Measures of association Homework

Tim Hulak

I produced the material below with no assistance

Exercises 3,4, 8, 9, and 10 on pages 155 and 156 of *Reasoning with Data: An Introduction to Traditional and Bayesian Statistics Using R*

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 "ꓱ*accessor to get at each of the two variables* (*like this* : *rock*area). Make sure that you interpret both the confidence interval and the p-value that is generated by cor.test().

```
options(scipen=999)
data("rock")
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.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
```

Answer: The output for the *Pearson's product-moment correlation* has a *lower-bound confidence interval* of **-0.6118206** and an *upper-bound confidence interval* of **-0.1267915**. Since both of these values are negative, it can be safely assumed that the correlation is skewed in a particular direction. The *p-value* is **0.005254**, which is less than the traditional alpha of **0.05**. This means that we can reject the null hypothesis that the correlation between the rock area and the rock perm is not 0.

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.

```
#Copied from the book:
# Reasoning with Data: An Introduction to Traditional and Bayesian Statistics Using R
# Stanton, Jeffrey M.. Reasoning with Data (p. 136). Guilford Publications. Kindle Edition.
library("BayesFactor")
```

```
## 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
##
## Type BFManual() to open the manual.
## ************
```

```
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(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
##      -0.348678        0.136155        0.001362       0.001498
##
## 2. Quantiles for each variable:
##
##     2.5%       25%       50%       75%     97.5%
## -0.62324 -0.43914 -0.34756 -0.25832 -0.08366
```

```
## Bayes factor analysis
```

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

Answer: The point estimate for rho is **is -0.3477**. The *95% Confidence Interval* has a lower-bound of **-0.61125** and an upper-bound of **-0.08239**. This range does not straddle zero. Finally, the Bayes Factor is approximately **8.07**. This means that there is a ratio of **0.1239157:1** odds in favor of the null hypothesis. In other words, there is strong evidence to reject the null hypothesis in favor if the alternative hypothesis (especially when the a rule-of-thumb odds cutoff value of 3:1 is considered).

---

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.

```
data('UCBAdmissions')
UCBAdmissions[, ,1]
```

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

```
chisq.test(UCBAdmissions[,,1], correct=F)
```

```
##
##  Pearson's Chi-squared test
##
## data:  UCBAdmissions[, , 1]
## X-squared = 17.248, df = 1, p-value = 0.0000328
```

Answer: The output of the *Pearson's Chi-squared test* shows a *chi-squared* value of **17.248** with only **1** *degree of freedom*. The *p-value* is **0.0000328**, which is less than the traditional alpha of **0.05**. This means that we can reject the null hypothesis and conclude that gender and admit are not independent of each other.

---

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

```
admission_table <- contingencyTableBF(UCBAdmissions[,,1],sampleType="poisson", posterior=F)
summary(admission_table)
```

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

Answer: There is a **0.0008996:1** ratio between the two factors. They are independent from one another (the two factors are not associated). Because Bayes factor exceeds the rule-of-thumb ratio of **3:1**, we can treat it as positive evidence in favor of independence.

---

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.

```
admission_table_posterior <- contingencyTableBF(UCBAdmissions[,,1], sampleType="poisson", posterior=T,
summary(admission_table_posterior)
```

```
##
## 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] 511.40 22.743  0.22743        0.22743
## lambda[2,1] 312.79 17.505  0.17505        0.17505
## lambda[1,2]  89.68  9.580  0.09580        0.09445
## lambda[2,2]  19.96  4.486  0.04486        0.04282
##
## 2. Quantiles for each variable:
##
##               2.5%    25%    50%    75%  97.5%
## lambda[1,1] 467.98 495.74 511.26 526.47 557.07
## lambda[2,1] 279.19 300.62 312.44 324.47 347.59
## lambda[1,2]  72.01  83.08  89.38  96.05 109.35
## lambda[2,2]  12.20  16.73  19.60  22.80  29.69
```

```
head(admission_table_posterior)
```

```
## Markov Chain Monte Carlo (MCMC) output:
## Start = 1
## End = 7
## Thinning interval = 1
##      lambda[1,1] lambda[2,1] lambda[1,2] lambda[2,2]
## [1,]    525.0880    330.2580    83.31960    28.89413
```

```
## [2,]      498.9203      326.0525       81.86942      21.71752
## [3,]      516.9967      318.7159       90.17500      17.14817
## [4,]      529.0354      337.2086       88.04313      16.56355
## [5,]      471.2954      314.7650      105.86022      18.53270
## [6,]      510.0820      311.8600       81.69840      22.10735
## [7,]      499.2450      342.2604       88.19060      17.82879
```

```
male <- admission_table_posterior[,"lambda[1,1]"]/admission_table_posterior[,"lambda[2,1]"]
hist(male)
```
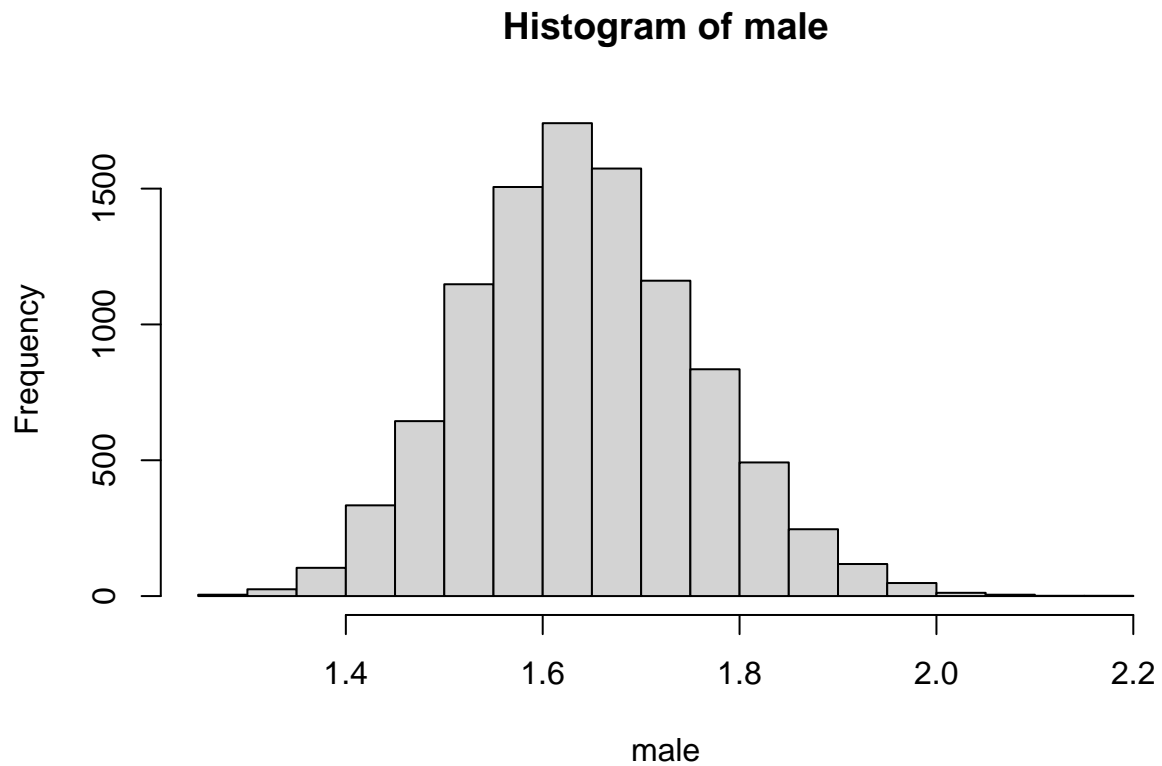
Histogram of male


```
female <- admission_table_posterior[,"lambda[1,2]"]/admission_table_posterior[,"lambda[2,2]"]
hist(female)
```

Histogram of female
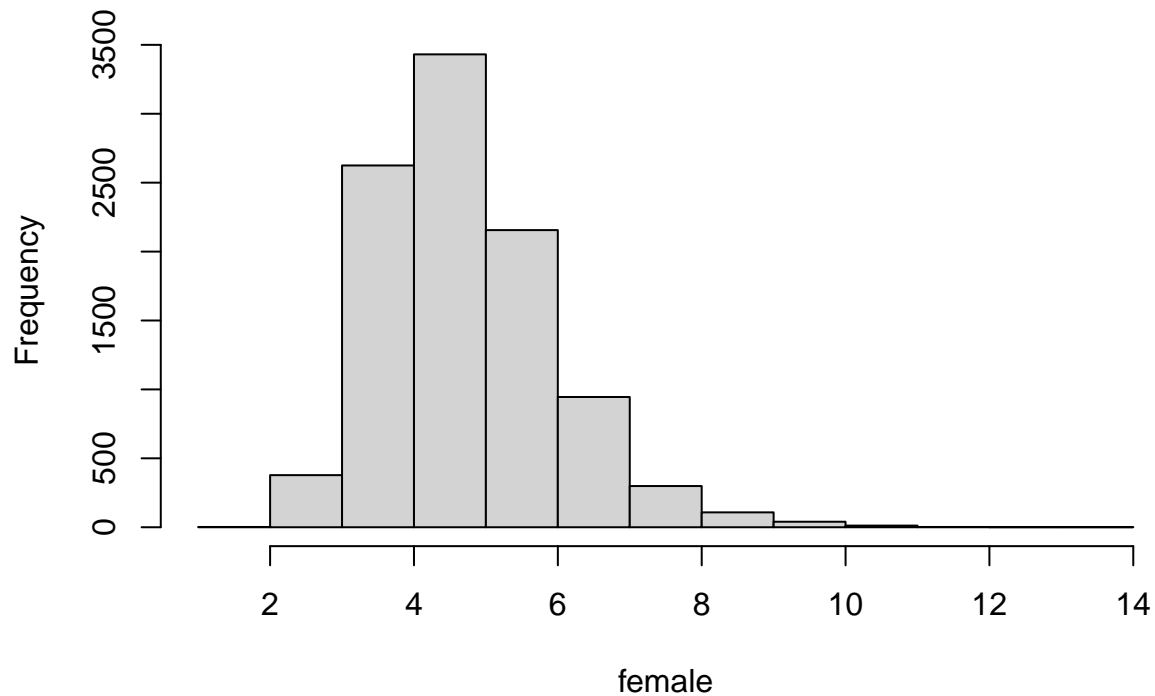

```
difference <- male - female
head(difference)
```

```
## Markov Chain Monte Carlo (MCMC) output:
## Start = 1
## End = 7
## Thinning interval = 1
## [1] -1.293684 -2.239557 -3.636454 -3.746608 -4.214784 -2.059920 -3.487858
```

```
c(quantile(difference,c(0.025)),quantile(difference,c(0.975)))
```

```
##      2.5%     97.5%
## -5.986787 -1.213336
```

Answer: The 95% HDI of the difference in proportions between the columns appears to between **-6.056048** and **-1.217753** .