

p8130_hw4_ps3194

Pangsibo Shen

11/5/2020

```
library(tidyverse)
library(arsenal)
library(multcomp)
```

Problem 1

In the context of ANOVA model, prove the partitioning of the total variability (sum of squares).

solution: We start the proof from the fact that difference between each observation and the grand mean equals the sum of within group variability and between group variability.

y_{ij} : denote the observation from the j^{th} subject from the i^{th} group.

$\bar{\bar{y}}$: denote the grand mean

\bar{y}_i : denote the mean from the i^{th} group.

$y_{ij} - \bar{\bar{y}}$: difference between each observation and the grand mean

$y_{ij} - \bar{y}_i$: within group variability

$\bar{y}_i - \bar{\bar{y}}$: between group variability

$$y_{ij} - \bar{\bar{y}} = (y_{ij} - \bar{y}_i) + (\bar{y}_i - \bar{\bar{y}})$$

We then square and take the total sum of both sides of the equation

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

We then expand the right side of the equation

$$\begin{aligned} \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{\bar{y}})^2 &= \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2 + 2 \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)(\bar{y}_i - \bar{\bar{y}}) + \sum_{i=1}^k \sum_{j=1}^{n_i} (\bar{y}_i - \bar{\bar{y}})^2 \\ &\quad \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)(\bar{y}_i - \bar{\bar{y}}) = 0 \\ &\Rightarrow \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{\bar{y}})^2 = \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2 + \sum_{i=1}^k \sum_{j=1}^{n_i} (\bar{y}_i - \bar{\bar{y}})^2 \end{aligned}$$

We thus complete the proof of partitioning of the total variability

Problem 2

load the data

```
knee_df = read_csv("../data/Knee.csv")
```

a) The descriptive statistics for each group are displayed below:

```
my_controls <- tableby.control(  
  total = T,  
  test = F,  
  numeric.stats = c("meansd", "medianq1q3", "iqr", "range", "Nmiss2"),  
  cat.stats = c("countpct", "Nmiss2"),  
  stats.labels = list(  
    meansd = "Mean (SD)",  
    medianq1q3 = "Median (Q1, Q3)",  
    iqr = "IQR",  
    range = "Min - Max",  
    Nmiss2 = "Missing",  
    countpct = "N (%)"  
  )  
  
tab1 <- tableby(~Below + Average + Above, data = knee_df, control = my_controls)  
  
knitr::kable(summary(tab1, title = "Descriptive Statistics"))
```

| | Overall (N=10) |
|-----------------|-------------------------|
| Below | |
| Mean (SD) | 38.000 (5.477) |
| Median (Q1, Q3) | 40.000 (36.000, 42.000) |
| IQR | 6.000 |
| Min - Max | 29.000 - 43.000 |
| Missing | 2 |
| Average | |
| Mean (SD) | 33.000 (3.916) |
| Median (Q1, Q3) | 32.000 (30.250, 35.000) |
| IQR | 4.750 |
| Min - Max | 28.000 - 39.000 |
| Missing | 0 |
| Above | |
| Mean (SD) | 23.571 (4.198) |
| Median (Q1, Q3) | 22.000 (21.000, 24.500) |
| IQR | 3.500 |
| Min - Max | 20.000 - 32.000 |
| Missing | 3 |

There are 2 missing value for below average group, 3 missing for above average group and no data missing for average. The below average group has the longest mean time required until rehabilitation among three groups and the above average group has the smallest mean time. Furthermore, the below average group also has the largest standard deviation for the mean time until rehabilitation and the average group has the

smallest standard deviation for the mean time until rehabilitation. The longest time required in physical therapy until successful rehabilitation is 42 days which also belongs to the below average group; the shortest time required in physical therapy until successful rehabilitation is 21 days which belongs to the above average group.

b) We are using $\alpha = 0.01$ for the ANOVA test and the test hypotheses are shown below:

$$H_0 : \mu_{below} = \mu_{average} = \mu_{above}$$

$$H_1 : \text{at least two means are not equal}$$

$$\text{Between } SS = \sum_{i=1}^k \sum_{j=1}^{n_i} (\bar{y}_i - \bar{y})^2$$

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

$$\text{Between Mean Square} = \frac{\sum_{i=1}^k \sum_{j=1}^{n_i} (\bar{y}_i - \bar{y})^2}{k - 1} = \frac{\sum_{i=1}^k n_i \bar{y}_i^2 - \frac{y_{..}^2}{n}}{k - 1}$$

$$\text{Within Mean Square} = \frac{\sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2}{n - k} = \frac{\sum_{i=1}^k (n_i - 1) s_i^2}{n - k}$$

$$F_{stat} = \frac{\text{Between Mean Square}}{\text{Within Mean Square}} \sim F(k - 1, n - k)$$

```
# Re-shape the data
mean_time = c(knee_df$Below,knee_df$Average,knee_df$Above)
group = c(rep("below",length(knee_df$Below)),rep("average",length(knee_df$Average)),rep("above",length(knee_df$Above)))
new_knee_df = as.data.frame(cbind(mean_time,group)) %>% drop_na()

# Perform an ANOVA test
# Function lm() is broader, including linear regression models
res = lm(mean_time~factor(group), data = new_knee_df)

# Coefficients of the ANOVA model with 'grand mean' and alpha effects.
res

##
## Call:
## lm(formula = mean_time ~ factor(group), data = new_knee_df)
##
## Coefficients:
##          (Intercept)  factor(group)average  factor(group)below
##              23.571              9.429              14.429

# Our regular ANOVA table with SS, Mean SS and F-test
anova(res)

## Analysis of Variance Table
##
## Response: mean_time
##              Df Sum Sq Mean Sq F value    Pr(>F)
## factor(group)  2  795.25   397.62    19.28 1.454e-05 ***
```

```
## Residuals      22 453.71    20.62
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The ANOVA table is shown below:

ANOVA Table:

| Source | Sum of Square (SS) | Degrees of freedom(df) | Mean Sum of Square | F-statistics |
|---------|--|------------------------|---|--|
| Between | Between SS = 795.25 | k-1 = 2 | $\frac{\text{Between SS}}{k-1} = 397.625$ | $F = \frac{\text{Between SS}/(k-1)}{\text{Within SS}/(n-k)} = 19.2805$ |
| Within | Within SS = 453.71 | n-k = 22 | $\frac{\text{Within SS}}{n-k} = 20.6232$ | |
| Total | Between SS + Within SS = 1248.96 | n-1 = 24 | | |

```
#critical value
qf(0.99,2,22)
```

```
## [1] 5.719022
```

$$F_{stat} = \frac{397.625}{20.6232} = 19.2805 \sim F(2, 22)$$

$$F_{k-1, n-k, 1-\alpha} = 5.719$$

$$F_{stat} > F_{2, 22, 1-0.01}$$

Since F (19.2805) is greater than $F_{2,22,0.99}$ (5.7190219) We rejected the null hypothesis and conclude that at α level of 0.01 that there is a significant difference in the mean time (days) required in physical therapy until successful rehabilitation between the groups with different physical status.

c) We performed a pairwise comparison with Bonferroni adjustments below:

```
pairwise.t.test(as.numeric(new_knee_df$mean_time), new_knee_df$group, p.adj = 'bonferroni')

##
## Pairwise comparisons using t tests with pooled SD
##
## data:  as.numeric(new_knee_df$mean_time) and new_knee_df$group
##
##      above      average
```

```
## average 0.0011 -
## below 1.1e-05 0.0898
##
## P value adjustment method: bonferroni
```

The p-value for pairwise comparison test between above average group and average group is 0.0011 which is less than the $\alpha = 0.01$. Hence we reject the null hypothesis and conclude that there is a significant difference in mean between the above average group and the average group; The p-value for pairwise comparison test between above group and below average group is $1.1e-05$ which is far less than the $\alpha = 0.01$. Hence we reject the null hypothesis and conclude that there is a significant difference in mean between the above group and the below average group. However, The p-value for pairwise comparison test between average group and below average group is 0.0898 which is greater than the $\alpha = 0.01$. Hence we fail to reject the null hypothesis and conclude that there is no difference in mean between the average group and below average group.

We performed a pairwise comparison with Tukey adjustments below:

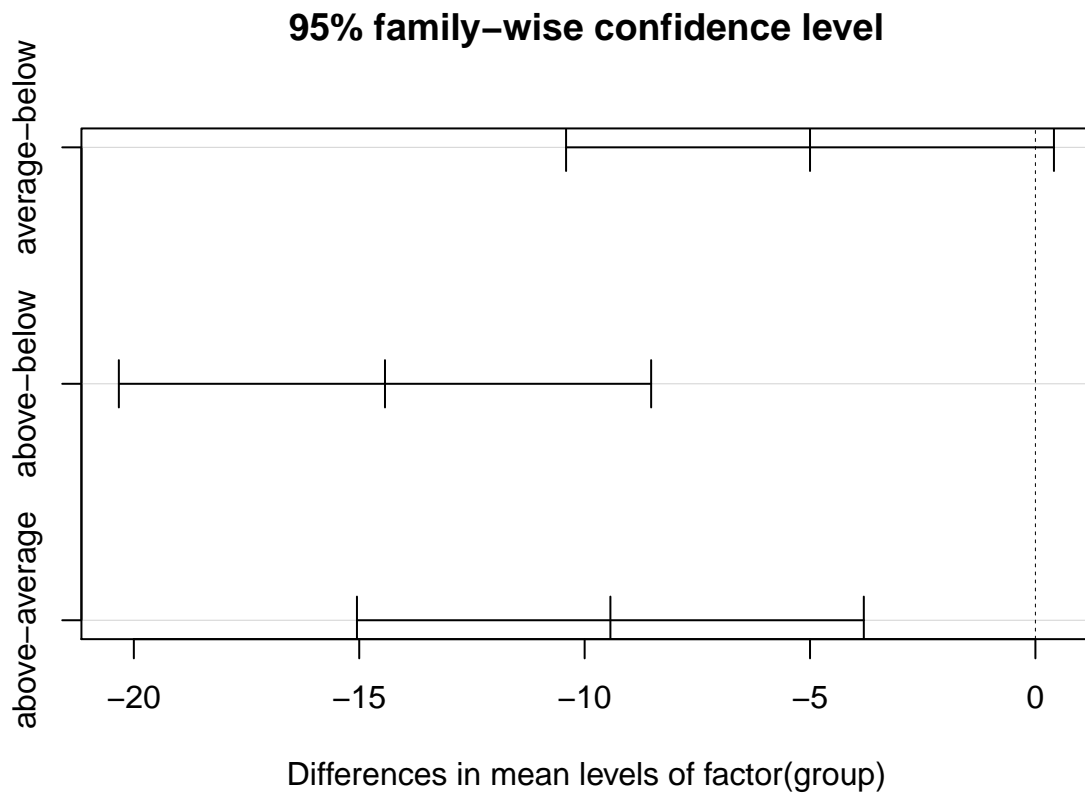
```
#refactor the group level so below average will be the reference
new_knee_df$group = factor(new_knee_df$group, levels = c("below", "average", "above"))
# Save the anova object for Tukey comparisons
res1 = aov(mean_time~factor(group), data = new_knee_df)

Tukey_comp = TukeyHSD(res1)
Tukey_comp
```

```
## Tukey multiple comparisons of means
## 95% family-wise confidence level
##
## Fit: aov(formula = mean_time ~ factor(group), data = new_knee_df)
##
## $'factor(group)'
```

| | diff | lwr | upr | p adj |
|---------------|------------|-----------|------------|-----------|
| average-below | -5.000000 | -10.41130 | 0.4113011 | 0.0736833 |
| above-below | -14.428571 | -20.33278 | -8.5243579 | 0.0000102 |
| above-average | -9.428571 | -15.05051 | -3.8066356 | 0.0010053 |

```
plot(Tukey_comp)
```



The p-value for pairwise comparison test between above average group and average group is 0.0010053 which is less than the $\alpha = 0.01$. Hence we reject the null hypothesis and conclude that there is a significant difference in mean between the above average group and the average group; The p-value for pairwise comparison test between above group and below average group is 0.0000102 which is far less than the $\alpha = 0.01$. Hence we reject the null hypothesis and conclude that there is a significant difference in mean between the above group and the below average group. However, The p-value for pairwise comparison test between average group and below average group is 0.0736833 which is greater than the $\alpha = 0.01$. Hence we fail to reject the null hypothesis and conclude that there is no difference in mean between the average group and below average group.

We performed a pairwise comparison with Dunnett adjustments below:

```
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 = mean_time ~ factor(group), data = new_knee_df)
##
## Linear Hypotheses:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) == 0      38.000     1.606  23.667 <0.001 ***
## factor(group)average == 0  -5.000     2.154  -2.321  0.068 .
## factor(group)above == 0  -14.429     2.350  -6.139 <0.001 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Adjusted p values reported -- single-step method)
```

We set the below average group as the reference for the Dunnett method. The p-value between below average group and average group is 0.0681 which is greater than the $\alpha = 0.01$. Hence we fail to reject the null hypothesis and conclude that there is no difference in mean between the average group and below average group; The p-value between the above average group and below average group is less than 0.001 which is less than the $\alpha = 0.01$. Hence we reject the null hypothesis and conclude that there is a significant difference in mean between the above average group and the below average group.

Similarity: The p-values between each pairwise comparison reported by each method are very close to each other. We can also interpret the same decision from these three methods: the mean time from below group is significantly different from that of above group; the mean time from average group is significantly different from that of above group; the mean time from below group is not significantly different from that of average group.

Differences: Bonferroni is the most conservative method and it is the most stringent in declaring significance; Tukey's method controls for all pairwise comparisons and it is less conservative than Bonferroni; Dunnett's method mainly focuses on comparisons with a pre-defined control arm, yet it can't do all pairwise comparisons like Tukey's and Bonferroni.

d) From the health records of patients 18-30 years old provided by the rehabilitation center, we performed a One-way ANOVA hypothesis testing to test whether there is a significant difference in the mean time (days) required in physical therapy until successful rehabilitation between the groups with different physical status ('below average', 'average' and 'above average'). The test statistics of ANOVA is greater than the critical value. Hence we reject the null hypothesis and conclude that there is a significant difference in the mean time (days) required in physical therapy until successful rehabilitation between the groups with different physical status. However, the ANOVA test didn't report which pairs of means are different. Therefore, we performed three Post-Hoc Tests: Bonferroni adjustment, Tukey's method and Dunnett's method. From the Post-Hoc analysis, we found out that the means is significantly different between below average group and above average group; similarly, the means is significantly different between average group and above average group; However, the mean time from below average group is not significantly different from that of the average group.

Problem 3

a) We are going to use Chi-squared: test of Homogeneity to evaluate whether the frequency counts distributed identically across different treatment groups, because the data from observations made on two different categorical variables with 2 or more levels. First, we are going to construct a 2 X 3 contingency table.

```
# 2 X 3 contingency table
Treatment_Outcome = c("Vaccine", "Placebo")
Major_Swelling = c(54, 16)
Minor_Swelling = c(42, 32)
No_Swelling = c(134, 142)
rc_table = data.frame("RC" = Treatment_Outcome, Major_Swelling, Minor_Swelling, No_Swelling)
knitr::kable(rc_table)
```

b)

| RC | Major_Swelling | Minor_Swelling | No_Swelling |
|---------|----------------|----------------|-------------|
| Vaccine | 54 | 42 | 134 |
| Placebo | 16 | 32 | 142 |

Then we are going to compute the expected values for each cell in the contingency table. The table for expected values is shown below:

```
# Table for expected value
rc_table_exp = data.matrix(data.frame(Major_Swelling, Minor_Swelling, No_Swelling))
expected_value = chisq.test(rc_table_exp)$expected
knitr::kable(data.frame(RC = Treatment_Outcome, expected_value))
```

| RC | Major_Swelling | Minor_Swelling | No_Swelling |
|---------|----------------|----------------|-------------|
| Vaccine | 38.33333 | 40.52381 | 151.1429 |
| Placebo | 31.66667 | 33.47619 | 124.8571 |

Seems like all assumptions are met: independent random samples; No expected cell counts are 0, and no more than 20% of the cells have an expected count level less than 5. Next we have our null hypothesis, alternative hypothesis and formula for test statistics.

c)

$$H_0 : p_{1j} = p_{2j} = \dots = p_{Rj} = p_{(\cdot j)}$$

$$H_1 : p_{ij} \neq p_{i'j}, j = 1, 2, \dots, C, i \neq i'$$

$$\chi^2_{stat} = \sum_i^R \sum_j^C \frac{(n_{ij} - E_{ij} - 0.5)^2}{E_{ij}} \sim \chi^2_{(R-1) \times (C-1)}, \text{ where } df = (R-1) \times (C-1)$$

```
chi_sq = ((54-38.33333 - 0.5)^2/38.33333 + (42 - 40.52381 - 0.5)^2/40.52381 + (134-151.1429 - 0.5)^2/151.1429 +
```

```
chi_sq
```

```
## [1] 18.67229
```

```
chi_cri = qchisq(0.95, 2)
```

```
chi_cri
```

```
## [1] 5.991465
```

```
chisq.test(rc_table_exp, correct = TRUE)
```



```
##
## Pearson's Chi-squared test
##
## data:  rc_table_exp
## X-squared = 18.571, df = 2, p-value = 9.277e-05
```

$$\chi^2_{stat} = \frac{(54 - 38.33333 - 0.5)^2}{38.33333} + \frac{(42 - 40.52381 - 0.5)^2}{40.52381} + \frac{(134 - 151.1429 - 0.5)^2}{151.1429} + \frac{(16 - 31.66667 - 0.5)^2}{31.66667} + \frac{(32 - 33.47619 - 0.5)^2}{33.47619} + \frac{(142 - 124.8571 - 0.5)^2}{124.8571} = 18.67229 \sim \chi^2_{(1) \times (2)}$$

$$\chi^2_{critical} = \chi^2_{2,0.95} = 5.991465$$

$$\chi^2_{stat} > \chi^2_{critical} \quad \text{reject the null hypothesis}$$

Plug in the number, we get $\chi^2_{stat} = 18.6722942$ which is greater than the critical value $\chi^2_{4,0.95} = 5.9914645$. Hence, we reject the null hypothesis and conclude that at least one of the singles outcomes differs in treatment and placebo groups.