Problem Set 2

applications of probability theory

Antonio 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 and 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.

The questions in this problem set use material from the slides on discrete and continuous probability spaces and the Rmds Discrete\_Probability\_Distributions\_2\_3\_3.Rmd and 02\_continuous\_probability\_distributions\_in\_r.rmd

# Load Data

data("PolioTrials")  
dat<-PolioTrials

# Question 1

Please carry out the analysis below and answer the questions that follow. For this assignment, please do all calculations in R and show the code and the results in the knit document.

## Context

Question 2 on problem set 1 addresses 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.

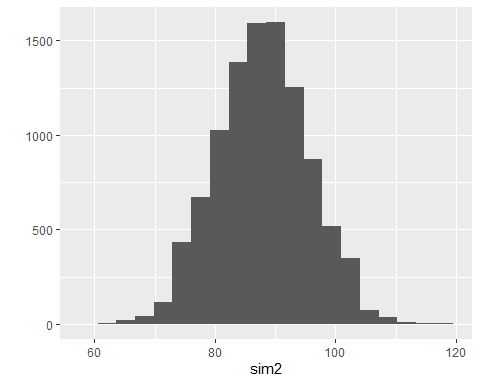
The approach, using the rbinom function, implemented the idea that populations in the NotInoculated and Placebo groups in the RandomizedControl experiment were the same in regards to paralytic polio cases by using the rbinom function to assign paralytic polio cases in the combined NotInoculated and Placebo 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.

Note that the function rbinom(x,size,prob) simulates drawing random samples from Binom(size,prob).

The computations for that analysis are reproduced here:

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)

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



# 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

## Q1, part 1

(10 points)

Using the same null model described above, please calculate the probability that the count of paralytic polio cases in the Placebo group under the null model is less than or equal to dat$Paralytic[2] directly rather than by simulating it. Recall that the function pbinom(x,size,prob) returns the probability of the event that the number of successes is in the set .

(Your answer here

# Please be sure that your computed probability shows in your knitted solutions  
res<-pbinom(dat$Paralytic[2],  
 ct,  
 prop)

With this result, we can see that the success rate that there are less than or equal paralytic polio cases as 115, with probability equal to prop, in the Placebo group, all assigned to the placebo group, is 0.999987. Which is very similar to the simulated value of 0.9997. )

## Q1, part 2

(10 points)

Using the same null model described above, please calculate the probability that the count of paralytic polio cases in the Placebo group under the null model is greater than or equal to dat$Paralytic[2] directly rather than by simulating it. Hint: Denote the value in part 1 by . This answer is not . The value is the probability of the event that count of paralytic polio cases in the Placebo group under the null model is strictly greater than dat$Paralytic[2].

(Your answer here

# Please be sure that your computed probability shows in your knitted solutions  
temp<-pbinom(dat$Paralytic[2]-1,  
 ct,  
 prop)  
res2 <- 1- temp

In order to solve this, we can reduce datParalytic[2]

By doing this, as our res2 is not equal to 1-P, p being res, then res + res2 should NOT be equal to 1. This is not confirmation that our result is correct, is just confirmation that our result is not immediately incorrect. )

## Q1, part 3

(10 points)

Is the value computed in part 2 strong evidence against the null model? (your yes or no answer here, with an explanation based on the calculations above) No, I do not believe this to be the case. In fact, I believe this is the opposite. We can see that, for Q1,P1, the probability for there being fewer paralytic cases with the probability being equal to the proportion of the placebo population in relation to the combined population of placebo and not inoculated is over 99% And the probability of it being equal to or more than the registered paralytic polio cases in the placebo group is less than 1%. This tells us that indeed, the placebo group is getting more paralytic polio cases that what it should. Now, it can be considered statistically relevant. If the probability were lower, like 80%, it could be considered chance. But 99% is not. We can also see this in the direct proportion of paralytic polio cases in relation to the placebo and not inoculated populations.

# Proportion of paralytic polio cases to the population of the placebo group  
dat$Paralytic[2]/ dat$Population[2]

## [1] 0.0005714882

# Proportion of paralytic polio cases to the population of the not inoculated group  
dat$Paralytic[3] / dat$Population[3]

## [1] 0.000357166

As we can see, the placebo group is almost double that of the not inoculated group.

# Question 2

## Context

This question concerns the uniform distribution on , the continuous probability space with and defined by for measurable sets as described in the week 2 slides. This distribution will be important in hypothesis testing.

## Q2

(10 points)

Are the events and independent. To answer this, please address the following questions:

1. What is ? (your answer here)

# P(A)  
pa<-punif(0.5,0,1)

1. What is ? (your answer here)

# P(3/4) - P(1/4) = P([1/4,3/4])  
pb<-punif(3/4,0,1)-punif(1/4,0,1)

1. What is ? (your answer here)

# Intersection of set A(0:0.5) and B(1/4:3/4) is R(1/4:1/2)  
panb<-punif(1/2,0,1)-punif(1/4,0,1)

1. Are the events and independent? (your answer here: please answer yes or no and explain your response using the calculations above.)

# These should be independent if panb == pa\*pb.   
panb == pa\*pb

## [1] TRUE

# As we can see, it fulfills the conditions, therefore, they are independent.

# Question 3

## Context

This question concerns the standard Normal distributions, the continuous probability space with and defined by for measurable sets as described in the week 2 slides. This distribution will be essential in future methods.

## Q3

(10 points)

Are the events and independent. To answer this, please address the following questions:

1. Please give a numerical approximation to . (your answer here

# We can omit the parameters for the mean and sd, as they are, by default, 0 and 1 respectively, which are the values for the standard normal distribution  
psa<-pnorm(0.5)-pnorm(0)

)

1. Please give a numerical approximation to ? (your answer here

psb<-pnorm(3/4)-pnorm(1/4)

)

1. Please give a numerical approximation to . (your answer here

# In this case, if we cconsider A as the interval from 0 to 1/2, and B as the interval from 1/4 to 3/4, then their intersection should be from 1/4 to 1/2.   
psanb<-pnorm(1/2)-pnorm(1/4)

)

1. Are the events and independent? (your answer here: please answer yes or no and explain your response using the calculations above.)

psanb == psa\*psb

## [1] FALSE

As we can see, the condition is not fulfilled, therefore, the events are not independent.