Problem Set 1 Practice

sample size demo

C. Durso

# Introduction

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> ).

On problem sets, you will generate your solutions in R markdown. From an Rmd file in RStudio, you can generate a word document by selecting the “Knit to Word” option next to the “Knit” icon in the toolbar above the edit window. Please upload both a knitted doc, docx, or pdf document in addition to the Rmd file.

Please put your name in the “author” section in the header.

RStudio may ask you to install packages when you run this code. Accepting the request will allow you to proceed.

In this practice problem set, most of the code is provided. The challenge is to interpret the results according to the principles introduced in the polio case study in week 1.

# Load Data

data("PolioTrials")  
dat<-PolioTrials  
kable(dat[,1:4])

| Experiment | Group | Population | Paralytic |
| --- | --- | --- | --- |
| RandomizedControl | Vaccinated | 200745 | 33 |
| RandomizedControl | Placebo | 201229 | 115 |
| RandomizedControl | NotInoculated | 338778 | 121 |
| RandomizedControl | IncompleteVaccinations | 8484 | 1 |
| ObservedControl | Vaccinated | 221998 | 38 |
| ObservedControl | Controls | 725173 | 330 |
| ObservedControl | Grade2NotInoculated | 123605 | 43 |
| ObservedControl | IncompleteVaccinations | 9904 | 4 |

# Reproduce Hypothesis Test

The code below, included as a convenience, reproduces the results in 01\_polio\_simulation\_binomial\_model.Rmd for the full data set. Subsequently, you will carry out a similar analysis on the Vaccinated and the Placebo group for data with the population sizes reduced by a factor of 20 and the Paralytic counts also reduced by a factor of 20.

## Paralytic Polio Proportions

# ratio of the "Paralytic" count to the "Population" for  
# the "Vaccination" group for the "RandomizedControl" experiment  
(vacc.paralytic.prop<-dat$Paralytic[1]/dat$Population[1])

## [1] 0.0001643877

# ratio of the "Paralytic" count to the "Population" for  
# the "Placebo" group for the "RandomizedControl" experiment  
(placebo.paralytic.prop<-dat$Paralytic[2]/dat$Population[2])

## [1] 0.0005714882

# ratio of the pooled "Paralytic" cases to   
# the pooled "Population" value for the "RandomizedControl" experiment  
(pooled.paralytic.prop<-sum(dat$Paralytic[1:2])/sum(dat$Population[1:2]))

## [1] 0.000368183

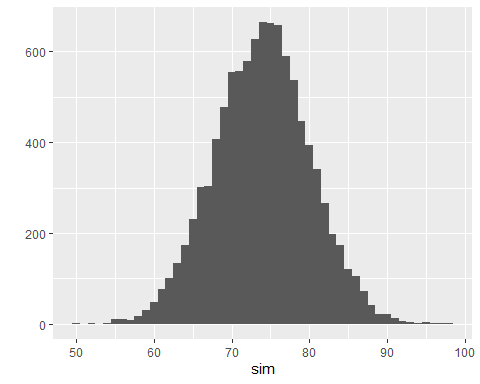
The null hypothesis is that the underlying proportion of Paralytic is the same in both populations.

Examination of the proportions and the historical context of the vaccine trials indicate that one can reasonably assume that if the Vaccinated and Placebo populations depart from having equal proportions of Paralytic, then the proportion of Paralytic is lower in the Vaccinated population.

## Hypothesis Test

The null model that there were going to be 148 cases and these were randomly assigned to the Vaccinated and Placebo groups is used to test the null hypothesis of equal proportions of Paralytic in the two populations. The test statistic is the number of Paralytic cases in the Vaccinated group.

n<-10000 # number of simulations  
  
# Calculate the number of paralytic polio cases in the pooled vaccination and placebo group.  
ct<-sum(dat$Paralytic[1:2])  
  
# Calculate the proportion "prop" of the the pooled vaccination and placebo group that are in the vaccination group.  
prop<-dat$Population[1]/sum(dat$Population[1:2])  
  
# Generate 10,000 counts of paralytic polio cases in the vaccination group under the model that each paralytic polio case in the pooled vaccination and placebo group has probability "prop" of being assigned to the vaccination group.  
set.seed(45678765)  
sim<-rbinom(n,ct,prop)  
  
# Plot a histogram of the simulated counts.  
qplot(sim,binwidth=1)



# Calculate the proportion of the simulated counts of paralytic polio in the "Vaccinated" group that are less than or equal to the observed count:  
mean(sim<=dat$Paralytic[1])

## [1] 0

The observed value of the count of Paralytic cases in the Vaccinated group, 33, is very unlikely under this null model. A value as small as or small that the observed value doesn’t occur in any of the 10,000 simulations according to the null model. This is strong evidence that the data are not consistent with this null model. Conclude that the data are not consistent with this model.

# Question 1

Please carry out the analysis below and answer the questions that follow.

## Context

The basic question “did the vaccine work?” was addressed in week 1 using the data from the randomized control trial of the Salk vaccine. The count of paralytic polio cases in the vaccinated group was compared to the counts that were produced under the null hypothesis that the vaccine had no effect on the incidences of paralytic polio. The second model examined in week 1 for this null hypothesis (01\_polio\_simulation\_binomial\_model.Rmd) was that each paralytic polio case in the pooled treatment and placebo group was assigned to the treatment group with probability equal to the ratio of the size of the treatment group to the size of pooled treatment and placebo group.  
Below, the computation used in that analysis is repeated in the hypothetical situation that the Vaccinated and Placebo groups in the RandomizedControl experiment were reduced by a factor of 20, with proportional reductions in the numbers of paralytic polio cases in the two groups.

Reduced data:

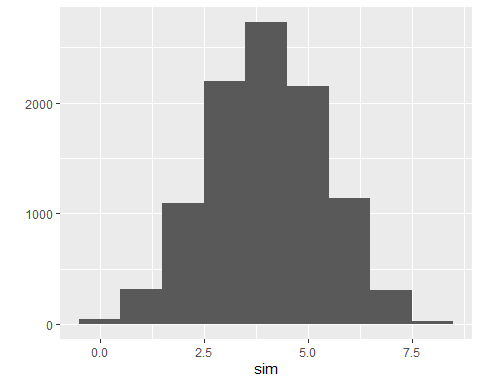
dat.reduced<-dat[1:2,1:4]  
dat.reduced$Population<-round(dat.reduced$Population/20,0)  
dat.reduced$Paralytic<-round(dat.reduced$Paralytic/20,0)  
kable(dat.reduced)

| Experiment | Group | Population | Paralytic |
| --- | --- | --- | --- |
| RandomizedControl | Vaccinated | 10037 | 2 |
| RandomizedControl | Placebo | 10061 | 6 |

Recall rbinom(n,ct,prop) is a function that models the number of random assignments to the distinguished group from a population of size ct if the probability of assignment to the distinguished group is prop. The value of n is the number of times to repeat the experiment.

### Test

n<-10000 # number of simulations  
  
# Calculate the number of paralytic polio cases in the pooled vaccination and placebo group.  
ct<-sum(dat.reduced$Paralytic[1:2])  
  
# Calculate the proportion "prop" of the the pooled vaccination and placebo group that are in the vaccination group.  
prop<-dat.reduced$Population[1]/sum(dat.reduced$Population[1:2])  
  
# Generate 10,000 counts of paralytic polio cases in the vaccination group under the model that each paralytic polio case in the pooled vaccination and placebo group has probability "prop" of being assigned to the vaccination group.  
set.seed(45678765)  
sim<-rbinom(n,ct,prop)  
  
# Plot a histogram of the simulated counts.  
qplot(sim,binwidth=1)



# Calculate the proportion of the simulated counts of paralytic polio in the "Vaccinated" group that are less than or equal to the observed count:  
mean(sim<=dat.reduced$Paralytic[1])

## [1] 0.146

# Save results for future reference  
reduction.factor<-c(20) # beginning of a vector to record the reduction factor used to generate the reduced data set  
extreme.proportion<-c(mean(sim<=dat.reduced$Paralytic[1])) # beginning of a vector to record the proportion of simulated counts less that or equal to the reduced "Paralytic" count in the reduced "Vaccinated" group

## Q1, part 1

1. What is the proportion of paralytic polio cases in the Vaccinated group in the reduced experiment? (your answer here)

0.0001992627 or 1.992627 × 10-4

1. What is the proportion of paralytic polio cases in the Placebo group in the reduced experiment? (your answer here)

0.0005963622 or 5.963622 × 10-4

1. What is the proportion of paralytic polio cases in the pooled Vaccinated and Placebo groups in the reduced experiment? (your answer here)

0.0003980496 or 3.980496 × 10-4

The following computations may be helpful.

# ratio of the "Paralytic" count to the "Population" for  
# the "Vaccination" group for the reduced data set  
dat.reduced$Paralytic[1]/dat.reduced$Population[1]

## [1] 0.0001992627

# ratio of the "Paralytic" count to the "Population" for  
# the "Placebo" group for the reduced data set  
dat.reduced$Paralytic[2]/dat.reduced$Population[2]

## [1] 0.0005963622

# ratio of the pooled "Paralytic" cases to   
# the pooled "Population" value for the reduced data:  
sum(dat.reduced$Paralytic[1:2])/sum(dat.reduced$Population[1:2])

## [1] 0.0003980496

## Q1, part 2

Is the observed number of paralytic polio cases in the Vaccinated group in the reduced data consistent with the probability model that each paralytic polio case in the pooled vaccinated and placebo group was assigned to the vaccinated group with probability equal to the ratio of the size of the vaccinated group to the size of pooled vaccinated and placebo group?

# some statistics of the simulated values under the probability model  
mean(sim)

## [1] 3.992

min(sim)

## [1] 0

max(sim)

## [1] 8

1. Please respond yes or no. (your answer here)

Yes

1. Please support your response with values generated in the test of the probability model above.

The paralytic polio cases in the vaccinated group in the reduced dataset is 2.

The proportion of paralytic polio cases in the vaccinated group in the reduced dataset is 0.25.

The proportion of paralytic polio cases in the pooled groups is 0.499398.

We can also see that the extreme proportion for the paralytic polio cases in the reduced dataset is 0.146. this is, the proportion that are less than or equal to 2.

## Q1, part 3

The code below repeats the simulation for a data set reduced by a factor of 2 from the original while preserving the Population proportions and the Paralytic proportions in the Vaccinated and the Placebo groups.

### factor of 2 calculations

Reduce the data:

dat.reduced<-dat[1:2,1:4]  
dat.reduced$Population<-round(dat.reduced$Population/2,0)  
dat.reduced$Paralytic<-round(dat.reduced$Paralytic/2,0)

Run the simulation:

n<-10000 # number of simulations  
  
# Calculate the number of paralytic polio cases in the pooled vaccination and placebo group.  
ct<-sum(dat.reduced$Paralytic[1:2])  
  
# Calculate the proportion "prop" of the the pooled vaccination and placebo group that are in the vaccination group.  
prop<-dat.reduced$Population[1]/sum(dat.reduced$Population[1:2])  
  
# Generate 10,000 counts of paralytic polio cases in the vaccination group under the model that each paralytic polio case in the pooled vaccination and placebo group has probability "prop" of being assigned to the vaccination group.  
set.seed(45678765)  
sim<-rbinom(n,ct,prop)  
  
# Calculate the proportion of the simulated counts of paralytic polio in the "Vaccinated" group that are less than or equal to the observed count:  
mean(sim<=dat.reduced$Paralytic[1])

## [1] 0

# Append the reduction factor to the vector of reduction factors.  
reduction.factor<-c(reduction.factor,2)  
# Append the proportion of sample statistics at least as extreme as the observed statistic to the vector proprtions of extreme values under the null hypothesis.  
extreme.proportion<-c(extreme.proportion,mean(sim<=dat.reduced$Paralytic[1]))

### factor of 5 calculations

Reduce the data:

dat.reduced<-dat[1:2,1:4]  
dat.reduced$Population<-round(dat.reduced$Population/5,0)  
dat.reduced$Paralytic<-round(dat.reduced$Paralytic/5,0)

Run the simulation:

n<-10000 # number of simulations  
  
ct<-sum(dat.reduced$Paralytic[1:2])  
  
prop<-dat.reduced$Population[1]/sum(dat.reduced$Population[1:2])  
  
set.seed(45678765)  
sim<-rbinom(n,ct,prop)  
  
# Calculate the proportion of the simulated counts of paralytic polio in the "Vaccinated" group that are less than or equal to the observed count:  
mean(sim<=dat.reduced$Paralytic[1])

## [1] 0.0032

# Append the reduction factor to the vector of reduction factors.  
reduction.factor<-c(reduction.factor,5)  
# Append the proportion of sample statistics at least as extreme as the observed statistic to the vector proprtions of extreme values under the null hypothesis.  
extreme.proportion<-c(extreme.proportion,mean(sim<=dat.reduced$Paralytic[1]))

### factor of 10 calculations

Reduce the data:

dat.reduced<-dat[1:2,1:4]  
dat.reduced$Population<-round(dat.reduced$Population/10,0)  
dat.reduced$Paralytic<-round(dat.reduced$Paralytic/10,0)

Run the simulation:

n<-10000 # number of simulations  
  
ct<-sum(dat.reduced$Paralytic[1:2])  
  
prop<-dat.reduced$Population[1]/sum(dat.reduced$Population[1:2])  
  
set.seed(45678765)  
sim<-rbinom(n,ct,prop)  
  
# Calculate the proportion of the simulated counts of paralytic polio in the "Vaccinated" group that are less than or equal to the observed count:  
mean(sim<=dat.reduced$Paralytic[1])

## [1] 0.0176

# Append the reduction factor to the vector of reduction factors.  
reduction.factor<-c(reduction.factor,10)  
# Append the proportion of sample statistics at least as extreme as the observed statistic to the vector proprtions of extreme values under the null hypothesis.  
extreme.proportion<-c(extreme.proportion,mean(sim<=dat.reduced$Paralytic[1]))

In all of the reduced data sets, the proportion of Paralytic cases in the Vaccinated population is approximately equal to the proportion of Paralytic cases in the Vaccinated population in the full data set. The proportion of Paralytic cases in the Placebo population is approximately equal to the proportion of Paralytic cases in the Vaccinated population in the full data set. Among the data set scaled down by a factor of 2, the data set scaled down by a factor of 5, the data set scaled down by a factor of 10, and the data set scaled down by a factor of 20 is the evidence, does the strength of the evidence against the null hypothesis increase with the size of the data set, decrease with the size of the data set, or show no obvious systematic relationship to the size of the data set?

(your answer here)