STAT 213: Pitfalls of Multiple Testing

Colin Reimer Dawson

Last Revised April 25, 2018

Let's do an experiment to examine what happens when we do lots of comparisons between pairs of means in the setting of a predictor variable with lots of categories. Work in pairs or threes for this activity.

First, let's suppose that we have a categorical predictor variable with 12 levels, and a quantitative response variable. In reality, the predictor is unrelated to the response; in other words, in the population / in the long run, the mean of the response variable is the same regardless of which of the ten categories the case is in.

Suppose we don't know this reality going in, and we want to test against the null hypothesis that the variables are unrelated; that the ten population means are identical. We collect data from this population, with 20 cases in each group.

Enter the following R code to create a synthetic dataset with this property. Set a random seed first (`set.seed(SOME_NUMBER)`) so that you can get the same results every time you re-run or re-Knit your code.

```
set.seed(1) # fill in your own seed value here
## create 12 groups of 20 observations of a response variable drawn from a single
## common Normally distributed population (in particular, there is
## one population mean in reality)
response <- rnorm(n = 20 * 12, mean = 50, sd = 25)
## Create a grouping variable with twelve values (groups), assigning 20
## observations to each group
groups <- rep(c("A", "B", "C", "D", "E", "F", "G", "H", "I", "J", "K", "L"),
              each = 20)
## Combine these into a dataset
FakeData <- data.frame(Y = response, X = groups)
```

Take a look at the dataset with `head()` to verify that it looks as expected.

Plot the data by group, highlighting the means, and get quantitative descriptive statistics as well.

```r
library("mosaic")
xyplot(Y ~ X, data = FakeData, type = c("p","a"))
favstats(Y ~ X, data = FakeData) %>% format(digits = 4)
```

Which groups have sample means that are farthest apart?

Filter the data to include only the two groups that have the *biggest* difference in means, and do a *t*-test for those groups to see whether we would reject the null hypothesis that the means are equal. For example, in my case, groups I and L had the largest and smallest sample means, so I will compare them via a *t*-test. (The groups that are farthest apart for you will depend on your random seed)

```r
filter(FakeData, X %in% c("I","L")) %>% t.test(Y ~ X, data = .)
```

In fact, we can get P-values for all possible pairwise comparisons of pairs of group means as follows:

```r
with(FakeData, pairwise.t.test(Y, X, p.adjust.method = 'none'))
```

Answer the following questions:

Questions

1. How many pairwise comparisons are there in total? How many of these appear to show statistically significant evidence of a difference at the population level (using a significance level of 0.05)?

2. Given that we know exactly the population model that generated the data, is it correct to reject H_0 for any pairs?

3. If you had a dataset like this handed to you and did all pairwise tests, how many times do you *expect* you would reject H_0 mistakenly? What is it called when this happens?

4. What is your conclusion if, instead of (or before) doing all of these pairwise comparisons, you do the overall F test? (You can use the following code to do this)

   ```r
   aov.model <- aov(Y ~ X, data = FakeData)
   summary(aov.model)
   ```

5. Why is it a bad idea to decide which differences you think will matter by looking at the data you've already collected?

6. Why might it be a good idea to do an overall F-test *before* doing any pairwise comparisons?