DSSH 6301 - HW 06 Solutions

Problem 1

Our hypotheses are as follows:

Null hypothesis: Age is independent of Party ID.

Alternative hypothesis: Age is not independent of Party ID.

First we construct our table in R. Note the use of `ftable` for contingency tables , which while not necessary, is convenient (see `?ftable` for more), and the use of `margin.table()`, which sums across rows, columns, or both.

```
table <- matrix(c(86, 52, 61, 72, 51, 74, 73, 55, 70, 71, 54, 73), nrow=3)

rownames(table) <- c("dem", "ind", "rep")
colnames(table) <- c("18-29", "30-44", "45-59", "60+")

ftable(table)
```

```
##      18-29 30-44 45-59 60+
##
## dem     86    72    73  71
## ind     52    51    55  54
## rep     61    74    70  73
```

```
margin.table(table, 1)
```

```
## dem ind rep
## 302 212 278
```

```
margin.table(table, 2)
```

```
## 18-29 30-44 45-59   60+
##   199   197   198   198
```

```
margin.table(table)
```

```
## [1] 792
```

Part a

Democrat calculations

Recall that the formula to calculate the chi-square is the following:

$$\chi^2 = \sum_{i=1}^{n} \frac{(f_o - f_e)^2}{f_e}$$

. This means that for every observation we need to know the observed and expected values. The observed values are included in the table. We calculate all

$$f_e$$

's in the following steps:

18-29:

$$f_{e_1} = \frac{302 * 199}{792^2} * 792 = \frac{302 * 199}{792} = 75.88131$$

30-44:

$$f_{e_2} = \frac{302 * 197}{792} = 75.11869$$

45-59:

$$f_{e_3} = \frac{302 * 198}{792} = 75.5$$

60+:

$$f_{e_4} = \frac{302 * 198}{792} = 75.5$$

Independent calculations

18-29:

$$f_{e_5} = \frac{212 * 199}{792} = 53.26768$$

30-44:

$$f_{e_6} = \frac{212 * 197}{792} = 52.73232$$

45-59:

$$f_{e_7} = \frac{212 * 198}{792} = 53$$

60+:

$$f_{e_8} = \frac{212 * 198}{792} = 53$$

Republican calculations

18-29:

$$f_{e_9} = \frac{278 * 199}{792} = 69.85101$$

30:44:

$$f_{e_10} = \frac{278 * 197}{792} = 69.14899$$

45-59:

$$f_{e_11} = \frac{278 * 198}{792} = 69.5$$

60+:

$$f_{e_12} = \frac{278 * 198}{792} = 69.5$$

$$\chi^2 = \sum_{i=1}^{n} \frac{(f_o - f_e)^2}{f_e}$$

$$\chi^2 = \frac{(86 - 75.88131)^2}{75.88131} + \frac{(72 - 75.11869)^2}{75.11869} + \frac{(73 - 75.5)}{75.5} + \frac{(71 - 75.5)}{75.5} +$$
$$\frac{(52 - 53.26768)^2}{53.26768} + \frac{(51 - 52.73232)^2}{52.73232} + \frac{(55 - 53)}{53} + \frac{(54 - 53)}{53} +$$

$$\frac{(61-69.85101)^2}{69.85101}+\frac{(74-69.14899)^2}{69.14899}+\frac{(70-69.5)}{69.5}+\frac{(73-69.5)}{69.5}$$

$$\chi^2 = 3.652908$$

```
matrix_dims <- dim(table)
fe <- matrix(NA,matrix_dims[1],matrix_dims[2])
for(i in 1:matrix_dims[1]){
  for(j in 1:matrix_dims[2]){
    fe[i,j] <- margin.table(table,1)[i] *  margin.table(table,2)[j] / margin.table(table)
  }
}
chi_sq <- sum((table-fe)^2/fe)
chi_sq
```

```
## [1] 3.652908
```

Degrees of Freedom

$$df = (n_{col} - 1) * (n_{row} - 1) = 3 * 2 = 6$$

```
df <- (nrow(table)-1) * (ncol(table)-1)
df
```

```
## [1] 6
```

95% Threshold Value

```
qchisq(0.95, df=df)
```

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

P-value

```
pchisq(chi_sq, df=df, lower.tail=F)
```

```
## [1] 0.7235272
```

Based on both the threshold value and the p-value, we fail to reject the null hypothesis, suggesting that age is not independent of Party ID.

Part b

```
chisq.test(table)
```

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

We fail to reject the null.

Problem 2

Part a

Test Statistic Calculation

$$n_1 = 302, \bar{y}_1 = 43.3, \sigma_1 = 9.1$$
$$n_2 = 212, \bar{y}_2 = 44.6, \sigma_2 = 9.2$$
$$n_3 = 278, \bar{y}_3 = 45.1, \sigma_3 = 9.2$$
$$G = 3, N = 792$$
$$\bar{y} = 44.2$$

Between Group Variance:

$$var_{between} = \sum_{i=1}^{k} \frac{n_i * (\bar{y}_i - \bar{y})^2}{G - 1}$$
$$var_{between} = \frac{n_1 * (\bar{y}_1 - \bar{y})^2 + n_2 * (\bar{y}_2 - \bar{y})^2 + n_3 * (\bar{y}_3 - \bar{y})^2}{G - 1}$$
$$var_{between} = \frac{302 * (43.3 - 44.2)^2 + 212 * (44.6 - 44.2)^2 + 278 * (45.1 - 44.2)^2}{2} = 251.86$$

Within Group Variance:

$$var_{within} = \sum_{i=1}^{k} \frac{s_i^2 * (n_i - 1)}{N - G}$$
$$var_{within} = \frac{s_1^2 * (n_1 - 1) + s_2^2 * (n_2 - 1) + s_3^2 * (n_3 - 1)}{N - G}$$
$$var_{within} = \frac{9.1^2 * (302 - 1) + 9.2^2 * (212 - 1) + 9.2^2 * (278 - 1)}{789} = 83.94186$$

F statistic:

$$F = \frac{between}{within} = \frac{251.86}{83.94} = 3.00041$$

```
n <- margin.table(table, 1)
n
```

```
## dem ind rep
## 302 212 278
```

```
N <- sum(n)
N
```

```
## [1] 792
```

```
G <- nrow(table)
G
```

```
## [1] 3
```

```r
names <- c("dems", "inds", "reps")
group_means <- c(43.3, 44.6, 45.1)
group_sds <- c(9.1, 9.2, 9.2)

names(group_means) <- names
names(group_sds) <- names

group_means
```

```
## dems inds reps
## 43.3 44.6 45.1
```

```r
group_sds
```

```
## dems inds reps
##  9.1  9.2  9.2
```

```r
mean_age <- 44.2

between_var <- sum(n*(group_means - mean_age)^2) / (G-1)
within_var <- sum((n-1)*group_sds^2) / (N-G)

between_var
```

```
## [1] 251.86
```

```r
within_var
```

```
## [1] 83.94186
```

```r
f <- between_var / within_var
f
```

```
## [1] 3.00041
```

Degrees of Freedom

```r
df1 <- G - 1
df2 <- N - G

df1
```

```
## [1] 2
```

```r
df2
```

```
## [1] 789
```

95% Threshold Value

```r
qf(0.95, df1, df2)
```

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

P-value

```
pf(f, df1, df2, lower.tail=F)
```

```
## [1] 0.05033486
```

Based on both the threshold value and the p-value, we fail to reject the null hypothesis (but just barely!), still suggesting that age is independent of Party ID>

Part b

Note that the simulation results will not necessarily come to the same conclusion as the results using the exact data. Since 2a was right on the threshold (p was nearly 0.05), the simulation may reject or fail to reject depending on the exact draw we get.

```
set.seed(1)
d <- cbind("d",rnorm(302,43.3,9.1))
i <- cbind("i",rnorm(212,44.6,9.2))
r <- cbind("r",rnorm(278,45.1,9.2))
agepol <- data.frame(rbind(d,i,r),stringAsFactors=FALSE)
colnames(agepol) <- c("party","age")
agepol$party <- as.factor(agepol$party)
agepol$age <- as.numeric(agepol$age)
summary(aov(age~party,data=agepol))
```

```
##              Df   Sum Sq Mean Sq F value Pr(>F)
## party         2    98449   49225    0.94  0.391
## Residuals   789 41300909   52346
```

In our simulation we also fail to reject the null hypothesis, again suggesting that age is not independent of Party ID.

Note that for the F test output, the first column is the between and within denomitors (respectively), the second column is the between and within numerators, the third column is the between and within variances (column 2 divided by column 1), and the F value is the between variance divided by the within variance. The P-value, as ever, is 1-pf(F value, df1, df2), or `1-pf(0.94,2,789)`.