Homework4-tk2886-fall2021

Loading the Crash data

```
crash_df <-
  read.csv("Crash.csv")
```

Tidying the data

```
crash_dfn <-
  crash_df %>%
  pivot_longer(everything(),
               names_to = "type_of_accidents",
               values_to = "Values")
```

Problem 2a

Generating Descriptive statistics for each group:

```
crash_df %>%
  summary() %>%
  knitr::kable(caption = "Mean, Median, Min, Max, 1st & 3rd Quartile Values for each type of crash")
```

Table 1: Mean, Median, Min, Max, 1st & 3rd Quartile Values for each type of crash

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

```
crash_df %>%
  summarize_if(is_numeric, sd, na.rm = T) %>%
  knitr::kable(caption = "*Standard deviation* for each type of crash")
```

Table 2: *Standard deviation* for each type of crash

pedestrian	bicycle	car
5.43632	4.062019	3.866831

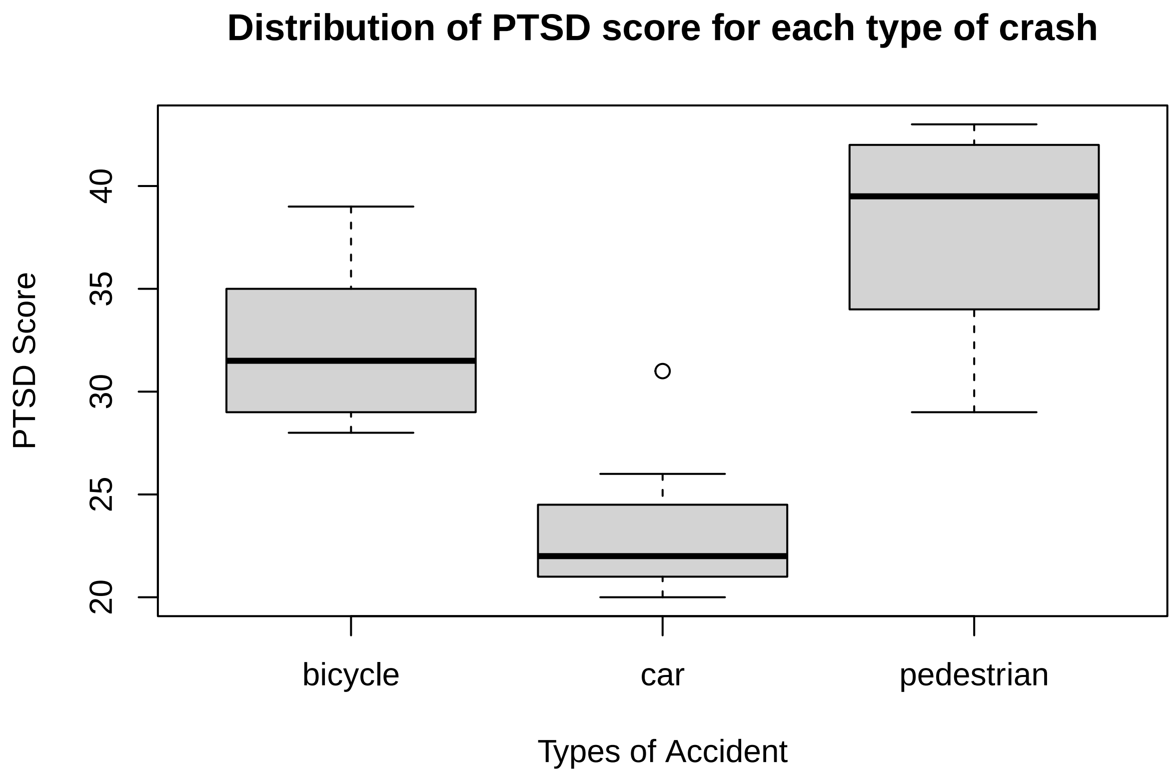
```
crash_dfn %>%
  group_by(type_of_accidents) %>%
  summarise(n = n(), mean = mean(Values, na.rm = T),
            sum = sum(Values, na.rm = T), variance = var(Values, na.rm = T)) %>%
  knitr::kable(caption = "Mean, Variance for each type of crash")
```

mean, sum, variance

Table 3: Mean, Variance for each type of crash

type_of_accidents	n	mean	sum	variance
bicycle	10	32.50000	325	16.50000
car	10	23.42857	164	14.95238
pedestrian	10	37.87500	303	29.55357

```
boxplot(Values ~ type_of_accidents, data = crash_dfn,
        main = "Distribution of PTSD score for each type of crash",
        xlab = "Types of Accident",
        ylab = "PTSD Score")
```



Analysis of Differences Observed:

In the data set that was provided, the mean of the PTSD Score for Pedestrian Incidents is the largest among the three types of accidents. The mean of the PTSD Score of bicycle incidents is the second highest and the mean of the PTSD Scores for car incidents is the lowest. The standard deviation for PTSD score for pedestrian incident is 5.44, the standard deviation for PTSD score in bicycle incident is 4.06, and the standard deviation for the PTSD score for car crash is 3.87. Pedestrian has a larger standard deviation than bicycle and car. This indicates that the PTSD score for this pedestrian incidents is more spread out compared to the other two types if crash. When pedestrians is involved in an incident, their PTSD in general is more varied compared to the other groups (bicycle and cars).

Problem 2b:

```
res1 = aov(Values ~ factor(type_of_accidents), data = crash_dfn)
summary(res1)
```

ANOVA TABLE

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

```
# Using R-code to obtain critical value
critical_value = qf(0.99, 2, 22)
```

Interpretation:

Hypothesis:

Ho: mu1 = mu2 = mu3

Ha: at least two means are not equal

Test-Statistics: F-Value: 19.53

Critical Value: Critical Value: 5.7190219

Interpretation in context to our problem: Our F-statistics (19.53) is bigger than our critical value (5.72), we reject the null hypothesis. At 0.01 significance level, we reject the null hypothesis and conclude that at least two of true mean PTSD score from the three type of crash groups are different.

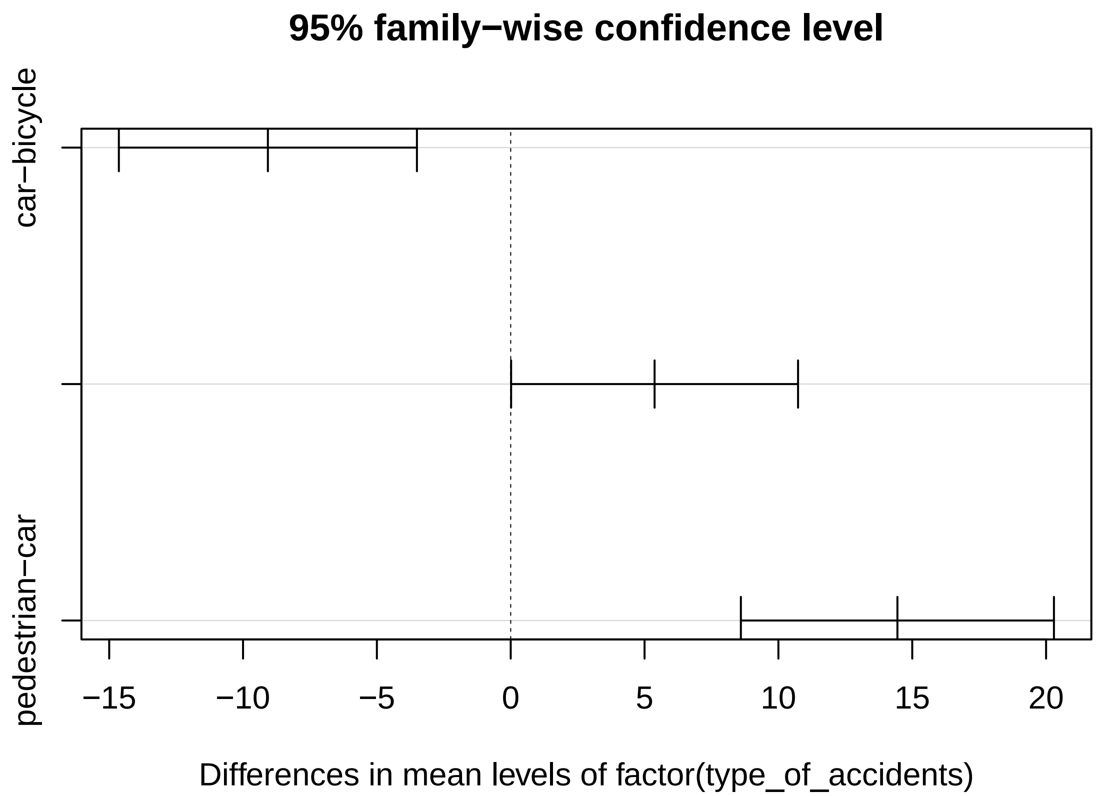
Problem 2c:

```r
Tukey_comp = TukeyHSD(res1)
Tukey_comp
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = Values ~ factor(type_of_accidents), data = crash_dfn)
##
## $`factor(type_of_accidents)`
##                         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
```

Analysis: We will be using be Tukey's adjustment since Bonferroni is the most conservative method and gives us less power while Tukey's method controls for all pairwise comparisons and it is less conservative. Based on our Tukey's adjustment method, it is indicated that the the true mean PTSD score for car and bicycle is different since the adjusted p-value for car-bicycle pairwise is 0.00134 which is smaller than our alpha 0.01. Also the the true mean PTSD score for pedestrian and car is different since the adjusted p-value for pedestrian-car pairwise is 0.0000088 which is smaller than our alpha 0.01. Based on the tukey's method, we do not have enough evidence to state that the true mean PTSD score for pedestrian-bicycle pairwise is different. *It is important to note that in lecture, we did not go over assumptions for these tukey and bonferroni and I emailed the professor and the professor stated that either tukey or bonferroni will be valid to use for this problem.*

```
plot(Tukey_comp)
```



Problem 2d:

Based on the data set provided for each type of crash (bicycle, car, pedestrian), our statistical analysis indicated that the mean PTSD score for pedestrian is 37.88 and the mean PTSD score bicycle crash is 32.5. It has been reported by the National Center for PTSD that a PTSD score of 31-33 or higher suggest the patient may benefit from PTSD treatment. Emergency Department physicians may provide additional resources or a better catered treatment plan to individuals involved in pedestrian and bicycle incidents/crashes for reducing their PTSD symptoms. In our statistical analysis, the mean PTSD score for car is 23.43 and it has been reported by the National Center for PTSD that scores lower than 31-33 may indicate the patient either has sub threshold symptoms of PTSD or does not meet criteria for PTSD. Emergency department physicians should be aware of which type of crash has higher average PTSD score and the physicians should give proper or more medical resources to those groups (which in our case is pedestrian and bicycle group). Furthermore, our statistical finding indicates that we do not have enough evidence to state that pedestrian and bicycle mean for PTSD score are different. In both groups, the calculated mean and median PTSD score is above 31 and the emergency department physicians need to have a provide better treatment towards these two groups (pedestrian and bicycle incidents) so that these individuals have less difficult time in recovering after being involved in or experiencing a terrifying crash.

Problem 3a:

The appropriate test I used to address this question of interest is **Chi-Squared of Independence**. Our observational units are collected at random from a population, we are not gathering the data by randomly sampling from each sub-group separately, which is the case of Chi-Squared Test for Homogeneity. Also, we have two categorical variables (relapse and non-relapse) that are being observed for each unit (desipramine users, lithium users, and placebo users). Also we're interested in whether the knowledge of one variable (antidepressants) value provides an information about the value of the other variable (relapse status), i.e. are these two variables independent. Chi-Squared of Homogeneity assesses whether the pattern of relapse was different between the three groups of antidepressants.

Equations that will be used for problem 3:

Under the null hypothesis:

Test Statistic Formula:

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

Critical Value Formula:

$$\chi^2_{(R-1)\times(C-1),\,1-\alpha}$$

We do not need to used Yates' Continuity Correction, since that correction is only for 2x2 tables.

Assumptions for Chi-Squared:

1. Independent Random Sample: Question states "randomly assigned"
2. No expected cell counts are 0, and no more than 20% of the cells have an expected count less than 5: *After creating the table, there is no cells with values of 0 and there is no cells with values less than 5.*

Problem 3b:

Chi-Squared: Test of Independence Table

	Relapse	Non-Relapse	Sum
Desipramine	15 $\frac{53 \cdot 33}{99} = 17.67$	18 $\frac{46 \cdot 33}{99} = 15.33$	33
Lithium	18 $\frac{53 \cdot 33}{99} = 17.67$	15 $\frac{46 \cdot 33}{99} = 15.33$	33
Placebo	20 $\frac{53 \cdot 33}{99} = 17.67$	13 $\frac{46 \cdot 33}{99} = 15.33$	33
Sum	53	46	99

We may use R-code to get the Chi-Squared: Test of Independence Table

```
antidepressant_df = matrix(c(15, 18, 18, 15, 20, 13), nrow=3,ncol=2,byrow=T,
                  dimnames = list(c("Desipramine","Lithium","Placebo"),
                                 c("Relapse","Non-Relpase")))

antidepressant_dfa = addmargins(antidepressant_df)
antidepressant_dfa %>% knitr::kable()
```

	Relapse	Non-Relpase	Sum
Desipramine	15	18	33
Lithium	18	15	33
Placebo	20	13	33
Sum	53	46	99

Problem 3c:

```r
#Using R code to get test-statistics, p-value, degrees od freedom
new_var = chisq.test(antidepressant_df)
new_var
```

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

```r
new_var$expected %>% knitr::kable()
```

	Relapse	Non-Relpase
Desipramine	17.66667	15.33333
Lithium	17.66667	15.33333
Placebo	17.66667	15.33333

```r
#Test Statistics
test_statistic = (15-17.67)^2/17.67 +
  (18-17.67)^2/17.67 + (20-17.67)^2/17.67 +
  (18-15.33)^2/15.33 + (15-15.33)^2/15.33 + (13-15.33)^2/15.33
```

```r
#Critical Value
critical_value = qchisq(.95, 2)
```

```r
#p-value
pval = pchisq(test_statistic, 2, lower.tail = FALSE)
```

Hypotheses:

Ho: Relapse Status and types of anti-depressant are independent (p1=p2=p3)

Ha: Relapse Status and types of anti-depressant are associated/dependent

Test Statistics: The test statistic is: 1.5431165

Critical Value: The critical value is: 5.9914645

P-Value: The p-value is: 0.4622921

Interpretation in context to our problem: At 0.05 significance level, we fail to reject the null hypothesis because the Chi-squared test value is smaller than our critical value. We conclude that we do not have enough evidence that the subject's relapse status is associated with the antidepressant drug the subject was assigned to.