Holsenbeck_S_6

Stephen Synchronicity

2017-10-19

Homework Outline

Load Data

```
# Create the Tibble
ep <- tibble::tribble(~Party, ~`18-29`, ~`30-44`, ~`45-59`, ~`60+`, "Democrat", 86L,
    72L, 73L, 71L, "Independent", 52L, 51L, 55L, 54L, "Republican", 61L, 74L, 70L,
    73L)
# Add sums for cols and rows
ep$Totals <- rowSums(ep[, c(2:5)])
c.sums <- c("Totals", as.vector(colSums(ep[, c(2:6)], dims = 1)))
as.numeric(c.sums[2:6])
```

```
## [1] 199 197 198 198 792
```

```
ep <- rbind(ep, c.sums)
# Apparently rbind converts all the values to characters, this next f(n) converts
# them back to numeric
ep <- transform(ep, `18-29` = as.numeric(`18-29`), `30-44` = as.numeric(`30-44`),
    `45-59` = as.numeric(`45-59`), `60+` = as.numeric(`60+`))
# Fix the column headers
names(ep) <- c("Party", "18-29", "30-44", "45-59", "60+", "Totals")
```

Use Google Sheets to Calculate Chi Squared by hand

```
# Use Google Sheets to Calculate the Chi Squared 'by hand'
library(googlesheets)
# Authorize a northeastern acct with the gs_auth() funvtion to have access to the
# sheet below

# gs_auth() ## register a google sheet object with the URL

gepURL <- gs_url("https://docs.google.com/spreadsheets/d/11AVqy2JoGA8YT8tSPTk3VJxEuxsuGBOFqqoNaAXx6SM/edit#gid=0")
# Next f(n) was used to write data to the sheet

# gep <- gep %>% gs_edit_cells(input = ep, col_names=T)

# read data from the sheet, store as gep data frame object
gep <- gs_read(gepURL, ws = 1)
```

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.

A. H_0H0: Age is independent of Party ID.

H_1H1: Party ID and age are dependent variables.

$f_e = \frac{\text{(row total)(column total)}}{\text{overall total}}$ fe=(row total)(column total)overall total

$\frac{(302)(199)}{792} = 75.88131$ (302)(199)792=75.88131

$\frac{(212)(199)}{792} = 53.26768$ (212(199)792=53.26768

$\frac{(278)(199)}{792} = 69.85101$ (278)(199)792=69.85101

$\frac{(302)(197)}{792} = 75.11869$ (302)(197)792=75.11869

$\frac{(212)(197)}{792} = 52.73232$ (212(197)792=52.73232

$\frac{(278)(197)}{792} = 69.14899$ (278)(197)792=69.14899 $\frac{(302)(198)}{792} = 75.5$ (302)(198)792=75.5

$\frac{(212)(198)}{792} = 53.0$ (212(198)792=53.0

$\frac{(278)(198)}{792} = 69.5$ (278)(198)792=69.5 $\frac{(302)(198)}{792} = 75.5$ (302)(198)792=75.5

$\frac{(212)(198)}{792} = 53.0$ (212(198)792=53.0

$\frac{(278)(198)}{792} = 69.5$ (278)(198)792=69.5

$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$ χ2=∑(fo−fe)2fe

$\chi^2 = \sum \frac{(86-75.88131)^2}{75.88131} = 1.34931539$ χ2=∑(86−75.88131)275.88131=1.34931539

$\chi^2 = \sum \frac{(52-53.26768)^2}{53.26768} = 0.03016847$ χ2=∑(52−53.26768)253.26768=0.03016847

$\chi^2 = \sum \frac{(61-69.85101)^2}{69.85101} = 1.12153539$ χ2=∑(61−69.85101)269.85101=1.12153539

$\chi^2 = \sum \frac{(72-75.11869)^2}{75.11869} = 0.1294779$ χ2=∑(72−75.11869)275.11869=0.1294779

$\chi^2 = \sum \frac{(51-52.73232)^2}{52.73232} = 0.0569090$ χ2=∑(51−52.73232)252.73232=0.0569090

$\chi^2 = \sum \frac{(74-69.14899)^2}{69.14899} = 0.3403130$ χ2=∑(74−69.14899)269.14899=0.3403130

$\chi^2 = \sum \frac{(73-75.5)^2}{75.5} = 0.082781457$ χ2=∑(73−75.5)275.5=0.082781457

$\chi^2 = \sum \frac{(55-53.0)^2}{53.0} = 0.075471698$ χ2=∑(55−53.0)253.0=0.075471698

$\chi^2 = \sum \frac{(70-69.5)^2}{69.5} = 0.003597122$ χ2=∑(70−69.5)269.5=0.003597122

$\chi^2 = \sum \frac{(71-75.5)^2}{75.5} = 0.26821192$ χ2=∑(71−75.5)275.5=0.26821192

$\chi^2 = \sum \frac{(54-53.0)^2}{53.0} = 0.01886792$ χ2=∑(54−53.0)253.0=0.01886792

$\chi^2 = \sum \frac{(73-69.5)^2}{69.5} = 0.17625899$ χ2=∑(73−69.5)269.5=0.17625899

$\chi^2 = 3.652908$ χ2=3.652908

$df = (3-1)(4-1)$ df=(3−1)(4−1)

$df = 6$ df=6

$\chi_{crit} = 12.59159$ χcrit=12.59159

$p_{value} = 0.7235272$ pvalue=0.7235272

Conclusion: Fail to reject the null hypothesis, age is independent of party affiliation.

b.

Verify your results using R to conduct the test.

1b

```
qchisq(0.95, 6)
```

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

```
1 - pchisq(3.652908, 6)
```

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

```
chisq.test(ep[c(1:3), c(2:5)])
```

```
##
##  Pearson's Chi-squared test
##
## data:  ep[c(1:3), c(2:5)]
## X-squared = 3.6529, df = 6, p-value = 0.7235
```

A. The values and conclusions are consistent between the ‘by hand’ and R χ^2χ2 test.

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.

$H_0 : \mu_D = \mu_I = \mu_R$ H0:μD=μI=μR The average age is the same across the affiliations. $H_1 : \neg(\mu_D = \mu_I = \mu_R)$ H1:¬(μD=μI=μR) The average age is different across the affiliations. F-Test

$$
\begin{aligned}
f_{stat} &= \frac{\text{average variance between groups}}{\text{average variance within groups}} \\
\text{between groups} &= \frac{n_1(\bar{y}_1 - \bar{y})^2 + \ldots + n_G(\bar{y}_G - \bar{y})^2}{df = G - 1} \\
\text{within groups} &= \frac{(n_1 - 1)s_1^2 + \ldots + (n_G - 1)s_G^2}{df = N - G} \\
\text{where } N &= \text{sum(n) in all, G = \# of Groups} \\
\text{compare } f_{stat} &\text{ to qf(cl}, df_1, df_2) \\
\text{or compare } p_{value} &= \text{1-pf}(f_{stat}, df_1, df_2) \text{ to } \alpha
\end{aligned}
$$

fstat=average variance between groupsaverage variance within groupsbetween groups=n1(y¯1−y¯)2+...+nG(y¯G−y¯)2df=G−1within groups=(n1−1)s12+...+

$$\text{between groups} = \frac{302(43.3-44.2)^2+212(44.6-44.2)^2+278(45.1-44.2)^2}{df=3-1}$$

between groups=302(43.3−44.2)2+212(44.6−44.2)2+278(45.1−44.2)2df=3−1

$$\text{within groups} = \frac{(302-1)1.28^2+(212-1)1.43^2+(278-1)1.1^2}{df=792-3}$$

within groups=(302−1)1.282+(212−1)1.432+(278−1)1.12df=792−3

2a - Computations with R

```r
# Summary data input vectors
mu <- c(43.3, 44.6, 45.1)
sd <- c(9.1, 9.2, 9.2)
n <- gep$Totals[1:3]

# Function for anova test with input values

# Inputs:

# y - Vector of means

# mu - Mean of means, finds it from input values if not declared

# sd - Vector of standard deviations

# n - Vector of group totals

# Returns - f-statistic

anva <- function(y, mu = mean(y), s, n) {
    wg.v <- vector("numeric")
    bg.v <- vector("numeric")
    for (i in 1:length(n)) {
        # Between Group Variance
        bgc <- n[i] * (y[i] - mu)^2
        bg.v <- append(bg.v, bgc, after = length(bg.v))
        i <- i + 1
    }
    bgvar <- sum(bg.v)/(length(n) - 1)
    i <- 1
    for (i in 1:length(n)) {
        # within Group Variance
        wgc <- (n[i] - 1) * s[i]^2
        wg.v <- append(wg.v, wgc, after = length(wg.v))
        i <- i + 1
    }
    wgvar <- sum(wg.v)/(sum(n) - length(n))
    fs <- bgvar/wgvar
    pv <- 1 - pf(fs, length(n) - 1, (sum(n) - length(n)))
    output <- tibble::tribble(~Param, ~Value, "Fstat", fs, "pValue", pv)
    return(output)
}
anva(mu, 44.2, sd, n)
```

```
## # A tibble: 2 x 2
##    Param      Value
##    <chr>      <dbl>
## 1  Fstat 3.00040993
## 2 pValue 0.05033486
```

```r
qf(0.95, 2, 789)
```

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

A)The results of the F-Test with the summary data provided indicate that we fail to reject the null at the 95% confidence level, and conclude that the average age is the same across the party affiliations.

b.

Check your results in R using simulated data. Generate a simulated dataset by creating three vectors: Democrats, Republicans, and Independents. Each vector should be a list of ages, each with a length equal to the number of Democrats, Independents, and Republicans in the table above, and the appropriate mean and sd based on 2.a (use rnorm to generate the vectors). Combine all three into a single dataframe with two variables: age, and a factor that specifies D, I, or R. Then conduct an F test using R's aov function on that data and compare the results to 2a. Do your results match 2a? If not, why not?

A)
Results The results from the simulated data experiment do not match the results from 2a. When we take a look at the actual data from the simulation (See 2nd chunk below). The f_{stat} fstat is 4.4 with the experimental data, indicating we reject the null at the 95% confidence level in favor of the research hypothesis which is that age is different between party affiliations, and they may be dependent variables.
Discussion We can see consistency in the data between the input vectors and the data from the dataframe used in the experiment. However, when we compare the actual mean and sd of the experimental data, to the given mean & sd inputs for the rnorm function, we can see that these values are slightly different. The difference between the actual experimental data mean & sd, and the given mean & sd used in the calculations likely accounts for the differing outcome between the anova test in R and the manual F test. **Conclusion** This would suggest that a savvy statistician should avoid using summary data for inputs and always run the test with the actual data to avoid significant accumulated error that results in different test outcomes.

2b

```r
# Create the data frames with rnorm
D <- data.frame(rnorm(n[1], mean = mu[1], sd[1]))
I <- data.frame(rnorm(n[2], mean = mu[2], sd[2]))
R <- data.frame(rnorm(n[3], mean = mu[3], sd[3]))
# I am unable to find a way to create a data frame with vectors of unequal length
# As a workaround, I'll use Google Sheets and then read the data back

# Create a new Worksheet named 2b gs_ws_new(gepURL,ws_title='2b')

# Put the dataframes in columns a b c respectively

# gs_edit_cells(gepURL,ws=2,input=D,anchor='A1')

# gs_edit_cells(gepURL,ws=2,input=I,anchor='B1')

# gs_edit_cells(gepURL,ws=2,input=R,anchor='C1')


ageParty <- gs_read(gepURL, ws = 2)  ## Read the data back into R
cn <- c("D", "I", "R")
colnames(ageParty) <- cn
ageParty <- ageParty %>% gather(key = "Party", value = "Age") %>% filter(is.na(Age) ==
    F)
aPaov <- aov(Age ~ Party, data = ageParty)
summary(aPaov)
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## Party         2    767   383.5     4.4 0.0126 *
## Residuals   789  68754    87.1
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

2b - Possible causes for discrepancy

```r
# Determine possible causes for the discrepancy between the two tests
De <- ageParty %>% filter(Party == "D")
Ie <- ageParty %>% filter(Party == "I")
Re <- ageParty %>% filter(Party == "R")
c(mean(De$Age), mean(Ie$Age), mean(Re$Age))
```

```
## [1] 43.06208 45.33170 44.83149
```

```r
c(mean(D[, 1]), mean(I[, 1]), mean(R[, 1]))  #experimental means
```

```
## [1] 43.59896 44.45389 44.54748
```

```r
mu  #means provided
```

```
## [1] 43.3 44.6 45.1
```

```r
c(sd(D[, 1]), sd(I[, 1]), sd(R[, 1]))  #experimental sd
```

```
## [1] 8.770424 9.994521 9.594601
```

```r
sd  #sd provided
```

```
## [1] 9.1 9.2 9.2
```