ST 411/511 Homework 6

Due on February 26

Aashish Adhikari

Winter 2020

Instructions

This assignment is due by 11:59 PM, February 26, 2020 on Canvas via Gradescope. **You should submit your assignment as a PDF which you can compile (should you choose – recommended) using the provide .Rmd (R Markdown) template.** If you opt to not use R Markdown, please format your solutions in a similar manner as provided in this document. Include your code in your solutions and indicate where the solutions for individual problems are located when uploading into Gradescope. You should also use complete, gramatically correct sentences for your solutions.

Problems (25 points total)

Question 1

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	0.0099
Within Groups	35088	24	1462		
Total	70907	31			

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

Given that

n-I=24 n-1=31

Solving gives us I = 8. Hence, the number of groups is 8.

(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.

SSB = SST - SSW = 70907 - 35088 = 35819

Between degrees of freedom = Total degrees of freedom - Within degrees of freedom = 31 - 24 = 7

Between Groups Mean Square = SSB / l-1 = 35819 / 7 = 5117

Within Groups Mean Square = SSW / n- l = 35088 / 24 = 1462

F-statistic = MSB / MSW = 5117 / 1462 = 3.5

```
pval=1-pf(3.5,7,24)
pval
```

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

The p-value is 0.00994.

(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.

H0: The means of all the populations are equal. H1: At least two population means are different.

Say, significance level alpha = 0.05 We reject the null hypothesis in favor of alternative hypothesis that there are at least two populations that have different population means.

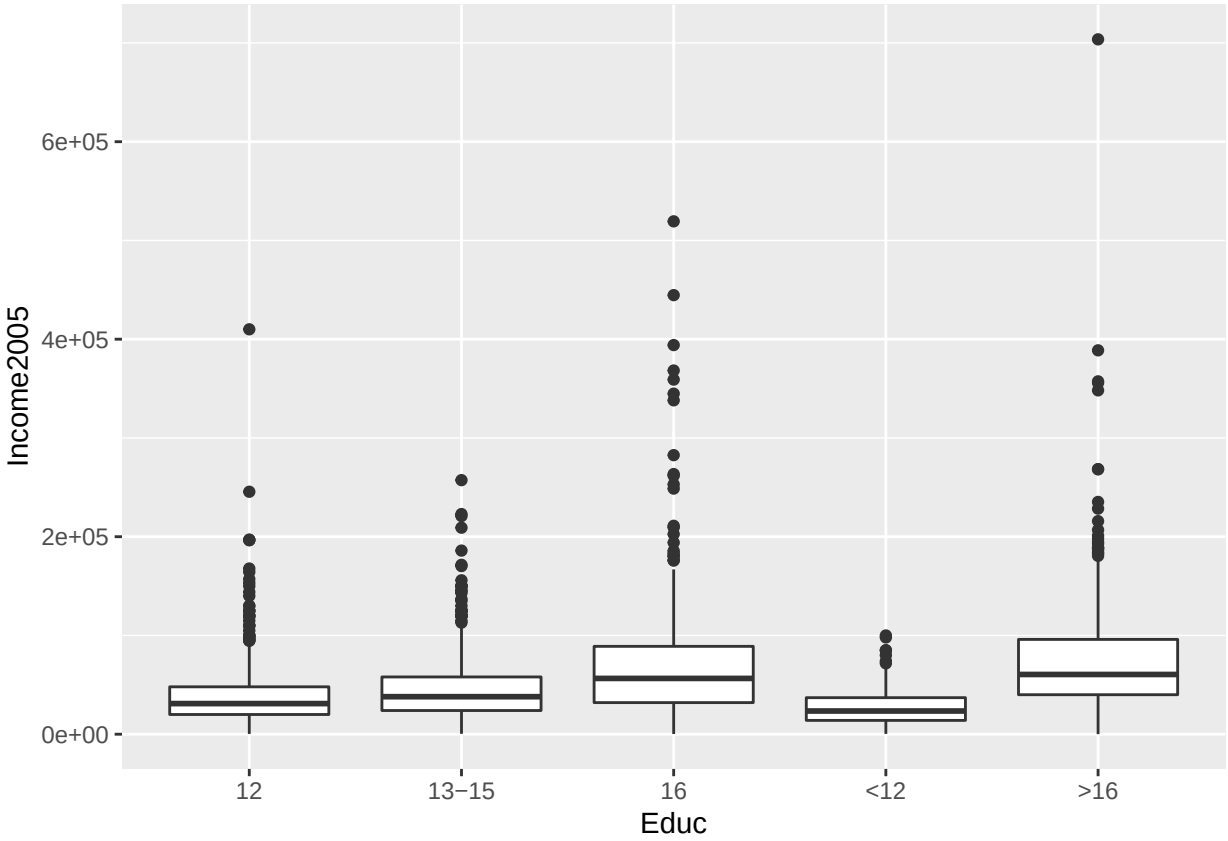
Question 2 (Modified from *Sleuth* 5.25)

The data file `ex0525` contains annual incomes in 2005 of a random sample of 2584 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)
```

```
ggplot(data = ex0525, aes(x=Educ, y=Income2005))+geom_boxplot()
```



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

```
grandmean=mean(ex0525$Income2005)
grandmean
```

```
## [1] 49417
```

```
edu12=which(ex0525$Educ=="12")
edu_13_to_15=which(ex0525$Educ=="13-15")
edu16=which(ex0525$Educ=="16")
edu_l_t_12=which(ex0525$Educ=="<12")
edu_g_t_16=which(ex0525$Educ==">16")

m1=mean(ex0525$Income2005[edu12])
m2=mean(ex0525$Income2005[edu_13_to_15])
m3=mean(ex0525$Income2005[edu16])
m4=mean(ex0525$Income2005[edu_l_t_12])
m5=mean(ex0525$Income2005[edu_g_t_16])

m1
```

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

```
m2
```

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

```
m3
```

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

```
m4
```

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

```
m5
```

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

The grand mean is 49417.

The mean of the education group 12 is 36864.9.

The mean of the education group 13 to 15 is 44875.96.

The mean of the education group 16 is 69996.97.

The mean of the education group less than 12 is 28301.45.

The mean of the education group greater than 16 is 76855.46.

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

```
SSW<-sum((ex0525$Income2005[edu12]-m1)^2)+  sum((ex0525$Income2005[edu_13_to_15]-m2)^2)+
  sum((ex0525$Income2005[edu16]-m3)^2)+  sum((ex0525$Income2005[edu_l_t_12]-m4)^2)+
  sum((ex0525$Income2005[edu_g_t_16]-m5)^2)



SST = sum((ex0525$Income2005-grandmean)^2)


SSB=SST-SSW

SSW
```

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

```
SSB
```

```
## [1] 688235137516
```

The sums of squares between groups is 688235137516. The sums of squares within groups is 4.951743e+12.

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

```
I = 5
n = nrow(ex0525)

dfW = n-I
dfB = I-1

MSW = SSW/dfW


MSB = SSB/dfB

MSW
```

```
## [1] 1920024320
```

```
MSB
```

```
## [1] 172058784379
```

The mean squares between groups is: 172058784379. The mean squares within groups is: 1920024320.

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

```
f_val=MSB/MSW

pval=(1-pf(f_val,df1=dfB,df2=dfW))

f_val
```

```
## [1] 89.61282
```

```
pval
```

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

The f-statistic is 89.61282. The p-value is 0.

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

Here, the p value is 0 which is less than the signifiicance level alpha. Thus, we reject the null hypothesis that the population mean incomes of all the groups is the same.

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

The extra sum of squares is equal to the sum of squares between groups.i.e., 688235137516 The pooled variance is equal to MSW.i.e., 1920024320.

Question 3 (Modified 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 nonconstant 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.

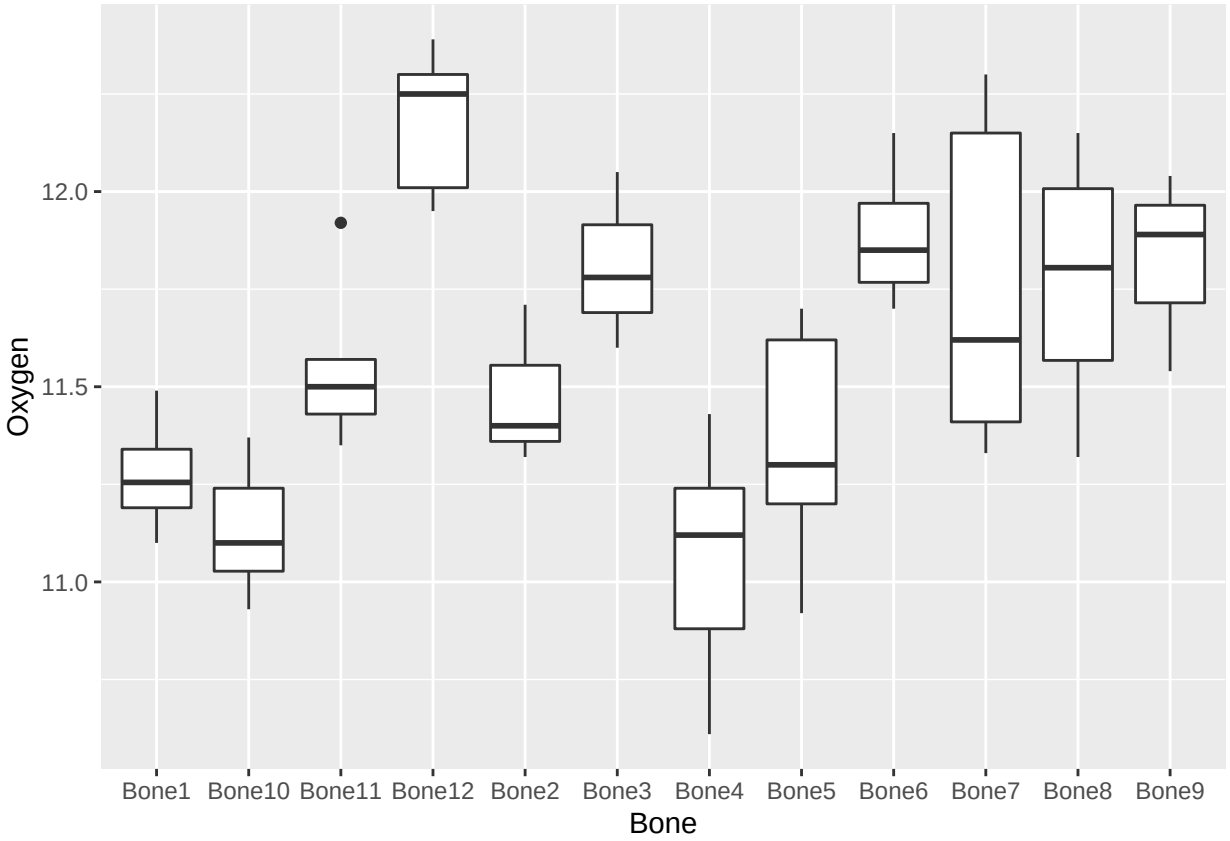
(a) (2 points) Plot the oxygen isotopic composition for each of the bones using a side-by-side boxplot. Comment on whether or not you think the population means are the same for all 12 bones based on your plot.

```
data(ex0523)
ex0523
```

```
##    Oxygen   Bone
## 1   11.10  Bone1
## 2   11.22  Bone1
## 3   11.29  Bone1
## 4   11.49  Bone1
## 5   11.32  Bone2
```

```
## 6   11.40  Bone2
## 7   11.71  Bone2
## 8   11.60  Bone3
## 9   11.78  Bone3
## 10  12.05  Bone3
## 11  10.61  Bone4
## 12  10.88  Bone4
## 13  11.12  Bone4
## 14  11.24  Bone4
## 15  11.43  Bone4
## 16  10.92  Bone5
## 17  11.20  Bone5
## 18  11.30  Bone5
## 19  11.62  Bone5
## 20  11.70  Bone5
## 21  11.70  Bone6
## 22  11.79  Bone6
## 23  11.91  Bone6
## 24  12.15  Bone6
## 25  11.33  Bone7
## 26  11.41  Bone7
## 27  11.62  Bone7
## 28  12.15  Bone7
## 29  12.30  Bone7
## 30  11.32  Bone8
## 31  11.65  Bone8
## 32  11.96  Bone8
## 33  12.15  Bone8
## 34  11.54  Bone9
## 35  11.89  Bone9
## 36  12.04  Bone9
## 37  10.93 Bone10
## 38  11.01 Bone10
## 39  11.08 Bone10
## 40  11.12 Bone10
## 41  11.28 Bone10
## 42  11.37 Bone10
## 43  11.35 Bone11
## 44  11.43 Bone11
## 45  11.50 Bone11
## 46  11.57 Bone11
## 47  11.92 Bone11
## 48  11.95 Bone12
## 49  12.01 Bone12
## 50  12.25 Bone12
## 51  12.30 Bone12
## 52  12.39 Bone12
```

```
n=nrow(ex0523)
I=12
ggplot(data=ex0523, aes(x=Bone, y=Oxygen))+geom_boxplot()
```



Looking at the box plot, we can see that the population means of all the groups, although sometimes very close, are different.

The variation between groups is much larger than the variation within groups and thus we have a strong evidence against a hypothesis that the population means are the same for all 12 bones based on the plot above.

(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.

```
summary(aov(Oxygen ~ Bone, data= ex0523))
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## Bone        11  6.067  0.5516   7.427 9.73e-07 ***
## Residuals   40  2.971  0.0743
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

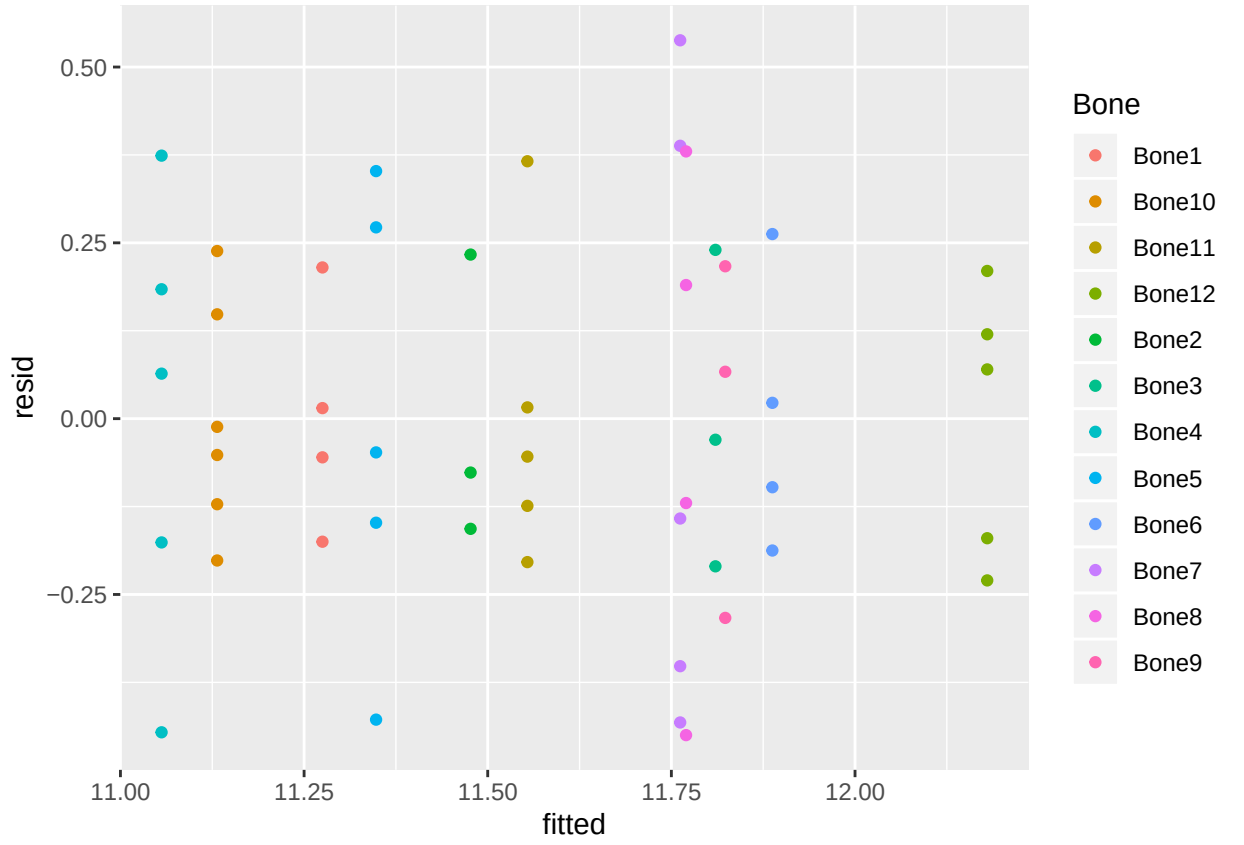
The F value is 7.427.

The p-value is 9.73e-07

The p value is less than the significance level alpha = 0.05. Hence, we reject the null hypothesis that all population means are the same in different bone types in favor of the alternative hypothesis.

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

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



The assumption seems to be not met since the residual spread in each bone population is not the same. Thus population variances are not the same for different groups.

(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=ex0523)
```

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

Kruskal-Wallis test tests whether the centers of the groups are the same or not using ranks.

The obtained p-value is very small.i.e., 0.000253.

Thus, considering the significance level, we still reject the null hypothesis that centers of the bone populations are the same.

The p value in the Kruskal-Wallis test is greater than the p-value from ANOVA. Nonetheless, we obtained the same conclusion after performing both the tests.