p8130_hw4_ps3194

Pangsibo Shen

11/5/2020

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --

## v ggplot2 3.3.2     v purrr   0.3.4
## v tibble  3.0.3     v dplyr   1.0.2
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.0

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()

## Warning: package 'arsenal' was built under R version 4.0.3
```

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

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

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\bar{y_i})(\bar{y_i}-\bar{\bar{y}}) = 0$$

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

We thus complete the proof of partitioning of the total variability

---

Problem 2

load the data

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

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   Below = col_double(),
##   Average = col_double(),
##   Above = col_double()
## )
```

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 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 hypotheseses are shown below:

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

$$H_1 : at\ least\ two\ means\ are\ not\ equal$$

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

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

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

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

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

```
# Re-shape the data
mean_time = c(knee_df$Below,knee_df$Average,knee_df$Above)
group = c(rep("blow",length(knee_df$Below)),rep("average",length(knee_df$Average)),rep("above",length(kn
```

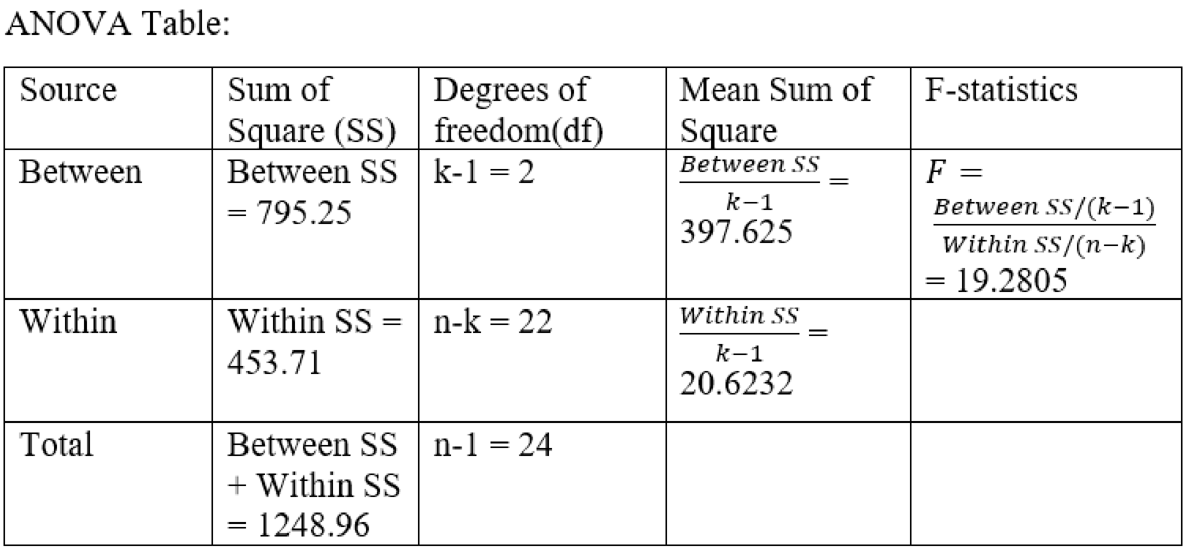
```r
new_knee_df = as.data.frame(cbind(mean_time,group)) %>% drop_na()

# 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.
# Will use them later in regression.
res
```

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

```r
# 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	$\frac{Between\ SS}{k-1} =$ 397.625	$F = \frac{Between\ SS/(k-1)}{Within\ SS/(n-k)}$ = 19.2805
Within	Within SS = 453.71	n-k = 22	$\frac{Within\ SS}{k-1} =$ 20.6232	
Total	Between SS + Within SS = 1248.96	n-1 = 24		

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

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

$$\begin{aligned} F_{stat} = \frac{397.625}{20.6232} &= 19.2805 \sim F(2,22) \\ F_{k-1,n-k,1-\alpha} &= 5.719 \\ F_{stat} &> F_{2,122,1-0.01} \end{aligned}$$

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 pysical status.