Problem Set 1

C. Durso

Introduction

One goal of the problem sets in this class is to encourage you to innovate with the techniques you have learned. This innovation can take to form of figuring out how a method applies to a new situation, or how a principle can be generalized to create a new method.

This level of creativity can be a very enjoyable aspect of the practice of data science. It can also be very hard to pull off on a tight deadline. Please try to start the problem set early to give yourself time to step away and come back with new ideas. If you are stuck as the deadline approaches, consulting with a colleague or mentor (instructor) may be an option.

"Take chances. Get messy. Make mistakes," (Magic School Bus)

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 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 separately in the "NotInoculated" group in the "RandomizedControl" trial.

```
library(HistData)
dat<-PolioTrials
```

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.

Part 2

In this problem, you will be asked to generalize the idea of a statistic and a null hypothesis of “no difference” for two groups with binary outcomes (Paralytic and not Paralytic) to a statistic and a null hypothesis of “no difference” for two groups with three outcomes (Paralytic,NonParalytic,FalseReport). The basic question is whether the proportions of each of those outcomes differed between the RandomizedControl Vaccinated and the RandomizedControl Placebo groups. This could be used to address the question of whether the appearance of symptoms and the severity of symptoms differed between the two groups.

```
dat.2<-t(dat[1:2,4:6])
dat.2<-data.frame(dat.2)
names(dat.2)<-c("Vaccinated","Placebo")
```

Question 3

Please use R to calculate the proportions in each category for the RandomizedControl Vaccinated and the RandomizedControl Placebo groups.

Question 4

Please describe a probability model for a simulation-based hypothesis test that addresses whether the two groups can reasonably be considered to come from populations with the same proportions of Paralytic,NonParalytic, and FalseReport

How is the test statistic computed?

What is the probability model that captures the null hypothesis?

How can the probability model be simulated?

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

Some possible variable manipulations are shown below.

```r
# Create a vector of the all the outcomes
# with the correct number of repetitions.

pop<-rep(row.names(dat.2),times=dat.2$Vaccinated+dat.2$Placebo)
table(pop) # view the results
```

```
## pop
## FalseReports NonParalytic    Paralytic
##           45           51          148
```

```r
# Draw a sample of size k from the entries in this vector, that is a
permutation of length k.
k<-10
samp.perm<-sample(pop,k)

# Create a vector of the proportion of times each outcome was
# observed in the two groups put together.

outcome.prop<-
  (dat.2$Vaccinated+dat.2$Vaccinated)/sum(dat.2$Vaccinated+dat.2$Vaccinated)
outcome.prop
```

```
## [1] 0.4024390 0.2926829 0.3048780
```

```r
# Sample the vector ("Paralytic","NonParalytic","FalseReports") k times
according to the
# probabilites in "rating.prop"

k<-10
set.seed(34567)
samp<-
sample(c("Paralytic","NonParalytic","FalseReports"),k,replace=TRUE,prob=outco
me.prop)
samp
```

```
##  [1] "FalseReports" "NonParalytic" "Paralytic"    "Paralytic"
"FalseReports"
##  [6] "Paralytic"    "Paralytic"    "FalseReports" "FalseReports"
"FalseReports"
```

```r
# Total the number of each type of outcome in the sample.
counts<-table(samp)
counts
```

```
## samp
## FalseReports NonParalytic    Paralytic
##            5            1            4
```

```r
# Calculate the proportion each type of outcome in the sample.
props<-counts/sum(counts)
props
```

```
## samp
## FalseReports NonParalytic    Paralytic
##          0.5          0.1          0.4

# Make two vectors of proportions
## Start by drawing two samples
k1<-10
k2<-20
set.seed(34567)
samp1<-sample(c("Paralytic","NonParalytic","FalseReports"),k1,
              replace=TRUE,prob=outcome.prop)
samp2<-sample(c("Paralytic","NonParalytic","FalseReports"),k2,
              replace=TRUE,prob=outcome.prop)

## Total the number of each type of outcome in each sample.
counts1<-table(samp1)
counts2<-table(samp2)
## Calculate the proportion each type of outcome in each sample.
props1<-counts1/sum(counts1)
props2<-counts2/sum(counts2)

# Calculate the Euclidean distance between two vectors.
dist.eu<-sqrt(sum((props1-props2)^2))

# Calculate the sum of the absolute differences in each position for
# two vectors.
dist.mann<-sum(abs(props1-props2))
```