Stat631 Midterm 2

Brandon Keck (netID qh9701)

1. (a) BK
2. (b) BK

```
library(car)
library(emmeans)
library(tidyverse)
library(readr)
```

Uploading various datasets

```
frog <- read.table("frog.txt")
head(frog)
```

```
   en    species
1 2.1 nepalensis
2 1.8 nepalensis
3 2.4 nepalensis
4 1.9 nepalensis
5 1.7 nepalensis
6 2.3 nepalensis
```

```
str(frog)
```

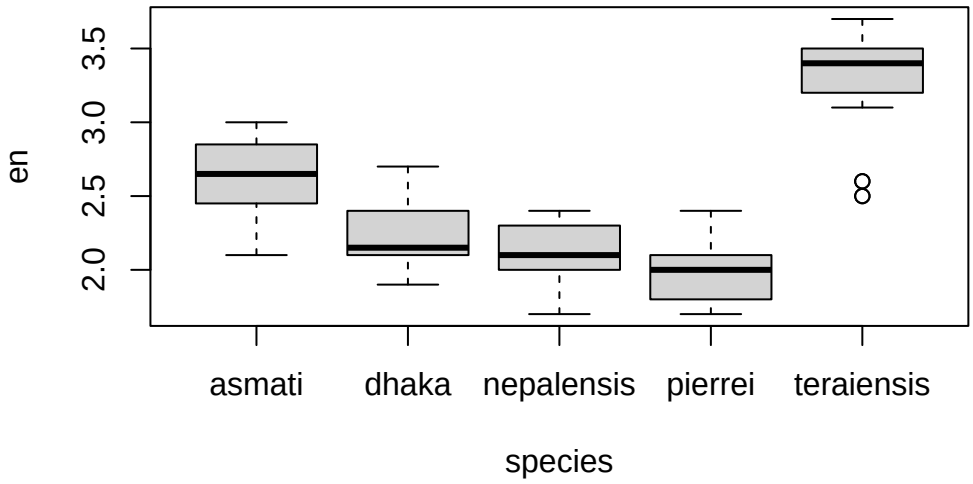
```
'data.frame':   59 obs. of  2 variables:
 $ en     : num  2.1 1.8 2.4 1.9 1.7 2.3 2.1 2.3 2.2 2.3 ...
 $ species: chr  "nepalensis" "nepalensis" "nepalensis" "nepalensis" ...
```

2.

(2a)

```
fit <- aov(en ~ species, data = frog)

boxplot(en ~ species, data = frog)
```



(2b)

From the boxplot we see that the species dhaka, nepalensis, pierrei are fairly similar. However, the species teraiensis's median en is the highest out of all of the species. Teraiensis also has two noticeable outliers around 2.5 en. Asmati has the second highest median around 2.6 en.

(2c)

```r
anova(fit)
```

```
Analysis of Variance Table

Response: en
          Df  Sum Sq Mean Sq F value    Pr(>F)
species    4 16.1235  4.0309  41.836 6.146e-16 ***
Residuals 54  5.2029  0.0964
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(2d)

$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$

vs

$H_A: at least one species mean en is different from the rest.

From the anova output we see a very small p-value of 6.146e-16 which is essentially 0. We also have an F-statistic of 41.836. From this we reject the null hypothesis and can conclude that there is a difference is mean en between at least one of the species of frogs.

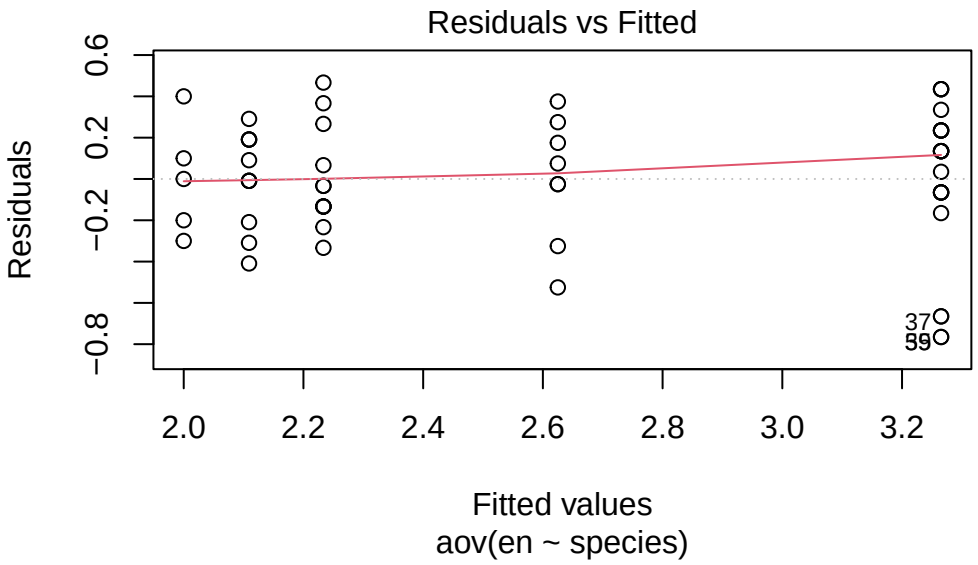
(2e)

(i)

The assumption of ANOVA that I would check with a Residuals vs Fitted Values plot is Equal variance. We are checking to ensure that the distribution of the residuals over the fitted values has randomization in the spread. That there is no pattern or fanning.

(ii)

```r
plot(fit, which = 1)
```




```
leveneTest(en ~ species, data = frog)
```

```
Warning in leveneTest.default(y = y, group = group, ...): group coerced to
factor.
```

```
Levene's Test for Homogeneity of Variance (center = median)
      Df F value Pr(>F)
group  4  0.4406 0.7787
      54
```

From the plot it seems that the equal variance assumption for ANOVA is satisfied. We can also check with the levene test for equal variance. With a p-value of 0.7787 which is larger than our significance level of 0.05 we can conclude equal variance.

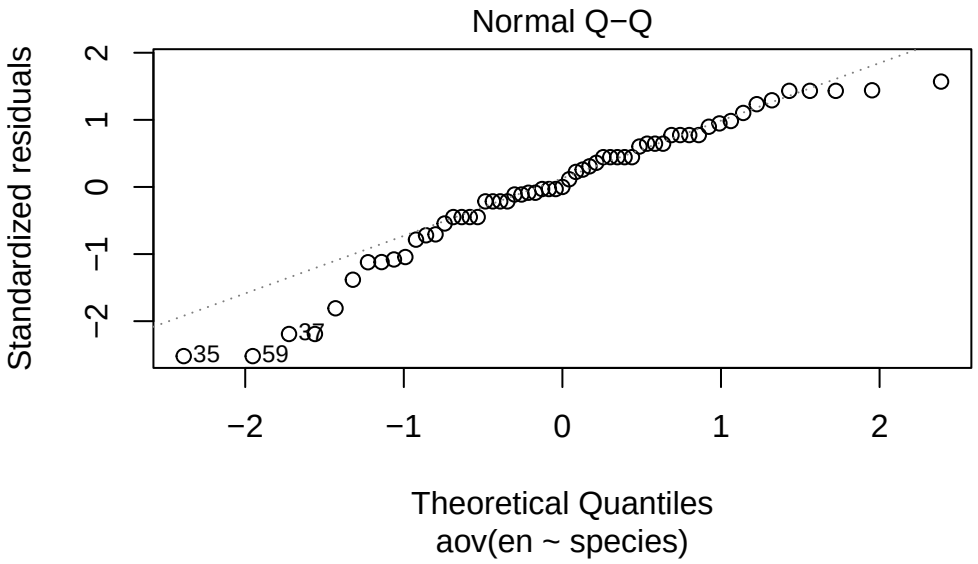
(2f)

(i)

The assumption of ANOVA I would be checking is the Normality of the residuals. Here we are checking to see if the residuals follow the normal QQ line to ensure that the Normality Assumption is satisfied.

(ii)

```
plot(fit, which = 2)
```




```
shapiro.test(fit$residuals)
```

```
    Shapiro-Wilk normality test

data:  fit$residuals
W = 0.94636, p-value = 0.01144
```

From the Normal QQ plot we witness major deviation towards the left side and there is some slight deviation of the right side of the plot. This means that the Normality assumption is not satisfied. We can double check with the shapior wilk test. WIth a p-value of 0.01144 which is less than our significance level of 0.05 we reject H_0 meaning that the data deviates from Normality.

(2g)

In order to proceed with the ANOVA test the other assumption needed is Independence within and between groups.

(2h)

```
ptf <- powerTransform(fit)
summary(ptf)
```

```
bcPower Transformation to Normality
   Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
Y1    0.2402           1      -0.7343       1.2148

Likelihood ratio test that transformation parameter is equal to 0
 (log transformation)
                            LRT df    pval
LR test, lambda = (0) 0.2344511  1 0.62824

Likelihood ratio test that no transformation is needed
                           LRT df    pval
LR test, lambda = (1) 2.282405  1 0.13085
```

```
fit$y_new <- bcPower(fit$en, ptf$roundlam)

#newfit <- aov(y_new ~ species, data = fit)
#anova(newfit)
```

(2i)

```
#lsmFrogs <- lsmeans(fit, ~ en)
#summary(contrast(lsmFrods, method = "pairwise", adjust = "tukey"), infer = c(T,T))
```

(2j)

```
#contrast(lsmFrogs, method = "trtvsctrl", adjust = "mvt", ref = 5,
        # infer = c(T,F))
```

(2k)

```
wi = c(1/4, 1/4, 1/4, 1/4, -1)

lsmfrog <- emmeans(fit, ~ species)
contrast(lsmfrog, method = list(ctrts = wi), infer = c(T,T))
```

```
 contrast estimate     SE df lower.CL upper.CL t.ratio p.value
 ctrts       -1.02 0.0848 54    -1.19   -0.853 -12.064  <.0001

Confidence level used: 0.95
```

(2l)

```
kruskal.test(en ~ species, data = frog)
```

```
    Kruskal-Wallis rank sum test

data:  en by species
Kruskal-Wallis chi-squared = 42.293, df = 4, p-value = 1.45e-08
```

Since the p-value is less than our significance level of 0.05 we reject the null hypothesis. This means that at least one species has a significantly different median than the others.

(2m)

What I can conclude from the non-parametric test is that at least one species has a significantly different median than that of the others. This agrees with my results in part (c)

3.

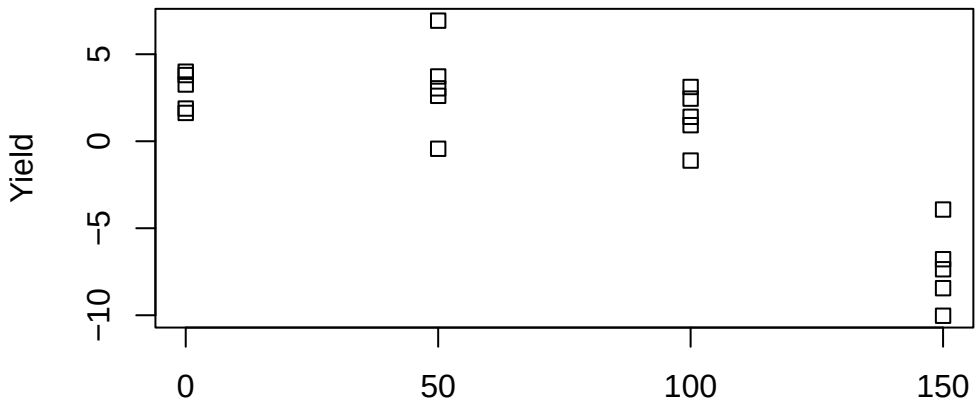
```
fertilizer <- read.table("fertilizer.txt")
head(fertilizer)
```

```
  Fertilizer     Yield
1          0  1.879049
2         50  3.039645
3        100  3.117417
4        150 -7.358983
5          0  3.258575
6         50  6.930130
```

(3a)

```
fit1 <- aov(Yield ~ Fertilizer, data = fertilizer)

stripchart(Yield ~ Fertilizer, data = fertilizer, vertical = T)
```



From the stripchart the variability between the first three levels of fertilizer 0, 50, and 100 all being approximately the same. While the fertilizer with 150kg/ha has the lowest yield of them all.

(3b)

```
anova(fit1)
```

```
Analysis of Variance Table

Response: Yield
           Df Sum Sq Mean Sq F value    Pr(>F)
Fertilizer  1 263.72 263.722   28.12 4.844e-05 ***
Residuals  18 168.81   9.379
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the ANOVA table we get a p-value of 4.844e-05 which is less than our significance level of 0.05. We also have a large F-statistic of 28.12. We have enough evidence to conclude that there is a difference in mean between crop yield between the four fertilizers.

(3c)

Since there are g = 4 treatments, the highest order we can consider at first is g - 1 =3.

(3d)

```
p0 <- lm(Yield ~ 1, data = fertilizer)
p1 <- lm(Yield ~ Fertilizer, data = fertilizer)
p2 <- lm(Yield ~ poly(Fertilizer, 2, raw = TRUE), data = fertilizer)
p3 <- lm(Yield ~ poly(Fertilizer, 3, raw = TRUE), data = fertilizer)
anova(p0,p1,p2,p3)
```

```
Analysis of Variance Table

Model 1: Yield ~ 1
Model 2: Yield ~ Fertilizer
Model 3: Yield ~ poly(Fertilizer, 2, raw = TRUE)
```

```
Model 4: Yield ~ poly(Fertilizer, 3, raw = TRUE)
  Res.Df    RSS Df Sum of Sq       F    Pr(>F)
1     19 432.54
2     18 168.81  1   263.722 66.3859  4.37e-07 ***
3     17  69.24  1    99.570 25.0644 0.0001292 ***
4     16  63.56  1     5.682  1.4304 0.2491180
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The ANOVA shows that terms up to $Fertilizer^2$ are needed.

(3e)

The polynomial (orthogonal) contrast can be used, because the quantitative treatment levels are all equally spaced and the sample sizes are all the same.

(3f)

```r
fertilizer$fFertilizer <- as.factor(fertilizer$Fertilizer)

mod <- aov(Yield ~ fFertilizer, data = fertilizer)

lsmeans(mod, poly ~ fFertilizer)
```

```
$lsmeans
 fFertilizer lsmean    SE df lower.CL upper.CL
 0             2.91 0.891 16    1.023     4.80
 50            3.17 0.891 16    1.284     5.06
 100           1.36 0.891 16   -0.534     3.25
 150          -7.31 0.891 16   -9.198    -5.42

Confidence level used: 0.95

$contrasts
 contrast  estimate   SE df t.ratio p.value
 linear      -32.48 3.99 16  -8.148  <.0001
 quadratic    -8.93 1.78 16  -5.006  0.0001
 cubic        -4.77 3.99 16  -1.196  0.2491
```

The quadratic model with first order is good enough for this dataset, since cubic p-value of the tests for that term are not significant. The highest order is 2. This confirms our finding earlier in part (d).

(3g)

```
mod2 <- lm(Yield ~ Fertilizer, data = fertilizer)
predict(mod2, data.frame(Fertilizer = c(75, 125)))
```

```
        1          2
0.0332476 -3.2146590
```

When the percentages of fertilizer used in the blend of materials is 75% and 125% the predicted mean yield are 0.0332 and -3.215 respectively.

(3h)

It is not reasonable to predict for 200kg/ha with the current model because from the earlier provided stripchart we can see that as the percentage of fertilizer is added to the crops the amount of yield also declines. From the previous prediction of 125kg/ha we see that we even start to loose crop yields. It would appear that anything from (0, 75kg/ha) is appropriate in terms of mean crop yield.