p8130 HW4

JingYao Geng

11/2/2020

Problem 2

2(a) There are a total of 25 observations in the "knee.csv" dataset with 8 observations in the 'below' group, 10 observations in the 'average' group, and 7 observations in the 'above' group. More statistics summaries are shown on the following table.

```
##
## Table: Descriptive Statistics: Knee Data
##
## |             | Overall (N=10)  |
## |:------------|:---------------:|
## |below        |                 |
## |-  N-Miss    |        2        |
## |-  Mean (SD) | 38.000 (5.477)  |
## |-  Range     | 29.000 - 43.000 |
## |average      |                 |
## |-  Mean (SD) | 33.000 (3.916)  |
## |-  Range     | 28.000 - 39.000 |
## |above        |                 |
## |-  N-Miss    |        3        |
## |-  Mean (SD) | 23.571 (4.198)  |
## |-  Range     | 20.000 - 32.000 |
```

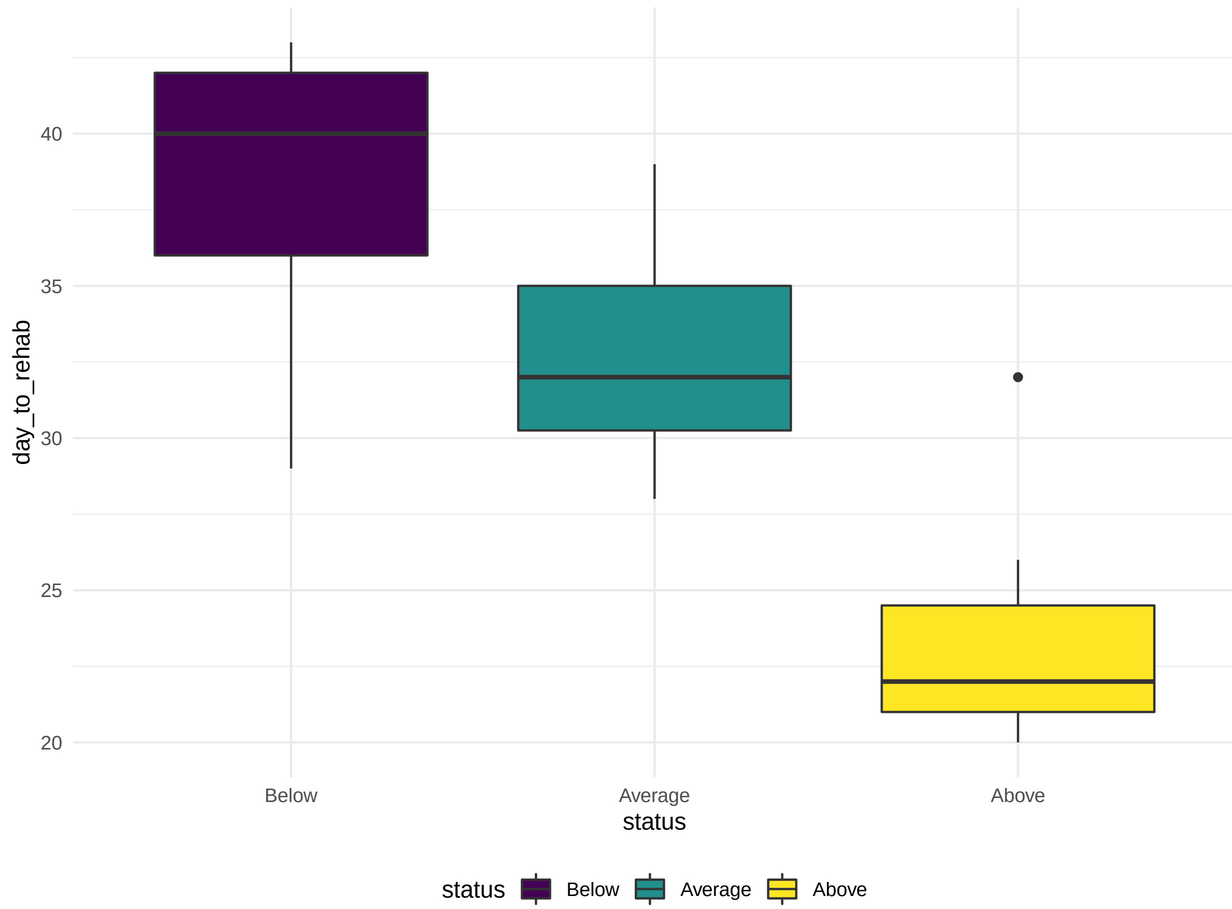
We tidy the original 'knee.csv' with 2 variables: 'status' and 'days'

- status: the physical status before therapy (3 levels: above, average, below)
- days: time required in physcial therapy until successful rehabilitation.) Therefore, the new dataset is call 'knee' with 30 observations and 5 missing values.

The mean required time in physical therapy until successful rehabilitation is longer in the physical status before therapy is categorized as 'below'.

The mean required time in physical therapy until successful rehabilitation is shorter in the physical status before therapy is categorized as 'above', except one observation.

Based on the box plots below, we see no overlapping between the 3 groups: below, average, above.



2(b) H_0: No significant difference among the population means for the 3 levels of status. H_1: At least one mean is different from the others.

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

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

$Total\ SS = Between\ SS\ +\ Within\ SS$

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

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

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

$Reject\ H_0\ if\ F > F_{k-1,n-k,1-\alpha};\ Fail\ reject\ H_0\ if\ F < F_{k-1,n-k,1-\alpha}$

$P - value:\ area\ to\ the\ right\ P(F_{k-1,n-k} > F).$

We obtain the ANOVA table as following:

```
##             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
```

At 0.1 significance level, the $F_{stat} = 19.3 > F_{crit} = 5.719$, we reject the null hypothesis and conclude that at least two of mean required time of the 3 levels are different.

2(c) Bonfferroni Adjustments: $\alpha^* = \frac{\alpha}{\binom{k}{2}}$

$Reject\ H_0 :\ if\ |t| > t_{n-k,\ 1-\frac{\alpha^*}{2}}$

$Fail\ to\ reject\ otherwise.$

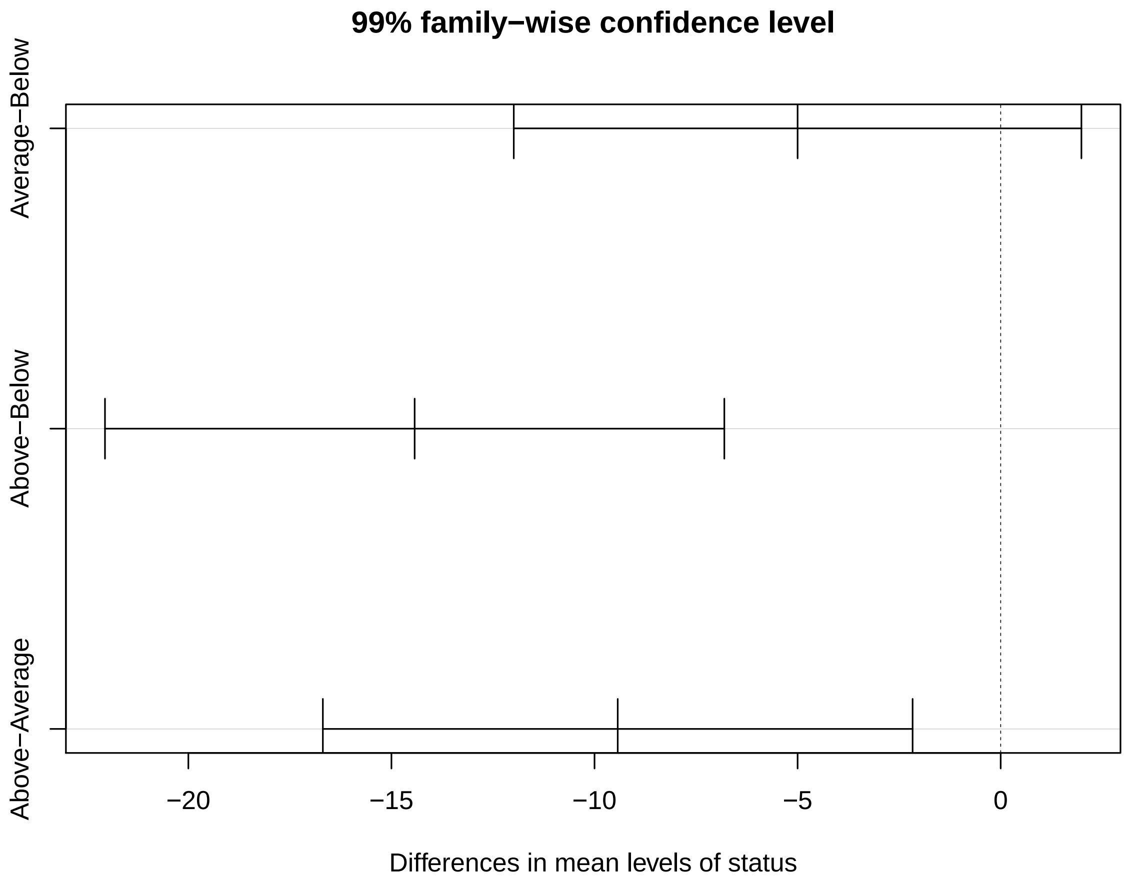
Bonferroni is the most conservative method, it is the most stringent in declaring significance (thus, less powerful).

```
##
##  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:

Tukey's method - controls for all pairwise comparisons and it is less conservative than Bonferroni. For Tukey, we need to use another function 'TukeyHSD' with an object created by aov(): 'knee_anova'

```
##   Tukey multiple comparisons of means
##     99% 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 -12.0  1.99 0.074
## Above-Below   -14.43 -22.1 -6.80 0.000
## Above-Average  -9.43 -16.7 -2.17 0.001
```



Dunnett's method: mainly focuses on comparisons wiht predefined control arms.

```
##
##   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)
```

After perfomming ANOVA and rejecting the null, it is often desired to know more about the specific groups and find out which ones are significantly different or similar. This step is usually referred to as "post-hoc analysis" Possible methods are: Bonfferroni, Tukey, and Dunnetts's methods. They all aim to control and preserve the overall (family-wise) error rate at the pre-specified alpha level.

2(d) Conclusion:

Based on the values of descriptive statitics, we find that the mean required time in physical therapy until successful rehabilitation is different based on the physical status before therapy. The mean required time is longer in the 'below' group and is shorter in the 'above' group. We further employ an ANOVA test to compare the mean required time for the 3 groups: below, average, and above. And we find that we are 99% confident that the mean required time in physcial therapy until successful rehabilitation of the 3 groups are different.

Problem 3

3(a) Identify Appropriate Test:

Since we have 2 categorical variables with more than 2 levels, we may employ the RxC Contingency Table and Chi-squared test. Moreover, since the distribution of swelling status is the same for the two treatment populations, it may suggest Chi-squared with honogemeity approach to evaluating the distribution/proportion between vaccine status and swelling symptom.

Assumption:

- independent random samples
- no expected cell counts are 0, and nor more than 20% of the cells have an expected counts less than 5.

3(b) Table: Observed values

	Major_Swelling	Minor_Swelling	No_Swelling	Total
Vaccine	54	42	134	230
Placebo	16	32	142	190
Total	70	74	276	420

Table: Expected Values

	Major_Swelling	Minor_Swelling	No_Swelling
Vaccine	38.3	40.5	151
Placebo	31.7	33.5	125

Pearson's Chi-squared test

data: table X-squared = 19, df = 2, p-value = 9e-05

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

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

Table 3: Expected Values

Major_Swelling	Minor_Swelling	No_Swelling
38.3	40.5	151
31.7	33.5	125

3(c) $H_0:$ the proportions of a 'major' swelling symptom in 'vaccine' and 'placebo' are equal ($p_{11} = p_{21}$); AND, the proportions of a 'minor' swelling symptom in 'vaccine' and 'placebo' are equal ($p_{12} = p_{22}$); AND, the proportions of a 'no' swelling symptom in 'vaccine' and 'placebo' are equal ($p_{13} = p_{23}$)

$H_1:$ not all proportions are equal.

$\chi^2 = \sum_i^R \sum_j^C \frac{(n_{ij}-E_{ij})^2}{E_{ij}}$ *under the null* $\sim \chi^2_{df=(R-1)\times(C-1)}$

$\chi^2 = \sum_i^2 \sum_j^3 \frac{(n_{ij}-E_{ij})^2}{E_{ij}}$
$= \frac{(54-38.3)^2}{38.3} + \frac{(42-40.5)^2}{40.5} + \frac{(134-151)^2}{151} + \frac{(16-31.7)^2}{31.7} + \frac{(32-33.5)^2}{33.5} + \frac{(142-125)^2}{125}$
$\cong 18.5601$
$\cong 19$

$\chi^2_{df=(R-1)\times(C-1),1-\alpha} = \chi^2_{(2-1)\times(3-1),1-0.05} = \chi^2_{2,\ 0.95} = 5.991$

Decision Rule:

$Reject\ H_0:\ if\ \chi^2 > \chi^2_{(R-1)*(C-1),1-\alpha}$

$Fail\ to\ reject\ H_0\ otherwise.$

Conclusion:

Because $\chi^2 > \chi^2_{2,0.95} = 5.991$, and the p_value is 9e-05, we reject the null hypothesis at 0.05 significance level, and conclude that the proportions of swell sympotoms in treatments: 'vaccine' and 'placebo' are not equal.