Homework 1

Joshua Oswari

14 April, 2019

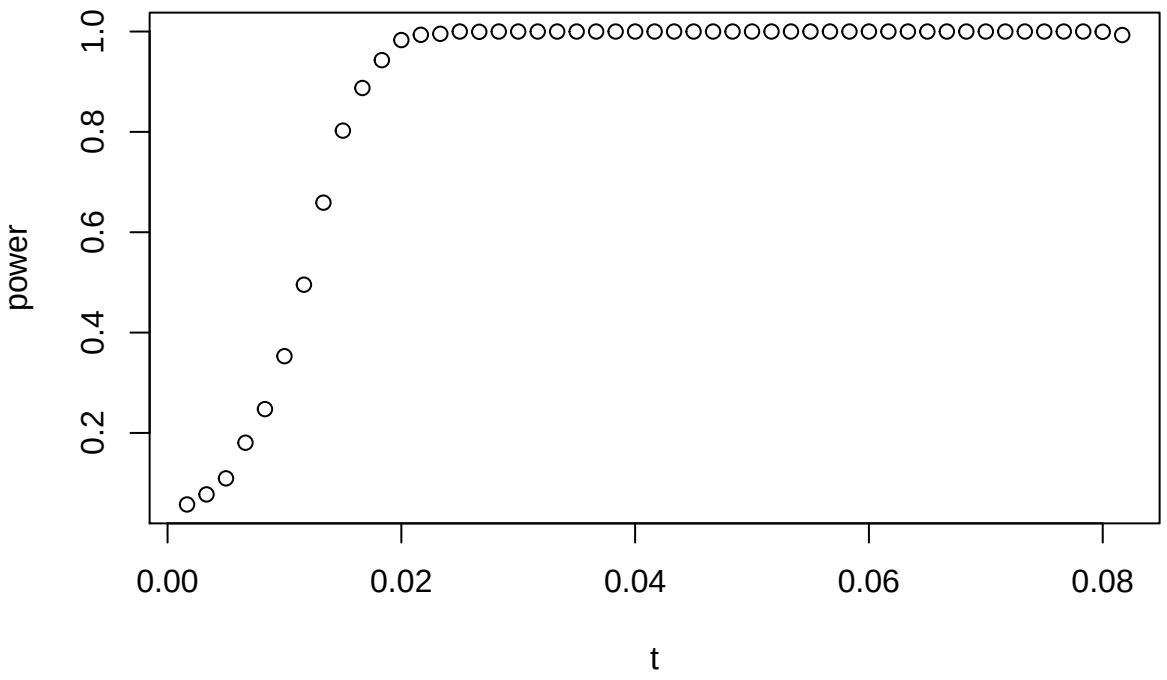
Write an R function chisq.power(k, t, n, B = 2000).

Problem 1 - Part A

```
chisq.power <- function(k, t, n, B=2000){
  R = numeric(B)
  #create the multinomial probaility vector
  probVector = numeric(2*k)
  for(i in 1:k){
    probVector[i] = ((1/(2*k))+t)
  }

  for(i in (k+1):(2*k)){
    probVector[i] = (1/(2*k)-t)
  }

  #for loop
  for(b in 1:B){
    ptvec = sample(1:(2*k), n, replace = TRUE, prob=probVector)
    tableloop = table(ptvec)
    if(chisq.test(tableloop)$p.value <= 0.05){
      R[b] = 1
      }
  }

  return(sum(R)/B)
}
```

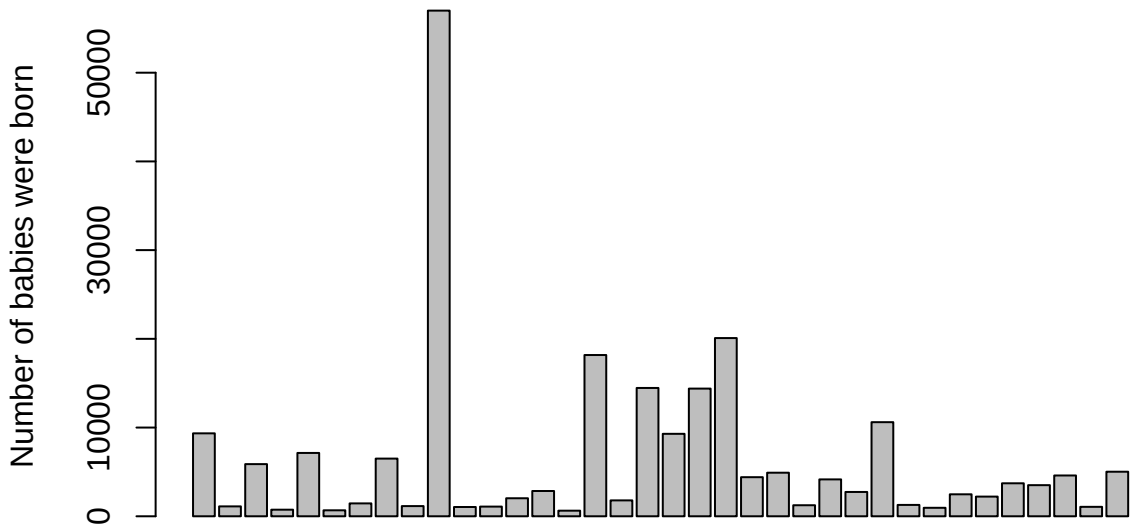
Problem 1 - Part B

```
#Fix k = 6, and using the function you just wrote, plot the (estimated) power curve of the
#chi-squared test as a function of t.

k = 6
n = 500
t = seq(1/(100*k), 1/(2*k)-1/(100*k), 1/(100*k))
ind = 1
power = numeric(length(t))
  for(i in t){
      power[ind] = chisq.power(k, i, n , B=2000)
        ind = ind +1
  }
plot(t, power, main = "Power Curve", xlab="t", ylab="power")
```



Problem 2
Hypotesis testing problem if the chance of a baby being born
a girl the same across counties in California? using the data taken from
http://wonder.cdc.gov/natality.html
Group by : Gender-County
State : California
Year : 2017

```
#read the table from the website
dat = read.table('~/Documents/Math185/natality-california-2017.txt',sep="\t", nrows=75, skip =1)
gender = substr(dat[,6], 1, 5)
Femalesdat = dat[1:36,6]

#conduct a hypotesis testing
#H0: The chance of baby girl being born is the same across the county of California
#H1: The chance of baby girl being born is not the same across the county of California

#first, draw the barplot so that we can know how the distribution on each county from female and male
barplot(Femalesdat, main = "Female baby born across California county",
        xlab = "County in California", ylab= "Number of babies were born")
```




```
#write down some helpful statistics
# Summary statistics
summary(Femalesdat)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    631    1220    3178    6382    6661   57000
```

```
mean(Femalesdat)
```

```
[1] 6382.222
```

```
median(Femalesdat)
```

```
[1] 3178.5
```

```
sd(Femalesdat) # same as sqrt(var(Femalesdat))
```

```
[1] 10038.91
```

```
mad(Femalesdat)
```

```
[1] 3078.619
```

```
# test against the uniform distribution
chisq.test(Femalesdat)
```

```
    Chi-squared test for given probabilities

data:  Femalesdat
X-squared = 552670, df = 35, p-value < 2.2e-16
```

```
# based on observation, we can conclude by the barplot that we drew that
# the chance of baby girl being born is not the same across the county of California
# it is very distributed on each counties and even from the chisq.test
# the df is 35, which means that the data is VERY distributed
```

Problem 3

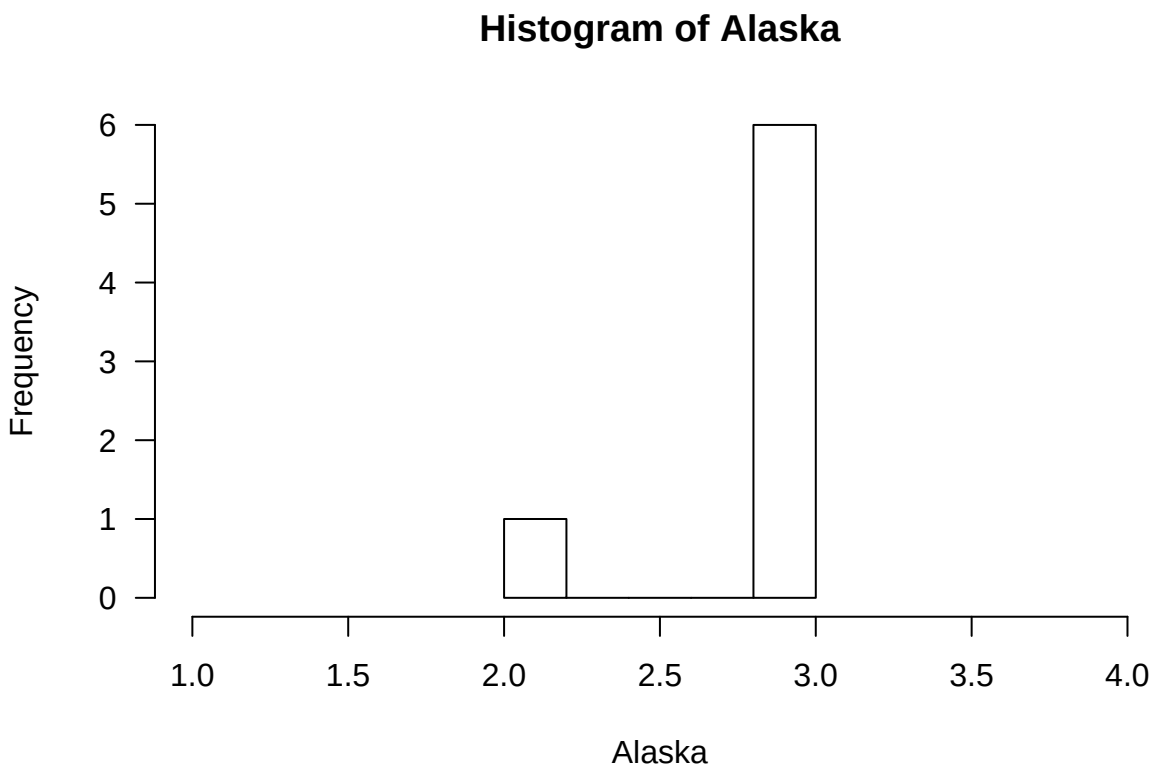
```
chisq.perm.test<- function(tab, B = 2000){
  D = chisq.test(tab)$statistic
  count = 0
  p.value = chisq.perm.test(tab)
  p.value
}

HairEyeColor
```

```
, , Sex = Male

       Eye
Hair    Brown Blue Hazel Green
  Black    32   11    10     3
  Brown    53   50    25    15
  Red      10   10     7     7
  Blond     3   30     5     8

, , Sex = Female

       Eye
Hair    Brown Blue Hazel Green
  Black    36    9     5     2
  Brown    66   34    29    14
  Red      16    7     7     7
  Blond     4   64     5     8
```

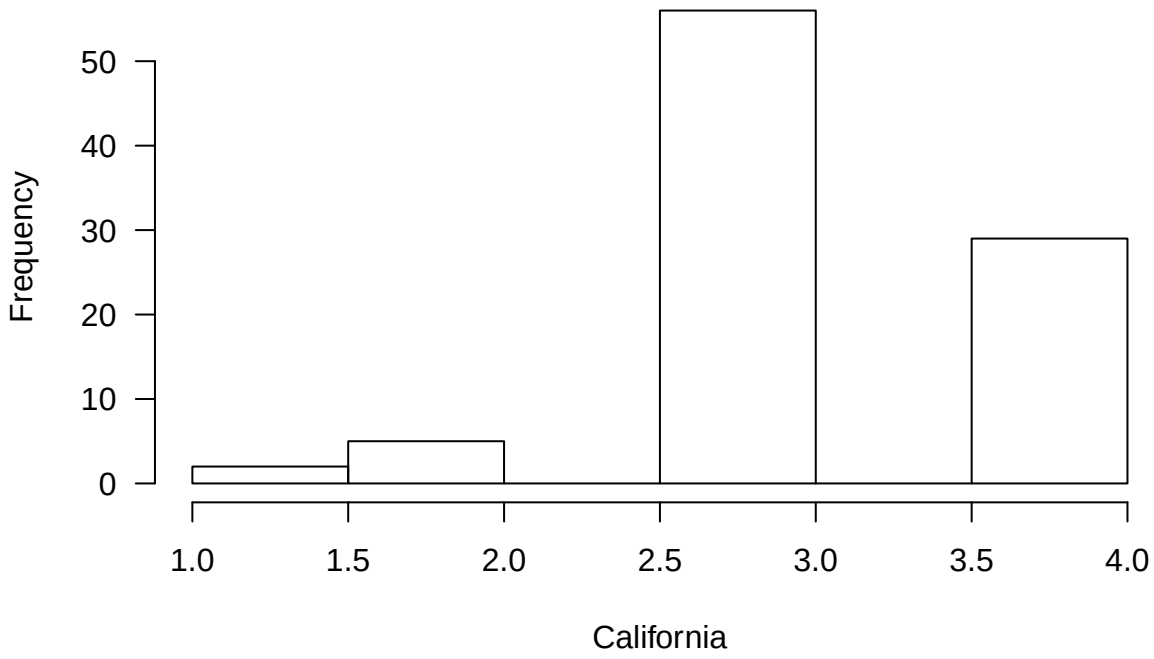
Problem 4 - Part A

```
#  Is there an association between the model that each
#  school selected and the state where the school was located at that time?
#  Explore this question with one or several appropriate plots. Then formulate the question into
#  a hypothesis testing problem and perform a test. Conclude with some brief comments.


#create the metadata
#turn csv->rda.
setwd("~/Documents/Math185")
dat = read.csv("~/Documents/Math185/school-improvement-2010.csv", header = TRUE, sep=",")
save(dat, file='school-improvement-2010.rda')
rm(dat)
load('~/Documents/Math185/school-improvement-2010.rda')

data = dat[,c(3,6)]
#omit Rhode Island Data
newData = data[-seq(662,667,1),]
colnames(newData)=c("States", "Model")

#transform each variable into numeric (i.e Closure: 1, Restart:2, Transformation:3, Turnaround:4)
newData$Model.num <- as.numeric(newData$Model)-1

#H0: There is an association between the model that each school selected
#    and the state where the school was located at that time
```

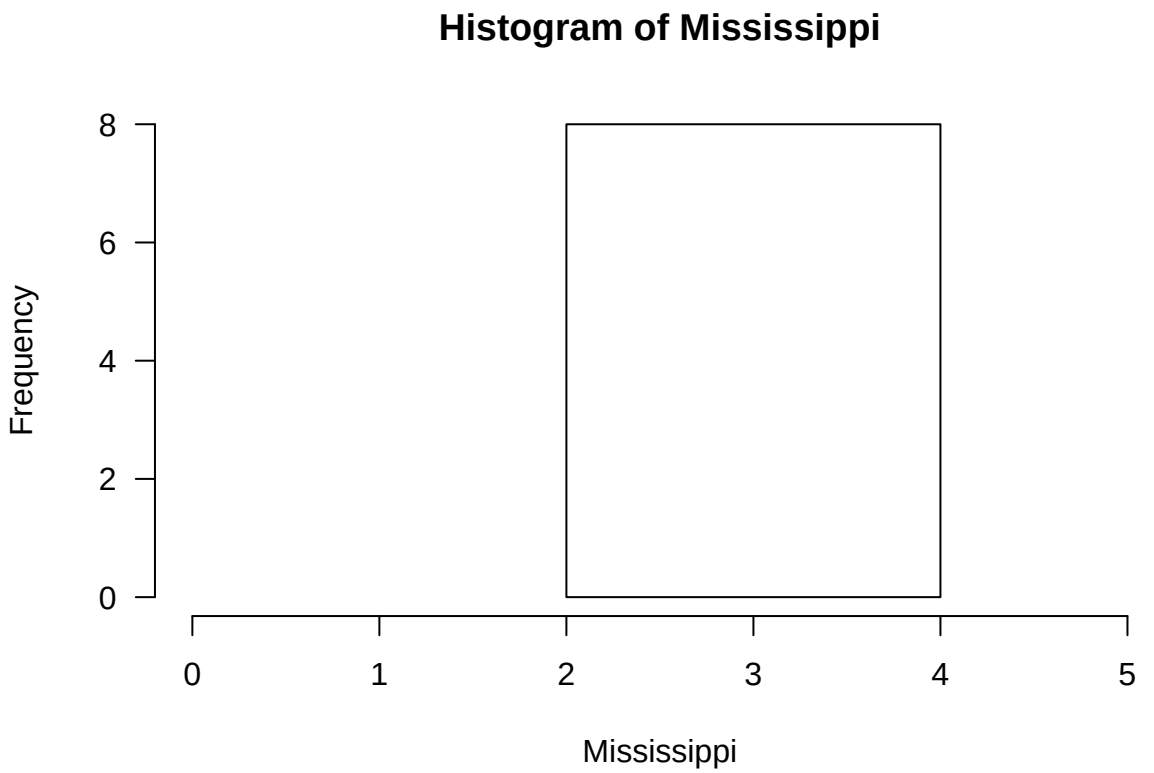
```r
#H1: There is no connection between the model that each school selected
#    and the state where the school was located at that time

#take sample from several states
Alaska = newData[1:7,3]
hist(Alaska, freq=TRUE, xlim=c(1,4), las=1 )
```



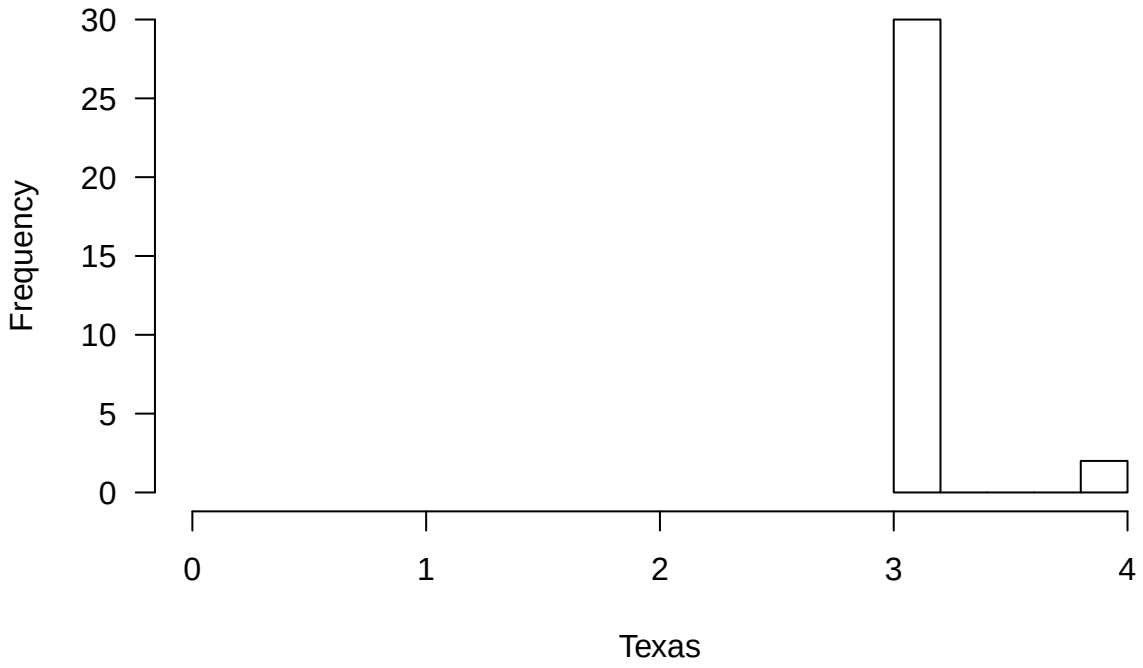

```r
California = newData[45:136,3]
hist(California, freq=TRUE, xlim=c(1,4), las=1 )
```




```
Mississippi = newData[434:441,3]
hist(Mississippi, freq=TRUE, xlim=c(0,5), las=1 )
```




```
Texas = newData[702:733,3]
hist(Texas, freq=TRUE, xlim=c(0,4), las=1 )
```




```
#based on the histogram, we can accept the H0 because the model are dependent
#to its States. Here, we can take an example from states Mississippi which
#only has one model which is the 3(Transformation) and we can take another example
#from texas which the majority of it is also 3, which is also Transformation.
#Therefore, we can conclude the Model are depend on the states.
```