Sean McLean

ALY 6015

Module 2 Assignment

Professor Goulding

01/21/2024

Introduction

The goal of this assignment is to practice using chi-square and one way and two-way ANOVA testing methods on a set of designated questions. The first half of the questions centered on testing the problems with chi-square goodness-of-fit and chi-square independence tests. The second half of the questions were solved by using the ANOVA one way and ANOVA two-way testing methods. The last two questions provided datasets to import and were employed to solve the problems using the Chi-Square Goodness-of-Fit test and two-way ANOVA testing as well as data visualizations to show the results.

Question 1

6. Blood Types: A medical researcher wishes to see if hospital patients in a large hospital have the same blood type distribution as those in the general population. The distribution for the general population is as follows: type A, 20%; type B, 28%; type O, 36%; and type AB = 16%. He selects a random sample of 50 patients and finds the following: 12 have type A blood, 8 have type B, 24 have type O, and 6 have type AB blood. At $\alpha = 0.10$, can it be concluded that the distribution is the same as that of the general population?

Hypothesis

#H0: Type A = 0.20, Type B = 0.28, Type O = 0.36, Type AB = 0.16

#H1: At least one of the blood types is different from the value stated in the null hypothesis.

Results

#df = 3, a = 0.10, Critical Value = 6.251

```
alpha <- 0.10

observed <- c(12, 8, 24, 6)

p <- c(0.20, 0.28, 0.36, 0.16)

result <- chisq.test(x = observed, p = p)

result$statistic  #Chi-square test value

result$p.value    #Chi-square p-value

result$parameter  #Degrees of freedom (# of categories - 1)

result
```

Analysis

Since our chi-square value is less than the critical value, we should fail to reject the null hypothesis. There is not enough evidence to reject the claim that the blood type distribution at the hospital is the same as that of the general population.

Question 2

8. Performance by Airlines: According to the Bureau of Transportation Statistics, on-time performance by the airlines is described as follows:

Action, % of Time: On time 70.8; National Aviation System delay 8.2; Aircraft arriving late 9.0; Other (because of weather and other conditions) 12.0.

Records of 200 randomly selected flights for a major airline company showed that 125 planes were on time; 40 were delayed because of weather, 10 because of a National Aviation System delay, and the rest because of arriving late. At $\alpha = 0.05$, do these results differ from the government's statistics?

Hypothesis

#H0: On time = 70.8, NAS Delay = 8.2, Late = 9.0, Other = 12.0

#H1: The on-time performance of the airline company is different from the government's statistics.

Results

#df = 3, a = 0.05, Critical Value = 7.815

```
alpha <- 0.05

observed1 <- c(125, 10, 25, 40)

p1 <- c(0.708, 0.082, 0.09, 0.12)

result1 <- chisq.test(x = observed1, p = p1)

result1$statistic  #Chi-square test value

result1$p.value    #Chi-square p-value

result1$parameter  #Degrees of freedom (# of categories - 1)

result1
```

Analysis

Since our chi-square value is more than the critical value, we should reject the null hypothesis. There is enough evidence to reject the claim that the on-time performance by airlines is 70.8% on time, 8.2% NAS delays, 9% arriving late, and 12% other.

Question 3

8. Ethnicity and Movie Admissions: Are movie admissions related to ethnicity? A 2014 study indicated the following numbers of admissions (in thousands) for two different years. At the 0.05 level of significance, can it be concluded that movie attendance by year was dependent upon ethnicity?

Hypothesis

#H0: There is no relationship between ethnicity and movie attendance.

#H1: There is a significant relationship between ethnicity and movie attendance.

Results

```
#df = 3, a = 0.05, Critical Value = 7.815

alpha <- 0.05

r1 <- c(724, 335, 174, 107)

r2 <- c(370, 292, 152, 140)

rows = 2

mtrx = matrix(c(r1, r2), nrow = rows, byrow = TRUE)

rownames(mtrx) = c("2013", "2014")

colnames(mtrx) = c("Caucasian", "Hispanic", "African American", "Other")

mtrx

result2 <- chisq.test(mtrx)

result2

result2$statistic  #Chi-square test value

result2$p.value    #Chi-square p-value

result2$parameter  #degrees of freedom (# of columns - 1* # rows - 1)
```

Analysis

Since our chi-square test value of 60.144 is greater than the critical value of 7.815 we should reject the null hypothesis. We will conclude that there is evidence that movie attendance by year is dependent on ethnicity.

Question 4

10. Women in the Military: This table lists the numbers of officers and enlisted personnel for women in the military. At $\alpha = 0.05$, is there sufficient evidence to conclude that a relationship exists between rank and branch of the Armed Forces?

Hypothesis

#H0: The distribution of women across different ranks is independent of the

#branch of the Armed Forces.

#H1: The distribution of women across different ranks is dependent on the branch

#of the Armed Forces.

Results

```
alpha <- 0.05
#df = 3, a = 0.05, Critical Value = 7.815
r1 <- c(10791, 62491)
r2 <- c(7816, 42750)
r3 <- c(932, 9525)
r4 <- c(11819, 54344)
rows = 4
mtrx1 = matrix(c(r1, r2, r3, r4), nrow = rows, byrow = TRUE)
rownames(mtrx1) = c("Army", "Navy", "Marine Corps", "Air Force")
colnames(mtrx1) = c("Officers", "Enlisted")
mtrx1
result3 <- chisq.test(mtrx1)
result3
result3$statistic  #Chi-square test value
result3$p.value    #Chi-square p-value
result3$parameter  #degrees of freedom (# of columns - 1* # rows - 1)
```

Analysis

Since our chi-square test value of 654.27 is greater than the critical value of 7.815 we should reject the null hypothesis. There is considerable evidence to reject the claim that the distribution of women across different ranks are independent of the branch of the Armed Forces.

Question 5

8. Sodium Contents of Foods: The amount of sodium (in milligrams) in one serving for a random sample of three different kinds of foods is listed. At the 0.05 level of significance, is there sufficient evidence to conclude that a difference in mean sodium amounts exists among condiments, cereals, and desserts?

Hypothesis

#H0: There is no difference in mean sodium amounts between condiments, cereals,

#and desserts.

#H1: There is at least one mean that is different from the others in the three

#three different food groups.

Results

#df = 2, 19, a = 0.05, Critical Value = 3.52

```
alpha <- 0.05

condiments <- data.frame('sodium' = c(270, 130, 230, 180, 80, 70, 200), 'food' =
rep('condiments', 7), stringsAsFactors = FALSE)

cereals <- data.frame('sodium' = c(260, 220, 290, 290, 200, 320, 140), 'food' = rep('cereals', 7),
stringsAsFactors = FALSE)

desserts <- data.frame('sodium' = c(100, 180, 250, 250, 300, 360, 300, 160), 'food' =
rep('desserts', 8), stringsAsFactors = FALSE)

sodium <- rbind(condiments, cereals, desserts)

sodium$food <- as.factor(sodium$food)

View(sodium)

anova <- aov(sodium ~ food, data = sodium)

summary(anova)

a.summary <- summary(anova)

df.numerator <- a.summary[[1]][1, "Df"]

df.numerator

df.denominator <- a.summary[[1]][2, "Df"]

df.denominator

F.value <- a.summary[[1]][[1, "F value"]]

F.value

p.value <- a.summary[[1]][[1, "Pr(>F)"]]

p.value

TukeyHSD(anova)
```

Analysis

Because the test value of 2.399 is less than the critical value of 3.52 we should not reject the null hypothesis. There is not enough evidence to conclude that a difference in mean sodium

amounts exist in the food groups, with the largest mean differences coming from the condiments being paired with the other food groups.

Question 6

10. Sales for Leading Companies: The sales in millions of dollars for a year of a sample of leading companies are shown. At $\alpha = 0.01$, is there a significant difference in the means?

Hypothesis

#H0: There is no significant difference in the means of sales between the three companies.

#H01: There is a significant difference in the means of sales between the three companies.

Results

#df = 2, 11, a = 0.01, Critical Value = 7.206

```
alpha <- 0.01

cereal <- data.frame('sales' = c(578, 320, 264, 249, 237), 'companies' = rep('cereal', 5), stringsAsFactors = FALSE)

choc <- data.frame('sales' = c(311, 106, 109, 125, 173), 'companies' = rep('choc', 5), stringsAsFactors = FALSE)

coffee <- data.frame('sales' = c(261, 185, 302, 689), 'companies' = rep('coffee', 4), stringsAsFactors = FALSE)

sales <- rbind(cereal, choc, coffee)

sales$companies <- as.factor(sales$companies)

View(sales)

anova1 <- aov(sales ~ companies, data = sales)

summary(anova1)

a.summary1 <- summary(anova1)

df.numerator <- a.summary1[[1]][1, "Df"]

df.numerator

df.denominator <- a.summary1[[1]][2, "Df"]

df.denominator

F.value1 <- a.summary1[[1]][[1, "F value"]]

F.value1

p.value1 <- a.summary1[[1]][[1, "Pr(>F)"]]
```

p.value1

TukeyHSD(anova1)

Analysis

Because the test value of 2.172 is less than the critical value of 7.206 we should not reject the null hypothesis. There is not enough evidence to conclude that there is a difference in mean sales between the three companies.

Question 7

12. Per-Pupil Expenditures: The expenditures (in dollars) per pupil for states in three sections of the country are listed. Using $\alpha = 0.05$, can you conclude that there is a difference in means?

Hypothesis

#H0: There is no difference in the means of expenditures between the three sections of the country.

#H01: There is a difference in at least one of the expenditures between the three sections of the country.

Results

#df = 2, 9, a = 0.05, Critical Value = 4.256

alpha <- 0.05

eastern <- data.frame('expenditures' = c(4946, 5953, 6202, 7243), 'country' = rep('eastern', 4), stringsAsFactors = FALSE)

middle <- data.frame('expenditures' = c(6149, 7451, 6000, 6479), 'country' = rep('middle', 4), stringsAsFactors = FALSE)

western <- data.frame('expenditures' = c(5282, 8605, 6528, 6911), 'country' = rep('western', 4), stringsAsFactors = FALSE)

expenditures <- rbind(eastern, middle, western)

expenditures$country <- as.factor(expenditures$country)

View(expenditures)

anova2 <- aov(expenditures ~ country, data = expenditures)

summary(anova2)

a.summary2 <- summary(anova2)

df.numerator <- a.summary2[[1]][1, "Df"]

df.numerator

```
df.denominator <- a.summary2[[1]][2, "Df"]
df.denominator
F.value2 <- a.summary2[[1]][[1, "F value"]]
F.value2
p.value2 <- a.summary2[[1]][[1, "Pr(>F)"]]
p.value2
TukeyHSD(anova2)
```

Analysis

Because the test value of 0.526 is less than the critical value of 4.256 we should not reject the null hypothesis. We did not reject the null hypothesis because there is not enough evidence to conclude that a difference in means is evident among the three sections of the country.

Question 8

10. Increasing Plant Growth: A gardening company is testing new ways to improve plant growth. Twelve plants are randomly selected and exposed to a combination of two factors, a "Grow-light" in two different strengths and a plant food supplement with different mineral supplements. After a number of days, the plants are measured for growth, and the results (in inches) are put into the appropriate boxes. Can an interaction between the two factors be concluded? Is there a difference in mean growth with respect to light? With respect to plant food? Use $\alpha = 0.05$.

Hypotheses

#H0: The means of the grow light groups are equal.

#H1: The means of the grow light groups are different

#H0: The means of the plant food supplement groups are equal.

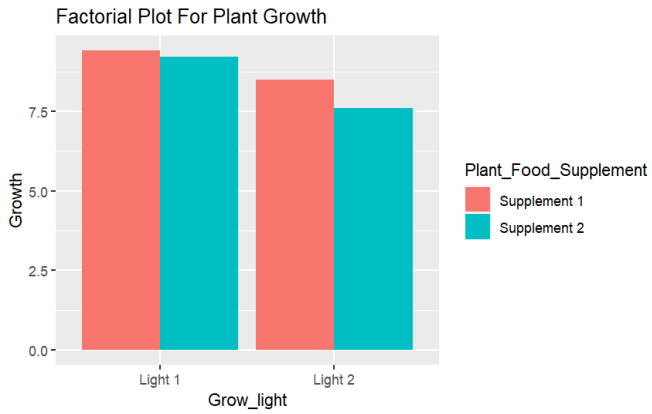
#H1: The means of the plant food supplement groups are different

#H0: There is no interaction between the grow light and plant food supplement.

#H1: There is an interaction between the grow light and plant food supplement.

Results

```
alpha <- 0.05
critical_value <- qf(1 - alpha, df1 = 1, df2 = 8)
print(critical_value)
#df = 1, 8, a = 0.05, Critical Value = 5.317655
growth <- data.frame(
  Grow_light = rep(c("Light 1", "Light 2"), each = 6),
  Plant_Food_Supplement = rep(c("Supplement 1", "Supplement 2"), times = 6),
  Growth = c(9.2, 8.5, 9.4, 9.2, 8.9, 8.9, 7.1, 5.5, 7.2, 5.8, 8.5, 7.6)
)
View(growth)
anova3 <- aov(Growth ~ Grow_light * Plant_Food_Supplement, data = growth)
summary(anova3)
a.summary3 <- summary(anova3)
df.numerator <- a.summary3[[1]][1, "Df"]
df.numerator
df.denominator <- a.summary3[[1]][2, "Df"]
df.denominator
F.value3 <- a.summary3[[1]][[1, "F value"]]
F.value3
p.value3 <- a.summary3[[1]][[1, "Pr(>F)"]]
p.value3
TukeyHSD(anova3)
ggplot(growth, aes(x = Grow_light, y = Growth, fill = Plant_Food_Supplement)) +
  geom_bar(stat = "identity", position = position_dodge()) +
  labs(title = "Factorial Plot For Plant Growth", x = "Grow_light", y = "Growth", fill =
"Plant_Food_Supplement")
```



Analysis

Because the f-test value of the factor interactions of 1.438 is less than the critical value of 5.318 we should fail to reject the null hypothesis. We failed to reject the null hypothesis because there is not enough evidence to conclude that there is a difference in mean growth with light, mean growth with plant food, and an interaction between the two factors on growth.

Question 9

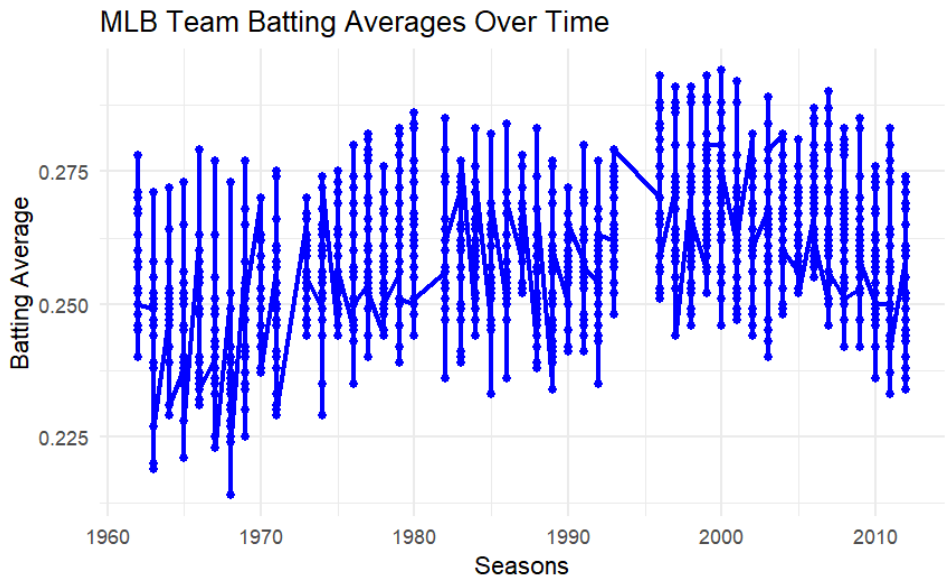
1. Baseball Question: Download the file 'baseball.csv' from the course resources and import the file into R. 2. Perform EDA on the imported data set. Write a paragraph or two to describe the data set using descriptive statistics and plots. Are there any trends or anything of interest to discuss? 3. Assuming the expected frequencies are equal, perform a Chi-Square Goodness-of-Fit test to determine if there is a difference in the number of wins by decade.

Initial Dataset Analysis

```
baseball <- read.csv("baseball.csv", header=TRUE)
baseball
summary(baseball)
View(baseball)
dim(baseball)
head(baseball)
tail(baseball)
str(baseball)
class(baseball)
range(baseball$SLG)
ggplot(baseball, aes(x = Year, y = BA)) +
```

```
  geom_line(color = "blue", size = 1) +
  geom_point(color = "blue", size = 1.5) +
  labs(title = "MLB Team Batting Averages Over Time",
       x = "Seasons",
       y = "Batting Average") +
  theme_minimal()
```

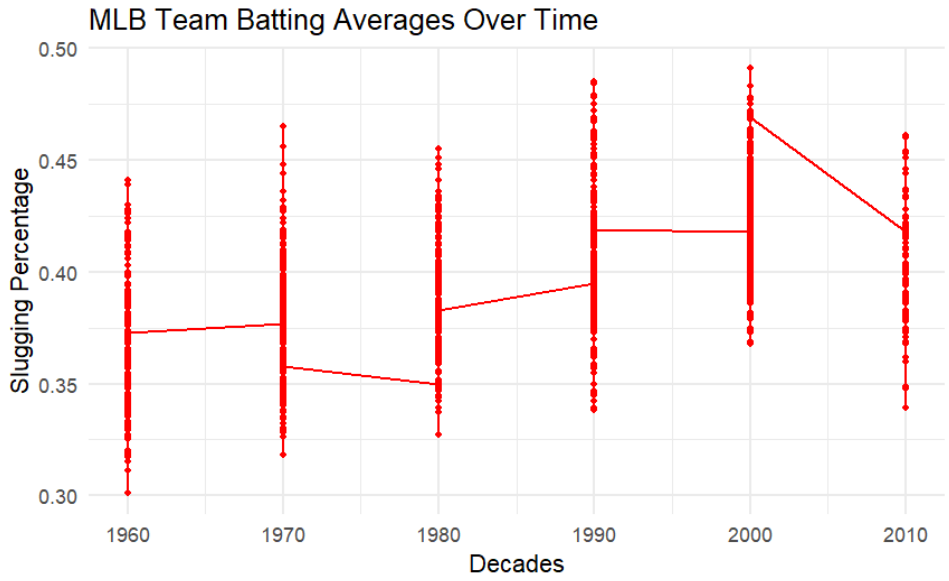
One consistent aspect of the baseball statistics is the increase in all the offensive categories overtime. There seems to be a slow but steady increase in the first three decades or so before a bigger increase starting around 1995. The category that is noticeably higher in the 1990's is slugging percentage, indicating that more offensive production like home runs began to heavily increase around this time. In the early 2000's all the offensive categories started to gradually decline, so there was a five-year period where they were at their apex before beginning to drop off.




```
ggplot(baseball, aes(x = Decade, y = SLG)) +
  geom_line(color = "red", size = 0.5) +
  geom_point(color = "red", size = 1) +
  labs(title = "MLB Team Batting Averages Over Time",
       x = "Decades",
       y = "Slugging Percentage") +
```

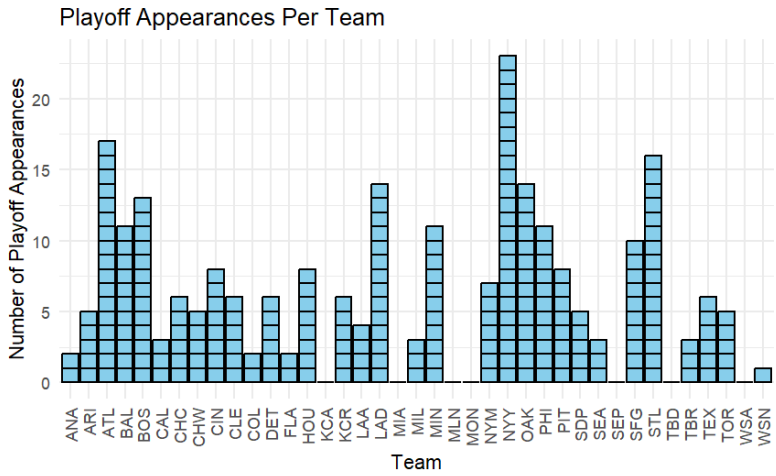
```
  theme_minimal()
```

Two offensive categories that were not measured until the late 1990's were opponents on-base percentage and opponents slugging percentage which both dropped over the next decade. This is consistent with the decline of offensive production during this time, indicating that some possible adjustments were made like better pitching and defense.




```
ggplot(baseball, aes(x = Team, y = Playoffs)) +
  geom_bar(stat = "identity", fill = "skyblue", color = "black") +
  labs(title = "Playoff Appearances Per Team",
       x = "Team",
       y = "Number of Playoff Appearances") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
```

About one third of the professional teams during the 50-year period made at least 10 playoff appearances with the New York Yankees having the highest amount by a considerable margin. About six teams did not have a single appearance and the other half were somewhere between zero and nine appearances. The numbers are misleading due to the number of teams not existing from the start of the dataset period. This would also be the reason the total number of wins per decade is different, even though each team plays 162 games per season in the dataset.




```
baseball$Decade <- baseball$Year - (baseball$Year %% 10)
wins <- baseball %>%
  group_by(Decade) %>% summarize(wins = sum(W)) %>%
    as.tibble()
wins
```

Hypothesis

#H0: There is no difference in the number of wins by decade for all teams.

#H1: There is at least one difference in the number of wins by decade for all teams.

Results

```
#df = 5, a = 0.05, Critical Value = 2.015
alpha <- 0.05
wins <- c(13267, 17934, 18926, 17972, 24286, 7289)
probabilities <- prop.table(wins)
for (i in seq_along(wins)) {
  print(paste("Decade:", i, "Wins:", wins[i], "Probability:", probabilities[i]))
}
 print(paste("Total Probability:", sum(probabilities)))
 decades <- c(0.133, 0.180, 0.190, 0.180, 0.244, 0.073)
 results <- chisq.test(x = wins, p = decades)
results$statistic  #Chi-square test value
```

```
results$p.value    #Chi-square p-value
results$parameter  #Degrees of freedom (# of categories - 1)
results
```

Final Analysis

Since our chi-square value is less than the critical value, we should fail to reject the null hypothesis. Our conclusion would be that the wins proportions by decade when comparing test and critical values and p-values and significant levels provides the same results in that it does not provide enough evidence that there is a difference in the number of wins per decade for the teams.

Question 10

2. Crop Question: Download the file 'crop_data.csv' from the course resources and import the file into R. 5. Perform a Two-way ANOVA test using yield as the dependent variable and fertilizer and density as the independent variables. Explain the results of the test. Is there reason to believe that fertilizer and density have an impact on yield?

Hypotheses

#H0: The fertilizer will not have an impact on the yield.

#H1: The fertilizer will have an impact on the yield.

#H0: The density will not have an impact on the yield.

#H1: The density will have an impact on the yield.

#H0: There will be no interaction between the fertilizer and density.

#H1: There will be an interaction between the fertilizer and density.

Results

```
alpha <- 0.05
a_factor1 <- 2
a_factor2 <- 3
N <- 96
df_factor1 <- a_factor1 - 1
df_factor1
```

```
df_factor2 <- a_factor2 - 1
df_factor2
df_interaction <- df_factor1 * df_factor2
df_interaction
df_error <- N - a_factor1 * a_factor2
df_error
df_total <- N - 1
df_total
critical_factor1 <- qf(1 - alpha/2, df_factor1, df_error)
critical_factor1
critical_factor2 <- qf(1 - alpha/2, df_factor2, df_error)
critical_factor2
critical_interaction <- qf(1 - alpha/2, df_interaction, df_error)
critical_interaction
#df = 95, a = 0.05, Critical Value = 3.844295
crop$density <- factor(crop$density)
crop$fertilizer <- factor(crop$fertilizer)
crop$block <- factor(crop$block)
anova4 <- aov(yield ~ fertilizer * density, data = crop)
summary(anova4)
a.summary4 <- summary(anova4)
df.numerator <- a.summary4[[1]][1, "Df"]
df.numerator
df.denominator <- a.summary4[[1]][2, "Df"]
df.denominator
F.value4 <- a.summary4[[1]][[1, "F value"]]
F.value4
p.value4 <- a.summary4[[1]][[1, "Pr(>F)"]]
p.value4
TukeyHSD(anova4)
```

Anaylsis

Because the f-test value of the factors interaction of 0.635 is less than the critical value of 3.844 we should fail to reject the null hypothesis. We failed to reject the null hypothesis because there is not enough evidence from the f-test and Tukey tests to conclude that density and fertilizer have significant impact on the yield.

Conclusions/Interpretations

The module assignment lab provided the necessary foundation for understanding how to properly code the questions with the correct testing. Having several examples to practice on with each test was especially useful and came away feeling like I had a better understanding of each of them. The two-way ANOVA testing produced the biggest challenges in formulating multiple hypotheses and analyzing the f-test and Tukey tests. All the problems were good examples that allowed for the opportunity to work on coding and analysis skills.

References:

Bluman, A. G. (2017). Elementary Statistics: A Step-by-Step Approach. 10th edition. McGraw-Hill Education.

Chat GPT. (2023, December 10th). Default (GPT 3.5)

Appendix:

```r
library(dplyr)
library(ggplot2)
library(corrplot)
library(RColorBrewer)
library(car)
library(MASS)

#Section 11-1
#Question 6: Blood Types
#State the hypothesis
#H0: Type A = 0.20, Type B = 0.28, Type O = 0.36, Type AB = 0.16
#H1: At least one of the blood types is not the same as
#its value stated in the null hypothesis.

#Find the Critical Value
#df = 3, a = 0.10, Critical Value = 6.251

#Compute the Test Value
#Set significance value
alpha <- 0.10

#Create a vector of the values
observed <- c(12, 8, 24, 6)

#Create a vector for the probabilities
p <- c(0.20, 0.28, 0.36, 0.16)

#Run the test and save the results to the result variable
result <- chisq.test(x = observed, p = p)

#View the test statistic and p-value
result$statistic  #Chi-square test value
result$p.value    #Chi-square p-value
result$parameter  #Degrees of freedom (# of categories - 1)

result

#Make the decision
#Since our chi-square value is less than the critical value we should fail to
#reject the null hypothesis.

#Summarize the results
#There is not enough evidence to reject the claim that the blood type
#distribution at the hospital is the same as that of the general population.

#Question 8: On-Time Performance by Airlines
#State the hypothesis
#H0: On time = 70.8, NAS Delay = 8.2, Late = 9.0, Other = 12.0
#H1: The on-time performance of the airline company is different from
#the government's statistics.

#Find the Critical Value
#df = 3, a = 0.05, Critical Value = 7.815

#Compute the Test Value
#Set significance value
alpha <- 0.05

#Create a vector of the values
observed1 <- c(125, 10, 25, 40)
```

```r

#Create a vector for the probabilities
p1 <- c(0.708, 0.082, 0.09, 0.12)

#Run the test and save the results to the result variable
result1 <- chisq.test(x = observed1, p = p1)

#View the test statistic and p-value
result1$statistic  #Chi-square test value
result1$p.value    #Chi-square p-value
result1$parameter  #Degrees of freedom (# of categories - 1)

result1

#Make the decision
#Since our chi-square value is more than the critical value we should
#reject the null hypothesis.

#Summarize the results
#There is enough evidence to reject the claim that the on-time performance
#by airlines is 70.8% on time, 8.2% NAS delays, 9% arriving late, and 12% other.

#Section 11-2
#Question 8: Ethnicity and Movie Admissions
#State the hypothesis
#H0: There is no relationship between ethnicity and movie attendance.
#H1: There is a significant relationship between ethnicity and movie attendance.

#Find the Critical Value
#df = 3, a = 0.05, Critical Value = 7.815
```

```r

#Compute the Test Value
alpha <- 0.05

#Create one vector for each row
r1 <- c(724, 335, 174, 107)
r2 <- c(370, 292, 152, 140)

#State the number of rows for the matrix
rows = 2

#Create a matrix from the rows
mtrx = matrix(c(r1, r2), nrow = rows, byrow = TRUE)

#Name the rows and columns (columns optional)
rownames(mtrx) = c("2013", "2014")
colnames(mtrx) = c("Caucasian", "Hispanic", "African American", "Other")

#View the matrix to ensure it matches your table
mtrx

#Run the test and save the results to the result variable
result2 <- chisq.test(mtrx)
result2

#View the test statistic and p-value
result2$statistic  #Chi-square test value
result2$p.value    #Chi-square p-value
result2$parameter  #degrees of freedom (# of columns - 1* # rows - 1)

```

```r
#Make the decision
#Since our chi-square test value of 60.144 is greater than the critical value of
#7.815 we should reject the null hypothesis.

#Summarize the results
#We will conclude that there is evidence that movie attendance by year is
#dependent on ethnicity.

#Question 10: Women in the Military
#State the hypothesis
#H0: The distribution of women across different ranks is independent of the
#branch of the Armed Forces.
#H1: The distribution of women across different ranks is dependent on the branch
#of the Armed Forces.

#Compute the Test Value
alpha <- 0.05

#Find the Critical Value
#df = 3, a = 0.05, Critical Value = 7.815

#Create one vector for each row
r1 <- c(10791, 62491)
r2 <- c(7816, 42750)
r3 <- c(932, 9525)
r4 <- c(11819, 54344)

#State the number of rows for the matrix
rows = 4

#Create a matrix from the rows
mtrx1 = matrix(c(r1, r2, r3, r4), nrow = rows, byrow = TRUE)

#Name the rows and columns (columns optional)
rownames(mtrx1) = c("Army", "Navy", "Marine Corps", "Air Force")
colnames(mtrx1) = c("Officers", "Enlisted")

#View the matrix to ensure it matches your table
mtrx1

#Run the test and save the results to the result variable
result3 <- chisq.test(mtrx1)
result3

#View the test statistic and p-value
result3$statistic  #Chi-square test value
result3$p.value    #Chi-square p-value
result3$parameter  #degrees of freedom (# of columns - 1* # rows - 1)

#Make the decision
#Since our chi-square test value of 654.27 is greater than the critical value of
#7.815 we should reject the null hypothesis.

#Summarize the results
#There is significant evidence to reject the claim that the distribution of
#women across different ranks is independent of the branch of the Armed Forces.

#Section 12-1
#Question 8: Sodium Contents of Foods
#State the hypotheses
```

```r
#H0:There is no difference in mean sodium amounts between condiments, cereals,
#and desserts.
#H1: There is at least one mean that is different from the others in the three
#three different food groups.

#Find the Critical Value
#df = 2, 19, a = 0.05, Critical Value = 3.52

#Set the significant level
alpha <- 0.05

#Create a data frame for condiments
condiments <- data.frame('sodium' = c(270, 130, 230, 180, 80, 70, 200), 'food' =

#Create a data frame for cereals
cereals <- data.frame('sodium' = c(260, 220, 290, 290, 200, 320, 140), 'food' =

#Create a data frame for desserts
desserts <- data.frame('sodium' = c(100, 180, 250, 250, 300, 360, 300, 160), 'fo

#Combine the data frames into one
sodium <- rbind(condiments, cereals, desserts)
sodium$food <- as.factor(sodium$food)

View(sodium)

#Run the ANOVA test
anova <- aov(sodium ~ food, data = sodium)

#View the model summary
```

Source on Save | Run | Source

```r
summary(anova)

#Save summary to an object
a.summary <- summary(anova)

#Degrees of freedom
# k-1: between group variance ~ numerator
df.numerator <- a.summary[[1]][1, "Df"]
df.numerator

#N ~ k: within group variance ~ denominator
df.denominator <- a.summary[[1]][2, "Df"]
df.denominator

#Extract the F test value from the summary
F.value <- a.summary[[1]][[1, "F value"]]
F.value

#Extract the p-value from the summary
p.value <- a.summary[[1]][[1, "Pr(>F)"]]
p.value

#See differences
TukeyHSD(anova)

#Make the decision
#Because the test value of 2.399 is less than the critical value of 3.52 we
#should not reject the null hypothesis.

#Summarize the results, and explain where the differences in the means are.
```

```r
#There is not enough evidence to conclude that a difference in mean sodium
#amounts exists in the food groups, with the largest mean differences coming
#from the condiments being paired with the other food groups.

#Section 12-2
#Question 10: Sales for Leading Companies
#State the hypotheses
#H0:There is no significant difference in the means of sales between the
#the three companies.
#H01:There is a significant difference in the means of sales between the
#the three companies.

#Find the Critical Value
#df = 2, 11, a = 0.01, Critical Value = 7.206

#Set the significant level
alpha <- 0.01

#Create a data frame for cereal
cereal <- data.frame('sales' = c(578, 320, 264, 249, 237), 'companies' = rep('ce

#Create a data frame for chocolate candy
choc <- data.frame('sales' = c(311, 106, 109, 125, 173), 'companies' = rep('choc

#Create a data frame for coffee
coffee <- data.frame('sales' = c(261, 185, 302, 689), 'companies' = rep('coffee'

#Combine the data frames into one
sales <- rbind(cereal, choc, coffee)
sales$companies <- as.factor(sales$companies)
```

```r

View(sales)

#Run the ANOVA test
anova1 <- aov(sales ~ companies, data = sales)

#View the model summary
summary(anova1)

#Save summary to an object
a.summary1 <- summary(anova1)

#Degrees of freedom
# k-1: between group variance ~ numerator
df.numerator <- a.summary1[[1]][1, "Df"]
df.numerator

#N ~ k: within group variance ~ denominator
df.denominator <- a.summary1[[1]][2, "Df"]
df.denominator

#Extract the F test value from the summary
F.value1 <- a.summary1[[1]][[1, "F value"]]
F.value1

#Extract the p-value from the summary
p.value1 <- a.summary1[[1]][[1, "Pr(>F)"]]
p.value1

#See differences
```

```
301 TukeyHSD(anova1)
302
303 #Make the decision
304 #Because the test value of 2.172 is less than the critical value of 7.206 we
305 #should not reject the null hypothesis.
306
307 #Summarize the results, and explain where the differences in the means are.
308 #There is not enough evidence to conclude that there is a difference in mean
309 #sales between the three companies.
310
311 #Question 12: Per-Pupil Expenditures
312 #State the hypotheses
313 #H0:There is no difference in the means of expenditures between the
314 #the three sections of the country.
315 #H01:There is a difference in at least one of the expenditures between the
316 #the three sections of the country.
317
318 #Find the Critical Value
319 #df = 2, 9, a = 0.05, Critical Value = 4.256
320
321 #Set the significant level
322 alpha <- 0.05
323
324 #Create a data frame for Eastern third
325 eastern <- data.frame('expenditures' = c(4946, 5953, 6202, 7243), 'country' = re
326
327 #Create a data frame for Middle third
328 middle <- data.frame('expenditures' = c(6149, 7451, 6000, 6479), 'country' = rep
329
330 #Create a data frame for Western third
```

```
Source on Save | Run | Source
331 western <- data.frame('expenditures' = c(5282, 8605, 6528, 6911), 'country' = re
332
333 #Combine the data frames into one
334 expenditures <- rbind(eastern, middle, western)
335 expenditures$country <- as.factor(expenditures$country)
336
337 View(expenditures)
338
339 #Run the ANOVA test
340 anova2 <- aov(expenditures ~ country, data = expenditures)
341
342 #View the model summary
343 summary(anova2)
344
345 #Save summary to an object
346 a.summary2 <- summary(anova2)
347
348 #Degrees of freedom
349 # k-1: between group variance ~ numerator
350 df.numerator <- a.summary2[[1]][1, "Df"]
351 df.numerator
352
353 #N ~ k: within group variance ~ denominator
354 df.denominator <- a.summary2[[1]][2, "Df"]
355 df.denominator
356
357 #Extract the F test value from the summary
358 F.value2 <- a.summary2[[1]][[1, "F value"]]
359 F.value2
360
```

```r
#Extract the p-value from the summary
p.value2 <- a.summary2[[1]][[1, "Pr(>F)"]]
p.value2

#See differences
TukeyHSD(anova2)

#Make the decision
#Because the test value of 0.526 is less than the critical value of 4.256 we
#should not reject the null hypothesis.

#Summarize the results, and explain where the differences in the means are.
#We did not reject the null hypothesis because there is not enough evidence to
#conclude that a difference in means is evident among the three sections of the
#country.

#Section 12-3
#Question 10: Increasing Plant Growth
#State the hypotheses
#H0: The means of the grow light groups are equal.
#H1: The means of the grow light groups are different

#H0: The means of the plant food supplement groups are equal.
#H1: The means of the plant food supplement groups are different

#H0: There is no interaction between the grow light and plant food supplement.
#H1: There is an interaction between the grow light and plant food supplement.

#Set the significant level
alpha <- 0.05
```

```r

#Find the Critical Value
critical_value <- qf(1 - alpha, df1 = 1, df2 = 8)
print(critical_value)
#df = 1, 8, a = 0.05, Critical Value = 5.317655

#Create a data frame for both factors
growth <- data.frame(
  Grow_light = rep(c("Light 1", "Light 2"), each = 6),
  Plant_Food_Supplement = rep(c("Supplement 1", "Supplement 2"), times = 6),
  Growth = c(9.2, 8.5, 9.4, 9.2, 8.9, 8.9, 7.1, 5.5, 7.2, 5.8, 8.5, 7.6)
)

View(growth)

#Perform two-way ANOVA
anova3 <- aov(Growth ~ Grow_light * Plant_Food_Supplement, data = growth)

#View the model summary
summary(anova3)

#Save summary to an object
a.summary3 <- summary(anova3)

#Degrees of freedom
# k-1: between group variance ~ numerator
df.numerator <- a.summary3[[1]][1, "Df"]
df.numerator

#N ~ k: within group variance ~ denominator
```

```r
df.denominator <- a.summary3[[1]][2, "Df"]
df.denominator

#Extract the F test value from the summary
F.value3 <- a.summary3[[1]][[1, "F value"]]
F.value3

#Extract the p-value from the summary
p.value3 <- a.summary3[[1]][[1, "Pr(>F)"]]
p.value3

#See differences
TukeyHSD(anova3)

# Factorial plot
ggplot(growth, aes(x = Grow_light, y = Growth, fill = Plant_Food_Supplement)) +
  geom_bar(stat = "identity", position = position_dodge()) +
  labs(title = "Factorial Plot For Plant Growth", x = "Grow_light", y = "Growth"

#Make the decision
#Because the f-test value of the factor interactions of 1.438 is less than the
#critical value of 5.318 we should fail to reject the null hypothesis.

#Summarize the results, and explain where the differences in the means are.
#We failed to reject the null hypothesis because there is not enough evidence to
#conclude that there is a difference in mean growth with light, mean growth
#with plant food, and an interaction between the two factors on growth.

#Baseball Question
#Importing the dataset
```

```r
baseball <- read.csv("baseball.csv", header=TRUE)
baseball

#Summary statistics of the data set
summary(baseball)
View(baseball)
dim(baseball)
head(baseball)
tail(baseball)
str(baseball)
class(baseball)
range(baseball$SLG)

# Line plot for batting average over different seasons
ggplot(baseball, aes(x = Year, y = BA)) +
  geom_line(color = "blue", size = 1) +
  geom_point(color = "blue", size = 1.5) +
  labs(title = "MLB Team Batting Averages Over Time",
       x = "Seasons",
       y = "Batting Average") +
  theme_minimal()

# Line plot for Slugging Percentage over different decades
ggplot(baseball, aes(x = Decade, y = SLG)) +
  geom_line(color = "red", size = 0.5) +
  geom_point(color = "red", size = 1) +
  labs(title = "MLB Team Batting Averages Over Time",
       x = "Decades",
       y = "Slugging Percentage") +
  theme_minimal()
```

```r

# Bar plot for playoff appearances per team
ggplot(baseball, aes(x = Team, y = Playoffs)) +
  geom_bar(stat = "identity", fill = "skyblue", color = "black") +
  labs(title = "Playoff Appearances Per Team",
       x = "Team",
       y = "Number of Playoff Appearances") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))

# Extract decade from year
baseball$Decade <- baseball$Year - (baseball$Year %% 10)
# Create a wins table by summing the wins by decade
wins <- baseball %>%
  group_by(Decade) %>% summarize(wins = sum(W)) %>%
    as.tibble()
wins

#State the hypothesis
#H0: There is no difference in the number of wins by decade for all teams.
#H1: There is at least one difference in the number of wins by decade for
#all teams.

#Find the Critical Value
#df = 5, a = 0.05, Critical Value = 2.015

#Compute the Test Value
#Set significance value
alpha <- 0.05

```

```r
#Create a vector for the wins
wins <- c(13267, 17934, 18926, 17972, 24286, 7289)

# Calculate probabilities per value in the vector
probabilities <- prop.table(wins)

# Display the values, their probabilities, and the total probability
for (i in seq_along(wins)) {
  print(paste("Decade:", i, "Wins:", wins[i], "Probability:", probabilities[i]))
}

print(paste("Total Probability:", sum(probabilities)))

#Create a vector for the probabilities
decades <- c(0.133, 0.180, 0.190, 0.180, 0.244, 0.073)

#Run the test and save the results to the result variable
results <- chisq.test(x = wins, p = decades)

#View the test statistic and p-value
results$statistic  #Chi-square test value
results$p.value    #Chi-square p-value
results$parameter  #Degrees of freedom (# of categories - 1)

results

#Make the decision
#Since our chi-square value is less than the critical value we should fail to
#reject the null hypothesis.

```

```r
#Summarize the results
#Our conclusion would be that the wins proportions by decade when comparing
#test and critical values and p-values and significant levels provides the same
#results in that it doesn't provide enough evidence that there is a difference
#in the number of wins per decade for the teams.


#Crop Question
#Importing the dataset
crop <- read.csv("crop_data.csv", header=TRUE)
crop

View(crop)

#Section 12-3
#Question 10: Increasing Plant Growth
#State the hypotheses
#H0: The fertilizer will not have an impact on the yield.
#H1: The fertilizer will have an impact on the yield.

#H0: The density will not have an impact on the yield.
#H1: The density will have an impact on the yield.

#H0: There will be no interaction between the fertilizer and density.
#H1: There will be an interaction between the fertilizer and density.

#Set the significant level
alpha <- 0.05

#Finding the degrees of freedom
```

```r
# Number of levels in each factor
a_factor1 <- 2
a_factor2 <- 3

# Total number of observations
N <- 96

# Degrees of freedom calculations
df_factor1 <- a_factor1 - 1
df_factor1
df_factor2 <- a_factor2 - 1
df_factor2
df_interaction <- df_factor1 * df_factor2
df_interaction
df_error <- N - a_factor1 * a_factor2
df_error
df_total <- N - 1
df_total

# Critical values for factors and interaction
critical_factor1 <- qf(1 - alpha/2, df_factor1, df_error)
critical_factor1
critical_factor2 <- qf(1 - alpha/2, df_factor2, df_error)
critical_factor2
critical_interaction <- qf(1 - alpha/2, df_interaction, df_error)
critical_interaction
#df = 95, a = 0.05, Critical Value = 3.844295

# Convert to factors
crop$density <- factor(crop$density)
```

```r
crop$fertilizer <- factor(crop$fertilizer)
crop$block <- factor(crop$block)

#Perform two-way ANOVA
anova4 <- aov(yield ~ fertilizer * density, data = crop)

#View the model summary
summary(anova4)

#Save summary to an object
a.summary4 <- summary(anova4)

#Degrees of freedom
# k-1: between group variance ~ numerator
df.numerator <- a.summary4[[1]][1, "Df"]
df.numerator

#N ~ k: within group variance ~ denominator
df.denominator <- a.summary4[[1]][2, "Df"]
df.denominator

#Extract the F test value from the summary
F.value4 <- a.summary4[[1]][[1, "F value"]]
F.value4

#Extract the p-value from the summary
p.value4 <- a.summary4[[1]][[1, "Pr(>F)"]]
p.value4

#See differences
TukeyHSD(anova4)

#Make the decision
#Because the f-test value of the factors interaction of 0.635 is less than the
#critical value of 3.844 we should fail to reject the null hypothesis.

#Summarize the results, and explain where the differences in the means are.
#We failed to reject the null hypothesis because there is not enough evidence
#from the f-test and Tukey tests to conclude that density and fertilizer have
#a significant impact on the yield.
```