Multiple Comparisons

This lab will discuss the problem of multiple comparisons, and introduce one method to deal with this problem – the Bonferroni correction.

Movie ratings

We will again be using the file moviesall.txt in this lab. This data contains various information on movies released in the year 2003. Although it is not really a random sample of movies, we will treat it as a random sample of movies for the purpose of this lab. You may think of it as representative of movies in general in the past decade.

For each movie, we have its rating (G, PG, PG-13, R), its genre, its box-office gross (in millions of dollars), its run time (in minutes), and its score on `rottentomatoes.com` (higher scores mean better movies). For this lab, you will investigate differences in run time between the movie ratings.

Pairwise comparisons

At this point we have three tests available for testing for difference between more than two groups – ANOVA, the permutation F-test, and the Kruskal-Wallis test. Each of these tests is what we call a global hypothesis test. In each case we test whether or not there are differences among the groups, but we do not determine which group(s) are different.

If our global test provides evidence that there are differences among the groups, the natural next step is to determine where these differences occur. That is, we want to know *which* of the groups are different. One way to determine this is to perform a series of two-sample tests for each pair of groups. For example, in the movie run time case, we would want to compare R movies to G movies, R movies to PG movies, R movies to PG-13 movies, etc. These two-sample tests are called *pairwise comparisons*. For these pairwise comparisons, we can choose any appropriate test from the two-sample tests we have available – the t-test, the permutation test (with either means or medians), or the Wilcoxon rank-sum test.

1. If we want to do a two-sample test on each pair of movie ratings, how many pairwise comparisons will we have to consider? In general, when we have K different groups, how many pairwise comparisons will we have to consider?

Multiple testing problem

Doing pairwise comparisons leads to what statisticians call the *multiple testing problem*. This problem arises when more than one statistical test is performed. When we perform a single statistical test, we can choose to set a significance level – quite often a level of $\alpha = .05$. If we set $\alpha = .05$, we choose to reject H_0 whenever we have a p-value less than .05. In doing this, we are saying that the probability of making a Type I error while performing the test is 5%. When we do multiple tests and use an $\alpha = .05$ for each individual test, the overall significance level may not be .05; instead, it may be much higher. This means that overall we may be making a Type I error more than 5% of the time.

To illustrate this point, we are going to take a random sample of 100 data points from a normal distribution with mean 10 and standard deviation 5. We will then perform a t-test to determine whether or not the mean is 10. (Notice that a t-test is appropriate here, because the population truly does have a normal distribution!)

2. Using the commands below, take a random sample of size 100 from a normal(10,5^2) distribution. Perform a t-test to test whether $\mu = 10$. Report your p-value. Using a significance level of $\alpha = .05$, would you reject the null hypothesis that $\mu = 10$?

```
data = rnorm(100, 10, 5)
t.test(data, mu = 10, alternative = "two.sided")
```

Note that in this case the null hypothesis is true, since the mean really is 10. So a correct decision would be to not reject the null hypothesis that $\mu = 10$.

3. Repeat the above two lines of code 19 more times, for a total of 20 tests. Report your p-value for each of these tests. How many and what proportion of times (out of the ten total) did you reject the null hypothesis that $\mu = 10$?

Since a significance level of $\alpha = .05$ means we have a 5% chance of making a Type I error, you would expect to incorrectly reject the null hypothesis in 1 of these 20 tests.

4. Now we want to repeat this test a total of 1000 times and calculate a p-value for each test. To do this, use the following code:

```
pval = rep(NA, 1000)

for(i in 1:1000) {

data = rnorm(100, 10, 5)
pval[i] = t.test(data, mu = 10, alternative = "two.sided")$p.value

}

sum(pval < .05)
sum(pval < .05)/1000
```

   How many and what proportion of times did you reject the null hypothesis that $\mu = 10$?

At a significance level of $\alpha = .05$ we expect to incorrectly reject the null hypothesis in 50 of these tests. As you can see, even though we perform each test at the $\alpha = .05$ level, we are virtually guaranteed to make a Type I error if we do enough tests. So although our probability of a Type I error *on any individual test* is 5%, our overall probability of Type I error *across all the tests together* may be much, much higher than 5%. As a result, we need to adjust our test to account for these multiple tests.

Bonferroni correction

There have been many methods developed to deal with the multiple-testing problem. The simplest such method is called the *Bonferroni correction*.

The Bonferroni correction says that if we want to perform a series of tests at an overall significance level of $\alpha = .05$, then we need to perform each individual test at a significance level of $\alpha' = \frac{.05}{\text{the \# of tests}}$.

5. In doing pairwise comparisons on the movie run time data, what significance level should you use for each individual test in order to end up with an overall $\alpha = .05$?

An alternative to adjusting the significance level for each individual test is to simply adjust the p-value for each individual test and then interpret the p-values as we usually do. To make a Bonferroni adjustment to a p-value, we multiply the original p-value by the number of tests we are performing. In this way, using a Bonferroni-adjusted p-value is equivalent to using a Bonferroni-adjusted significance level.

6. If you want to calculate Bonferroni-adjusted p-values for the movie run-time data, you should multiply your original p-values by what value?

In Lab 7, you performed a permutation F-test to determine whether there were differences in run times among the rating groups. Your test resulted in a p-value that was quite small – around 0.005 – and so you had strong evidence that at least one group had a different run time distribution than the others. Now you will perform pairwise comparisons for this data.

6. Perform a permutation test (using a difference in means) to determine whether there is a difference in run times between G and PG movies. Report both your original p-value and your Bonferroni-adjusted p-value. State a conclusion based on your Bonferroni-adjusted p-value.
7. Repeat #6, but test for a difference between G and PG-13 movies.
8. Repeat #6, but test for a difference between G and R movies.
9. Repeat #6, but test for a difference between PG and PG-13 movies.
10. Repeat #6, but test for a difference between PG and R movies.
11. Repeat #6, but test for a difference between PG-13 and R movies.
12. Using the results of #6 through #11, write a short paragraph summarizing the differences in run times between movies of difference ratings.

In this case, we used the permutation F-test as our overall test, followed by permutation tests for the individual pairs. We could also have used a Kruskal-Wallis test for our overall test, followed by Wilcoxon rank-sum tests for the individual pairs. Any appropriate K-sample method could be used for the overall test, and any appropriate two-sample method could be used for the individual pairs.

Lab Summary

Complete your Lab Summary in a separate Word document. Please write clearly and in complete sentences. Include any `R` code at the end of your summary, after you have written up your solution.

1. Perform an analysis of the relationship between movie rating and score on `rottentomatoes.com`. To do this, perform an overall test for differences in box office gross using the Kruskal-Wallis test. Then, provided you find at least some evidence of a difference between the groups, perform pairwise tests of each rating pair using the Wilcoxon rank-sum test and Bonferroni-adjusted p-values. Summarize your results about the relationship between movie rating and box office gross in a short paragraph.
2. Perform an analysis of the relationship between movie rating and box office gross. To do this, perform an overall test for differences in box office gross using the Kruskal-Wallis test. Then, provided you find at least some evidence of a difference between the groups, perform pairwise tests of each rating pair using the Wilcoxon rank-sum test and Bonferroni-adjusted p-values. Summarize your results about the relationship between movie rating and box office gross in a short paragraph.