HW4

Anna Ma

11/16/2021

Problem 1

Problem 1

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

Proof.

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

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

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

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

Since $\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_i) = n_i\bar{y}_i - n_i\bar{y}_i = 0$, then

$$2\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_i)(\bar{y}_i-\bar{\bar{y}}) = 2\cdot\sum_{i=1}^{k}(\bar{y}_i-\bar{\bar{y}})\cdot\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_i) = 0$$

Thus

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

$$= \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_i)^2+\sum_{i=1}^{k}\sum_{j=1}^{n_i}(\bar{y}_i-\bar{\bar{y}})^2+0$$

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

□

Problem 2

```
library(tidyverse)
score_df = read_csv("./Crash.csv")
```

a) Descriptive Statistics

```
var_sd = function(score){

  score_sd = sd(score, na.rm = TRUE)

  score_var = var(score, na.rm = TRUE)

  output_df =
    tibble(
      sd = score_sd,
      variance = score_var
)
  return(output_df)
}


var_df =
  map(score_df, var_sd) %>%
  bind_rows(.id = "Crash Type")

desp_stat_result =
  map(score_df, summary) %>%
  bind_rows(.id = "Crash Type") %>%
  left_join(var_df, by = "Crash Type") %>%
  select(-"NA's")

desp_stat_result %>% knitr::kable(digits = 2)
```

Crash Type	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	sd	variance
pedestrian	29	36.0	39.5	37.87500	42.0	43	5.44	29.55
bicycle	28	29.5	31.5	32.50000	34.5	39	4.06	16.50
car	20	21.0	22.0	23.42857	24.5	31	3.87	14.95

From the table, we can observe that patients from car crashes have the lowest mean of 23.43, lowest median of 22, and the lowest standard deviation of 3.87 in their PTSD scores out of the three groups. On the other hand, patients from pedestrian crashes have the highest mean of 37.88, the highest median of 39.5, and the highest standard deviation 5.44 in their PTSD scores. As for patients from bicycle crashes, they have a mean of 32.5, median of 31.5, and standard deviation of 4.06, all of which are higher than the car crash group and lower than the pedestrian group. Generally, the score of car crash patients is lower than the other two groups and it has a more compact distribution where as the distribution of the other two groups are more spread.

b) Hypotheses Test with ANOVA

```
score_tidy_df =
  score_df %>%
  pivot_longer(
    cols = pedestrian:car,
    names_to = "crash_type",
```

```
    values_to = "PTSD_score"
  ) %>%
  mutate(
    crash_type = factor(crash_type))
```

Hypotheses:

$H_0 : \mu_{bicycle} = \mu_{car} = \mu_{pedestrian}$;

H_a : not all mean PTSD scores are equal across the three groups

```
anova_table =
  aov(PTSD_score ~ crash_type, data = score_tidy_df) %>%
  summary()

anova_table
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## crash_type   2  790.4   395.2   19.53 1.33e-05 ***
## Residuals   22  445.1    20.2
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 5 observations deleted due to missingness
```

The anova table can be written as the following:

Source	Sum of Square(SS)	DF	Mean Sum of Square	F statistics
Between	Between SS = 790.4	k-1 = 2	Between SS/(k-1)= 395.2	F = 19.53
Within	Within SS = 445.1	n-k = 22	Within SS/(n-k) = 20.2	
Total	Between SS+Within SS = 1235.5	n-1 = 24		

Here, $F_{stat} = \frac{\text{Between SS}/(k-1)}{\text{Within SS}/(n-k)} = \frac{395.2}{20.2} = 19.53$

The critical value $F_{critical} = F_{k-1,n-k,1-\alpha} = F_{2,22,0.99} = 5.72$. Since $F_{stats} = 19.53 > F_{crit} = 5.72$, then at $\alpha = 0.01$ significance level, we reject the null hypothesis and conclude that not all the group means of PTSD score are equal.

c) Pair-wise Comparison

The modified critical region $\alpha^* = \frac{\alpha}{\binom{k}{2}} = 0.01/3 = 0.0033$. We would reject the null if $|t| > t_{n-k,1-\frac{\alpha^*}{2}}$. In our case, $t_{critical} = t_{n-k,1-\frac{\alpha^*}{2}} = t_{22,1-\frac{0.0033}{2}} = t_{22,0.9983} = 3.2825323$

Use the Bonferroni adjustment for pairwise t test:

```
pairwise.t.test(score_tidy_df$PTSD_score, score_tidy_df$crash_type, p.adj = 'bonferroni')
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  score_tidy_df$PTSD_score and score_tidy_df$crash_type
##
##            bicycle car
## car        0.0014  -
## pedestrian 0.0586  9.1e-06
##
## P value adjustment method: bonferroni
```

From the result we observe that $p_{car-bicycle} = 0.0014 < \alpha^*$. Therefore, we can reject the null hypothesis and conclude that the mean PTSD score of patients from car crashes and bicycle crashes are different. Similarly, since $p_{car-pedestrian} = 9.1 * 10^{-6} < \alpha^*$, we can reject the null and conclude that the mean PTSD score of patients from car crashes and pedestrian crashes are different. Finally, since $p_{pedestrian-bicycle} = 0.0586 > \alpha^*$, we fail to reject the null hypothesis and conclude that the mean PTSD score of patients from bicycle crashes and pedestrian crashes is not significantly different.

d) Summary of Result

From the analysis we can observe that patients suffering from car crashes generally have the lowest PTSD score with a mean of 23.428; whereas patients from pedestrian crashes have the highest PTSD score with a mean of 37.875. Patients from bicycle crashes have a average PTSD score of 32.5. Moreover, patients from the car group has the lowest standard deviation of 3.87; patients from bicycle crashes has a standard deviation of 4.06; whereas the patients from pedestrian crashes has the highest standard deviation of 5.44.This indicates that the patients from pedestrian crashes has the biggest variance among the three groups.

After conducting ANOVA test, we are 99% confident that the mean PTSD scores of the three groups is different. After the Bonferroni adjustment, the pair wise t test shows uthe mean PTSD score of patients from car crashes and bicycle crashes, and patients from car crashes and pedestrian crashes are different. However, we do not have sufficient evidence to show that the mean PTSD score of patients from bicycle crashes and pedestrian crashes is significantly different.

Problem 3

a) Appropriate Test

We can use the Chi-square test for homogeneity. This is because we want to compare the three independent proportions, namely the proportion of relapse among subjects who used desipramine, lithium, and placebo to break their drug habit. We can use this test because the subjects are evenly assigned to the groups so the row totals are fixed. Also, in the problem, samples are random and independent; and there's no expected cell counts of 0 and no more than 20% of the cells have an expected count less than 5.

b) Test Table

The observed data of the study can be shown as:

```
relapse_df =
  tibble(
    drug = c("desipramine","lithium","placebo"),
    n_relapsed = c(15,18,20),
    n_nonrelapse = c(18,15,13),
    r_total = n_relapsed + n_nonrelapse,
  )

c_total = c("c_total", colSums(relapse_df[,-1]))

rbind(relapse_df,c_total) %>% knitr::kable()
```

drug	n_relapsed	n_nonrelapse	r_total
desipramine	15	18	33
lithium	18	15	33
placebo	20	13	33
c_total	53	46	99

The expected data of the study can be shown as:

Drug	Relapsed	No relapse	Total
desipramine	$E_{11} = 53 * 33/99 = 17.67$	$E_{12} = 46 * 33/99 = 15.33$	33
lithium	$E_{21} = 53 * 33/99 = 17.67$	$E_{22} = 46 * 33/99 = 15.33$	33
placebo	$E_{31} = 53 * 33/99 = 17.67$	$E_{32} = 46 * 33/99 = 15.33$	33
Total	53	46	99

c) Test

Hypothesis:

$H_0 : p_{11} = p_{21} = p_{31}$, the proportions of relapse among desipramine, lithium and placebo groups are equal; and $p_{12} = p_{22} = p_{32}$, the proportions of non-relapse among desipramine, lithium and placebo groups are equal.

H_a: not all the proportions are equal.

Test Statistics:

$$\begin{aligned}\chi^2 = \sum_{i=1}^{3}\sum_{j=1}^{2}\frac{(O_{ij}-E_{ij})^2}{E_{ij}} =& \frac{(15-17.67)^2}{17.67} + \frac{(18-17.67)^2}{17.67} + \frac{(20-17.67)^2}{17.67} \\ &+ \frac{(18-15.33)^2}{15.33} + \frac{(15-15.33)^2}{15.33} + \frac{(13-15.33)^2}{15.33} \\ &= 1.543\end{aligned}$$

Critical value: $\chi^2_{(R-1)(C-1),1-\alpha} = \chi^2_{(3-1)(2-1)} = \chi^2_{2,0.95} = 5.99$

P value: The p-value is 0.4630131

Conclusion: Since the $\chi^2_{stat} < \chi^2_{2,0.95}$, we fail to reject the null hypothesis. Therefore, we conclude that at a significant level of $\alpha = 0.05$, the proportions of relapse among the desipramine, lithium, and placebo groups are equal; and that the proportions of non-relapse among these three groups are equal as well.