HW6.Zunqiu.Wang

Zunqiu Wang

10/8/2021

```r
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

Q1 a
H_0 =Age is independent of party
H_a =Age is dependent of party

$$total = 86 + 52 + 61 + 72 + 51 + 74 + 73 + 55 + 70 + 71 + 54 + 73 = 792$$

$$total_D = 86 + 72 + 73 + 71 = 302$$

$$total_I = 52 + 51 + 55 + 54 = 212$$

$$total_R = 61 + 74 + 70 + 73 = 278$$

18_29= $86 + 52 + 61 = 199$
30_44= $72 + 51 + 74 = 197$
45_59= $73 + 55 + 70 = 198$
60_= $71 + 54 + 73 = 198$

f_e(18_29 & D)= $total_D * 18\_29/total = 75.88$
f_e(30_44 & D)= $total_D * 30\_44/total = 75.11$
f_e(45_59 & D)= $total_D * 45\_59/total = 75.5$
f_e(60_ & D)= $total_D * 60\_/total = 75.5$

f_e(18_29 & I)= $total_I * 18\_29/total = 53.26$
f_e(30_44 & I)= $total_I * 30\_44/total = 52.73$
f_e(45_59 & I)= $total_I * 45\_59/total = 53$
f_e(60_ & I)= $total_I * 60\_/total = 53$

f_e(18_29 & R)= $total_R * 18\_29/total = 69.85$
f_e(18_29 & R)= $total_R * 18\_29/total = 69.15$
f_e(18_29 & R)= $total_R * 18\_29/total = 69.5$
f_e(18_29 & R)= $total_R * 18\_29/total = 69.5$

$\chi^2 = \Sigma \frac{(f_o - f_e)^2}{f_e}$ $\chi^2 = \frac{(86-75.88)^2}{75.88} + \frac{(72-75.11)^2}{75.11} + \frac{(73-75.5)^2}{75.5} + \frac{(71-75.5)^2}{75.5} + \frac{(52-53.26)^2}{53.26} + \frac{(51-52.73)^2}{52.73} + \frac{(55-53)^2}{53} + \frac{(54-53)^2}{53} + \frac{(61-69.85)^2}{69.85} + \frac{(74-69.15)^2}{69.15} + \frac{(70-69.5)^2}{69.5} + \frac{(73-69.5)^2}{69.5} = 1.35 + 0.13 + 0.083 + 0.27 + 0.029 + 0.057 + 0.075 + 0.02 + 1.12 + 0.34 + 0.036 + 0.18 = 3.69$ $df = (r-1)(c-1) = 2 * 3 = 6$

```r
#chi critic
qchisq(.95, df=6)
```

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

Since $3.69 < 12.59159$, we fail to reject null hypothesis.

```r
#p val
1-pchisq(3.69,6)
```

```
## [1] 0.7185431
```

```r
# create a df storing all info
df <- data.frame(age_18_29=c(86,52,61), age_30_44=c(72,51,74),
                 age_45_59=c(73,55,70), age_60=c(71,54,73))
rownames(df) <- c("D", "I", "R")
# conduct Chisq test
chisq.test(df)
```

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

```r
# p val > 0.05, so fail to reject null hypothesis.
```

Q2 a
$H_0 = \mu_D = \mu_I = \mu_R$
H_a =at least one is different

F-stat=$\frac{\text{average variance between groups}}{\text{average variance within groups}}$
N=total number of observations
G=Groups

Between variance $= \frac{n_1(\bar{y}_1 - \bar{y})^2 + \ldots + n_G(\bar{y}_G - \bar{y})^2}{G-1}$

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

$df_1 = G - 1$ $df_2 = N - G$

D: (43.3, 9.1, 302), I: (44.6, 9.2, 212), R: (45.1, 9.2, 278)
$\bar{y} = 44.2$ $N = 792, G = 3$

Between variance $= \frac{302(43.3-44.2)^2+212(44.6-44.2)^2+278(45.1-44.2)^2}{3-1} = 251.86$
Within variance $= \frac{(302-1)9.1^2+(212-1)9.2^2+(278-1)9.2^2}{792-3} = 83.94$

F-stat$= \frac{251.86}{83.94} = 3.00048$

```
#create a function to calculate F stat, p val based on formula
y <- c(43.3, 44.6, 45.1)
s <- c(9.1, 9.2, 9.2)
n <- c(302, 212, 278)
mu <- 44.2
anov <- function(y, mu, s, n) {
  bvec <- numeric()
  wvec <- numeric()
  for (i in 1:length(n)) {
    bvec[i] <- n[i] * (y[i]-mu)^2
    wvec[i] <- (n[i]-1) * s[i]^2
  }
  BV <- sum(bvec)/(length(n) - 1)
  WV <- sum(wvec)/(sum(n) - length(n))
  fstat <- BV/WV
  pval <- 1 - pf(fstat, length(n) - 1, (sum(n)-length(n)))
  df <- data.frame(param = c("f_stat", "p_val"), stats = c(fstat, pval))
  return(df)
}
anov(y, mu, s, n)
```

```
##    param      stats
## 1 f_stat 3.00040993
## 2  p_val 0.05033486
```

```
# F crit
qf(0.95, 2, 789) # 3.0004 is slightly smaller than 3.007
```

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

```
# p val
1-pf(3.00048, 2, 789)
```

```
## [1] 0.05033136
```

b

```
# simulate normal distribution of age and construct a df
set.seed(1234)
D <- cbind(rnorm(n[1], y[1], s[1]), "Democrat")
I <- cbind(rnorm(n[2], y[2], s[2]), "Independent")
R <- cbind(rnorm(n[3], y[3], s[3]), "Republican")
df <- rbind(D, I, R)
colnames(df) <- c("age", "party")
df <- as.data.frame(df)
df$age <- as.numeric(df$age)
df$party <- as.factor(df$party)
head(df)
```

```
##        age    party
## 1 32.31570 Democrat
## 2 45.82461 Democrat
## 3 53.16841 Democrat
## 4 21.95415 Democrat
## 5 47.20503 Democrat
## 6 47.90511 Democrat
```

```
# conduct F test
aov.test <- aov(df$age ~ df$party)
summary(aov.test)
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## df$party      2    231  115.40   1.365  0.256
## Residuals   789  66698   84.53
```

```
# compare F test results using simulated distribution with original
# case by calculating descriptive stats
Dem <- df %>% filter(party == "Democrat")
Ind <- df %>% filter(party == "Independent")
Rep <- df %>% filter(party == "Republican")

mean(Dem$age)
```

```
## [1] 43.37157
```

```
sd(Dem$age)
```

```
## [1] 9.128099
```

```
mean(Ind$age)
```

```
## [1] 44.5512
```

```
sd(Ind$age) # this simulated sd differs a lot from 2a provided
```

```
## [1] 9.878794
```

```
mean(Rep$age) # differs by age = 1
```

```
## [1] 44.42253
```

```
sd(Rep$age) # this simulated sd differs a lot from 2a provided
```

```
## [1] 8.712471
```

```
## They are different. resulted simulated descriptive stats will differ from the given stats to
# perform simulation thus summary stats of F stat and p val will also differ.
```