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.,

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

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)

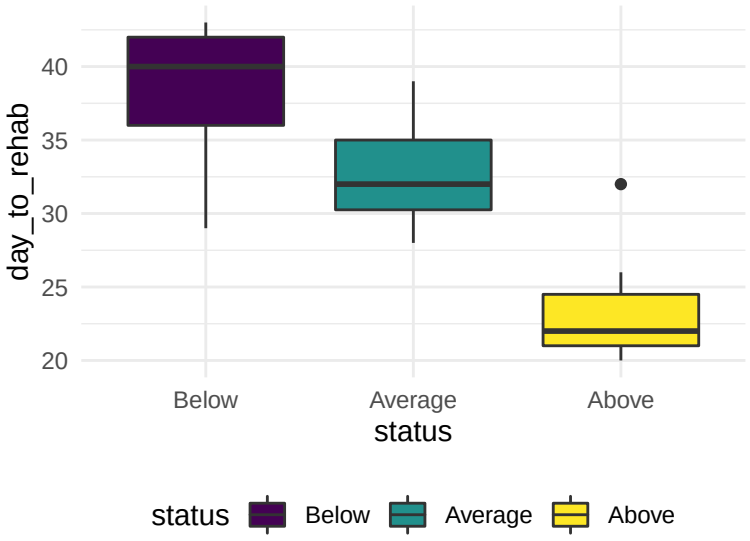
d) Write a short paragraph summarizing your overall results as if you were presenting to the rehabilitation center director. (1p)

PROOF

a)

Table 1: Descriptive statistics

	Below (N=8)	Average (N=10)	Above (N=7)	Total (N=25)
day_to_rehab				
Mean (SD)	38.000 (5.477)	33.000 (3.916)	23.571 (4.198)	31.960 (7.214)
Median (Q1, Q3)	40.000 (36.000, 42.000)	32.000 (30.250, 35.000)	22.000 (21.000, 24.500)	31.000 (28.000, 39.000)
Min - Max	29.000 - 43.000	28.000 - 39.000	20.000 - 32.000	20.000 - 43.000



We see that the mean rehabilitate days in group-ABOVE and groip-Below might be different.

b)

H_0 : there's no difference between groups

H_1 : at least one group is different from the other groups

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

$Within\ Sum\ of\ Square = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y_i})^2 = \sum_{i}^{k}(n_i - 1)s_i^2$

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

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

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

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

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

```
##             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 99% confidence level, the F-statistics is greater than the critical value(5.719), we reject the null hypothesis that there's no difference between groups.

c)

Bonferroni adjusted pairswide t-test

Bonfferroni adjusts the confidence level for each pairwise test with $\alpha^* = \frac{\alpha}{\binom{k}{2}}$.

```
##
##  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 Test

"Tukey's method – controls for all pairwise comparisons and it is less conservative than Bonferroni."

```
##   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 Test

"Dunnett's method – mainly focuses on comparisons with a pre-defined control arm."

```
##
##   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)
```

All three tests are posy-hoc analysis followed multi-group comparison. All adress the issue of control and preserve the overall (family-wise) error rate at the pre- specified alpha level. All test if one group is different/same with the other groups.

d)

The total mean time required in therapy is 31.9 days, whereas whom begin with below average physical status takes 38 days, compared with 33 and 23 days with groups average status and above average status. After ANOVA test, we don't have enough evidence that these difference are cause by chance at a 99% confidence level. Further testing between groups confirm this finding as we reject that the patients begin with Above average need the same recover time as those from below average or average physical status.

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

a)

we are examining the distribution/proportion between Vaccine status and Swelling symptom. And there're more than 2 groups, so Contingency table with Chi-sq test for homogeneity would be the testing method we consider.

b)

Table 2: Observed Values

vaccine_status	Major_Swelling	Minor_Swelling	No_Swelling
Vaccine	54	42	134
Placebo	16	32	142

Table 3: Expected Values

vaccine_status	Major_Swelling	Minor_Swelling	No_Swelling
Vaccine	38.3	40.5	151
Placebo	31.7	33.5	125

All expected values in the cells are greater than 5, the normality for Chi-sq test is fitted. We continue applying Chi-sq test.

c)

H_0 :p_{11} =p_{21};…p_{13}=p_{23}$ the proportion/distribution of swelling symptom among vaccine and not vaccined are equal …

H_1 : For at least one column there're two rows i and i' where the proportion are not the same.

$\mathcal{X}^2 = \sum_i^{row} \sum_j^{col} \frac{(n_{ij}-E_{ij})^2}{E_{ij}} \sim \mathcal{X}^2_{df=(row-1)\times(col-1)}$

$Reject\ H_0\ if\ \mathcal{X}^2 > \mathcal{X}^2_{(r-1))*(c-1),1-\alpha}$

$Fail\ reject\ H_0\ if\ \mathcal{X}^2 < \mathcal{X}^2_{(r-1))*(c-1),1-\alpha}$

$$\begin{aligned} p\ value &= \int_{x^2}^{\infty} Chi\_sq(\mathcal{X}, k=2) \\ &= \int_{x^2}^{\infty} 2 * Z^2 = \int_{x^2}^{\infty} \frac{1}{\pi} e^{-s^2} ds \\ &= e^{-x^2/2} = 9.277 \times 10^{-5} \end{aligned}$$

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

The Chi-sq statistics value is greater than the critical value(5.991) at 95% confidence level, with the P-Value of 9.277×10^{-5} compared to 0.05', so we reject the null hypothesis that the the swelling symptom is independent to vaccine status.