STAT 213: Exploring the Familywise Error Rate

Colin Reimer Dawson

Last Revised March 3, 2016

Let's do an experiment to examine what happens when we do lots of pairwise tests.

First, let's suppose that we have a dataset with 10 groups of 20 observations each, where all the means are equal.

Enter the following R code to create a synthetic dataset with this property. If you are running this in a script or a Markdown document, first type `set.seed(SOME_NUMBER)` so that you can get the same results every time you re-run or re-Knit your code.

```r
## create a categorical predictor variable with 20 repetitions of 10 levels
x <- rep(c("A", "B", "C", "D", "E", "F", "G", "H", "I", "J"), each = 20)
## create a response variable drawn from a single common Normal population
y <- rnorm(n = 20 * 10, mean = 50, sd = 10)
## Combine these into a dataset
FakeData <- data.frame(X = x, Y = y)
```

Now let's fit the ANOVA model.

```r
TheModel <- lm(Y ~ X, data = FakeData)
```

We can get P-values for all possible pairwise comparisons of X levels as follows:

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

Exercises

1. How many comparisons are there in total? How many of these suggest a significant difference (at the 0.05 level)?
2. Given that we know exactly the population distribution, 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?

The Family-Wise Error Rate

When a particular pair of population means are identical, the significance level α controls the probability that we incorrectly reject H_0 and mistakenly conclude that those two population means are not equal.

If we have many means, however, there are many more pairs, and each one has a probability of α of yielding a Type I Error. Taken together, the probability that we make *at least one* Type I Error is called the **family-wise error rate** (FWER), and may be much higher than α. It is often desireable to control the FWER directly. Unfortunately, it is not possible without knowing more about the population than we know to exactly control this rate, and so different approaches exist that make different tradeoffs between allowing the FWER to exceed the desired α, and allowing higher-than-necessary Missed Discovery (Type II Error) rate. Approaches that are more strict about the FWER at the expense of Missed Discoveries are called **conservative**; those that strike the balance in the other direction aree called **liberal**.

Three popular approaches (from most liberal to most conservative) are

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