Homework 4

Peyton Hall

02/14/2025

Load Necessary Libraries

```
library(readxl)
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.4     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(ggplot2)
```

1. An obstetrician wanted to determine the impact that two experimental diets (A and B) and age of moms had on the birth weights of the infants. She randomly selected 18 pregnant mothers in the first trimester of whom 6 were 20 to 29 years old, 6 were 30 to 39 years old, and 6 were 40 or older. For each age group, she randomly assigned the mothers to one of the two diets. After delivery she measured the birth weights (in grams) of the babies and obtained the following data. Assume that the data follow the normal distribution. Age, Diet A B 20 - 29, 4473, 3667 3878, 3139 3936, 3356 30 - 39, 3886, 2762 4147, 3551 3693, 3272 40 or older, 3878, 2781 4002, 3138 3382, 3435

a) What is the null and alternative hypothesis, in symbols, for testing the main effect of age? H_0 : $\mu_{20-29} = \mu_{30-39} = \mu_{40+}$ vs H_a : At least one differs
b) What is the test statistic and p-value for testing the main effect of age?
c) What conclusion in the context can you draw based on the p-value in b)? (is there any significant mean weight difference between the three age groups?)
d) What is the null and alternative hypothesis, in symbols, for testing the main effect of diet? H_0 : $\mu_{\text{DietA}} = \mu_{\text{DietB}}$ vs H_a : $\mu_{\text{DietA}} \neq \mu_{\text{DietB}}$
e) What is the test statistic and p-value for testing the main effect of diet?
f) What conclusion in the context can you draw based on the p-value in e)? (is there any significant mean weight difference between the A and B diet groups?)
g) Is there a significant interaction between diet and age? List the p-value to explain.

Question 1 Code

```r
# b) find test statistic and p-value
age_group <- rep(c("20 - 29", "30 - 39", "40 or older"), each = 3)
diet_a <- c(4473, 3878, 3936, 3886, 4147, 3693, 3878, 4002, 3382)
diet_b <- c(3667, 3139, 3356, 2762, 3551, 3272, 2781, 3138, 3435)
diet_data <- data.frame(Age = age_group, Diet_A = diet_a, Diet_B = diet_b)

# reshape data to long format
data_long <- pivot_longer(diet_data, cols = c(Diet_A, Diet_B),
                          names_to = "Diet", values_to = "Birth_Weight")
data_long$Diet <- factor(data_long$Diet)
data_long$Age <- factor(data_long$Age)

anova_result1 <- aov(Birth_Weight~Age * Diet, data = data_long)
summary(anova_result1)
```

```
##             Df  Sum Sq Mean Sq F value   Pr(>F)
## Age          2  285442  142721   1.416 0.280557
## Diet         1 2117682 2117682  21.005 0.000629 ***
## Age:Diet     2    5646    2823   0.028 0.972449
## Residuals   12 1209822  100819
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

f = 1.416; p-value = 0.280557 Fail to reject H0; there is no evidence to support the claim that there is significant mean weight difference between the three age groups. f = 21.005; p-value = 0.000629 Reject H0; there is significant evidence to support the claim that significant mean weight difference between the A and B diet groups exist. answer part g f = 0.028; p-value = 0.972449 Fail to reject H0; there is no significant evidence to support the claim that there is a significant interaction between diet and age.

2. A researcher took 64 water samples from a stream running through a forest. The samples were collected over the course of a year. Each sample was analyzed for concentration of dissolved organic carbon (mg/L) in it. Each sample was categorized according to the type of water collected: groundwater from organic soil (Organic) or groundwater from mineral soil (Mineral), and categorized according to the location of the sample: the surface of the stream or the bottom of the stream. The researchers wanted to determine if the mean concentration of dissolved organic carbon was the same between the two types of water collected, and between the two locations. The researchers want to know which type of water has the highest dissolved organic carbon. The data is on D2L. Answer the following questions:

a) What are the null and alternative of testing the mean difference between the two types of water? $H_0 : \mu_{\text{Organic}} = \mu_{\text{Mineral}}$ vs $H_a : \mu_{\text{Organic}} \neq \mu_{\text{Mineral}}$
b) What is the test statistic and p-value for a)?
c) What conclusion in the context from the p-value in a)?
d) What are the null and alternative of testing the mean difference between the two locations? $H_0 : \mu_{\text{Surface}} = \mu_{\text{Bottom}}$ vs $H_a : \mu_{\text{Surface}} \neq \mu_{\text{Bottom}}$
e) What is the test statistic and p-value for d)?
f) What conclusion in the context from the p-value in d)?
g) What is the test statistic and p-value for testing the interaction? What can you conclude from that p-value?
h) Generate an interaction plot. Based on the plot, which type of water at which location reached the highest dissolved organic carbon?

Question 2 Code

```r
SoilTesting <- read_excel("~/Desktop/STAT 301/Week 4/SoilTesting.xls")
SoilTesting$Carbon<-as.numeric(SoilTesting$Carbon)
# b)
organic <- SoilTesting[SoilTesting$Source == "organic", ]
mineral <- SoilTesting[SoilTesting$Source == "mineral", ]
t.test(organic$Carbon, mineral$Carbon, alternative = "two.sided", var.equal = TRUE)
```

```
##
##  Two Sample t-test
##
## data:  organic$Carbon and mineral$Carbon
## t = 2.7642, df = 62, p-value = 0.007504
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  1.146101 7.133899
## sample estimates:
## mean of x mean of y
##  14.66781  10.52781
```

```r
# e)
surface <- SoilTesting %>% filter(Location == "surface")
bottom <- SoilTesting %>% filter(Location == "bottom")
t.test(surface$Carbon, bottom$Carbon, alternative = "two.sided", var.equal = TRUE)
```

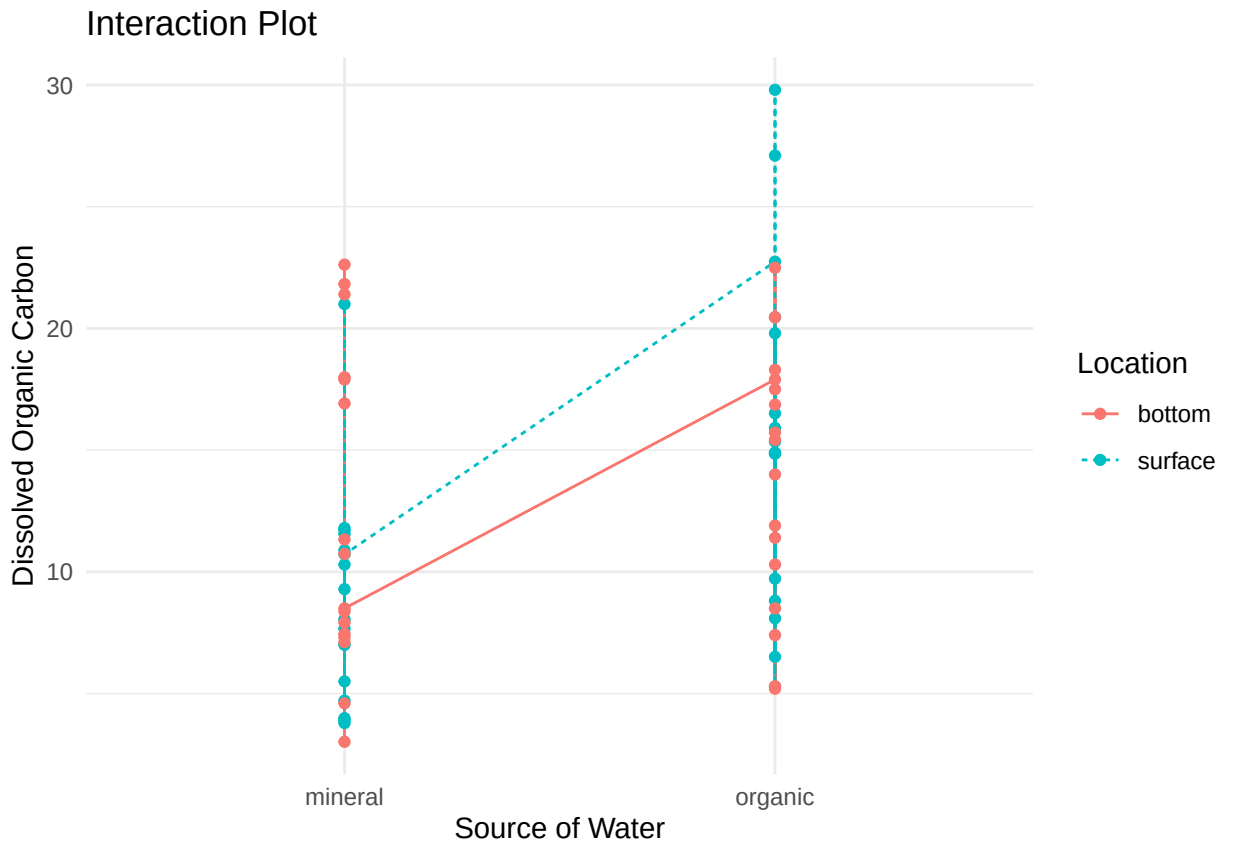
```
##
##  Two Sample t-test
##
## data:  surface$Carbon and bottom$Carbon
## t = -0.41519, df = 62, p-value = 0.6794
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -3.826746  2.510496
## sample estimates:
## mean of x mean of y
##  12.26875  12.92688
```

```r
# g)
interaction_result <- aov(Carbon~Source * Location, data = SoilTesting)
summary(interaction_result)
```

```
##                 Df Sum Sq Mean Sq F value  Pr(>F)
## Source           1  274.2  274.23   7.817 0.00694 **
## Location         1    6.9    6.93   0.198 0.65832
## Source:Location  1  113.3  113.32   3.230 0.07734 .
## Residuals       60 2105.0   35.08
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# h)
ggplot(SoilTesting, aes(x = Source, y = Carbon, color = Location, group = Location)) +
  geom_line(aes(linetype = Location)) +
```

```
  geom_point() +
  labs(title = "Interaction Plot", x = "Source of Water", y = "Dissolved Organic Carbon") +
  theme_minimal()
```



b and c) f = 2.7642; p-value = 0.007504 Reject h0; there is significant evidence to support the claim that the mean of the surface does not equal the mean of the bottom e and f) t = -0.41519; p-value = 0.6794 Fail to reject H0; there is significant evidence to support the claim that the mean of the surface equals the mean of the bottom. Question 1 Part g): f = 3.230; p = 0.07734 There is no significant evidence to support the claim that the interaction between the source of the water (organic vs. mineral) and the location (surface vs. bottom) significantly affects the dissolved organic carbon levels.

3. Researchers have sought to examine the effect of various types of music on agitation levels in patients who are in the early and middle stages of Alzheimer's disease. Patients were selected to participate in the study based on their stage of Alzheimer's disease. Three forms of music were tested: Easy listening, Mozart, and Piano Interludes. While listening to music, agitation levels were recorded for the patients with a high score indicating a higher level of agitation. Scores are recorded below. Use the data to answer the following questions: Piano Interlude, Mozart, Easy Listening Early State Alzheimer's, 21, 9, 29 24, 12, 26 22, 10, 30 18, 5, 24 20, 9, 26 Middle State Alzheimer's, 22, 14, 15 20, 18, 18 25, 11, 20 18, 9, 13 20, 13, 19

a) What is the test statistic and p-value for testing the mean score difference between the early state Alzheimers and the middle state Alzheimers?
b) What do you conclude in the context from the p-value in a)? (Is there any significant mean score difference between the two states?)
c) What is the test statistic and p-value for testing the mean score difference between the three music types?

d) What can you conclude from the p-value in c)? (Is there any significant mean score difference between the three types of music?)
e) Is there a significant interaction between music type and the state of Alzheimer? List the p-value.
f) If there is a significant interaction between the music type and the state of Alzheimer, separate the data based on the Alzheimer state and answer the following:

   a. For the Early State Alzheimer's, what is the test statistic and p-value to compare the mean scores from the three music types? What can you conclude from the p-value?
   b. For the Middle State Alzheimer's, what is the test statistic and p-value to compare the mean scores from the three music types? What can you conclude from that p-value?

Question 3 Code

```
piano_interlude_early <- c(21, 24, 22, 18, 20)
mozart_early <- c(9, 12, 10, 5, 9)
easy_listening_early <- c(29, 26, 30, 24, 26)

piano_interlude_middle <- c(22, 20, 25, 18, 20)
mozart_middle <- c(14, 18, 11, 9, 13)
easy_listening_middle <- c(15, 18, 20, 13, 19)

early_state <- data.frame(
  Piano_Interlude = piano_interlude_early,
  Mozart = mozart_early,
  Easy_Listening = easy_listening_early
)

middle_state <- data.frame(
  Piano_Interlude = piano_interlude_middle,
  Mozart = mozart_middle,
  Easy_Listening = easy_listening_middle
)

# a)
t.test(piano_interlude_early, piano_interlude_middle, var.equal = FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  piano_interlude_early and piano_interlude_middle
## t = 0, df = 7.7838, p-value = 1
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -3.589793  3.589793
## sample estimates:
## mean of x mean of y
##        21        21
```

```
t.test(mozart_early, mozart_middle, var.equal = FALSE)
```

```
##
##  Welch Two Sample t-test
##
```

```
## data:  mozart_early and mozart_middle
## t = -2.1082, df = 7.4269, p-value = 0.07072
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -8.4348021  0.4348021
## sample estimates:
## mean of x mean of y
##         9        13
```

```
t.test(easy_listening_early, easy_listening_middle, var.equal = FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  easy_listening_early and easy_listening_middle
## t = 5.8722, df = 7.7691, p-value = 0.0004172
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   6.052604 13.947396
## sample estimates:
## mean of x mean of y
##        27        17
```

```
# b)
# combine the existing data into one dataset
combined_scores <- c(piano_interlude_early, mozart_early, easy_listening_early,
                     piano_interlude_middle, mozart_middle, easy_listening_middle)
music_type <- factor(rep(c("Piano_Interlude", "Mozart", "Easy_Listening"), each = 10))
stage <- rep(c("Early", "Middle"), each = 15)
# create a data frame for the ANOVA
anova_data <- data.frame(Score = combined_scores, Music_Type = music_type, Stage = stage)

# c) perform ANOVA
anova_results <- aov(Score~Music_Type, data = anova_data)
summary(anova_results)
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## Music_Type   2    540  270.00   10.91 0.000336 ***
## Residuals   27    668   24.74
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# e)
anova_results_interaction <- aov(Score~Music_Type * Stage, data = anova_data)
summary(anova_results_interaction)
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## Music_Type   2    540  270.00  12.145 0.000189 ***
## Stage        1     90   90.00   4.048 0.054681 .
## Residuals   26    578   22.23
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

a) and b): Piano Interlude: t = 0.0; p-value = 1.0 Fail to reject H0; there is no evidence to support the claim that there is a significant mean agitation score difference between the early and middle state Alzheimer's patients when listening to Piano Interlude. Mozart: t = -2.1082; p-value = 0.07072 Fail to reject H0; there is no significant evidence to support the claim that there is a significant mean agitation score difference between the early and middle state Alzheimer's patients when listening to Mozart. Easy Listening: t = 5.8722; p-value = 0.0004172 Reject H0; there is significant evidence to support the claim that there is a significant mean agitation score difference between the early and middle state Alzheimer's patients when listening to Easy Listening music.

b) and d): f = 10.91; p-value = 0.000336 Reject H0; there is significant evidence to support the claim that significant mean score differences between the three music types exist. e): Music_Type p-value = 0.000189 Stage: p-value = 0.054681 The interaction term between music type and Alzheimer's stage is not significant.

4. Use the "job and gender" data on d2l. Test the main effects of gender, main effects of job category and the interaction between job and gender. Note that this is an unbalanced design.

a) Perform a two-way ANOVA to test the main effect of gender

   i) What is the test statistic and p-value for testing the main effect of gender?
   ii) What do you conclude? Is there significant mean salary difference between male and female?

b) Test the main effect of job category

   i) What is the test statistic and p-value for testing the main effect of job category?
   ii) What do you conclude? Is there significant mean salary difference between the janitors, clerks, and managers?

c) If there is a significant interaction, separate the data by job category and answer the following:

   a. For the janitors, what is the test statistic and p-value for comparing the mean salary difference between males and females? Is there any significant difference between the average salary from male janitors and female janitors?
   b. For the clerks, what is the test statistic and p-value for comparing the mean salary difference between males and females? Is there any significant difference between the average salary from male clerks and female clerks?
   c. For the managers, what is the test statistic and p-value for comparing the mean salary difference between males and females? Is there any significant difference between the average salary from male managers and female managers?

Question 4 Code

```
job_and_gender <- read_excel("~/Desktop/STAT 301/Week 4/job and gender.xlsx")

# a)
anova_results <- aov(Salary~Jobcat * Gender, data = job_and_gender)
summary(anova_results)
```

```
##                Df    Sum Sq   Mean Sq F value   Pr(>F)
## Jobcat          2 3.030e+10 1.515e+10  921.51  < 2e-16 ***
## Gender          1 1.467e+09 1.467e+09   89.23 1.62e-11 ***
## Jobcat:Gender   2 5.858e+08 2.929e+08   17.82 3.48e-06 ***
## Residuals      38 6.247e+08 1.644e+07
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# c)
janitors <- subset(job_and_gender, Jobcat == "janitor")
clerks <- subset(job_and_gender, Jobcat == "clerk")
managers <- subset(job_and_gender, Jobcat == "manager")

# Perform t-tests comparing male vs. female salaries
t_test_janitors <- t.test(Salary~Gender, data = janitors, var.equal = FALSE)
t_test_clerks <- t.test(Salary~Gender, data = clerks, var.equal = FALSE)
t_test_managers <- t.test(Salary~Gender, data = managers, var.equal = FALSE)
t_test_janitors
```

```
##
##  Welch Two Sample t-test
##
## data:  Salary by Gender
## t = -1.4309, df = 10.635, p-value = 0.1812
## alternative hypothesis: true difference in means between group Female and group Male is not equal to
## 95 percent confidence interval:
##  -3199.9064   684.9064
## sample estimates:
## mean in group Female   mean in group Male
##              20372.5              21630.0
```

```r
t_test_clerks
```

```
##
##  Welch Two Sample t-test
##
## data:  Salary by Gender
## t = -7.2276, df = 9.6247, p-value = 3.471e-05
## alternative hypothesis: true difference in means between group Female and group Male is not equal to
## 95 percent confidence interval:
##  -14881.059  -7839.491
## sample estimates:
## mean in group Female   mean in group Male
##             31338.60             42698.88
```

```r
t_test_managers
```

```
##
##  Welch Two Sample t-test
##
## data:  Salary by Gender
## t = -7.4818, df = 14.319, p-value = 2.581e-06
## alternative hypothesis: true difference in means between group Female and group Male is not equal to
## 95 percent confidence interval:
##  -24631.78 -13673.78
## sample estimates:
## mean in group Female   mean in group Male
##             71609.56             90762.33
```

a) f = 89.23; p-value = 1.62e-11 There is a significant mean salary difference between male and female.

b) f = 921.51; p-value < 2e-16 There is a significant mean salary difference between janitors, clerks, and managers.

c)

a. t = -1.4309; p-value = 0.1812 There is no significant mean salary difference between male janitors and female janitors.
b. t = -7.2276; p-value = 3.471e-05 There is a significant mean salary difference between male clerks and female clerks.
c. t = -7.4818; p-value = 2.581e-06 There is a significant mean salary difference between male managers and female managers.