MATH 185 - Computational Statistics - Homework 1

Dukki Hong

15 April, 2019

Problem 1

Part A

```
chisq.power <- function(k,t,n,B = 2000){
  R <- vector(mode='numeric', length = B)     #create an empty binary vector of length B

  null <- numeric(2*k)                        #data under the null, uniformly distributed

  for (i in 1:k){                             #Generate data for Observed counts
    null[i] <- 1/(2*k) + t
  }
  for (i in (k+1):(2*k)){
    null[i] <- 1/(2*k) - t
  }

  #Monte Carlo Simulation
  for(i in 1:B){
    sim = sample(1:(2*k), n, replace = T, prob = null)
    sim = table(sim)
    if (chisq.test(sim)$p.value <= 0.05){
      R[i] = 1
    }
    else{
      R[i] = 0
    }
  }

  return(sum(R)/B) #return proportion of correctly rejecting H_0 over H1
}
```

Part B

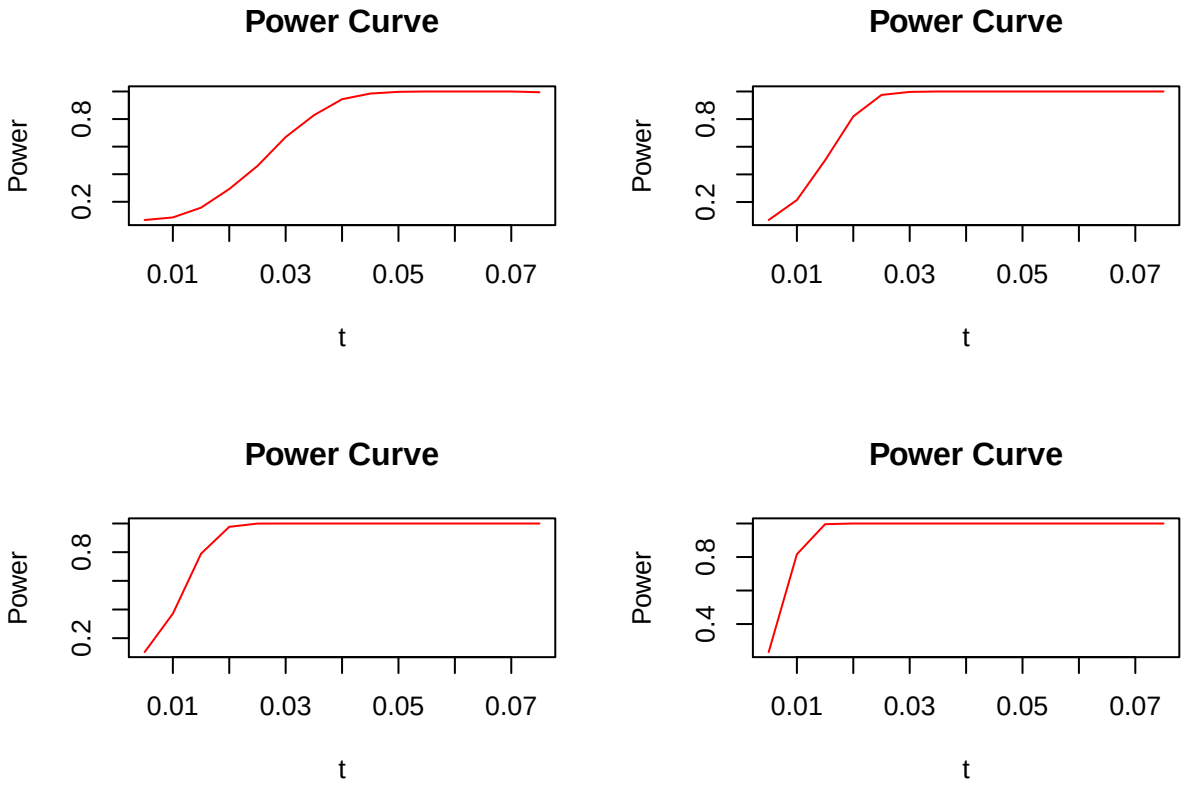
```
k = 6
t = seq(0+0.005, 1/(2*k) - 0.005, by =0.005) #range of values for t
n = c(100,300,500,1200) #set different values for n

pwr = numeric(length(t)) #powers

ylab = ('Power')
main = ('Power Curve')
col = ('red')

par(mfrow = c(2,2))
for (j in 1:length(n)) { #for loop to calculate over all n values
  for (i in 1:length(t)) {
```

```
    pwr[i] <- chisq.power(k, t[i], n[j])
    i = i + 1
    }
  plot(t, pwr, type = 'l', xlab = 't', ylab = ylab, main = main, col=col)
}
```



Problem 2

Are the chance of a baby being born a girl the same across counties in California? Go to http://wonder.cdc.gov/natality.html. In Section 1, choose Gender and County. In Section 2, choose the state to be California. In Section 4, choose the year 2017. Then click on Send" anywhere. Enter the data in R. You can do so by hand or click on Export“, edit the resulting text

le and then use the function read.table. (You will have to edit the TXT

le in order to read it into R.) Save the dataset as an RDA

le named natality-california-2017.rda. Do this as a preprocessing. Start your code by loading the dataset using the function load.

Use these data to answer the question the best you can. Start by formalizing the question into a hypothesis testing problem, display relevant summary statistics and graphics, and then perform an appropriate test. Conclude with a sentence or two.

```
mydata <- read.table("~/Desktop/UCSD MATH/Math.185/Natality, 2007-2017.txt", header = TRUE)
save(mydata, file="~/Desktop/UCSD MATH/Math.185/natality-california-2017.rda")

load("~/Desktop/UCSD MATH/Math.185/natality-california-2017.rda")

mydata = data.frame(mydata) #change mydata into a dataframe
mydata = subset(mydata, select=c('Gender', 'County.Code','Births')) #remove unnecessary columns
```

```r
#Are the chance of a baby being born a girl the same across counties in California?
#Parameter of interest: chance of baby being born a girl (p_i)
#H_0: chance of baby being born a girl SAME across counties in CA
#H_1: chance of baby being born a girl DIFFERENT across counties in CA

mydata = mydata[order(mydata$County.Code),] #sort out 'County.Code' column in an increasing order

female = mydata[seq(1,72,2),] #sort out mydata with females only
male = mydata[seq(2,72,2),] #sort out mydata with males only

f = matrix(female$Births, nrow = 1, ncol=36) #turn data into matrices
m = matrix(male$Births, nrow = 1, ncol=36)

X = rbind(f,m) #combine the two matrices into one matrix
row.names(X) = c('Female','Male') #Label row names

chisq.test(X)
```

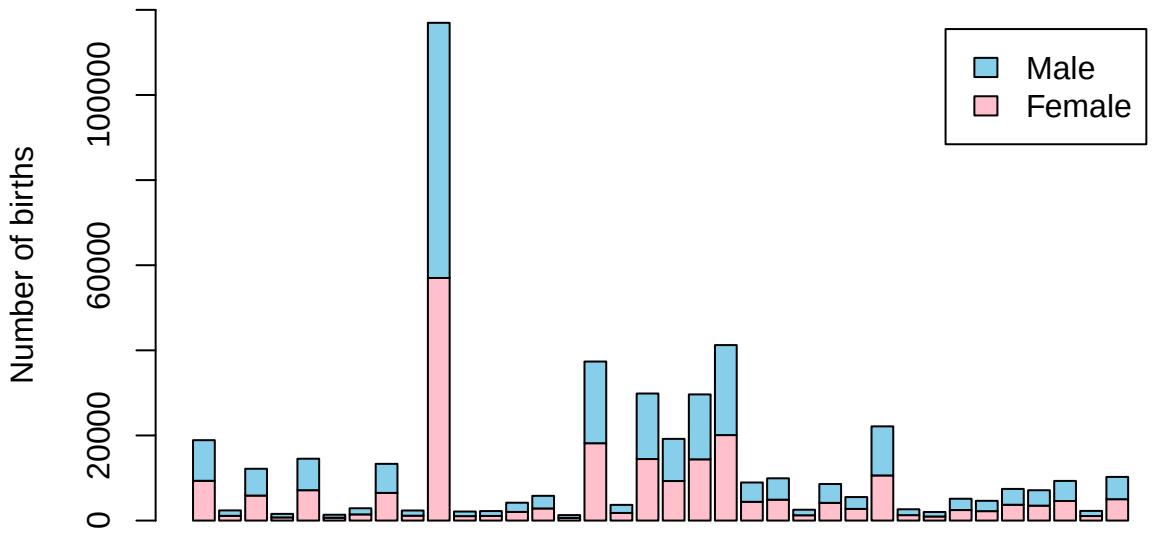
```
##
##  Pearson's Chi-squared test
##
## data:  X
## X-squared = 41.285, df = 35, p-value = 0.215
```

```r
cat('value of X-Squared is', '41.285,', 'df is 35', 'and p-value is:', '0.215.')
```

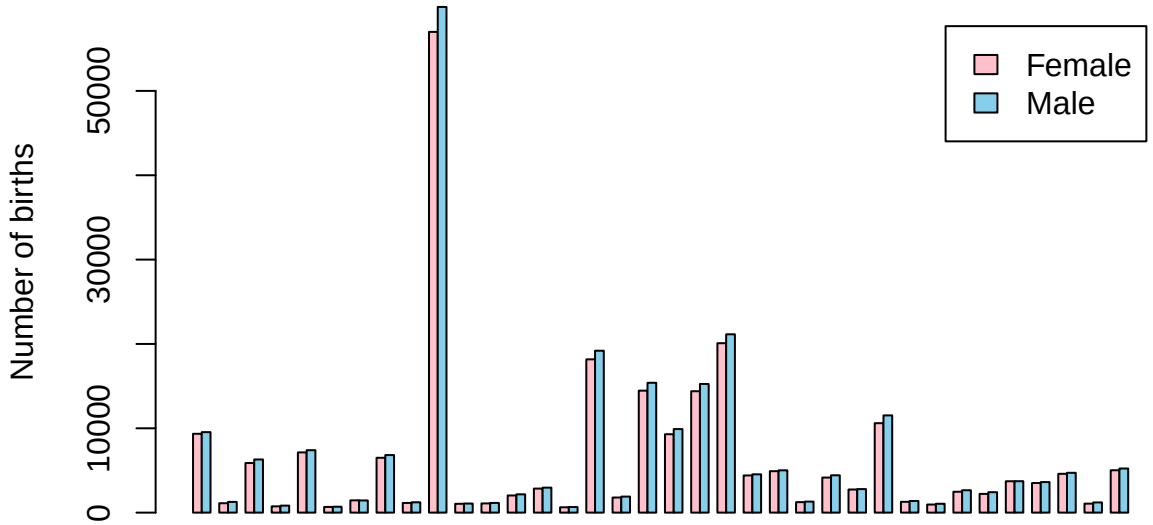
```
## value of X-Squared is 41.285, df is 35 and p-value is: 0.215.
```

Now we proceed to showing a graphical representation of the data

```r
col = c('pink','skyblue') #set color
ylab = ('Number of births') #Y-label name
barplot(X, legend=T, col=col, ylab=ylab, ylim=c(0,120000))
```




```r
barplot(X, , beside= T, legend=T, col=col, ylab=ylab)
```



We see similarities between Male and Female groups in the graph above

Problem 3

Write a function *chisq.perm.test*(tab;B = 2000) implementing the chi-squared test of independence calibrated by permutation. The inputs are a table of joint counts (without totals) and number of permutations to be done. Have the test return the p-value. Test your function on the *HairEyeColor* dataset.

```
chisq.perm.test <- function(tab, B=2000){
  obs = chisq.test(tab)$stat #observed statistic from the Chisqd test

  nrow = nrow(tab) #Number of rows
  ncol = ncol(tab) #Number of columns

  rows = numeric(nrow) #create a 0 vector with n number of rows
  totrow = numeric(0) #create a vector with 0's

  for (i in 1:nrow){ #for-loop to fill up the vectors defined above
    rows[i] = sum(tab[i,]) #summing up each of the rows
    totrow = c(totrow, rep.int(i, times=rows[i]))
  }

  #do the same with the columns
  cols = numeric(ncol) #empty vector with n columns
  totcol = numeric(0) #0 vector

  for (i in 1:ncol){
    cols[i] = sum(tab[,i]) #summing up each of the columns
    totcol = c(totcol, rep.int(i, times=cols[i]))
  }

  D = numeric(0) #vector for permutations for the Chisqd test statistic
  tot = sum(tab[,]) #total number of counts

  ######################################

  for (b in 1:B){
    x = totrow
    y = sample(totcol, tot, replace=FALSE) #sample column totals 'tot' times without replacement
```

```
    matperm = cbind(x,y) #combine columns of x and y

    matpermcount = matrix(0, nrow=nrow, ncol=ncol) #create a 0 matrix with n number of rows and columns

    for(x in 1:nrow){ #rows from 1 - 4
      for(y in 1:ncol){ #rows from 1 - 4
        for(m in 1:tot){ #loop for going through all the rows in matperm
          if(all(matperm[m,] == c(x,y)) == TRUE){ #ordered pairs such as (1,1), (2,2) and etc...
            matpermcount[x,y] = matpermcount[x,y] + 1
          }
        }
      }
    }

    D[b] = chisq.test(matpermcount)$stat #chisq.test statistic for a given permutation

    #Check if chisq.test stat >= observed stat
    count = 0 #counts to check (chisq.test stat > observed)
    if (D[b] >= obs){
      count = count + 1
    }
    return((count+1)/(B+1))
  }
}
```

Now we check if the function works

```
tab = apply(HairEyeColor, c(1,2), sum)
tab
```

```
##        Eye
## Hair    Brown Blue Hazel Green
##   Black    68   20    15     5
##   Brown   119   84    54    29
##   Red      26   17    14    14
##   Blond     7   94    10    16
```

```
chisq.perm.test(tab)
```

```
## [1] 0.0004997501
```

```
cat('We get that the Chi-squared test statistic is equal to :',chisq.perm.test(tab))
```

```
## We get that the Chi-squared test statistic is equal to : 0.0004997501
```

Problem 4

Go to the following webpage: https://catalog.data.gov/dataset/school-improvement-2010-grants Download the dataset. The description is on the webpage. To read it into R, use the function read.csv. Remove the schools from Rhode Island as their selected models are missing. Do this on your own to practice reading datasets into R. Save the dataset as an RDA file named school-improvement-2010.rda. Do this as a preprocessing. Start your code by loading the dataset using the function load.

```r
#Loading and cleaning data
data = read.csv('~/Desktop/UCSD MATH/Math.185/userssharedsdfschoolimprovement2010grants.csv', head=T)

head(data)
```

```
##                         School.Name     City State
## 1 HOGARTH KINGEEKUK MEMORIAL SCHOOL SAVOONGA    AK
## 2                   AKIACHAK SCHOOL AKIACHAK    AK
## 3                    GAMBELL SCHOOL  GAMBELL    AK
## 4              BURCHELL HIGH SCHOOL  WASILLA    AK
## 5                      AKIAK SCHOOL    AKIAK    AK
## 6                    MIDVALLEY HIGH  WASILLA    AK
##                               District.Name X2010.11.Award.Amount
## 1             BERING STRAIT SCHOOL DISTRICT            $471014.00
## 2                    YUPIIT SCHOOL DISTRICT            $520579.00
## 3             BERING STRAIT SCHOOL DISTRICT            $449592.00
## 4 MATANUSKA-SUSITNA BOROUGH SCHOOL DISTRICT            $641184.00
## 5                    YUPIIT SCHOOL DISTRICT            $399686.00
## 6 MATANUSKA-SUSITNA BOROUGH SCHOOL DISTRICT            $697703.00
##   Model.Selected
## 1 Transformation
## 2 Transformation
## 3 Transformation
## 4 Transformation
## 5 Transformation
## 6        Restart
##                                                        Location
## 1            200 MAIN ST\nSAVOONGA, AK 99769\n(63.6687, -170.603)
## 2         AKIACHAK 51100\nAKIACHAK, AK 99551\n(60.8911, -161.376)
## 3             169 MAIN ST\nGAMBELL, AK 99742\n(63.7413, -171.689)
## 4     1775 WEST PARKS HWY\nWASILLA, AK 99654\n(61.5794, -149.495)
## 5                   AKIAK 5227\nAKIAK, AK 99552\n(60.8879, -161.2)
## 6 7362 WEST PARKS HWY 725\nWASILLA, AK 99654\n(61.5023, -149.796)
```

```r
#data[sample(nrow(data),5),] 5 random sample of the data
data[which(data$State=='RI'),] #we see that schools in RI have missing models
```

```
##                                          School.Name          City State
## 662                  CHARLOTTE WOODS ELEMENTARY SCHOOL    PROVIDENCE    RI
## 663       LILLIAN FEINSTEIN ELEMENTARY, SACKETT STREET    PROVIDENCE    RI
## 664                       ROGER WILLIAMS MIDDLE SCHOOL    PROVIDENCE    RI
## 665                             FEINSTEIN HIGH SCHOOL    PROVIDENCE    RI
## 666 WILLIAM B. COOLEY/HEALTH AND SCIENCE TECH. ACADEMY    PROVIDENCE    RI
## 667                   CENTRAL FALLS SENIOR HIGH SCHOOL CENTRAL FALLS    RI
##     District.Name X2010.11.Award.Amount Model.Selected
## 662    PROVIDENCE
## 663    PROVIDENCE
## 664    PROVIDENCE
## 665    PROVIDENCE
## 666    PROVIDENCE
## 667 CENTRAL FALLS
##                                                    Location
## 662  674 PRAIRIE AVE\nPROVIDENCE, RI 2905\n(41.7957, -71.4106)
## 663    159 SACKETT ST\nPROVIDENCE, RI 2907\n(41.794, -71.4193)
## 664 278 THURBERS AVE\nPROVIDENCE, RI 2905\n(41.7984, -71.4109)
```

```
## 665  544 ELMWOOD AVE\nPROVIDENCE, RI 2907\n(41.7977, -71.4254)
## 666 182 THURBERS AVE\nPROVIDENCE, RI 2905\n(41.7981, -71.4078)
## 667  24 SUMMER ST\nCENTRAL FALLS, RI 2863\n(41.8872, -71.3916)
```

```
data = data[which(data$State != 'RI'),] #remove all the schools from RI (Rhode Island)

save(data, file="~/Desktop/UCSD MATH/Math.185/school-improvement-2010.rda") #save dataset as an rda file
```

Problem 4

We are interested in the following question: Is there an association between the model that each school selected and the state where the school was located at that time?

Part A

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.

H_0: There is no association between the model that each school selected and the state where the school was located at that time, i.e. independence.

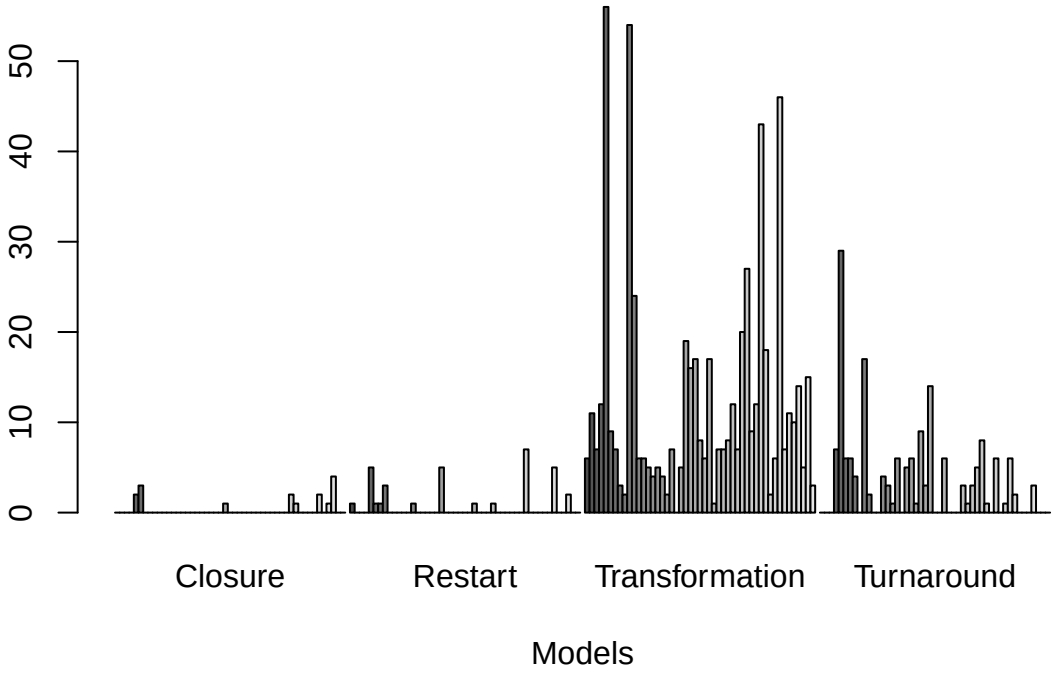
H_1: There is association between the model that each school selected and the state where the school was located at that time, i.e. dependence.

Since significance level was not mentioned, we proceed and set a =0.05.

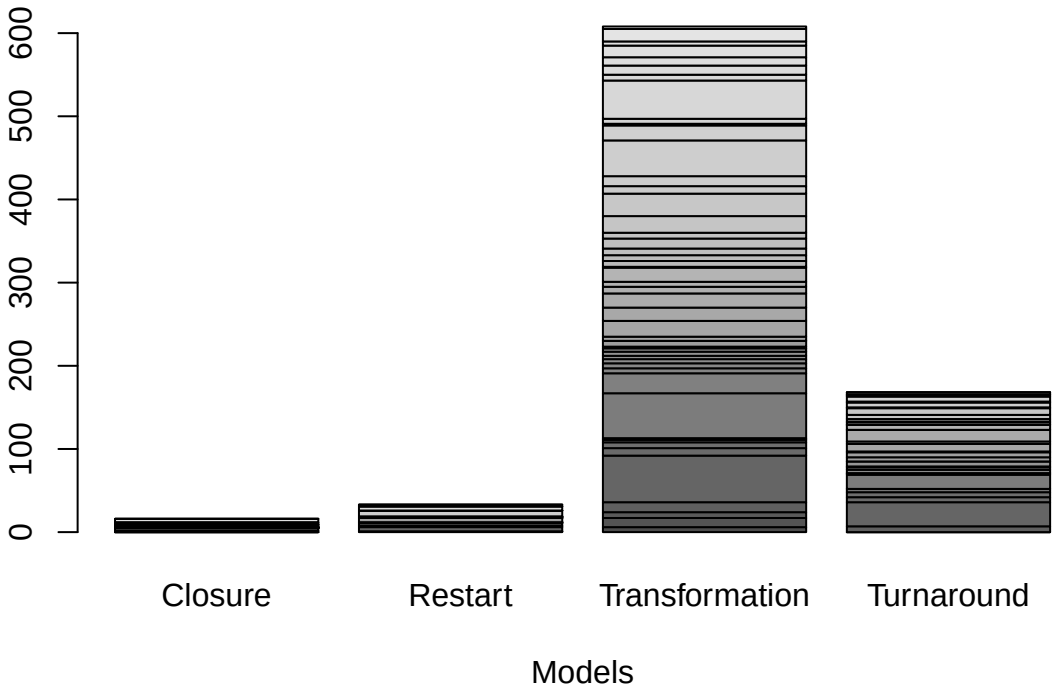
```
load("~/Desktop/UCSD MATH/Math.185/school-improvement-2010.rda")
data = data[,c('State','Model.Selected')] #sort out the state and Model.selected columns
table = table(data$State, data$Model.Selected) #create a table using the two columns
table = table[c(1:38,40:50), c(2:5)]

xlab = ('Models')

barplot(table, beside=T, xlab=xlab)
```




```
barplot(table, xlab=xlab)
```




```
chisq.test(table)
```

```
##
##  Pearson's Chi-squared test
##
## data:  table
## X-squared = 378.37, df = 144, p-value < 2.2e-16
```

```r
cat('p-value is',chisq.test(table)$p.value, '< 0.05=a', 'so we reject the null hypothesis. This means that
```

```
## p-value is 5.885856e-23 < 0.05=a so we reject the null hypothesis. This means that there is association
```

Part B

In this particular case, is the method of Problem 3 applicable? If so, apply it and compare with the previous test that you performed.

It is applicable to use the method from Problem 3 in this case.

```r
chisq.perm.test(table)
```

```
## [1] 0.0004997501
```

```r
cat('Despite the fact that the p-value we obtain is:', chisq.perm.test(table), "which is larger than the p
```

```
## Despite the fact that the p-value we obtain is: 0.0004997501 which is larger than the p-value from the
```

```r
fisher.test(table, simulate.p.value=TRUE)
```

```
##
##  Fisher's Exact Test for Count Data with simulated p-value (based
##  on 2000 replicates)
##
## data:  table
## p-value = 0.0004998
## alternative hypothesis: two.sided
```

```r
cat("Using 'fisher.test' function, we also get a p-value similar to that of the 'chisq.test':", fisher.tes
```

```
## Using 'fisher.test' function, we also get a p-value similar to that of the 'chisq.test': 0.0004997501
```

Despite the fact that the p-value from 'chisq.perm.test' is larger than the p-value from the Chi-Squared test but it is not too large to keep the null hypothesis. Hence we still have the fact that there is no association between the state and the model.