**IST-772 Quantitative Reasoning in Data science**

Week7/HW-7: Measures of association

**Associations between variables (Page 155-156: Problems 3-4, 8-10)**

Sharat Sripada ([vssripad@syr.edu](mailto:vssripad@syr.edu))

1. **Run a *cor.test()* on variables area and perm in the rock data set. Interpret the probability and the confidence interval**

Output of the cor.test() for area and perm from the rock data set:

> 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

First, stating NULL Hypothesis for the correlation test:

***The correlation between variables area and perm is zero***

Note here that the p-value of 0.005254 is < alpha (0.05) and we therefore reject the NULL Hypothesis.

The confidence interval (CI) is in the range (-0.61, -0.13) which provides a 95% chance of where the point estimate for correlation can exist. Importantly also, the CI does not straddle 0 and we therefore have a sense of certainty that the correlation is negative.

Finally, R provides the correlation co-efficient, r = -0.39

1. **Create a copy of the *bfCorTest()* custom function. Conduct a Bayesian analysis of the correlation between ‘area’ and ‘perm’ in the rock data set.**

R by default does not offer a function or package to do the Bayesian test to determine correlation coefficient between variables. And so, we use the following custom function to achieve this:

bfCorTest <- function(x, y) {

zx <- scale(x) # scale() normalizes or standardizes the data - Essentially: (xi - x-bar) / SD

zy <- scale(y)

zData <- data.frame(x=zx, rhoNot0=zy)

bfOut <- generalTestBF(x ~ rhoNot0, data=zData)

# Run the MCMC simulation

mcmcOut <- posterior(bfOut, iterations=10000)

print(summary(mcmcOut[,'rhoNot0']))

return(bfOut)

}

Using function bfCorTest we run a correlation test:

> 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

-0.343681 0.136483 0.001365 0.001510

2. Quantiles for each variable:

2.5% 25% 50% 75% 97.5%

-0.61467 -0.43431 -0.34266 -0.25410 -0.07725

Bayes factor analysis

--------------

[1] rhoNot0: 8.072781 ±0%

Against denominator:

Intercept only

---

Bayes factor type: BFlinearModel, JZS

From the output, see under Bayes Factor analysis, rhoNot0 with odds 8.072:1 in favor of the alternate hypothesis.

1. **Conduct a Chi-sq test on the UCBAdmissions data set. Interpret the results**

The 3-dim contingency table UCBAdmissions is subsetted to 2-dim using the [,,1] as below:

> UCBAdmissions[,,1]

Gender

Admit Male Female

Admitted 512 89

Rejected 313 19

Run the Chi-sq test on the corresponding 2-dim contingency table:

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

Pearson's Chi-squared test

data: UCBAdmissions[, , 1]

X-squared = 17.248, df = 1, p-value = 3.28e-05

As always, it is important to first state the NULL Hypothesis (note here the variables are Categorical in nature):

***Admission outcomes and Gender are independent***

From the Chi-sq test we see that the p-value is 3.28e-05 which is < alpha (0.05). And therefore, we reject the null hypothesis.

NOTE: Also, since this is 2-dim (2x2) contingency table the DF = (row-1) \* (col-1) = 1

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

Use the following R code to do this:

ctBFout <- contingencyTableBF(UCBAdmissions[,,1], sampleType='poisson', posterior = F)

Note the nuance of having to specify a sampleType when using the *contingencyTableBF()*. There are namely 4-types: poisson, jointMulti, indepMulti, hypergeom

In the simplest case when we collect data with no notion of how many cases we will end up with and no special plans for the marginal proportions we use the poisson type.

The result of the Bayesian Chi-sq test:

> ctBFout

Bayes factor analysis

--------------

[1] Non-indep. (a=1) : 1111.64 ±0%

Against denominator:

Null, independence, a = 1

---

Bayes factor type: BFcontingencyTable, poisson

Null Hypothesis:

***Admission outcomes and Gender are independent***

Based on the Bayes Factor analysis, we see a high 11111:1 odds in favor of the alternate hypothesis and so, we reject the NULL (hypothesis).

1. **Run the contingencyTableBF() on the UCBA data set with posterior sampling. Calculate the 95% HDI of the difference in proportions between the columns.**

The following R code serves with running contingencyTableBF() with posterior sampling and subsequently obtain the difference in proportions:

ctMCMCout <- contingencyTableBF(UCBAdmissions[,,1], sampleType='poisson', posterior = T,

iterations=10000)

summary(ctMCMCout)

# Male proportions

maleProp <- ctMCMCout[,"lambda[1,1]"]/ctMCMCout[,"lambda[2,1]"]

# Female proportions

femaleProp <- ctMCMCout[,"lambda[1,2]"]/ctMCMCout[,"lambda[2,2]"]

# Diff proportions

diffProp <- maleProp - femaleProp

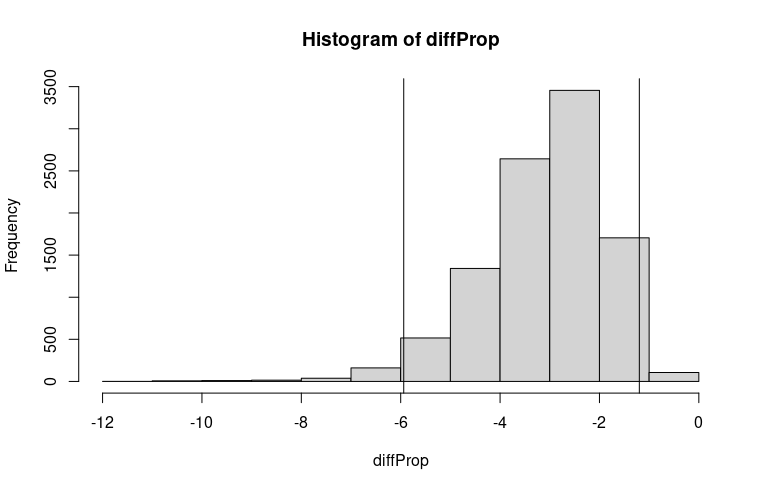
hist(diffProp)

abline(v=quantile(diffProp, c(0.025), color='black'))

abline(v=quantile(diffProp, c(0.975), color='black'))

mean(diffProp)

The corresponding histogram with 95% HDI is as below:



This histogram shows the difference in proportions for male versus female as we navigate between admissions and rejections across 10000 iterations.