ST 411/511 Homework 6

Chimdi Chikezie

Winter 2022

Instructions

This assignment is due by 11:59 PM, February 25th on Canvas via Gradescope. **You should submit your assignment as a PDF which you can compile using the provide .Rmd (R Markdown) template.**

> Note: Create a PDF by either compiling a PDF directly (via LaTeX) or from a Word Doc. Do not submit a PDF of the HTML output as they're cumbersome and difficult to read.

Include your code in your solutions and indicate where the solutions for individual problems are located when uploading into Gradescope (Failure to do so will result in point deductions). You should also write complete, grammatically correct sentences for your solutions.

Once you've completed the assignment, and before you submit it to Gradescope, you should read the document to ensure that (1) The computed values show up in the document, (2) the document "looks nice" (i.e. doesn't have extraneous code/outputs and includes *just* the essentials), and (3) ensure the document is "easy" to read.

Goals:

1. Practice hypothesis testing for equality of "scales" using Levene's test.
2. Practice computing the different "pieces" we need to compute to conduct an ANOVA test.
3. Demonstrate the connections between ANOVA as a test of population means and as a method for model comparison.

Question 1 (5 points)

(a) (2 points) Generate two samples using the `rnorm()` function. Combine the two samples into one vector, and create another vector that indicates which group the observations belong to (See Lab 6 for help with creating the data for this question). The two samples should be drawn as follows:

- Sample A: $m = 10$ observations from a Normal($\mu = 0, \sigma^2 = 1$) distribution.
- Sample B: $n = 20$ observations from a Normal($\mu = 0, \sigma^2 = 4$) distribution. Note: You might also consider using the `set.seed()` function.

```
m <- 10
n <- 20
set.seed(234334)
samp1 <- rnorm(m, mean=0, sd=1)
samp1
```

```
##  [1] -1.3994052  0.8879045 -1.4424399 -0.3276531  0.6894410 -0.6553014
##  [7]  0.1042037 -0.7610561 -0.8832022  1.7402062
```

```
samp2 <- rnorm(n, mean=0, sd=2)
samp2
```

```
##  [1] -2.03435987 -1.53543331  1.59399903 -0.54327877 -1.72111795  1.26079126
##  [7] -1.42053213  0.28459513 -0.04160066 -0.67949654  1.41527979  3.04972928
## [13] -0.72441231  2.54065693 -0.23623323  0.04339000 -1.60908039  0.32903115
## [19]  0.38324414  0.93304408
```

```
sampComb <- c(samp1, samp2)
sampGrp <- as.factor(rep(c(1,2), c(m,n)))
df <- data.frame(sampComb, sampGrp)
```

(b) (3 points) Perform Levene's test in R using the `leveneTest()` function in the `car` library. Note: you will need to install/load the `car` package using `install.packages("car")` [Run the `install.packages()` function in the console and not the R Markdown] `library(car)` [Include this command in the R Markdown]. Report the resulting p-value and summarize your findings (state the hypothesis tested, the results of your analysis, and your conclusions).

```
leveneTest(sampComb, group=sampGrp)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value Pr(>F)
## group  1  0.9407 0.3404
##       28
```

```
dists1 <- abs(samp1 - median(samp1))
dists2 <- abs(samp2 - median(samp2))
t.test(dists1, dists2, var.equal=TRUE)
```

```
##
##  Two Sample t-test
##
## data:  dists1 and dists2
## t = -0.9699, df = 28, p-value = 0.3404
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.9193229  0.3284937
## sample estimates:
## mean of x mean of y
## 0.8235507 1.1189653
```

$H_0 : Sx^2 = Sy^2$ $H_A : Sx^2 \neq Sy^2$

I fail to reject the null hypothesis at 5% significance level for the alternative hypothesis because the p-value is greater than alpha. We do not have enough evidence to show that the spread between samp1 and samp2 are not equal.

Question 2 (7 points)

The table below shows a partially completed ANOVA table. (Note: if you are looking at this in RStudio it may be helpful to knit the file to properly view the table.)

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F-statistic	p-value
Between Groups	35819	7	5117	3.5	0.00994
Within Groups	35088	24	1462		
Total	70907	31			

(a) (1 point) How many groups were there?

There are 8 groups.

(b) (4 points) Fill in the rest of the table. Values to be calculated are indicated by a "." Please show how you compute the values for your calculations.

```
SSW <- 35088
SST <- 70907
dfW <- 24
dfT <- 31

SSB <- SST  - SSW
sprintf("Sum of Squares Between Groups: %f", SSB)
```

```
## [1] "Sum of Squares Between Groups: 35819.000000"
```

```
dfB <- dfT  - dfW
sprintf("Degrees of Freedom Between Groups: %f", dfB)
```

```
## [1] "Degrees of Freedom Between Groups: 7.000000"
```

```
MSB <- SSB/dfB
sprintf("Mean Square Between Groups: %f", MSB)
```

```
## [1] "Mean Square Between Groups: 5117.000000"
```

```
MSW <- SSW/dfW
sprintf("Mean Square Within Groups: %f", MSW)
```

```
## [1] "Mean Square Within Groups: 1462.000000"
```

```
MST <- MSB + MSW
sprintf("Mean Square Total: %f", MST)
```

```
## [1] "Mean Square Total: 6579.000000"
```

```
Fstat <- MSB/MSW
sprintf("F_state: %f", Fstat)
```

```
## [1] "F_state: 3.500000"
```

```
pvalue <- 1 - pf(Fstat, df1 = dfB, df2 = dfW)
sprintf("P-value: %f", pvalue)
```

```
## [1] "P-value: 0.009942"
```

(c) (2 points) What is your conclusion from the one-way ANOVA analysis? State the hypothesis you are testing and what your decision/strength of evidence are.

Reject the null hypothesis at 5% significance level. We do have enough evidence to show that there are at least two population means that are different.

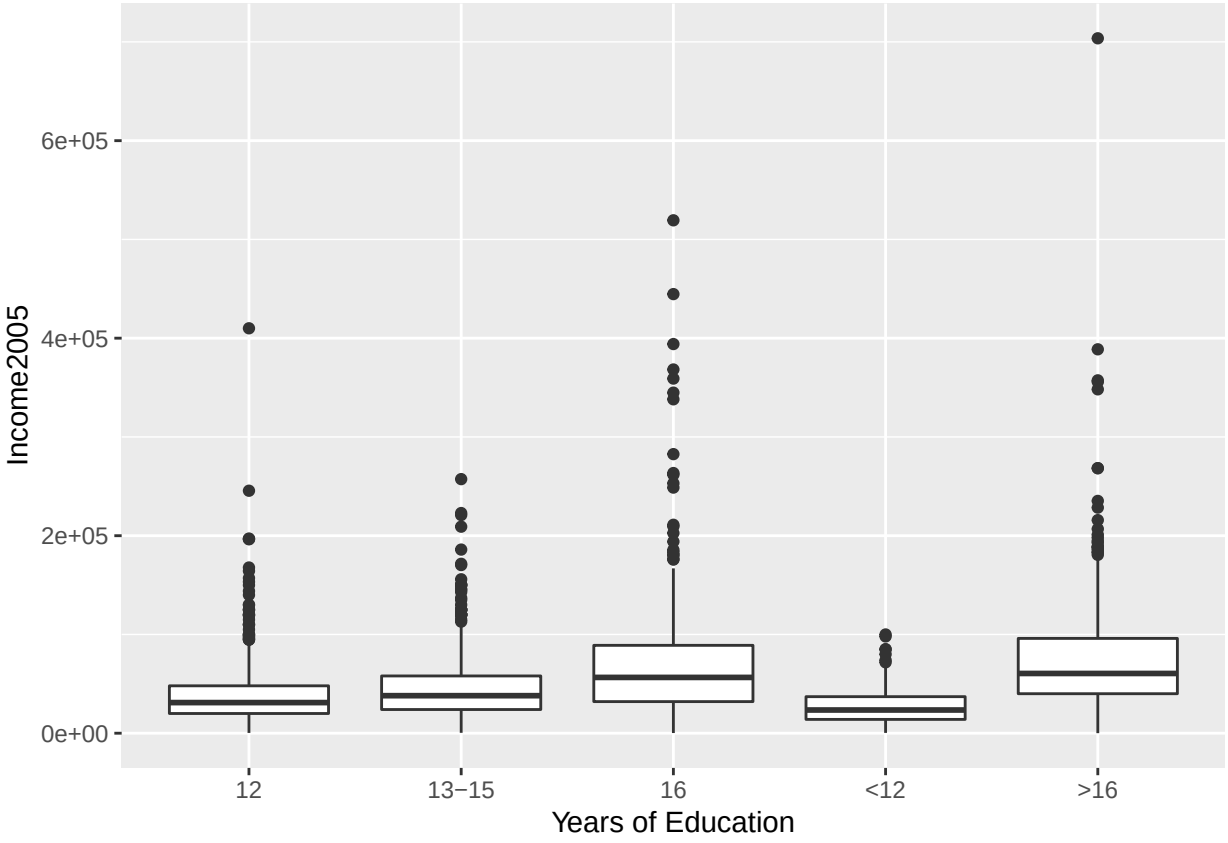
Question 3 (9 points) - Modified from *Sleuth* 5.25

The data file `ex0525` contains annual incomes in 2005 of a random sample of 2,584 Americans who were selected for the National Longitudinal Survey of Youth in 1979 and who had paying jobs in 2005. The data set also includes a code for the number of years of education that each individual had completed by 2006: <12, 12, 13-15, 16, and >16. Perform an analysis of variance *by hand* (i.e. not using the built-in anova functions like `lm()` and `anova()`) to assess whether or not the population mean 2005 incomes were the same in all five education groups. Work through the following steps:

(a) (1 point) Create a side-by-side boxplot of 2005 income grouped by education category.

```
library(tidyverse)
```

```
data(ex0525)
ggplot(ex0525, aes(x=as.factor(Educ), y=Income2005)) + geom_boxplot() +
  xlab("Years of Education")
```



(b) (2 points) Find the grand mean and the mean of each of the five education groups.

```
grandmean <- mean(ex0525$Income2005)
sprintf("grand mean: %f", grandmean)
```

```
## [1] "grand mean: 49416.998839"
```

```r
levels(ex0525$Educ)
```

```
## [1] "12"    "13-15" "16"    "<12"   ">16"
```

```r
sprintf("mean of group <12: %f", mean(ex0525$Income2005[ex0525$Educ == "<12"]))
```

```
## [1] "mean of group <12: 28301.448529"
```

```r
sprintf("mean of group 12: %f", mean(ex0525$Income2005[ex0525$Educ == "12"]))
```

```
## [1] "mean of group 12: 36864.896078"
```

```r
sprintf("mean of group 13-15: %f", mean(ex0525$Income2005[ex0525$Educ == "13-15"]))
```

```
## [1] "mean of group 13-15: 44875.956790"
```

```r
sprintf("mean of group 16: %f", mean(ex0525$Income2005[ex0525$Educ == "16"]))
```

```
## [1] "mean of group 16: 69996.972906"
```

```r
sprintf("mean of group >16: %f", mean(ex0525$Income2005[ex0525$Educ == ">16"]))
```

```
## [1] "mean of group >16: 76855.462567"
```

(c) (2 points) Find the sums of squares between and within groups.

```r
ex0525 %>%
  group_by(Educ) %>%
  summarise(SW = sum((Income2005 - mean(Income2005))^2)) %>%
  summarise(SSW = sum(SW))
```

```
## # A tibble: 1 x 1
##       SSW
##     <dbl>
## 1 4.95e12
```

```r
sprintf("SSW: %f", SSW)
```

```
## [1] "SSW: 35088.000000"
```

```r
sprintf("SST: %f", ex0525 %>%
  summarise(SST = sum((Income2005 - mean(Income2005))^2)))
```

```
## [1] "SST: 5639977858618.996094"
```

```r
SSW <- 4.951743e+12
SST <- 5.639978e+12
SSB <- SST - SSW

sprintf("SSB: %f", SSB)
```

```
## [1] "SSB: 688235000000.000000"
```

(d) (1 point) Find the mean squares between and within groups.

```r
n <- nrow(ex0525)
I <- length(unique(ex0525$Educ))
dfW <- n - I
dfT <- n - 1
dfB <- I - 1
MSB <- SSB/dfB
MSW <- SSW/dfW

sprintf("MSB: %f", MSB)
```

```
## [1] "MSB: 172058750000.000000"
```

```r
sprintf("MSW: %f", MSW)
```

```
## [1] "MSW: 1920024428.072896"
```

(e) (1 point) Find the F-statistic and p-value.

```r
F_stat <- MSB/MSW
sprintf("F_stat: %f", F_stat)
```

```
## [1] "F_stat: 89.612792"
```

```r
pvalue <- 1 - pf(F_stat, df1 = dfB, df2 = dfW)
sprintf("P_value: %f", pvalue)
```

```
## [1] "P_value: 0.000000"
```

(f) (1 point) State the conclusion of your test.

```r
summary(aov(Income2005 ~ Educ, data = ex0525))
```

```
##               Df    Sum Sq   Mean Sq F value Pr(>F)
## Educ           4 6.882e+11 1.721e+11   89.61 <2e-16 ***
## Residuals   2579 4.952e+12 1.920e+09
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
anova(lm(Income2005 ~ Educ, data=ex0525))
```

```
## Analysis of Variance Table
##
## Response: Income2005
##             Df     Sum Sq    Mean Sq F value    Pr(>F)
## Educ         4 6.8824e+11 1.7206e+11  89.613 < 2.2e-16 ***
## Residuals 2579 4.9517e+12 1.9200e+09
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Reject the null hypothesis for the alternative at 5% significance level. We have enough evidence to show that at least 2 population mean are different.

(g) (1 point) We can also state things we have calculated in the "model testing/comparison" framework (You should not need to calculate anything new for this part). What is the extra sum of squares? What is the pooled variance?

Extra sum of squares(SSB): 688235000000.000000 Pooled Variance(MSW): 1920024428.072896