Homework 7

Maya Mileva

due date: Nov 21, 2019

I did this homework by myself, with help from the book and the professor.

Exercises

3. Run cor.test() on the correlation between "area" and "perm" in the rock data set and interpret the results. Note that you will have to use the "$\jmath accessor to get at each of the two variables (like this: rock$**area). Make sure that you interpret both the confi- dence interval and the p-value that is generated by cor.test().**

```
## Explore the data
str(rock) # 48 obs and 4 variables
```

```
## 'data.frame':    48 obs. of  4 variables:
##  $ area : int  4990 7002 7558 7352 7943 7979 9333 8209 8393 6425 ...
##  $ peri : num  2792 3893 3931 3869 3949 ...
##  $ shape: num  0.0903 0.1486 0.1833 0.1171 0.1224 ...
##  $ perm : num  6.3 6.3 6.3 6.3 17.1 17.1 17.1 17.1 119 119 ...
```

```
head(rock)
```

```
##   area    peri     shape perm
## 1 4990 2791.90 0.0903296  6.3
## 2 7002 3892.60 0.1486220  6.3
## 3 7558 3930.66 0.1833120  6.3
## 4 7352 3869.32 0.1170630  6.3
## 5 7943 3948.54 0.1224170 17.1
## 6 7979 4010.15 0.1670450 17.1
```

```
cor(rock)
```

```
##              area       peri      shape       perm
## area   1.0000000  0.8225064 -0.1821611 -0.3966370
## peri   0.8225064  1.0000000 -0.4331255 -0.7387158
## shape -0.1821611 -0.4331255  1.0000000  0.5567208
## perm  -0.3966370 -0.7387158  0.5567208  1.0000000
```

When we calculated Parson's r correlation from the "rock" data set, we noticed that the output is a correlation matrix. The 1.0 values of the diagonal are the correlations between each variable and itself. There are two triangles of correlation data, one above diagonal and one below. The two triangles are transposed versions of each other: they contain the same information, so we really need to look at the lower triangle.

We can notice .82 correlation between "area" and "peri" - high value , suggesting the possibility that these two variables might in some senses be redundant with each other. There is also stong negative correlation between "perm" and "peri", and moderate correlation between "shape" and "perm".

Null hypothesis testing on the correlation – the procedure for testing the significance of the correlation coefficient assumes a null hypothesis of rho = 0.

```
## Run cor.test() on the correlation between "area" and "perm"
cor.test(rock$area, rock$perm)
```

```
##
##  Pearson's product-moment correlation
##
## data:  rock$area and rock$perm
## t = -2.9305, df = 46, p-value = 0.005254
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.6118206 -0.1267915
## sample estimates:
##       cor
## -0.396637
```

The output above has three sections: the first three lines are the conventional null hypothesis test with an assumption of rho = 0. For our example the null hypothesis will be that rho, the population correlation coefficient between area and perm, is zero. The alternative hypothesis will be simply the logical opposite and incorporates the possibility of nonzero correlation that is either positive or negative. We have detected statistical significance (t-value = -2.9305 with associated p-value of p = 0.0052). The test statistic is a t-test on a transformed version of the correlation coefficient. The test yields a t-value of -2.9305. Df are the degrees of freedom – how many elements are free to vary in a statistical system. In our case we started with 48 observations, and one degree of freedom was lost for the calculation of the mean in each of the two samples. The value of t = -2.9305 is outside the central region of the t-distribution for df=46, with a corresponding probability of 0.0052. One way of thinking about p-value is to say that there is .0052 chance of observing an absolute value of t this high or higher under the assumption that the population value of rho = 0. Using the conventional p < .05 threshold for alpha to evaluate this result, we can reject the null hypothesis of **rho=0**. The cor.test() also provides 95% confidence interval around the point estimate of r = -0.39. We can define CI as follows: if we repeated this sampling process many times and each time constructed a confidence interval around the calculated value of r, about 95% of those constructed confidence intervals would contain the true population value of rho. In conclusion we cam say that 95% CI for rho ranged from -.61 to -.12. This is a very tight range. Importantly CI does not straddle with 0, result that concur with the result from the significance test and we have a sense of certainty that the correlation is negative.

4. Create a copy of the bfCorTest() custom function presented in this chapter. Don't forget to "source" it (meaning that you have to run the code that defines the function one time to make R aware of it). Conduct a Bayesian analysis of the correlation between "area" and "perm" in the rock data set.

```
## Creating a funcion for Bayesian test of cor coef
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
  }
```

```
# Bayesian analysis of the correlation between "area" and "perm"
bfCorTest(rock$area, rock$perm)
```

```
##
## 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
##      -0.342453       0.136227       0.001362       0.001503
##
## 2. Quantiles for each variable:
##
##    2.5%     25%     50%     75%   97.5%
## -0.6122 -0.4343 -0.3411 -0.2500 -0.0732
```

```
## Bayes factor analysis
## --------------
## [1] rhoNot0 : 8.072781 ±0%
##
## Against denominator:
##   Intercept only
## ---
## Bayes factor type: BFlinearModel, JZS
```

The point estimate (mean correlation in the posterior distribution of rho) is -.34, similar to what we observed earlier (-.39). The 95% HDI ranges from -.61 to -.06, wider range that still doesn't straddle with 0, we have a credible notion that rho is negative. The result is confirmed by the Bayes factor output, which shows odds of 8.07:1 in favor of the alternative hypothesis that the population correlation, rho, between area and perm, is not equal to 0.

Taking off this evidence together, we can say with some credibility that the population correlation is e negative value lying somewhere in the range of -.61 up to -.06, and probably close to the central value of -.34.

8. Not unexpectedly, there is a data set in R that contains these data. The data set is called UCBAdmissions and you can access the department mentioned above 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 this subset of the data set and make sense of the results.

UCBAdmissions data set is frequently used for illustrating Simpson's paradox, see Bickel et al (1975). At issue is whether the data show evidence of sex bias in admission practices.

In our subset we have 825 males and 108 females. 62% of the males and 82% of the females were admitted.

```
# UCBAdmissions[ , ,1]  # two dimentional contigency table
usba <- UCBAdmissions[ , ,1]
usba
```

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

```
chiOut <- chisq.test(usba)
chiOut
```

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

We have two categorical factors (gender and admit), and we want to see if they are related to one another. By subletting we have created 2x2 contingency table that was used to call chisq.test() which calculates chi-square from the table and obtains the associated p-value.

The observed chi-square calculated from this table is 16.37,on one degree of freedom, with an associated p-value of 5.205e-05. Because that p-value is really small, smaller than the typical alpha threshold of p <.05, we reject the null hypothesis that gender and admit are not associated, there is no relationship between them, and they are independent. df = 1 because we have 2x2 table((2-1)(2-1)). Looking at the 2x2 contingency table we can see that the proportion of admitted among males is lower than the proportions of addmited among females.

```
chiOut$residuals # how far is the obsereud from expected
```

```
##           Gender
## Admit           Male    Female
##   Admitted -0.842886  2.329614
##   Rejected  1.134063 -3.134383
```

Residuals represent how far an observed value was from the expected value. A large positive residual means that the observation for that cell was too much lower than expected. Large residuals (negative or positive) indicate the cells that made the most powerful contribution to the value of chi-square. In our example that would be Rejected/Female; Admitted/Female. This cells showed where the "action" is with respect to non-independance.

9. Use contingencyTableBF() to conduct a Bayes factor analysis on the UCB admissions data. Report and interpret the Bayes factor.

The BayesFactor packge produce posterior distributions for the fequencies (or proportions)in the cell of the contingency table.

```
ctBFOut <- contingencyTableBF(usba, sampleType="poisson", posterior=FALSE)

ctBFOut
```

```
## 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 in favor of the alternative hypothesis that the two factors are not independent from one another (they are associated). Because the reported Bayes factor excess 150:1, we can treat it as a strong evidence in favor of the alternative hypothesis(nonindependence). Therefore, in the situation the Bayes factor and the null hypothesis concur with each other.

10. Using the UCBA data, run contingencyTableBF() with posterior sampling. Use the results to calculate a 95% HDI of the difference in proportions between the columns.

```
ctMCMCOut <- contingencyTableBF(usba, sampleType="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] 510.85 22.790  0.22790        0.22790
## lambda[2,1] 312.54 17.554  0.17554        0.17554
## lambda[1,2]  89.71  9.594  0.09594        0.09594
## lambda[2,2]  19.89  4.438  0.04438        0.04438
##
## 2. Quantiles for each variable:
##
##               2.5%    25%    50%    75%  97.5%
## lambda[1,1] 467.32 495.25 510.64 525.72 556.35
## lambda[2,1] 279.07 300.56 312.20 324.20 348.32
## lambda[1,2]  72.28  82.90  89.41  95.88 109.42
## lambda[2,2]  12.14  16.73  19.52  22.61  29.52
```

The resulting object, ctMCMCOut, contains the result of the 10000 samples in the form of means and HDIs for each of the cell counts.

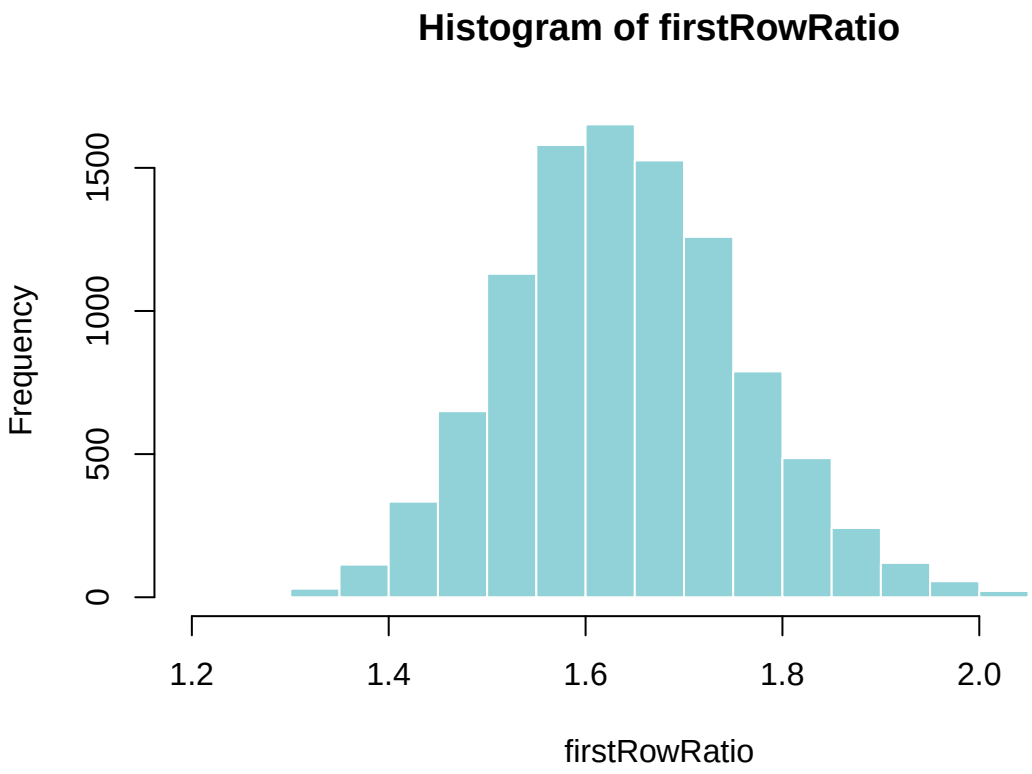
The means at the first section closely match the content of the cells in the original data.

```
usba
```

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

We don‘t learn anything new from those means. The second table shows quantiles for each variable, including the boundaries of the 95% highest density interval (HDI)in the fist and the last column.

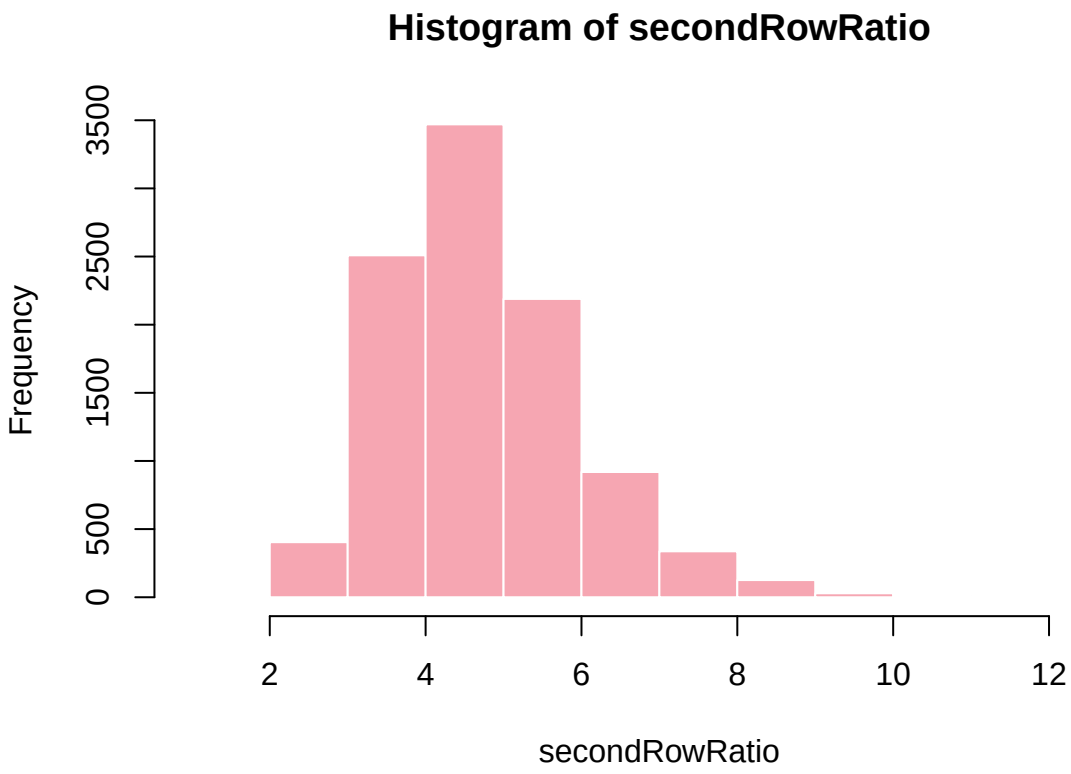
```
firstRowRatio <- ctMCMCOut[,"lambda[1,1]"] / ctMCMCOut[,"lambda[2,1]"]
hist(firstRowRatio
     , border = "white"
     , col = "#90d2d8")
```




```
# [1,1] admitted/male; [1,2] rejected/male - (512/313)
```

Males admitted have higher proportion than males rejected. Most common value is around 1.7.

```
secondRowRatio <- ctMCMCOut[,"lambda[1,2]"] / ctMCMCOut[,"lambda[2,2]"]
hist(secondRowRatio
     , border = "white"
     , col = "#f6a6b2")
```

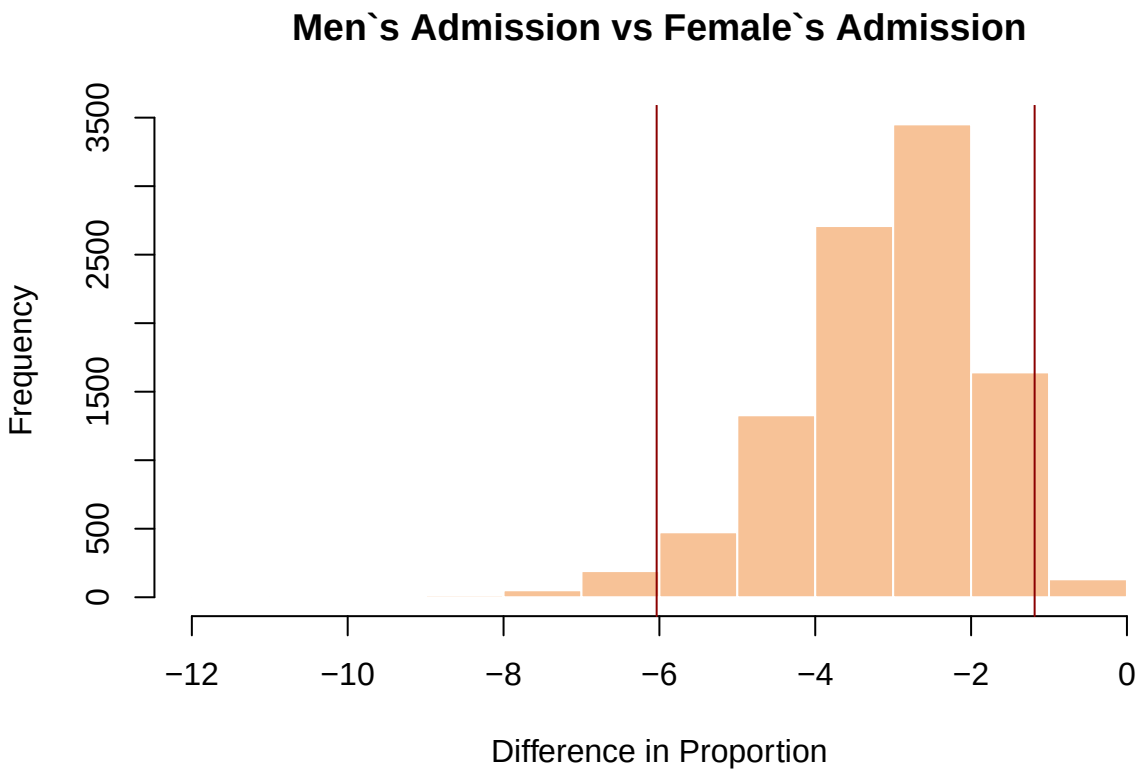



```
# [1,2] admitted/female; [2,2] rejected/female - (89/19)
```

Females admitted have higher proportion than females rejected. The center of the distribution a little bit higher than 4.5.

This two posterior distributions represent the two columns of our original table. We will try find the difference between them by substracting them.

```
## Substract males and females
subs <- firstRowRatio - secondRowRatio
#range(subs)
hist(subs
     , xlab =  c("Difference in Proportion")
     , main = "Men`s Admission vs Female`s Admission"
     , border = "white"
     , col = "#f7c297"
     )
## Show HDI
abline(v=quantile(subs, c(0.025)), col = "darkred") # low end
abline(v=quantile(subs, c(0.975)), col = "darkred") # high end
```




```
mean(subs)
```

```
## [1] -3.107785
```

The center of this distribution is a difference in proportions of -3.1. HDI is marked by ablines. In the population, the proportion shifts about -3.1, although there is small likelihood that the difference in proportions could be as little as about -6.1 or as much as about -1.6.

HDI does not overlap with 0. There is definitely relationship between those two categorical variables and we modeled it with our posterior estimates.

Women admission is way better than men admission in the subset we explored.