Multiple Comparisons

STAT3022 Applied Linear Models Lecture 16

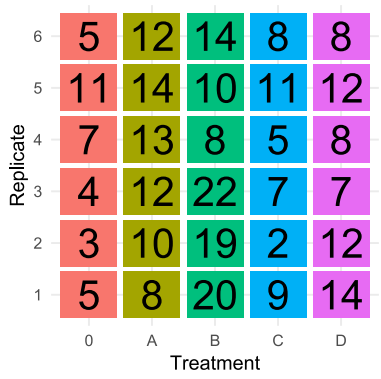
2020/02/17

THE UNIVERSITY OF SYDNEY

Today

1. Multiple comparisons
2. Tukey, Scheffe and Bonferroni corrections
3. Data snooping

Example: Fertilizer



- 5 treatments
- 6 replications each

```
skimr::skim(dat)
```

```
 Skim summary statistics
  n obs: 30
  n variables: 3

 -- Variable type:factor ------------------------------------------------------
   variable missing  n n_unique ordered
  Replicate       0 30        6   FALSE
        Trt       0 30        5   FALSE

 -- Variable type:numeric -----------------------------------------------------
  variable missing complete  n mean   sd p0 p25 p50 p75 p100     hist
         Y       0       30 30   10 4.81  2   7 9.5  12   22 <U+2583><U+2586>
```

```
round(c(tapply(Y,Trt,mean),mean(Y)),3)
```

```
     0      A      B      C      D
 5.833 11.500 15.500  7.000 10.167 10.000
```

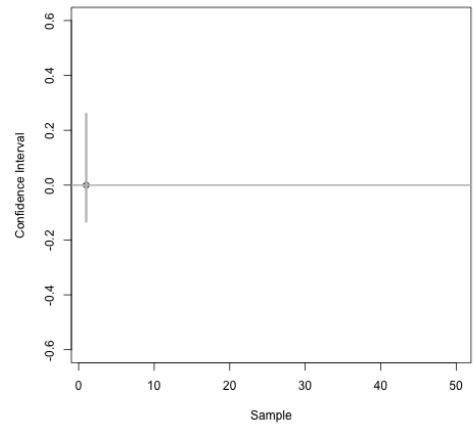
B and 0 have the largest difference and next come B and C.

All pairwise differences

- When no single group is "special" or notable, we can consider each pairwise difference as a contrast of interest.
- In this case,
  - a t-statistic can be constructed for each pairwise difference;
  - a t-based confidence interval can be constructed for each pairwise population difference (contrast).
- Let us focus on confidence intervals and revise the definition.

These lecture slides make grateful use of Michael Stewart's notes.

Recall confidence interval



- A 95% CI for the population mean of 0 is constructed repeatedly from 50 samples.
- We expect 95% of these CIs contain 0. In this run, 48/50 or 96% of these CIs contain 0.

Confidence interval for pairwise differences

```
library(emmeans)
M1 <- lm(Y ~ Trt, data = dat)
plot(pairs(emmeans(M1, "Trt"), adjust="none")) +
  theme_bw(base_size=18) +
  geom_vline(xintercept=0, size=2, color="red")
```

Construct a 95% CI for the contrast C - D

```
c(mean(dat$Y[dat$Trt=="C"]),mean(dat$Y[dat$Trt=="D"]))
```

```
[1]  7.00000 10.16667
```

```
c(sigma(M1), qt(0.975, M1$df.residual),M1$df.residual)
```

```
[1]  3.570247  2.059539 25.000000
```

- So for 95% CI we have

$$\hat{\alpha}_C - \hat{\alpha}_D \pm t_{n-t,0.975} \times SE(\hat{\alpha}_C - \hat{\alpha}_D)$$

$$(7 - 10.1667) \pm 2.0595 \times 3.5702 \times \sqrt{\frac{1}{6} + \frac{(-1)^2}{6}}$$

$$(-7.41, 1.08)$$

- This includes 0 so the difference is not significant.

95% CIs for pairwise differences (unadjusted)

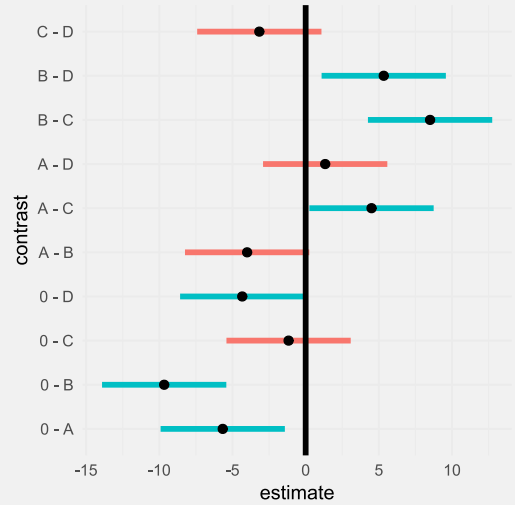
```
confint(pairs(emmeans(M1, "Trt"), adjust="none"))
```

```
 contrast estimate   SE df lower.CL upper.CL
 0 - A       -5.67 2.06 25   -9.912   -1.421
 0 - B       -9.67 2.06 25  -13.912   -5.421
 0 - C       -1.17 2.06 25   -5.412    3.079
 0 - D       -4.33 2.06 25   -8.579   -0.088
 A - B       -4.00 2.06 25   -8.245    0.245
 A - C        4.50 2.06 25    0.255    8.745
 A - D        1.33 2.06 25   -2.912    5.579
 B - C        8.50 2.06 25    4.255   12.745
 B - D        5.33 2.06 25    1.088    9.579
 C - D       -3.17 2.06 25   -7.412    1.079

Confidence level used: 0.95
```

So individually the significantly different pairs are

- B & D,
- B & C,
- A & C,
- 0 & D,
- 0 & B,
- 0 & A.



- However, we construct each of these $\binom{5}{2} = 10$ CIs without taking any regard of the others.

Controlling the family-wise error rate

- Each interval is constructed using a procedure so that when the model is correct, the "probability" that the correct population contrast is covered is 0.95. . . individually.
- But, what is the "probability" that all intervals cover their corresponding true values simultaneously?

Notations

- Let $A_1, A_2, \ldots, A_{10}$ denote the events where each of the 10 intervals above cover the corresponding "true" value.
- Then, under our model, we have

$$P(A_1) = P(A_2) = \ldots = P(A_{10}) = 0.95.$$

- What is $P(A_1 \cap A_2 \cap \ldots \cap A_{10})$?
- This is a bit hard, but we can derive a lower bound a bit more easily using the relation

$$(A_1 \cap A_2 \cap \ldots \cap A_{10})^c = A_1^c \cup A_2^c \cup \ldots \cup A_{10}^c$$

- Recall that $P(A \cup B) \leq P(A) + P(B)$.

$$\begin{aligned} 1 - P(A_1 \cap A_2 \cap \ldots \cap A_{10}) &= P\left((A_1 \cap A_2 \cap \ldots \cap A_{10})^c\right) \\ &= P\left(A_1^c \cup A_2^c \cup \ldots \cup A_{10}^c\right) \\ &\leq P(A_1^c) + P(A_2^c) + \ldots + P(A_{10}^c) \\ &= 10 \times 0.05 = 0.5 \\ \Rightarrow P(A_1 \cap A_2 \cap \ldots \cap A_{10}) &\geq 1 - 10 \times 0.05 = 0.5 \end{aligned}$$

- Therefore $P(A_1 \cap A_2 \cap \ldots \cap A_{10}) \geq 0.5$.
- The simultaneous coverage probability of all 10 intervals is therefore at least 50%.

Bonferroni Procedure

- To ensure a higher simultaneous coverage, the individual coverage should also be higher, eg $1 - 0.05/10 = 0.995$. Then

$$P(A_1 \cap A_2 \cap \ldots \cap A_{10}) \geq 1 - 10(1 - 0.995) = 0.95!$$

- For each individual $100(1-\alpha)$% CI, the critical value is $t_{n-t,1-\alpha/(2m)}$ where m is the number of pairwise comparisons.
- So for the Bonferroni adjusted 95% CI we have

$$\hat{\alpha}_C - \hat{\alpha}_D \pm t_{n-t,0.9975} \times \mathrm{SE}(\hat{\alpha}_C - \hat{\alpha}_D)$$

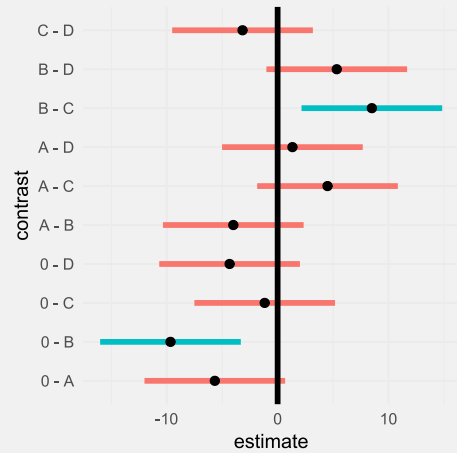
$$(7 - 10.1667) \pm 3.0782 \times 3.5702 \times \sqrt{\frac{1}{6} + \frac{(-1)^2}{6}}$$

$$(-9.51, 3.18)$$

which is wider and only BC & OB are significant.

```
confint(pairs(emmeans(M1, "Trt"),
              adjust="bonferroni"))
```

```
 contrast estimate   SE df lower.CL upper.CL
 0 - A       -5.67 2.06 25   -12.01    0.678
 0 - B       -9.67 2.06 25   -16.01   -3.322
 0 - C       -1.17 2.06 25    -7.51    5.178
 0 - D       -4.33 2.06 25   -10.68    2.012
```



Tukey's Procedure

- The Bonferroni adjustment which fulfills the lower bounds may not be exact.
- John Tukey derived the **exact multiplier** needed for simultaneous confidence intervals for all pairwise comparisons when the sample sizes for each group are equal*.
- It was later shown that when the sample sizes are unequal, Tukey's procedure is conservative, thus yielding valid simultaneous intervals that may be narrower than those using the Bonferroni method.
- Tukey named his method "Honest Significant Differences (HSD)" and use this distribution to replace the t distribution in the CI.

- Suppose $\bar{Y}_{1\bullet}, \ldots, \bar{Y}_{t\bullet}$ distributed as $NID(\mu, \sigma^2/n_i)$.
- $\hat{\sigma}^2$ is estimated with ν degrees of freedom.
- Then, the ratio

$$\left(\max_{1 \leq i \leq t} \bar{Y}_{i\bullet} - \min_{1 \leq i \leq t} \bar{Y}_{i\bullet}\right) / \hat{\sigma}$$

  is called the studentized range.
- The $(1-\alpha)$ quantile of the distribution of studentised range is denoted by $\mathrm{HSD}_\nu(1-\alpha; t)$.
- This multiplier $\mathrm{HSD}_\nu(1-\alpha; t)$ is calculated in R as `qtukey(1-alpha, t, df)`
- A $(1-\alpha)100$% Tukey confidence interval is given by

$$(\hat{\alpha}_i - \hat{\alpha}_{i'}) \pm \frac{\mathrm{HSD}_{n-t}(1-\alpha; t)}{\sqrt{2}} \times \mathrm{SE}(\hat{\alpha}_i - \hat{\alpha}_{i'})$$

Tukey's Honest Significant Differences

```
nlevels(dat$Trt)
```

```
[1] 5
```

```
nrow(dat)
```

```
[1] 30
```

```
qtukey(1 - 0.05, 5, 25)
```

```
[1] 4.153363
```

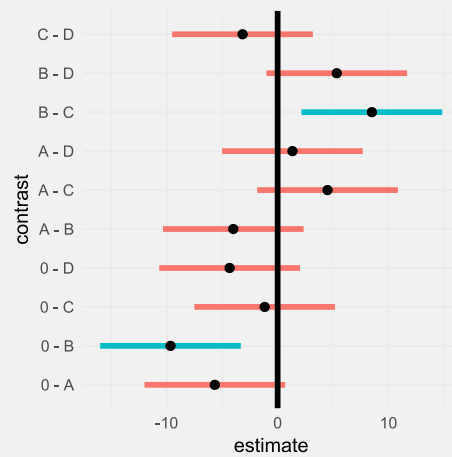
$$(7 - 10.1667) \pm \frac{4.1534}{\sqrt{2}} \times 3.5702 \times \sqrt{\frac{1}{6} + \frac{(-1)^2}{6}}$$

$$(-9.22, 2.89)$$

which is narrower than the Bonferroni adjusted CI $(-9.51, 3.18)$ and again, only BC & OB are significant.

```
confint(pairs(emmeans(M1, "Trt"),
              adjust="tukey"))
```

```
 contrast estimate   SE df lower.CL upper.CL
 0 - A       -5.67 2.06 25   -11.72    0.387
 0 - B       -9.67 2.06 25   -15.72   -3.613
 0 - C       -1.17 2.06 25    -7.22    4.887
 0 - D       -4.33 2.06 25   -10.39    1.720
```



The "most significant sample contrast"

- It can be shown (using the Cauchy-Schwarz inequality) that the "most significant sample contrast" (which is data dependent) is given by choosing the coefficients according to

$$c_i = n_i(\bar{Y}_{i\bullet} - \bar{Y}_{\bullet\bullet}).$$

- The corresponding contrast is given by (since $\sum_{i=1}^t n_i(\bar{Y}_{i\bullet} - \bar{Y}_{\bullet\bullet})\bar{Y}_{\bullet\bullet} = 0$)

$$\hat{c} = \sum_{i=1}^t c_i \bar{Y}_{i\bullet} = \sum_{i=1}^t n_i(\bar{Y}_{i\bullet} - \bar{Y}_{\bullet\bullet})\bar{Y}_{i\bullet} = \sum_{i=1}^t n_i(\bar{Y}_{i\bullet} - \bar{Y}_{\bullet\bullet})^2.$$

- An underestimate of $\mathrm{SE}(\hat{c})$ assuming c_i are known (not random) is

$$\widehat{\mathrm{SE}}(\hat{c}) = \sqrt{\widehat{\mathrm{Var}}\left(\sum_{i=1}^t n_i(\bar{Y}_{i\bullet} - \bar{Y}_{\bullet\bullet})\bar{Y}_{i\bullet}\right)} = \sqrt{\sum_{i=1}^t n_i^2(\bar{Y}_{i\bullet} - \bar{Y}_{\bullet\bullet})^2 \widehat{\mathrm{Var}}(\bar{Y}_{i\bullet})}$$

So the test statistic for testing $H_0 : c = 0$ is

$$\begin{aligned} t &= (\hat{c} - 0)/\widehat{\mathrm{SE}}(\hat{c}) \\ &= \sqrt{\frac{\sum_{i=1}^t n_i(\bar{Y}_{i\bullet} - \bar{Y}_{\bullet\bullet})^2}{\hat{\sigma}^2}} \\ &= \sqrt{(t-1)\frac{\mathrm{TSS}/(t-1)}{\mathrm{RSS}/(n-t)}} \\ &= \sqrt{(t-1)F_{t-1,n-t}}. \end{aligned}$$

Hence the multiplier

$$\sqrt{(t-1)F_{t-1,n-t}(1-\alpha)}$$

is chosen to construct the Scheffe simultaneous CI (again to replace t quantile).

Scheffe adjustment

```
sqrt(4 * qf(0.95, 4, 25))
```

```
[1] 3.321873
```

- So for the Scheffe adjusted 95% CI we have

$$\hat{\alpha}_C - \hat{\alpha}_D \pm \sqrt{(t-1)F_{t-1,n-t}(0.95)} \times \widehat{\mathrm{SE}}(\hat{\alpha}_C - \hat{\alpha}_D)$$

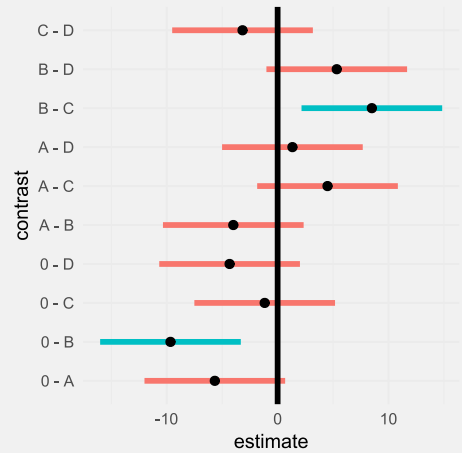
$$(7 - 10.1667) \pm 3.3219 \times 3.5702 \times \sqrt{\frac{1}{6} + \frac{(-1)^2}{6}}$$

$$(-10.01, 3.68)$$

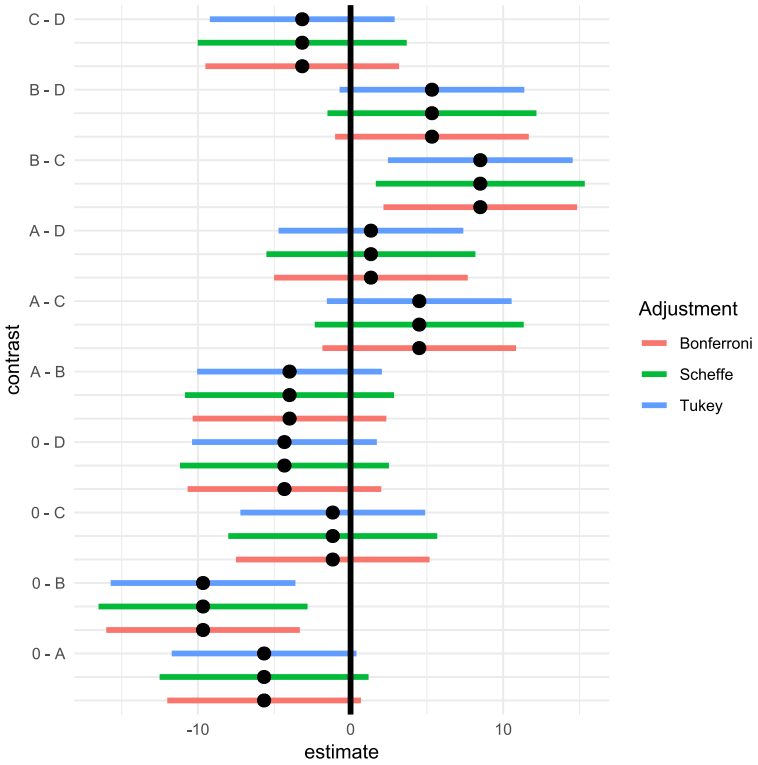
which is wider than both Bonferroni adjusted CI $(-9.51, 3.18)$ and Tukey adjusted CI $(-9.22, 2.89)$ (may due to the underestimation of $\mathrm{SE}(\hat{c})$).

```
confint(pairs(emmeans(M1, "Trt"),
              adjust="scheffe"))
```

```
 contrast estimate   SE df lower.CL upper.CL
 0 - A       -5.67 2.06 25   -12.51     1.18
 0 - B       -9.67 2.06 25   -16.51    -2.82
 0 - C       -1.17 2.06 25    -8.01     5.68
 0 - D       -4.33 2.06 25   -11.18     2.51
```



Conclusion - Multiple Testing



- If all pairwise comparisons are of interest: Tukey is superior to Bonferroni.
- If not all pairwise comparisons are of interest: Bonferroni can be better than Tukey.
- Bonferroni is better than Scheffe when number of contrasts of interest about the same as the number of factor levels.
- In practice: choose smallest of the three!