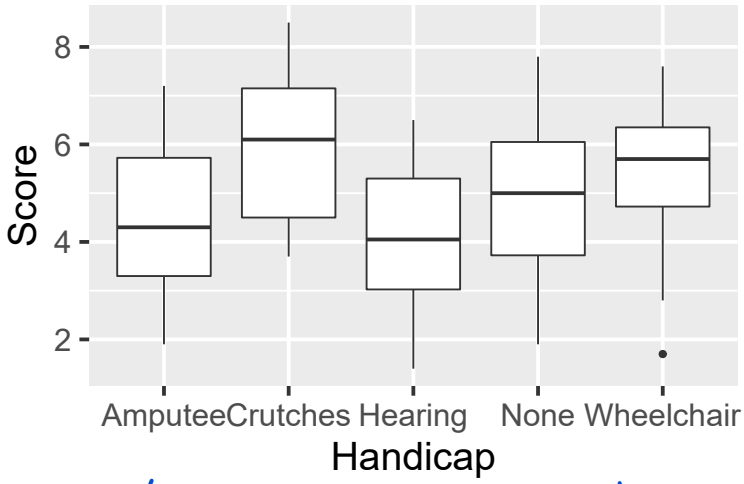
ST 411/511 Outline 6

Reading assignment: Chapter 6. This chapter discusses inference on linear combinations (usually differences or differences of averages) of population means, introduces the multiple comparison problem, and gives some solutions.

Chapter 6 Linear Combinations and Multiple Comparisons

Case Study 6.1.1: Discrimination against disabled job applicants.



General research question: Are popn mean scores different?

μ_i = popn mean of i^{th} group.

$H_0: \mu_1 = \ldots = \mu_5$

H_A: at least one μ_i is different

or $H_A: \mu_i \neq \mu_j$ for some $i \neq j$

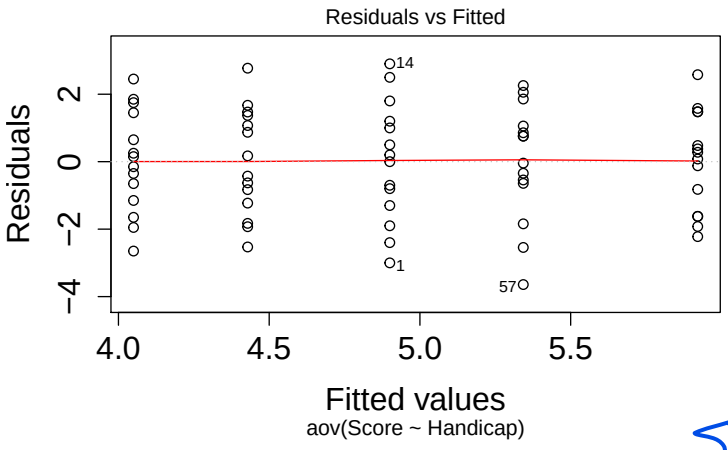
```
> head(case0601, n=3)
  Score Handicap
1   1.9     None
2   2.5     None
3   3.0     None
```

```
> case0601_aov<- aov(Score~Handicap, data=case0601)
> anova(case0601_aov)
Analysis of Variance Table

Response: Score
           Df  Sum Sq Mean Sq F value  Pr(>F)
Handicap    4  30.521  7.6304  2.8616 0.03013 *
Residuals 65 173.321  2.6665
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

s_p^2

```
> plot(case0601_aov, which=1)
```



Good tool to assess equal var. assumption.

Equal var OK here because spread is very similar for the 5 groups.

Normality OK too since points are approx. symmetric across 0.

Statistical conclusion reporting ANOVA F-test:

Moderate evidence that pop'n mean scores differ among types of disability ($p \approx 0.03$, one-way ANOVA F-test)

More specific research questions:

- Could compare "none" to each other disability in 4 "pairwise" comparisons
- Compare "none" to "disabled"
  This is not a pairwise comparison.

```
> # Find sample means.
> with(case0601, unlist(lapply(split(Score, Handicap), mean)))
   Amputee   Crutches    Hearing       None Wheelchair
  4.428571   5.921429   4.050000   4.900000   5.342857

> # Find sample sizes.
> with(case0601, unlist(lapply(split(Score, Handicap), length)))
   Amputee   Crutches    Hearing       None Wheelchair
        14         14         14         14         14
```

Same code as in Ch. 5

One specific question: How does "none" compare with "disabled?"

Compare μ_4 (popn mean for "none") with average of $\mu_1, \mu_2, \mu_3, \mu_5$. (Need to average, so "disabled" doesn't get too much weight.)

Definition: A *linear combination* of population means has the form

Comparison $\rightarrow$ $\gamma = C_1\mu_1 + C_2\mu_2 + \ldots + C_I\mu_I$

The C_i's are numbers called "coefficients"

$$\gamma = \mu_4 - \frac{\mu_1 + \mu_2 + \mu_3 + \mu_5}{4}$$

$$1\cdot\mu_4 - \frac{1}{4}\mu_1 - \frac{1}{4}\mu_2 - \frac{1}{4}\mu_3 - \frac{1}{4}\mu_5$$

C_i's are either 1 or $-\frac{1}{4}$.

Recall: Comparing means μ_1 vs. μ_2 in the two-sample case.

$$\begin{aligned}\gamma &= \mu_1 - \mu_2 \\ &= 1\cdot\mu_1 + -1\cdot\mu_2\end{aligned}$$

$C_1 = 1 \qquad C_2 = -1$

Point estimate:

hat means "estimate of"

$$\hat{\gamma} = \overline{Y}_1 - \overline{Y}_2 \quad \leftarrow \text{substitute sample means}$$
$$= 1 \cdot \overline{Y}_1 + -1 \cdot \overline{Y}_2$$

c_1 c_2

Standard error:

$$\text{SE}(\hat{\gamma}) = s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \quad \text{ci's}$$
$$= s_p \sqrt{\frac{(1)^2}{n_1} + \frac{(-1)^2}{n_2}}$$

Book's notation

Point estimate for $\gamma = C_1\mu_1 + C_2\mu_2 + \ldots + C_I\mu_I$

$g = \hat{\gamma} = C_1\overline{Y}_1 + C_2\overline{Y}_2 + \ldots + C_I\overline{Y}_I$

$g = -\frac{1}{4}4.429 - \frac{1}{4}5.921 - \frac{1}{4}4.05 + 1 \cdot 4.9 - \frac{1}{4}5.343$

≈ -0.03575

Standard error for g: $\text{SE}(g) = s_p \cdot \sqrt{\frac{C_1^2}{n_1} + \frac{C_2^2}{n_2} + \ldots + \frac{C_I^2}{n_I}}$

$$SE(g) = \sqrt{2.6665}\sqrt{\frac{(-1/4)^2}{14} + \frac{(-1/4)^2}{14} + \frac{(-1/4)^2}{14} + \frac{1^2}{14} + \frac{(-1/4)^2}{14}}$$

$s_p = \sqrt{MSE}$ from ANOVA table on p. 1

≈ 0.4879

90% confidence interval for γ

0.95

$\gamma = 0.1$

pt est $\pm\, t_{\text{df}}(1-\alpha/2)$ SE(pt est)

-0.03575 ↑ 0.4879

$t_{65}(0.95)$

```
> qt(0.95, 65)
[1] 1.668636
```

CI: $-0.03575 \pm 1.668636 \cdot 0.4879$

$(-0.85, 0.78)$

Could instead use pt. est.

Statistical conclusion: We estimate (no diff.) in pop'n mean score between disabled and non-disabled job applicants (90% CI -0.85 to 0.78).

Testing $H_0 : \gamma = 0$:

$H_A: \gamma \neq 0$

or

$H_A: \gamma > 0$

$\gamma < 0$

$$t\text{-stat} = \frac{\text{pt.est.} - \text{value under } H_0}{\text{SE(pt. est.)}} = \frac{-0.03575 - 0}{0.4879}$$

Another specific comparision: amputee and hearing vs. crutches and wheelchair

$$\gamma = \frac{\mu_1 + \mu_3}{2} - \frac{\mu_2 + \mu_5}{2}$$

$$= \frac{1}{2}\mu_1 - \frac{1}{2}\mu_2 + \frac{1}{2}\mu_3 + 0\cdot\mu_4 - \frac{1}{2}\mu_5$$

Caution: If this comparison was *data-suggested*, usual statistical conclusion of this confidence interval is invalid. A little later, we will see a procedure that works for data-suggested comparisons. Need to use Scheffés procedure.

DISPLAY 6.4 Confidence interval construction for the linear combination $\gamma = (\mu_3 + \mu_5)/2 - (\mu_2 + \mu_4)/2$ in the handicap study

(1) *Summary statistics* → $s_p = 1.6329$; 65 *d.f.*

order different from R's

	Feigned Handicap				
	None	Amputee	Crutches	Hearing	Wheelchair
n:	14	14	14	14	14
Average:	4.900	4.4286	5.9214	4.0500	5.3429
C:	0	−1/2	+1/2	−1/2	+1/2

(2) *Specify the coefficients for the linear combination.* ← Coefficients

(3) *Estimate the linear combination.*

$$g = \frac{(\overline{Y}_3 + \overline{Y}_5)}{2} - \frac{(\overline{Y}_2 + \overline{Y}_4)}{2} = \frac{(5.9214 + 5.3429)}{2} - \frac{(4.4286 + 4.0500)}{2}$$

$$= 1.3929$$

(4) *Find the standard error of the estimate.*

$\sqrt{2.6665}$ from ANOVA table

$$\text{SE}(g) = 1.6329\sqrt{\frac{(0)^2}{14} + \frac{(-1/2)^2}{14} + \frac{(+1/2)^2}{14} + \frac{(-1/2)^2}{14} + \frac{(+1/2)^2}{14}}$$

$$= 0.4364$$

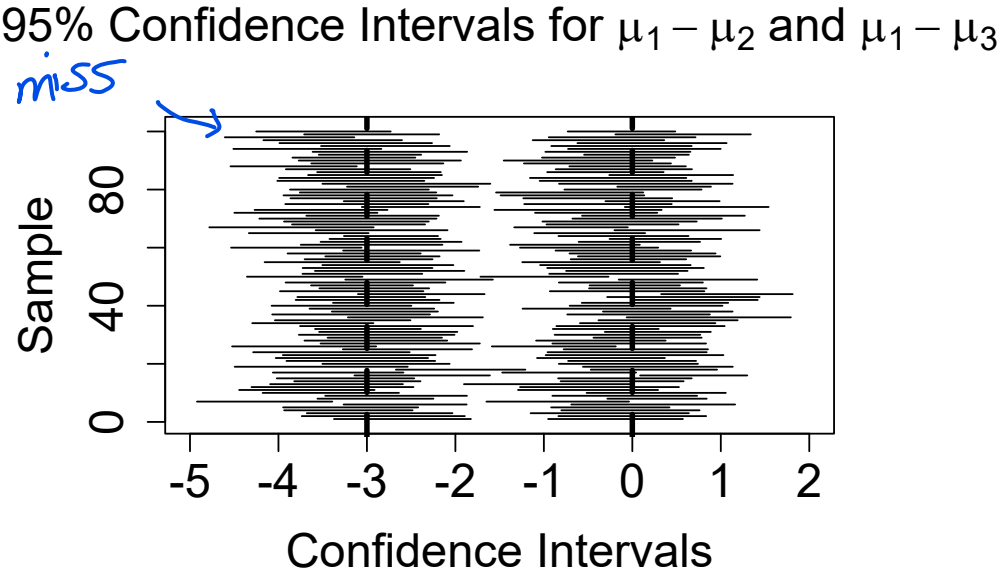
(5) *Construct the 95% confidence interval.*

$t_{65}(0.975) = 1.9971$ ← *from the t-distribution with 65 d.f.*

$1.3929 \pm (1.9971) \times (0.4364)$ → **from 0.521 to 2.264**

The Simultaneous Inference Problem

Illustration via simulation: Simulated 100 samples of size 15 from three normal populations with $\sigma = 1$. Population means were $\mu_1 = 2$, $\mu_2 = 5$, and $\mu_3 = 2$.




```
> # Count successful CIs for mu1-mu2
> length(which(CI12[,1] < -3 & CI12[,2] > -3)
+ )
[1] 94
> # Count successful CIs for mu1-mu3
> length(which(CI13[,1] < 0 & CI13[,2] > 0))
[1] 94
> # Count samples where both CIs were successful
> length(which(CI12[,1] < -3 & CI12[,2] > -3
+              & CI13[,1] < 0 & CI13[,2]  >0))
[1] 90
```

= 94% coverage

= 94% coverage

Would be 95% if many more than 100 simulations.

← Only 90% of the time do both intervals cover the true diff.

Success rate goes down as we do more CI's. Each CI has its own chance to miss.

Individual confidence level: Over repeated sampling from the population, the proportion of confidence intervals for a parameter that contain the true parameter.

95% for a 95% CI

Familywise confidence level: Over repeated sampling from the population, the proportion of times that **all** confidence intervals in the family contain the true parameters.

Less than 95% for a family of 95% CI's unless we use a specialized procedure.

Distinguishing between **planned** and **unplanned** comparisons

Planned comparisons determined when study was designed. (Like mouse study)
Unplanned comparisons are suggested by the data. Must use Scheffé's procedure for valid inference with unplanned comparisons.

Four of the many techniques for simultaneous confidence intervals:

- Tukey-Kramer
- Dunnett
- Scheffé
- Bonferroni

} For HW calculate these. For final, know which is appropriate in a given situation, and interpret R output.

Recall: General form of a confidence interval: pt est $\pm$ multiplier $\cdot$ SE(pt est)

t-quantile up to now

Each of our 4 procedures will have its own multiplier.

Tukey-Kramer procedure:
For making all possible pairwise comparisons when these were pre-planned.

Pairwise comparisons: $\gamma = \mu_i - \mu_j$

aov object from

```
> TukeyHSD(case0601_aov)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = Score ~ Handicap, data = case0601)

$`Handicap`
                           diff        lwr        upr     p adj
Crutches-Amputee      1.4928571 -0.2388756  3.2245899 0.1232819
Hearing-Amputee      -0.3785714 -2.1103042  1.3531613 0.9724743
None-Amputee          0.4714286 -1.2603042  2.2031613 0.9399911
Wheelchair-Amputee    0.9142857 -0.8174470  2.6460185 0.5781165
Hearing-Crutches     -1.8714286 -3.6031613 -0.1396958 0.0277842
None-Crutches        -1.0214286 -2.7531613  0.7103042 0.4686233
Wheelchair-Crutches  -0.5785714 -2.3103042  1.1531613 0.8812293
None-Hearing          0.8500000 -0.8817328  2.5817328 0.6442517
Wheelchair-Hearing    1.2928571 -0.4388756  3.0245899 0.2348141
Wheelchair-None       0.4428571 -1.2888756  2.1745899 0.9517374
```

conf. limits of Tukey-Kramer CI's.

* Hearing-Crutches

Statistical conclusion: [show CI's in a table]

The only two popn means estimated to be different are hearing & crutches (95% Tukey-Kramer CI)

* CI for diff in popn means for hearing & crutches doesn't contain 0, so 0 is not a plausible value for this diff.

Background (OK to ignore): Tukey procedure based on *studentized range distribution*, the sampling distribution under $H_0 : \mu_1 = \mu_2 = \ldots = \mu_I$ of

$$q = \frac{\max(\overline{Y}_i) - \min(\overline{Y}_i)}{\mathrm{SE}(\overline{Y}_i)}$$

Assumes sample sizes all equal. Kramer's contribution allows diff. sample sizes.

Tukey 95% CI for $\mu_i - \mu_j$:

Tukey-Kramer multiplier

$$\text{pt est} \pm q_{I,n-I}(1-\alpha)/\sqrt{2} \cdot \mathrm{SE}(\text{pt est})$$

not $1-\alpha/2$

Don't forget to divide by $\sqrt{2}$.

Calculating the Tukey-Kramer multiplier in R:

$1-\alpha$ ↘ I ↓ ← res. df from ANOVA table

```
> qtukey(0.95, 5, 65)/sqrt(2)
[1] 2.805824
```

Tukey 95% CI for $\mu_1 - \mu_2$:

```
# Sample means and sample sizes

> with(case0601, unlist(lapply(split(Score, Handicap), mean)))
   Amputee   Crutches    Hearing       None Wheelchair
  4.428571   5.921429   4.050000   4.900000   5.342857

> with(case0601, unlist(lapply(split(Score, Handicap), length)))
   Amputee   Crutches    Hearing       None Wheelchair
        14         14         14         14         14
```

$$(4.428571 - 5.921429) \pm 2.805824 \cdot \sqrt{2.6664}\sqrt{\tfrac{1}{14} + \tfrac{1}{14}}$$

$\approx (-3.22, 0.24)$

cf. 1st interval in TukeyHSD( ) output

s_p^2 = res. MS from ANOVA table on p. 1

Dunnett's procedure:
For comparing each treatment to a control (also must be preplanned)

$\gamma = \mu_i - \mu_{control}$ You decide which is "control" but this has to be preplanned.

```
> # Package multcomp does Dunnett's and others.
> library(multcomp)

> # The function in the multcomp package assumes
> # the control is the first level.

> summary(case0601$Handicap)
   Amputee    Crutches     Hearing        None  Wheelchair
        14          14          14          14          14
```

want "none" to be control

```
> case0601$Handicap <- relevel(case0601$Handicap, "None")
> summary(case0601$Handicap)
      None    Amputee   Crutches    Hearing Wheelchair
        14          14          14          14          14
```

Move "none" to front

```
> # Rerun aov() because we releveled.
> case0601_aov <- aov(Score~Handicap, data=case0601)
```

```
> library(multcomp) # For Dunnett's
>
> # Do Dunnett's procedure.
> case0601_glht <- glht(case0601_aov, linfct=mcp(Handicap="Dunnett"))
```

"general linear hypothesis test"

glht object

linear function

multiple comparison procedure

```
> # Get Dunnett's confidence intervals.
> confint(case0601_glht)
```

gives us confidence intervals

```
 Simultaneous Confidence Intervals

Multiple Comparisons of Means: Dunnett Contrasts

Fit: aov(formula = Score ~ Handicap, data = case0601)

Quantile = 2.5031
95% family-wise confidence level

Linear Hypotheses:
                         Estimate lwr     upr
Amputee - None == 0      -0.4714 -2.0163  1.0735
Crutches - None == 0      1.0214 -0.5235  2.5663
Hearing - None == 0      -0.8500 -2.3949  0.6949
Wheelchair - None == 0    0.4429 -1.1020  1.9877
```

Dunnett * multiplier

Dunnett's CIs

Statistical conclusion: We estimate no differences between pop'n mean scores for any of the disabilities and "none" (95% Dunnett's CI's). (Refer to table.)

* Note. Dunnett's multiplier is smaller than Tukey-Kramer. Dunnett's does fewer intervals.

Scheffé procedure: For all possible comparisons, including data-suggested ones, e.g. amputee and hearing vs. crutches and wheelchair.

Lots of CIs!

Not just pairwise.

$$\gamma = \frac{\mu_1 + \mu_3}{2} - \frac{\mu_2 + \mu_5}{2} = \frac{1}{2}\mu_1 + \frac{1}{2}\mu_3 - \frac{1}{2}\mu_2 - \frac{1}{2}\mu_5$$

Scheffé 95% CI for γ: pt est $\pm\, M \cdot$ SE(pt est)

Scheffé multiplier:

$$M = \sqrt{(I-1) \cdot F_{I-1,\mathrm{df}}(1-\alpha)}$$

I = # groups

quantile from F-distn with $I-1$ & df degrees of freedom

(df = residual df from ANOVA table)

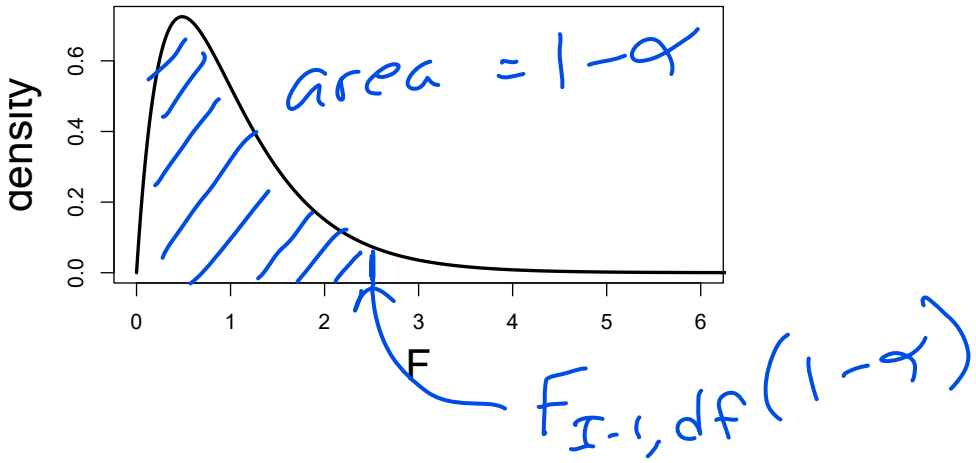
```
> # Scheffe multiplier for 95% CI with I=5 and residual df=65
> (M <- sqrt(4 * qf(0.95, 4, 65)))
[1] 3.170514
```

4 = I-1

65 = res. df

for a 95% Scheffé CI

Background (OK to ignore): If $H_0 : \mu_1 = \ldots = \mu_I$ is true, F-stat from ANOVA F-test has an F distribution with $I-1$ and $n_1 + \ldots + n_I - I$ degrees of freedom.



Calculating a Scheffé confidence interval

```
> # Sample means and sample sizes
> with(case0601, unlist(lapply(split(Score, Handicap), mean)))
   Amputee   Crutches    Hearing       None Wheelchair
  4.428571   5.921429   4.050000   4.900000   5.342857
> with(case0601, unlist(lapply(split(Score ,Handicap), length)))
   Amputee   Crutches    Hearing       None Wheelchair
        14         14         14         14         14

> # Point estimate
> (g <- (4.428571+4.05)/2 - (5.921429+5.342857)/2)
[1] -1.392858

> # Standard error
> (SE <- sqrt(2.6665) * sqrt(0.5^2/14 + 0.5^2/14 + 0.5^2/14 + 0.5^2/14))
[1] 0.4364221

> # Lower Scheffe confidence limit
> g - M*SE
[1] -2.776183

> # Upper Scheffe confidence limit
> g + M*SE
[1] -0.008817396
```

s_p

$4 \cdot (0.5)^2/14$

$\frac{(-0.5)^2}{14} + \frac{(-0.5)^2}{14}$

Statistical conclusion: We estimate that the average popn mean score for crutches & wheelchair is 0.0088 to 2.776 points higher than average popn mean score for amputee & hearing (95% Scheffé CI).

Note: 0 isn't in CI, but it's very close to an endpoint. Related to the idea of practical significance.

Bonferroni procedure: For a small number (K) of preplanned inferences, e.g. (1) none vs. others, (2) amputee vs. crutches, and (3) amputee vs. wheelchair.

```
> summary(case0601$Handicap)
   Amputee    Crutches     Hearing        None Wheelchair
        14          14          14          14          14
```

(1) $\mu_4 - \frac{\mu_1 + \mu_2 + \mu_3 + \mu_5}{4}$

(2) $\mu_1 - \mu_2$

(3) $\mu_1 - \mu_5$

Bonferroni "correction" for k comparisons:

Replace α with α/k and proceed with t-based CI's.

Bonferroni multiplier:

$$t_{\text{df}}(1 - (\alpha/k)/2)$$

↑ adjusted α

This will be large if k is large, and in that case, it might be better to use Scheffé. (Can compare Bonferroni & Scheffé multipliers and use smaller one.)

Calculating Bonferroni confidence intervals

```
> # 3 Bonferroni intervals
> k <- 3

> (alpha <- 0.05/k) # Set Bonferroni alpha to nominal alpha divided by k.
[1] 0.01666667

> (M <- qt(1-alpha/2, 65)) # Bonferroni multiplier for k=3
[1] 2.457515


> # Sample means
> with(case0601, unlist(lapply(split(Score, Handicap), mean)))
   Amputee   Crutches    Hearing       None Wheelchair
  4.428571   5.921429   4.050000   4.900000   5.342857

> # First Bonferroni CI (none vs. others):
> # point estimate
> (g1 <- 4.9 - (4.429 + 5.9219 + 4.05 + 5.343)/4)
[1] -0.035975
```

$C_4 = 1$ all other C_i's are $1/4$

```
> # standard error
> (SE1 <- sqrt(2.6665) *
+     sqrt(1/14 + (0.25)^2/14 + (0.25)^2/14 + (0.25)^2/14 + (0.25)^2/14))
[1] 0.4879348


> # Lower Bonferroni confidence limit
> (g1 - M*SE1)
[1] -1.235082


> # Upper Bonferroni confidence limit
> (g1 + M*SE1)
[1] 1.163132
```

```
> # Second Bonferroni CI (amputee vs. crutches):
> # point estimate
> (g2 <- 4.429 - 5.921)
[1] -1.492
> # standard error
> (SE2 <- sqrt(2.6665) * sqrt(1/14 + 1/14))
[1] 0.6171941
> # Lower Bonferroni confidence limit
> (g2 - M*SE2)
[1] -3.008764
> # Upper Bonferroni confidence limit
> (g2 + M*SE2)
[1] 0.02476383

> # Third Bonferroni CI (amputee vs. wheelchair)
> # point estimate
> (g3 <- 4.429 - 5.343)
[1] -0.914
> # standard error
> SE3 <- (sqrt(2.6665) * sqrt(1/14 + 1/14))
> # Lower Bonferroni confidence limit
> (g3 - M*SE3)
[1] -2.430764
> # Upper Bonferroni confidence limit
> (g3 + M*SE3)
[1] 0.6027638
```

$\sqrt{\text{Res. MS}}$ from ANOVA table

Coefficients are 1 & −1

Comparison	Lower limit	Upper limit
none vs. others	-1.24	1.16
amputee vs. crutches	-3.01	0.02
amputee vs. wheelchair	-2.43	0.60

Statistical conclusion:

We estimate the comparisons among pop'n means are as shown in the table (95% Bonferroni CI's).

Summary of Simultaneous Confidence Interval Procedures

$$\text{pt est} \pm \text{multiplier} \cdot \text{SE(pt est)}$$

res. MS

For $\gamma = C_1\mu_1 + \ldots + C_I\mu_I$,

$$\text{pt est} = C_1\overline{Y}_1 + \ldots + C_I\overline{Y}_I \quad \text{and} \quad \text{SE(pt est)} = s_p\sqrt{\frac{C_1^2}{n_1} + \ldots + \frac{C_I^2}{n_I}}$$

Procedure	Multiplier	Comments
Tukey–Kramer	$\frac{q_{I,n-I}(1-\alpha)}{\sqrt{2}}$ use R's qtukey() or TukeyHSD()	Useful for making all possible ***pairwise*** comparisons. preplanned
Dunnett	Use `glht()` in the `multicomp` R package. Use relevel() to make control group the first one.	Compare all treatments to control treatment. You get to decide which is control, but before looking at data.
Scheffé	$\sqrt{(I-1)F_{(I-1),\mathrm{df}}(1-\alpha)}$ qf() I = # groups df = res. df from ANOVA table	For all possible comparisons. including data-suggested ones. (not just pairwise)
Bonferroni	$t_{\mathrm{df}}(1-(\alpha/k)/2)$ usual t-quantile but with new α.	For k preplanned comparisons Must be preplanned Comparisons need not be pairwise

What to know about simultaneous inference/multiple comparisons for an exam:

- Understand the simultaneous inference problem.

Every CI has a chance to miss what it's estimating. More CI means more chances. (see p7 of outline 6.)

- Understand what it means to "snoop" data.

Looking at data to decide on a particular comparison or inference.

- Be able to select an appropriate technique for a given situation.

Refer to table on p. 18 on Outline 6.

- Interpret a given a set of simultaneous confidence intervals for a data set.

In particular, for pairwise comparisons, which pairs of pop'n means differ? (Look at which CI's exclude 0.)

DISPLAY 6.7 Common fallacies of reasoning from statistical hypothesis tests

Fallacy name	The fallacy	Avoiding the fallacy
False Causality Fallacy	Incorrectly interpreting statistical significance (i.e., a small *p*-value) from an observational study as evidence of causation	Use the word *association* to indicate a relationship that is not necessarily a causal one.
Fallacy of Accepting the Null	Incorrectly interpreting a lack of statistical evidence that a null hypothesis is false (i.e., a large *p*-value) as statistical evidence that the null hypothesis is true	Avoid this incorrect wording: "the study provides evidence that there is no difference." Say instead: "there is no evidence from this study of a difference." Also, report a confidence interval to emphasize the many possible hypothesized values (in addition to 0) that are consistent with the observed data.
Confusing Statistical for Practical Significance	Interpreting a "statistically significant" effect (which has to do with the strength of evidence that there's an effect) as a practically important one, which it may or may not be	If you must use the term *statistically significant*, don't abbreviate it. Also, report a confidence interval so that the *size* of an effect can be evaluated for its practical importance.
Data Dredging (Fishing for Significance, Data Snooping)	Incorrectly drawing conclusions from an unadjusted *p*-value that emerged from a process of sifting through many possible *p*-values Note: "Publication Bias" is the de facto data dredging that results if journals only accept research papers with statistically significant findings.	For multiple comparisons of means, use the adjustments in this chapter. For identifying a few from many possible predictor variables, use the variable selection methods in Chapter 12. For tests based on many different response variables (data mining), use the False Discovery Rate methods of Chapter 16.
Good Statistics from Bad Data	Incorrectly accepting conclusions based on sound statistical technique when there are problems with data collection, such as biased sampling or data contamination	Critically evaluate the potential biases from non–randomly selected samples.