IST772 Problem Set 7

Shivani Sanjay Mahaddalkar

The homework for week 8 is based on exercises 3, 4, 8, 9, 10 on pages 155-156 but with changes as noted in this notebook (i.e., follow the problems as given in this document and not the textbook).

Attribution statement: (choose only one) 1. I did this homework by myself, with help from the book and the professor

Chapter 7, Exercise 3

Run cor.test() on the correlation between "speed" and "dist" in the cars data set (type "? cars" to see the documentation) and interpret the results. (1 pt) Note that you will have to use the "$" accessor to get at each of the two variables (like this: cars$speed, but without the backslash, needed since the dollar sign is a special character in R markdown). Make sure that you interpret both the confidence interval and the p-value that is generated by cor.test(). (1 pt)

```
cor.test(cars$speed, cars$dist)
```

```
##
##  Pearson's product-moment correlation
##
## data:  cars$speed and cars$dist
## t = 9.464, df = 48, p-value = 1.49e-12
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.6816422 0.8862036
## sample estimates:
##       cor
## 0.8068949
```

The null hypothesis is rho=0, and alternative is that the rho is not equal to zero. t-value is 9.464 which is much greater than 2, which means that it is likely to be significant. The p-value is much less than alpha=0.05 therefore we reject the null hypothesis that the correlation coefficient is zero. The confidence interval does not contain 0, which supports the rejection of the null hypothesis.

Chapter 7, Exercise 4

Below is a copy of the bfCorTest() custom function presented in this chapter; you can instead use the correlationBF function from the BayesFactor library. Conduct a Bayesian analysis of the correlation between "speed" and "dist" in the cars data set. (1 pt) Report the results. (1 pt)

```r
library("BayesFactor")
```

```
## Warning: package 'BayesFactor' was built under R version 4.0.4
```

```
## Loading required package: coda
```

```
## Loading required package: Matrix
```

```
## ************
## Welcome to BayesFactor 0.9.12-4.2. If you have questions, please contact
Richard Morey (richarddmorey@gmail.com).
##
## Type BFManual() to open the manual.
## ************
```

```r
bfCorTest <- function (x,y) # Get r from BayesFactor
{
  zx <- scale(x)  # standardize X
  zy <- scale(y)  # standardize Y
  zData <- data.frame(x=zx,rhoNot0=zy) # put in a data frame
  bfOut <- generalTestBF(x ~ rhoNot0, data=zData) # linear coefficient
  mcmcOut <- posterior(bfOut,iterations=10000) # posterior samples
  print(summary(mcmcOut[,"rhoNot0"])) # Show the HDI for r
  return(bfOut) # Return Bayes factor object
}

bfCorTest(cars$speed,cars$dist)
```

```
##
## Iterations = 1:10000
## Thinning interval = 1
## Number of chains = 1
## Sample size per chain = 10000
##
## 1. Empirical mean and standard deviation for each variable,
##    plus standard error of the mean:
##
##           Mean             SD       Naive SE Time-series SE
##      0.7895272      0.0881286      0.0008813      0.0009087
##
## 2. Quantiles for each variable:
##
##   2.5%    25%    50%    75%  97.5%
## 0.6136 0.7318 0.7901 0.8474 0.9616
```

```
## Bayes factor analysis
## --------------
## [1] rhoNot0 : 3486525337 ±0.01%
##
## Against denominator:
##   Intercept only
```

```
## ---
## Bayes factor type: BFlinearModel, JZS
```

The point estimate for rho is 0.789, with a HDI of 0.6105-0.9607 which does not contain zero, which supports the previous test result that the rho is not zero. The rhoNot() signifies that the odds are heavily against the null hypothesis and it means that speed and distance are correlated # Chapter 7, Exercise 8

The data set called UCBAdmissions (see "? UCBAdmissions" for documentation) contains data on applicants to graduate school at Berkeley for the six largest departments in 1973 classified by admission and sex. You can access the data for the first department like this: UCBAdmissions[, , 1]. Make sure you put two commas before the 1: this is a three dimensional contingency table that we are subsetting down to two dimensions. Run chisq.test() on the subset of the data set for department 1 (1 pt) and make sense of the results. (1 pt)

```
contingencytable1 <- UCBAdmissions[ , , 1]
contingencytable1
```

```
##           Gender
## Admit      Male Female
##   Admitted  512     89
##   Rejected  313     19
```

```
chisq.test(contingencytable1)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  contingencytable1
## X-squared = 16.372, df = 1, p-value = 5.205e-05
```

The reported chi squared value is 16.372 with p-value of 5.2 * 10^-5, which is less than 0.05. Therefore, we reject the null hypothesis that the rho value is zero, which means there is correlation between the decision results and the gender in department 1. # Chapter 7, Exercise 9

Use contingencyTableBF() to conduct a Bayes factor analysis on the UCB admissions data for department 1. (1 pt) Report and interpret the Bayes factor. (1 pt)

```
contingencyTableBF(contingencytable1, sample="poisson",posterior=FALSE)
```

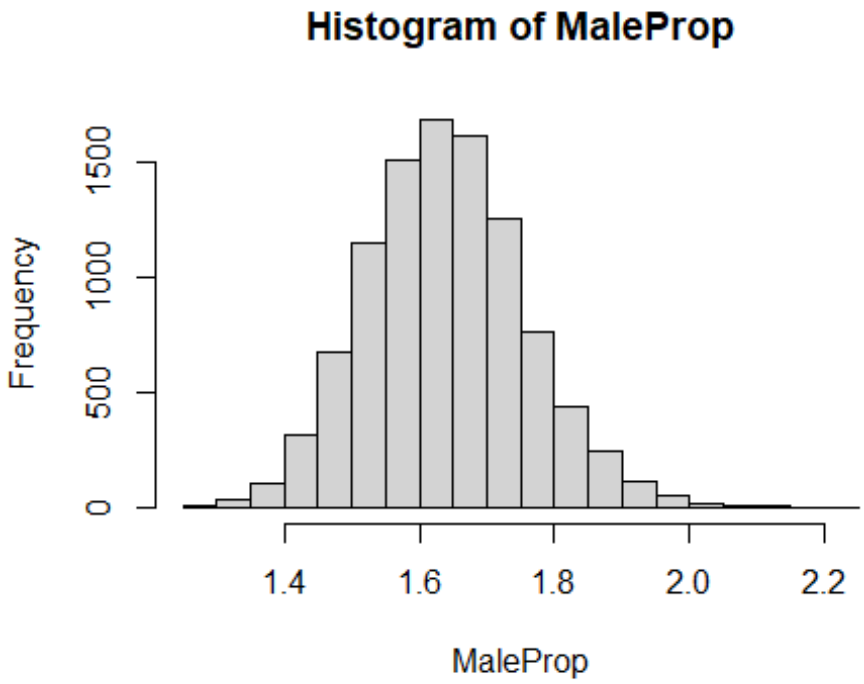
```
## Bayes factor analysis
## --------------
## [1] Non-indep. (a=1) : 1111.64 ±0%
##
## Against denominator:
##   Null, independence, a = 1
## ---
## Bayes factor type: BFcontingencyTable, poisson
```

The Bayes factor of 1111.64:1 is in favor of the alternative hypothesis that the two factors are not independent from one another. There is some correlation between the decision results and the gender in department 1.

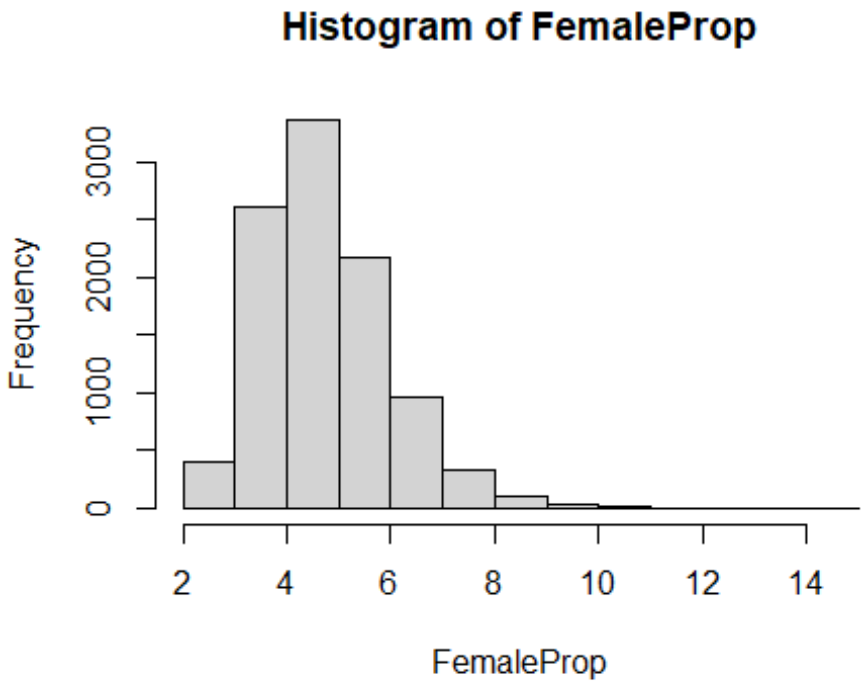
Chapter 7, Exercise 10

Using the UCBAdmissions data for department 1, run contingencyTableBF() with posterior sampling. (1 pt) Use the results to calculate a 95% HDI of the difference in proportions between the columns. (1 pt for extracting proportions, 1 pt for HDI, 1 pt for interpretation)

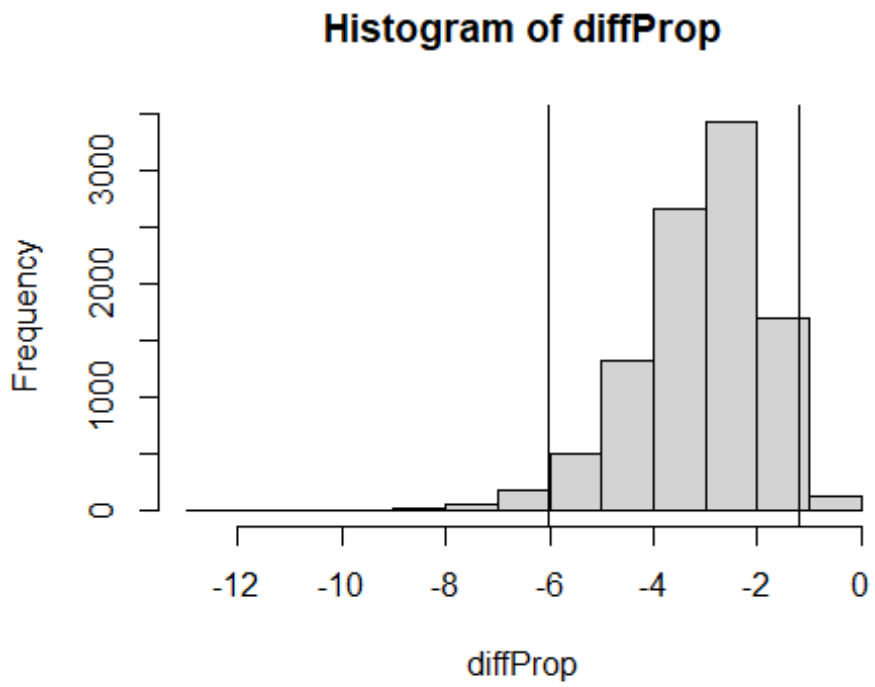
```
ctMCMCout <- contingencyTableBF(contingencytable1,
sample="poisson",posterior=TRUE, iterations=10000)
summary(ctMCMCout)

##
## Iterations = 1:10000
## Thinning interval = 1
## Number of chains = 1
## Sample size per chain = 10000
##
## 1. Empirical mean and standard deviation for each variable,
##    plus standard error of the mean:
##
##               Mean     SD Naive SE Time-series SE
## lambda[1,1] 511.12 22.385  0.22385        0.21808
## lambda[2,1] 312.76 17.691  0.17691        0.17691
## lambda[1,2]  89.65  9.393  0.09393        0.09393
## lambda[2,2]  19.91  4.474  0.04474        0.04474
##
## 2. Quantiles for each variable:
##
##               2.5%    25%    50%    75%  97.5%
## lambda[1,1] 467.90 495.93 511.05 526.05 555.24
## lambda[2,1] 278.63 300.69 312.77 324.63 347.60
## lambda[1,2]  72.15  83.23  89.27  95.84 109.01
## lambda[2,2]  12.12  16.75  19.61  22.79  29.66

MaleProp <- ctMCMCout[,'lambda[1,1]'] / ctMCMCout[,'lambda[2,1]']
hist(MaleProp)
```




```
FemaleProp <- ctMCMCout[,'lambda[1,2]'] / ctMCMCout[,'lambda[2,2]']
hist(FemaleProp)
```



```
diffProp <- MaleProp-FemaleProp
hist(diffProp)
abline(v=quantile(diffProp,c(0.025)), col='black')
abline(v=quantile(diffProp,c(0.975)), col='black')
```



The diffProp histogram of the posterior distribution of differences in proportion between the two columns. It is how much the admitted:rejected ratio decreases as we switch columns. The HDI is -6 as the lower bound and around -1.1 as the upper bound. Since zero is not present in the HDI we have credible evidence to say that gender and the admission decision were not independent of one another.