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

## Warning: package 'multcomp' was built under R version 4.0.3

## Loading required package: mvtnorm

## Warning: package 'mvtnorm' was built under R version 4.0.3

## Loading required package: survival

## Loading required package: TH.data

## Warning: package 'TH.data' was built under R version 4.0.3

## Loading required package: MASS

##
## Attaching package: 'MASS'

## The following object is masked from 'package:dplyr':
##
##     select

##
## Attaching package: 'TH.data'

## The following object is masked from 'package:MASS':
##
##     geyser
```

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.

$$
\begin{aligned}
y_{ij}: \ & \textit{denote the observation from the } j^{th} \textit{ subject from the } i^{th} \textit{ group.} \\
\bar{\bar{y}}: \ & \textit{denote the grand mean} \\
\bar{y}_i: \ & \textit{denote the mean from the } i^{th} \textit{ group.} \\
y_{ij} - \bar{\bar{y}}: & \textit{difference between each observation and the grand mean} \\
y_{ij} - \bar{y}_i: & \textit{within group variability} \\
\bar{y}_i - \bar{\bar{y}}: & \textit{between group variability} \\
y_{ij} - \bar{\bar{y}} &= (y_{ij} - \bar{y}_i) + (\bar{y}_i - \bar{\bar{y}})
\end{aligned}
$$

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
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
```

```
# 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		

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

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

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

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  -
## blow    1.1e-05 0.0898
##
## P value adjustment method: bonferroni
```

The p-value for t test between above 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 group and the average group; The p-value for t test between above group and blow 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 blow group. However, The p-value for t test between average group and blow 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 blow group.

We performed a pairwise comparison with Tukey adjustments below:

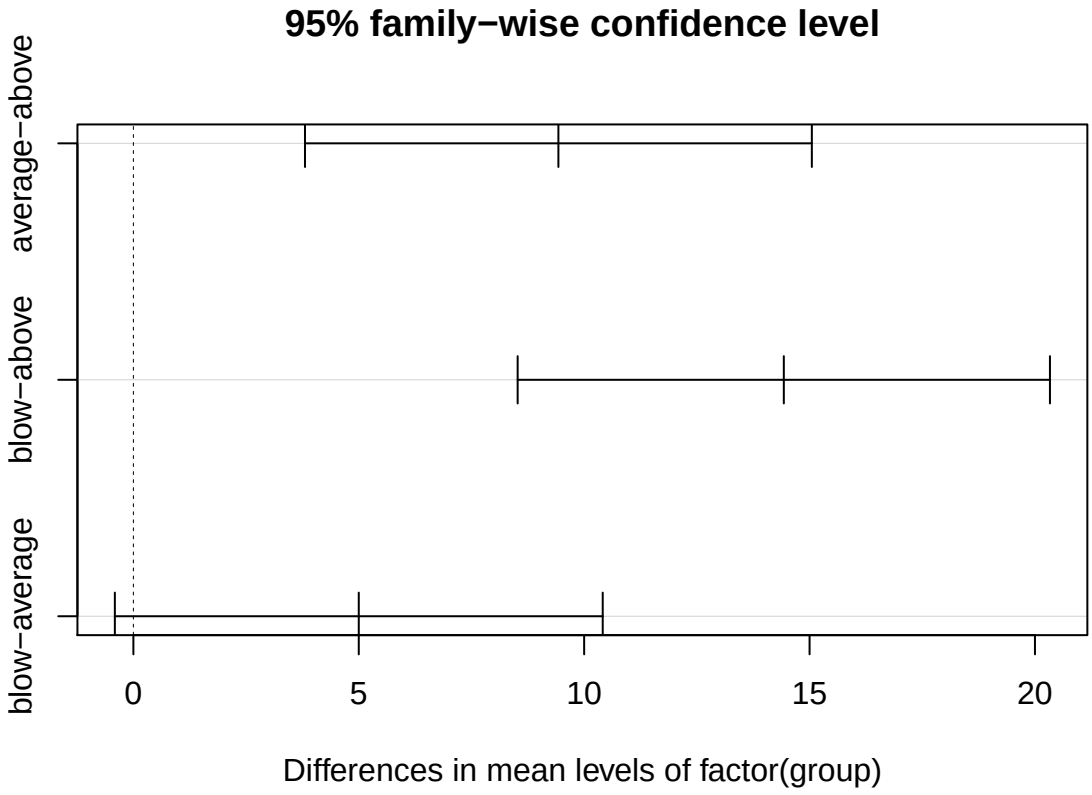
```
# Another option using aov();
# Save the anova object for Tukey comparisons
res1 = aov(mean_time~factor(group), data = new_knee_df)
summary(res1)
```

```
##               Df Sum Sq Mean Sq F value   Pr(>F)
## factor(group)  2  795.2   397.6   19.28 1.45e-05 ***
## Residuals     22  453.7    20.6
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
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-above  9.428571  3.8066356 15.05051 0.0010053
## blow-above    14.428571  8.5243579 20.33278 0.0000102
## blow-average   5.000000 -0.4113011 10.41130 0.0736833
```

```
plot(Tukey_comp)
```



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            23.571      1.716  13.733  < 1e-04 ***
## factor(group)average == 0    9.429      2.238   4.213 0.000746 ***
## factor(group)blow == 0      14.429      2.350   6.139  < 1e-04 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Adjusted p values reported -- single-step method)
```

d) summary

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)

$$H_0 : p_{1j} = p_{2j} = ... = p_{Rj} = p_{(.}j)$$
$$H_1 : p_{ij} \neq p_{i'j},\ j = 1, 2, ..., C, i \neq i'$$
$$\chi^2_{stat} = \sum_i^R \sum_j^C \frac{(n_{ij} - E_{ij} - 0.5)^2}{E_{ij}} \sim \chi^2_{(R-1)\times(C-1)}, where\ df = (R-1) \times (C-1)$$

```r
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

chi_sq
```

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

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

chi_cri
```

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

```r
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
```

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

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

$$\chi^2_{stat} > \chi^2_{critical} \quad \textit{rejcet the null hypothesis}$$

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.