STAT 215
Pairwise Comparisons and the Family-wise Error Rate

Colin Reimer Dawson

Oberlin College

November 28, 2017

Outline

Review: Pairwise Comparisons

The Family-wise Error Rate

Overall Test of the Model

Null Population Model:

$$Y_i = \mu + \varepsilon$$

Groups Population Model:

$$Y_i = \mu + \alpha_k + \varepsilon$$

$H_0 : \alpha_k \equiv 0$ for all k H_1 : some $\alpha_k \neq 0$

Individual and Pairwise Inference

Items of Interest...

1. CIs for individual μ_ks
2. CIs for pairwise differences, $\mu_A - \mu_B$
3. t-tests for pairwise differences, $H_0 : \mu_A = \mu_B$, $H_1 : \mu_A \neq \mu_B$

In general...

Do these as we normally would, but use the "pooled within groups variance", estimated by MS_{Within}, in place of s_A, s_B, etc.

Intervals and Tests to Compare Two Means

- Normally:

$$\text{CI for } \mu : \bar{Y} \pm t^* \cdot SE \text{ where } SE = \sqrt{\frac{\hat{\sigma}^2}{n}}$$

$$\text{CI for } \mu_1 - \mu_2 : \bar{Y} \pm t^* \cdot SE \text{ where } SE = \sqrt{\frac{\hat{\sigma}_A^2}{n_A} + \frac{\hat{\sigma}_B^2}{n_B}}$$

$$t_{obs} \text{ to test } H_0 : \mu_1 - \mu_2 = 0 \text{ is } t_{obs} = \frac{\bar{Y} - 0}{SE}$$

- For the ANOVA model, we assume, among other things, that there is one σ_ε^2 common to all groups, estimated by $\hat{\sigma}_\varepsilon^2 = MS_{Error}$.

So...

CI for μ_k : $\bar{Y} \pm t^* \cdot SE$ where $SE = \sqrt{\frac{MS_{Error}}{n_k}}$

CI for $\mu_A - \mu_B$: $\bar{Y} \pm t^* \cdot SE$ where $SE = \sqrt{\frac{MS_{Error}}{n_A} + \frac{MS_{Error}}{n_B}}$

t_{obs} to test $H_0 : \mu_1 - \mu_2 = 0$ is $t_{obs} = \frac{\bar{Y} - 0}{SE}$

How many df for t^* and t_{obs}? Use df_{Error}, since this represents number of pieces of information about σ^2_ε

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		

Let's compute a CI for $\mu_{Athlete} - \mu_{NoPrime}$.

Pairwise Comparison

We have

$$\bar{x}_{Athlete} = 66.97 \qquad n_{Athlete} = 12$$
$$\bar{x}_{NoPrime} = 82.46 \qquad n_{NoPrime} = 13$$
$$\bar{x}_{Athlete} - \bar{x}_{NoPrime} = -15.49 \qquad MS_{Error} = 25.72$$
$$\widehat{SE} = \sqrt{\frac{MSE}{n_{Athlete}} + \frac{MSE}{n_{NoPrime}}} = \sqrt{\frac{25.72}{12} + \frac{25.72}{13}} = 2.03$$

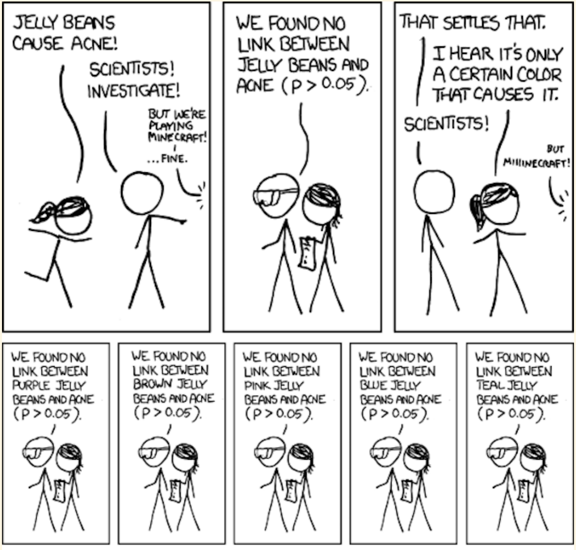
```
tstar <- qt(c(0.025, 0.975), df = 37 - 3)
CI <- -15.49 + tstar * 2.03; CI

    [1] -19.61546 -11.36454
```

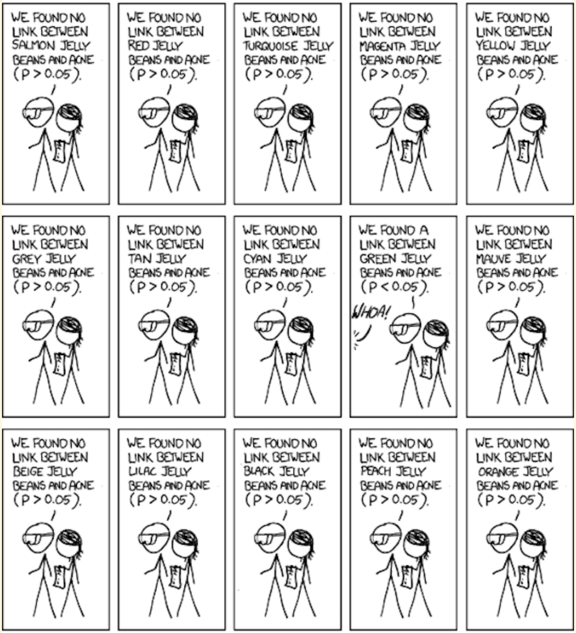
$$t_{obs} = \frac{-15.49}{2.03} = -7.63$$

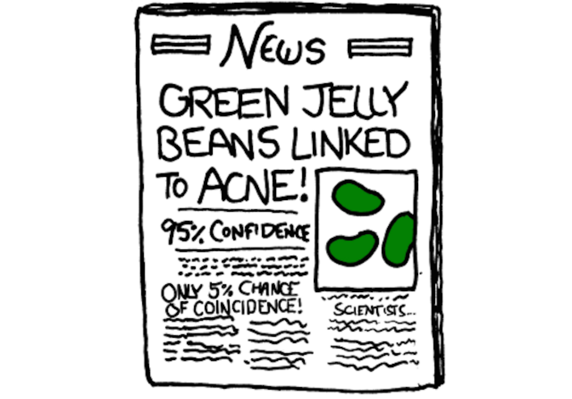
```
P.value <- pt(-7.73, df = 37 - 3); P.value

    [1] 2.71998e-09
```

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 MIIIINECRAFT!
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.

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

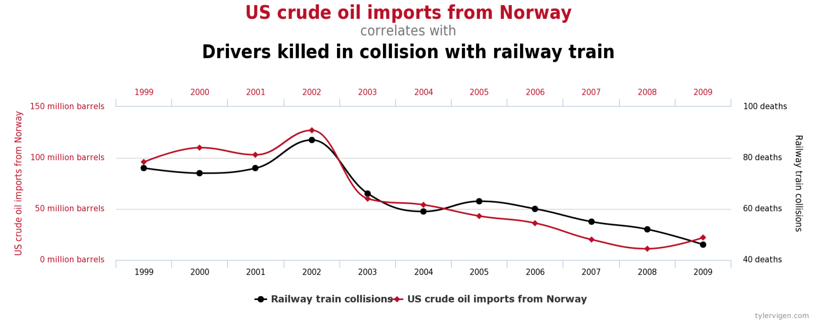
Chronological Rejuvenation, Honestly

Using the same method as in Study 1, we asked 34 University of Pennsylvania undergraduates to listen only to either "When I'm Sixty-Four" by The Beatles or "Kalimba" or "Hot Potato" by the Wiggles. We conducted our analyses after every session of approximately 10 participants; we did not decide in advance when to terminate data collection. Then, in an ostensibly unrelated task, they indicated only their birth date (mm/dd/yyyy) and how old they felt, how much they would enjoy eating at a diner, the square root of 100, their agreement with "computers are complicated machines," their father's age, their mother's age, whether they would take advantage of an early-bird special, their political orientation, which of four Canadian quarterbacks they believed won an award, how often they refer to the past as "the good old days," and their gender.

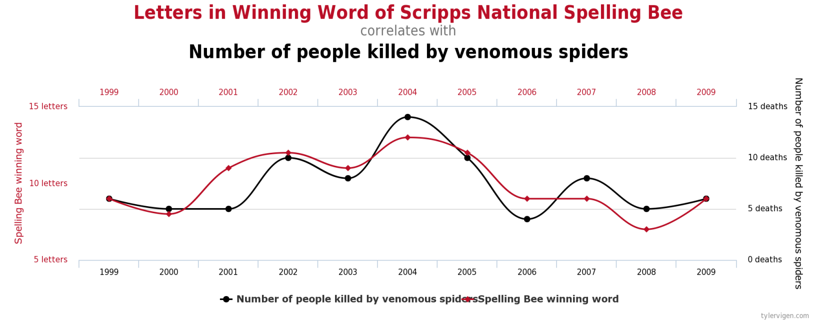
Chronological Rejuvenation, Honestly

We used father's age to control for variation in baseline age across participants. A multiple 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" ($F(1, 17) = 4.92, p = .040$). Without controlling for father's age, the age difference was smaller and did not reach significance ($F(1, 18) = 1.01, p = .33$).

Statistically Significant Correlation



Statistically Significant Correlation



Statistically Significant Correlation