Problem Set 1

hello world

A. Dehesa

# 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> ).

Please 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 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 |

# 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.

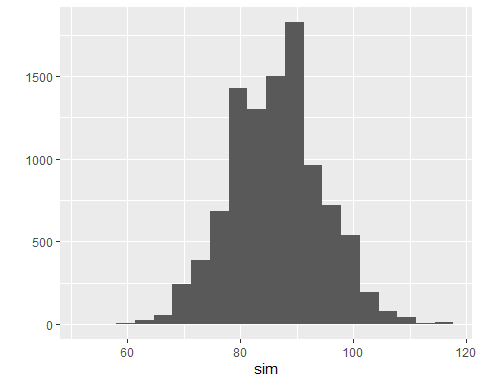
Below, the computation used in that analysis is repeated for the vaccinated group and the control group in the observed control trial. (Vaccinated and Controls in the ObservedControl experiment)

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 control group.  
ct<-sum(dat$Paralytic[5:6])  
  
# Calculate the proportion "prop" of the the pooled vaccination and control group that are in the vaccination group.  
prop<-dat$Population[5]/sum(dat$Population[5:6])  
  
# 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 control 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,bins=20)

## Warning: `qplot()` was deprecated in ggplot2 3.4.0.



# 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[5])

## [1] 0

## Q1, part 1

(10 points)

* What is the proportion of paralytic polio cases in the Vaccinated group in the ObservedControl experiment? (Your answer here) 0.0001711727
* What is the proportion of paralytic polio cases in the Controls group in the ObservedControl experiment? (Your answer here) 0.0004550638
* What is the proportion of paralytic polio cases in the pooled Vaccinated and Controls groups in the ObservedControl experiment? (Your answer here) 0.0003885254 The following computations may be helpful.

# ratio of the value in the 5th row of the "Paralytic"  
# column of "dat" to the value in the 5th row of the "Population"  
# column of "dat":  
dat$Paralytic[5]/dat$Population[5]

## [1] 0.0001711727

# ratio of the value in the 6th row of the "Paralytic"  
# column of "dat" to the value in the 6th row of the "Population"  
# column of "dat":  
dat$Paralytic[6]/dat$Population[6]

## [1] 0.0004550638

# ratio of the sum of the values in the 5th and 6th row of   
# the "Paralytic" column of "dat" to the sum of the values in the   
# 5th and 6th row of the "Population" column of "dat":  
sum(dat$Paralytic[5:6])/sum(dat$Population[5:6])

## [1] 0.0003885254

## Q1, part 2

(15 points)

Is the observed number of paralytic polio cases in the Vaccinated group in the ObservedControl experiment consistent with the probability model that each paralytic polio case in the pooled vaccinated and control 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 control group?

# ratio of the value in the 5th row of the "Paralytic"  
# column of "dat" to the value in the 5th row of the "Population"  
# column of "dat":  
dat$Paralytic[5]/dat$Population[5]

## [1] 0.0001711727

# ratio of the value in the 6th row of the "Paralytic"  
# column of "dat" to the value in the 6th row of the "Population"  
# column of "dat":  
dat$Paralytic[6]/dat$Population[6]

## [1] 0.0004550638

mean(sim)

## [1] 86.2878

min(sim)

## [1] 54

max(sim)

## [1] 117

# Actual observed value  
dat$Paralytic[5]

## [1] 38

(Your answer here. Please explain your yes or no response and support it with values generated in the test of the probability model above.)

No, it is not consistent. We can see in the mean for the simulation, which is 86, that most of the simulated values did not even approach the actual observed value, which is 38. This means that 38 was not obtained only by chance. There must have been something that caused this result. Therefore, the it is not consistent.

## Q1, part 3

(10 points)

Using your conclusion in part 2, can the data from the ObservedControl experiment be interpreted as evidence that the vaccination *causes* a reduction in the likelihood of contracting paralytic polio? Please explain. Recall that the Vaccinated group consists of second graders whose parents consented to vaccination while the Controls group consists of first and third graders.

(Your answer here. Please explain your yes or no response as it relates to the data and the design of the ObservedControl experiment.) I believe that this cannot be interpreted as evidence that the vaccination causes a reduction in the likelihood of contracting paralytic polio. This data can only be used to refute the null hypothesis that the vaccine does nothing. The Observed and Control groups consist of first and third graders for which the parents gave consent. This is a big bias, for age, possibly social groups, and, as the parents consented, possibly even education level of the parents, which may indicate a possible bias in the income level of families, which may indicate a bias in the health of the children, such as diet habits.

# Question 2

Please carry out the analysis below and answer the question that follows.

## Context

In this section, you will address the question of whether the NotInoculated and Placebo groups in the Randomized Control experiment had statistically significantly different rates of paralytic polio.

Recall that the NotInoculated and Placebo groups differ in that the children in the Placebo group had been enrolled in the vaccine trial while the parents of the children in the NotInoculated group did not enroll their children. If the home environments in these groups differed systematically in ways related to susceptibility to paralytic polio, rates in the two groups could differ beyond the amount readily explained by chance. If the enrolled and not enrolled populations didn’t differ in relevant ways, the differences should be fairly consistent with chance.

The approach, using the rbinom function again, implements the idea that populations in the NotInoculated and Placebo groups in the Randomized Control experiment were the same in regards to paralytic polio cases by using the rbinom function to assign paralytic polio cases in the combined NotInoculated and Placebo groups of the RandomizedControl experiment to the Placebo group with probability equal to the ratio of the size of the Placebo group to the size of pooled Placebo group and NotInoculated group.

### Data summary

# proportion of paralytic polio cases in "Placebo"  
dat$Paralytic[2]/dat$Population[2]

## [1] 0.0005714882

# proportion of paralytic polio cases in "NotInoculated"  
dat$Paralytic[3]/dat$Population[3]

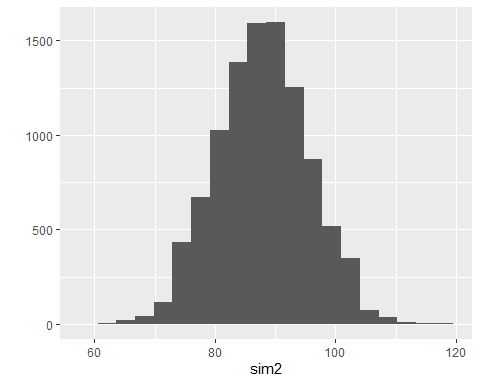
## [1] 0.000357166

# proportion of paralytic polio cases in "Placebo" and "NotInoculated" combined  
sum(dat$Paralytic[2:3])/sum(dat$Population[2:3])

## [1] 0.0004370314

### Test

n<-10000 # number of simulations  
  
# Calculate the number of paralytic polio cases in the pooled "Placebo" and "NotInoculated" group.  
ct<-sum(dat$Paralytic[2:3])  
  
# Calculate the proportion "prop" of the the pooled "Placebo" and "NotInoculated" group that are in the "Placebo" group.  
prop<-dat$Population[2]/sum(dat$Population[2:3])  
  
# Generate 10,000 counts of paralytic polio cases in the "Placebo" group under the model that each paralytic polio case in the pooled pooled "Placebo" and "NotInoculated" group has probability "prop" of being assigned to the "Placebo" group.  
set.seed(45678765)  
sim2<-rbinom(n,ct,prop)   
qplot(sim2,bins=20)



# proportion of the simulated counts of paralytic polio in the "Placebo" that are less than or equal to the observed count:  
mean(sim2<=dat$Paralytic[2])

## [1] 0.9997

# proportion of the simulated counts of paralytic polio in the "Placebo" that are greater than or equal to the observed count:  
mean(sim2>=dat$Paralytic[2])

## [1] 4e-04

# Q2

(15 points)

Consider the null hypothesis that children who receive the placebo and children whose parents don’t consent to participation have the same susceptibility to paralytic polio. One possible probability model for this is that each case in the pooled Placebo and NotInoculated paralytic polio cases can be viewed as being assigned to the Placebo group with probability equal to the proportion of the Placebo population in the pooled Placebo and NotInoculated population. Are these data consistent with the null hypothesis? Please explain your conclusion and support it by referring to the values generated in simulation above.

# some statistics of the simulated values under the probability model  
mean(sim2)

## [1] 87.9619

min(sim2)

## [1] 59

max(sim2)

## [1] 118

(Your answer here. Please explain your yes or no response.) No, it is not consistent. As we can see, in the simulations most of the results were lower than the actual observed results. If the susceptibility to paralytic polio were the same, then the actual observed results would have been lower. We can actually see this in the results themselves, as the number of the placebo population is only 201229, and the paralytic polio cases within this population is 115. A proportion of 0.00057, while the not inoculated population is 338778, and the paralytic polio cases within this population is 121. A proportion of 0.000357166, which is lower. Therefore, we can reject the null hypothesis. Logically, I would assume that they should have the same susceptibility, or even the non inoculated polution shoud have a higher sussceptibility than the placebo population, but the data does not show that.