Statistical Computing Project 3

Mario Ibanez

March 14, 2016

Introduction

In this report, a randomization test is used on the following data:

A	B	C
6, 8	9, 11, 9	17, 15, 16, 16

to test the null hypothesis $H_0 : \mu_A = \mu_B = \mu_C$. A randomization test can be used to test this hypothesis when the assumptions for an ANOVA test are not satisfied. Two methods will be used: systematic permutation and random permutation.

Discussion

In order to conduct a one-way ANOVA test, there are a few assumptions that need to be met in order for the test to be valid. The first is the assumption of independence. The values in each of the treatments A, B, and C must be independent, and the values across treatments must also be independent. The second assumption that is required is that the samples in each treatment must come from populations that are normally distributed. The third assumption is that the populations from which each sample is selected must have equal variance. If any of these three assumptions are violated, the validity of the ANOVA test is compromised.

Since the sample sizes in each treatment are so small, it is difficult to test the assumptions of normality and equal variance. It may also be suspect that the samples are independent. Though an ANOVA test could be performed, a randomization test is a good option because it does not rely on any of these assumptions.

Algorithm

The randomization test relies on calculating a certain test statistic for each different permutation of the original data. In this case, the test statistic is:

$$T = \sum_{i=1}^{3} \frac{T_i^2}{n_i}$$

where T_i is the sum of the values in treatment i and n_i is the number of values in treatment i. This test statistic is a simplified version of the F statistic from the ANOVA test that would be used in this situation. The p-value is calculated by:

$$\text{p-value} = \frac{\text{number of permutations that have a test statistic greater than the original}}{\text{total number of permutations}}$$

Systematic Permutation

This method involves calculating the test statistic for every possible permutation of the data. There are repititions in the data, but since the total number of permutations is not large, the repititions do not need to be dealt with. In total, the number of permutations is:

$$\frac{9!}{2!3!4!} = 1260$$

The code works by finding the 1260 unique permutations of the string "AABBBCCCC". Each permutation of the string can be used to create a permutation of the numbers 6, 8, 9, 11, 9, 17, 15, 16, 16 where position does not matter in the first two positions, in the following three positions, or in the last four positions. For example, the string "ABCCABCBC" is looked at as "AB - CCA - BCBC". This means that the first number, 6, is put into treatment A, the second second number, 8, is put into treatment B, the third number, 9, is put into treatment C, and so on. The *R* function *permn* from the package *combinat* is used to find the 1260 permutations. The p-value is calculated by calculating the test statistic for each permutation and counting the number that are at least as large as the test statistic for the original permutation. The code below does the calculation (the rest of the code is in the appendix):

```
# Initialize results vector
results <- rep(0, 1260)

# Loop to calculate test statistics
for(i in 1:1260){
  results[i] <- test_stat(data, permutations[[i]])
}
```

The test statistic for the original permutation is 1402.3333333. The number of test statistics that are greater than or equal to this is:

```
sum(results >= results[1])
```

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

It turns out that the original permutation had the largest test statistic. Therefore, the p-value is $\frac{1}{1260}$. There is sufficient evidence to reject the null hypothesis at the 0.05% significance level.

Random Permutation

Immediately it can be seen that since the original permutation has the largest test statistic, the only result that can occur if a random sample (without replacement) of 800 permutations is tested is a p-value of either 0/800 if the original is not selected in the 800 or 1/800 if the original is selected in the random sample of 800.

In a normal question where randomization would be used, the original permutation would not have the largest test statistic and the number of total permutations would be higher. With more total permutations, sampling with replacement would be reasonable. The following code calculates the test statistic for 800 randomly chosen (with replacement) permutations of the string "AABBBCCCC". This method is almost identical to what was done in the systematic case except that the permutations are being chosen randomly and with replacement rather than systematically.

```r
# Initialize results vector
results <- rep(0, 800)

# Loop to calculate test statistics
for(i in 1:800){
  results[i] <- test_stat(data, sample(x = permutations[[1]], size = 9))
}
```

The number of test statistics in this set of 800 greater than or equal to 1402.3333 is:

```r
sum(results >= 1402.3333)
```

```
## [1] 0
```

This number will fluctuate from simulation to simulation because of the random aspect. Regardless, as a proportion it is so low that the same conclusion can be reached as in the systematic case. There is sufficient evidence to reject the null hypothesis at the 0.05% significance level.

Conclusion

In the original data set, all of the highest values were in treatment C, all the second highest values were in treatment B, and all of the smallest values are in treatment A. Due to the way the test statistic is calcuated, this is the reason that the highest test statistic belonged to the original permutation. Random permutation is used when the total number of permutations is so large that they cannot all be systematically checked. In that case, sampling random permutations with replacement would be reasonable, and selecting the same permutation more than once would have a low probability. In this report, with 1260 permutations, it does not make sense to use the random permutation method. Regardless, using either method, there is sufficient evidence to reject the null hypothesis and conclude that it is not the case that the three population means are equal.

Appendix

Below is the code that was used in this report:

```
# Package with permn function
library(combinat)

# Original data set
data <- c(6, 8, 9, 11, 9, 17, 15, 16, 16)

# Treatments vector that will be permuted
treats <- c("A", "A", "B", "B", "B",
            "C", "C", "C", "C")

# Calculates the 1260 unique permutations
permutations <- unique(permn(treats))

# Function that takes the data vector and the
# permutations, returns the test statistic
test_stat <- function(data, treatments){
  return(sum((sum(data[treatments=="A"])^2)/2,
             (sum(data[treatments=="B"])^2)/3,
             (sum(data[treatments=="C"])^2)/4))

# Initialize results vector
results <- rep(0, 1260)

# Loop to calculate test statistics for systematic case
for(i in 1:1260){
  results[i] <- test_stat(data, permutations[[i]])
}

# P-value
sum(results >= 1402.3333)

# Initialize results vector
results <- rep(0, 800)

# Loop to calculate test statistics for random permutation case
for(i in 1:800){
  results[i] <- test_stat(data,
                          sample(x = permutations[[1]],
                                 size = 9))
}

# P-value for random case
sum(results >= 1402.3333)
```