STAT 213
Multiple Comparisons and the Family-Wise Error Rate

Colin Reimer Dawson

Oberlin College

April 25, 2018

Outline

Outline

Overall Test of the Model

Null Population Model:

$$Y_i = \mu + \varepsilon$$

Groups Population Model:

$$Y_i = \mu + \alpha_{X_i} + \varepsilon$$

$H_0 : \alpha_X \equiv 0$ for all X H_1 : some $\alpha_X \neq 0$

Example: Stereotype Threat and Student Athletes

The term "stereotype threat" refers to a phenomenon whereby reminders of particular components of an individual's identity (race, gender, ethnicity) can result in the individual conforming to stereotypes about that group. For example, women perform worse on a math test after being reminded of their gender (Spencer et al., 1999). Some researchers (Steele, 1997) believe this is due to anxiety about the possibility of confirming negative stereotypes. Yopyk and Prentice (2005) administered a math test to student-athletes after either (A) reminding them of their athlete status, (B) reminding them of their student status, or (C) not reminding them of either component of their identity. The test scores had the following mean and standard deviations.

Example: Stereotype Threat and Student Athletes

	Athlete Prime	No Prime	Student Prime
n	12	13	12
$\bar{x}$	66.97	82.46	86.17
s	5.60	4.99	4.58

Source	df	SS	MS	F	P-value
Prime	2	2504.38	1252.19	48.68	1.05e-10
Residuals	34	874.5	25.72	–	–
Total	36	3378.88	–	–	–

Post-Hoc Comparisons

Once we determine that there is evidence for differences, we want to be able to say which groups differ.

- Could just refit separate models on each pair of groups...
- But if we take seriously the idea that the variability is the same for each group, we can gain efficiency by using that fact
- Instead of estimating within group variability separately each time, estimate it once using all the data: this is our MSE
- This governs our standard error for each comparison/interval

Post-Hoc Comparison of Pairs of Groups

```
library(DescTools) # may need to install this
aov.model <- aov(Score ~ Prime, data = StudentAthletes)
summary(aov.model)
```

```
             Df Sum Sq Mean Sq F value   Pr(>F)
Prime         2 2504.4  1252.2   48.68 1.05e-10 ***
Residuals    34  874.5    25.7
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
PostHocTest(aov.model, conf.level = 0.95, method = "lsd")
```

```
  Posthoc multiple comparisons of means : Fisher LSD
    95% family-wise confidence level

$Prime
                  diff     lwr.ci    upr.ci    pval
None-Athlete    15.49 11.3640448 19.615955 7.2e-09 ***
Student-Athlete 19.20 14.9923348 23.407665 7.8e-11 ***
Student-None     3.71 -0.4159552  7.835955  0.0764 .
```

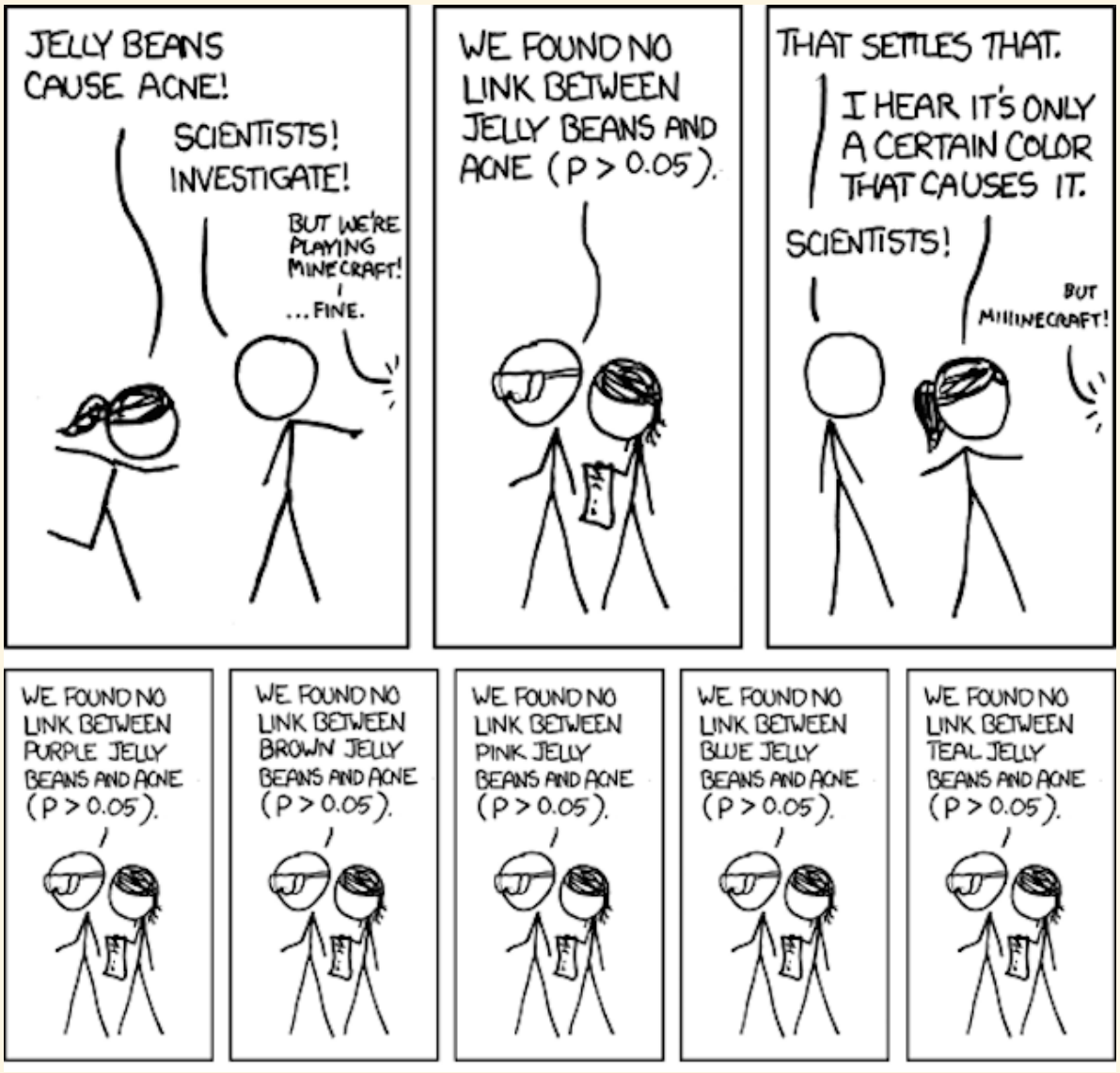
Many Comparisons

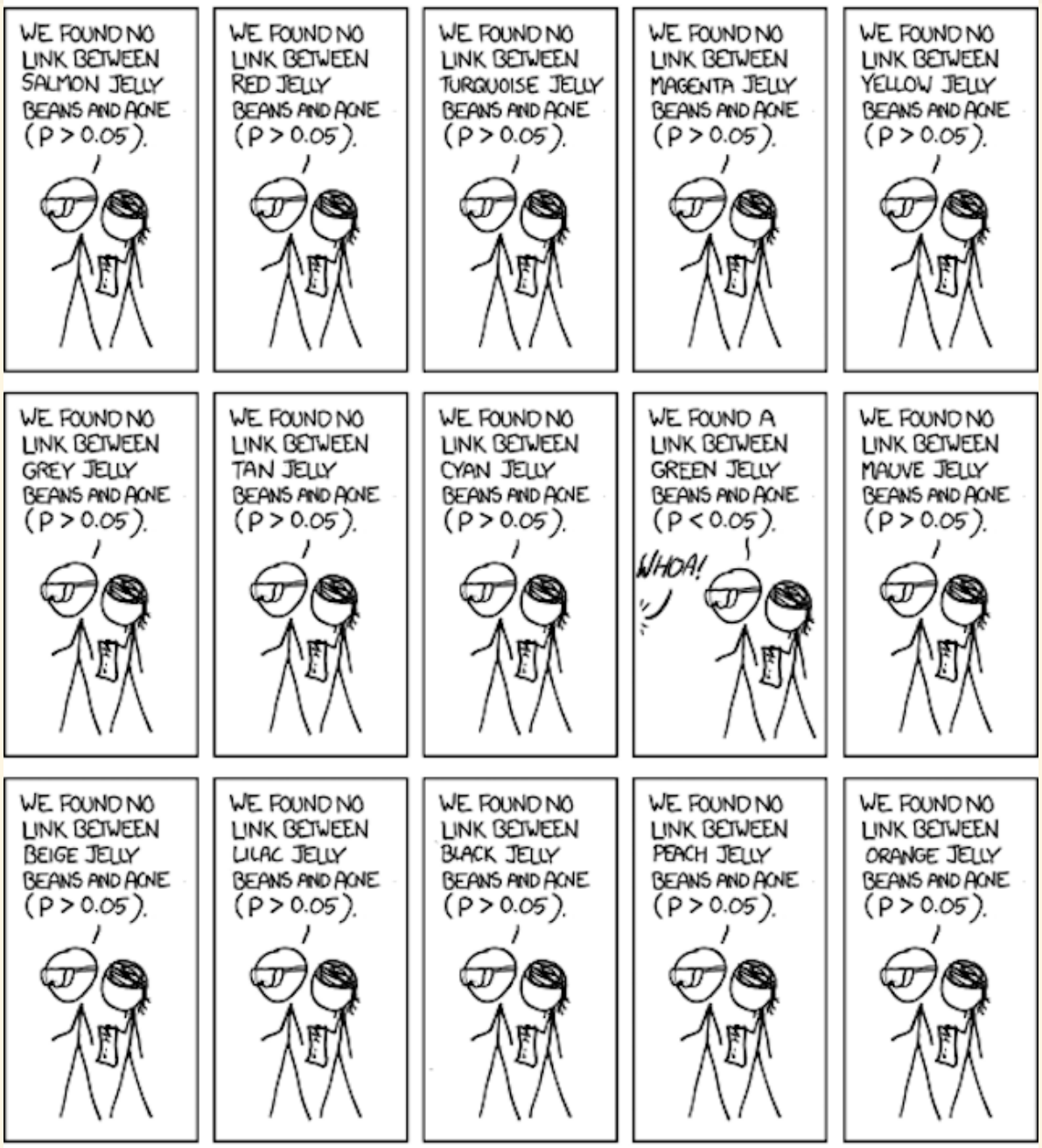
- What happens if we have a lot of possible comparisons we can make?
- Handout

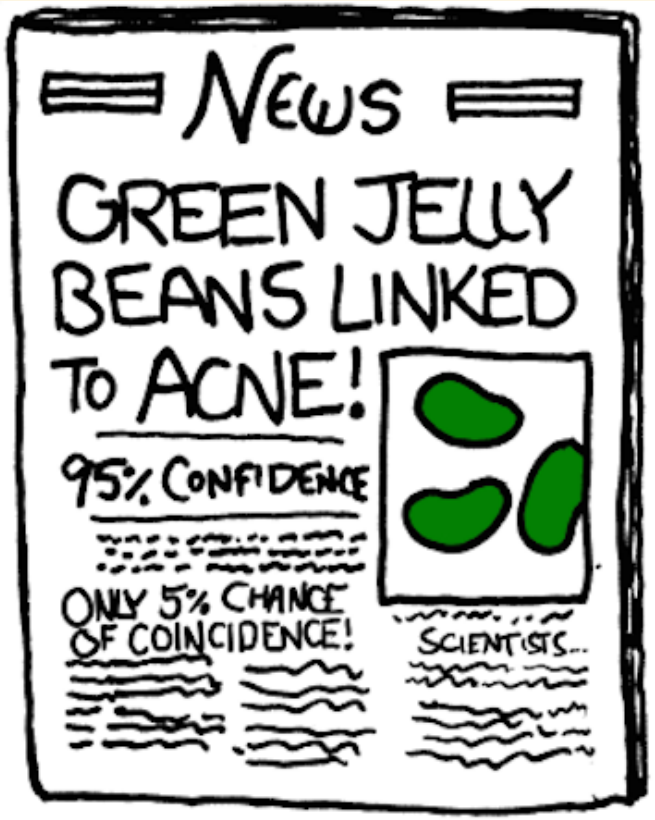
Outline

Familywise Error Rate

- Each test has a probability α of yielding a Type I Error.
- The probability that we make *at least one* Type I Error is called the **family-wise error rate** (FWER).
- Can be much greater than α if no adjustment is made.


JELLY BEANS CAUSE ACNE!
SCIENTISTS! INVESTIGATE!
BUT WE'RE PLAYING MINECRAFT!
... FINE.
WE FOUND NO LINK BETWEEN JELLY BEANS AND ACNE (P > 0.05).
THAT SETTLES THAT.
I HEAR IT'S ONLY A CERTAIN COLOR THAT CAUSES IT.
SCIENTISTS!
BUT MIIINECRAFT!
WE FOUND NO LINK BETWEEN PURPLE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BROWN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PINK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLUE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TEAL JELLY BEANS AND ACNE (P > 0.05).


WE FOUND NO LINK BETWEEN SALMON JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN RED JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TURQUOISE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN MAGENTA JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN YELLOW JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN GREY JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN CYAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND A LINK BETWEEN GREEN JELLY BEANS AND ACNE (P < 0.05).
WHOA!
WE FOUND NO LINK BETWEEN MAUVE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BEIGE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN LILAC JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLACK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PEACH JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN ORANGE JELLY BEANS AND ACNE (P > 0.05).


NEWS
GREEN JELLY BEANS LINKED TO ACNE!
95% CONFIDENCE
ONLY 5% CHANCE OF COINCIDENCE!
SCIENTISTS...

Controlling Family-wise Error rate

Three methods:

1. Fisher's Least Significant Difference (LSD)
2. Tukey's Honestly Significant Difference (HSD)
3. Bonferroni adjustment

Fisher's LSD

- Idea: Use F-test as a "filter"; don't do any pairwise comparisons if F-test is nonsignificant.
- If F is significant, proceed with tests/intervals as discussed, using MSE.
- The most "liberal" of the three methods (more false discoveries/Type I Errors, fewer missed discoveries/Type II Errors)
- Controls probability of finding some difference when there are none, but not probability of finding *too many* differences.

Bonferroni Correction

- Idea: Divide α by the number of comparisons, M being made, then report significant differences for $P < \alpha/M$ (equivalently, multiply P by M and use original α as threshold) and use $1 - \alpha/M$ confidence intervals for differences.
- The most "conservative" of the three methods (guarantees probability of at least one Type I Error does not exceed α, but may be much less, at the cost of more Type II Errors)

Tukey's HSD

- Idea: Use the distribution of $\bar{y}_{max} - \bar{y}_{min}$ under H_0 to see how big the biggest pairwise difference is likely to be by chance alone.
- Any difference bigger than the $1 - \alpha$ quantile of this distribution is declared significant.
- Has exact FWER α if sample sizes are equal (and standard conditions all satisfied); otherwise is somewhat conservative.

Anxiety and Cognitive Functioning

Is there a relationship between anxiety levels and cognitive functioning? A collection of variables pertaining to cognitive and emotional status, sleep habits, and academic habits were collected from 253 college students. One of these, `AnxietyStatus` classifies students according to whether they have `Normal`, `Moderate`, or `Severe` anxiety. Another, `CognitionZscore`, measures (standardized) performance on a test of cognitive skills.

In R

```
library("Lock5Data"); library("mosaic")
data("SleepStudy")
m <- aov(CognitionZscore ~ AnxietyStatus, data = SleepStudy)
summary(m)
```

```
                 Df Sum Sq Mean Sq F value Pr(>F)
AnxietyStatus     2   2.87  1.4368    2.92 0.0558 .
Residuals       250 123.03  0.4921
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Tukey's HSD

```
library("DescTools") ## Need to install first
PostHocTest(m, conf.level = 0.90, method = "hsd", ordered = TRUE)
```

```
  Posthoc multiple comparisons of means : Tukey HSD
    90% family-wise confidence level
    factor levels have been ordered

$AnxietyStatus
                      diff      lwr.ci    upr.ci   pval
normal-moderate 0.2371281  0.01596592 0.4582902 0.0713 .
severe-moderate 0.3579464 -0.05205195 0.7679448 0.1717
severe-normal   0.1208184 -0.25640947 0.4980462 0.7867

---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Fisher's LSD

```
library("DescTools") ## Need to install first
PostHocTest(m, conf.level = 0.90, method = "lsd", ordered = TRUE)
```

```
  Posthoc multiple comparisons of means : Fisher LSD
    90% family-wise confidence level
    factor levels have been ordered

$AnxietyStatus
                      diff      lwr.ci    upr.ci   pval
normal-moderate 0.2371281  0.06003120 0.4142249 0.0280 *
severe-moderate 0.3579464  0.02963786 0.6862550 0.0731 .
severe-normal   0.1208184 -0.18124900 0.4228857 0.5096

---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Bonferroni

```
library("DescTools") ## Need to install first
PostHocTest(m, conf.level = 0.90, method = "bonferroni", ordered = TRUE)
```

```
  Posthoc multiple comparisons of means : Bonferroni
    90% family-wise confidence level
    factor levels have been ordered

$AnxietyStatus
                      diff       lwr.ci    upr.ci   pval
normal-moderate 0.2371281  0.007587509 0.4666686 0.0839 .
severe-moderate 0.3579464 -0.067584165 0.7834770 0.2192
severe-normal   0.1208184 -0.270700212 0.5123370 1.0000

---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Chronological Rejuvenation

Simmons, et al. (2011)

Having demonstrated [in Study 1] that listening to a children's song makes people feel older, Study 2 investigated whether listening to a song about older age makes people actually younger.

Using the same method as in Study 1, we asked 20 University of Pennsylvania undergraduates to listen to either "When I'm Sixty-Four" by The Beatles or "Kalimba". Then, in an ostensibly unrelated task, they indicated their birth date (mm/dd/ yyyy) and their father's age. We used father's age to control for variation in baseline age across participants.

An regression revealed the predicted effect: According to their birth dates, people were nearly a year-and-a-half younger after listening to "When I'm Sixty-Four" rather than to "Kalimba"

$F(1, 17) = 4.92, p = .040.$