Assignment 6

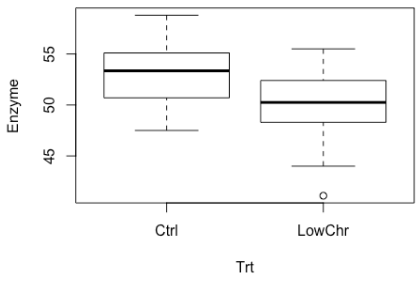
Natalie Schmer

1 Ratliver Variance Tests

```
library(car)

## Loading required package: carData

library(tidyverse)

## ── Attaching packages ─────────────────────────────────────── tidyverse
1.2.1 ──

## ✓ ggplot2 3.2.1     ✓ purrr   0.3.3
## ✓ tibble  2.1.3     ✓ dplyr   0.8.3
## ✓ tidyr   1.0.0     ✓ stringr 1.4.0
## ✓ readr   1.3.1     ✓ forcats 0.4.0

## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ✖ dplyr::recode() masks car::recode()
## ✖ purrr::some()   masks car::some()

RatLiver <-
read.csv("/Users/natalieschmer/Desktop/GitHub/stats_511/data/RatLiver.csv")
str(RatLiver)

## 'data.frame':    24 obs. of  2 variables:
##  $ Trt   : Factor w/ 2 levels "Ctrl","LowChr": 2 2 2 2 2 2 2 2 2 2 ...
##  $ Enzyme: num  44 48.5 50.7 45 53 52.7 51.8 49.8 48.3 55.5 ...
```

1A.Construct side-by-side boxplots of the data.

```
boxplot(Enzyme ~ Trt, data = RatLiver)
```



1B. Use the F-test to test for equality of variances. Give the null hypothesis, test statistic, p-value and conclusion. (4 pts)

F test for $H_0: \sigma_{ctrl}/\sigma_{LowChr} = 1$ vs $HA: \sigma_{ctrl}/\sigma_{LowChr} \neq 1$

```
var.test(Enzyme ~ Trt, data = RatLiver)

##
##  F test to compare two variances
##
## data:  Enzyme by Trt
## F = 0.78978, num df = 9, denom df = 13, p-value = 0.7373
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.2384571 3.0253182
## sample estimates:
## ratio of variances
##          0.7897775

#Or find by Hand
#break up data into two vectors then make calculations
Ctrl <- subset(RatLiver, Trt=="Ctrl")
LowChr <- subset(RatLiver, Trt=="LowChr")
```

The null hypothesis is that the true ratio of variances in enzymes relative to treatment is equal to 1. The F- statistic is 0.78, and the p = 0.73, and since F > p-value, we reject the null hypothesis that the true ratio of variances is equal to 1.

1C. Use Levene's test (with center="median") to test for equality of variances. Give the p-value and conclusion.

```
car::leveneTest(Enzyme ~ Trt, data = RatLiver)

## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value Pr(>F)
## group  1   0.176 0.6789
##       22
```

The p-value = 0.67, which is still less than the F value, so still reject the null hypothesis

1D. Based on your conclusions from parts B and C, would the pooled variance t-test or Welch-Satterthwaite t-test be preferred?

The Welch-Satterthwaite t-test would be preferred because the variances are not equal.

1E. Regardless of your answer to part D, run a 2sided two-sample t-test assuming equal variances. Give the null hypothesis, test statistic, p-value and conclusion. (4 pts)

$H_0: \mu_1 = \mu_2$

```
t.test(Enzyme ~ Trt, data = RatLiver)
```

```
##
##  Welch Two Sample t-test
##
## data:  Enzyme by Trt
## t = 2.2157, df = 20.84, p-value = 0.03798
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.2055579 6.5344421
## sample estimates:
##   mean in group Ctrl mean in group LowChr
##                52.87                49.50
```

The null hypothesis is that the true means of each treatment type are the same. The test statisitc is t = 2.2157, and p-value = 0.03798, and since p < 0.05, we reject the null hypothesis.

1F Rerun the analysis as a one-way ANOVA. Give the ANOVA table in your assignment. Compare your results to part E and notice that the p-value is the same and F = t2.

```
Fit = lm(Enzyme ~ Trt, data = RatLiver)
(Ftest = anova(Fit))
```

```
## Analysis of Variance Table
##
## Response: Enzyme
##           Df  Sum Sq Mean Sq F value Pr(>F)
## Trt        1  66.249  66.249  4.7127  0.041 *
## Residuals 22 309.261  14.057
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Ftest$statistic^2
```

```
## numeric(0)
```

```
Ftest$`F value`
```

```
## [1] 4.712747       NA
```

```
Ftest$`Pr(>F)`
```

```
## [1] 0.04100496         NA
```

2. Meningitis

2A. State the four assumptions for conducting a one-way analysis of variance to compare the population mean antigen concentrations for infants with late onset sepsis, infants with late onset meningitis, and asymptomatic infants. (4 pts)

The assumptions are that this is a random sample, there are independent observations, there are normally distributed residuals, and there is the equality of variances

2B Conduct a one-way analysis of variance test to determine if at least one of the infant groups has a different average antigen concentration from the others. Use a significance level of 0.05.

###i. State the hypotheses. Define the parameters. **The null hypothesis is that the mean antigen concentrations for infants with late onset sepsis, infants with late onset meningitis, and asymptomatic infants are equal, and the alternative is that they are not equal in that there is one or more difference. Rejection of H0 would be if F > F(alpha, df1, df2)**

2B ii Provide the missing components of the ANOVA table. Use R or calculator to plug appropriate summary statistics into equations from Slide 6 & Slide 8. Hint: (y_(..))‾ is weighted average of group ȳ's. (1 point for each SS's, dfs, and MS's, 6 pts total),

```
overall_mean <- (1.5 + 1.67 +1.14)/3

#Between
##SStrt
{
  group1t <- 19*(1.5 - overall_mean)^2
  group2t <- 22*(1.67 - overall_mean)^2
  group3t <- 18*(1.14- overall_mean)^2
}

SStrt <- sum(group1t, group2t, group3t)

#df = t-1
(dftrt <- 3-1)

## [1] 2

#MS
(MStrt <- SStrt / dftrt)

## [1] 1.429094
```

```r
#Within
##SS residual
{
  group1r <- (19 - 1)*0.12
  group2r <- (22 - 1)*0.18
  group3r <- (18 - 1)*0.04
}

SSres <- sum(group1r, group2r, group3r)

#df = nt - t
(dfres <- ((19 + 22 + 18) - 3))
```

```
## [1] 56
```

```r
#MS
(MSres <- SSres / dfres)
```

```
## [1] 0.1182143
```

2B iii Verify that your ratio of (s_B$^{2)/(s}$_W2) from values in the MS column above is relatively close to the F test statistic. Then provide a p-value from R for F = 12.02. (recall this is a "non-directional" alternative, p-value is area of F distribution to right of test static with appropriate degrees of freedom. Use the pf() function with correct df's)

```r
(ratio <- MStrt / MSres) #very close to F statistic, 12.08 > 12.02
```

```
## [1] 12.08902
```

```r
fcrit <- qf(0.95, 2, 56)
pf(fcrit, 2, 56, 12.02)
```

```
## [1] 0.1348321
```

2B iv Is it appropriate to conduct a follow-up analysis to determine which mean(s) significantly differ from the others? Explain in a single sentence.

It would not be appropriate since our calculated F is very very close to the given F

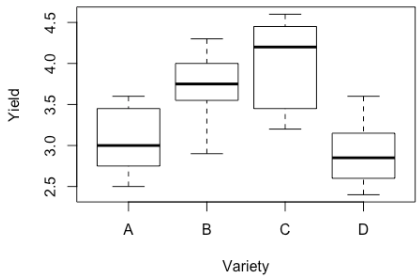
3. Corn Yield

```r
library(car)
library(dplyr)
library(emmeans)
Corn <-
read.csv("/Users/natalieschmer/Desktop/GitHub/stats_511/data/CornYield.csv")
str(Corn)
```

```
## 'data.frame':    32 obs. of  2 variables:
##  $ Variety: Factor w/ 4 levels "A","B","C","D": 1 1 1 1 1 1 1 1 2 2 ...
##  $ Yield  : num  2.5 3.6 2.8 2.7 3.1 3.4 2.9 3.5 3.6 3.9 ...
```

3A.Provide a boxplot of the corn yields by the four treatment levels. Also, provide summary statistics (sample size, mean, spread, SE) for each treatment level. There is the summarise() function from the dplyr package from the example or there are certainly many other functions and packages that can do this.

```
boxplot(Yield ~ Variety, data = Corn)
```




```
Corn %>%
  group_by(Variety) %>%
  summarise(n = n(),
            mean = mean(Yield),
            sd = sd(Yield),
            se = sd / sqrt(n))
```

```
## # A tibble: 4 x 5
##   Variety     n  mean    sd    se
##   <fct>   <int> <dbl> <dbl> <dbl>
## 1 A           8  3.06 0.403 0.143
## 2 B           8  3.72 0.423 0.150
## 3 C           8  4    0.550 0.195
## 4 D           8  2.9  0.393 0.139
```

3B. Looking at the plot and summary statistics, do you expect any problems with the assumptions for performing an ANOVA? If assumptions are met, do you expect a “large” F-statistic (and a “small” p-value) when performing an ANOVA? (For these questions, there is no need to justify responses to both. The intent is to serve as a reminder to simply look at the summary stats and plots and think about expected outcomes.)

Some boxplots have overlap but some do not, so there may be issues

3C. Carry out a one-way ANOVA to determine whether there is a significant difference (using α=0.05) in the mean yield for the different varieties. State the null hypothesis, give the F test statistic, p-value and conclusion. (4 pts)

```
corn = lm(Yield ~ Variety, data = Corn)
(Ftest = anova(corn))
```

```
## Analysis of Variance Table
##
## Response: Yield
##           Df Sum Sq Mean Sq F value   Pr(>F)
## Variety    3 6.6209 2.20698  11.047 5.85e-05 ***
## Residuals 28 5.5938 0.19978
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The null hypothesis is that the mean yields for each variety is equal. The F = 11.047 and p = 5.85e-05. Since F >> p, we reject the null hypothesis

3D. Run (unadjusted) pairwise comparisons of means. State the pairs that appear to be significantly different. Hint: The p-values for comparing group means should be consistent with expectations from examining the box plot from part (A)

```
(corn_3d <- emmeans::emmeans(corn, pairwise ~ Variety, adjust = "none"))
```

```
## $emmeans
##  Variety emmean    SE df lower.CL upper.CL
##  A         3.06 0.158 28     2.74     3.39
##  B         3.73 0.158 28     3.40     4.05
##  C         4.00 0.158 28     3.68     4.32
##  D         2.90 0.158 28     2.58     3.22
##
## Confidence level used: 0.95
##
## $contrasts
##  contrast estimate    SE df t.ratio p.value
##  A - B      -0.662 0.223 28 -2.964  0.0061
##  A - C      -0.938 0.223 28 -4.195  0.0002
##  A - D       0.163 0.223 28  0.727  0.4732
##  B - C      -0.275 0.223 28 -1.231  0.2287
##  B - D       0.825 0.223 28  3.692  0.0010
##  C - D       1.100 0.223 28  4.922  <.0001
```

```
corn_3d
```

```
## $emmeans
##  Variety emmean    SE df lower.CL upper.CL
##  A         3.06 0.158 28     2.74     3.39
##  B         3.73 0.158 28     3.40     4.05
##  C         4.00 0.158 28     3.68     4.32
```

```
##  D         2.90 0.158 28     2.58     3.22
##
## Confidence level used: 0.95
##
## $contrasts
##  contrast estimate    SE df t.ratio p.value
##  A - B      -0.662 0.223 28 -2.964  0.0061
##  A - C      -0.938 0.223 28 -4.195  0.0002
##  A - D       0.163 0.223 28  0.727  0.4732
##  B - C      -0.275 0.223 28 -1.231  0.2287
##  B - D       0.825 0.223 28  3.692  0.0010
##  C - D       1.100 0.223 28  4.922  <.0001
```

The pairs that seem to be significantly different are A and B (p = 0.0061), A and C (p = 0.0002), B and D (p = 0.0010), and C and D (<.0001). This is also consistent with the boxplots, in that there is little overlap between plots.