p8130_hw4_rw2844

Renjie Wei

10/31/2020

Problem 1

Proof:

$$\because \textit{by the definition, } y_{ij} - \bar{\bar{y}} = (y_{ij} - \bar{y_i}) + (\bar{y_i} - \bar{\bar{y}})$$

$$\therefore \sum_j (y_{ij} - \bar{\bar{y}})^2 = \sum_j [(y_{ij} - \bar{y_i}) + (\bar{y_i} - \bar{\bar{y}})]^2 = \sum_j [(y_{ij} - \bar{y_i})^2 + (\bar{y_i} - \bar{\bar{y}})^2 + 2 \times (y_{ij} - \bar{y_i})(\bar{y_i} - \bar{\bar{y}})]$$

$$\because \sum_j y_{ij}/n_j = \bar{y_i}, \textit{ and } \sum_j 1 = n_j$$

$$\therefore \sum_j (y_{ij} - \bar{y_i})(\bar{y_i} - \bar{\bar{y}}) = \sum_j [y_{ij} \times \bar{y_i} - y_{ij} \times \bar{\bar{y}} - \bar{y_i}^2 + \bar{y_i} \times \bar{\bar{y}}]$$

$$= n_j \times \bar{y_i}^2 - n_j \times \bar{y_i} \times \bar{\bar{y}} - n_j \times \bar{y_i}^2 + n_j \times \bar{y_i} \times \bar{\bar{y}}$$

$$= (n_j \times \bar{y_i}^2 - n_j \times \bar{y_i}^2) + (n_j \times \bar{y_i} \times \bar{\bar{y}} - n_j \times \bar{y_i} \times \bar{\bar{y}}) = 0$$

Then, we can prove that:

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

Problem 2

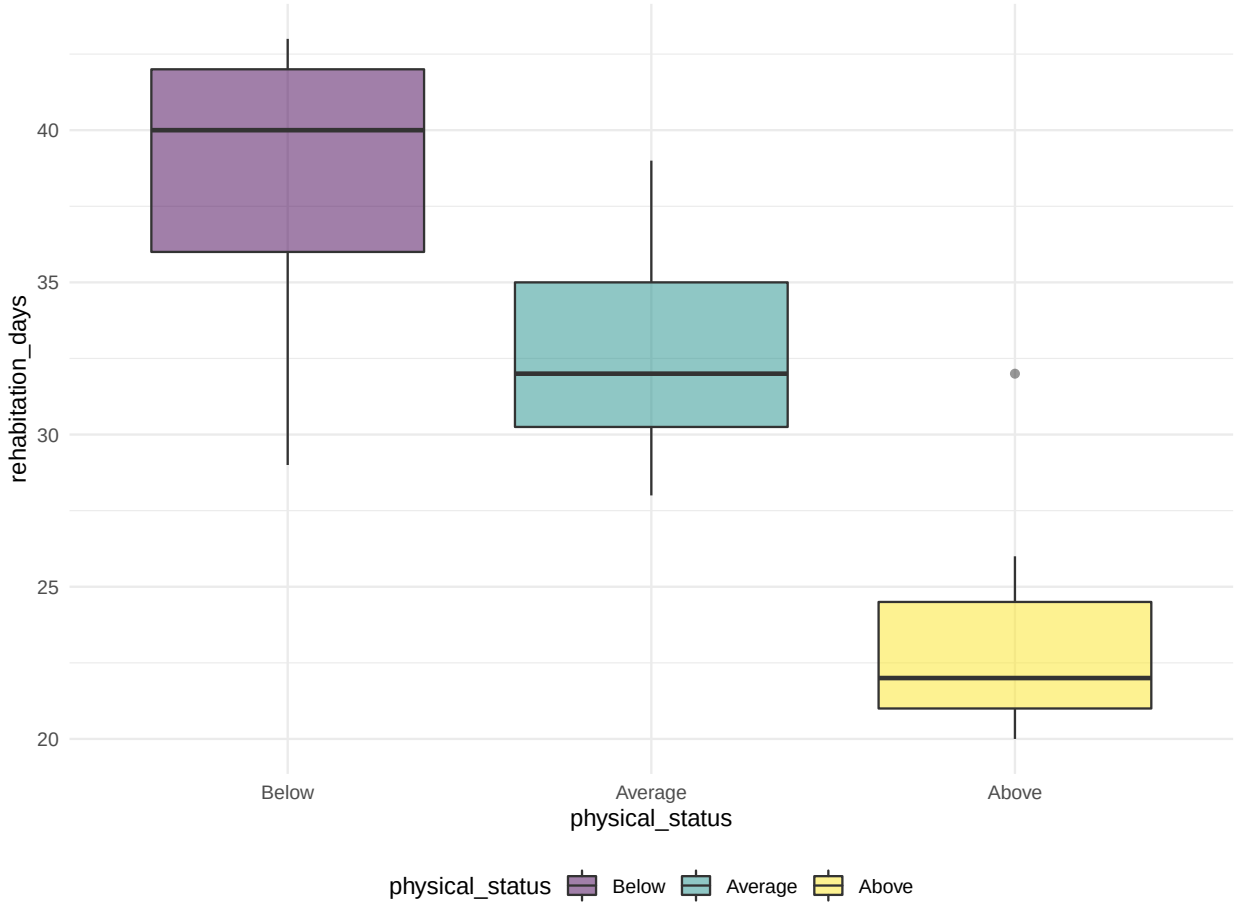
1

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

Comments:

From the table, we can see that the Mean of rehabilitation days are different in the three different physical condition groups.

To make it more clear, I made a box plot.



2

$H_0 : \mu_1 = \mu_2 = \mu_3$, *there is no difference in mean of the three physical condition groups*

H_1 : *at least two means are not equal*

```
##                 Df Sum Sq Mean Sq F value  Pr(>F)
## physical_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
```

$$Between\ SS = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(\bar{y_i} - \bar{\bar{y}})^2 = \sum_{i=1}^{k} n_i \bar{y_i}^2 - \frac{y_{..}^2}{n} = (38 - 31.96)^2 + (33 - 31.96)^2 + (23.57 - 31.96)^2 = 796$$

$$Within\ SS = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y_i})^2}{n-k} = \sum_{i=1}^{k}(n_i - 1)s_i^2 = 7 \times (5.48)^2 + 9 \times (3.92)^2 + 6 \times (4.20)^2 = 454$$

$$Between\ Mean\ Square = \frac{Between\ SS}{k-1} = 398$$

$$Within\ Mean\ Square = \frac{Within\ SS}{n-k} = 21$$

$$F_{stat} = \frac{Between\ Mean\ Square}{Within\ Mean\ Square} = 19.3 \sim F(k-1, n-k) = F(2, 22)$$

$$F_{crit} = F_{2,22,0.99} = 5.719$$

Decision Rule:

Reject H_0 if $F_{stat} > F_{crit}$

Fail to reject H_0 otherwise.

Conclusion:

Since $F_{stat} = 19.3 > F_{crit} = 5.791$, at 1% significance level, we reject the null hypothesis and conclude that at least two of mean rehabilitation days from the 3 physical condition groups are different.

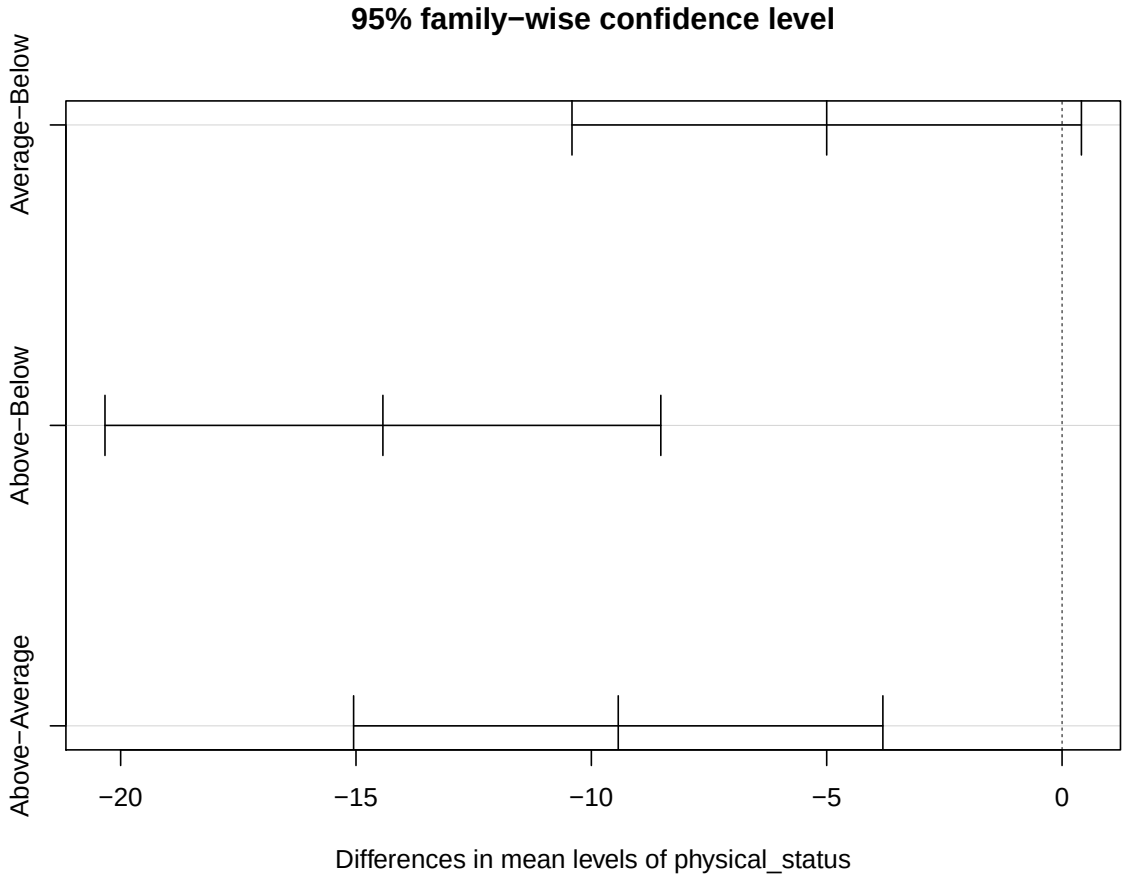
3

This is the Bonferroni adjust pairwise t-test:

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

This is the Tukey's test:

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = rehabitation_days ~ physical_status, data = knee_df, alpha = 0.01)
##
## $physical_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
```



This is the Dunnett's test:

```
##
##   Simultaneous Tests for General Linear Hypotheses
##
## Fit: aov(formula = rehabitation_days ~ physical_status, data = knee_df,
##     alpha = 0.01)
##
## Linear Hypotheses:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept) == 0               38.00       1.61   23.67   <0.001 ***
## physical_statusAverage == 0    -5.00       2.15   -2.32    0.068 .
## physical_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)
```

Comments:

- **Similarities:** All these three test are multiple comparisons adjustment methods. They are used to find out which two groups are different in mean.
- **Differences:** Tukey's method is less conservative than Bonferroni, which means it has a less p-values. And Dunnett's method need to define a control arm before the test.

4

The total mean of the days of rehabilitation is 31.96 days. And by using an ANOVA, we found that the rehabilitation days are not all the same between three different physical condition groups at 1% significance level, whereas whom with below average physical status takes 38 days, whom with average takes 33 days compared with 23 days in above average group. After multiple comparisons, we confirm the conclusion and we reject that the patients begin with different physical conditions need the same recover time.

Problem 3

1

We are going to compare the proportions between 3 different groups—- Major swelling, Minor swelling and No swelling. We should create a $R \times C$ Contingency table and conduct a Chi-Squared test of homogeneity.

2

Table 2: Observed Values

Groups	Major	Minor	No
Vaccine	54	42	134
Placebo	16	32	142

Table 3: Expected Values

Groups	Major	Minor	No
Vaccine	38.3	40.5	151
Placebo	31.7	33.5	125

3

Now we are going to use Chi-Square test:

$$H_0 : p_{1j} = p_{2j} = ... = p_{5j} = p_{.j}, i = 1, 2,\ j = 1, 2, 3, \textit{the proportions of different swelling conditions are the same among}$$

$$H_1 : p_{ij} \neq p_{i'j},\ j = 1, 2, 3, i \neq i',$$

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

$$\chi^2_{stat} = (54 - 38.3)^2/38.3 + (42 - 40.5)^2/40.5 + (134 - 151)^2/151 + (16 - 31.7)^2/31.7 + (32 - 33.5)^2/33.5 +$$

$$\chi^2_{crit} = \chi^2_{(R-1)}$$

Decision Rule:

RejectH_0 if $\chi^2_{stat} > \chi^2_{crit}$.

Fail to reject H_0 otherwise.

And the p-value is:

$$p - value = \int_{x^2}^{\infty} 2 * Z^2 = \int_{x^2}^{\infty} \frac{1}{\pi} e^{-s^2} ds$$
$$= e^{-x^2/2} = 9.277 \times 10^{-5}$$

Conclusion:

Since in our situation $\chi^2_{stat} > \chi^2_{crit}$, and the p-value is 9.277×10^{-5}we reject the null hypothesis at 5% significance level, and conclude that not all the proportions of different swelling conditions are the same among vaccine and placebo group. Or the swelling condition is significantly different between vaccine and placebo group.