Elliot Smith
STAT 616 - HW #2

Problem 1

a. Effects Model: $y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$

Using Dummy Coding:

$$\begin{array}{cc} & \begin{array}{cc} \mu & \alpha_2 \end{array} \\ y \begin{array}{c} M \\ F \end{array} & \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix} \end{array} \qquad \alpha_1 = 0$$

So: $y = \mu + \alpha_2 x_2$

$\rightarrow \hat{\mu} = \bar{y}_{..}$

$\rightarrow \hat{\alpha}_i = \bar{y}_{i.} - \hat{\mu} = \bar{y}_{i.} - \bar{y}_{..}$

b. Effects Model - Dummy Coding

$$\begin{array}{cc} & \begin{array}{cc} \mu & \alpha_2 \end{array} \\ y \begin{array}{c} M \\ F \end{array} & \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix} \end{array}$$

$\hat{\alpha}_1 = \bar{y}_{1.} - \hat{\mu} = \bar{y}_{1.} - \frac{1}{2}(\bar{y}_{1.} + \bar{y}_{2.})$

$\hat{\alpha}_2 = \bar{y}_{2.} - \hat{\mu} = \bar{y}_{2.} - \frac{1}{2}(\bar{y}_{1.} + \bar{y}_{2.})$

$\hat{\mu} = \frac{1}{2}(\bar{y}_{1.} + \bar{y}_{2.})$

$$\hat{\sigma}^2_{LS} = \frac{\sum_i \sum_j (y_{ij} - \bar{y}_{i.})^2}{N-k}$$

k = # of treatments

$N = n_1 + n_2$

$SE(\text{intercept}) \rightarrow Var(\text{intercept}) = Var(\hat{\mu}) = Var(\bar{y}_{1.}) = \frac{\sigma^2}{n}$

So $\hat{SE}(\text{intercept}) = \frac{\hat{\sigma}}{\sqrt{n}}$

$SE(\hat{\alpha}_1) \rightarrow Var(\hat{\alpha}_1) = \bar{y}_{1.} - \bar{y}_{1.} = Var(\bar{y}_{1.}) + Var(\bar{y}_{1.}) = \frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_1} = \frac{2\sigma^2}{n}$

So: $SE(\hat{\alpha}_1) = \frac{\sqrt{2}}{\sqrt{n}}\hat{\sigma}$

$SE(\hat{\alpha}_2) \rightarrow Var(\hat{\alpha}_2) = \bar{y}_{1.} - \bar{y}_{2.} = Var(\bar{y}_{1.}) + Var(\bar{y}_{2.}) = \frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2} = \frac{2\sigma^2}{n}$

So: $SE(\hat{\alpha}_2) = \frac{\sqrt{2}}{\sqrt{n}}\hat{\sigma}$

Since a model is defined as $E[Y \mid X_2]$ (since μ takes on X_1 when $X_1 = 0$) we should exactly n_2 fitted values.

c. Effects Model - Zero Sum Coding

$\alpha_1 + \alpha_2 = 0 \qquad \alpha_1 = -\alpha_2$

$$y \quad \begin{array}{c|cc} & \mu & \alpha_1 \\ \hline M & 1 & -1 \\ F & 1 & 1 \end{array}$$

So: $y = \mu + \alpha_1 x_1 - \alpha_2 x_2$

$\rightarrow \hat{\mu} = \frac{1}{2}(\bar{y}_{1\cdot} + \bar{y}_{2\cdot})$

$\rightarrow \hat{\alpha}_1 = \bar{y}_{1\cdot} - \hat{\mu} = \bar{y}_{1\cdot} - \frac{1}{2}(\bar{y}_{1\cdot} + \bar{y}_{2\cdot})$

$\rightarrow \hat{\alpha}_2 = \bar{y}_{2\cdot} - \hat{\mu} = \bar{y}_{2\cdot} - \frac{1}{2}(\bar{y}_{1\cdot} + \bar{y}_{2\cdot})$

$$\hat{\sigma}_{LS}^2 = \frac{\sum_i \sum_j (y_{ij} - \bar{y}_{i\cdot})^2}{N-k}$$

$k = \#$ of treatments

$N = n_1 + n_2$

$\hat{\mu} = \bar{y}_{\cdot\cdot} \quad SE(\hat{\mu}) = \sqrt{\frac{\sigma^2}{N}} = \sqrt{\frac{\sigma^2}{kn}}$

$Var(\hat{\mu}) = Var(\bar{y}_{\cdot\cdot}) = \frac{\sigma^2}{N}$

$\hat{\alpha}_i = \bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot} \quad SE(\hat{\alpha}_i) = \sqrt{\frac{2}{3}\frac{\sigma^2}{n}}$

$Var(\hat{\alpha}_i) = Var(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot}) = Var(\bar{y}_{i\cdot}) + Var(\bar{y}_{\cdot\cdot}) - 2cov(\bar{y}_{i\cdot}, \bar{y}_{\cdot\cdot})$

$= \frac{\sigma^2}{n} + \frac{\sigma^2}{N} - \frac{2}{3}\frac{\sigma^2}{n}$

Since a model is defined as $E[Y|X_1, X_2]$ we should have exactly n_1 values if $n_1 < n_2$.

d. The estimates are different because of the coding. For example, with dummy coding, we establish a baseline β_1 meaning that we set this estimator to zero and interpret all results based on this "new zero". On the contrary, with zero-sum coding, we set a different baseline, with the sum of all the estimators equal to zero and thus establishing a different "new zero". In summary, different coding methods decide on a particular baseline, which determines the value of the estimators.

Scanned by CamScanner

e. To find the MLEs, perform the following steps:

① Find the joint probability function of the data

② Simplify with logs

③ Maximize function with respect to μ_i, σ^2 $\left(\frac{\partial l}{\partial \mu_i} = 0, \frac{\partial l}{\partial \sigma^2} = 0\right)$

The MLEs come out to:

$$\hat{\mu}_i = \bar{y}_{i\cdot}$$

$$\hat{\sigma}_{ML}^2 = \frac{\sum_i \sum_j (y_{ij} - \bar{y}_{i\cdot})^2}{N}$$

$y_{ij} = \mu_i + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim N(0, \sigma^2)$

① pdf: $f(\underset{\sim}{y}; \mu_1, \mu_2, \sigma^2) = \prod_{ij} f(y_{ij}; \underset{\sim}{\mu}, \sigma^2)$

$= \prod_{ij} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(y_{ij} - \mu_i)^2}$

② $f(\underset{\sim}{y}) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^{\frac{N}{2}} e^{-\frac{1}{2\sigma^2}\sum_{ij}(y_{ij} - \mu_i)^2}$

$\log[f(\underset{\sim}{y})] = -\frac{N}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{ij}(y_{ij} - \mu_i)^2$

③ $\underset{\mu_i, \sigma^2}{\arg\max}\ l(\underset{\sim}{\mu}, \sigma^2; \underset{\sim}{y})$

$\frac{\partial l}{\partial \mu_i} = 0, \quad \frac{\partial l}{\partial \sigma^2} = 0$

$\hat{\sigma}_{LS}^2 = \frac{\sum_i \sum_j (y_{ij} - \bar{y}_{i\cdot})^2}{N - k} \longrightarrow$ unbiased $\quad E[\hat{\sigma}^2_{LS}] = \sigma^2 \quad$ * k = # of μ's

$\hat{\sigma}^2_{ML} = \frac{\sum_i \sum_j (y_{ij} - \bar{y}_{i\cdot})^2}{N} \longrightarrow$ biased $\quad E[\hat{\sigma}^2_{ML}] \neq \sigma^2$

As $N \gg k$, these will converge!

Problem 2

a.

Source	SS	df	MS	F-ratio
Between (SS_B)	258	4	64.5	10.184
Within (SS_W)	76	12	6.333	N/A
Total (SS_T)	334	16		

We may conclude that $\frac{258}{334} = \sim 77\%$ of the variance is explained by the group-to-group variance

Since the F-statistic is significantly greater than 1, we may conclude that at least one of the population means differs from the others.

Based on only the F-ratio, we cannot say if the variation is statistically significant. We will first want to design a hypothesis test and calculate a p-value to test our hypothesis of all equal means.

b. I tested my null hypothesis of equal means and got a p-value of 0.0008 and thus rejected my null hypothesis. My methodology was to randomly permute the labels of the data and then generate an F-statistic for each run through (I performed 5000). To calculate my p-value I found the number of generated F-statistics that were greater or equal to my observed F-statistic (there were 4). I then divided this number by the total number of generated F-statistics (5000) to obtain a p-value of 0.0008 and rejected my null hypothesis.

*See code appendix

c. Please see my code appendix for my construction of the model and the accompanying plots.

Based on my QQ Plot and Histogram of the residuals, I conclude that the errors are mostly normal. The QQ Plot shows that the residuals mostly follow the line, telling us that they are normally distributed. The histogram as well shows us that the residuals are mostly normal as the middle value is tallest and the fringe values tend to be less represented as we approach the extremes.

Based on my test for equal variances, I conclude this test passes as well. As we can see, the residual values are equally and fairly distributed around the zero-axis line, showing that the variances are equally distributed and mostly equal.

I conclude that all tests pass.

d. The ANOVA table shows us the sources and contributions of variance and whether or not, through hypothesis testing using the F-statistic, all of the means are equal; it does NOT tell us which groups are different. To answer this question, we will fit an ANOVA model using regression. So in summary, the ANOVA table will tell us if all means are equal and then our ANOVA model via regression may tell us which means are not equal.

e. To compare groups B and C and B and E, we will use a t-test. Please see my code appendix for R calculations

B and C	B and E
t-statistic: 0.7071	t-statistic: -2.1669
df = 5	df = 5
p-value = 0.5123	p-value = 0.0983
I will NOT reject $H_0: \mu_B = \mu_C$ in favor of $H_1: \mu_B > \mu_C$	I will NOT reject $H_0: \mu_B = \mu_E$ in favor of $H_1: \mu_B > \mu_E$

STAT 616

Problem 1

After some trial and error, the lowest possible p-value that I could obtain was ~ 0.02 (see code appendix for values chosen). I am not entirely sure what this suggest, however, a fair consideration might be that, with a certain confidence level, we may not reject H_0 with a p-value of 0.02 (for example if $\alpha = 0.01$). So we may say that having only 2 observations for each of the 3 treatments is not enough and we must add more for best results.

Code Appendix

Elliot Smith

2/13/2018

```
options(scipen = 999)
```

Problem 2

Part b

```
# Problem 2

## Part b

values <- c(1,3,5,9,5,5,5,6,6,3,3,3,0,6,14,10,18)
labels <- c("A","A","A","B","B","B","B","C","C","C","D","D","D","D","E","E","E")

data <- data.frame(labels, values)

aov_out <- aov(values ~ labels, data = data)
summary(aov_out)
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## labels       4    258   64.50   10.18 0.000782 ***
## Residuals   12     76    6.33
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
f_stat <- as.vector(summary.lm(aov_out)$fstatistic["value"])



f_stats <- numeric()
n <- 5000

for (i in 1:n) {

    labels_temp <- sample(labels)
    data_temp <- data.frame(labels_temp, values)
    aov_out_temp <- aov(values ~ labels_temp, data = data_temp)
    f_stats[i] <- as.vector(summary.lm(aov_out_temp)$fstatistic["value"])

}

count_greater <- sum(f_stats >= f_stat)

count_greater/n
```

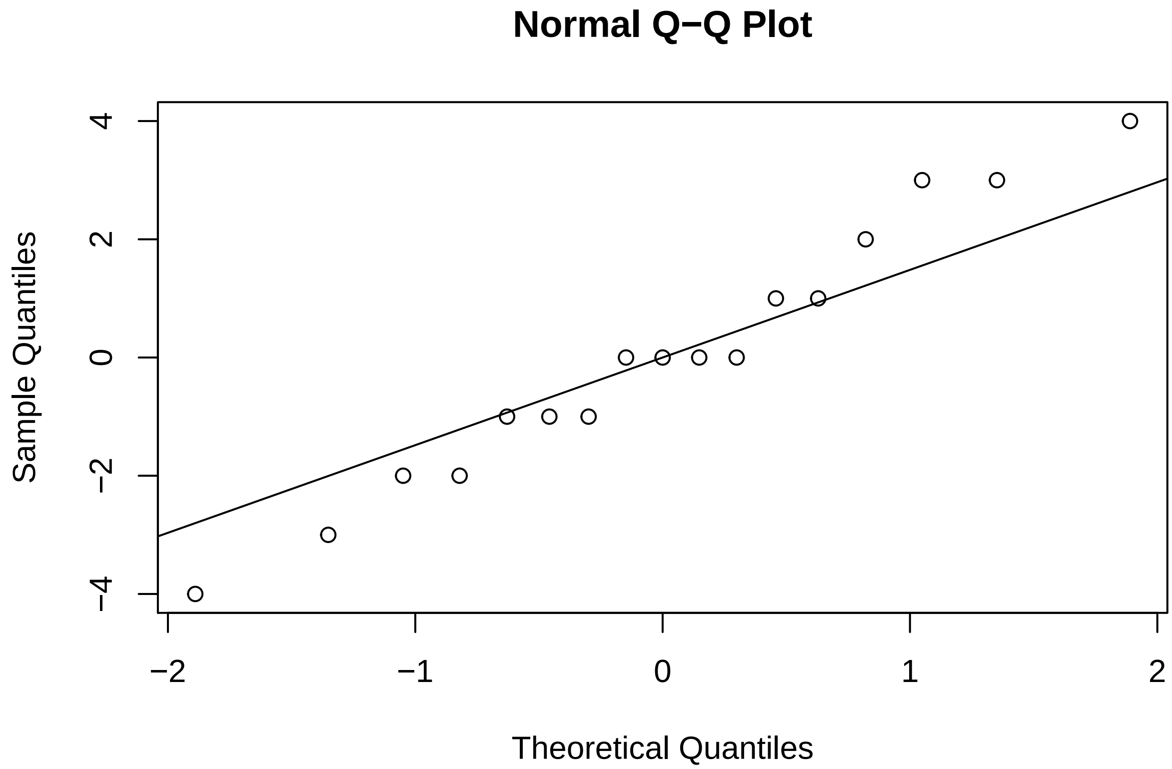
```
## [1] 0.0004
```

Part c

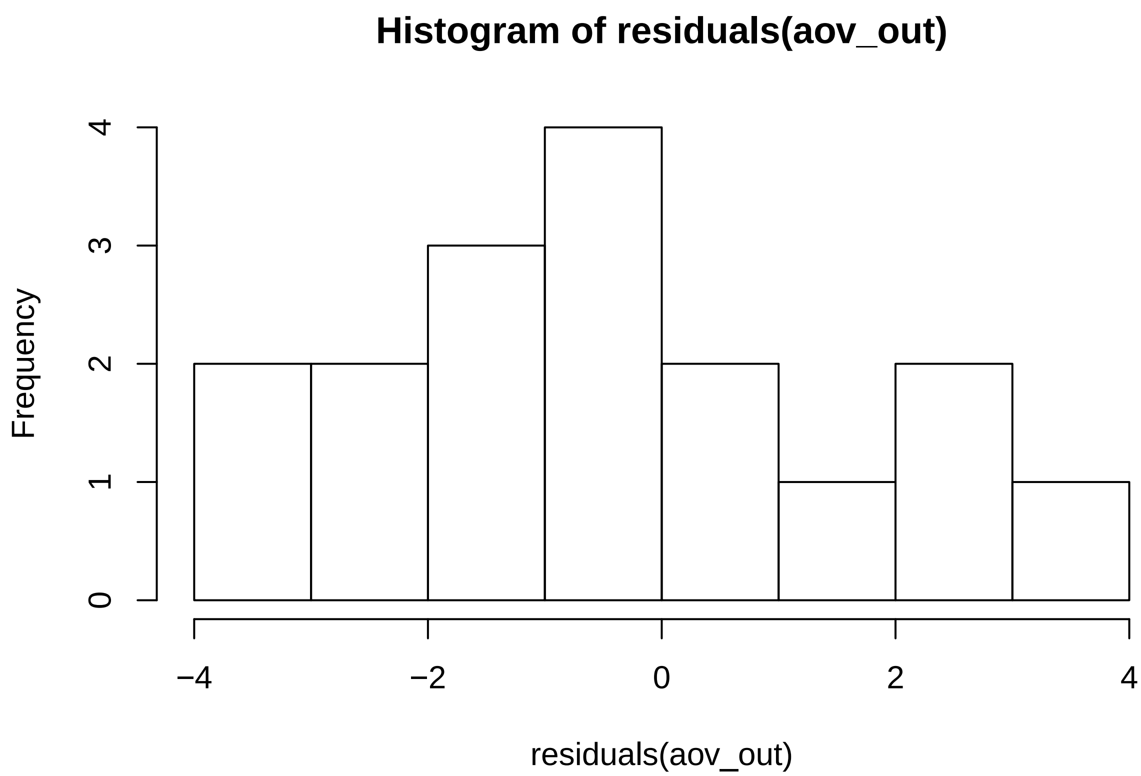
```
## Part c

### Check for error normality

qqnorm(residuals(aov_out))
qqline(residuals(aov_out))
```

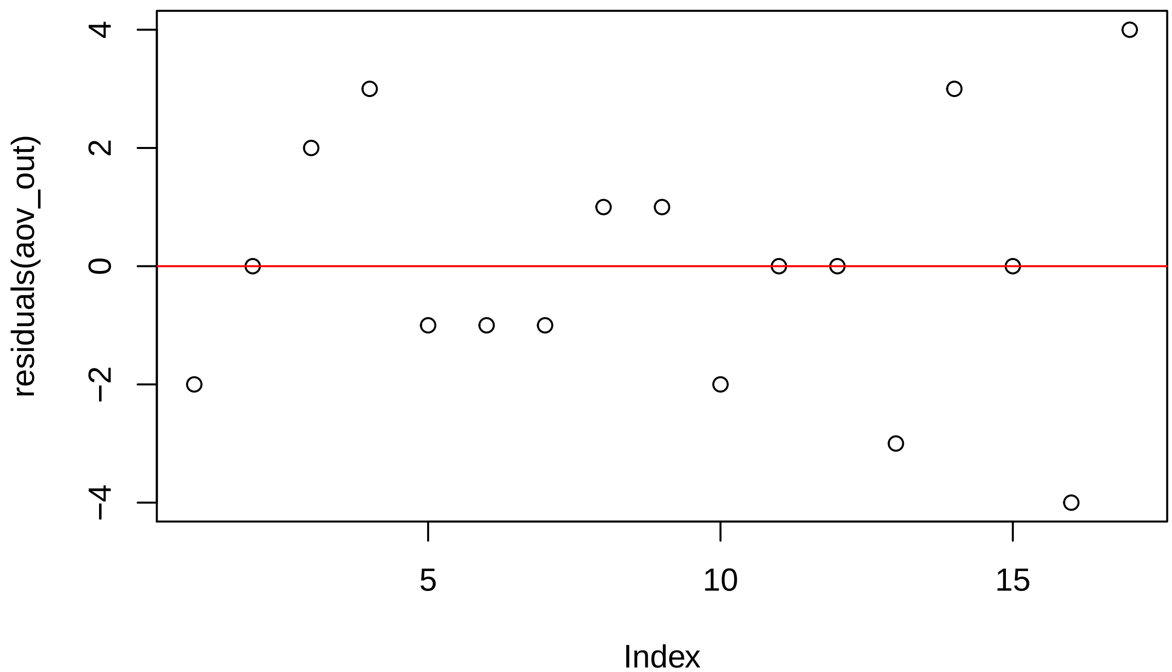



```
hist(residuals(aov_out))
```




```r
### Check residuals for equal variance

plot(residuals(aov_out))
abline(h = 0, col = "red")
```



Part e

```r
## Part e

### Compare B and C

t.test(x = data[4:7,2], y = data[8:10,2])
```

```
##
##  Welch Two Sample t-test
##
## data:  data[4:7, 2] and data[8:10, 2]
## t = 0.70711, df = 4.8, p-value = 0.5123
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -2.681387  4.681387
## sample estimates:
## mean of x mean of y
##         6         5
```

```r
### Compare B and E

t.test(x = data[4:7,2], y = data[14:17,2])
```

```
##
##  Welch Two Sample t-test
##
## data:  data[4:7, 2] and data[14:17, 2]
## t = -2.1669, df = 3.8802, p-value = 0.09825
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -13.78215   1.78215
## sample estimates:
## mean of x mean of y
##         6        12
```

STAT616 - Problem 1

```r
# STAT616 - Problem 1

values <- c(1,2,3,4,60000001,60000000)
labels <- c("A","A","B","B","C","C")

data <- data.frame(labels, values)

aov_out <- aov(values ~ labels, data = data)
summary(aov_out)
```

```
##             Df           Sum Sq          Mean Sq          F value
## labels       2 4799999680000010 2399999840000005 4799999696203265
## Residuals    3                1                0
##                          Pr(>F)
## labels      <0.0000000000000002 ***
```

```
## Residuals
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
f_stat <- as.vector(summary.lm(aov_out)$fstatistic["value"])



f_stats <- numeric()
n <- 5000

for (i in 1:n) {

    labels_temp <- sample(labels)
    data_temp <- data.frame(labels_temp, values)
    aov_out_temp <- aov(values ~ labels_temp, data = data_temp)
    f_stats[i] <- as.vector(summary.lm(aov_out_temp)$fstatistic["value"])

}

count_greater <- sum(f_stats >= f_stat)

count_greater/n
```

```
## [1] 0.024
```