Stat 411/511

Multiple Comparisons

Nov 16 2015

Charlotte Wickham

stat511.cwick.co.nz

Thanksgiving week

No lab material **next week 11/24 & 11/25**.

Labs as usual this week.

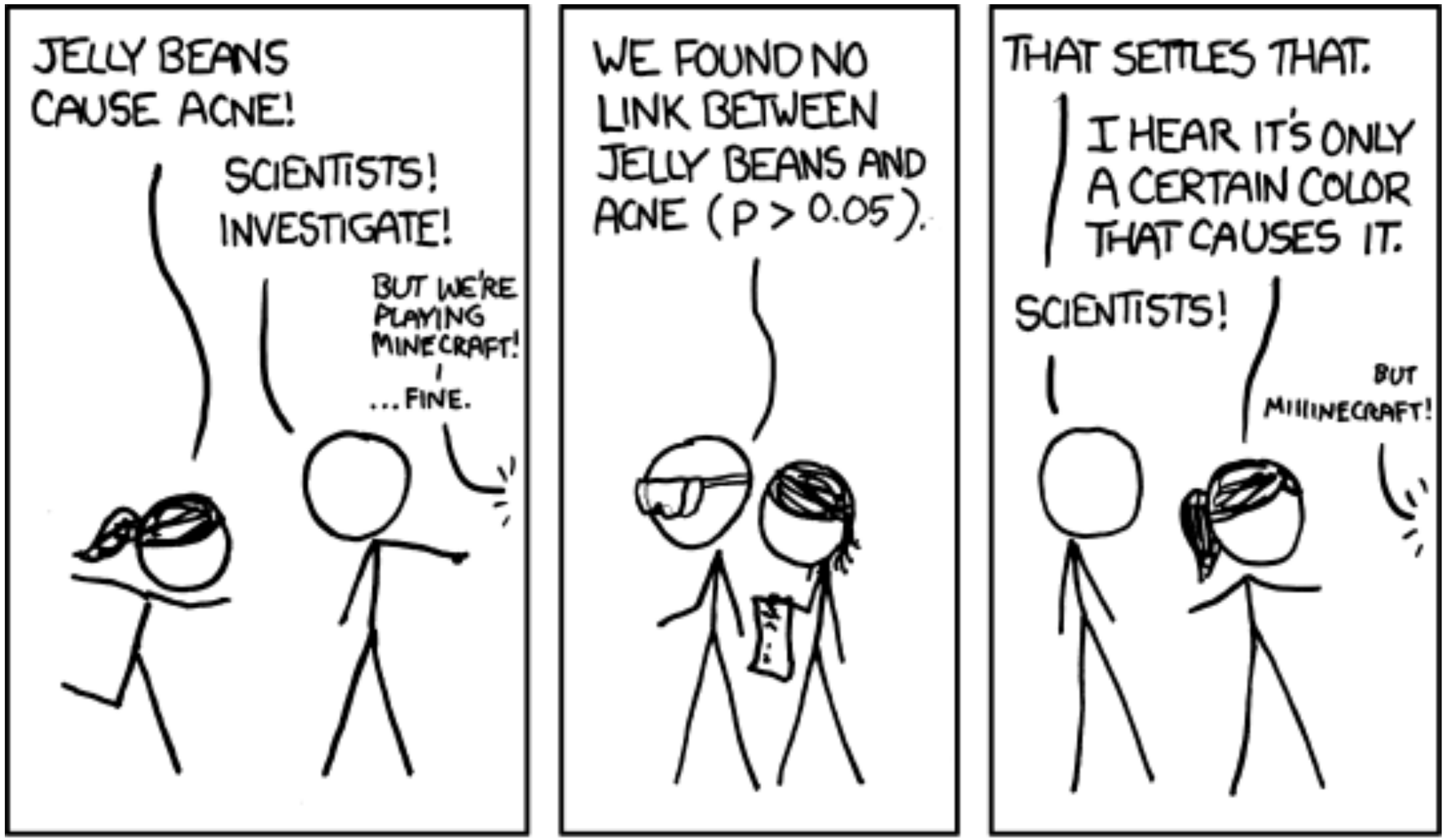
Lectures as usual Mon & Weds next week.

Preplanned comparisons: a few comparisons that directly answer the questions of interest.

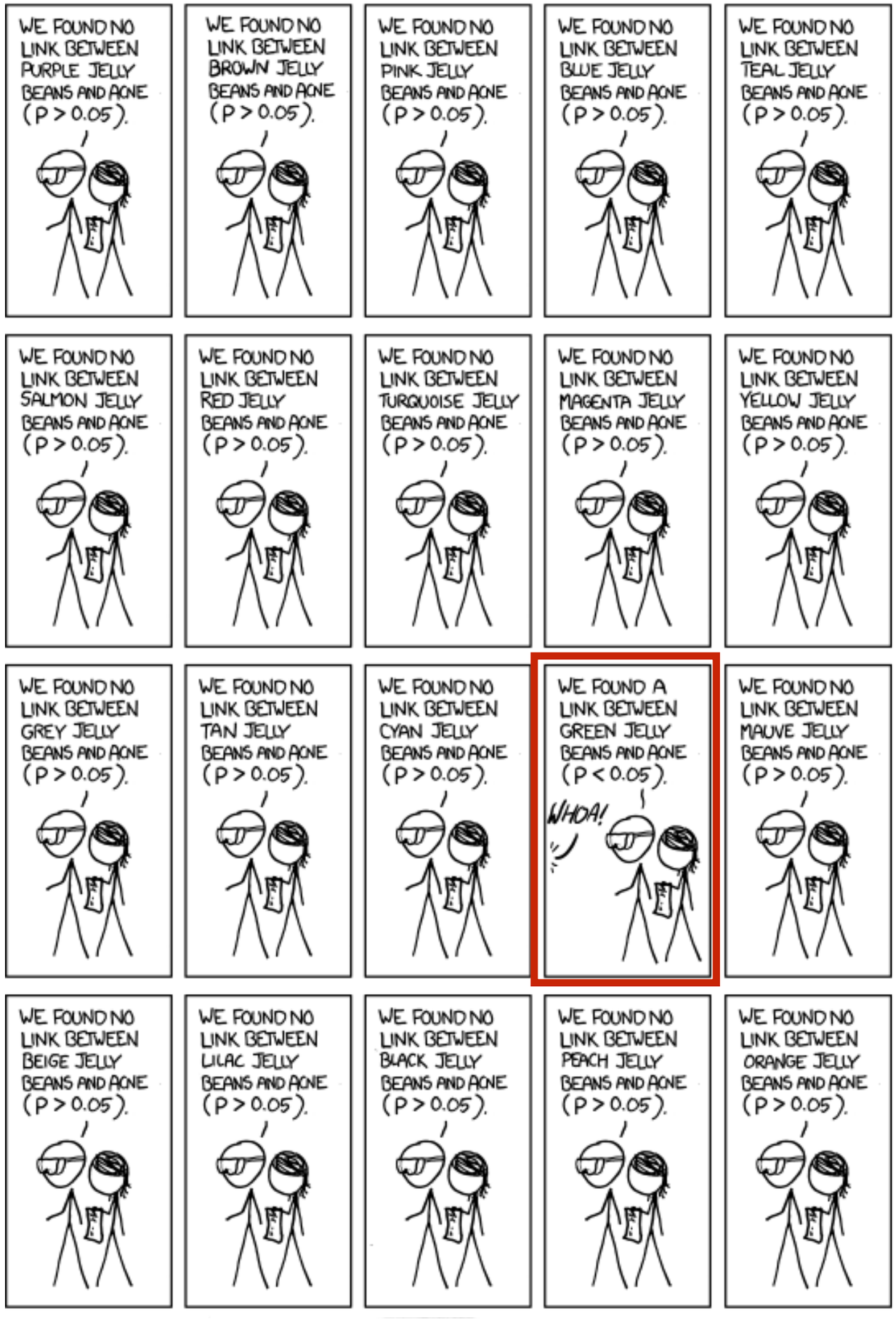
Last Friday

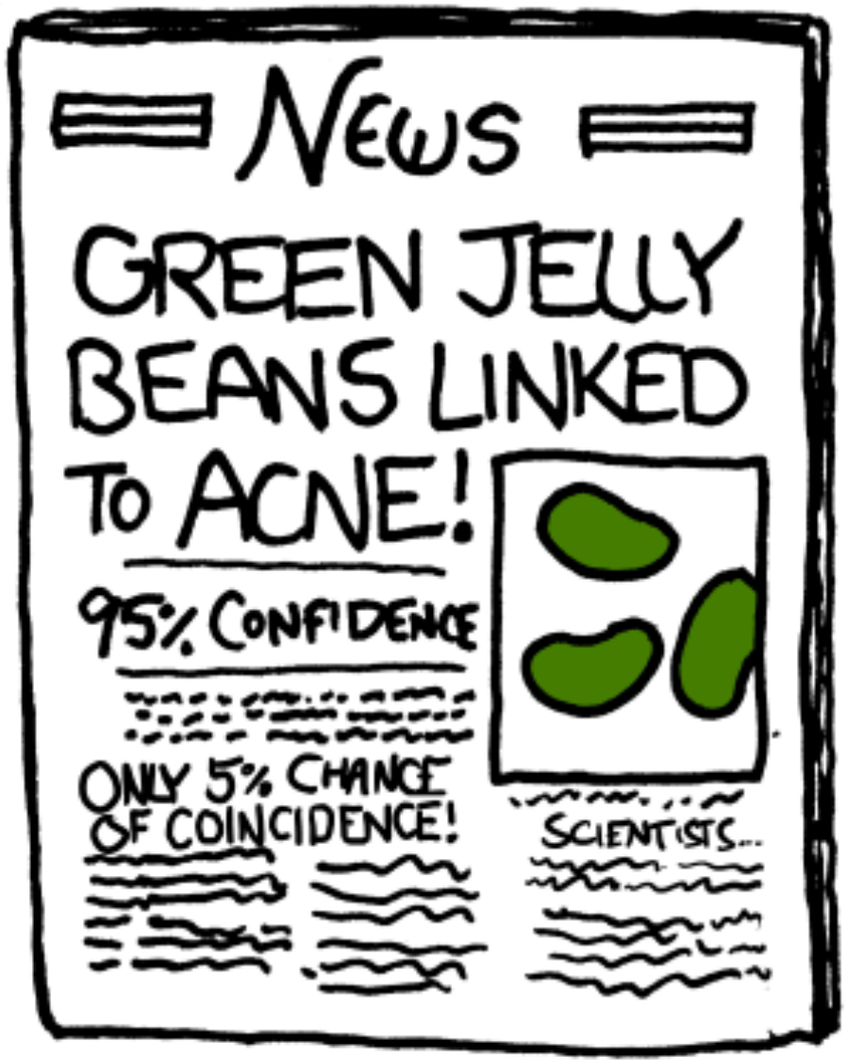
Unplanned comparisons: many comparisons are of interest, you often aren't sure which until you see the data.

Today



http://xkcd.com/882/





Even when there is nothing to find, the probability we find **something** increases the more comparisons we make.

AKA p-hacking

http://www.nature.com/news/scientific-method-statistical-errors-1.14700

http://www.slate.com/articles/health_and_science/science/2013/07/statistics_and_psychology_multiple_comparisons_give_spurious_results.html

We will focus on a very small area:
Multiple comparisons that arise from performing many tests and only reporting significant ones, particularly in ANOVA models (or regression models).

What is the problem?

Individual error rate: the probability of incorrectly rejecting the null hypothesis in a single test, α.

Familywise (or experimentwise) error rate: the probability of incorrectly rejecting **at least one** null hypothesis in a family of tests, α_E.

If $\alpha = 0.05$, $\alpha_E >= 0.05$, and α_E gets bigger the more comparisons you make.

Testing all two group comparisons with `I` groups

when all nulls are true.

I.e. data is simulated to have no differences in means

I	No. of tests $[\ I(I-1)/2\ ]$	Prob. at least one significant $(\ \alpha_E\ )$
2	1	0.05
3	3	0.11
5	10	0.28
10	45	0.64

With 5 groups, even though there are no true differences in mean, we will get at least one significant test about 28% of the time

All pairwise differences in means for the disability study

```
95% confidence level                                  Lower   Upper
                                  Estimate lwr        upr
Amputee - None == 0              -0.47143 -1.70405  0.76119
Crutches - None == 0              1.02143 -0.21119  2.25405
Hearing - None == 0              -0.85000 -2.08262  0.38262
Wheelchair - None == 0            0.44286 -0.78976  1.67548
Crutches - Amputee == 0           1.49286  0.26024  2.72548
Hearing - Amputee == 0           -0.37857 -1.61119  0.85405
Wheelchair - Amputee == 0         0.91429 -0.31833  2.14690
Hearing - Crutches == 0          -1.87143 -3.10405 -0.63881
Wheelchair - Crutches == 0       -0.57857 -1.81119  0.65405
Wheelchair - Hearing == 0         1.29286  0.06024  2.52548
```

Same idea for confidence intervals

Intro to multiple comparisons

Individual confidence level: success rate of the CI procedure for a single interval.

For a 95% CI, individual success rate **= 95%**

Familywise (or experimentwise) confidence level: success rate of the CI procedure for a family of intervals, where success is all intervals capture their true parameter.

For a collection of 95% CIs, family success rate **< 95%**

Multiple comparison procedures

Attempt to control the familywise error rate and familywise confidence level.

For tests: increase the p-value the more tests we do, "adjusted p-values"

OR

decrease the significance level, the more tests we do.

For confidence intervals: make the intervals wider, the more comparisons we make.

Still of the form: estimate ± multiplier x SE

change the multiplier

Least significant difference

No adjustment

These are what we obtain from applying the usual t-tools and confidence intervals.

95%CI: $(\bar{Y}_2 - \bar{Y}_1) \pm qt(0.975, d.f.) \times SE_{\bar{Y}2 - \bar{Y}1}$

Estimate ± Multiplier x $SE_{estimate}$

```
library(multcomp)
full_model <- lm(Score ~ Handicap, data
= case0601)
comparisons <- glht(full_model,
  linfct = mcp(Handicap = "Tukey"))

# LSD tests, usual two group comparisons
summary(comparisons,
    test = adjusted("none"))

# LSD confidence intervals
confint(comparisons,
    calpha = univariate_calpha())
```

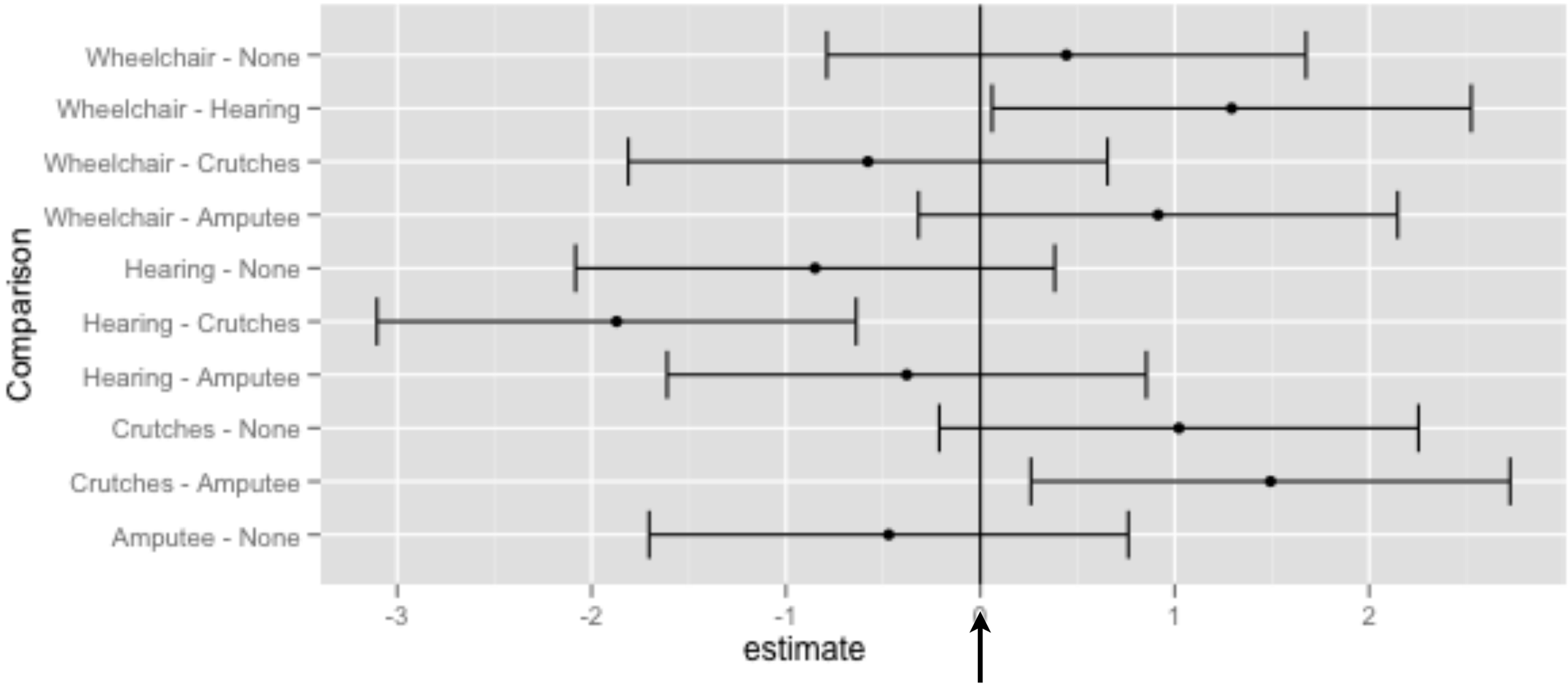
set up
all pairwise comparisons

To make sure we get LSD
i.e. do no adjustment

```
Linear Hypotheses:
                             Estimate Std. Error t value Pr(>|t|)
Amputee - None == 0           -0.4714     0.6172  -0.764  0.44773
Crutches - None == 0           1.0214     0.6172   1.655  0.10275
Hearing - None == 0           -0.8500     0.6172  -1.377  0.17317
Wheelchair - None == 0         0.4429     0.6172   0.718  0.47561
Crutches - Amputee == 0        1.4929     0.6172   2.419  0.01838 *
Hearing - Amputee == 0        -0.3786     0.6172  -0.613  0.54177
Wheelchair - Amputee == 0      0.9143     0.6172   1.481  0.14334
Hearing - Crutches == 0       -1.8714     0.6172  -3.032  0.00349 **
Wheelchair - Crutches == 0    -0.5786     0.6172  -0.937  0.35201
Wheelchair - Hearing == 0      1.2929     0.6172   2.095  0.04010 *
```

```
95% confidence level
                              Estimate lwr      upr
Amputee - None == 0          -0.47143 -1.70405  0.76119
Crutches - None == 0          1.02143 -0.21119  2.25405
Hearing - None == 0          -0.85000 -2.08262  0.38262
Wheelchair - None == 0        0.44286 -0.78976  1.67548
Crutches - Amputee == 0       1.49286  0.26024  2.72548
Hearing - Amputee == 0       -0.37857 -1.61119  0.85405
Wheelchair - Amputee == 0     0.91429 -0.31833  2.14690
Hearing - Crutches == 0      -1.87143 -3.10405 -0.63881
Wheelchair - Crutches == 0   -0.57857 -1.81119  0.65405
Wheelchair - Hearing == 0     1.29286  0.06024  2.52548
```

The LSD confidence intervals in a plot

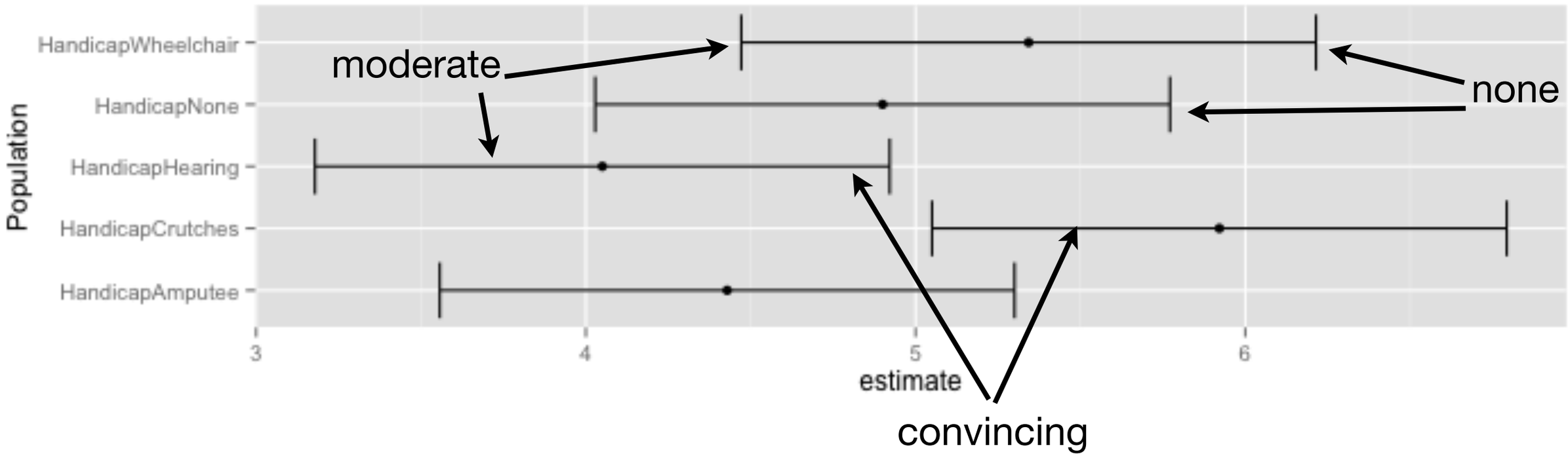


If the interval **doesn't** include zero, the p-value for the test for equal means would be < 0.05

i.e. moderate/convincing evidence of different means

An aside

Sometimes people just show the estimates of the group means,
rather than all the possible differences



The only way to tell how much evidence there is
against equal means
for two overlapping intervals
is to do the test on the difference

http://www.nature.com/nmeth/journal/v10/n10/full/nmeth.2659.html?WT.ec_id=NMETH-201310

Procedures specifically for adjusting comparisons between means of multiple groups

Dunnett

Designed to control the familywise error rate when making difference in mean comparisons between the **one group and all the other groups**.

Tukey-Kramer

Designed to control the familywise error rate when making **all pairwise difference in mean comparisons.**

Scheffé

Designed to control the familywise error rate when making all possible **linear contrasts of means.**

Your turn

The adjustments make the confidence intervals wider the more comparisons we make.

Which adjustment would you expect to give wider intervals?

Tukey-Kramer, Dunnett or Scheffe?

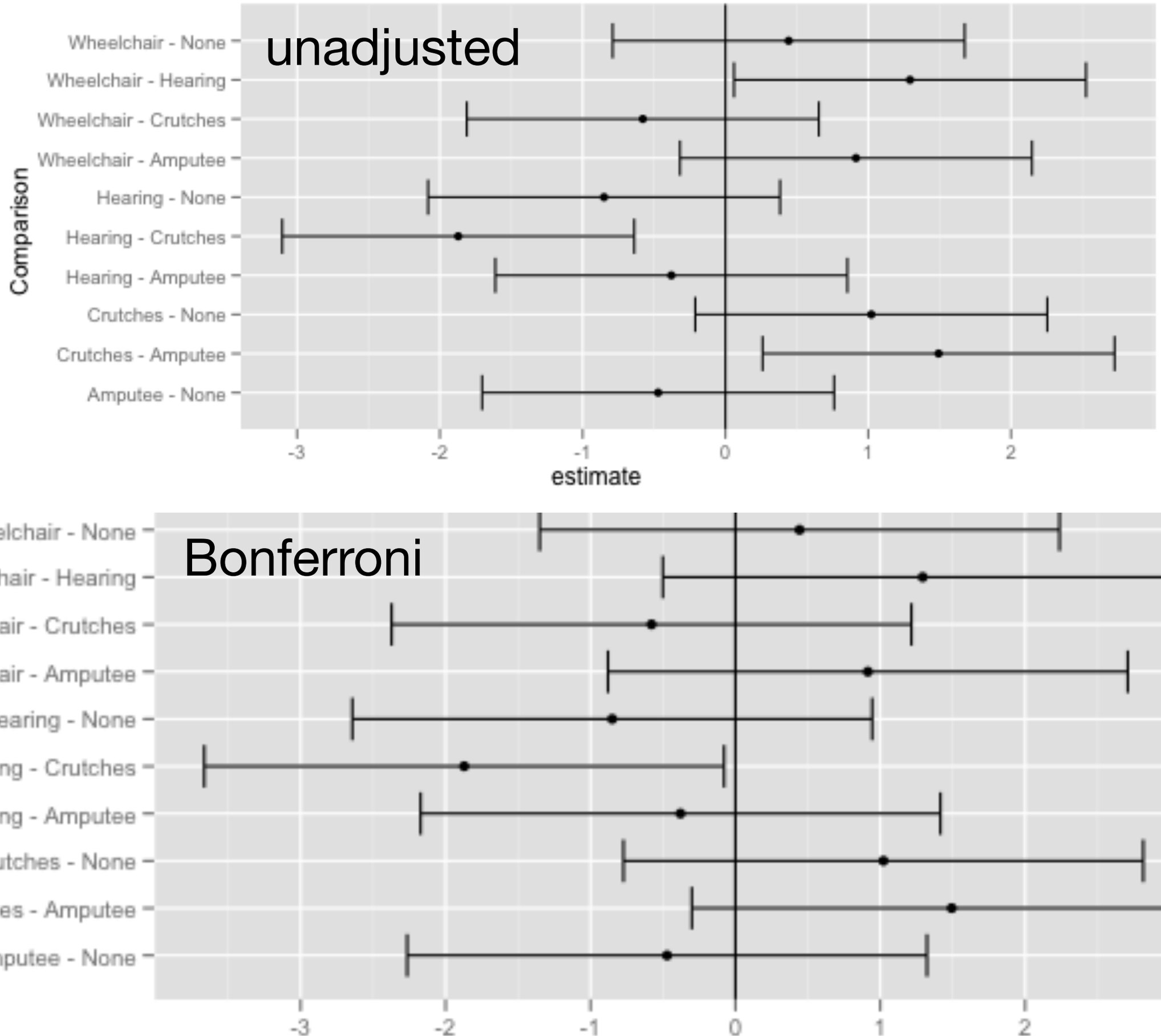
Bonferroni Adjustment

for **any** set of k comparisons

Application of a mathematical result places a bound on the familywise error rate.

Bonferroni adjustment guarantees our familywise error rate is at most 5%.

For k comparisons, adjust the significance level down to 0.05/k, and confidence level up to (100*(1- 0.05/k)).


unadjusted
Wheelchair - None
Wheelchair - Hearing
Wheelchair - Crutches
Wheelchair - Amputee
Hearing - None
Hearing - Crutches
Hearing - Amputee
Crutches - None
Crutches - Amputee
Amputee - None
Comparison
-3
-2
-1
0
1
2
estimate
Bonferroni
Wheelchair - None
Wheelchair - Hearing
Wheelchair - Crutches
Wheelchair - Amputee
Hearing - None
Hearing - Crutches
Hearing - Amputee
Crutches - None
Crutches - Amputee
Amputee - None
Comparison
-3
-2
-1
0
1
2
3

Choosing an adjustment

In the multiple group setting when you are interested in means (or linear combinations of means), the appropriate adjustment depends on the set of **interesting comparisons before you see the data**, not the comparisons you actually report.

For example, imagine you have five treatments and you are interested in which are most different. You calculate all possible two group comparisons and find treatment 1 and treatment 3 are the most different and report only that CI in your report. Tukey-Kramer would be an appropriate adjustment, because you considered all pairwise comparisons.

If you examine your data for the linear combination that gives the smallest p-value, Scheffe would be the appropriate adjustment.

The problems with multiple comparisons

How do you define an experiment?

Do you want to control the familywise error rate:

- in this experiment?
- in all research you do on this topic?
- in all tests in your career?

This is a controversial area of statistics.

Always report how many comparisons you planned to do (This includes the case where you look at your data first, and only choose to test the differences that look big)

There are alternatives to controlling familywise error rate, e.g. control the false discovery rate.

In R

If you can specify all the comparisons you are interested in beforehand, the `multcomp` package will do the adjustment for you.

There are shortcut's for Tukey-Kramer, and Dunnet.

```
# all pairwise comparisons
comparisons <- glht(full_model,
                    linfct = mcp(Handicap = "Tukey"))

# LSD tests, usual two group comparisons
summary(comparisons,
        test = adjusted("none"))

# LSD confidence intervals, aka Unadjusted intervals
confint(comparisons,
        calpha = univariate_calpha())
qplot(lhs, estimate, data = confint(comparisons,
        calpha = univariate_calpha()),
      geom = "pointrange", ymin = lwr, ymax = upr) +
  coord_flip() + geom_hline(yintercept = 0)

# Tukey Kramer
summary(comparisons)
confint(comparisons)
qplot(lhs, estimate, data = confint(comparisons),
      geom = "pointrange", ymin = lwr, ymax = upr) +
  coord_flip()+ geom_hline(yintercept = 0)

# Dunnett
dunnett <- glht(full_model,
                    linfct = mcp(Handicap = "Dunnett"))
summary(dunnett)
confint(dunnett)
qplot(lhs, estimate, data = confint(dunnett),
      geom = "pointrange", ymin = lwr, ymax = upr) +
  coord_flip() + geom_hline(yintercept = 0)
```