Multiple Comparisons (Sleuth3 Sections 6.3 and 6.4)

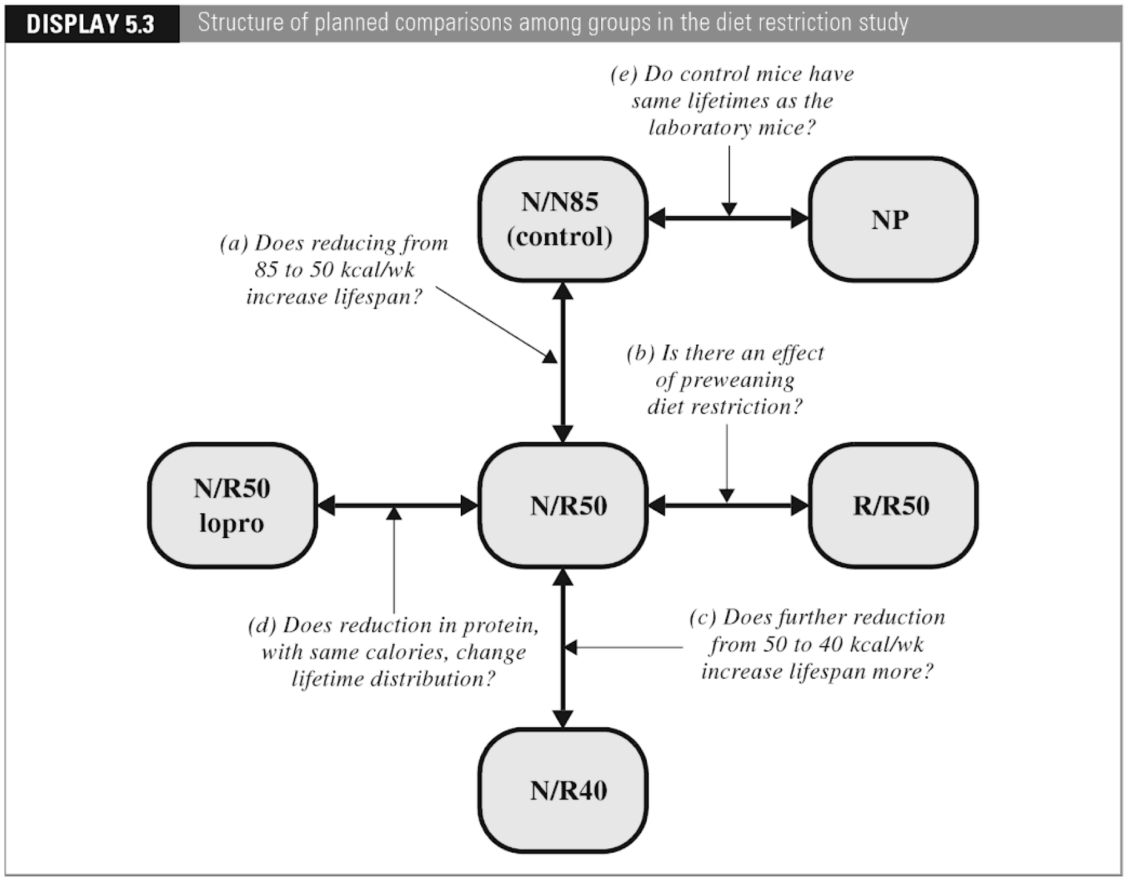
Example 1: Diet restriction and longevity in mice (Sleuth3 Case study 5.1.1)

Mice were randomly assigned to one of 6 treatment groups with different diets to investigate relationships between diet and lifetime. The life span of each mouse was recorded in months.

1. **NP**: Mice ate as much as they wanted of standard food for lab mice
2. **N/N85**: Control group. **N**: no intervention before weaning; ate as normal. **N85**: no intervention after weaning; fed weekly diet of 85kcal/week (standard diet for lab mice)
3. **N/R50**: **N**: no intervention before weaning. **R50**: after weaning, restricted diet of 50 kcal/week
4. **R/R50**: **R**: restricted diet of 50 kcal/week before weaning. **R50**: after weaning, restricted diet of 50 kcal/week
5. **N/R50 lopro**: **N**: no intervention before weaning. **R50**: after weaning, restricted diet of 50 kcal/week. Dietary protein decreased with mouse age.
6. **N/R40**: **N**: no intervention before weaning. **R40**: after weaning, restricted diet of 40 kcal/week

Denote the mean life spans in the population of mice fed each of these diets under laboratory conditions by μ_1 through μ_6.

Planned Comparisons: Before data were collected, researchers decided on the comparisons below:



These comparisons determine 5 hypothesis tests:

(a) $H_0 : \mu_2 = \mu_3$ vs $H_A : \mu_2 \neq \mu_3$. Are the population mean lifetimes the same for the **N/N85** and **N/R50** groups?

(b) $H_0 : \mu_3 = \mu_4$ vs $H_A : \mu_3 \neq \mu_4$. Are the population mean lifetimes the same for the **N/R50** and **R/R50** groups?

(c) $H_0 : \mu_3 = \mu_6$ vs $H_A : \mu_3 \neq \mu_6$. Are the population mean lifetimes the same for the **N/R50** and **N/R40** groups?

(d) $H_0 : \mu_3 = \mu_5$ vs $H_A : \mu_3 \neq \mu_5$. Are the population mean lifetimes the same for the **N/R50** and **N/R50** lopro groups?

(e) $H_0 : \mu_2 = \mu_1$ vs $H_A : \mu_2 \neq \mu_1$. Are the population mean lifetimes the same for the **N/N85** and **NP** groups?

Example 2: Handicaps and hiring (Sleuth3 Case Study 6.1.1 in Sleuth 3)

A 1990 study conducted a randomized experiment to explore how physical handicaps affect people's perception of employment qualifications. The researchers prepared five videotaped job interviews using the same two male actors for each. A set script was designed to reflect an interview with an applicant of average qualifications. The videos differed only in that the applicant appeared with a different handicap:

1. in one, he appeared to have no handicap;
2. in a second, he appeared to have one leg amputated;
3. in a third, he appeared on crutches;
4. in a fourth, he appeared to have impaired hearing;
5. and in a fifth, he appeared in a wheelchair.

Seventy undergraduate students from a US university were randomly assigned to view the videos, fourteen to each video. After viewing ther video, each subject rated the qualifications of the applicant on a 0 to 10 point applicant qualification scale.

Denote by μ_1 through μ_5 the mean qualification score in the population of ratings that might be given by US undergraduate students from the US university in this study for each of the 5 handicaps groups.

"Unplanned" Comparisons: Maybe we want to compare the mean qualification score for every pair of groups

- $H_0 : \mu_1 = \mu_2$ vs $H_A : \mu_1 \neq \mu_2$
- $H_0 : \mu_1 = \mu_3$ vs $H_A : \mu_1 \neq \mu_3$
- $H_0 : \mu_1 = \mu_4$ vs $H_A : \mu_1 \neq \mu_4$
- $H_0 : \mu_1 = \mu_5$ vs $H_A : \mu_1 \neq \mu_5$
- $H_0 : \mu_2 = \mu_3$ vs $H_A : \mu_2 \neq \mu_3$
- $H_0 : \mu_2 = \mu_4$ vs $H_A : \mu_2 \neq \mu_4$
- $H_0 : \mu_2 = \mu_5$ vs $H_A : \mu_2 \neq \mu_5$
- $H_0 : \mu_3 = \mu_4$ vs $H_A : \mu_3 \neq \mu_4$
- $H_0 : \mu_3 = \mu_5$ vs $H_A : \mu_3 \neq \mu_5$
- $H_0 : \mu_4 = \mu_5$ vs $H_A : \mu_4 \neq \mu_5$

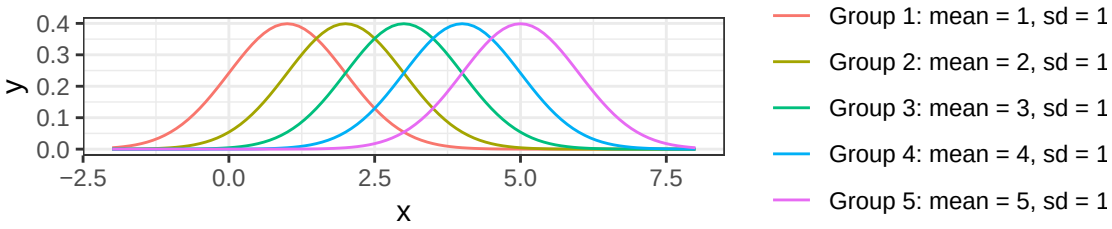
There are 10 different comparisons to do.

Individual Confidence Level vs. Familywise Confidence Level

- Individual confidence level: the proportion of samples for which a single confidence interval contains the parameter it is estimating
- Familywise confidence level: the proportion of samples for which every one of several different confidence intervals contain the parameters they are estimating

Example (simulation study)

Suppose I have 5 groups with means $\mu_1 = 1$, $\mu_2 = 2$, $\mu_3 = 3$, $\mu_4 = 4$, $\mu_5 = 5$ and standard deviation $\sigma = 1$.



Results for 1 simulation

- Simulated a data set with 100 observations from each of the 5 groups
- Calculated 95% confidence intervals for differences in group means, for each pair of means (10 intervals total)

Groups	Difference in Means	95% CI lower bound	95% CI upper bound	Contains true difference?
2, 1	2 - 1 = 1	0.99	1.54	Yes
3, 1	3 - 1 = 2	1.60	2.16	Yes
4, 1	4 - 1 = 3	2.87	3.42	Yes
5, 1	5 - 1 = 4	3.55	4.10	Yes
3, 2	3 - 2 = 1	0.34	0.89	No
4, 2	4 - 2 = 2	1.60	2.15	Yes
5, 2	5 - 2 = 3	2.28	2.84	No
4, 3	4 - 3 = 1	0.99	1.54	Yes
5, 3	5 - 3 = 2	1.67	2.22	Yes
5, 4	5 - 4 = 1	0.41	0.96	No

For this particular sample, 7 out of 10 of the confidence intervals contain the difference in means they are estimating.

Repeated for 1000 simulations:

- Repeated the process above for 1000 different simulated data sets. Table shows:
  - percent of samples for which each CI comparing 2 groups succeded
  - percent of samples for which all 10 CIs succeeded

Groups	Percent of Samples Successful
2, 1	95.1%
3, 1	94.5%
4, 1	95.0%
5, 1	94.5%
3, 2	95.5%
4, 2	95.1%
5, 2	94.8%
4, 3	94.9%
5, 3	95.7%
5, 4	94.4%
All 10 comparisons	71.1%

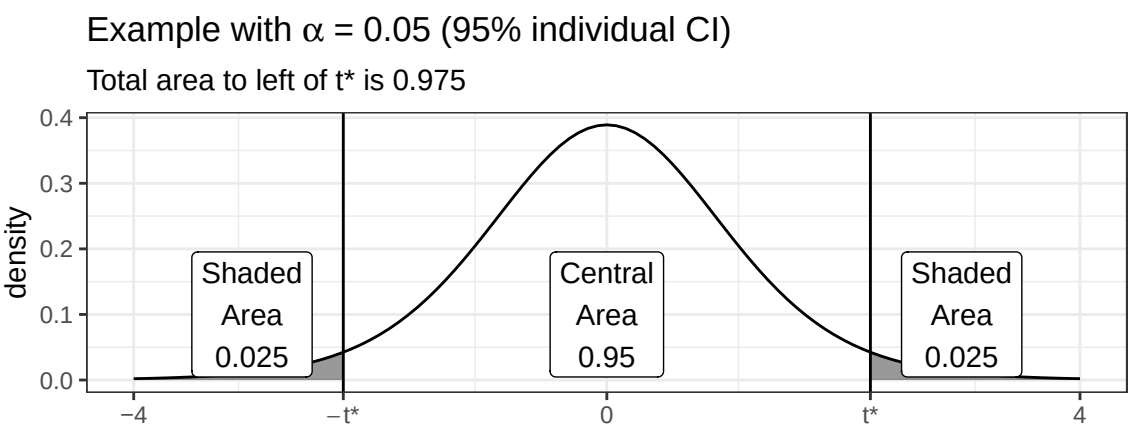
Basic idea: Make individual confidence levels larger to get desired familywise confidence level.

Adjustments for Confidence Intervals

We will discuss just a couple of ideas that are most generally applicable; see the book for many more.

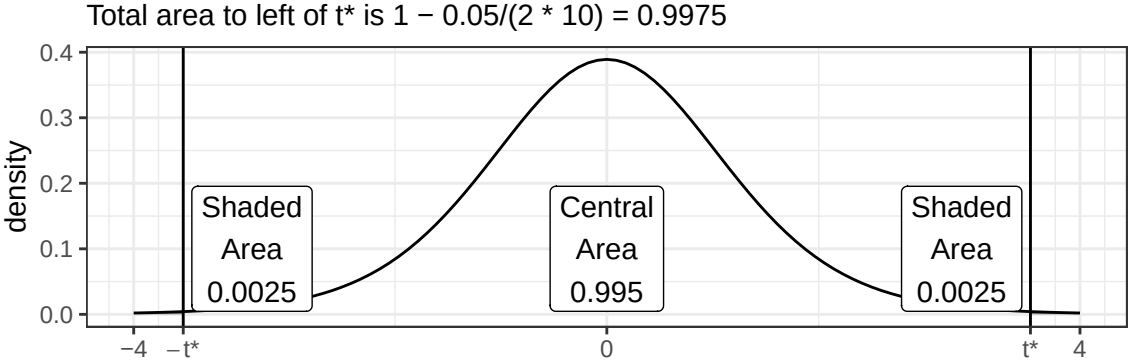
Reminder of standard procedure

- In this class, all confidence intervals are calculated as Estimate $\pm$ Multiplier $\times$ SE(Estimate)
- So far, the Multiplier is $t_{df}(1-\alpha/2)$.
  - For a 95% CI, $\alpha = 0.05$, and $1-\alpha/2 = 0.975$



Bonferroni adjustment

- If there are k confidence intervals to compute, use Multiplier $= t_{df}(1-\alpha/2k)$
- Example: Above, we had $k = 10$ confidence intervals



- Intuition:
  - Each individual CI misses for 0.5% of samples
  - Across all 10, at least one misses for 5% of samples
  - Across all 10, all succeed for 95% of samples

Scheffe adjustment

- Use Multiplier $= \sqrt{(I-1)F_{(I-1),(n-1)}(1-\alpha)}$
- Generally a larger multiplier (wider CIs) than the Bonferroni adjustment
- Works for familywise inferences about every possible linear combination of group means $\gamma = C_1\mu_1 + \cdots + C_I\mu_I$
- Usually not useful for ANOVA, but very useful for regression models, coming soon!

R code, Handicaps Study

```
handicaps <- read_csv("http://www.evanlray.com/data/sleuth3/ex0601_handicaps.csv") %>%
  mutate(
    Handicap = factor(Handicap, levels = c("None", "Amputee", "Crutches", "Hearing", "Wheelchair"))
  )
nrow(handicaps)
```

```
## [1] 70
```

```
head(handicaps)
```

```
## # A tibble: 6 x 2
##   Score Handicap
##   <dbl> <fct>
## 1   1.9 None
## 2   2.5 None
## 3   3   None
## 4   3.6 None
## 5   4.1 None
## 6   4.2 None
```

```
anova_fit <- lm(Score ~ Handicap, data = handicaps)
```

Individual 95% CIs

```
fit.contrast(anova_fit, "Handicap", c(-1, 1, 0, 0, 0), conf.int = 0.95)
```

```
##                             Estimate Std. Error    t value  Pr(>|t|)  lower CI  upper CI
## Handicap c=( -1 1 0 0 0 ) -0.4714286  0.6171922 -0.7638278 0.4477337 -1.704047 0.7611894
## attr(,"class")
## [1] "fit_contrast"
```

```
fit.contrast(anova_fit, "Handicap", c(-1, 0, 1, 0, 0), conf.int = 0.95)
```

```
##                            Estimate Std. Error t value  Pr(>|t|)   lower CI upper CI
## Handicap c=( -1 0 1 0 0 ) 1.021429  0.6171922 1.65496 0.1027537 -0.2111894 2.254047
## attr(,"class")
## [1] "fit_contrast"
```

. . . and so on.

Confirming how the CIs were calculated from the estimates and standard errors:

```
t_star <- qt(0.975, df = 70 - 5)
t_star
```

```
## [1] 1.997138
```

```
-0.4714286 - t_star * 0.6171922
```

```
## [1] -1.704047
```

```
-0.4714286 + t_star * 0.6171922
```

```
## [1] 0.7611893
```

```
1.021429 - t_star * 0.6171922
```

```
## [1] -0.2111889
```

```
1.021429 + t_star * 0.6171922
```

```
## [1] 2.254047
```

Bonferroni 95% familywise CIs

First, find multiplier (note that it's larger!)

```
bonferroni_multiplier <- qt(0.9975, df = 70 - 5)
bonferroni_multiplier
```

```
## [1] 2.906015
```

```
-0.4714286 - bonferroni_multiplier * 0.6171922
```

```
## [1] -2.264999
```

```
-0.4714286 + bonferroni_multiplier * 0.6171922
```

```
## [1] 1.322141
```

```
1.021429 - bonferroni_multiplier * 0.6171922
```

```
## [1] -0.772141
```

```
1.021429 + bonferroni_multiplier * 0.6171922
```

```
## [1] 2.814999
```

Scheffe 95% familywise CIs

First find the multiplier - note that it's even larger!

```
scheffe_multiplier <- sqrt((5 - 1) * qf(0.95, df1 = 5 - 1, df2 = 70 - 5))
scheffe_multiplier
```

```
## [1] 3.170514
```

```
-0.4714286 - scheffe_multiplier * 0.6171922
```

```
## [1] -2.428245
```

```
-0.4714286 + scheffe_multiplier * 0.6171922
```

```
## [1] 1.485388
```

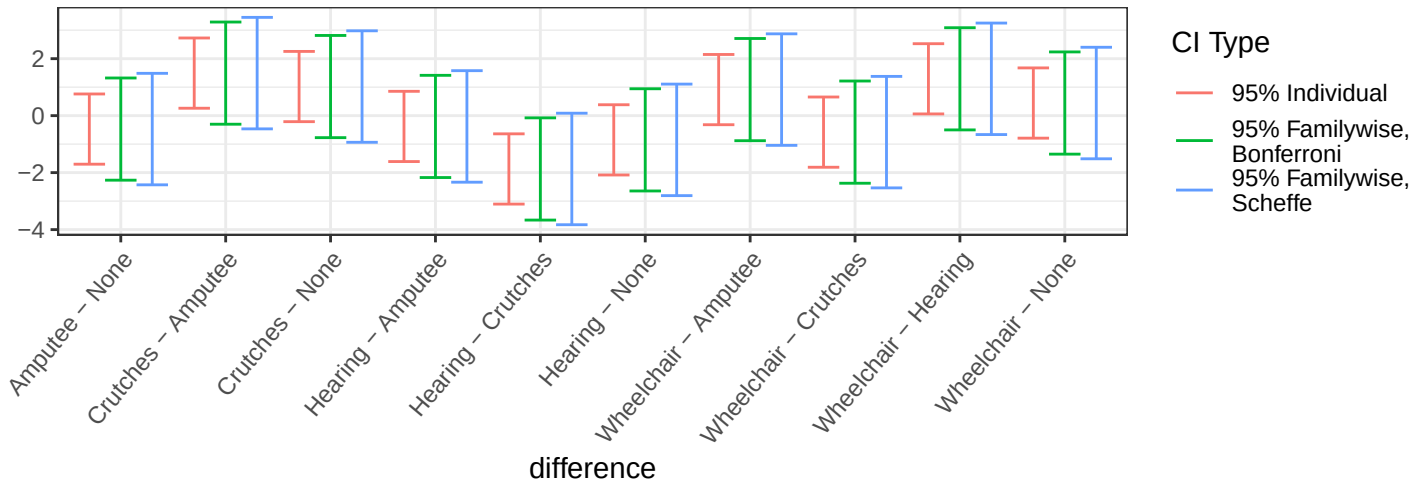
```
1.021429 - scheffe_multiplier * 0.6171922
```

```
## [1] -0.9353876
```

```
1.021429 + scheffe_multiplier * 0.6171922
```

```
## [1] 2.978246
```

All 10 CIs plotted for each method



Similar ideas for hypothesis tests

- p-value = probability of obtaining a test statistic at least as extreme as the value of that statistic we got in our sample data, if H_0 is true **in a single test**
  - Provides a measure of strength of evidence against H_0 for that test.
- Imagine that we do 10 hypothesis tests.
- The chance of obtaining a small p-value ("statistically significant result") in at least one of the tests is larger than the chance of obtaining a small p-value in a single test.

Idea 1

- Adjust our standard of evidence (require smaller p-values to believe something is going on)
  - Could do a Bonferroni like adjustment, require 10 times as much evidence for each test if doing 10 tests.
- In practice
  - Skip calculation of p-values
  - Get familywise confidence intervals for all parameters/comparisons of interest
  - Declare a result "statistically significant" if the familywise CI does not contain hypothesized values

Idea 2 (very common, not perfect)

- Conduct an F test of $H_0 : \mu_1 = \mu_2 = \ldots = \mu_I$ vs H_A : at least one mean is different from the others
  - If p-value from the F test is small (say, less than 0.05), proceed to look at individual results, typically using unadjusted intervals/p-values
  - If p-value from the F test is large (say, greater than 0.05), declare no individual results statistically significant, even if some individual t tests had small p-values.

When to bother?

Opinions differ

- Book says:
  - if tests are "planned", no need to adjust for multiple comparisons
  - if tests are "unplanned", adjust
- Some people say you should always adjust for multiple comparisons
- I say you need to understand the issues and report what you are doing:
  - **Familywise confidence levels can be much less than individual confidence levels**
  - **Report whether or not you have adjusted for multiple comparisons**
  - **Report all confidence intervals/hypothesis tests you perform**, whether or not the results are "statistically significant". **Reporting only statistically significant results is cheating.**
  - To the extent possible, **plan your analysis before collecting data**, and keep number of planned comparisons small