M440B HW 9

Erick Castillo

3/15/2021

The following is a function I will use throughout the HW when performing χ^2 tests.

```
pears_chi <- function(obs, exp){
  return(sum((obs - exp)^2/exp))
}
```

It's best to define it early to make the work less tedious. Now…

Problem 5: The question asks if there is a relationship between a response to a question and ethnic groups. Data will be imported here.

```
#input of columns from book
obs.yes = c(78,56,43,53,43,36,42,29)
obs.no = c(47,29,29,32,30,22,23,7)
ethnic.total = obs.yes + obs.no

#this adds the last row which includes the sum of the respective col.
obs.yes = c(obs.yes, sum(obs.yes))
obs.no = c(obs.no, sum(obs.no))
ethnic.total = c(ethnic.total, sum(ethnic.total))

#following calculates the expected values for yes/no
exp.yes = tail(obs.yes,1)*ethnic.total/tail(ethnic.total, 1)
exp.no = tail(obs.no, 1)*ethnic.total/tail(ethnic.total, 1)

#calculating test statistic and p-value for each for obs/exp pair
ts1 = pears_chi(obs.yes, exp.yes) + pears_chi(obs.no, exp.no)
ts1
```

```
## [1] 6.029392
```

```
pchisq(ts1, df = 7, lower.tail = FALSE)
```

```
## [1] 0.5363217
```

The null hypothesis can be stated as, any ethnicity has an equal chance of selecting yes or no. From the above test statistic and corresponding p-value, there is good evidence to suggest that ethnicity is independent from the response to the question posed.

```
#this is what the data looks like all together
data.frame(obs.yes,obs.no,ethnic.total,exp.yes,exp.no)
```

```
##   obs.yes obs.no ethnic.total   exp.yes    exp.no
## 1      78     47          125  79.29883  45.70117
## 2      56     29           85  53.92321  31.07679
## 3      43     29           72  45.67613  26.32387
## 4      53     32           85  53.92321  31.07679
## 5      43     30           73  46.31052  26.68948
## 6      36     22           58  36.79466  21.20534
## 7      42     23           65  41.23539  23.76461
## 8      29      7           36  22.83806  13.16194
## 9     380    219          599 380.00000 219.00000
```

Problem 9: This question asks whether Austen was consistent across the works she created. Then asks if her imitator was successful in copying her style. I will first compare Sense and Sensibility with Emma.

```
totalsse = obs.ss + obs.emma
exp.ss1 = tail(obs.ss, 1)*totalsse/tail(totalsse, 1)
exp.emma1 = tail(obs.emma, 1)*totalsse/tail(totalsse, 1)
ts2 = pears_chi(obs.ss,exp.ss1) + pears_chi(obs.emma,exp.emma1)
ts2
```

```
## [1] 6.174354
```

```
pchisq(ts2, df=5, lower.tail = FALSE)
```

```
## [1] 0.2896214
```

The above output is the test statistic and the p-value for the homogeneity test. Both show that there is good evidence that Austen was consistent across her works Sense and Sensibility and Emma.

I will now test if she was consistent across the previously tested books (pooled data) and Sanditon I.

```
obs.sse = totalsse
tot2 = obs.sse + obs.sand1 #sum of pooled data and sand1

#calculate expected values for each:
exp.sse = tail(obs.sse,1)*tot2/tail(tot2,1)
exp.sand1 = tail(obs.sand1,1)*tot2/tail(tot2,1)
ts3 = pears_chi(obs.sse,exp.sse) + pears_chi(obs.sand1,exp.sand1)
ts3
```

```
## [1] 17.34488
```

```
pchisq(ts3, df = 5, lower.tail = FALSE)
```

```
## [1] 0.003890107
```

The above outputs for the test statistic and the p-value suggest very strongly that Austen was not consistent across her own books.

Now I'll check whether if Sanditon I and Sandition II were consistent.

```
tot.sss = obs.sand1+obs.sand2
exp.sand1 = tail(obs.sand1,1)*tot.sss/tail(tot.sss,1) #reassignment
exp.sand2 = tail(obs.sand2,1)*tot.sss/tail(tot.sss,1)
ts4 = pears_chi(obs.sand1, exp.sand1) + pears_chi(obs.sand2, exp.sand2)
ts4
```

```
## [1] 17.77381
```

```
pchisq(ts4,df=5,lower.tail=FALSE)
```

```
## [1] 0.00324366
```

The above test statistic and p-value indicates that Austen's imitator was not consistent with the writing style found in Sanditon I.

Probelm 16: The following problem includes a 3 × 3 table of data. The columns seem to sum up to 100, so this appears to be a test of independence.

```
##     attitudes obs.cautious obs.midroad obs.explorer att.total
## 1   favorable           79          58           49       186
## 2     neutral           10           8            9        27
## 3 unfavorable           10          34           42        86
## 4       total           99         100          100       299
```

```
exp.cautious <- tail(obs.cautious,1)*att.total/tail(att.total,1)
exp.midroad <- tail(obs.midroad,1)*att.total/tail(att.total,1)
exp.explorer <- tail(obs.explorer,1)*att.total/tail(att.total,1)

ts5 = pears_chi(obs.cautious,exp.cautious)+pears_chi(obs.midroad,exp.midroad)+pears_chi(obs.explorer,exp
ts5
```

```
## [1] 27.2892
```

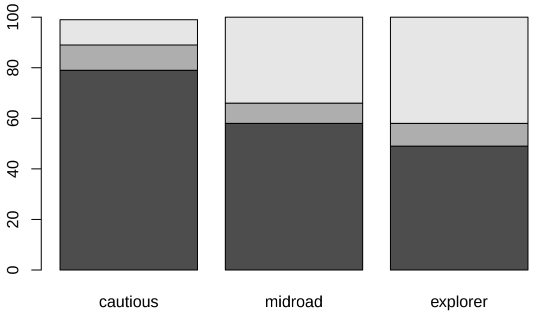
```
pchisq(ts5,df=4,lower.tail = FALSE)
```

```
## [1] 1.737411e-05
```

The above test statistic and corresponding p-value are strong evidence suggesting that there is a relationship between personality type and small cars. The relationship appears to go as follows:

- Cautious people tend to have a more favorable attitude towards individuals driving in small cars.
- Midroad people appear to be split between favorable and unfavorable attitudes, though more of these people appear to favor small cars. Very few people fall in the neutral ground.
- Explorer people, are very split, with few people falling in the neutral category.

```r
rownames(data16) <- attitudes1
barplot(data16)
```



Here is the stacked bar plot of the data.

Problem 29: This question is asking, given the scenario, would I, the company's statistician, perform a χ^2 test of independence or homogeneity.

The type of test to be utilized would completely depend on the way the sampling was made. For example:

- If the sample of 100 was made by taking random people from population of employees, then dividing them into categories of male/female, then this would require a test of independence.
- If the amount of people to be selected from each sex were fixed, that is 50 men and 50 women, then this would require a test of homogeneity.

Problem 14: I will begin by inputting the necessary data for both of the tables in the problem:

```r
#first table. No high school:
under45.nohs.obs = c(71,305)
over45.nohs.obs = c(217,652)
under45.nohs.obs = c(under45.nohs.obs, sum(under45.nohs.obs))
over45.nohs.obs = c(over45.nohs.obs, sum(over45.nohs.obs))
total.nohs.obs = under45.nohs.obs + over45.nohs.obs

#second table. Some high school:
under45.some.obs = c(305,869)
over45.some.obs = c(180,259)
under45.some.obs = c(under45.some.obs, sum(under45.some.obs))
over45.some.obs = c(over45.some.obs, sum(over45.some.obs))
total.some.obs = under45.some.obs + over45.some.obs
```

A. This question asks to analyze the dependence of interest in political elections on age and education. The following code calculates the expected values for each of the tables, obtains a test statistics, and p-value.

```r
#expected values for no high school:
under45.nohs.exp = tail(under45.nohs.obs,1)*total.nohs.obs/tail(total.nohs.obs,1)
over45.nohs.exp = tail(over45.nohs.obs,1)*total.nohs.obs/tail(total.nohs.obs,1)

#expected values for some high school:
under45.some.exp = tail(under45.some.obs,1)*total.some.obs/tail(total.some.obs,1)
over45.some.exp = tail(over45.some.obs,1)*total.some.obs/tail(total.some.obs,1)

#test statistic for independence:
ts6 = pears_chi(under45.nohs.obs,under45.nohs.exp) + pears_chi(over45.nohs.obs,over45.nohs.exp) +
  pears_chi(under45.some.obs,under45.some.exp) + pears_chi(over45.some.obs, over45.some.exp)
ts6
```

```
## [1] 39.76451
```

```r
#p-value for independence:
pchisq(ts6, df = 2*1*1, lower.tail = FALSE)
```

```
## [1] 2.31871e-09
```

The above output indicates that given interest in political elections, there appears to be strong evidence to suggest that there is a relationship between age and education.

B. This question asks to test the following hypothesis:

- H_1 : given education, age and interest in politics are unrelated.
- H_2 : given age, education and interest are unrelated.

This first wall of code will test H_1.

```r
#observed values for under 45:
lit.und.obs = c(305,869)
great.und.obs = c(71,305)
lit.und.obs = c(lit.und.obs, sum(lit.und.obs))
great.und.obs = c(great.und.obs, sum(great.und.obs))
tot.und = lit.und.obs + great.und.obs

#observed values for over 45:
lit.ov.obs = c(652,259)
great.ov.obs = c(217,180)
lit.ov.obs = c(lit.ov.obs, sum(lit.ov.obs))
great.ov.obs = c(great.ov.obs, sum(great.ov.obs))
tot.ov = lit.ov.obs + great.ov.obs

#expected values for under 45 table:
lit.und.exp = tail(lit.und.obs,1)*tot.und/tail(tot.und,1)
great.und.exp = tail(great.und.obs,1)*tot.und/tail(tot.und,1)

#expected values for over 45 table:
lit.ov.exp = tail(lit.ov.obs,1)*tot.ov/tail(tot.ov,1)
great.ov.exp = tail(great.ov.obs,1)*tot.ov/tail(tot.ov,1)
```

```
#test statistic for the above values:
ts7 = pears_chi(lit.und.obs, lit.und.exp) + pears_chi(great.und.obs, great.und.exp) +
  pears_chi(lit.ov.obs, lit.ov.exp) + pears_chi(great.ov.obs, great.ov.exp)
ts7
```

```
## [1] 43.26342
```

```
pchisq(ts7, df = 2, lower.tail = FALSE)
```

```
## [1] 4.031504e-10
```

The above test statistic and p-value indicate that there is strong evidence to suggest that given education, there appears to be a relationship between age and the degree of interest in politics.

The following test statistic and p-value output will pertain to tables distinguishing between little and great interest in politics. I will refrain from outputting the code I used to prevent this assignment from being any longer.

```
## [1] 535.0899
```

```
## [1] 6.407734e-117
```

This test statistic is huge! It's 535.08. This alone would provide enough evidence to suggest that given age, educational level and degree of interest in politics do have a strong relationship with one another.

To conclude, I rejected H_0, H_1, and H_2. This means that given an one of the three parameters recorded, there is strong evidence to suggest that the other two parameters are related.