Homework 06

Christina Lee

10/13/2021

1.

a. Based on the exit poll results, is age independent of Party ID or not? Conduct a chi-squared test by hand, showing each step in readably-formatted latex.

Percent 18-29 = 199/792 = 0.251
Percent 30-44 = 197/792 = 0.249
Percent 45-59 = 198/792 = 0.25
Percent 60+ = 198/792 = 0.25
Percent Democrat = 302/792 = 0.381
Percent Independent = 212/792 = 0.267
Percent Republican = 278/792 = 0.351

p(Democrat&18-29) = 0.251 * 0.381, thus the total number of 18-29 Democrats we would expect to see is 75.74
p(Democrat&30-44) = 0.249 * 0.381, thus the total number of 30-44 Democrats we would expect to see is 75.14
p(Democrat&45-59) = 0.25 * 0.381, thus the total number of 45-59 Democrats we would expect to see is 75.44
p(Democrat&60+) = 0.25 * 0.381, thus the total number of 60+ Democrats we would expect to see is 75.44
p(Indep&18-29) = 0.251 * 0.267, thus the total number of 18-29 Indep we would expect to see is 53.08
p(Indep&30-44) = 0.249 * 0.267, thus the total number of 30-44 Indep we would expect to see is 52.65
p(Indep&45-59) = 0.25 * 0.267, thus the total number of 45-59 Indep we would expect to see is 52.87
p(Indep&60+) = 0.25 * 0.267, thus the total number of 60+ Indep we would expect to see is 52.87
p(Rep&18-29) = 0.251 * 0.351, thus the total number of 18-29 Rep we would expect to see is 69.78
p(Rep&30-44) = 0.249 * 0.351, thus the total number of 30-44 Rep we would expect to see is 69.22
p(Rep&45-59) = 0.25 * 0.351, thus the total number of 45-59 Rep we would expect to see is 69.50
r(Rep&60+) = 0.25 * 0.351, thus the total number of 60+ Rep we would expect to see is 69.50

$$X^2 = \sum \frac{(f_o - f_e)^2}{f_e}$$

$$X^2 = \frac{(86 - 75.74)^2}{75.74} + \frac{(52 - 53.08)^2}{53.08} ... + \frac{(73 - 69.50)^2}{69.50} = 3.7$$

$$X^2 = 3.7$$

$$df = (r - 1)(c - 1)$$

$$df = (4 - 1)(3 - 1) = 6$$

```
qchisq(0.95,df=6) # threshold value
```

```
## [1] 12.59159
```

```r
1-pchisq(3.7,df=6) # p-value
```

```
## [1] 0.717198
```

1.

$$H_a = Variables\ are\ not\ independent$$

$$H_o = Variables\ are\ independent$$

2. Since X^2= 3.7 < 12.6 threshold value, we fail to reject Null.
3. Since p-value 0.717 > 0.05, we again fail to reject Null.

Therefore, we fail to reject the null hypothesis and accept our null hypothesis. This suggests that age is independent of Party ID.

b. Verify your results using R to conduct the test.

Results from R matches with the results in 1a.

```r
# save data into a data.frame and save it under a variable for chi-square test.
ageparty <- data.frame(democrat=c(86,72,73,71), independent=c(52,51,55,54),
                       republican=c(61,74,70,73),row.names=c("18-29","30-44","45-59","60+"))
ageparty
```

```
##       democrat independent republican
## 18-29       86          52         61
## 30-44       72          51         74
## 45-59       73          55         70
## 60+         71          54         73
```

```r
chisq.test(ageparty) # conduct chi-square test.
```

```
##
##  Pearson's Chi-squared test
##
## data:  ageparty
## X-squared = 3.6529, df = 6, p-value = 0.7235
```

2.

a. Now test for independence using ANOVA (an F test). Your three groups are Democrats, Independents, and Republicans. The average age for a Democrat is 43.3, for an Independent it's 44.6, and for a Republican it's 45.1. The standard deviations of each are D: 9.1, I: 9.2, R: 9.2. The overall mean age is 44.2. Do the F test by hand, again showing each step.

$$H_a = The\ average\ age\ is\ different\ across\ the\ PartyID$$

$$H_o = The\ average\ age\ is\ the\ same\ across\ the\ PartyID$$

$$f_{stat} = \frac{average\ variance\ between\ groups}{average\ variance\ within\ groups}$$

$$Between\ Groups = \frac{n_1(y_1 - y)^2 + ...n_G(y_G - y)^2}{df = G - 1}$$

$$Within\ Groups = \frac{(n_1 - 1)s_1^2 + ...(n_G - 1)s_G^2}{df = N - G}$$

$$Between\ Groups = \frac{302(43.3-44.2)^2 + 212(44.6-44.2)^2 + 278(45.1-44.2)^2}{3-1} = 251.86$$

$$Within\ Groups = \frac{(302-1)(9.1)^2 + (212-1)(9.2)^2 + (278-1)(9.2)^2}{792-3} = 83.94$$

$$F\ Statistic = \frac{251.86}{83.94} = 3.0$$

$$df_1 = 3-1 = 2$$

$$df_2 = 792-3 = 789$$

```
qf(0.95,2,789)  #threshold value for F test.
```

```
## [1] 3.007136
```

```
1-pf(3,2,789) #p-value
```

```
## [1] 0.05035534
```

1. Since F-stat= 3.0 (before rounding F-stat= 3.00047653) is not greater than 3.007136 threshold value, we fail to reject the null (very close values, but still slightly smaller than the threshold value).
2. Since p-value= 0.05035534 $>$ 0.05, we fail to reject the null (although the values are very close, but it is still greater than 0.05).

b. Check your results in R using simulated data. Your simulated dataset should have two variables: age, and a factor indicating for each row whether it is Democrat, Independent, or Republican. There should be 302 Democrats drawn from a normal distribution with mean 43.3 and sd 9.1, and likewise for Independents and Republicans. One way to construct this dataset is to first use cbind to create the Democratic, Independent, and Republican datasets, and then use rbind to stack them together: eg, use democrats <- cbind(rnorm(302,43.3,9.1),"democrat") and likewise for the other two groups; then use rbind to glue the democrat, indpendent, and republican datasets into one data frame; and then be sure to name your variables and make sure that the first is numeric and the second a factor. Once this dataset has been constructed, conduct an F test using R's aov function on the ages and compare the results to 2a. Do your results match 2a? If not, why not?

```
# Create the data frame.
set.seed(1)
Dem <- cbind(rnorm(302,43.3,9.1),"Democrat")
Indep <- cbind(rnorm(212,44.6,9.2), "Independent")
Repub <- cbind(rnorm(278,45.1,9.2), "Republican")

data_ageparty <- data.frame(rbind(Dem,Indep,Repub))
colnames(data_ageparty) <- c("Age","PartyID")
data_ageparty$Age<- as.numeric(as.character(data_ageparty$Age))
head(data_ageparty)
```

```
##        Age  PartyID
## 1 37.59927 Democrat
## 2 44.97115 Democrat
## 3 35.69578 Democrat
## 4 57.81706 Democrat
## 5 46.29852 Democrat
## 6 35.83374 Democrat
```

```r
# Conduct the ANOVA test.

anova_ageparty <- aov(Age~PartyID, data=data_ageparty)
summary(anova_ageparty)
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## PartyID       2    155   77.55   0.878  0.416
## Residuals   789  69730   88.38
```

Results:
1. Since F-stat= 0.87 is not greater than 3.0 threshold value, we fail to reject the null.
2. Since p-value= 0.416 > 0.05, we fail to reject the null.

No, the Fstat and P-value do not match with the results that was calculated in 2a. If you don't set.seed and continue to run the ANOVA test and the dataframe a couple more times, you will realize, some times the Fstat and P-value are much greater/smaller than the calculated ones in 2a. Such that, the results from R's F test are not consistent, even though we have already input the n, mean, and sd for our data set in the rnorm function (each time you run the dataframe with the rnorm function, the mean and sd will still change due to the different data that's been generated each time). This suggests that it is much better to conduct the F test by using the "actual" data, rather than summary data to avoid inaccuracies in tests.