Problem Set 1

Adam Ten Hoeve

Introduction

Questions are 10 points each.

These questions were rendered in R markdown through RStudio (https://www.rstudio.com/wp-content/uploads/2015/02/rmarkdown-cheatsheet.pdf, http://rmarkdown.rstudio.com).

Please generate your solutions in R markdown and upload both a knitted doc, docx, or pdf document in addition to the Rmd file.

Part 1

The goal of questions 1 and 2 is to investigate whether the polio rate among the non-vaccinated children in a randomized control trial is significantly different from the polio rate in the placebo group. If participation in the trial is unrelated to contracting polio, these populations shouldn't differ significantly in their experience of the disease.

The code and simulation methods from 01_polio_simulation_binomial_model.Rmd and 01_polio_simulation_shuffle_model.Rmd may be helpful.

Question 1

Please calculate and display the proportion of paralytic polio cases in the "Placebo" group and separetely in the "NotInoculated" group in the "RandomizedControl" trial.

```
library(HistData)
dat<-PolioTrials

# Note that the placebo group is the second row in the dataframe.
n.placebo <- dat[2, 3]
n.placebo.paralytic <- dat[2, 4]
prop.placebo <- n.placebo.paralytic / n.placebo

# The "NotInoculated" group in the "RandomizedControl" is the third row of the dataframe.
n.notinoc <- dat[3, 3]
n.notinoc.paralytic <- dat[3, 4]
prop.notinoc <- n.notinoc.paralytic / n.notinoc

cat("The proportion of paralytic polio cases in the 'Placebo' group was ",
    prop.placebo, ".\n", sep="")
```

```
## The proportion of paralytic polio cases in the 'Placebo' group was 0.0005714882.
```

```
cat("The proportion of paralytic polio cases in the 'NotInoculated' group was ",
    prop.notinoc, ".\n", sep="")
```

```
## The proportion of paralytic polio cases in the 'NotInoculated' group was 0.000357166.
```

Question 2

Under the hygiene hypothesis, the "Placebo" group could be more vulnerable to polio than the "NotInoculated" group.

Consider the probability model that the number of paralytic polio cases in the "Placebo" group of the "RandomizedControl" experiment is a draw from the binomial distribution with the number of trials equal to the number of children in the "Placebo" group and the probability of "success" is equal to the proportion of paralytic polio cases in the "Placebo" and "NotInoculated" groups of the "RandomizedControl" combined. Without simulation, calculate the probability of a draw that is greater than or equal to the observed value.

```r
n <- n.placebo
# Calculate the combined proportion by dividing the combined paralytic cases
# by the combined population size.
p <- (n.placebo.paralytic + n.notinoc.paralytic) / (n.placebo + n.notinoc)
# Sum the binomial probabilities from n.placebo.paralytic to n to get the probability
# that the number of occurrences is greater than or equal to n.placebo.paralytic.
prob.Q2 <- sum(dbinom(n.placebo.paralytic:n, size=n, prob=p))
cat("The probability of having ", n.placebo.paralytic,
    " or more paralytic observations is ", prob.Q2, ".\n", sep="")
```

```
## The probability of having 115 or more paralytic observations is 0.003240697.
```

Part 2

The data below represent reviews of two facilities on a 1-5 scale with the worst rating being 1 and the best being 5. The variable "value" is the rating. The column "fac1" gives the number of each rating that the first facility received. For example, 7 raters gave the first facility the rating 1, the lowest rating. The column "fac2" gives the number of each rating that the second facility received. For example, 11 raters gave the second facility the rating 5, the highest rating.

```r
reviews<-data.frame(value=5:1,
                    fac1=c(4,0,1,0,7),
                    fac2=c(11,2,4,2,3))
```

Question 3

Please use R to calculate the mean rating for each facility in the "reviews" data.

```r
# Create a vector for the actual scores of fac1 and fac2, then take the means of those.
fac1.scores <- rep(reviews$value, times=reviews$fac1)
fac1.mean <- mean(fac1.scores)

fac2.scores <- rep(reviews$value, times=reviews$fac2)
fac2.mean <- mean(fac2.scores)

cat("Facility 1 had mean rating ", fac1.mean, " and Facility 2 had mean rating ",
    fac2.mean, ".\n", sep="")
```

```
## Facility 1 had mean rating 2.5 and Facility 2 had mean rating 3.727273.
```

Question 4

Please describe a probability model for a simulation-based hypothesis test that addresses whether the two facilities can reasonably be considered to be equivalent in the sense that the rating differences are consistent with chance. Please be sure to address the following questions:

How is the test statistic computed?

The test statistic will be the difference between the means of the scores of fac1 and fac2.

What is the probability model that captures the null hypothesis?

The null hypothesis is that fac1 and fac2 come from the same populations. In other words, that the mean scores of the two factories are statistically the same. The alternative hypothesis is that there is a statistically significant difference between the two factories, i.e. their means are not the same.

How can the probability model be simulated?

We can simulate this by resampling from the observed data. Under the null, we assume that the two groups come from the same population so we can merge the two together into a larger sample group. With the full population, we can calculate the proportion that each score value occured in the observed data. Then we can resample two vectors, one the size of fac1 and one the size of fac2, and calculate the difference between their means. If we repeat this process many times, we can create a simulated distribution for difference in means, assuming that the null hypothesis is true.

What comparison of the observed statistic and the values of the test statistics from the simulations addresses the question?

We want to determine how likely the observed statistic is to have occured, assuming the null hypothesis is true. Using our simulated distribution, we can calculate the proportion of means that were as or more extreme than our observed statistic. This tells us the probability of getting the observed statistic, assuming the null is true. If this probability is very low, then we can assume that the null hypothesis is not true, and that the two factories are different.

Some possible variable manipulations are shown below.

```r
# Create a vector of the values assigned by all the reviewers
# with the correct number of repetitions.
pop<-rep(reviews$value,times=reviews$fac1+reviews$fac2)

# Create a vector of the proportion of times each value was
# awarded.
rating.prop<-
  (reviews$fac1+reviews$fac2)/sum(reviews$fac1+reviews$fac2)

# Sample the vector (5,4,3,2,1) k times according to the
# probabilites in "rating.prop"

# set.seed(0)
k<-10
samp<-sample(5:1, k, replace=TRUE, prob=rating.prop)
```

Question 5

Please carry out the test you designed in Question 4 and state your conclusion about the extent to which the data are consistent with the null hypothesis.

```r
# What is the observed difference between the groups?
obs.mean.diff <- fac1.mean - fac2.mean

# Create a vector of the values assigned by all the reviewers
# with the correct number of repetitions.
pop<-rep(reviews$value,times=reviews$fac1+reviews$fac2)

# Create a vector of the proportion of times each value was
# awarded.
rating.prop<-
```

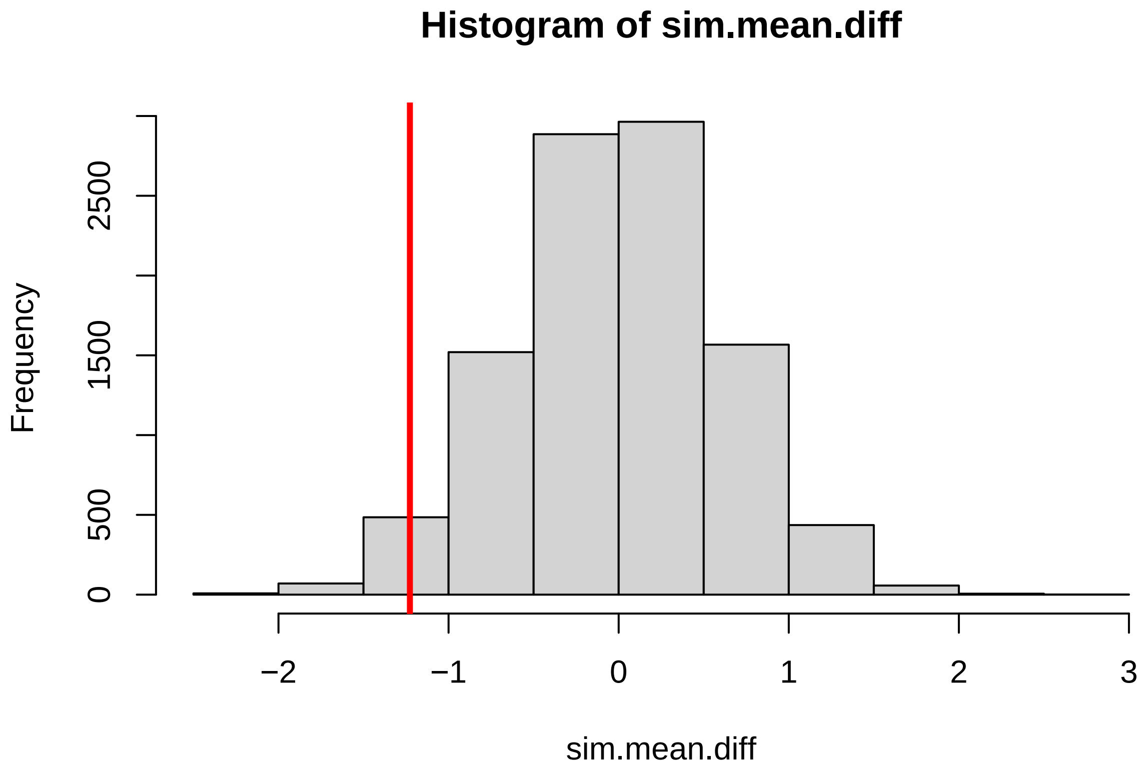
```
  (reviews$fac1+reviews$fac2)/sum(reviews$fac1+reviews$fac2)

# Get the lengths of the two observed factory vectors
fac1.size = length(fac1.scores)
fac2.size = length(fac2.scores)

# Now we can actually resample the data.
set.seed(0)
k <- 10000  # Number of iterations
sim.mean.diff <- numeric(k)
for(i in 1:k){
  # Sample a vector to resemble fac1, using the combined value proportions.
  sample1 <- sample(5:1, fac1.size, replace=T, prob=rating.prop)
  # Sample a vector to resemble fac2, using the combined value proportions.
  sample2 <- sample(5:1, fac2.size, replace=T, prob=rating.prop)
  # Calculate and save the difference in means between the two simulated samples.
  sim.mean.diff[i] <- mean(sample1) - mean(sample2)
}
```

Now that we've completed our resampling, let's take a look at where the observed data compares to the simulated data.

```
hist(sim.mean.diff)
abline(v=obs.mean.diff, col="red", lw=3)
```



Now we need to calculate the p-value. We can do this by calculating the proportion of simulated statistics that were as or more extreme than the observed data.

```
# Calculate the proportion of data that is as or more extreme than the observed data.
# This is the p-value.
pval <- mean(abs(sim.mean.diff) >= abs(obs.mean.diff))
pval
```

```
## [1] 0.047
```

Our resulting p-value was 0.047, which is less than 0.05, so we can reject the null and say that the means of the two factories are different. However, the p-value is on the cusp of significance, and solid conclusions probably can't be made. Let's check our work with a different method (the t-test) to see if we get the same results.

```
t.test(fac1.scores, fac2.scores)
```

```
## 
##  Welch Two Sample t-test
## 
## data:  fac1.scores and fac2.scores
## t = -1.9044, df = 18.56, p-value = 0.07247
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -2.5782474  0.1237019
## sample estimates:
## mean of x mean of y 
##  2.500000  3.727273
```

The t-test gives us a p-value of 0.07247, which disagrees with our simulated results. But it is also very close to the 0.05 significance, so we can be sure that the conclusion was on the edge either way.