p8130_hw4

Jeffrey Liang

10/26/2020

Problem 1

In the context of ANOVA model, prove the partitioning of the total variability (sum of squares), i.e., =1 = 1(−)2 = =1 = 1(−)2 + =1 = 1(−)2

PROOF we have by definition, the

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

Fixing within group i, we have within group variance:

$$\begin{aligned} \sum_j (y_{ij} - \bar{\bar{y}})^2 \\ = \sum_j [(y_{ij} - \bar{y_i}) + (\bar{y_i} - \bar{\bar{y}})]^2 \\ = \sum_j (y_{ij} - \bar{y_i})^2 + (\bar{y_i} - \bar{\bar{y}})^2 + 2 * (y_{ij} - \bar{y_i})(\bar{y_i} - \bar{\bar{y}}) \end{aligned}$$

With $\sum_j y_{ij}/n_j = \bar{y_i}$ and $\sum_j 1 = n_j$

$$\begin{aligned} \sum_j 2 * (y_{ij} - \bar{y_i})(\bar{y_i} - \bar{\bar{y}}) \\ = 2 \sum_j y_{ij} * \bar{y_i} - y_{ij} * \bar{\bar{y}} - \bar{y_i}^2 + \bar{y_i} * \bar{\bar{y}} \\ = 2 * n_j * \bar{y_i}^2 - 2 * n_j * \bar{y_i} * \bar{\bar{y}} - 2 * n_j * \bar{y_i}^2 + 2 * n_j * \bar{y_i} * \bar{\bar{y}} \\ rearrange = 2 * n_j * \bar{y_i}^2 - 2 * n_j * \bar{y_i}^2 + 2 * n_j * \bar{y_i} * \bar{\bar{y}} - 2 * n_j * \bar{y_i} * \bar{\bar{y}} \\ = 0 \end{aligned}$$

Now sum over group i we have

$$\sum_i \sum_j (y_{ij} - \bar{\bar{y}})^2 = \sum_i \sum_j (y_{ij} - \bar{y_i})^2 + (\bar{y_i} - \bar{\bar{y}})^2$$

Problem 2

A rehabilitation center is interested in examining the relationship between physical status before therapy ('below average', 'average' and 'above average') and the time (days) required in physical therapy until successful rehabilitation. Records from patients 18-30 years old were collected and provided to you for statistical analysis (dataset "Knee.csv").

Assuming that data are normally distributed, answer the questions below:

a) Generate descriptive statistics for each group and comment on the differences observed. (4p)

b) Using a type I error of 0.01, obtain the ANOVA table. State the hypotheses, test statistic, critical value, and decision interpreted in the context of the problem. (5p)

c) Based on your response in part b), perform pairwise comparisons with the appropriate adjustments (Bonferroni, Tukey, and Dunnett – 'below average' as reference). Report your findings and comment on the differences/similarities between these three methods. (5p)

d) Write a short paragraph summarizing your overall results as if you were presenting to the rehabilitation center director. (1p)

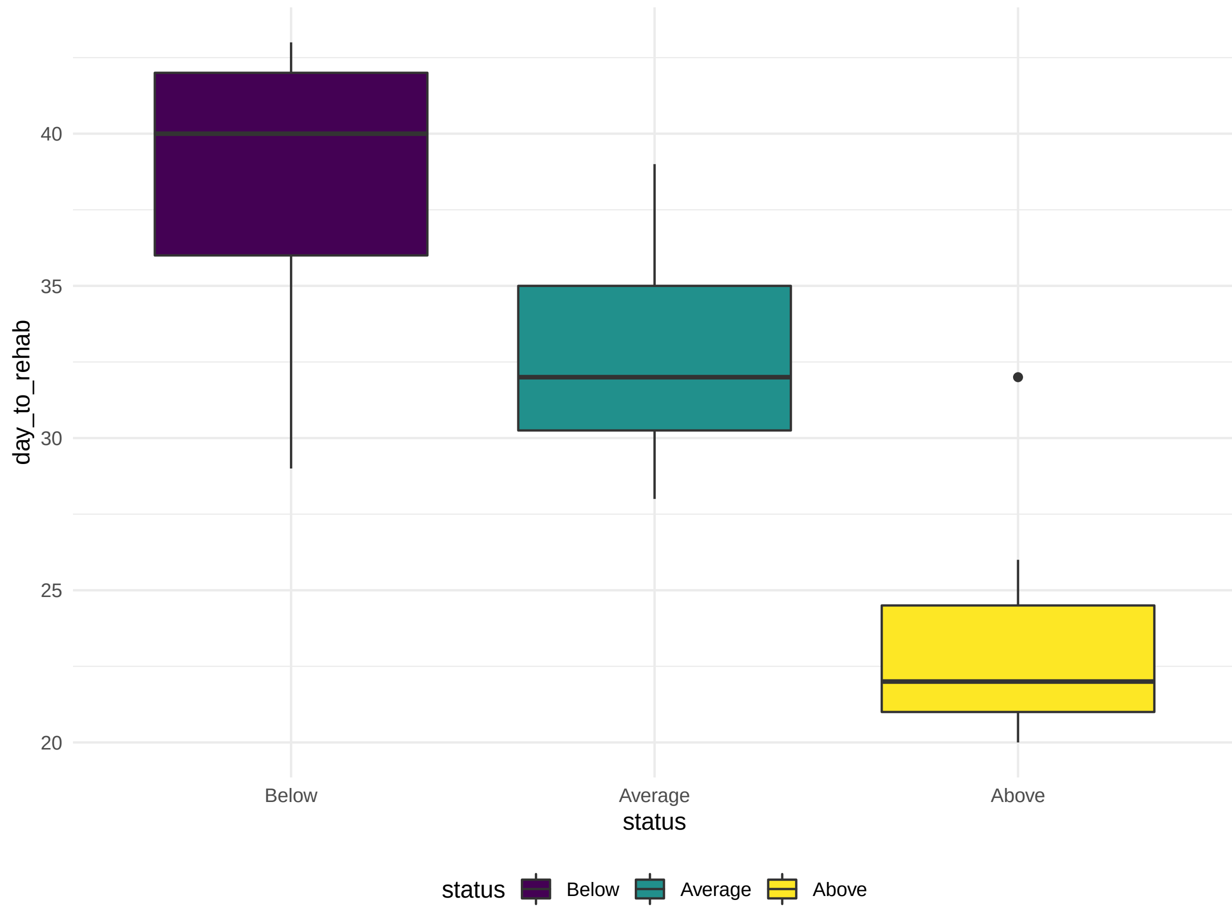
PROOF

Table 1: Data summary

Name	knee_data %>% group_by(st...
Number of rows	30
Number of columns	2
Column type frequency:	
numeric	1
Group variables	status

Variable type: numeric

skim_variable	status	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100
day_to_rehab	Below	2	0.8	38.0	5.48	29	36.0	40	42.0	43
day_to_rehab	Average	0	1.0	33.0	3.92	28	30.2	32	35.0	39
day_to_rehab	Above	3	0.7	23.6	4.20	20	21.0	22	24.5	32



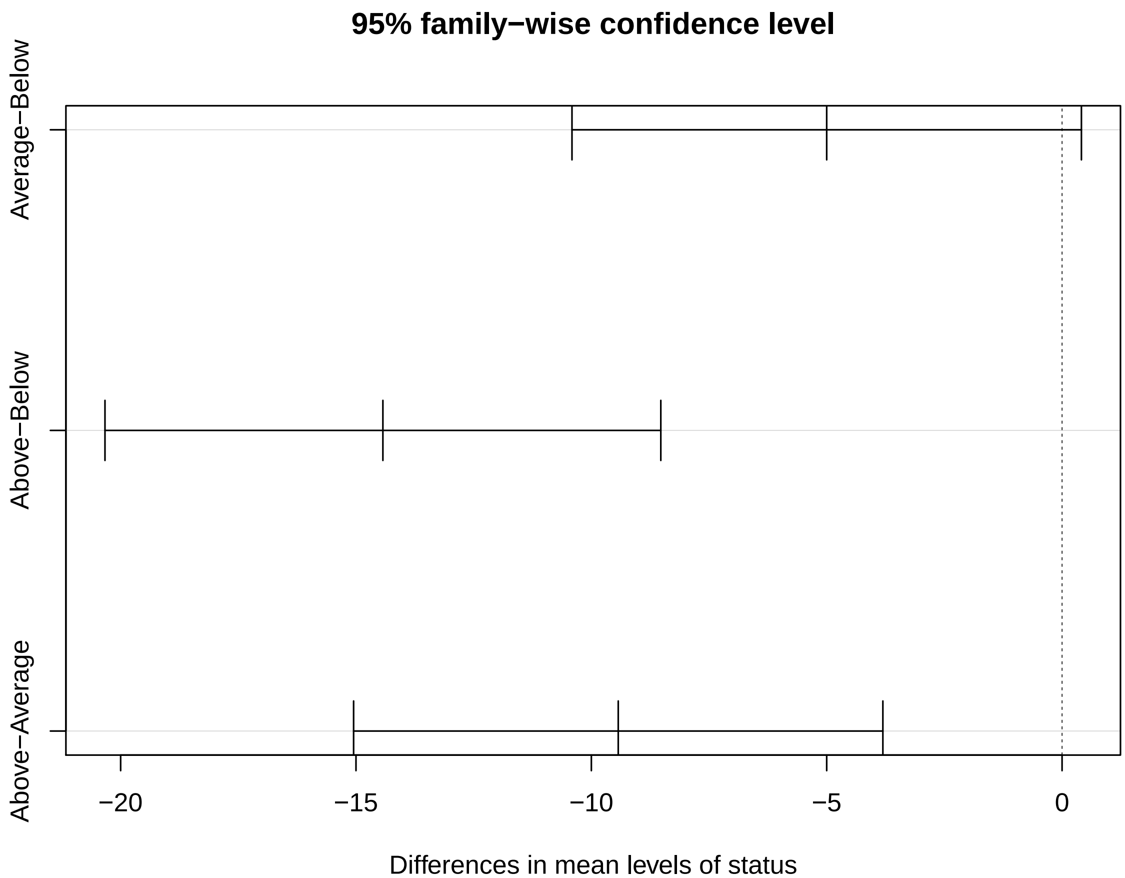

```
##             Df Sum Sq Mean Sq F value  Pr(>F)
## status       2    795     398    19.3 1.5e-05 ***
## Residuals   22    454      21
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 5 observations deleted due to missingness
```

```
## [1] 5.72
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  knee_data$day_to_rehab and knee_data$status
##
##         Below Average
## Average 0.090 -
## Above   1e-05 0.001
##
## P value adjustment method: bonferroni
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = day_to_rehab ~ status, data = knee_data, alpha = 0.01)
##
```

```
## $status
##                  diff   lwr    upr p adj
## Average-Below  -5.00 -10.4  0.411 0.074
## Above-Below   -14.43 -20.3 -8.524 0.000
## Above-Average  -9.43 -15.1 -3.807 0.001
```




```
##
##   Simultaneous Tests for General Linear Hypotheses
##
## Fit: aov(formula = day_to_rehab ~ status, data = knee_data, alpha = 0.01)
##
## Linear Hypotheses:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept) == 0      38.00       1.61   23.67   <0.001 ***
## statusAverage == 0    -5.00       2.15   -2.32    0.068 .
## statusAbove == 0     -14.43       2.35   -6.14   <0.001 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Adjusted p values reported -- single-step method)
```

Problem 3

A research article was published with the following headline “For adults, chicken pox vaccine may stop shingles”. The findings were based on a randomized clinical trial with a total of 420 adults being randomized to receive either chicken pox vaccine or placebo. While the results were intriguing, some side effects emerged and required further investigation. The table below summarizes the frequencies of one of the most frequent and concerning side effect - swelling around the injection site.

	Major Swelling	Minor Swelling	No swelling
Vaccine	54	42	134
Placebo	16	32	142

Use a significance level of 0.05 to assess if the distribution of swelling status is the same for the two treatment populations.

a) Justify the appropriate test to be used for addressing the question of interest. (2p)

b) Provide the table with all values necessary for calculating the test statistic. (4p)

c) State the hypotheses, test statistic, critical value, p value and decision rule interpreted in the context of the problem. (4p)

PROOF

```
##         Major_Swelling Minor_Swelling No_Swelling
## Vaccine             54             42         134
## Placebo             16             32         142
```

```
##         Major_Swelling Minor_Swelling No_Swelling
## Vaccine           38.3           40.5         151
## Placebo           31.7           33.5         125
```

```
##
##  Pearson's Chi-squared test
##
## data:  Prob3_table
## X-squared = 19, df = 2, p-value = 9e-05
```