Problem Set 1

Megan Hoeksema

September 16, 2021

### 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)

## Warning: package 'HistData' was built under R version 3.6.3

dat<-PolioTrials  
dat$Paralytic[2:3]/dat$Population[2:3]

## [1] 0.0005714882 0.0003571660

### 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.

size.mytrials<-dat$Population[2]  
size.mytrials

## [1] 201229

prob.this<-(dat$Paralytic[2]+dat$Paralytic[3])/(dat$Population[2] + dat$Population[3])  
prob.this

## [1] 0.0004370314

1-pbinom(dat$Paralytic[2]-1, size=size.mytrials, prob=prob.this)

## [1] 0.003240697

pbinom(dat$Paralytic[2]-1, size=size.mytrials, prob=prob.this, lower.tail=FALSE)

## [1] 0.003240697

## 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.

dat.2$Vaccinated

## [1] 33 24 25

sum(dat.2$Vaccinated)

## [1] 82

row.names(dat.2)

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

dat.2$Vaccinated/sum(dat.2$Vaccinated)

## [1] 0.4024390 0.2926829 0.3048780

### 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.

# Create a vector of the all the outcomes with the correct number of repetitions.  
row.names(dat.2)

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

colnames(dat.2)

## [1] "Vaccinated" "Placebo"

names(dat.2)

## [1] "Vaccinated" "Placebo"

dat.2$Vaccinated+dat.2$Placebo

## [1] 148 51 45

rep(row.names(dat.2),c(3,5,2))

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

pop<-rep(row.names(dat.2),times=dat.2$Vaccinated+dat.2$Placebo)  
table(pop)

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

# 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

# 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=outcome.prop)  
samp

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

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

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

# 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))  
  
# Look at the Euclidean distance between vectors of proportions of outcomes  
# We expect that the proportions of the three outcomes in the populations  
# are equal to the pooled proportions.  
  
prop.vacc<-dat.2$Vaccinated/sum(dat.2$Vaccinated)  
prop.vacc

## [1] 0.4024390 0.2926829 0.3048780

prop.plac<-dat.2$Placebo/sum(dat.2$Placebo)  
prop.plac

## [1] 0.7098765 0.1666667 0.1234568

test.stat<-sqrt(sum((prop.vacc-prop.plac)^2))  
test.stat

## [1] 0.3785652

n<-100000  
  
k1<-sum(dat.2$Vaccinated)  
k1

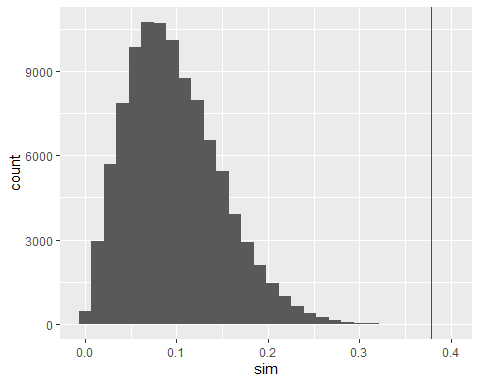
## [1] 82

k2<-sum(dat.2$Placebo)  
k2

## [1] 162

sim = rep(NA, n)  
  
for(i in 1:n){  
 samp1<-sample(c("Paralytic","NonParalytic","FalseReports"), k1, replace=TRUE, prob = outcome.prop)  
 samp2<-sample(c("Paralytic","NonParalytic","FalseReports"), k2, replace=TRUE, prob = outcome.prop)  
 counts1<-table(samp1)   
 counts2<-table(samp2)  
 props1<-counts1/sum(counts1)  
 props2<-counts2/sum(counts2)  
 dist.eu<-sqrt(sum((props1-props2)^2))  
   
 sim[i]<-dist.eu  
   
}  
  
dat.temp<-data.frame(sim=sim)  
  
ggplot(data = dat.temp, aes(x=sim))+geom\_histogram()+geom\_vline(xintercept = test.stat, color="red")

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



mean(sim>=test.stat)

## [1] 1e-05