Assignment 4

Emil Hafeez (eh2928)

10/28/2020

Problem 1 (5p)

In the context of one-way ANOVA models, prove the partitioning of the total variability (sum of squares).

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(y_{ij}-\overline{\overline{y}}\right)^2=\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(y_{ij}-\bar{y}_i\right)^2+\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(\bar{y}_i-\overline{\overline{y}}\right)^2 \tag{1}$$

We use a one-way ANOVA model specified as:

$y_{ij}=\mu+\alpha_i+e_{ij}, i=1,2,...k, j=1,2,...n_i,$

where μ is a constant representing the underlying mean of all groups taken together (the grand mean), α_i is a constant for the difference between the i-th group mean and the grand mean, and e_{ij} represent the normally distribution error terms.

We assume that the samples are drawn independently from the underlying populations, that the distributions of the error terms are normal, and that the variances of the k populations are equal (and thus that the variance of the outcome does not depend on the sample).

Let n be the total number of observations

Let's write the total variability as:

$$y_{ij}-\bar{y}$$

Now, we can also write that the total variation is equal to $\sum_{i=1}^{n}(y_i-\overline{y})^2$, giving the sum of squares for these data where y_i is the ith data point, and where $\bar{y}$ is the estimate of the mean. Now, more specifically,

Let $\bar{y}$ be the mean from each of the the i^{th} group where i = 1,2,3,... k is the number of groups and let $\overline{\overline{y}}$ be the grand mean, such that $\overline{\overline{y}}=\frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i}y_{ij}}{n}$.

Now, we square the variability term and simplify the summation signs and find:

$\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(y_{ij}-\bar{y}\right)^2=$

$$\begin{aligned}
&=\sum_{i=1}^{n}\left(\left(\hat{y}_i-\bar{y}\right)^2+2\hat{\varepsilon}_i\left(\hat{y}_i-\bar{y}\right)+\hat{\varepsilon}_i^2\right)\\
&=\sum_{i=1}^{n}\left(\hat{y}_i-\bar{y}\right)^2+\sum_{i=1}^{n}\hat{\varepsilon}_i^2+2\sum_{i=1}^{n}\hat{\varepsilon}_i\left(\hat{y}_i-\bar{y}\right)\\
&=\sum_{i=1}^{n}\left(\hat{y}_i-\bar{y}\right)^2+\sum_{i=1}^{n}\hat{\varepsilon}_i^2+2\sum_{i=1}^{n}\hat{\varepsilon}_i\left(\hat{\beta}_0+\hat{\beta}_1x_{i1}+\cdots+\hat{\beta}_px_{ip}-\bar{y}\right)\\
&=\sum_{i=1}^{n}\left(\hat{y}_i-\bar{y}\right)^2+\sum_{i=1}^{n}\hat{\varepsilon}_i^2+2\left(\hat{\beta}_0-\bar{y}\right)\sum_{i=1}^{n}\hat{\varepsilon}_i+2\hat{\beta}_1\underbrace{\sum_{i=1}^{n}\hat{\varepsilon}_ix_{i1}}_{0}+\cdots+2\hat{\beta}_p\underbrace{\sum_{i=1}^{n}\hat{\varepsilon}_ix_{ip}}_{0}\\
&=\sum_{i=1}^{n}\left(\hat{y}_i-\bar{y}\right)^2+\sum_{i=1}^{n}\hat{\varepsilon}_i^2
\end{aligned}$$

where we can assume those terms are zero based on the normal distribution of those error terms.

So, $= \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n} \hat{\varepsilon}_i^2 = ESS + RSS$,

Thus, $\sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2 = ESS + RSS$ for each individual of each group, such that

$$\sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2 = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2 + \sum_{i=1}^{k} \sum_{j=1}^{n_i} (\bar{y}_i - \bar{y})^2$$

which proves

$$\sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(y_{ij} - \overline{\bar{y}}\right)^2 = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2 + \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(\bar{y}_i - \overline{\bar{y}}\right)^2 \tag{2}$$

Problem 2

Part a)

```
summary(knee_df)
```

```
##      Below       Average          Above
##  Min.   :29   Min.   :28.00   Min.   :20.00
##  1st Qu.:36   1st Qu.:30.25   1st Qu.:21.00
##  Median :40   Median :32.00   Median :22.00
##  Mean   :38   Mean   :33.00   Mean   :23.57
##  3rd Qu.:42   3rd Qu.:35.00   3rd Qu.:24.50
##  Max.   :43   Max.   :39.00   Max.   :32.00
##  NA's   :2                    NA's   :3
```

Regarding participants' physical status before therapy, the "Above Average" group had a mean of 23.57 days (standard deviation = 4.20), and a median of 22 days. The "Average" group had a mean of 33 days to recover (standard deviation = 3.92 days), and a median of 32 days. The "Below Average" group had a mean of 38 days to recover (standard deviation = 5.478 days) and a median of 40 days to recover. Thus, it appears that worse prior physical status is associated with more days to recovery. Interestingly, it also appears that a worse prior physical status seems to have higher dispersion (as per the SD, as well). This can be explored visually, as below.

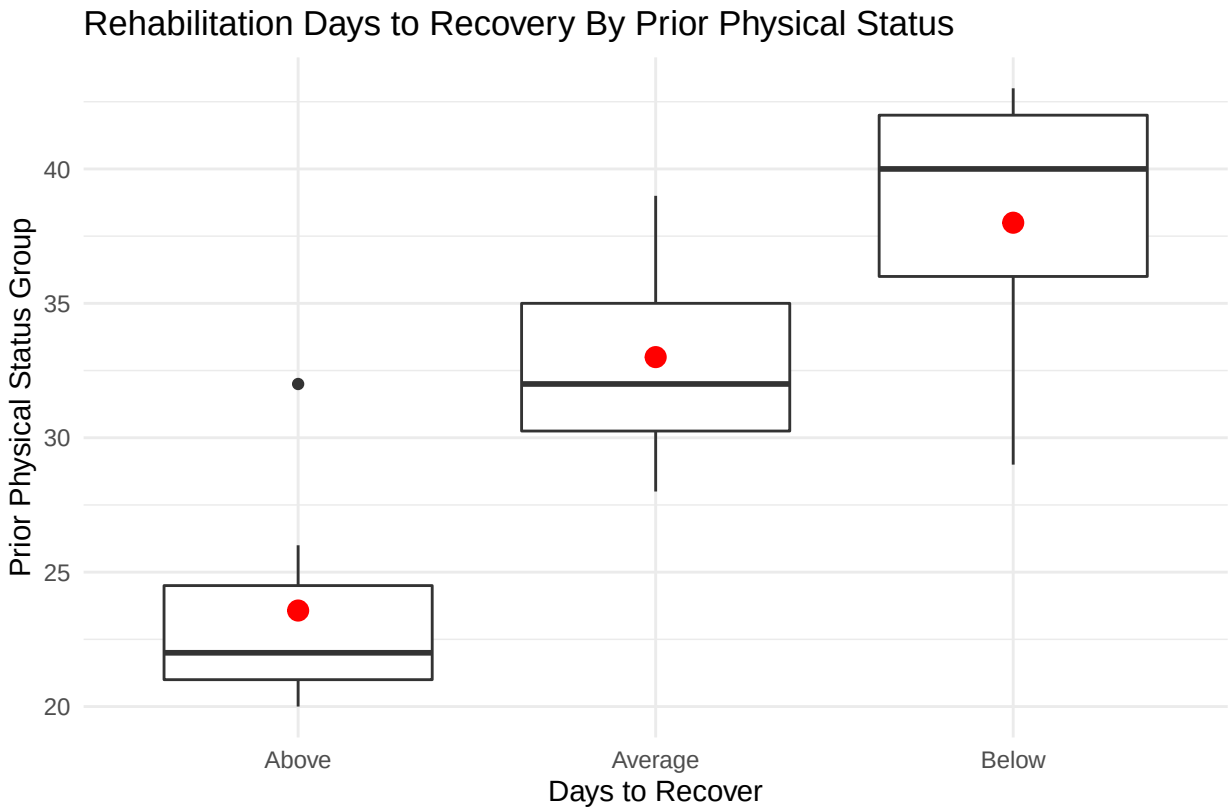


Figure 1, Biostatistics P8130 Assignment 4

Part b)

Let $\alpha = 0.01$

$H_0 : \mu_1 = \mu_2 = \mu_3$, ie that all group means (the Below, Average, and Above groups respectively) are equal

H_A : at least two of the population means differ, ie not all group means are equal

Source	Sum of Square (SS)	Degrees of Freedom (df)	Mean Sum of Square	F-Statistic
Between	Between SS	k-1	Between SS / (k-1)	F = $\frac{(MeanBetweenSS/k-1)}{(WithinSS/(n-k)}$
Within	Within SS	n-k	Within SS / (n-k)	
Total	Between SS + Within SS	n-1		

Source	Sum of Square (SS)	Degrees of Freedom (df)	Mean Sum of Square	F-Statistic
Between	795.25	2	397.62	F = $\frac{(397.62)}{(19.28)}$
Within	453.71	22	20.62	
Total	1248.96	24		

$F = \frac{(MeanBetweenSS/k-1)}{(WithinSS/(n-k)} = 19.28$

$F_{critical} = F_{k-1,n-k,1-\alpha}$

`qf(0.99, 2, 22)` = 5.719022

Seeing as the decision rule is that we reject the null hypothesis if $F > F_{k-1,n-k,1-\alpha}$ and fail to reject the null hypothesis if $F \leq F_{k-1,n-k,1-\alpha}$, we find evidence to reject the null hypothesis and conclude that there is at least one group mean that is not equal to the other group means. In context, we find evidence to conclude that there is at least one difference in group mean days to recovery until successful rehabilitation among the Below, Average, and Above Average prior physical fitness status.

Part c)

Seeing as we find evidence to reject the null hypothesis in the omnibus test, we can proceed to implementing a pairwise comparison method. We opt to perform each of three post-hoc comparisons, including the Bonferroni comparison, Tukey, and Dunnett.

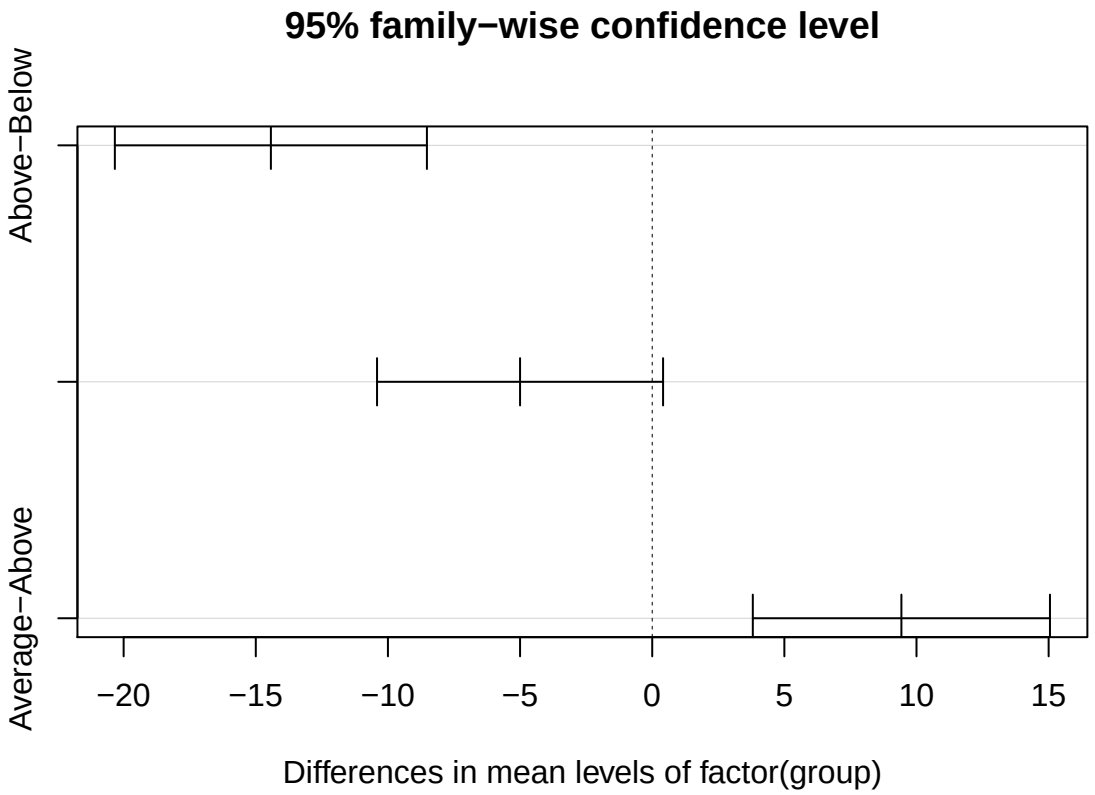
```
pairwise.t.test(knee_tidied_df$days_to_recover, knee_tidied_df$group, p.adj='bonferroni')
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  knee_tidied_df$days_to_recover and knee_tidied_df$group
##
##         Below   Above
## Above   1.1e-05 -
## Average 0.0898  0.0011
##
## P value adjustment method: bonferroni
```

```
# For Tukey, we need to use another function with an object created by aov()
Tukey_comp<-TukeyHSD(res1)
Tukey_comp
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = days_to_recover ~ factor(group), data = knee_tidied_df)
##
## $`factor(group)`
##                      diff        lwr        upr     p adj
## Above-Below   -14.428571 -20.332785 -8.5243579 0.0000102
## Average-Below  -5.000000 -10.411301  0.4113011 0.0736833
## Average-Above   9.428571   3.806636 15.0505072 0.0010053
```

```
plot(Tukey_comp)
```




```
# Dunnett's test: multiple comparisons with a specified control (here group #1)
summary(glht(res1), linfct=mcp(Group="Dunnett"))
```

```
## Warning in chkdots(...): Argument(s) 'linfct' passed to '...' are ignored
```

```
##
##   Simultaneous Tests for General Linear Hypotheses
##
## Fit: aov(formula = days_to_recover ~ factor(group), data = knee_tidied_df)
##
## Linear Hypotheses:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept) == 0             38.000      1.606  23.667   <0.001 ***
## factor(group)Above == 0     -14.429      2.350  -6.139   <0.001 ***
## factor(group)Average == 0    -5.000      2.154  -2.321   0.0677 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Adjusted p values reported -- single-step method)
```

Here we see our results reported. Summarily, we find that each of the 3 methods we implement (Bonferroni, Tukey, Dunnett) each all individually find that there are significant differences between the mean recovery time between the Average and Above Average group means, and between the Below Average and Above Average group means. The Bonferroni is the most conservative method (and the least powerful), whereas Tukey's is less conservative than Bonferroni, and Dunnett's primarily focuses on comparisons using the declared reference group of ("Below").

Part d)

Dear Dr. Rehab Hafeez,

I am writing to you to summarize the results of my analysis regarding your rehabilitation study. In order

to investigate whether there are differences between the mean rehabilitation days to recovery time between the specified Below Average prior physical status, the Average prior physical status, and Above Average physical status groups, we implemented an ANOVA model and found evidence to reject the null hypothesis and conclude there is at least one difference between the group means $F = \frac{(MeanBetweenSS/k-1)}{(WithinSS/(n-k)} = 19.28$ and $F_{critical} = 5.719022$. Then, we implemented pairwise comparisons using several different methods (Bonferroni, Tukey, and Dunnett's), and found converging evidence in order to support a significant difference between the mean recovery time between the Average and Above Average group means, and between the Below Average and Above Average group means. Best of luck on your rehabilitation work.

Problem 3

Part a)

We consider a situation in which we are examining proportions, not means, which already limits our initial test choices. Further, McNemar's test for binomial proportions is not appropriate given we are not utilizing paired data, and one- or two-sample tests for binomial proportions are also not appropriate considering the number of categories we wish to compare. A Fisher's Exact Test is not wise because this is not a 2x2 table.

Finally, considering that we are interested in testing whether distribution of swelling status is the same across the levels of the intervention group (vaccine / placebo), we opt to use a Chi-Squared test of homogeneity.

Problem b)

The table required to calculate the Chi Squared value is as follows.

	Major [Expected]	Minor [Expected]	No Swelling [Expected]	Total
Vaccine	54 [38.33]	42 [40.52]	134 [151.14]	230
Placebo	16 [31.66]	32 [33.48]	142 [124.85]	190
Total	70	74	276	142

Problem c)

We use $\alpha = 0.05$

$H_0 : p_{1,1} = p_{2,1} = p_{3,1}$, that the proportions of swelling category are equal among the vaccine status, AND that $H_0 : p_{1,2} = p_{2,2} = p_{3,2}$ the proportion of swelling category are equal among the placebo status.

The H_0 is that not all proportions are equal.

The decision rule is that, at the 0.05 level, $\chi^2 > \chi^2_{df,\alpha}$ where the degrees of freedom are given by $df = (r-1)(3-1) = 2$ and the alpha has been set at 0.05. We fail to reject the null hypothesis if $\chi^2 \leq \chi^2_{df,\alpha}$.

The test statistic is given by $\chi^2 = \sum_{i=1}^{2} \sum_{j=1}^{3} \frac{(O_{ij}-E_{ij})^2}{E_{ij}} = 18.571$.

The critical value, as mentioned, is given by $\chi^2_{(2-1)*(4-1),0.95)}$ computed by `qchisq(0.95,2)` $= 5.991465$.

Thus, we find evidence to reject the null hypothesis and conclude that there is sufficient evidence that the extent of swelling is significantly different by vaccine status.