Homework 3

Peyton Hall

02/07/2025

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
```

1. Dr. Brown would like to compare his students' performance after he taught three classes with three different pedagogies. He used his students' final exam scores as the measurements. He took a sample of 12 students' final scores from pedagogy A, 12 from pedagogy B and 12 from C. Use the 0.05 significance level to test the claim that there is a significant difference between the mean final scores of the three pedagogies. The data is on D2L, named as Student final scores.

a) Step 1: Formulate the null and alternative hypotheses in symbols H_0 : $\mu_A = \mu_B = \mu_C$ vs H_a : At least one differs Step 2: Choose the one-way ANOVA test
b) Step 3: Find the correct test statistic and p-value for testing the means from R output
c) Step 4: Make a decision based on the significance level of 0.05; and explain your decision in the context.
d) If there are significant differences, perform a pairwise comparison to determine where the differences lie. Question 1 Code

```
Student_final_scores <- read_excel("~/Desktop/STAT 301/Week 3/Student final scores.XLSX")

# Step 3
anova_result1 <- aov(Scores ~ Pedagogy, data = Student_final_scores)
summary(anova_result1)
```

```
##             Df Sum Sq Mean Sq F value  Pr(>F)
## Pedagogy     2    672   336.0   5.926 0.00632 **
## Residuals   33   1871    56.7
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# Step 4 d)
TukeyHSD(anova_result1) # pairwise comparison test
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = Scores ~ Pedagogy, data = Student_final_scores)
##
## $Pedagogy
##     diff        lwr       upr     p adj
## B-A    8  0.4570281 15.542972 0.0357471
## C-A   10  2.4570281 17.542972 0.0072393
## C-B    2 -5.5429719  9.542972 0.7933188
```

f = 5.926; p-value = 0.00632 Reject H0; there is evidence to support the claim that there is a significant difference between the mean final scores of the three pedagogies. In the pairwise comparison test, it is found that the mean of B significantly differs from that of A, the mean of C significantly differs from that of A, and the mean of C does not significantly differ from that of B.

2. The data CO2 on d2l recorded the CO2 emission from the burning of fossil fuels (metric tonnes of CO2 per person) (Source: https://cdiac.ess-dive.lbl.gov/). The data recorded the CO2 emission per person from three countries: USA, Russia and India. Use the 0.05 significance level to test the claim that the average CO2 emission per person are significantly different between the three counties. Note, the format of the data is wide format. You will need to use the gather()OR the pivot_longer()function in R to convert it to a long format with one independent variable and one dependent variable so the format is analysis ready.

a) Step 1: Formulate the null and alternative hypotheses in symbols $H_0 : \mu_{USA} = \mu_{Russia} = \mu_{India}$ vs H_a : At least one differs Step 2: Choose the one-way ANOVA test
b) Step 3: Find the correct test statistic and p-value for testing the means from R output
c) Step 4: Make a decision based on the significance level of 0.05; and explain your decision in the context.
d) If you discover the significant difference, Where the differences lie? Question 2 Code

```r
CO2 <- read_excel("~/Desktop/STAT 301/Week 3/CO2.XLSX")
# Step 3
CO2_long <- pivot_longer(
  CO2,
  cols = c(`United States`, Russia, India),
  names_to = "Country",
  values_to = "Emissions"
)
anova_result2 <- aov(Emissions ~ Country, data = CO2_long)
summary(anova_result2)
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## Country      2   4652  2326.0    1524 <2e-16 ***
## Residuals   81    124     1.5
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# 2 d)
TukeyHSD(anova_result2) # pairwise comparison test
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = Emissions ~ Country, data = CO2_long)
##
## $Country
##                             diff       lwr       upr p adj
## Russia-India          10.018750  9.230436 10.807064     0
## United States-India   18.197679 17.409365 18.985993     0
## United States-Russia   8.178929  7.390615  8.967243     0
```

f = 1524; p-value = <2e-16 Reject H0; there is evidence to support the claim that the average CO2 emission per person are significantly different between the three counties. In the pairwise comparison test, it is found that the mean CO2 emissions per person significantly differ between each pair of countries analyzed. Specifically, the mean emissions from the United States are significantly higher than those from India by approximately 18.2 metric tonnes and also higher than those from Russia by about 8.2 metric tonnes. Additionally, the mean emissions from Russia are significantly higher than those from India by approximately 10.0 metric tonnes. These results indicate substantial differences in CO2 emissions per capita among the three countries.

3. A clinical research was conducted to compare three treatments (drug A, B and

C) on treating depression. The researchers recruited 24 volunteers and randomly assigned 8 to take drug A, 8 to take B, and 8 to take C. The baseline characteristics of the 24 volunteers are the same. The researchers measured the depression scores after the volunteers took the medications. The data is listed in the table below. Suppose the data follow normal distribution. The lower depression score, the more effective the drug is. Use the 0.05 significance level to test the claim that the average depression scores after taking the drugs are significantly different. Drug A, 51, 45, 33, 41, 36, 38, 39, 33 Drug B, 23, 31, 23, 20, 19, 26, 28, 17 Drug C, 22, 18, 29, 32, 41, 20, 16, 23

a) Enter the data in R with the correct format (one IV and one DV)
b) Step 1: Formulate the null and alternative hypotheses in symbols $H_0 : \mu_A = \mu_B = \mu_C$ vs H_a : At least one differs Step 2: Choose the one-way ANOVA test
c) Step 3: Find the correct test statistic and p-value for testing the means from R output
d) Step 4: Make a decision based on the significance level of 0.05; and explain your decision in the context.
e) Which two drugs had significantly different average depression scores? Question 3 Code

```r
# a) Enter data
Drug <- rep(c("A", "B", "C"), each = 8)
DepressionScore <- c(51, 45, 33, 41, 36, 38, 39, 33,  # Drug A scores
                     23, 31, 23, 20, 19, 26, 28, 17,  # Drug B scores
                     22, 18, 29, 32, 41, 20, 16, 23)  # Drug C scores
depressiondf <- data.frame(Drug, DepressionScore)

# Step 3, use aov function to perform analysis of variance
anova_result3 <- aov(DepressionScore ~ Drug, data = depressiondf)
summary(anova_result3)
```

```
##              Df Sum Sq Mean Sq F value   Pr(>F)
## Drug          2 1252.6   626.3   14.44 0.000113 ***
## Residuals    21  910.8    43.4
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# 3 e)
TukeyHSD(anova_result3) # pairwise comparison test
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = DepressionScore ~ Drug, data = depressiondf)
##
## $Drug
##        diff       lwr      upr     p adj
## B-A -16.125 -24.42463 -7.82537 0.0002172
## C-A -14.375 -22.67463 -6.07537 0.0007607
## C-B   1.750  -6.54963 10.04963 0.8568554
```

f = 14.44; p-value = 0.000113 Reject H0; there is evidence to support the claim that the average depression scores after taking the drugs are significantly different. In the pairwise comparison test, it is found that the mean depression scores for Drug B and Drug C are significantly lower than those for Drug A. Specifically, the mean depression score for Drug B is significantly lower than that for Drug A by approximately 16.1 points, and the mean score for Drug C is also lower than Drug A by about 14.4 points. However, there is no significant difference between the depression scores for Drug B and Drug C.

4. A representative from a seed-producing company has been on a seed-collecting trip in the hope of finding improved varieties of millet. She wants to discover whether the yields from different local seed sources are significantly different. An experiment is set up with five different seed sources (A, B, C, D and E). For each seed source, she planted seven replications of plots. She measured the yield from each plot at the end of the growing season. She is interested in testing the main effect of seed sources with the significance level of 0.05. The data is as following. Perform an appropriate test to answer the following questions: A, B, C, D, E 1.4, 1.5, 0.3, 1.5, 1.1 1.2, 1.8, 0.5, 1.9, 1.5 1.1, 1.6, 0.8, 1.2, 0.7 1.0, 1.3, 0.2, 1.6, 1.4 0.9, 1.2, 0.6, 1.8, 1.2 1.8, 1.9, 0.5, 1.5, 1.3 1.2, 1.7, 0.3, 0.6, 1.7

a) Enter the data in R with the correct format.
b) Step 1: State the null and alternative hypotheses H_0 : $\mu_A = \mu_B = \mu_C = \mu_D = \mu_E$ vs H_a : At least one differs Step 2: Choose the one-way ANOVA test
c) Step 3: Find the test statistic and p-value for testing the significant difference in mean
d) Step 4: What conclusion can you draw based on the p-value and the significance level of 0.05?
e) Step 5: Which two seed sources are significantly different from each other based on the significance level of 0.05? Question 4 Code

```r
# a) Enter data in R
A <- c(1.4, 1.2, 1.1, 1.0, 0.9, 1.8, 1.2)
B <- c(1.5, 1.8, 1.6, 1.3, 1.2, 1.9, 1.7)
C <- c(0.3, 0.5, 0.8, 0.2, 0.6, 0.5, 0.3)
D <- c(1.5, 1.9, 1.2, 1.6, 1.8, 1.5, 0.6)
E <- c(1.1, 1.5, 0.7, 1.4, 1.2, 1.3, 1.7)
yields <- data.frame(A, B, C, D, E)
yields
```

```
##     A   B   C   D   E
## 1 1.4 1.5 0.3 1.5 1.1
## 2 1.2 1.8 0.5 1.9 1.5
## 3 1.1 1.6 0.8 1.2 0.7
## 4 1.0 1.3 0.2 1.6 1.4
## 5 0.9 1.2 0.6 1.8 1.2
## 6 1.8 1.9 0.5 1.5 1.3
## 7 1.2 1.7 0.3 0.6 1.7
```

```
# Step 3: Find test statistic
yields <- data.frame(
  value = c(A, B, C, D, E),
  seed_source = factor(rep(c("A", "B", "C", "D", "E"), each = 7))
) # data frame in long format created

anova_result4 <- aov(value ~ seed_source, data = yields)
summary(anova_result4)
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## seed_source  4  5.282  1.3204   13.49 2.11e-06 ***
## Residuals   30  2.937  0.0979
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Step 5
TukeyHSD(anova_result4) # pairwise comparison test
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = value ~ seed_source, data = yields)
##
## $seed_source
##            diff        lwr        upr     p adj
## B-A  0.34285714 -0.1422715  0.8279858 0.2678492
## C-A -0.77142857 -1.2565572 -0.2862999 0.0006209
## D-A  0.21428571 -0.2708429  0.6994143 0.7043831
## E-A  0.04285714 -0.4422715  0.5279858 0.9989935
## C-B -1.11428571 -1.5994143 -0.6291571 0.0000021
## D-B -0.12857143 -0.6137001  0.3565572 0.9375947
## E-B -0.30000000 -0.7851286  0.1851286 0.3957282
## D-C  0.98571429  0.5005857  1.4708429 0.0000176
## E-C  0.81428571  0.3291571  1.2994143 0.0003059
## E-D -0.17142857 -0.6565572  0.3137001 0.8418369
```

f = 13.49; p-value = 2.11e-06 Reject H0; there is evidence to support the claim that the yields from different local seed sources are significantly different. In the pairwise comparison test, it is found that the mean of C significantly differs from that of A, B, D, and E.

5. Suppose that we have a different data for the chest deceleration for three different size categories (small, midsize and large) of cars. In the sample, we have a total of 21 cars. We still want to use the significance level of 0.05 and test the claim that the different size categories have the different chest

deceleration in the standard crash test. Suppose the equal variance assumption is satisfied. Part of the output from SPSS is listed below. Using the information from the SPSS output, calculate the MS_B, MS_W, df_B, df_W, and the F statistic.

Sources, Sum of Squares, Df, Mean Square, F Statistic, P-Value Between Groups, 200.857, ?, ?, ?, 0.061 Within Groups, 549.714, ?, ?, ?, 0.061

Show your calculations:

MS_B = Mean Square Between Groups MS_W = Mean Square Within Groups df_b = degrees of freedom between groups df_w = degrees of freedom within groups

df_b = # of groups - 1 = k - 1 = 3 - 1 = 2

df_w = total observations - # of groups = n - k = 21 - 3 = 18

MS_B = (SS_B) / (df_B) = 200.857 / 2 = 100.4285

MS_W = (SS_W) / (df_W) = 549.714 / 18 = 30.5396

F = (MS_B) / (MS_W) = 100.4285 / 30.5396 = 3.28846809