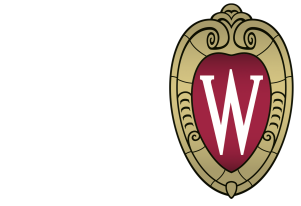
Lecture 20: ANOVA, Post-hoc/Multiple tests

STAT 324

Ralph Trane
University of Wisconsin–Madison

Spring 2020



WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON

Multiple Comparisons Following Significant ANOVA

```
library(tidyverse); library(distributions3); theme_set(theme_bw())

rat_poison <- tibble(treatment = as.character(rep(1:4, c(4, 6, 6, 8))),
                     coagulation_time = c(62, 60, 63, 59,
                                          63, 67, 71, 64, 65, 66,
                                          68, 66, 71, 67, 68, 68,
                                          56, 62, 60, 61, 63, 64, 63, 59))

summary(aov(coagulation_time ~ treatment, data = rat_poison))
```

```
            Df Sum Sq Mean Sq F value   Pr(>F)
treatment    3    228    76.0   13.57 4.66e-05 ***
Residuals   20    112     5.6
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The ANOVA showed us that there seem to be a difference somewhere between the treatment groups. More often than not, this is only somewhat interesting. We would much rather know where the differences are.

Multiple Comparisons Following Significant ANOVA

The first, and arguably simplest, is actually (a slightly modified version of) our first attempt to test this hypothesis: pairwise t-tests/confidence intervals. This approach is called **Fisher's Least Significant Difference (LSD)**.

Recall, for us to be able to perform an ANOVA, we need equal variance of all groups. So, the version of the t-test we will use is a two-sample t-test with equal variance.

Say we want to find out if the means in groups i and i' are different. I.e. we want to test $H_0 : \mu_i = \mu_{i'}$ vs. $H_A : \mu_i \neq \mu_{i'}$. We already established that we are willing to assume $\sigma_i = \sigma_{i'}$.

The test statistic we previously used is $T = \frac{\bar{y}_{i\cdot} - \bar{y}_{i'\cdot}}{s_p\sqrt{1/n_i + 1/n_{i'}}}$, where s_p is our best guess for σ, the common standard deviation of the data from the two populations

Now we have more populations, and we assume they all have same variance. We have a sample from each population. It would be silly not to utilize all the data available. Remember that $MSE = SSE/df_{\mathrm{E}}$ is pooled variance, i.e. our best guess of σ.

Multiple Comparisons Following Significant ANOVA

So, to check if $H_0 : \mu_i = \mu_{i'}$ vs. $H_A : \mu_i \neq \mu_{i'}$, we will use the test statistic $T = \frac{\bar{y}_{i\cdot} - \bar{y}_{i'\cdot}}{\sqrt{MSE(1/n_i + 1/n_{i'})}}$

.

If the null hypothesis is true, $T \sim t_{df_E}$. So, a $(1-\alpha)\cdot 100\%$ CI is given by

$$(\bar{y}_{i\cdot} - \bar{y}_{i'\cdot}) \pm t_{\alpha/2, df_E}\sqrt{MSE(1/n_i + 1/n_{i'})}.$$

Example: a $95/$ confidence interval for the difference between groups 1 and 3 is given by:

$$\begin{aligned}(\bar{y}_{1\cdot} - \bar{y}_{3\cdot}) \pm t_{0.025,20}\sqrt{MSE(1/n_1 + 1/n_3)} &= (61 - 68) \pm 2.086\sqrt{5.6 \cdot (1/4 + 1/6)} \\ &= [-10.186, -3.814]\end{aligned}$$

Similarly, we can test the hypothesis: $T_{\text{obs}} = \frac{61-68}{\sqrt{5.6(1/4+1/6)}} = -4.583$, which gives us a p-value = $2 \cdot P(T < T_{\text{obs}} \mid H_0 \text{ true}) = 1.8 \times 10^{-4}$.

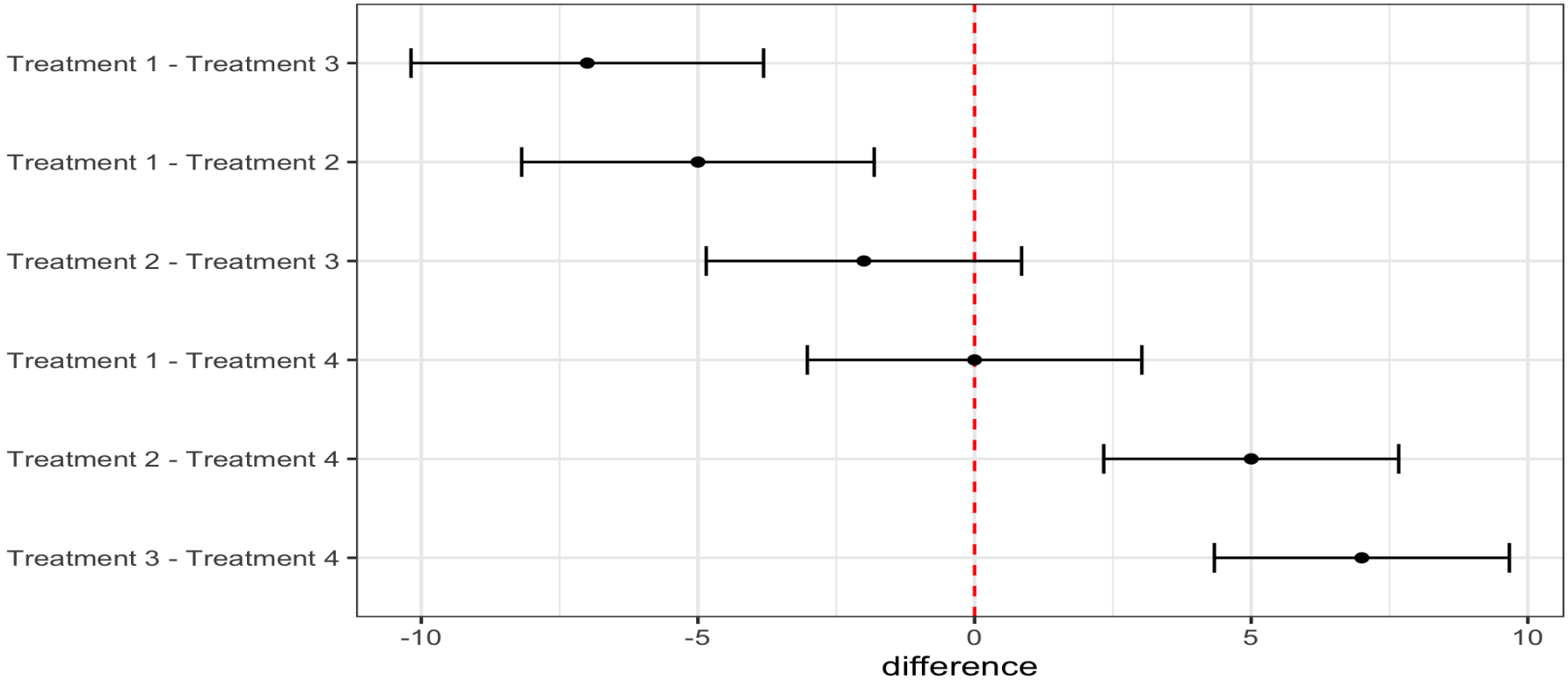
Multiple Comparisons Following Significant ANOVA

Here's a table with all of the pairwise comparisons.

group_A	group_B	mean_A	n_A	mean_B	n_B	difference	CI_lower	CI_
1	2	61	4	66	6	-5	-8.186	
1	3	61	4	68	6	-7	-10.186	
1	4	61	4	61	8	0	-3.023	
2	3	66	6	68	6	-2	-4.85	
2	4	66	6	61	8	5	2.334	
3	4	68	6	61	8	7	4.334	

Multiple Comparisons Following Significant ANOVA

And a nice plot:

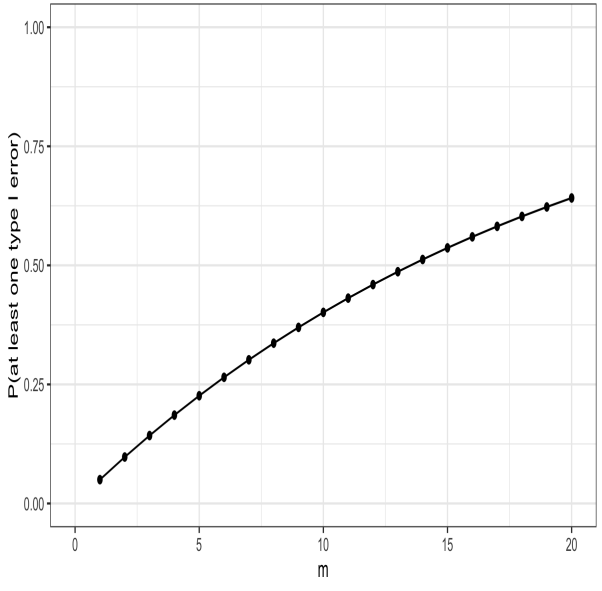


If the vertical line (0) intersects a CI, then that difference is not significantly different from 0.

Multiple Comparisons Following Significant ANOVA

We still have the multiple testing problem here: each test we perform has a 5% chance of rejecting the null even if the null is true. So, when performing six tests, we have a $1 - 0.95^6 = 0.2649$.

In general, when performing m tests using α for each of them, there's a $1 - (1 - \alpha)^m$ chance we make at least one type I error. Say $\alpha = 0.05$. Then $P(\text{at least one type I error})$ for different values of m:



Technically only accurate if all tests are independent, which might not be the case here. If test are dependent, same pattern, but exact probabilities will be somewhat different.

The problem is that we reject too many null hypotheses. Fortunately, there are techniques to adjust when we reject the null hypothesis. We will discuss two popular choices here.

Bonferroni Correction

The Bonferroni correction is in my opinion the simplest/most intuitive approach.

Say we are just testing two hypotheses, H_0^1 and H_0^2. We want to make sure that the probability that we falsely reject at least one of them is α.

$$P(\text{reject } H_0^1 \text{ OR reject } H_0^2 \mid H_0^1, H_0^2 \text{ true}) \leq P(\text{reject } H_0^1 \mid H_0^1 \text{ true}) + P(\text{reject } H_0^2 \mid H_0^1, H_0^2 true)$$

If we decide to adjust our criteria for when we reject H_0^1 and H_0^2 such that

$$P(\text{reject } H_0^1 \mid H_0^1 \text{ true}) = P(\text{reject } H_0^2 \mid H_0^2 \text{ true}) = \alpha/2,$$

then

$$P(\text{reject } H_0^1 \text{ OR reject } H_0^2 \mid H_0^1, H_0^2 \text{ true}) \leq \alpha.$$

In general, if we do m tests, the *Bonferroni correction* is to reject only when the p-value is less than α/m. This ensures that $P(\text{make at least one type I error}) \leq \alpha$. (Sometimes people will calculate the *Bonferroni corrected* p-values: this is simply m times the original p-value.)

For confidence intervals, we then use $t_{(\alpha/m)/2, df_{\text{E}}}$ as the multiplier.

Bonferroni Correction

For example, to find a 95% Bonferroni Corrected Confidence Interval, we would use $t_{(0.05/6)/2,df_{\mathrm{E}}} = t_{0.0042,df_{\mathrm{E}}} = 2.927$. So, to get the CI for the difference between treatment 2 and 4:

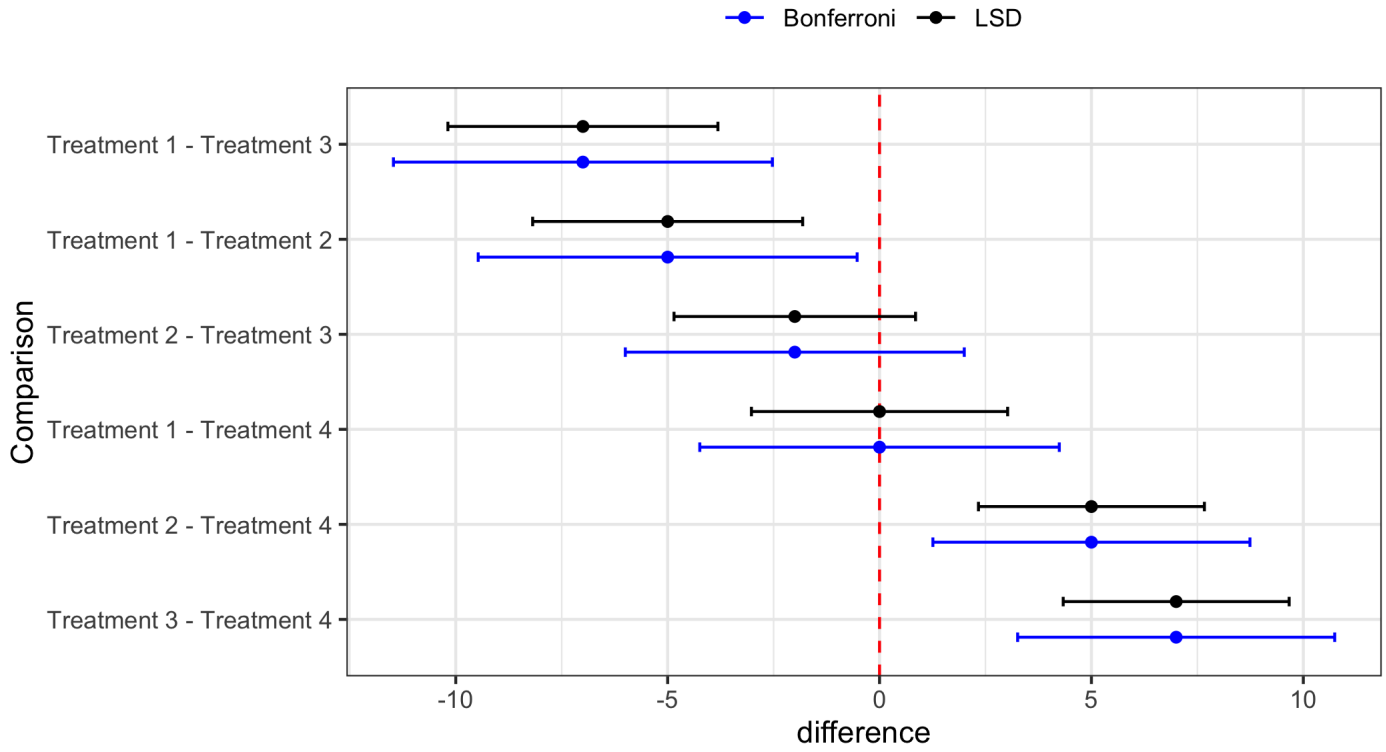
$$(66 - 61) \pm 2.927 \cdot \sqrt{5.6(1/6 + 1/8)} = [1.259, 8.741]$$

Table with all Bonferroni corrected CIs:

group_A	group_B	mean_A	n_A	mean_B	n_B	difference	CI_lower	CI_
1	2	61	4	66	6	-5	-9.471	
1	3	61	4	68	6	-7	-11.471	
1	4	61	4	61	8	0	-4.242	
2	3	66	6	68	6	-2	-5.999	
2	4	66	6	61	8	5	1.259	
3	4	68	6	61	8	7	3.259	

Bonferroni Correction

Plot with both types of CIs:



Notice how the Bonferroni corrected intervals are wider. This makes sense: the whole point is to reject the null less frequently. Wider intervals are more likely to contain 0, which means we will reject less frequently.

Bonferroni Correction

Although we have presented the Bonferroni Correction in the context of post-hoc analyses of an ANOVA, it is a general approach that can be used anytime you have done multiple tests, and want to control the overall (also called **familywise**) type I error rate. You simply replace α with α/m, where m is the number of tests.

Arguably the biggest benefit of the Bonferroni Correction is it's simplicity. Unfortunately, it is very conservative, especially if m is large. I.e. many hypothesis that should be rejected, will not be rejected. Unfortunately, we can never determine which ones were wrongly not rejected...

Tukey's Method

Tukey's Honest Significant Diffference (HSD) is another approach to adjusting confidence intervals. It can also be used to find adjusted p-values, but we will only consider CIs.

Just like how the Bonferroni method uses a different multiplier for the CIs, so does Tukey's method. However, Tukey's method uses an entirely different distribution. It's rather complicated, so we will not go into details. We will simply show how this is done.

Say we want to find a 95% CI for the difference between treatments 2 and 3. The multiplier used in Tukey's method is

$$\frac{Q_{\alpha,t,df_{\mathrm{E}}}}{\sqrt{2}}.$$

```
## NOTE!! If you want to use distributions3 for this, another update is required.
## Restart R (Session -> Restart R), and before doing anything else, run this line:
# devtools::install_github("rmtrane/distributions3")

tukey_dist <- Tukey(nmeans = 4, df = 20, nranges = 1) # note: nranges will always be 1

# multipler:
quantile(tukey_dist, 0.95)/sqrt(2)
```

```
[1] 2.798936
```

Tukey's Method

So the 95% CI is:

$$(66 - 68) \pm 2.799 \cdot \sqrt{5.6(1/6 + 1/6)} = [-5.824, 1.824].$$

```
## If you refuse to update distributions3: you can get the numerator by using qtukey
# qtukey(0.95, 4, 20)
```

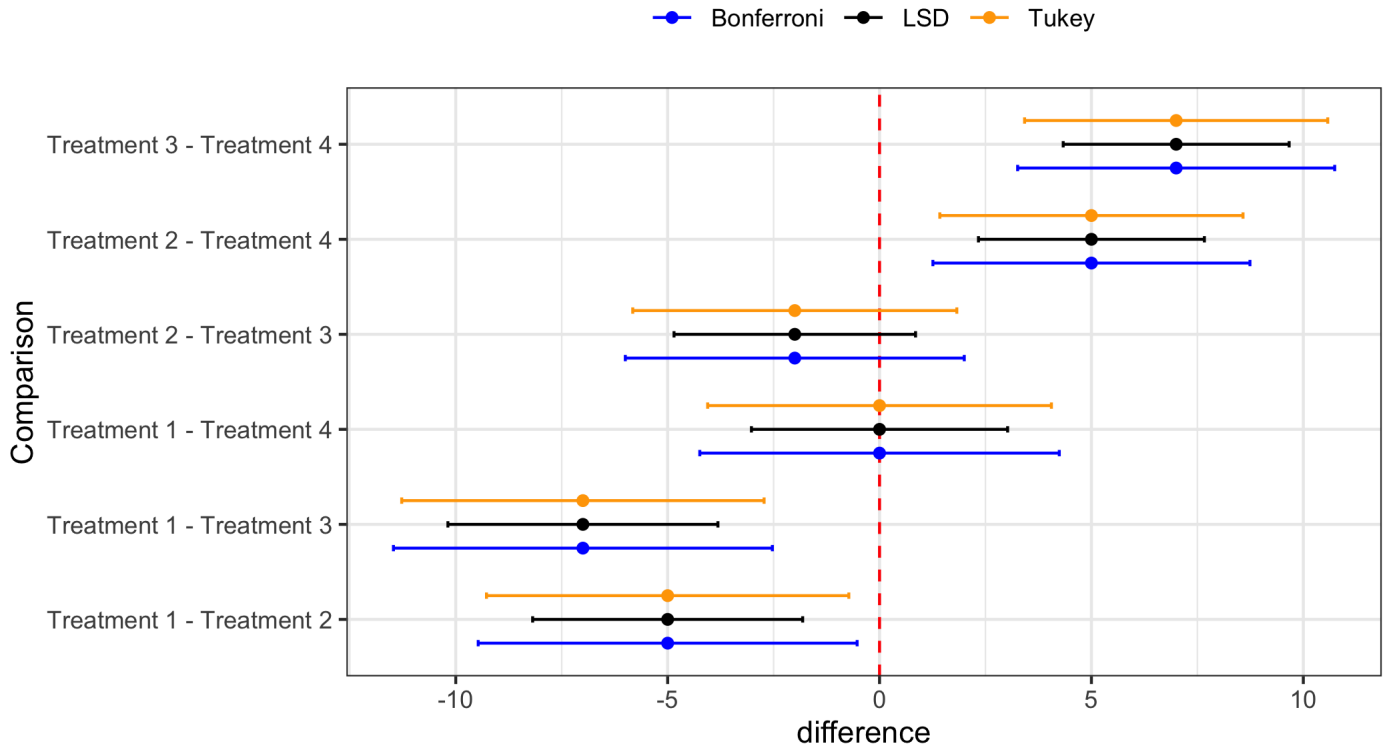
Tukey's Method

All Tukey Intervals:

Comparison	difference	CI_lower	CI_upper
Treatment 1 - Treatment 2	-5	-0.724554411356119	-9.27544558864388
Treatment 1 - Treatment 3	-7	-2.72455441135612	-11.2754455886439
Treatment 1 - Treatment 4	0	4.0560438216702	-4.0560438216702
Treatment 2 - Treatment 3	-2	1.82407478812373	-5.82407478812373
Treatment 2 - Treatment 4	5	8.57709441963978	1.42290558036021
Treatment 3 - Treatment 4	7	10.5770944196398	3.42290558036021

Tukey's Method

Same figure, including new CIs.



Notice how Tukey intervals are slightly more narrow than Bonferroni intervals. With more comparisons, this difference grows.

Tukey's Method

Personally, not a fan of Bonferroni. However, you might consider using it if:

1. You are doing a relatively limited number of tests
2. You are much more worried about type I errors than type II errors

Fisher's LSD is useful when if you don't care about type I errors at all, and simply want the most power.

Tukey's HSD is my preferred method in the context of ANOVA (although not as powerful as Bonferroni when number of tests is small).

A fourth option is the socalled Benjamini-Hochberg procedure, which you will see on the next homework.