Homework 7

Kwasi Mensah

2022-11-25

Author: Kwasi Mensah Homwork\_Number: 7 Output: pdf\_document Attribution statement:

I did the homework by myself, with help from the book and the professor #R Markdown #Run these three functions to get a clean test of homework code

dev.off() #Clear the graph window

## null device   
## 1

cat('\014') #Clear the console

rm(list = ls()) #Clear user objects from the environment

#R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see <http://rmarkdown.rstudio.com>.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this: ############################################################################

1. 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 “$” accessor to get at each of the two variables. Make sure you interpret both the confidence interval and the p-value that is generated by cor.test().

library(tidyverse)

## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.2 ──  
## ✔ ggplot2 3.3.6 ✔ purrr 0.3.5   
## ✔ tibble 3.1.8 ✔ dplyr 1.0.10  
## ✔ tidyr 1.2.1 ✔ stringr 1.4.1   
## ✔ readr 2.1.3 ✔ forcats 0.5.2   
## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()

data(rock)  
glimpse(rock)

## Rows: 48  
## Columns: 4  
## $ area <int> 4990, 7002, 7558, 7352, 7943, 7979, 9333, 8209, 8393, 6425, 9364…  
## $ peri <dbl> 2791.90, 3892.60, 3930.66, 3869.32, 3948.54, 4010.15, 4345.75, 4…  
## $ shape <dbl> 0.0903296, 0.1486220, 0.1833120, 0.1170630, 0.1224170, 0.1670450…  
## $ perm <dbl> 6.3, 6.3, 6.3, 6.3, 17.1, 17.1, 17.1, 17.1, 119.0, 119.0, 119.0,…

rock\_perms <- rock$perm  
rock\_areas <- rock$area  
  
# Correlation test between "area" and "perm"  
cor.test(rock\_areas, rock\_perms)

##   
## Pearson's product-moment correlation  
##   
## data: rock\_areas and rock\_perms  
## 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 t value is equal to -2.9305 and our degree of freedom is equal to 46. The p-value is 0.005254 which is below the threshold of 0.05, meaning that it is statically significant. As a result we can reject the null hypothesis and that rho or the correlation between the perms and areas variable is not equal to 0. The 95% Confidence Interval does not pass 0 which also supports our alternative hypothesis. The 95% Confidence Interval for rho ranges from -0.6118206 to -0.1267915, which may possibly indicate the correlation's direction being negative. We have a possible point estimate for the correlation of -0.396637.

1. Create a copy of the bfCorTest() custom function presented in this chapter. Don’t forget to “source” it. Conduct a Bayesian analysis of the correlation between “area” and “perm” in the rock data set.

# Below is the bfCorTest function from the textbook, "Reasoning with Data"  
library(BayesFactor)

## Warning: package 'BayesFactor' was built under R version 4.2.2

## Loading required package: coda

## Loading required package: Matrix

##   
## Attaching package: 'Matrix'

## The following objects are masked from 'package:tidyr':  
##   
## expand, pack, unpack

## \*\*\*\*\*\*\*\*\*\*\*\*  
## Welcome to BayesFactor 0.9.12-4.4. If you have questions, please contact Richard Morey (richarddmorey@gmail.com).  
##   
## 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"])) # Get the HDI for rho  
 return(bfOut) # Return Bayes factor object  
}  
  
# Bayesian analysis of the correlation between "area" and "perm"  
bfCorTest(rock\_areas, rock\_perms)

##   
## 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   
## -0.343437 0.136009 0.001360 0.001515   
##   
## 2. Quantiles for each variable:  
##   
## 2.5% 25% 50% 75% 97.5%   
## -0.61405 -0.43344 -0.34137 -0.25222 -0.07724

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

# The 95% HDI ranges -0.61216 to -0.08402, mean of Bayesian correlation or the point estimate of rho is -0.345621. The rhoNot0 is equal to 8.072781, which is higher than the odds cutoff value which is 3:1. This analysis shows that we will reject the null hypothesis and that it favors the alternative null hypothesis. As a result we can say with confidence that our two variables are dependent on each other, not independent from each other.

1. 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 chisqp.test() on this subset of the data set and make sense of the results.

?UCBAdmissions

## starting httpd help server ... done

# Below is the 2 x 2 contingency table in R function from the textbook, "Reasoning with Data"  
make2x2table <- function(ul)  
{  
 ll <- 50 - ul  
 ur <- 30 - ul  
 lr <- 50 - ur  
 matrix(c(ul,ur,ll,lr),nrow=2, ncol=2, byrow = TRUE)  
}  
  
# Below is the function that computes the chi-square value from the textbook, "Reasoning with Data"  
calcChiSquared <- function(actual,expected)  
{  
 diffs <- actual - expected  
 diffsSq <- diffs^2  
 diffsSqNorm <- diffsSq/expected  
 sum(diffsSqNorm)  
}  
# TEST THE TWO FUNCTIONS CREATED  
make2x2table(43)

## [,1] [,2]  
## [1,] 43 -13  
## [2,] 7 63

calcChiSquared(make2x2table(43), make2x2table(43))

## [1] 0

# Run a chi-squared test on the subset of the data.   
chisq.test(UCBAdmissions[,,1], correct=FALSE)

##   
## Pearson's Chi-squared test  
##   
## data: UCBAdmissions[, , 1]  
## X-squared = 17.248, df = 1, p-value = 3.28e-05

# We can assume that the null hypothesis is that gender and admit are independent variables. The chi-squared is 17.248 and our degree of freedom is 1. The p-value is equal to 3.28e-05, which is extremely lower than .05, so we will decide to reject the null hypothesis. Thus the results of the analysis would support an alternative hypothesis of gender and admit not being independent variables.

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

bayes\_chisqt <- contingencyTableBF(UCBAdmissions[,,1],sampleType="poisson", posterior=FALSE)  
summary(bayes\_chisqt)

## 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 which is well over the odds cut off of 3:1. This analysis results support our alternative hypothesis that the two variables are not independent from each other. Because the reported Bays factor is in excess of 3:1 we can treat it as positive evidence in favor of nonindependce.

1. 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.

# contingencyTableBF() with posterior sampling  
bayes\_chisqt2 <- contingencyTableBF(UCBAdmissions[,,1],sampleType="poisson", posterior=TRUE, iterations=10000)  
  
summary(bayes\_chisqt2) #Review the posterior distributions

##   
## 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] 510.89 22.717 0.22717 0.23154  
## lambda[2,1] 312.31 17.416 0.17416 0.17416  
## lambda[1,2] 89.54 9.455 0.09455 0.09455  
## lambda[2,2] 20.02 4.499 0.04499 0.04499  
##   
## 2. Quantiles for each variable:  
##   
## 2.5% 25% 50% 75% 97.5%  
## lambda[1,1] 467.23 495.40 510.56 525.42 557.18  
## lambda[2,1] 278.52 300.40 312.01 324.07 347.25  
## lambda[1,2] 71.72 83.01 89.29 95.90 108.80  
## lambda[2,2] 12.28 16.80 19.73 22.84 29.69

# Calculate admission status proportions for men  
maleProp <- bayes\_chisqt2[,"lambda[1,1]"]/bayes\_chisqt2[,"lambda[2,1]"]  
summary(maleProp)

##   
## 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   
## 1.64085 0.11601 0.00116 0.00116   
##   
## 2. Quantiles for each variable:  
##   
## 2.5% 25% 50% 75% 97.5%   
## 1.424 1.562 1.636 1.717 1.880

# Calculate admission status proportions for women  
femaleProp <- bayes\_chisqt2[,"lambda[1,2]"]/bayes\_chisqt2[,"lambda[2,2]"]  
summary(femaleProp)

##   
## 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   
## 4.71342 1.23615 0.01236 0.01236   
##   
## 2. Quantiles for each variable:  
##   
## 2.5% 25% 50% 75% 97.5%   
## 2.848 3.845 4.511 5.388 7.575

# Calculate the difference between males and females  
diffProp <- maleProp - femaleProp  
summary(diffProp)

##   
## 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   
## -3.07257 1.24181 0.01242 0.01242   
##   
## 2. Quantiles for each variable:  
##   
## 2.5% 25% 50% 75% 97.5%   
## -5.951 -3.752 -2.870 -2.200 -1.184

# Lower Bound of 95% HDI   
quantile(diffProp,c(0.025))

## 2.5%   
## -5.951127

# Upper Bound of 95% HDI  
quantile(diffProp,c(0.975))

## 97.5%   
## -1.183621

# Center of the distribution   
mean(diffProp)

## [1] -3.072572

# Histogram of the posterior distribution of differences in proportions between the two columns.  
hist(diffProp,   
 main="Histogram of Posterior Distribution of Proportion Differences - Male vs. Female",   
 col="red", xlab="DIfference")  
abline(v=quantile(diffProp,c(0.025)), col="black")  
abline(v=quantile(diffProp,c(0.975)), col="black")

