Homework 2

Miles Tweed

2/13/2021

Problem #1

1

```
my.chisq.test <-function(ct=NULL) {
  d <- dim(ct)
  total <- sum(ct)
  ect <- ct
  df <- (d[1]-1) * (d[2]-1)
  for(i in 1:d[1]){
    r.tot <- sum(ct[i,])
    for(j in 1:d[2]) {
      c.tot <- sum(ct[,j])
      ect[i,j] <- r.tot * c.tot / total
    }
  }
  chi <- 0
  for(i in 1:d[1]){
    for(j in 1:d[2]) {
      chi <- chi + (ct[i,j] - ect[i,j])^2 / ect[i,j]
    }
  }
  p = pchisq(chi, df = df, lower.tail = FALSE)

  print(paste("The chi-squared statistic is: ",chi, 3))
  if(p != 0) {
    print(paste("The p-value is: ", p))
  } else {
    print("p-value << 0.0001")
  }
}
```

2

```
airbnb <- read.csv('../../Data/listings.csv')

airbnb %>% select(price) %>% summary()
```

```
##      price
##  Min.   :    0.0
##  1st Qu.:   69.0
##  Median :  105.0
##  Mean   :  151.5
```

```
##   3rd Qu.:  175.0
##   Max.    :10000.0
```

a

I decided to look at the association of price on borough, but because price is a quantitative variable I first had to bin the values into categorical groups which I determined using the quantiles and labelled "Low Priced", "Moderately Low Priced", "Moderately High Priced", and "High Priced".

b

$$H_0: \text{There is no association between the price range of the listings and the borough it is located in.}$$

$$H_a: \text{The price range of the listings is associated with the borough it is located in.}$$

c

```r
encode <- function(price){
  result <- vector()
  for(i in 1:length(price)){
    if(price[i] <= 69) {
      result <- append(result, 'Low Priced')
    } else if(price[i] > 69 & price[i] <= 105){
      result <- append(result,'Moderately Low Priced')
    } else if(price[i] > 105 & price[i] <= 175) {
      result <- append(result,'Moderately High Priced')
    } else {
      result <- append(result,'High Priced')
    }
  }
  result
}

airbnb$price_cat <- encode(airbnb$price)

airbnb$price_cat <-  fct_relevel(airbnb$price_cat, c('Low Priced',
                                                     'Moderately Low Priced',
                                                     'Moderately High Priced',
                                                     'High Priced'))

conTable <-
table(airbnb[,c("neighbourhood_group","price_cat")])

conTable
```

```
##                      price_cat
## neighbourhood_group Low Priced Moderately Low Priced Moderately High Priced
##       Bronx                589                   296                    152
##       Brooklyn            6569                  5426                   4858
##       Manhattan           2481                  4582                   6189
##       Queens              2651                  1629                    982
##       Staten Island        158                   117                     64
##                      price_cat
## neighbourhood_group High Priced
##       Bronx                  68
##       Brooklyn             3261
##       Manhattan            8204
```

```
##       Queens                 549
##       Staten Island           39
```

The expected count for low priced listings in the Bronx is:

```
tot <- sum(conTable)
bronx.tot <-  sum(conTable["Bronx",])
low.p.tot <-  sum(conTable[,"Low Priced"])
exp.val <- bronx.tot*low.p.tot/tot
exp.val
```

```
## [1] 281.4964
```

The expected count for low priced listings in Brooklyn is:

```
tot <- sum(conTable)
brook.tot <-  sum(conTable["Brooklyn",])
low.p.tot <-  sum(conTable[,"Low Priced"])
exp.val <- brook.tot*low.p.tot/tot
exp.val
```

```
## [1] 5123.999
```

d Results of *my.chisq.test()* on the airbnb data:

```
conTable <- as.matrix(conTable)

my.chisq.test(conTable)
```

```
## [1] "The chi-squared statistic is:  6760.98516245231 3"
## [1] "p-value << 0.0001"
```

Originally my function returned 0 for the p-values because it was such a small number. I adjusted the function so that if *pchisq()* returned 0 the function would print that the p-value was much less than 0.0001.

Results of the *chisq.test()* function:

```
chisq.test(conTable)
```

```
##
##  Pearson's Chi-squared test
##
## data:  conTable
## X-squared = 6761, df = 12, p-value < 2.2e-16
```

This confirmed that the p-value was extremely small. This indicated the results would be extremely unlike if the null hypothesis were true. Therefore, we would use these results to reject the null hypothesis indicating that the may be an association between the borough of the listing and the price range of the listing (the variable are more likely to be dependant than independant).

e I will characterize the strength of the association using the highest priced listings.

```
conTable
```

```
##                    price_cat
## neighbourhood_group Low Priced Moderately Low Priced Moderately High Priced
##       Bronx                589                   296                    152
##       Brooklyn            6569                  5426                   4858
##       Manhattan           2481                  4582                   6189
##       Queens              2651                  1629                    982
##       Staten Island        158                   117                     64
```

```
##                     price_cat
## neighbourhood_group High Priced
##       Bronx                  68
##       Brooklyn             3261
##       Manhattan            8204
##       Queens                549
##       Staten Island          39
```

```r
pTable <-  conTable[,"High Priced"]
pTable["Bronx"] <-  pTable["Bronx"]/sum(conTable['Bronx',])
pTable["Brooklyn"] <-  pTable["Brooklyn"]/sum(conTable['Brooklyn',])
pTable["Manhattan"] <-  pTable["Manhattan"]/sum(conTable['Manhattan',])
pTable["Queens"] <-  pTable["Queens"]/sum(conTable['Queens',])
pTable["Staten Island"] <-  pTable["Staten Island"]/sum(conTable['Staten Island',])
pTable
```

```
##         Bronx      Brooklyn     Manhattan        Queens Staten Island
##    0.06153846    0.16212588    0.38236391    0.09447599    0.10317460
```

Considering the borough with the highest percentage of high priced listing and the borough with the lowest percentage of high priced listing (Manhattan versus Bronx):

```r
# Difference of proportions
abs(pTable['Manhattan'] - pTable['Bronx'])
```

```
## Manhattan
## 0.3208254
```

```r
# Relative risk
pTable['Manhattan'] / pTable['Bronx']
```

```
## Manhattan
##  6.213413
```

The proportion of listing in Manhattan that are high priced is 32 percentage points higher than the proportion of listings in the Bronx that are high priced. In other words, it is 6 times more likely that a listing in Manhattan would be considered high priced than a listing in the Bronx.

Problem 2

1 11.84

a

```r
pol.view.1 <-
matrix(c(56,490,NA,604,NA,NA,NA,24,58,509,61,628), nrow = 4, ncol = 3)
rownames(pol.view.1) <- c("Liberal","Moderate","Conservative","Total")
colnames(pol.view.1) <- c("Yes","No","Total")
pol.view.1
```

```
##              Yes No Total
## Liberal       56 NA    58
## Moderate     490 NA   509
## Conservative  NA NA    61
## Total        604 24   628
```

```r
#Filling in values
pol.view.1[1,2] <- 58-56
pol.view.1[2,2] <- 509-490
```

```
pol.view.1[3,1] <- 604 - (56 + 490)
pol.view.1[3,2] <- 61 - pol.view.1[3,1]

pol.view.1
```

```
##              Yes No Total
## Liberal       56  2    58
## Moderate     490 19   509
## Conservative  58  3    61
## Total        604 24   628
```

b

```
pol.view.2 <-
matrix(c(NA,NA,NA,604,NA,19,3,24,58,509,61,628), nrow = 4, ncol = 3)
rownames(pol.view.2) <- c("Liberal","Moderate","Conservative","Total")
colnames(pol.view.2) <- c("Yes","No","Total")
pol.view.2
```

```
##              Yes No Total
## Liberal       NA NA    58
## Moderate      NA 19   509
## Conservative  NA  3    61
## Total        604 24   628
```

```
#Filling in values
pol.view.2[3,1] <- 61-3
pol.view.2[2,1] <- 509-19
pol.view.2[1,2] <- 24 - (3 + 19)
pol.view.2[1,1] <- 58 - pol.view.2[1,2]

pol.view.2
```

```
##              Yes No Total
## Liberal       56  2    58
## Moderate     490 19   509
## Conservative  58  3    61
## Total        604 24   628
```

2 11.9

a

$$H_0 : \text{There is no association between gender and happiness}$$

$$H_a : \text{There is an association between gender and happiness}$$

b

The p-value is very high (> 0.50) which would be in favor of the null hypothesis being true. This would indicated that there is, indeed, no association between gender and hapiness.

Calculation of expected cell counts:

```
tot <- 154 + 592 + 336 + 123 + 502 + 257
N.tot <- 154 + 123
P.tot <- 592 + 336
V.tot <- 336 + 257
F.tot <- 154 + 592 + 336
M.tot <- 123 + 502 + 257
```

```r
#Female - Not Happy
N.tot*F.tot/tot
```

```
## [1] 152.6039
```

```r
#Female - Pretty Happy
P.tot*F.tot/tot
```

```
## [1] 511.2505
```

```r
#Female - Very Happy
V.tot*F.tot/tot
```

```
## [1] 326.6935
```

```r
#Male - Not Happy
N.tot*M.tot/tot
```

```
## [1] 124.3961
```

```r
#Male - Pretty Happy
P.tot*M.tot/tot
```

```
## [1] 416.7495
```

```r
#Male - Very Happy
V.tot*M.tot/tot
```

```
## [1] 266.3065
```

3 11.16

a

```r
Csd <- matrix(nrow = 2, ncol = 4)
rownames(Csd) <- c("Hancock","Trafford")
colnames(Csd) <- c("Fish","Invertebrates","Birds & Reptiles","Other")
Csd[1,1] <- round(30/55,2)
Csd[1,2] <- round(4/55,2)
Csd[1,3] <- round(8/55,2)
Csd[1,4] <- round(13/55,2)
Csd[2,1] <- round(13/53,2)
Csd[2,2] <- round(18/53,2)
Csd[2,3] <- round(12/53,2)
Csd[2,4] <- round(10/53,2)
Csd
```

```
##          Fish Invertebrates Birds & Reptiles Other
## Hancock  0.55          0.07             0.15  0.24
## Trafford 0.25          0.34             0.23  0.19
```

b

$$H_0 : \text{The primary food choice of alligators is independant of the lake.}$$

$$H_a : \text{The primary food choice of alligators is dependant on the lake.}$$

c The mean value of the chi-squared distribution will be approximately equal to the degrees of freedom and in this case that is $df = (2-1)*(4-1) = 3$. The chi-squared values of 16.79 for this table is more that 5 times the df indicating that it will be far out on the right tail, which would lead to a small p-value, and should be considered large.

d Because this is a small p-value we would choose the reject the null hypothesis, that there is no association with the primary food choice of alligators and the specific lake, in favor of the alternative hypothesis, that the primary choice of food for alligators depends on the specific lake in which they live.