Stat511 Hw6

Amy Fox

10/14/2019

```
library(readr)
library(dplyr)
library(ggplot2)
library(car)
library(emmeans)
library(multcomp)
```

Question 1

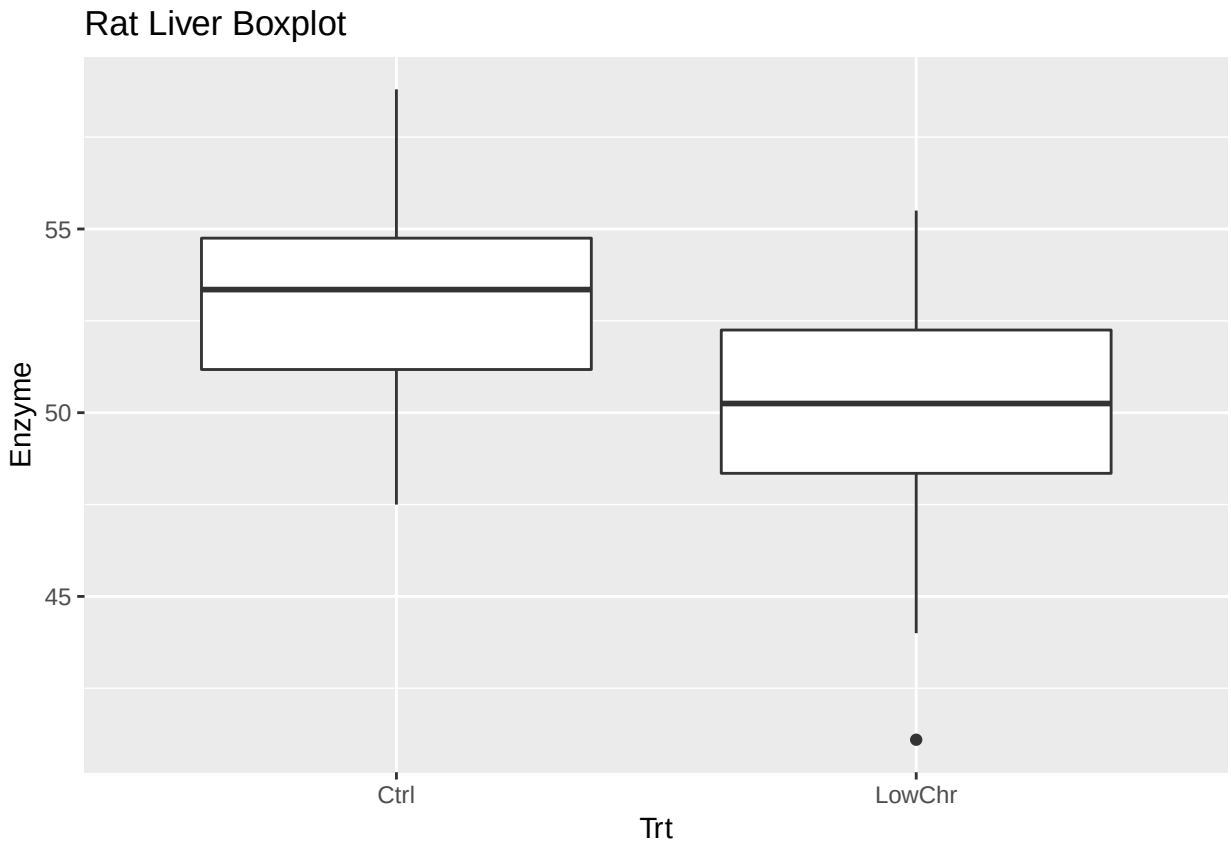
In an investigation of the possible influence of dietary chromium on diabetic symptoms, 14 rats were fed a low-chromium diet and 10 were fed a control diet. One response variable was activity of the liver enzyme GITH. The data is available as "RatLiver.csv".

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

```
rat_liver_data <- read_csv("../Data/RatLiver.csv") %>%
  mutate(Trt = as.factor(Trt))
```

```
## Parsed with column specification:
## cols(
##   Trt = col_character(),
##   Enzyme = col_double()
## )
```

```
ggplot(rat_liver_data, aes(x = Trt, y = Enzyme)) +
  geom_boxplot() +
  ggtitle("Rat Liver Boxplot")
```



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

Null Hypothesis

H0: $\sigma_1{}^2 = \sigma_2{}^2$

HA: $\sigma_1{}^2 \neq \sigma_2{}^2$

df1 = Ctrl = 10

df2 = LowChr = 14

```
rat_liver_data %>%
  group_by(Trt) %>%
  summarise(sd = sd(Enzyme))
```

```
## # A tibble: 2 x 2
##   Trt       sd
##   <fct>  <dbl>
## 1 Ctrl    3.49
## 2 LowChr  3.92
```

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

```
##
##  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
```

Test statistic: 0.78978

The p-value = $0.74 > 0.05 = \alpha$. Therefore, we fail to reject H0. We can thus conclude that the true variances are most likely equal.

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

```
leveneTest(Enzyme ~ Trt, data = rat_liver_data, center = "median")
```

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

The p-value is 0.68. Because the p-value = $0.68 > 0.05 = \alpha$ we fail to reject H0. Therefore, the true variances are most likely equal.

D. Based on your conclusions from the two previous questions, would the pooled variance t-test or Welch-Satterthwaite t-test be preferred?

Because the two previous questions have shown that the true variances are most likely equal, a pooled variance test would be preferred. (The Welch-Satterthwaite is used in cases of unequal variances.)

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

Null Hypothesis

H0: $\mu_{\text{Ctrl}} = \mu_{\text{LowChr}}$

HA: $\mu_{\text{Ctrl}} \neq \mu_{\text{LowChr}}$

```
t.test(Enzyme ~ Trt, data = rat_liver_data, alternative = "two.sided", var.equal = TRUE)
```

```
##
##  Two Sample t-test
##
## data:  Enzyme by Trt
## t = 2.1709, df = 22, p-value = 0.041
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.1505995 6.5894005
## sample estimates:
##   mean in group Ctrl mean in group LowChr
##                52.87                49.50
```

Test statistic: 2.17

p-value: 0.041

Conclusion: Because p-value = $0.041 < 0.05 = \alpha$, we reject H0. Thus, we conclude that the true difference in means is not equal to 0.

F. Rerun the analysis as a one-way ANOVA. Give the ANOVA table in your assignment. Compare your results to the previous question and notice that the p-value is the same and F

= t2.

```
library(broom)
aov(Enzyme ~ Trt, data = rat_liver_data)
```

```
## Call:
##    aov(formula = Enzyme ~ Trt, data = rat_liver_data)
##
## Terms:
##                       Trt Residuals
## Sum of Squares   66.24858 309.26100
## Deg. of Freedom         1        22
##
## Residual standard error: 3.749309
## Estimated effects may be unbalanced
```

```
# or
lm(Enzyme ~ Trt, data = rat_liver_data) %>%
  anova() %>%
  tidy()
```

```
## # A tibble: 2 x 6
##   term         df sumsq meansq statistic p.value
##   <chr>     <int> <dbl>  <dbl>     <dbl>   <dbl>
## 1 Trt           1  66.2   66.2      4.71  0.0410
## 2 Residuals    22 309.    14.1     NA    NA
```

The p-values are the same between the ANOVA and two-sample t test. Further, the F statistic in this anova table is equal to t-statistic^2 in the t-test.

$F = t^2$

$4.71 = 2.17\hat{}2$

Question 2

Read Problem 8.32 which concerns corn yield. The data is available as "CornYield.csv".

A. Construct a bar plot showing means and SEs for each variety. (4 pts)

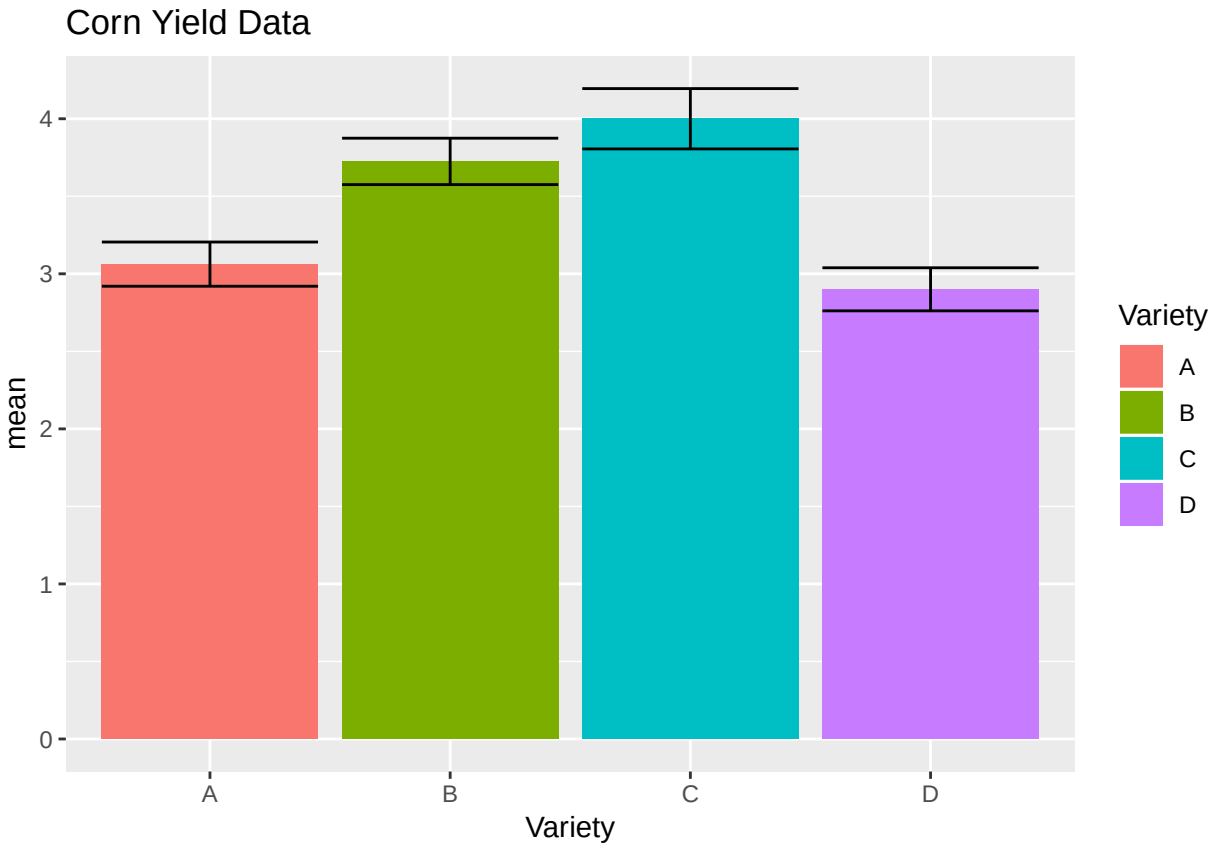
```
corn_data <- read_csv("../Data/CornYield.csv")
```

```
## Parsed with column specification:
## cols(
##   Variety = col_character(),
##   Yield = col_double()
## )
```

```
Sum_stats <- corn_data %>%
  group_by(Variety) %>%
  summarise(n  = n(),
            mean = mean(Yield),
            sd = sd(Yield),
            se = sd/sqrt(n))

ggplot(Sum_stats, aes(x = Variety, y = mean, fill = Variety)) +
  geom_col() +
```

```
  geom_errorbar(aes(ymin = mean - se, ymax = mean + se)) +
  ggtitle("Corn Yield Data")
```



B. Carry out a one-way ANOVA analysis to determine whether there is evidence of differences (using $\alpha = 0.05$) in the mean yield for the different varieties. State the null hypothesis, give the test statistic, p-value and conclusion. (4 pts)

Null Hypothesis

H0: $\mu_A = \mu_B = \mu_C = \mu_D$

HA: there is one or more difference (not all of the μ are equal)

```
lm(Yield ~ Variety, data = corn_data) %>%
  anova()
```

```
## 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
```

Test Statistic: 11.05

p-value: 5.85×10^{-5}

Conclusion: Because the p-value is $< \alpha$, we reject H0. Thus, there is at least one difference between the true means.

C. Run (unadjusted) pairwise comparisons of means. Give the estimated difference and p-value for each comparison. (4 pts)

```
lm(Yield ~ Variety, data = corn_data) %>%
  emmeans(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
```

The estimated differences and p-values are in the $contast part of the table in the columns "estimate" and "p-value" respectively.

D. Calculate the LSD(0.05) value. Recall that this is the 95% ME for pairwise comparisons of means.

S_W = sqrt(MSResid) = sqrt(0.19978) (taken from anova analysis)

$LSD_{0.05}$ = $t_{\alpha/2}$ x S_W x sqrt(2/n) $t_{0.05/2}$ x sqrt(0.19978) x sqrt(2/8)

```
df_resid <- sum(Sum_stats$n) - nrow(Sum_stats)
t_alpha <- qt(1-0.05/2, df = df_resid)

t_alpha * sqrt(0.19978) * sqrt(2/8)
```

```
## [1] 0.4577858
```

The LSD(0.05) is 0.458

E. Construct an (unadjusted) "cld" display including the mean for each variety and assigning number groups (or underlining) varieties that are not "significantly" different. (4 pts)

```
corn_lm <- lm(Yield ~ Variety, data = corn_data)
corn_pairwise <-  emmeans(corn_lm, pairwise ~ Variety,  adjust = "none")

CLD(corn_pairwise$emmeans, adjust = "none")
```

```
##  Variety emmean    SE df lower.CL upper.CL .group
##  D         2.90 0.158 28     2.58     3.22  1
##  A         3.06 0.158 28     2.74     3.39  1
##  B         3.73 0.158 28     3.40     4.05   2
##  C         4.00 0.158 28     3.68     4.32   2
##
## Confidence level used: 0.95
## significance level used: alpha = 0.05
```

F. Summarize your findings from parts C and E.

The findings from C and E agree that there is not a difference in means between varieties D-A and varietes B-C.

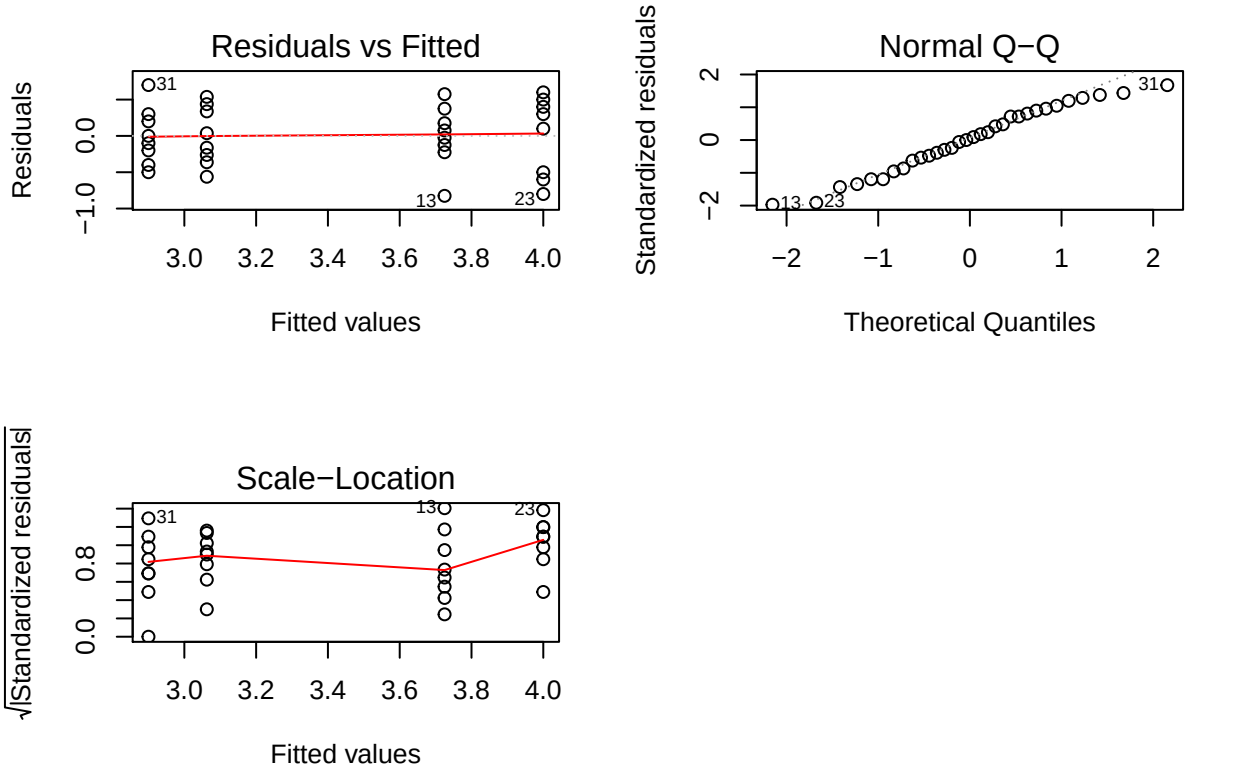
Using part C, the only p-values greater than 0.05 are the ones comparing D to A and comparing B to C. This means that we fail to reject H0 and assume that there is not a difference between means between these groups.

Using CLD in part E, we can see there there is not a "significant" difference (less than $\alpha = 0.05$) between 1) variety D and A and between 2) variety B and C.

B & C have heigher yields than D & A based on the boxplots and this data.

G. Use the plot() function to generate the diagnostic plots from the model from part B. You do not have to include the graphs in your assignment, but discuss the plots of (1) Residuals vs Fitted values and (2) qqplot of residuals and whether assumptions appear to be satisfied based on each plot. (4 pts)

```
par(mfrow = c(2,2))
plot(corn_lm)
```



Based on the residuals vs. fitted values plot, the scatter looks about equal, which supports the assumption of equal variances. (data has equal variances)

Based on the qqplot of residuals, it looks like most of the points fall on the linear line, therefore, this supports the assumption of normality. (data is normal)