hw4_bm_wq2160

Wenshan Qu (wq2160)

11/11/2021

Problem 1

We want to prove $\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\bar{\bar{y}})^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\bar{y_l})^2 + \sum_{i=1}^{k}\sum_{j=1}^{n_i}(\bar{y_l}-\bar{\bar{y}})^2$

Which could also be seen as $TotalSS = WithinSS + BetweenSS$

Note that:

k: number of groups;

n_i: number of observations in the i^{th} group;

y_{ij}: denotes the observation of the j^{th} subject from the i^{th} group;

$\bar{\bar{y}}$: grand mean;

$\bar{y_l}$: mean from the i^{th} group.

Proof:

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

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

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

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

(Becasue $\bar{y_l} = \frac{\sum_{j=1}^{n_i} y_{ij}}{n_i}$, then $\sum_{j=1}^{n_i}(y_{ij}-\bar{y_l}) = \sum_{j=1}^{n_i} y_{ij} - n_i\bar{y_l} = 0$,)

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

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

$= WithinSS + BetweenSS$

Problem 2

Read data

```
raw_df =
  read_csv("./Crash.csv")
```

(a) Descriptive Statistics

First we take a quick look on the raw data:

```
raw_df %>%
  knitr::kable(align = "c")
```

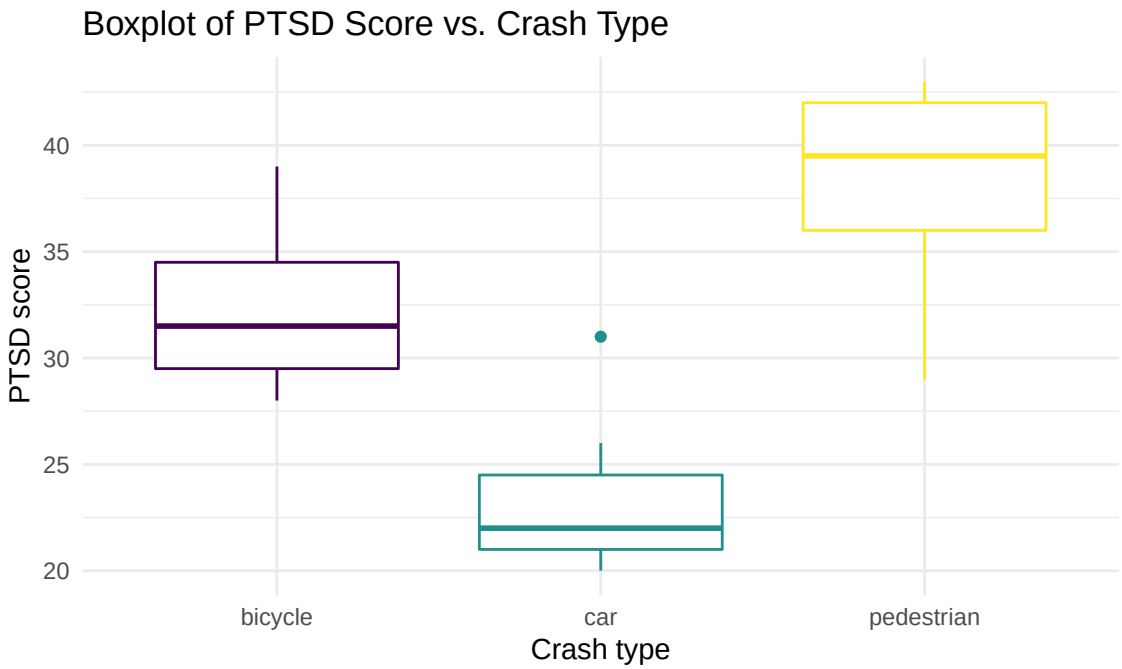
pedestrian	bicycle	car
29	29	26
42	32	31
38	39	21
40	28	20
43	31	23
39	31	22
30	28	21
42	35	NA
NA	39	NA
NA	33	NA

Tidy data:

```
crash_df =
  raw_df %>%
    pivot_longer(
      pedestrian:car,
      names_to = "crash_type",
      values_to = "PTSD_score"
    ) %>%
    mutate(
      crash_type = as.factor(crash_type)
    ) %>%
  arrange(crash_type) %>%
  drop_na(PTSD_score)
```

Boxplot the data:

```
crash_df %>%
  group_by(crash_type) %>%
  ggplot(aes(x = crash_type, y = PTSD_score, color = crash_type)) +
  geom_boxplot() +
  theme(legend.position = "none") +
  labs(
    title = "Boxplot of PTSD Score vs. Crash Type",
    x = "Crash type",
    y = "PTSD score"
  )
```



Measures of **Location** and **Dispersion** for each group:

```
raw_df %>%
  summary() %>%
  knitr::kable(align = "c")
```

pedestrian	bicycle	car
Min. :29.00	Min. :28.0	Min. :20.00
1st Qu.:36.00	1st Qu.:29.5	1st Qu.:21.00
Median :39.50	Median :31.5	Median :22.00
Mean :37.88	Mean :32.5	Mean :23.43
3rd Qu.:42.00	3rd Qu.:34.5	3rd Qu.:24.50
Max. :43.00	Max. :39.0	Max. :31.00
NA's :2	NA	NA's :3

Also we can manually calculate `mean`, `sd` like this:

```
crash_df %>%
  group_by(crash_type) %>%
  summarize(
    n = n(),
    mean = mean(PTSD_score),
    sd = sd(PTSD_score),
    sum = sum(PTSD_score)
  ) %>%
  knitr::kable(align = "c")
```

crash_type	n	mean	sd	sum
bicycle	10	32.50000	4.062019	325
car	7	23.42857	3.866831	164
pedestrian	8	37.87500	5.436320	303

And manually calculate `median`, `quartiles` like this:

For `pedestrian`:

```
## arranged data: 29 30 38 39 40 42 42 43

## Median
(39 + 40)/2
## [1] 39.5

## Q1 (25th percentile)
(30 + 38)/2
## [1] 34

## Q3 (75th percentile)
(42 + 42)/2
## [1] 42
```

For `bicycle`:

```
## arranged data: 28 28 29 31 31 32 33 35 39 39

## Median
(31 + 32)/2
## [1] 31.5

## Q1 (25th percentile)
29
## [1] 29

## Q3 (75th percentile)
35
## [1] 35
```

For `car`:

```
## arranged data: 20 21 21 22 23 26 31

## Median
22
## [1] 22

## Q1 (25th percentile)
(21 + 21)/2
## [1] 21

## Q3 (75th percentile)
(23 + 26)/2
## [1] 24.5
```

Comment: based on the boxplot and mean values, we found that the average PTSD scores of 3 crash types are *car* < *bicycle* < *pedestrian* in our samples.

(b) ANOVA Test

Hypothesis:

$H_0 : \mu_1 = \mu_2 = \mu_3, where\ \mu_1, \mu_2, \mu_3\ represents\ the\ mean\ of\ PTSD\ scores\ of\ 3\ crash\ types.$

$H_1 : at\ least\ two\ means\ are\ not\ equal.$

```
res1 = aov(PTSD_score ~ factor(crash_type), data = crash_df)
summary(res1)
```

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

Based on the test, we can get ANOVA table:

Source	Sum of Square (SS)	Degrees of freedom	Mean Sum of Square	F-Statistics
Between	790.4	k – 1 = 2	395.2	19.53
Within	445.1	n – k = 22	20.2	
Total	1235.5	n – 1 = 24		

Figure 1: anova

The `test statistic` is 19.53.

And for $n = 10 + 7 + 8 = 25$, $k = 3$, $\alpha = 0.01$.

Then the F critical value is:

```
qf(0.99, df1 = 2, df2 = 22)
```

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

$F_{k-1,n-k,1-\alpha} = F_{2,22,0.99} = 5.719$

Because $F = 19.53 > F_{2,22,0.99} = 5.719$, reject the null hypothesis, and conclude that with pre-specified significant level of 0.01, at least two types of crash (among `car`, `bicycle` and `pedestrian`) have different mean PTSD scores.

(c) Multiple Comparison

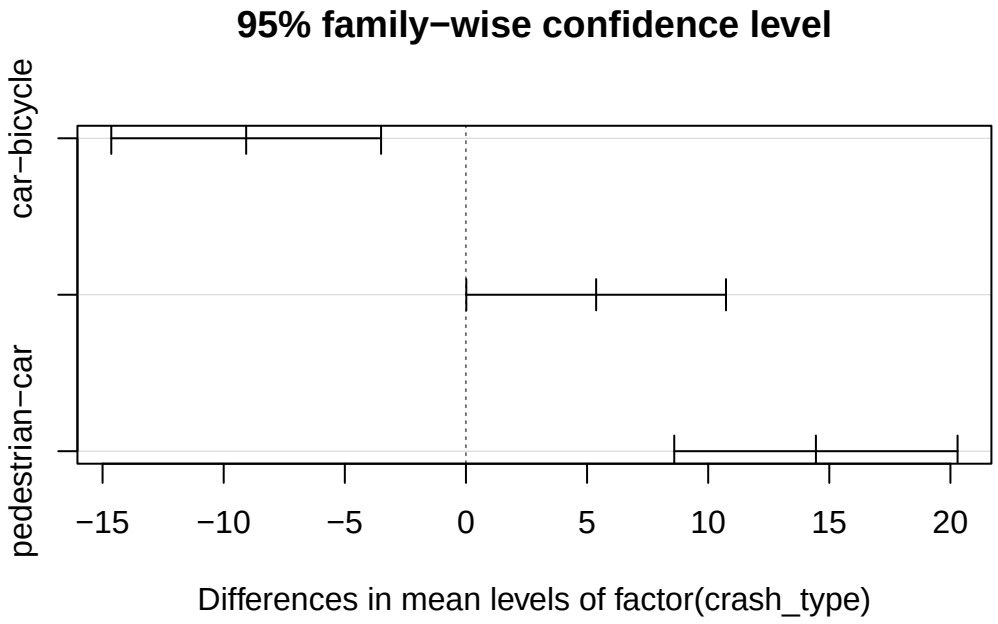
```r
# use pairwise.t.test() for Bonferroni
pairwise.t.test(crash_df$PTSD_score, crash_df$crash_type, p.adj = 'bonferroni')
```

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

```r
# use Tukey
Tukey_comp = TukeyHSD(res1)
Tukey_comp
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = PTSD_score ~ factor(crash_type), data = crash_df)
##
## $`factor(crash_type)`
##                          diff          lwr       upr     p adj
## car-bicycle        -9.071429 -14.63967214 -3.503185 0.0013441
## pedestrian-bicycle  5.375000   0.01537946 10.734621 0.0492580
## pedestrian-car     14.446429   8.59860314 20.294254 0.0000088
```

```r
plot(Tukey_comp)
```



```
# use Dunnett's
```

Though the accurate p-value may be slightly different among different methods, we can still conclude that:

1) The p-value between `pedestrian` and `car` group is around $9^{-6} \ll 0.05$, thus the mean PTSD scores between pedestrian crash and car crash are highly significantly different.
2) The p-value between `pedestrian` and `bicycle` group is around $0.05 \approx 0.05$, thus the mean PTSD scores between pedestrian crash and bicycle crash have NO significant difference.
3) The p-value between `car` and `bicycle` group is around $0.0014 \ll 0.05$, thus the mean PTSD scores between bicycle crash and car crash are highly significantly different.

(d) Summary

After analyzing the PTSD scores of 25 patients aging from 18-30 suffering from different types of crash (8 pedestrian crash, 10 bicycle crash and 7 car crash), we found that the mean PTSD scores are NOT statistically same among 3 crash types. Specifically, significant differences of mean PTSD scores have shown between `car crash` and `bicycle crash`, as well as between `car crash` and `pedestrian crash`, while there are NO significant difference between `pedestrian crash` and `bicycle crash`. Generally, we can conclude that the mean PTSD score of `car crash` is lower than other crash types.

Problem 3

(a) Justify Test

I choose **Chi-Squared Test** to address this problem.

Since

1) Three treatment groups are independent random samples;
2) No expected cell counts are 0, and expected values in each cell ≥ 5.

(b) Table

The table is given below:

Drug	Relapse	NOT_Relapse	Total
Desipramine	15 (E = 17.67)	18 (E = 15.33)	33
Lithium	18 (E = 17.67)	15 (E = 15.33)	33
Placebo	20 (E = 17.67)	13 (E = 15.33)	33
Total	53	46	99

About the `E` calculation process, for `Relapse` group, both are $\frac{(53\times33)}{99}$; and for `NOT Relapse` group, both are $\frac{(46\times33)}{99}$.

(c) Hypothesis Test

State the Hypothesis:

H_0 : *the probability of a subject's relapse has NO association with the drug the subject was assigned to. i.e.* $p_1 = p_2 = p_3$.

H_1 : *the probability of a subject's relapse is associated with the drug the subject was assigned to.*

With significant level $\alpha = 0.05$ pre-specified, compute the test statistic:

$\chi^2 = \sum_{i=1}^{3}\sum_{i=1}^{2}\frac{(O_{ij}-E_{ij})^2}{E_{ij}}$

$= \frac{(15-17.67)^2}{17.67} + \frac{(18-17.67)^2}{17.67} + \frac{(20-17.67)^2}{17.67} + \frac{(18-15.33)^2}{15.33} + \frac{(15-15.33)^2}{15.33} + \frac{(13-15.33)^2}{15.33}$

$= 1.543$

$follows\ \chi^2_{(3-1)\times(2-1)=2}\ under\ H_0$

Critical Value:

```
qchisq(0.95, df = 2)
```

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

$\chi^2_{2,0.95} = 5.991$

For p-value:

```
## Define data
drug_data = matrix(c(15, 18, 18, 15, 20, 13), nrow = 3,ncol = 2,byrow = T,
                   dimnames = list(c("Desipramine", "Lithium", "Placebo"),
                                   c("Relapse","NOT Relapse")))

# Perform test
chisq.test(drug_data)
```

```
##
##  Pearson's Chi-squared test
##
## data:  drug_data
## X-squared = 1.5431, df = 2, p-value = 0.4623
```

Thus the $p-value = 0.4623$.

Decision and Interpretation:

Given $\chi^2 = 1.543 < \chi^2_{2,0.95} = 5.991$, and $p-value = 0.4623 > 0.05$, we fail to reject H_0, and conclude that at the significance level of 0.05, there is no significant association between the probability of a subject's relapse and the drug the subject was assigned to.