ST 411/511 Homework 5

E. ALex Soderquist

Summer 2022

Instructions

This assignment is due by 11:59 PM, August 9th 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 computing the different "pieces" we need to compute to conduct an ANOVA test.
2. Demonstrate the connections between ANOVA as a test of population means and as a method for model comparison.
3. Get hands on experience conducting ANOVA and Kruskal-Wallis tests.
4. Practice using linear combinations of means to answer scientific questions of interest.
5. Learn how to perform different methods of multiple comparison procedures with R.

Question 1 (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	0.009941
Within Groups	35088	24	1462		
Total	70907	31			

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

There are "I" groups, where I-1 is the between degrees of freedom. Therefore we can find this value using the degrees of freedom column: $(n-1)-(n-I) = 31-24 = 7 \iff I-1 = 7$. So 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.

Order of code: SSB = SST - SSW MSB = SSB/(I-1) MSW = SSW/(n-I) F-stat = MSB/MSW p-value = 1-F(df1=7,df2=24) at F = MSB/MSW = 5117/1462

```
70907 - 35088
```

```
## [1] 35819
```

```
35819/7
```

```
## [1] 5117
```

```
35088/24
```

```
## [1] 1462
```

```
5117/1462
```

```
## [1] 3.5
```

```
1-pf(5117/1462,7,24)
```

```
## [1] 0.009941808
```

(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.

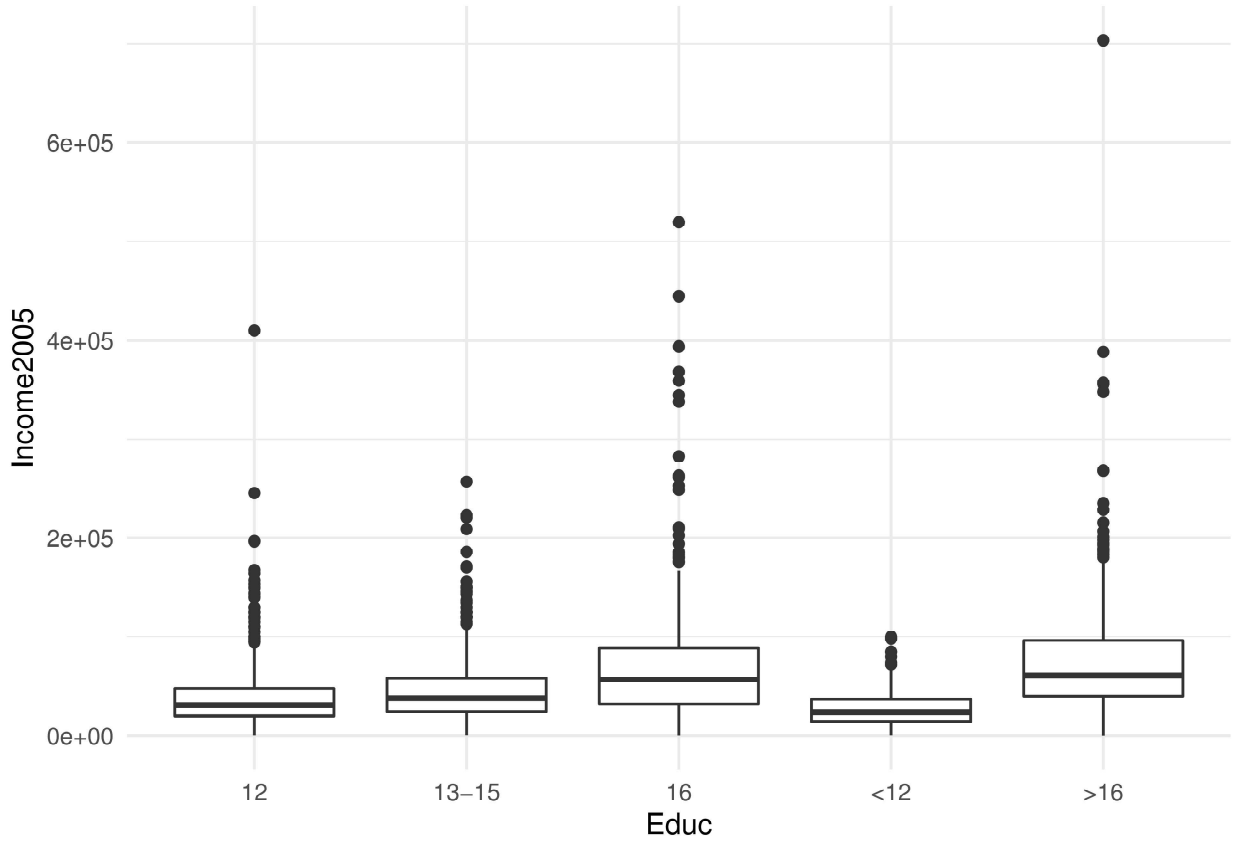
The hypotheses tests: $H_0 : \mu_1 = \mu_2 = ... = \mu_8$. H_A: At least two different population means are different ($u_i \neq u_j$ for some $i \neq j$ and $i, j \in \{1, 2, ..., 8\}$). Our test statistic F = 3.5 and our associated p-value of the F-distribution comes out to be 0.009942. At a significance level as small as $\alpha = 0.01$, we reject the null hypothesis in favor of the alternative hypothesis. That is, we have enough statistical evidence to say at least two of the population means of the groups are different.

Question 2 (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.

```
data(ex0525)
df <- data.frame(ex0525)
ggplot(data=df, aes(x=Educ,y=Income2005))+geom_boxplot() +theme_minimal()
```



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

```
n <- 333
grand <- (1/n)*sum(ex0525$Income2005)
n1 <- sum(ex0525$Educ == "<12")
aggregate(df$Income2005,list(df$Educ),FUN=sum)
```

```
##   Group.1        x
## 1      12 37602194
## 2   13-15 29079620
## 3      16 28418771
## 4     <12  3848997
## 5     >16 28743943
```

```
s1 <- (1/n1)*3848997
n2 <- sum(ex0525$Educ == "12")
s2 <- (1/n2)*37602194
n3 <- sum(ex0525$Educ == "13-15")
s3 <- (1/n3)*29079620
n4 <- sum(ex0525$Educ == "16")
s4 <- (1/n4)*28418771
n5 <- sum(ex0525$Educ == ">16")
s5 <- (1/n5)*28743943
n1
```

```
## [1] 136
```

```
s1
```

```
## [1] 28301.45
```

```
n2
```

```
## [1] 1020
```

```
s2
```

```
## [1] 36864.9
```

```
n3
```

```
## [1] 648
```

```
s3
```

```
## [1] 44875.96
```

```
n4
```

```
## [1] 406
```

```
s4
```

```
## [1] 69996.97
```

```
n5
```

```
## [1] 374
```

```
s5
```

```
## [1] 76855.46
```

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

```
IDless <- which(ex0525$Educ == "<12")
ID12 <- which(ex0525$Educ == "12")
ID13 <- which(ex0525$Educ == "13-15")
ID16 <- which(ex0525$Educ == "16")
IDover <- which(ex0525$Educ == ">16")

groupmeans <- c(mean(ex0525$Income2005[IDless]),
                mean(ex0525$Income2005[ID12]),
                mean(ex0525$Income2005[ID13]),
                mean(ex0525$Income2005[ID16]),
                mean(ex0525$Income2005[IDover]))

SSW <- sum((ex0525$Income2005[IDless]-groupmeans[1])^2)+
  sum((ex0525$Income2005[ID12]-groupmeans[2])^2)+
  sum((ex0525$Income2005[ID13]-groupmeans[3])^2)+
  sum((ex0525$Income2005[ID16]-groupmeans[4])^2)+
  sum((ex0525$Income2005[IDover]-groupmeans[5])^2)
SST <- sum((ex0525$Income2005 - grand)^2)

SSW
```

```
## [1] 4.951743e+12
```

```
SST
```

```
## [1] 2.939819e+14
```

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

```
MSB <- (SST-SSW)/4
MSW <- SSW/(2584-4)
MSB
```

```
## [1] 7.225754e+13
```

```
MSW
```

```
## [1] 1919280124
```

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

```
F <- MSB/MSW
F
```

```
## [1] 37648.25
```

```
p <- 1-pf(F, df1=4,df2=2580)
p
```

```
## [1] 0
```

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

Our hypotheses tests are: $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$, H_A :At least two of our group population means are different, that is, at least two different groups of years of education had different mean annual incomes in 2005. We reject the null hypothesis that they are all equal in this case with a p-value of 0 (does this mean the actual p-value is less than machine epsilon, i.e. very small?) and an associated F-statistic of 37,648.25 USD. Therefore, there is a noticeable difference in the variability of our group distributions.

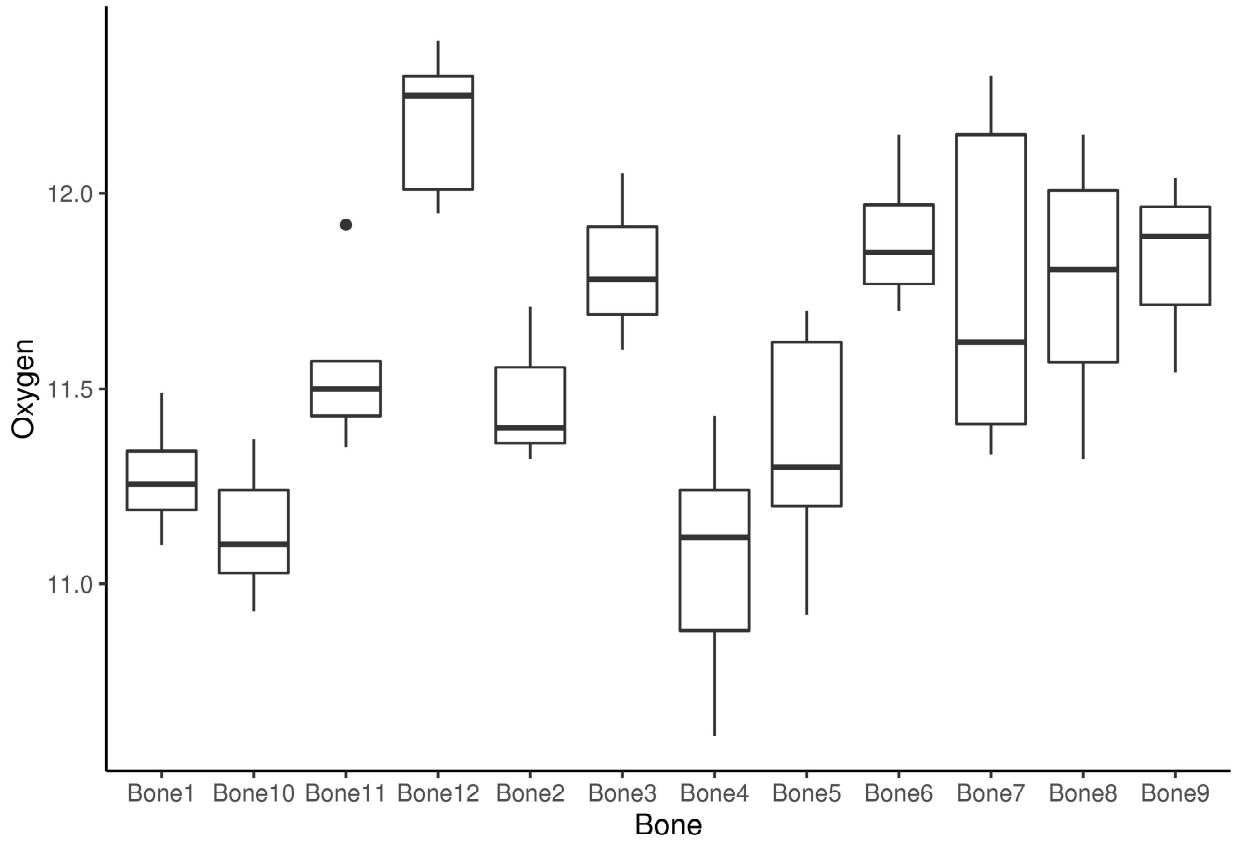
(g) (1 point) We can also state things we have calculated in the "model testing/comparison" framework (You should not need to calculate anything new for this part). What is the extra sum of squares? What is the pooled variance?

Question 3 (9 points) - from *Sleuth* 5.23

Was Tyrannosaurus Rex warm-blooded? Several measurements of the oxygen isotopic composition of bone phosphate in each of 12 bone specimens from a single *Tyrannosaurus rex* skeleton were taken. It is known that the oxygen isotopic composition of vertebrate bone phosphate is related to the body temperature at which the bone forms. Differences in means at different bone sites would indicate non-constant temperatures throughout the body. Minor temperature differences would be expected in warm-blooded animals. Is there evidence that the means are different for the different bones? The data are in `ex0523` in the `Sleuth3` library.

(a) (2 points) Plot the oxygen isotopic composition for each of the bones using a side-by-side boxplot. Comment on whether or not you think the population means are the same for all 12 bones based on your plot.

```
data(ex0523)
df <- data.frame(ex0523)
ggplot(df, aes(x=Bone,y=Oxygen))+geom_boxplot()+theme_classic()
```



I think it seems fairly obvious the group mean is different for at least two of these bones. Consider Bone 10 and Bone 12 for one of the more extreme cases. There is no overlap of the IQRs, and the difference of means is over 1 unit in terms of their oxygen isotopic composition.

(b) (2 points) Perform an analysis of variance to test whether or not all the population mean oxygen isotopic compositions are the same in the 12 bone types. State your p-value and conclusion of the test. You may use the built-in ANOVA functions in R.

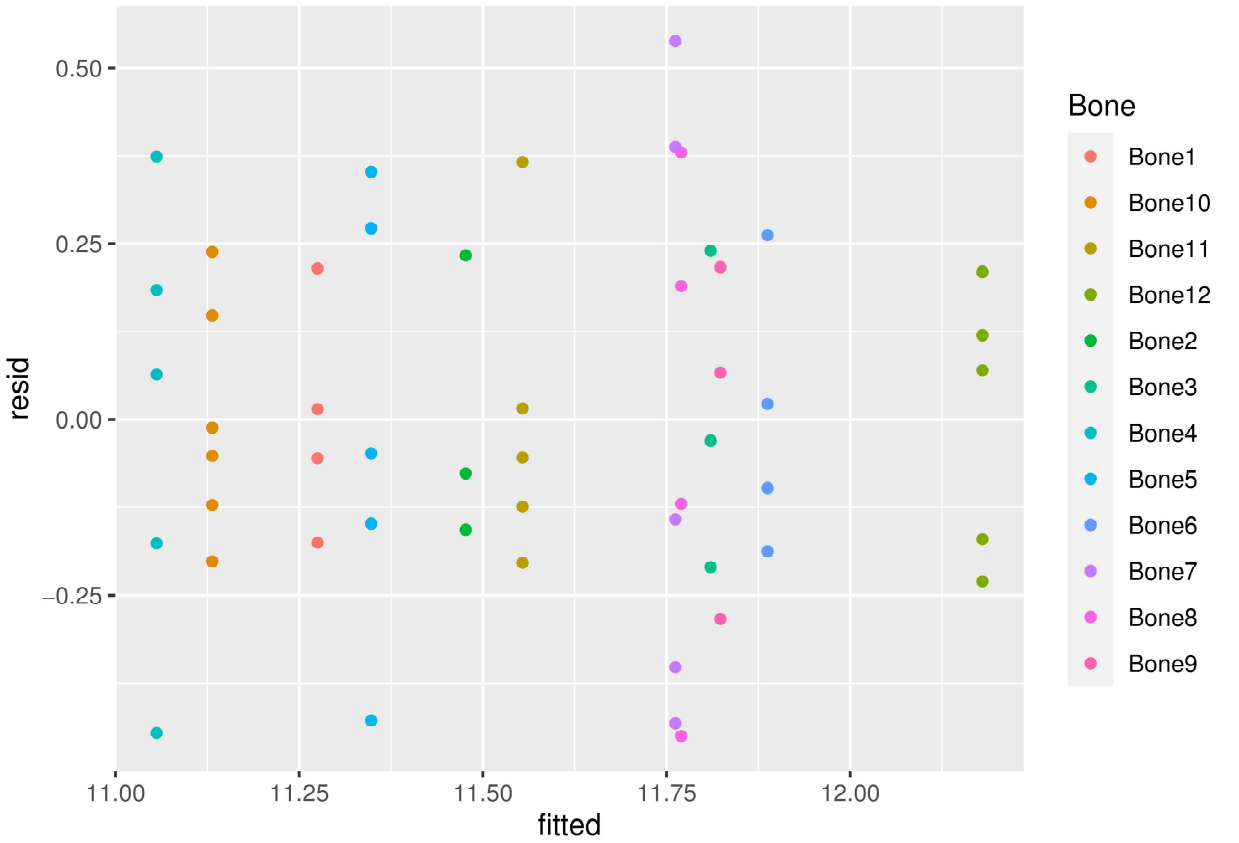
```
anova(lm(Oxygen~Bone,data=df))
```

```
## Analysis of Variance Table
##
## Response: Oxygen
##           Df Sum Sq Mean Sq F value   Pr(>F)
## Bone      11 6.0675 0.55159  7.4268 9.73e-07 ***
## Residuals 40 2.9708 0.07427
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Hypothesis tests: $H_0 : \mu_1 = \mu_2 = ... = \mu_{12}$ and H_A : At least two of the bone group means are different. Our F statistic has a value of 7.427 with an associated p-value of 9.73e-07 (very small), and so we have enough statistical evidence to reject the null hypothesis and say that at least two of the bone group means appear different. Therefore, there are different amounts of variability between at least two of the group distributions.

(c) (2 points) Assess the assumption that the population variances are the same in each group by creating a diagnostic plot using the residuals (See Lecture 18 for help with this). Does this assumption appear to have been met?

```
mod <- lm(Oxygen~Bone,data=df)
ex0523$fitted <- mod$fitted
ex0523$resid <- mod$resid
ggplot(ex0523,aes(x=fitted,y=resid,color=Bone))+geom_point()
```



The equal population variance assumption seems to be met; There are no extreme or disproportionate outliers, the variance appears relatively consistent and constant across all 12 bone groups, and there are no trends or 'shapes' which may indicate some kind of pattern which breaks our assumption.

(d) (3 points) Perform a Kruskal-Wallis test using the `kruskal.test()` function. What do you conclude from this test? Compare your conclusion with your result from the analysis of variance in part (b).

```
kruskal.test(Oxygen~Bone,data=df)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  Oxygen by Bone
## Kruskal-Wallis chi-squared = 34.938, df = 11, p-value = 0.0002537
```

Hypotheses tests: H_0 : The group 'centers' are equal. H_A : At least two of the group centers are different. Our test statistic comes out to be the Chi-Squared value of 34.938 with 11 degrees of freedom, giving us an associated p-value of 0.0002537. Because this p-value is so small, at a significance level of as small as $\alpha = 0.01$ we reject the null hypothesis. Meaning, the variability of the group distributions is "different" enough that at least two of the bone group centers have different centers.

It appears both tests give us the same sort of conclusion, only Kruskal-Wallis in terms of 'center' more abstractly than the group means discussed in part b). I see no reason either test can't be used for this

problem, however we have all the assumptions for our usual ANOVA test which gives us more accuracy and precision.

Question 4 (4 points)

(a) (2 points) In comparing 10 groups, a researcher notices that the sample mean of group 7 is the largest and the sample mean of group 3 is the smallest. The researcher then decides to test the hypothesis that $\mu_7 - \mu_3 = 0$. Why should a multiple comparison procedure be used even though there is only one comparison being made?

The key to multiple comparisons is that even though only one comparison is being made, by re-testing with a different hypothesis you are increasingly likely to make a Type I Error in the study. (The tests are related and lack total independence!) Therefore you should adjust for this possibility by using a multiple comparison procedure.

b) (2 points) When choosing coefficients for a contrast, does the choice of $\{C_1, C_2, \ldots, C_I\}$ give a different t-statistic than the choice of $\{4C_1, 4C_2, \ldots, 4C_I\}$? Explain why or why not.

No, there is no difference in the t-statistic if the coefficients are all changed to be the same (positive?) value, or scaled/multiplied by the same constant if each constant is different in the question of interest. Because the parameter is simply a measurement of interpreting how each group relates to the others in question via use of constants, it would not give a different t-statistic.