Stats. Inference Assignment 4

Naeem Chowdhury

5/25/2020

##1. Setup ### options Set up global options

libraries

Load in needed libraries

2. File management

Create variables for directories

3. Importing Data

The goal of this homework is to have you 1) Familiarized with the χ^2 test of independence for contingency tables; 2) measuring practical strength of dependence between categorical variables; 3) familiarized with permutation test for contingency tables; 4) practice your R coding.

Problem #1

1. For all three examples on slide #24, proceed to

```
case.A <- matrix(c(51, 49, 49, 51), ncol = 2, byrow = TRUE)
colnames(case.A) <- c("Yes", "No")
rownames(case.A) <- c("Female", "Male")
case.A <- as.table(case.A)

case.B <- matrix(c(102, 98, 98, 102), ncol = 2, byrow = TRUE)
colnames(case.B) <- c("Yes", "No")
rownames(case.B) <- c("Female", "Male")
case.B <- as.table(case.B)

case.C <- matrix(c(5100, 4900, 4900, 5100), ncol = 2, byrow = TRUE)
colnames(case.C) <- c("Yes", "No")
rownames(case.C) <- c("Female", "Male")
case.C <- as.table(case.C)
```

a. Use the *my.chisq.test()* function you've defined in the previous HW in order to confirm the X^2 and p-values. Hint: Make sure to convert the %'s into counts first.

Establishing *my.chisq.test()*:

```
my.chisq.test <- function(ctable) {
  rows <- nrow(ctable)
  cols <- ncol(ctable)

  rowsums <- rowSums(ctable)
  colsums <-colSums(ctable)

  total <- sum(ctable)
  df <- (rows-1)*(cols-1)

  expected <- matrix(0, nrow = rows, ncol = cols)

  for(i in 1:rows){

    for(j in 1:cols){

      expected[i,j] <- (rowsums[i]*colsums[j])/ total

    }
  }
  chi_sq <- ((ctable - expected)^2)/expected
  chi_sq <- sum(chi_sq)

  p <- pchisq(chi_sq, df, lower.tail = FALSE)


  print(c("X-Squared: ",  chi_sq))
  print(c("Degrees of Freedom: ", df))
  print(c("p-value: ", p))
  # print(c("Matrix of expected counts: ", expected))
}
```

```
#Case A
my.chisq.test(case.A)
```

```
## [1] "X-Squared: " "0.08"
## [1] "Degrees of Freedom: " "1"
## [1] "p-value: "         "0.777297410789522"
```

```
#Case B
my.chisq.test(case.B)
```

```
## [1] "X-Squared: " "0.16"
## [1] "Degrees of Freedom: " "1"
## [1] "p-value: "         "0.689156516779352"
```

```
#Case C
my.chisq.test(case.C)
```

```
## [1] "X-Squared: " "8"
## [1] "Degrees of Freedom: " "1"
## [1] "p-value: "           "0.00467773498104727"
```

The results of all 3 tests line-up with the results of the tests presented in the slides!

b. Calculate the difference in proportion between males and females that attend religious services weekly. Calculate the risk ratio between males and females that attend religious services weekly.

The data in the original tables is already set up as a proportion.

```
#Convert case A table into a table of conditional proportions
props <- case.A*0.01
#Compute the difference of proportions
diffprop <- props[1]-props[3]
print(c("The difference of proportions is ", diffprop))
```

```
## [1] "The difference of proportions is " "0.02"
```

```
riskratio <- props[1]/props[3]
riskratio
```

```
## [1] 1.040816
```

c. Based on your answers to parts (a) - (b), as n increases, what do you notice with respect to statistical significance? Practical significance?

As n increases, the statistical signifance of the χ^2-test increases. However, the practical significance remains precisely the same for each experiment, showing very little practical significance.

2. Do exercises 11.32 and 11.33 from the Agresti book.

11.32 Marital Happiness

The table shows 2012 GSS data on marital and general happiness for married respondents.

```
marital.happy <- matrix(c(11, 11, 4, 31, 215, 34, 20, 231, 337), ncol = 3, byrow = TRUE)
colnames(marital.happy) <- c("Not Too Happy", "Happy", "Very Happy")
rownames(marital.happy) <- c("Not Too Happy", "Happy", "Very Happy")
marital.happy <- as.table(marital.happy)

marital.happy
```

```
##               Not Too Happy Happy Very Happy
## Not Too Happy            11    11          4
## Happy                    31   215         34
## Very Happy               20   231        337
```

a. The chi-squared test of independence has $X^2 = 214$. What conclusion would you make using a significance level of 0.05? Interpret.

```
df <- (3-1)*(3-1)
pchisq(214, df, lower.tail = FALSE)
```

```
## [1] 3.663652e-45
```

The value $X^2 = 214$ is very deep into the tail of the χ^2 distribution under the null hypothesis of independence with $df = 4$. In fact, this results in an extremely low p-value, and we can conclude by rejecting the null hypothesis. In context, it means that either Marital Happiness is dependent on General Happiness, or General Happiness is dependent on Marital Happiness.

b. Does this large chi-squared value imply there is a strong association between marital and general happiness? Explain.

The large chi-squared value alone does not imply any strength of association, only a dependence relation between the variables. Instead we look to strength of relationship metrics.

c. Find the difference in the proportion of being not too happy between those that are not too happy in their marriage and those that are very happy in their marriage. Interpret that difference.

```
marginal1 <- sum(marital.happy[1,])
marginal3 <- sum(marital.happy[3,])

prop1 <- marital.happy[1,1]/marginal1
prop2 <- marital.happy[3,1]/marginal3

diffprop <- prop1 - prop2
diffpoints <- diffprop*100

print(c("The proportion of people who are generally 'Not Too Happy' is ", diffpoints, "percentage points
```

```
## [1] "The proportion of people who are generally 'Not Too Happy' is "
## [2] "38.9063317634746"
## [3] "percentage points higher for those who are 'Not Too Happy' in their marriage than 'Very Happy'
```

d. Find and interpret the relative risk of being not too happy, comparing the lowest and highest marital happiness group. Interpet.

```
relativerisk <- prop1/prop2

relativerisk
```

```
## [1] 12.43846
```

```
print(c("People who are 'Not Too Happy' in their marriage are ", relativerisk, "times more likely to be
```

```
## [1] "People who are 'Not Too Happy' in their marriage are "
## [2] "12.4384615384615"
## [3] "times more likely to be 'Not Too Happy' in general than those who are 'Very Happy' with their ma
```

11.33 Party ID and gender

The table shows the 2012 GSS data on gender and political party identification from Exercise 11.1 (The row totals are slightly different from the second table in Exercise 11.24 because selecting Independent is ignored.) The chi-squared test of independence has $X^2 = 10.04$ with a p-value of 0.0066, indicating a significant association. Let's describe this association.

```
gender.party <- matrix(c(421,398,244,278,367,198), ncol = 3, byrow = TRUE)
colnames(gender.party) <- c("Democrat", "Independent", "Republican")
rownames(gender.party) <- c("Female", "Male")
gender.party <- as.table(gender.party)

gender.party
```

```
##        Democrat Independent Republican
## Female      421         398        244
## Male        278         367        198
```

a. Estimate the difference between females and males in the proportion who identify themselves as Republicans. Interpret.

```
female.total <- sum(gender.party[1,])
male.total <- sum(gender.party[2,])

gender.props <- gender.party
gender.props[1,] <- gender.party[1,]/female.total
gender.props[2,] <- gender.party[2,]/male.total



gender.props
```

```
##         Democrat Independent Republican
## Female 0.3960489   0.3744120  0.2295390
## Male   0.3297746   0.4353499  0.2348754
```

```
prop.diff <- gender.props[2,3] - gender.props[1,3]
prop.diff <- prop.diff*100

print(c("The proportion of male respondents who are Republican is", prop.diff, "percentage points highe
```

```
## [1] "The proportion of male respondents who are Republican is"
## [2] "0.533640438830543"
## [3] "percentage points higher than the proportion of female respondents who are Republican. We conclu
```

b. Estimate the difference between females and males in the proportion who identify themselves as Democrat. Interpret.

```
prop.diff <- gender.props[1,1] - gender.props[2,1]
prop.diff <- prop.diff*100

print(c("The proportion of female respondents who are Democrat is", prop.diff, "percentage points highe
```

```
## [1] "The proportion of female respondents who are Democrat is"
## [2] "6.62743036840385"
## [3] "percentage points higher than the proportion of male respondents who are Democrat. We conclude
```

Problem #2

1. Code up your own *my.permutation.test()* function to conduct permutation tests on contingency tables. As inputs, it should take a dataframe with two categorical variables as columns (first one - explanatory, second one - response), and the number of randomly generated permutations to be executed. As output, it should provide:

```
- contingency table for the data frame,

- permutation $p$-value,

- plot the histogram of permutation distribution for $X^2$ statistic.
```

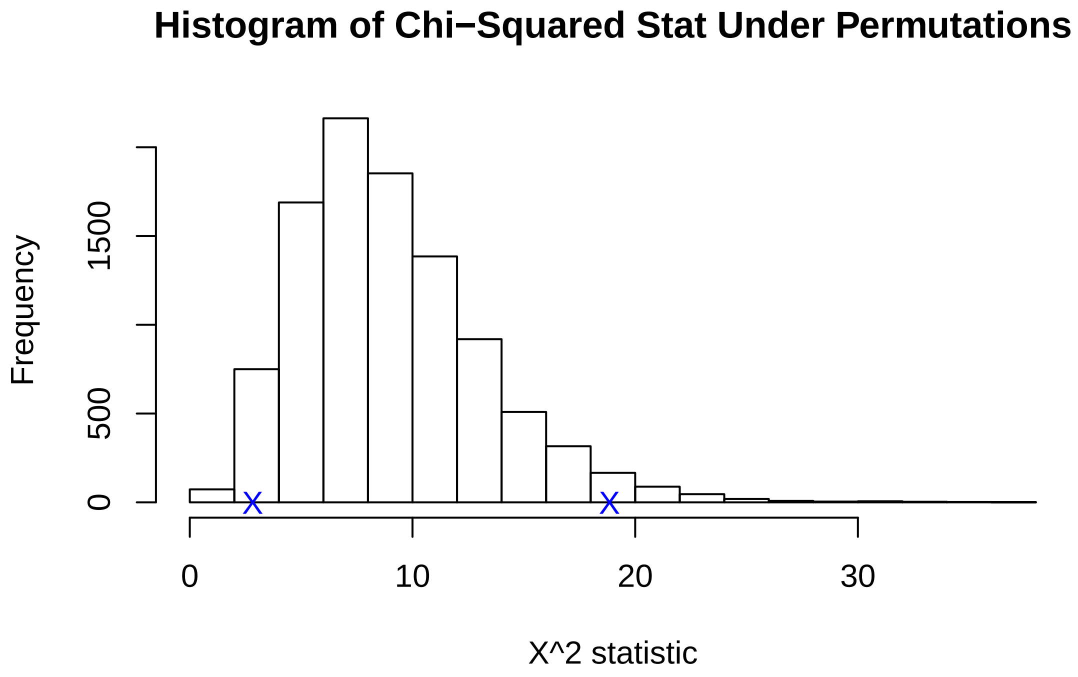
To obtain X^2 values for each permutation, you are allowed to use R's *chisq.test()* function.

2. Proceed to apply the *my.permutation.test()* function (and subsequently interpret the results) to:

```r
 # Arguments: Dataset, number of times to permute, column of desired variable
my.permutation.test <- function(data, permutations, colnum){
  # Get the length of the data
  nrows <- nrow(data)
  #indices to be permuted by sample
  index <- 1:nrows
  #empty matrix to be filled by statistic values
  chis <- matrix(NA, nrow = permutations)
  for(i in 1:permutations){
    index.perm <- sample(index, replace = FALSE)
    perm.data <- data
    perm.data[colnum] <- data[colnum][index.perm, ]
    datatable.perm <- table(perm.data)
    chis[i] <- chisq.test(datatable.perm)$statistic
  }
  return(chis)
}
```

a. CPU_data, with 10,000 permutations. What's the conclusion?

```r
# Plotting the distribution of the values
chiquant <- quantile(vert_stat.chi, c(0.025, 0.975))
hist(vert_stat.chi, main='Histogram of Chi-Squared Stat Under Permutations', xlab='X^2 statistic')
points(x = Xsq, y = 0, pch = 19, cex = 2, col = "red")
points(x = chiquant, y = c(0,0), pch = 'X', col = 'blue')
```

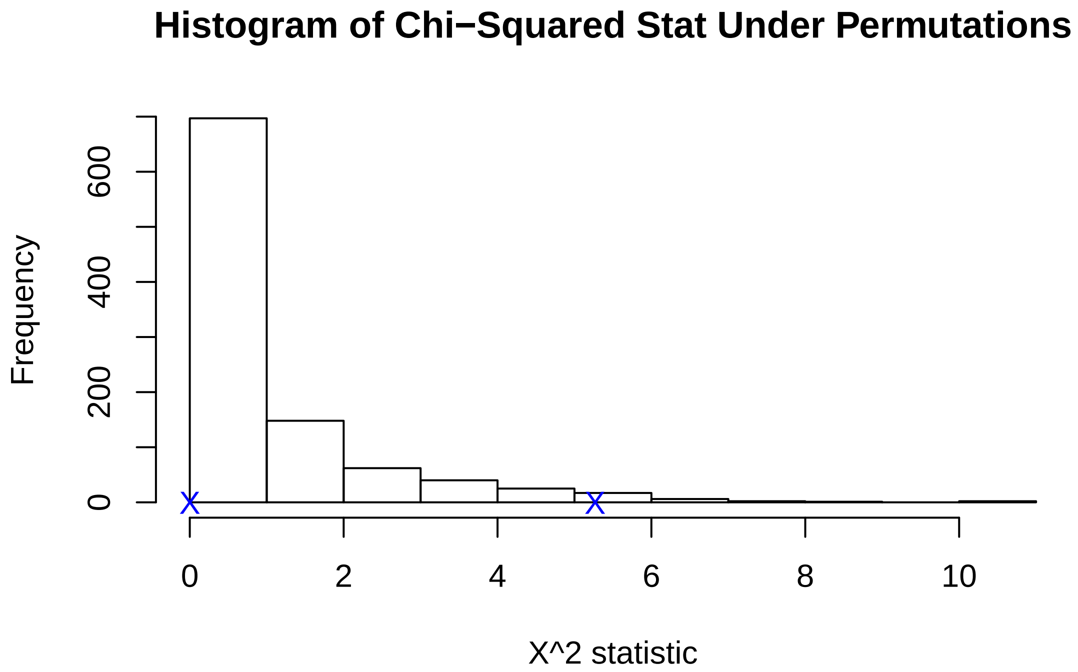


The X^2 value for the table is so high that it does not appear on the graph.

b. GPU data, with just 1,000 permutations. What's the conclusion? Compare the shape of the resulting histogram with the density of χ^2 distribution with appropriate degrees of freedom. What does it tell us about whether χ^2-test results from HW3 were appropriate for GPU data?

The dataset below is a from the GPU dataset, taking the two binary response variables Dedicated and SLI_Crossfire.

```r
# Plotting the distribution of the values
chiquant <- quantile(dedi.chis, c(0.025, 0.975))
hist(dedi.chis, main='Histogram of Chi-Squared Stat Under Permutations', xlab='X^2 statistic')
points(x = Xsq, y = 0, pch = 19, cex = 2, col = "red")
points(x = chiquant, y = c(0,0), pch = 'X', col = 'blue')
```



The observed X^2 for this particular experiment is particularly high, at $X^2 = 488$. This is well outside of the 95% confidence interval. We can conclude by rejecting the null hypothesis, that is, we reject the hypothesis that the variables Dedicated and SLI_Crossfire are independent.