Road Traffic Accidents Data Analysis using R

Student ID: 201081646

1 Introduction

This report is the second assessment of the **MATH5741M Statistical Theory and Methods** module. Its aim is to answer statistically three questions regarding a road traffic accidents dataset collected by the UK Department for Transport in 2005.

All the analysis has been done using **R** (programming language) and is code reproducible. To see the complete coding process for data preparation and analysis visit https://github.com/eugenividal/Road-Traffic-Accidents-Data-Analysis

2 Results

2.1 Question 1

In this question, first, we are asked to draw a boxplot to compare the number of vehicles involved in urban areas with the number invlolved in rural areas.

To plot the boxplot, we first remove the unallocated values, and then we transform the data due to the fact that it is not symetric. It is very skewed to the left (see histogram 1). The transofrmation that we use is log10.

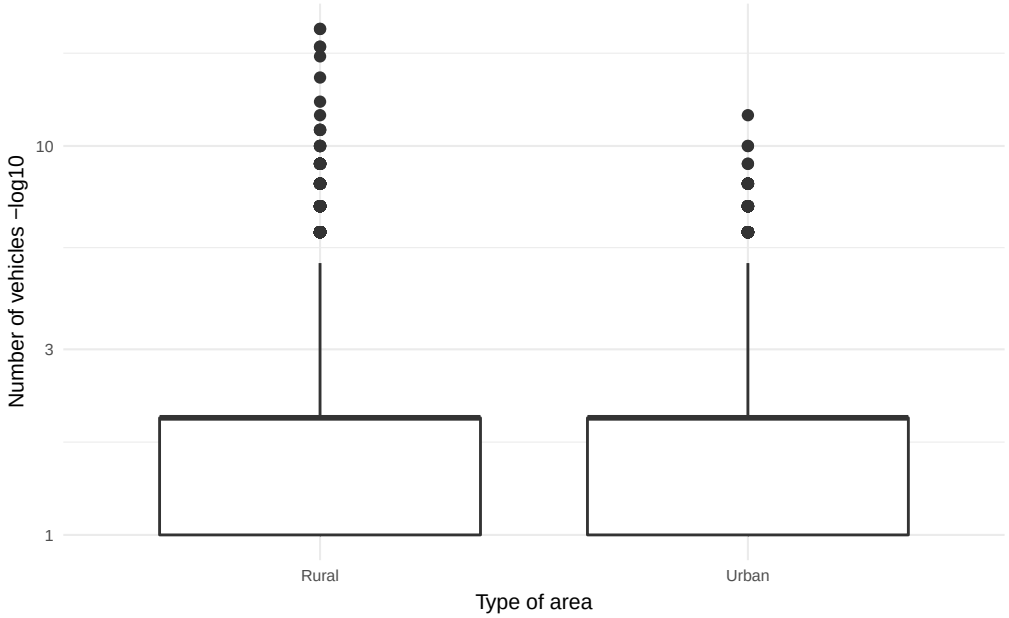


Figure 1: Number of vehicles involved grouped by type of area

Secondly, we have to carry out a test to investigate wheter the average number of vehicles in an accident differs per type of area.

For this, using the transformation we perform a pooled variance test (https://statistics.laerd.com/statistical-guides/independent-t-test-statistical-guide.php)

$$H_0 : \mu_{urban} = \mu_{rural} \quad vs. \quad H_1 : \mu_{urban} \neq \mu_{rural}$$

They are not about the same, should we continue with a two sample test which assumes equal variance? The independent t-test assumes the variances of the two groups you are measuring are equal in the population. If your variances are unequal, this can affect the Type I error rate. The assumption of homogeneity of variance can be tested using Levene's Test of Equality of Variances.

```
# pooled variance test
#t.test(xx[])
```

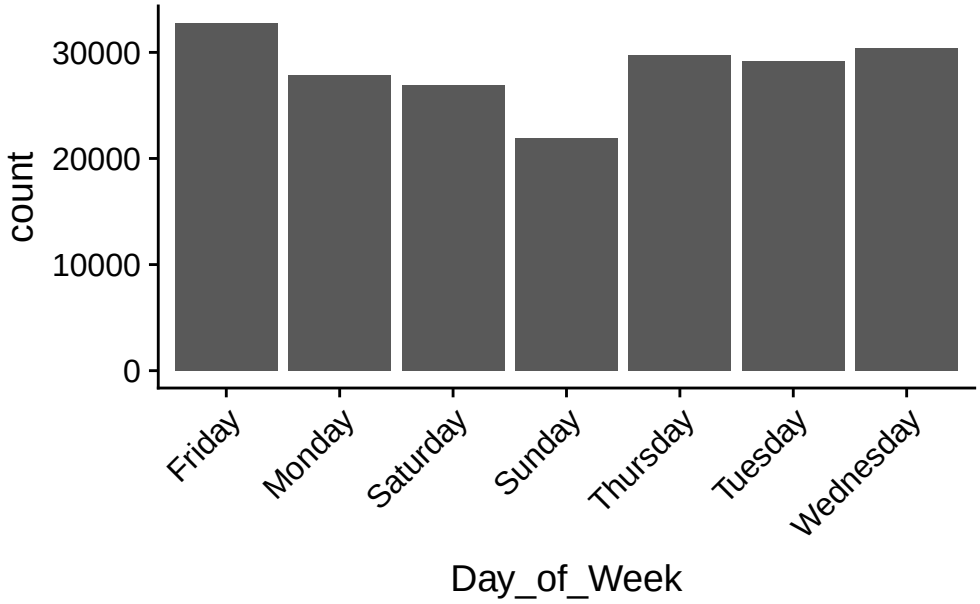
In conclusion, we can say that

2.2 Question 2

In this question, first, we have to investigate whether the frequency of accidents varies by day of the week using the suitable statistical hypothesis test.

For this, we apply a Chi-squared test.

```
# hypothesis per each day
ggplot(xx, aes(x=Day_of_Week))+geom_bar() +
  theme(axis.text.x = element_text(angle = 45, hjust=1))
```



Next, we are required to do the same test using only week-days (excluding Saturday and Sunday).

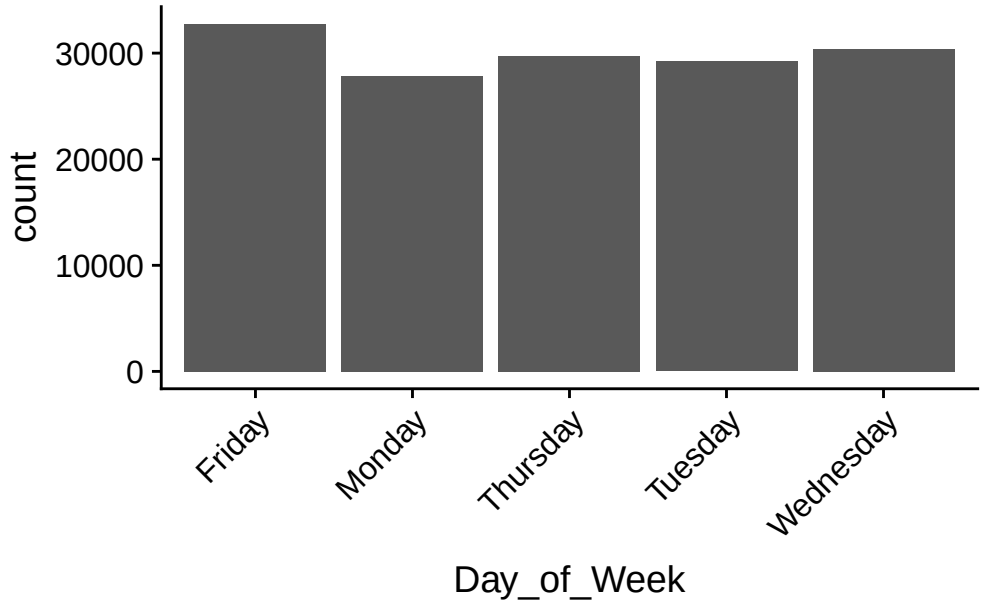
```
# Select data using only week-days
week_days <- xx%>%
  select(Day_of_Week)%>%
  filter(!Day_of_Week %in% c("Saturday", "Sunday"))
```

```
## Warning: package 'bindrcpp' was built under R version 3.4.4
```

```
table(week_days)
```

```
## week_days
##    Friday    Monday  Thursday   Tuesday Wednesday
##     32738     27812     29738     29219     30373
```

```
ggplot(week_days, aes(x=Day_of_Week))+geom_bar() +
  theme(axis.text.x = element_text(angle = 45, hjust=1))
```




```
# hypothesis per each day using only week-days
```

2.3 Question 3

Finally, in question 3, we are asked to compute a 95% confidence interval for the expected (mean) number of accidents which occur on a Monday.

```
# Select data using only week-days
week_days<- xx%>%
  select(Day_of_Week)%>%
  filter(Day_of_Week == "Monday")
```

3 Bibliography

The resources used to carry out this project are:

- Balka, J. 2013. JBStatistics: Making Statistics Make Sense. Available from: http://www.jbstatistics.com.
- Lane, D.M. 2018. Online Statistics Education: An Interactive Multimedia Course of Study. Available from: http://onlinestatbook.com/.
- Taylor, C. 2017. MATH5741M: Statistical Theory and Methods. Outline Lecture Notes.
- Yau, C. 2018. R tutorial: an R introduction to statistics. Available from: http://www.r-tutor.com