Multiple comparisons

When we carry out an ANOVA on k treatments, we test

$$H_0 : \mu_1 = \cdots = \mu_k \quad \text{versus} \quad H_a : H_0 \text{ is false}$$

Assume we reject the null hypothesis, i.e. we have some evidence that not all treatment means are equal. Then we could for example be interested in which ones are the same, and which ones differ. For this, we might have to carry out some more hypothesis tests.

This procedure is referred to as multiple comparisons.

Key issue

We will be conducting, say, **T** different tests, and we become concerned about the overall error rate (sometimes called the "family-wise" error rate).

Overall error rate = $\Pr(\text{reject at least one } H_0 \mid \text{all } H_0 \text{ are true})$

$$\begin{cases} = 1 - [1 - \Pr(\text{reject first} \mid \text{first } H_0 \text{ is true})]^T & \text{if independent} \\ \leq T \times \Pr(\text{reject first} \mid \text{first } H_0 \text{ is true}) & \text{generally} \end{cases}$$

Types of multiple comparisons

There are two different types of multiple comparisons procedures:

Sometimes we already know in advance what questions we want to answer. Those comparisons are called planned (or a priori) comparisons.

Sometimes we do not know in advance what questions we want to answer, and the judgement about which group means will be studied the same depends on the ANOVA outcome. Those comparisons are called unplanned (or a posteriori) comparisons.

The distinction

Planned comparisons:

adjust for just those tests that are planned.

Unplanned comparisons:

adjust for all possible comparisons.

Former example

We previously investigated whether the mean blood coagulation times for animals receiving different diets (A, B, C or D) were the same.

Imagine A is the standard diet, and we wish to compare each of diets B, C, D to diet A.

$\rightarrow$ planned comparisons!

After inspecting the treatment means, we find that A and D look similar, and B and C look similar, but A and D are quite different from B and C. We might want to formally test the hypothesis $\mu_A = \mu_D \neq \mu_B = \mu_C$.

$\rightarrow$ unplanned comparisons!

Another example

A plant physiologist recorded the length of pea sections grown in tissue culture with auxin present. The purpose of the experiment was to investigate the effects of various sugars on growth. Four different treatments were used, plus one control (no sugar):

- No sugar
- 2% glucose
- 2% fructose
- 1% glucose + 1% fructose
- 2% sucrose

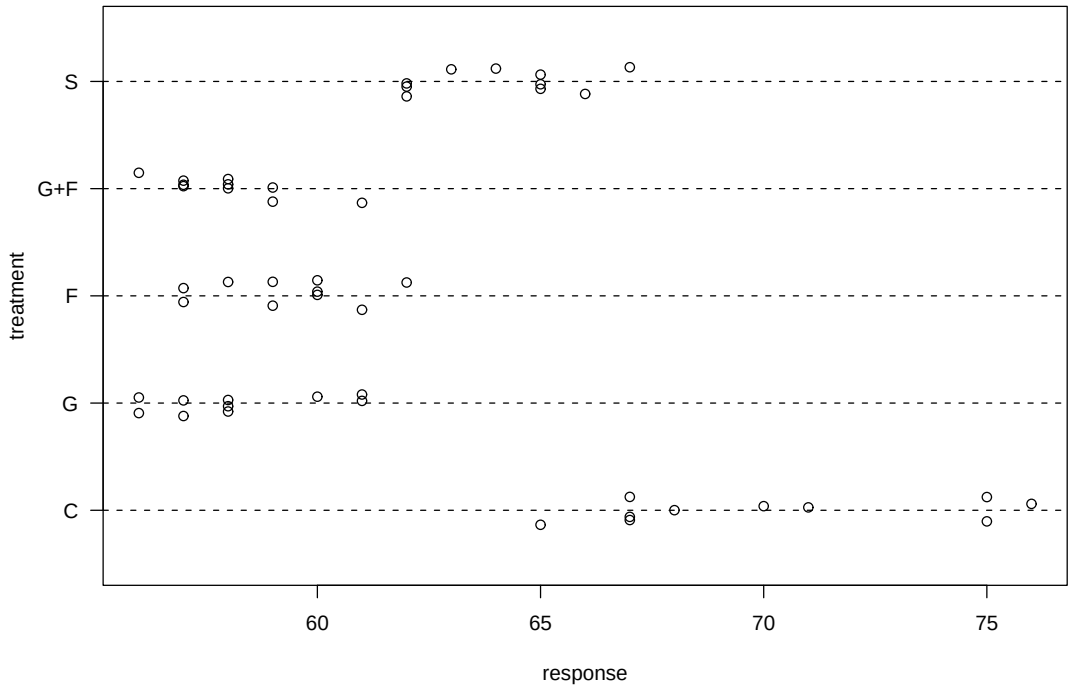
Specific questions

The investigator wants to answer three specific questions:

- Does the addition of sugars have an effect on the lengths of the pea sections?
- Are there differences between the pure sugar treatments and the mixed sugar treatment?
- Are there differences among the pure sugar treatments?

Planned comparisons!

The plant physiologist's data



ANOVA table

Source	SS	Df	MS	F-value	p-value
Between treatment	1077.3	4	269.3	49.4	< 0.001
Within treatment	245.5	45	5.5		

The first comparison

Compare the control to the others.

Test

$$\mu_C - \frac{\mu_G + \mu_F + \mu_{G+F} + \mu_S}{4} = 0$$

Look at

$$\bar{y}_C - \frac{\bar{y}_G + \bar{y}_F + \bar{y}_{G+F} + \bar{y}_S}{4} = 0$$

Contrasts

$\sum_i c_i \bar{y}_i$ where $\sum_i c_i = 0$

If the assumptions for the ANOVA model are correct,

$$E(\sum_i c_i \bar{y}_i) = \sum_i c_i \mu_i$$

$$\text{var}(\sum_i c_i \bar{y}_i) = \sum_i c_i^2 \text{var}(\bar{y}_i) = \sigma^2 \sum_i c_i^2 / n_i$$

$\sum_i c_i \bar{y}_i$ is normally distributed

Tests of contrasts

- Use $\sqrt{\text{MSE}}$ from the ANOVA as the estimate of σ
- Use N $-$ k degrees of freedom (concerns the precision of the estimate of σ)

- SS = $(\sum_i c_i \bar{y}_i)^2 \,/\, (\sum_i c_i^2 / n_i)$

ANOVA table

Source	SS	Df	MS	F-value	p-value
Between treatment	1077.3	4	269.3	49.4	< 0.001
Within treatment	245.5	45	5.5		

Control versus sugars

Source	SS	Df	MS	F-value	p-value
Between treatment	832.3	1	832.3	152.4	< 0.001
Within treatment	245.5	45	5.5		

Pure sugars versus mixed sugar

Source	SS	Df	MS	F-value	p-value
Between treatment	48.1	1	48.1	8.82	0.005
Within treatment	245.5	45	5.5		

Among pure sugars

Source	SS	Df	MS	F-value	p-value
Between treatment	196.9	2	98.4	18.0	< 0.001
Within treatment	245.5	45	5.5		

Orthogonal comparisons

Two comparisons $c_1 = (c_{1,1}, \ldots, c_{1,k})$ and $c_2 = (c_{2,1}, \ldots, c_{2,k})$ are orthogonal if and only if

$$\sum_{t=1}^{k} n_t c_{t,1} c_{t,2} = 0$$

where n_t is the number of observations in treatment group t.

Summary

source	SS	df	MS	F-value
Treatment	1077.3	4	269.3	49.4
Control versus sugars	832.3	1	832.3	152.4
Pure sugars versus mixed sugar	48.1	1	48.1	8.82
Among pure sugars	196.9	2	98.4	18.0
Within treatment	245.5	45	5.5	
Total	1322.8	49		

Adjusting the significance level

Assume the investigator plans to make T independent significance tests, all at the significance level α'. If all the null hypothesis are true, the probability of making no Type I error is $(1 - \alpha')^T$. Hence the overall significance level is

$$\alpha = 1 - (1 - \alpha')^T$$

Solving the above equation for α' yields

$$\alpha' = 1 - (1 - \alpha)^{\frac{1}{T}}$$

The above adjustment is called the Dunn – Sidak method.

An alternative method

In the literature, investigators often use

$$\alpha'' = \frac{\alpha}{T}$$

where T is the number of planned comparisons.

This adjustment is called the Bonferroni method.

"Unplanned" comparisons

Suppose we are comparing k treatment groups.

Suppose ANOVA indicates that you reject $H_0 : \mu_1 = \cdots = \mu_k$

What next?

Which of the μ's are different from which others?

Consider testing $H_0 : \mu_i = \mu_j$ for all pairs i,j.

There are $\binom{k}{2} = \frac{k\,(k-1)}{2}$ such pairs.

k = 5 $\longrightarrow$ $\binom{k}{2} = 10$.

Bonferroni correction

Suppose we have 10 treatment groups, and so 45 pairs.

If we perform 45 t-tests at the significance level α = 0.05, we'd expect to reject 5% $\times$ 45 $\approx$ 2 of them, even if all of the means were the same.

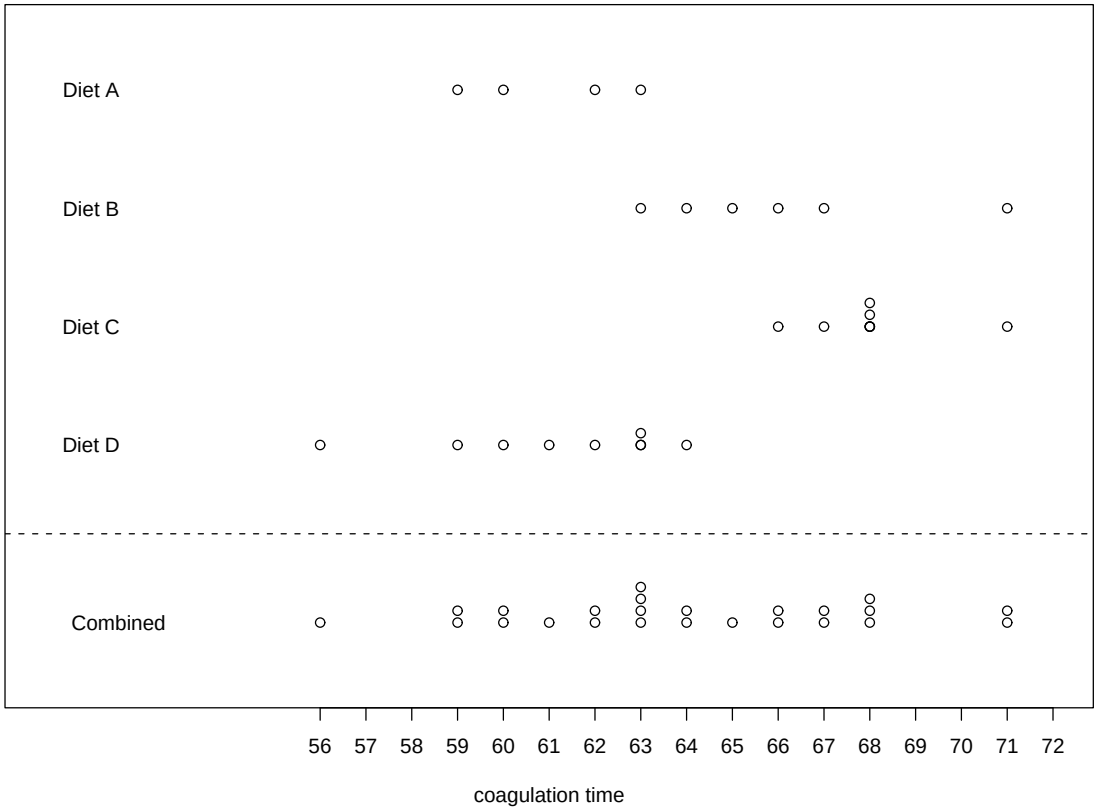
Let α = Pr(reject at least one pairwise test | all μ's the same)

$\leq$ (no. tests) $\times$ Pr(reject test #1 | μ's the same)

The Bonferroni correction:

Use $\alpha' = \alpha/(\text{no. tests})$ as the significance level for each test.

For example, with 10 groups and so 45 pairwise tests, we'd use α' = 0.05 / 45 $\approx$ 0.0011 for each test.

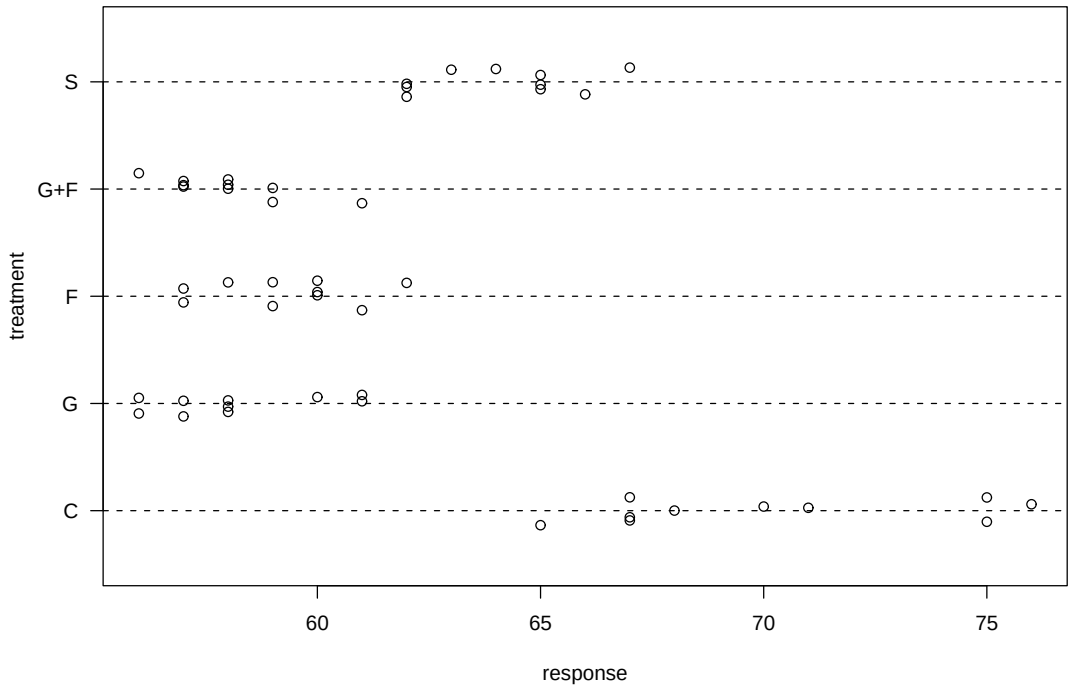


Pairwise comparisons

Comparison	p-value
A vs B	0.004
A vs C	< 0.001
A vs D	1.000
B vs C	0.159
B vs D	< 0.001
C vs D	< 0.001

$$\alpha'' = \frac{\alpha}{k} = \frac{0.05}{6} = 0.0083$$

The other example



ANOVA table

Source	SS	Df	MS	F-value	p-value
Between treatment	1077.3	4	269.3	49.4	< 0.001
Within treatment	245.5	45	5.5		

$\binom{5}{2}$ = 10 pairwise comparisons $\longrightarrow$ α' = 0.05/10 = 0.005

For each pair, consider $T_{i,j} = (\bar{Y}_{i\cdot} - \bar{Y}_{j\cdot}) \,/\, \left(\hat{\sigma}\sqrt{\frac{1}{n_i} + \frac{1}{n_j}}\right)$

Use $\hat{\sigma} = \sqrt{M_W}$ (M_W = within-group mean square)
and refer to a t distribution with df = 45.

Results

$\hat{\sigma}$ = 2.34

n = 10 for each group

SE = $2.34 \times \sqrt{2/10}$ = 1.05 for each comparison.

df = 45, α' = 0.005 $\longrightarrow$ t = 2.69

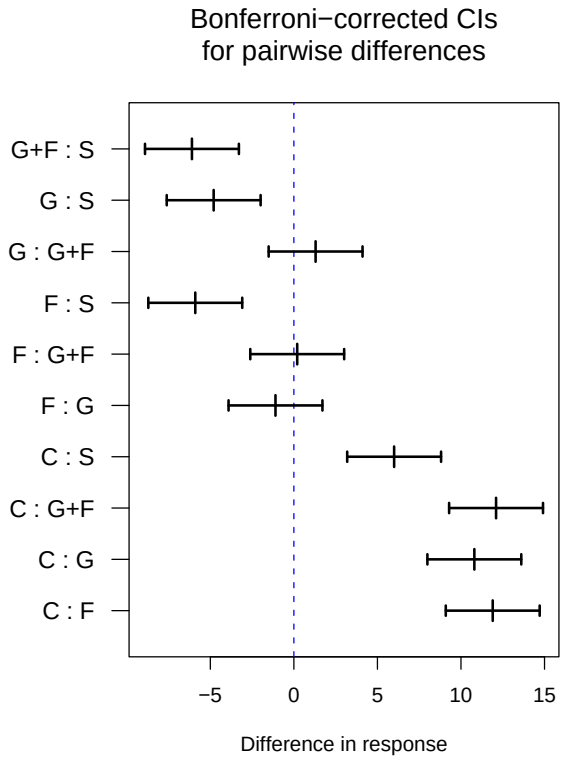
Groups with

$$|\bar{Y}_{i\cdot} - \bar{Y}_{j\cdot}| > 2.69 \times 1.05 = 2.81$$

are deemed different.

Bonferroni-corrected CIs:

$$(\bar{Y}_{i\cdot} - \bar{Y}_{j\cdot}) \pm 2.81$$



Tukey's HSD

HSD = "Honest significant difference"

Reject $H_0 : \mu_i = \mu_j$ if

$$|\bar{Y}_{i\cdot} - \bar{Y}_{j\cdot}| > Q_\alpha(k, df) \times \sqrt{M_W/n}$$

We're assuming equal sample sizes (n) for the treatment groups.

k = no. treatment groups; df = n · k – k

$Q_\alpha(k, df)$ = 1 – α quantile of the "Studentized range distribution."

We won't go into where $Q_\alpha(k, df)$ comes from. Suffice it to say: it's an adjustment not unlike the Bonferroni correction, and it can be calculated using `qtukey()` in R. Alternatively, the function `TukeyHSD()` will do the whole thing.

Results

Taking α = 0.05, k = 5, df = 45,

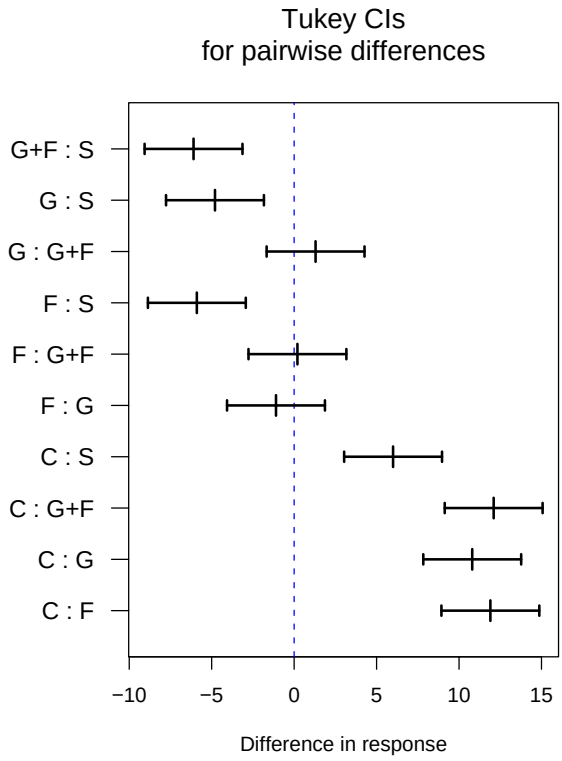
$Q_\alpha(k, df)$ = 4.02.

```
qtukey(0.95, 5, 45)
```

Groups with

$$|\bar{Y}_{i\cdot} - \bar{Y}_{j\cdot}| > 4.02 \times \sqrt{5.46/10} = 2.97$$

are deemed different.



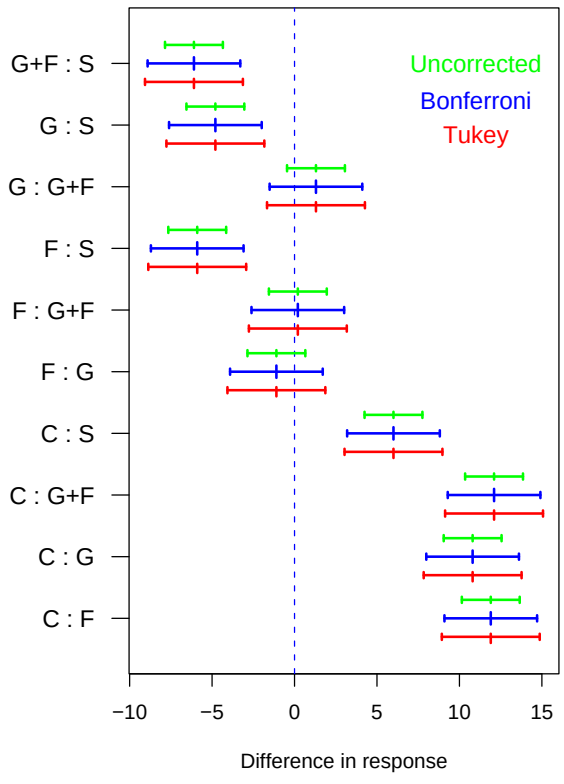
A comparison

Uncorrected:

Each interval, individually, had (in advance) a 95% chance of covering the true mean difference.

Corrected:

(In advance) there was a greater than 95% chance that all of the intervals would cover there respective parameters.



Newman-Keuls procedure

Goal: Identify sets of treatments whose mean responses are not significantly different.

(Assuming equal sample sizes for the treatment groups.)

Procedure:

1. Calculate the group sample means.
2. Order the sample means from smallest to largest.
3. Calculate a triangular table of all pairwise sample means.
4. Calculate $q_i = Q_\alpha(i, df)$ for i = 2, 3, . . . , k.
   (Use `qtukey()` in R.
5. Calculate $R_i = q_i \times \sqrt{M_W/n}$.

Newman-Keuls procedure (continued)

Procedure:

6. If the difference between the biggest and the smallest means is less than R_k, draw a line under all of the means and stop.
7. Compare the second biggest and the smallest (and the second-smallest and the biggest) to R_{k-1}. If observed difference is smaller than the critical value, draw a line between these means.
8. Continue to look at means for which a line connecting them has not yet been drawn, comparing the difference to R_i with progressively smaller i's.

Example

Sorted sample means:

G+F	F	G	S	C
58.0	58.2	59.3	64.1	70.1

Table of differences:

	F	G	S	C
G+F	0.2	1.3	6.1	12.1
F		1.1	5.9	11.9
G			4.8	10.0
S				6.0

Example (continued)

From the ANOVA table:

$M_W = 5.46$ n = 10 for each group $\sqrt{M_W/10} = 0.739$ df = 45

The q_i (using df=45 and $\alpha = 0.05$):

q_2	q_3	q_4	q_5
2.85	3.43	3.77	4.02

$R_i = q_i \times \sqrt{M_W/10}$:

R_2	R_3	R_4	R_5
2.10	2.53	2.79	2.97

Example (continued)

Table of differences:

	F	G	S	C
G+F	0.2	1.3	6.1	12.1
F		1.1	5.9	11.9
G			4.8	10.0
S				6.0

$R_i = q_i \times \sqrt{M_W/10}$:

R_2	R_3	R_4	R_5
2.10	2.53	2.79	2.97

Results

Sorted sample means:

G+F	F	G	S	C
58.0	58.2	59.3	64.1	70.1

Interpretation:

$$G+F \approx F \approx G < S < C$$

Another example

Sorted sample means:

D	C	A	B	E
29.6	32.9	40.0	40.7	48.8

Table of differences:

	C	A	B	E
D	3.3	10.4	11.1	19.2
C		7.1	7.8	15.9
A			0.7	8.8
B				8.1

Example (continued)

From the ANOVA table:

$M_W = 21.29$ n = 4 for each group $\sqrt{M_W/4} = 2.31$ df = 15

The q_i (using df=15 and $\alpha = 0.05$):

q_2	q_3	q_4	q_5
3.01	3.67	4.08	4.37

$R_i = q_i \times \sqrt{M_W/4}$:

R_2	R_3	R_4	R_5
6.95	8.47	9.40	10.07

Example (continued)

Table of differences:

	C	A	B	E
D	3.3	10.4	11.1	19.2
C		7.1	7.8	15.9
A			0.7	8.8
B				8.1

$R_i = q_i \times \sqrt{M_W/4}$:

R_2	R_3	R_4	R_5
6.95	8.47	9.40	10.07

Results

Sorted sample means:

D	C	A	B	E
29.6	32.9	40.0	40.7	48.8

Interpretation:

$\{D, C, A, B\} < E$ and $D < \{A, B\}$

Varying sample sizes

For the Tukey and Newman-Keuls methods, we assumed that the numbers of responses in each treatment group were the same.

What to do if they vary?

- If they don't vary too much, use $1/n_i + 1/n_j$ in place of $2/n$.
- If they are quite different, it's probably best to just stick with the Bonferroni correction.

Final words on multiple comparisons