypotheses of the form

$$H_0\colon \lambda_1\mu_1 + \lambda_2\mu_2 + \cdots + \lambda_k\mu_k = 0,$$

where $\lambda_1 + \lambda_2 + \cdots + \lambda_k = 0$.

For example, the first question asks whether the mean of group A1, $\mu_1$, differes from the average of the means for the groups A2, A3, and A4, $(\mu_2 + \mu_3 + \mu_4)/3$. That is, we wish to test the null hypothesis

$$H_0(1) : \mu_1 - \frac{1}{3}\mu_2 - \frac{1}{3}\mu_3 - \frac{1}{3}\mu_4 = 0.$$

A contrast is a combination of population means of the form

$$\psi = \sum \lambda_i \mu_i, \text{ where } \sum \lambda_i = 0.$$

The corresponding sample contrast is

$$L = \sum \lambda_i \bar{x_i}.$$

In our example:

$$L_1 = \bar{X}_1 - \frac{1}{3}\bar{X}_2 - \frac{1}{3}\bar{X}_3 - \frac{1}{3}\bar{X}_4.$$

Now, since each individual observation $X_{ij}$ is distributed as $N(\mu_i, \sigma^2)$ and all observations are independent of each other, each sample mean $\bar{X}$ is distributed as $N(\mu_i, \sigma^2/n_i)$, such that

$$L \sim N\left(\sum \lambda_i \mu_i, \sigma^2 \sum \frac{\lambda_i^2}{n_i}\right),$$

which implies that under the null hypothesis

$$\frac{L}{\sigma\sqrt{\sum \frac{\lambda_i^2}{n_i}}} \sim N(0,1).$$

Now recalling that the square of a standard normally distributed random variable is always $\chi^2$-distributed with $df = 1$, we obtain denoting $Q = L^2/(\sum \lambda_i^2/n_i)$:

$$Q/\sigma^2 \sim \chi^2(1).$$

Furthermore, using SSW/$\sigma^2 \sim \chi^2(N-k)$, we obtain that:

$$Q/\text{MSW} \sim F(1, N-k),$$

the square root of which has a $t$-distribution, that is:

$$t = \frac{L}{s(L)} \sim t(N-k),$$

where

$$s(L) = s\sqrt{\sum \frac{\lambda_i^2}{n_i}} \quad \text{and} \quad s = \sqrt{\text{MSW}},$$

which simplifies to

$$s(L) = s\sqrt{\frac{\sum \lambda_i^2}{n}} \text{ for equal sample sizes } n.$$

$s(L)$ is called the <u>standard error of the contrast</u>.

This suggests to test the null hypothesis

$$H_0\colon \lambda_1\mu_1 + \lambda_2\mu_2 + \cdots + \lambda_k\mu_k = 0$$

with a conventional $t$-test of the form

$$t = \frac{L}{s(L)} \sim t(N-k).$$

Example: (continued.)

$$\begin{aligned} L &= \bar{X}_1 - \frac{1}{3}\bar{X}_2 - \frac{1}{3}\bar{X}_3 - \frac{1}{3}\bar{X}_4 \\ &= 6.75 - \frac{1}{3}(10.375 + 8.625 + 13.75) \\ &= -4.167, \\ s(L) &= \sqrt{\frac{MSW\sum\lambda_i^2}{n}} = \sqrt{\frac{7.38393\,(1+3/9)}{8}} \\ &= 1,109 \quad \text{such that} \\ & \quad t = \frac{-4.167}{1.109} = -3.76. \end{aligned}$$

The associated $p$-value in a two-sided test is T.DIST.2T(3.76;28)=0.08%, and in a one-sided test against $H_1\colon \mu_1 < (\mu_2 + \mu_3 + \mu_4)/3$, T.DIST.RT(3.76;28)=0.04%, which provides clear statistical evidence that the drugs considered increase error rates.

The remaining questions may be tackled in an analogous way:

1. Do subjects make more errors if given both drugs than if given only one?
   $H_0(2) : \mu_4 - \frac{1}{2}(\mu_2 + \mu_3) = 0$
2. Do the two drugs differ in the number of errors they produce?
   $H_0(3) : \mu_2 - \mu_3 = 0$

The computational steps in conducting the test are summarized in the tables below:

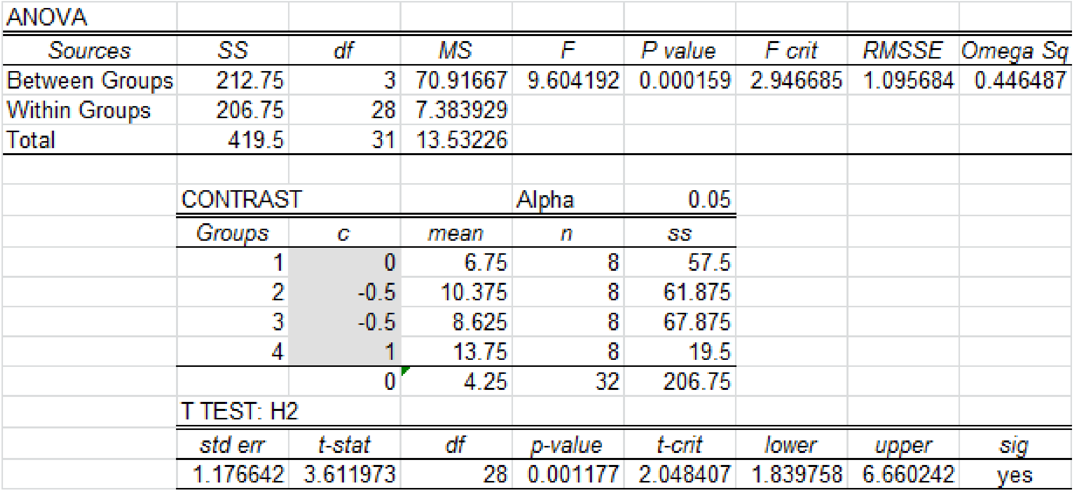
| | A1 | A2 | A3 | A4 | $\sum \lambda_i^2$ |
|---|---|---|---|---|---|
| $\bar{X}_i$ | 6.750 | 10.375 | 8.625 | 13.750 | |
| $\lambda_i(1)$ | 1 | -1/3 | -1/3 | -1/3 | 4/3 |
| $\lambda_i(2)$ | 0 | -1/2 | -1/2 | 1 | 3/2 |
| $\lambda_i(3)$ | 0 | 1 | -1 | 0 | 2 |

| | L | $s(L)$ | $t$ | $p$ |
|---|---|---|---|---|
| $H_0(1)$ | -4.167 | 1.109 | -3.76 | 0.0008 |
| $H_0(2)$ | 4.250 | 1.177 | 3.61 | 0.0012 |
| $H_0(3)$ | 1.750 | 1.359 | 1.29 | 0.208 |

## *Contrasts in Excel*

The Real Statistics Single Factor ANOVA tool has an option to calculate contrasts. After entering your contrast weights (the $\lambda$'s) into the grey shaded area labeled ‘c’ you will get the sample contrast $L$, its standard error $s(L)$, the corresponding $t$-statistics and its $p$-value.

Below is the contrast for testing $H_0(2)$ as an example:



ANOVA

| Sources | SS | df | MS | F | P value | F crit | RMSSE | Omega Sq |
|---|---|---|---|---|---|---|---|---|
| Between Groups | 212.75 | 3 | 70.91667 | 9.604192 | 0.000159 | 2.946685 | 1.095684 | 0.446487 |
| Within Groups | 206.75 | 28 | 7.383929 | | | | | |
| Total | 419.5 | 31 | 13.53226 | | | | | |

CONTRAST Alpha 0.05

| Groups | c | mean | n | ss |
|---|---|---|---|---|
| 1 | 0 | 6.75 | 8 | 57.5 |
| 2 | -0.5 | 10.375 | 8 | 61.875 |
| 3 | -0.5 | 8.625 | 8 | 67.875 |
| 4 | 1 | 13.75 | 8 | 19.5 |
| | 0 | 4.25 | 32 | 206.75 |

T TEST: H2

| std err | t-stat | df | p-value | t-crit | lower | upper | sig |
|---|---|---|---|---|---|---|---|
| 1.176642 | 3.611973 | 28 | 0.001177 | 2.048407 | 1.839758 | 6.660242 | yes |

*Orthogonal hypotheses*

The 3 hypotheses in the preceding section were special in the sense that the truth of each null hypothesis is unrelated to the truth of any others. If $H_0(1)$ is false, we know that drugs have some effect upon errors, although $H_0(1)$ says nothing about whether the effect of taking both drugs simultaneously is different from the average of the effects of the drugs taken seperately. Similarly, if $H_0(2)$ is false, we know that the effect of taking both drugs is different from taking only one, but we don't know which of the two drugs taken individually is more effective.

A set of contrasts such that the truth of any one of them is unrelated to the truth of any other is called orthogonal. For equal sample sizes in each group the orthogonality of two contrasts $L_1 = \sum \lambda_{1i} \bar{x_i}$ and $L_2 = \sum \lambda_{2i} \bar{x_i}$ may be assessed by checking the orthogonality condition

$$\sum \lambda_{1i} \lambda_{2i} = 0.$$

If the group sizes $n_i$ involved in the contrast are not all identical, the orthogonality condition becomes

$$\sum \frac{\lambda_{1i}\lambda_{2i}}{n_i} = 0.$$

In an ANOVA with $k$ groups, no more than $k - 1$ orthogonal contrasts can be tested. Such a set of $k - 1$ orthogonal contrasts is however not unique.

In practice the researcher will select a set of orthogonal contrasts such that those contrasts of particular interest in the research question are included. Once such a set is found, it exploits all available information that can be extracted for answering the maximum amount of $k - 1$ independent questions, that can be asked in form of contrasts.

To illustrate the importance of orthogonal hypothesis, assume that instead of testing the orthogonal hypotheses $H_0(1) - H_0(3)$ we would have tested the original hypothesis

$$H_0(1) : \mu_1 - \frac{1}{3}\mu_2 - \frac{1}{3}\mu_3 - \frac{1}{3}\mu_4 = 0$$

together with

$$H_0(4) : \mu_4 - \mu_1 = 0,$$

that is that error rates are the same when taking both drugs or taking no drugs, upon the data below:

Anova: Single Factor

| A1 | A2 | A3 | A4 |
|---|---|---|---|
| 5 | 0 | 10 | 4 |
| 4 | 8 | 3 | 5 |
| 4 | 3 | 5 | 9 |
| 5 | 9 | 8 | 10 |
| 8 | 7 | 4 | 10 |
| 7 | 5 | 2 | 7 |
| 9 | 7 | 5 | 13 |
| 2 | 7 | 9 | 14 |

SUMMARY

| *Groups* | *Count* | *Sum* | *Average* | *Variance* |
|---|---|---|---|---|
| A1: no drug | 8 | 44 | 5.50 | 5.429 |
| A2: drug A | 8 | 46 | 5.75 | 8.786 |
| A3: drug B | 8 | 46 | 5.75 | 8.500 |
| A4: both drugs | 8 | 72 | 9.00 | 12.571 |

ANOVA

| *Source of Variation* | *SS* | *df* | *MS* | *F* | *P-value* |
|---|---|---|---|---|---|
| Between Groups | 67 | 3 | 22.333 | 2.53171 | 0.07731 |
| Within Groups | 247 | 28 | 8.821 | | |
| | | | | | |
| Total | 314 | 31 | | | |

The two hypotheses are not orthogonal:

| | A1 | A2 | A3 | A4 |
|---|---|---|---|---|
| $\bar{X}_i$ | 5.50 | 5.75 | 5.75 | 9.00 |
| $\lambda_i(1)$ | 1 | -1/3 | -1/3 | -1/3 |
| $\lambda_i(4)$ | -1 | 0 | 0 | 1 |

We may therefore get conflicting results from both hypothesis tests:

| | L | $s(L)$ | $t$ | $p$ |
|---|---|---|---|---|
| $H_0(1)$ | -1.33 | 1.21 | -1.10 | 0.281 |
| $H_0(4)$ | 3.50 | 1.49 | 2.36 | 0.026 |

In this example, we safely accept $H_0(1)$ that taking drugs has no impact on error rates, while we strongly reject $H_0(4)$ that taking both drugs has no impact.

*Constructing Orthogonal Tests*

If all sample sizes are equal ($n_i = n$=const.), orthogonal tests may be constructed based on the principle that the test on the differences among a given set of means is orthogonal to any test involving their average.

Consider our original set of hypotheses:

$$\begin{aligned} H_0(1) &: \mu_1 - \frac{1}{3}(\mu_2 + \mu_3 + \mu_4) = 0 \\ H_0(2) &: \mu_4 - \frac{1}{2}(\mu_2 + \mu_3) = 0 \\ H_0(3) &: \mu_2 - \mu_3 = 0 \end{aligned}$$

$H_0(1)$ involves the average of $\mu_2$, $\mu_3$, and $\mu_4$, while $H_0(2)$ and $H_0(3)$ involve differences among them. Similarly, $H_0(2)$ involves the average of $\mu_2$ and $\mu_3$, while $H_0(3)$ involves the difference between them.

## *A Word of Caution*

Contrasts are a special form of planned comparisons, that is, the hypotheses with the contrasts to be tested must be set up before taking a look at the data, otherwise the associated $p$-values will not be valid.

The situation is similar to whether applying a one-sided or a two-sided test. Assume you want to test whether the means in two samples are the same using the conventional significance level of $\alpha = 5\%$. You apply a two-sided test and get a $p$-value of 8%, so you may not reject. Let's say as a step in your calculations you figured out that $\bar{x}_1 > \bar{x}_2$. You may be tempted then to replace your original two-sided test against $H_1 \colon \mu_1 \neq \mu_2$ by a one-sided test against $H_1 \colon \mu_1 > \mu_2$, which, technically, would allow you to divide your $p$-value by 2 and get a significant result at 4%. But that $p$-value of 4% is fraud, because the reason that it is only one half of the two sided $p$-value is exactly that without looking at the data, you also had a 50% chance that the sample means would have come out as $\bar{x}_1 < \bar{x}_2$.

*3) Multiple Comparisons (Post hoc Tests)*

As seen above, contrasts must be formulated in advance of the analysis in order for the $p$-values to be valid. Multiple-comparisons procedures, on the other hand, are designed for testing effects suggested by the data after the ANOVA $F$-test led to a rejection of the hypothesis that all sample means are equal. For that reason they are also called post hoc tests.

All multiple comparisons methods we will discuss consist of building $t$-ratios of the form

$$t_{ij} = \frac{\bar{x}_i - \bar{x}_j}{SE_{ij}}, \qquad \text{or, equivalently,}$$

setting up confidence intervals of the form

$$[(\bar{x}_i - \bar{x}_j) \pm t^* SE_{ij}]$$

for all $\binom{k}{2} = \frac{k(k-1)}{2}$ pairs that can be built within the $k$ groups.

The methods differ in the calculation of the standard errors $SE_{ij}$ and the distributions and critical levels used in the determination of $t^*$.

*Fisher's LSD = least-significant difference*

One obvious choice is to extend the 2 independent sample $t$-test to $k$ independent samples with test statistics

$$t_{ij} = \frac{\bar{x}_i - \bar{x}_j}{s\sqrt{\frac{1}{n_i} + \frac{1}{n_j}}}, \text{ where } s = \sqrt{\text{MSW}},$$

and to declare $\mu_i$ and $\mu_j$ different whenever $|t_{ij}| > t_{\alpha/2}(\text{DFW})$, or, equivalently, whenever

$$0 \notin \left[(\bar{x}_i - \bar{x}_j) \pm t_{\alpha/2}(\text{DFW}) \cdot s\sqrt{\frac{1}{n_i} + \frac{1}{n_j}}\right],$$

where DFW$=$ $N-k$ in the case of one-way ANOVA. This procedure fixes the probability of a false rejection <u>for each single pair of means</u> being compared as $\alpha$.

This is a problem if the numbers of means being compared is large. For example, if we use LSD with a significance of $\alpha$=5% to compare $k$=20 means, then there are $\frac{k(k-1)}{2} = 190$ pairs of means and we expect $5\% \cdot 190 = 9.5$ false rejections!

*Bonferroni Method*

The Bonferroni method uses the same test statistic as Fisher's LSD, but replaces the significance level $\alpha$ for each single pair with $\alpha' = \alpha/\binom{k}{2}$ or equivalently the original $p$-value with $p' = \binom{k}{2}p$ as a conservative estimate of the probability that any false rejection among all $\frac{k(k-1)}{2}$ comparisons will occur. This is also called the experimentwise error rate. An obvious disadvantage is that the test becomes weak when $k$ is large.

Example:
Comparison of A1 and A2 (original data):

$$|t_{ij}| = \left| \frac{\bar{x}_i - \bar{x}_j}{s\sqrt{\frac{1}{n_i} + \frac{1}{n_j}}} \right| = \left| \frac{6.75 - 10.375}{\sqrt{7.38393 \cdot \frac{2}{8}}} \right| = 2.668$$

LSD:

$$p_{\text{LSD}} = \text{T.DIST.2T}(2.668; 28) = 1.25\%.$$

Bonferroni:

$$p_B = \frac{4 \cdot 3}{2} \cdot p_{\text{LSD}} = 6 \cdot 1.25\% = 7.5\%.$$

*Tukey's HSD = honestly significant difference*

Tukey's method for multiple comparisons delivers the most precise estimate of the probability that <u>any false rejection</u> among all paired comparisons will occur. In its original form it is only applicable in the case of equal sample sizes in each group ($n_i = n$), but corrections for unequal sample sizes have been suggested and are implemented in Real Statistics.

<u>Tukey's honestly significant difference</u> is:

$$HSD = q_\alpha(k, \mathsf{DFW}) \cdot \frac{s}{\sqrt{n}},$$

where $q_\alpha$ denotes the $\alpha$-critical value from the so called *Studentized range distribution* tabulated e.g. in table 6 of Aczel, and $s$ is estimated by $\sqrt{\mathsf{MSW}}$, as usual.

Two group means $\mu_i$ and $\mu_j$ are declared different at significance level $\alpha$ if

$$0 \notin \left[(\bar{x}_i - \bar{x}_j) \pm HSD\right],$$

or, equivalently, if

$$\left|\frac{\bar{x}_i - \bar{x}_j}{s/\sqrt{n}}\right| > q_\alpha(k, \mathsf{DFW}).$$

Example:
Comparison of A1 and A2 (continued).

LSD and Bonferroni gave conflicting results whether the difference between $\mu_1$ and $\mu_2$ is significant at 5% or not (recall $p_{\text{LSD}} = 1.25\%$ and $p_B = 7.5\%$), because LSD underestimates the probability of any false rejection among all paired comparisons, whereas Bonferroni overestimates it.

Applying Tukey's procedure yields

$$\left|\frac{\bar{x}_i - \bar{x}_j}{s/\sqrt{n}}\right| = \left|\frac{6.75 - 10.375}{\sqrt{7.38393/8}}\right| = 3.77.$$

Looking up in a table or using the QCRIT function from Real Statistics reveals that

$$q_{0.05}(4; 28) = \text{QCRIT(4;28;0.05;2)} = 3.86,$$

which is larger than the statistic calculated above. The difference between $\mu_1$ and $\mu_2$ is therefore not significant at 5% level, if the test was first suggested by the data.

Note: A planned comparison before looking at the data in form of a contrast would have found a significant difference with a $p$-value of 1.25%(=$p_{\text{LSD}}$).

*Simultaneous Confidence Intervals*

The idea with multiple comparisons in post-hoc tests is usually to determine, which groups may be combined into larger groups, that are homogeneous in the sense that their group means are not significantly different.

For that purpose, one constructs confidence intervals of the form

$$[(\bar{x}_i - \bar{x}_j) \pm \text{smallest significant difference}],$$

where the smallest significant differences are:

$$LSD\colon\ t_{\frac{\alpha}{2}}(\text{DFW}) \cdot s\sqrt{\frac{1}{n_i} + \frac{1}{n_j}}$$

$$\text{Bonferroni:}\ t_{\frac{\alpha}{2}/\binom{k}{2}}(\text{DFW}) \cdot s\sqrt{\frac{1}{n_i} + \frac{1}{n_j}}$$

$$HSD\colon\ q_\alpha(k, \text{DFW}) \cdot \frac{s}{\sqrt{n}}$$

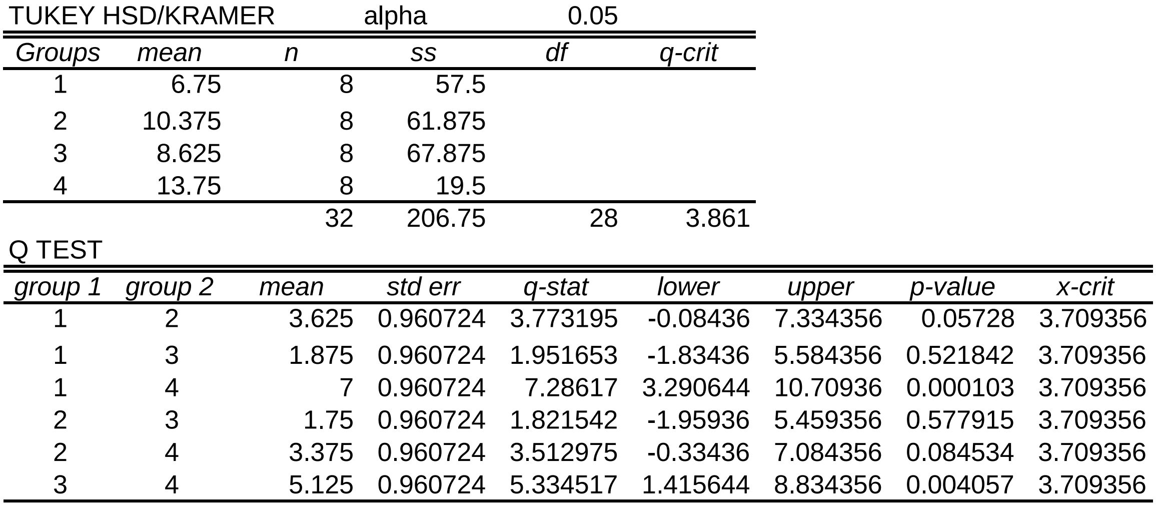
Combinations of groups are considered homogeneous at confidence level $(1-\alpha)$, if none of their paired comparisons lies outside the corresponding confidence intervals; that is, pairs of means, the confidence intervals of which include the value 0 will not be declared significantly different, and vice versa.

*Multiple Comparisons in Excel*

The Real Statistics toolpack implements post hoc tests as contrasts restricted to have all weights coefficients $\lambda_{1,\dots,k} = 0$, except for the two groups $i$ and $j$ which are currently compared with weights $\lambda_i = 1$ and $\lambda_j = -1$. Choosing ‘Contrasts’ with ‘No correction’ under ‘Alpha correction for contrasts’ within the Single Factor ANOVA tool corresponds then to Fisher’s LSD.

Note that Real Statistics implements the ‘Bonferroni correction’ under ‘Alpha correction for contrasts’ as dividing $\alpha$ by the maximal number of orthogonal contrasts $k-1$ rather than the number of all possible comparisons $\binom{k}{2}$. Generally the Bonferroni correction should not be used, because it overestimates the experimentwise error rate, making the test too conservative.

The best way to do multiple comparisons is Tukey's HSD, which is implemented as its own option in the Single Factor ANOVA tool.



| TUKEY HSD/KRAMER | | | alpha | 0.05 | |
|---|---|---|---|---|---|
| *Groups* | *mean* | *n* | *ss* | *df* | *q-crit* |
| 1 | 6.75 | 8 | 57.5 | | |
| 2 | 10.375 | 8 | 61.875 | | |
| 3 | 8.625 | 8 | 67.875 | | |
| 4 | 13.75 | 8 | 19.5 | | |
| | | 32 | 206.75 | 28 | 3.861 |

Q TEST

| *group 1* | *group 2* | *mean* | *std err* | *q-stat* | *lower* | *upper* | *p-value* | *x-crit* |
|---|---|---|---|---|---|---|---|---|
| 1 | 2 | 3.625 | 0.960724 | 3.773195 | -0.08436 | 7.334356 | 0.05728 | 3.709356 |
| 1 | 3 | 1.875 | 0.960724 | 1.951653 | -1.83436 | 5.584356 | 0.521842 | 3.709356 |
| 1 | 4 | 7 | 0.960724 | 7.28617 | 3.290644 | 10.70936 | 0.000103 | 3.709356 |
| 2 | 3 | 1.75 | 0.960724 | 1.821542 | -1.95936 | 5.459356 | 0.577915 | 3.709356 |
| 2 | 4 | 3.375 | 0.960724 | 3.512975 | -0.33436 | 7.084356 | 0.084534 | 3.709356 |
| 3 | 4 | 5.125 | 0.960724 | 5.334517 | 1.415644 | 8.834356 | 0.004057 | 3.709356 |

For unequal group sizes $n_i$ and $n_j$, Real Statistics implements Tukey's honestly significant difference as

$$HSD = q_\alpha(k, \text{DFW}) \cdot \sqrt{\frac{\text{MSW}}{2}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}.$$