Simultaneous Inference

David Gerard

2018-12-07

Objectives

- Learn about issues when running many tests.
- Learn about solutions when running many tests.
- Implement these solutions in R.

P-values

- Probability of seeing data as extreme or more extreme than what we saw **if** H_0 **were true**.
- Suppose we are running *many* tests.
- Suppose we reject when the p-value is less than 0.05.
- Then even if H_0 is **true** in all tests, we would **reject** 5% of them.

Illustration

```r
tvec <- rt(1000, df = 20) ## distrubiton under H0
pvalue <- 2 * pt(-abs(tvec), df = 20) ## p-values
mean(pvalue < 0.05) ## proportion of p-vals less than 0.05
```

```
## [1] 0.033
```

Illustration

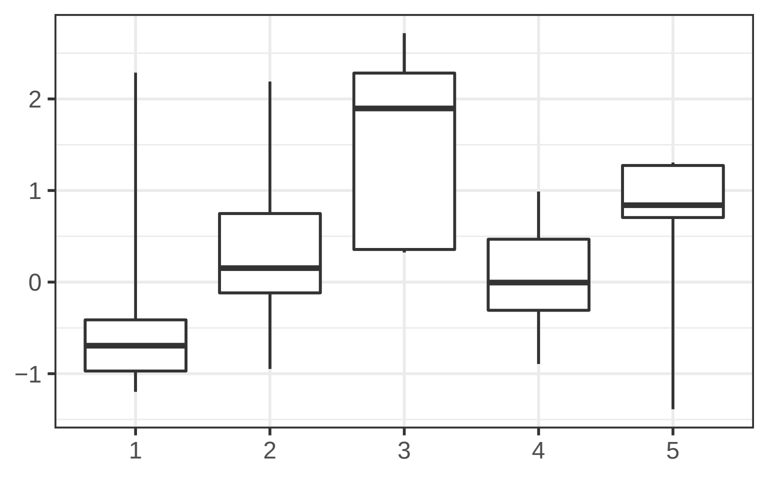
https://xkcd.com/882/

Data Snooping

- Suppose you run 20 tests, get one significant result, and only report that significant result. This is a form of **data snooping**.
- More generally, **data snooping** is where you look at the data before choosing the hypotheses to test.
- A **planned comparison** is a hypothesis test chosen before looking at the data.

Data Snooping

- Exercise: Rank the below pairwise comparisons in decreasing order of what you think would be the largest effect.



Data Snooping

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  df_temp$y and df_temp$x
##
##   1    2    3    4
## 2 0.41 -    -    -
## 3 0.03 0.14 -    -
## 4 0.73 0.62 0.05 -
## 5 0.31 0.84 0.19 0.49
##
## P value adjustment method: none
```

Data Snooping

- Actual ordering

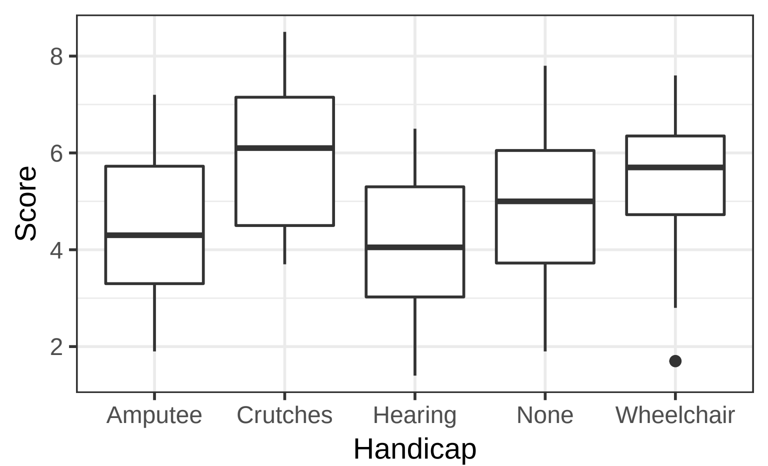
```
##   Var1 Var2  pvalue
##      1    3 0.02637
##      3    4 0.05355
##      2    3 0.13589
##      3    5 0.19046
##      1    5 0.30960
##      1    2 0.40873
##      4    5 0.49431
##      2    4 0.62424
##      1    4 0.73272
##      2    5 0.84452
```

Handicap Study

- How do physical handicaps affect people's perception of employment qualifications?
- Randomly assigned 70 undergrads to view videos of interviews containing actors performing with different handicaps.
- Undergrads rated the qualifications of the applicant on a 10-point scale.

EDA

```r
library(Sleuth3)
library(ggplot2)
data("case0601")
qplot(Handicap, Score, data = case0601, geom = "boxplot")
```



All Pairwise Tests

- Run all tests for $H_0 : \mu_i = \mu_j$ vs $H_A : \mu_i \neq \mu_j$.

```
pairwise.t.test(x = case0601$Score,
                g = case0601$Handicap,
                p.adjust.method = "none")
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  case0601$Score and case0601$Handicap
##
##            Amputee Crutches Hearing None
## Crutches   0.018   -        -       -
## Hearing    0.542   0.003    -       -
## None       0.448   0.103    0.173   -
## Wheelchair 0.143   0.352    0.040   0.476
##
## P value adjustment method: none
```

Question

- Are those moderate p-values (0.018 and 0.04) meaningful?
- Or are they there because all hypotheses are null and these just happened to be less than 0.05?
- Running 10 tests, so on average 0.5 should be rejected.

Definition

- The **family-wise error rate** is the probability of a false postitive (Type I error) among a family of hypothesis tests.
- I.e. the probability of making at least one Type I Error
- Recall: Type I error = rejecting H_0 when it is true.

Adjusted *p*-value

- Given a family of hypothesis tests, the **adjusted *p*-value** of a test is less than α if and only if the probability of at least one Type I error (among all tests) is at most α.
- That is, if you reject when the adjusted *p*-value is less than α, then the probability (prior to sampling) of any test producing a Type I error is less than α.

Bonferroni Procedure

- Multiply the p-value by the number of tests.
- Works for **any** family of **preplanned** hypothesis tests.
- p-values tend to be much larger than other corrections.

Proof of Bonferroni Correction

- $m =$ Total number of tests.
- $m_0 =$ Number tests where the null hypothesis is correct.
- $p_i = p$-value for test i.
- Suppose (unknown to us) that the first m_0 tests are the ones where the null is true.

Family-wise error rate

Proof of Bonferroni Correction

- $m =$ Total number of tests.
- $m_0 =$ Number tests where the null hypothesis is correct.
- $p_i = p$-value for test i.
- Suppose (unknown to us) that the first m_0 tests are the ones where the null is true.

Family-wise error rate

$= Pr(\text{Type I error among the } m_0 \text{ tests})$

Proof of Bonferroni Correction

- $m =$ Total number of tests.
- $m_0 =$ Number tests where the null hypothesis is correct.
- $p_i = p$-value for test i.
- Suppose (unknown to us) that the first m_0 tests are the ones where the null is true.

Family-wise error rate

$= Pr(\text{Type I error among the } m_0 \text{ tests})$

$= Pr(mp_1 \leq \alpha \text{ or } mp_2 \leq \alpha \text{ or } \cdots \text{ or } mp_{m_0} \leq \alpha)$

Proof of Bonferroni Correction

- m = Total number of tests.
- m_0 = Number tests where the null hypothesis is correct.
- p_i = p-value for test i.
- Suppose (unknown to us) that the first m_0 tests are the ones where the null is true.

Family-wise error rate

$= Pr(\text{Type I error among the } m_0 \text{ tests})$

$= Pr(mp_1 \leq \alpha \text{ or } mp_2 \leq \alpha \text{ or } \cdots \text{ or } mp_{m_0} \leq \alpha)$

$= Pr(p_1 \leq \alpha/m \text{ or } p_2 \leq \alpha/m \text{ or } \cdots \text{ or } p_{m_0} \leq \alpha/m)$

Proof of Bonferroni Correction

- $m =$ Total number of tests.
- $m_0 =$ Number tests where the null hypothesis is correct.
- $p_i = p$-value for test i.
- Suppose (unknown to us) that the first m_0 tests are the ones where the null is true.

Family-wise error rate

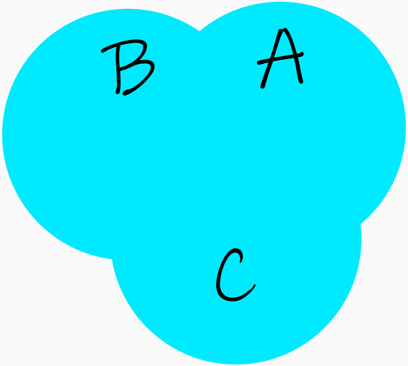
$$= Pr(\text{Type I error among the } m_0 \text{ tests})$$

$$= Pr(mp_1 \leq \alpha \text{ or } mp_2 \leq \alpha \text{ or } \cdots \text{ or } mp_{m_0} \leq \alpha)$$

$$= Pr(p_1 \leq \alpha/m \text{ or } p_2 \leq \alpha/m \text{ or } \cdots \text{ or } p_{m_0} \leq \alpha/m)$$

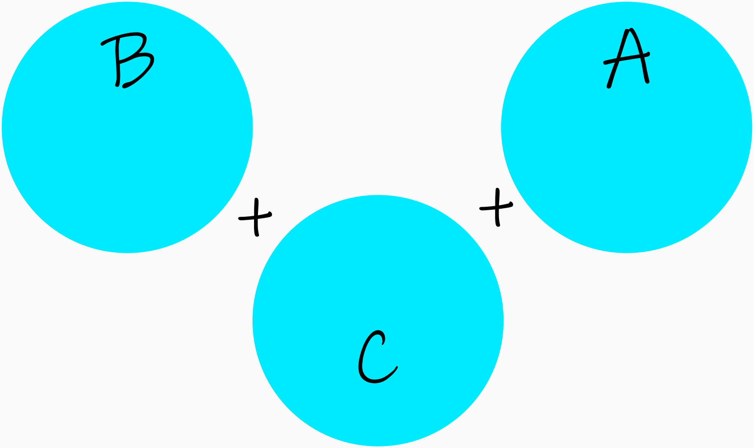
$$\leq Pr(p_1 \leq \alpha/m) + Pr(p_2 \leq \alpha/m) + \cdots + Pr(p_{m_0} \leq \alpha/m)$$

$$\Pr(A \text{ or } B \text{ or } C)$$



Bonferroni Inequality

$$\Pr(A) + \Pr(B) + \Pr(C)$$



Proof of Bonferroni Correction

Family-wise error rate

$= Pr(\text{Type I error among the } m_0 \text{ tests})$

$= Pr(mp_1 \leq \alpha \text{ or } mp_2 \leq \alpha \text{ or } \cdots \text{ or } mp_{m_0} \leq \alpha)$

$= Pr(p_1 \leq \alpha/m \text{ or } p_2 \leq \alpha/m \text{ or } \cdots \text{ or } p_{m_0} \leq \alpha/m)$

$\leq Pr(p_1 \leq \alpha/m) + Pr(p_2 \leq \alpha/m) + \cdots + Pr(p_{m_0} \leq \alpha/m)$

Proof of Bonferroni Correction

Family-wise error rate

$= Pr(\text{Type I error among the } m_0 \text{ tests})$

$= Pr(mp_1 \leq \alpha \text{ or } mp_2 \leq \alpha \text{ or } \cdots \text{ or } mp_{m_0} \leq \alpha)$

$= Pr(p_1 \leq \alpha/m \text{ or } p_2 \leq \alpha/m \text{ or } \cdots \text{ or } p_{m_0} \leq \alpha/m)$

$\leq Pr(p_1 \leq \alpha/m) + Pr(p_2 \leq \alpha/m) + \cdots + Pr(p_{m_0} \leq \alpha/m)$

$= \alpha/m + \alpha/m + \cdots + \alpha/m$ (m_0 summations)

Proof of Bonferroni Correction

Family-wise error rate

$= Pr(\text{Type I error among the } m_0 \text{ tests})$

$= Pr(mp_1 \leq \alpha \text{ or } mp_2 \leq \alpha \text{ or } \cdots \text{ or } mp_{m_0} \leq \alpha)$

$= Pr(p_1 \leq \alpha/m \text{ or } p_2 \leq \alpha/m \text{ or } \cdots \text{ or } p_{m_0} \leq \alpha/m)$

$\leq Pr(p_1 \leq \alpha/m) + Pr(p_2 \leq \alpha/m) + \cdots + Pr(p_{m_0} \leq \alpha/m)$

$= \alpha/m + \alpha/m + \cdots + \alpha/m$ (m_0 summations)

$= m_0\alpha/m$

Proof of Bonferroni Correction

Family-wise error rate

$= Pr(\text{Type I error among the } m_0 \text{ tests})$

$= Pr(mp_1 \leq \alpha \text{ or } mp_2 \leq \alpha \text{ or } \cdots \text{ or } mp_{m_0} \leq \alpha)$

$= Pr(p_1 \leq \alpha/m \text{ or } p_2 \leq \alpha/m \text{ or } \cdots \text{ or } p_{m_0} \leq \alpha/m)$

$\leq Pr(p_1 \leq \alpha/m) + Pr(p_2 \leq \alpha/m) + \cdots + Pr(p_{m_0} \leq \alpha/m)$

$= \alpha/m + \alpha/m + \cdots + \alpha/m$ (m_0 summations)

$= m_0\alpha/m$

$\leq m\alpha/m$

Proof of Bonferroni Correction

Family-wise error rate

$= Pr(\text{Type I error among the } m_0 \text{ tests})$

$= Pr(mp_1 \leq \alpha \text{ or } mp_2 \leq \alpha \text{ or } \cdots \text{ or } mp_{m_0} \leq \alpha)$

$= Pr(p_1 \leq \alpha/m \text{ or } p_2 \leq \alpha/m \text{ or } \cdots \text{ or } p_{m_0} \leq \alpha/m)$

$\leq Pr(p_1 \leq \alpha/m) + Pr(p_2 \leq \alpha/m) + \cdots + Pr(p_{m_0} \leq \alpha/m)$

$= \alpha/m + \alpha/m + \cdots + \alpha/m$ (m_0 summations)

$= m_0\alpha/m$

$\leq m\alpha/m$

$= \alpha$

Bonferroni Continued

```r
pairwise.t.test(x = case0601$Score,
                g = case0601$Handicap,
                p.adjust.method = "bonferroni")
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  case0601$Score and case0601$Handicap
##
##            Amputee Crutches Hearing None
## Crutches   0.18    -        -       -
## Hearing    1.00    0.03     -       -
## None       1.00    1.00     1.00    -
## Wheelchair 1.00    1.00     0.40    1.00
##
## P value adjustment method: bonferroni
```

Holm Procedure

- Slightly better than Bonferroni, and is the default in R.
- Same conditions as Bonferroni (pre-planned tests, any type of tests)

Holm Continued

```r
pairwise.t.test(x = case0601$Score,
                g = case0601$Handicap,
                p.adjust.method = "holm")
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  case0601$Score and case0601$Handicap
##
##            Amputee Crutches Hearing None
## Crutches   0.17    -        -       -
## Hearing    1.00    0.03     -       -
## None       1.00    0.72     0.87    -
## Wheelchair 0.86    1.00     0.32    1.00
##
## P value adjustment method: holm
```

Tukey-Kramer Procedure

- Use when you want **all** pairwise comparisons.
- Smaller p-values than Bonferroni.
- Tests need to be **preplanned**.
- Needs `aov()` object as input.

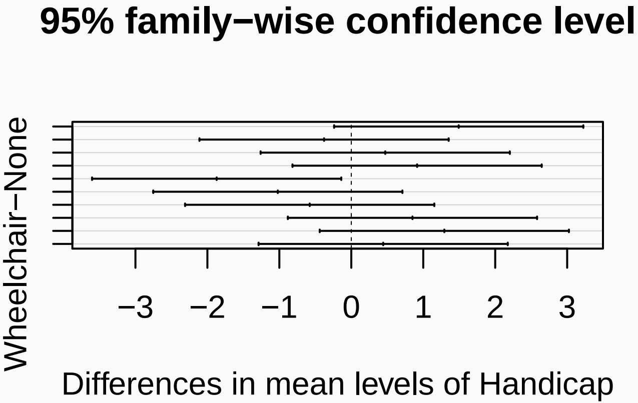
Tukey Continued

```
aout <- aov(Score ~ Handicap, data = case0601)
tout <- TukeyHSD(aout)
tout
```

```
##   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     1.4929 -0.2389  3.2246 0.1233
## Hearing-Amputee     -0.3786 -2.1103  1.3532 0.9725
## None-Amputee         0.4714 -1.2603  2.2032 0.9400
## Wheelchair-Amputee   0.9143 -0.8174  2.6460 0.5781
## Hearing-Crutches    -1.8714 -3.6032 -0.1397 0.0278
## None-Crutches       -1.0214 -2.7532  0.7103 0.4686
## Wheelchair-Crutches -0.5786 -2.3103  1.1532 0.8812
```

Cool plotting

```r
plot(tout)
```



Many others

- There are *many* other adjustment methods.
- Each of these specialize in certain testing scenarios.
- Read the help-page of `p.adjust()` for more information.

Adjusted Confidence Intervals

All Confidence Inverals for Means

$$\text{estimate} + \text{multiplier} * \text{standard error}$$

Different Multipliers

- Original multiplier $= t_{n-1}(1 - \alpha/2)$
- Bonferroni multiplier $= t_{n-1}(1 - \alpha/(2m))$, where m is the number of tests.
- Tukey has its own multiplier (get those CI's automatically from `TukeyHSD()`).