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 three questions through inferential statistical analysis regarding a road traffic accidents dataset from the UK Department for Transport.

All the analysis has been done using **R** (programming language) and is code reproducible. To see the complete **R** coding process and outputs 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.

http://www.jbstatistics.com/pooled-variance-t-tests-and-confidence-intervals-an-example/

To plot the boxplot, we first remove the unallocated values. Next, we visualise the data to analyse if it is normal (histogram 1). Because it is very skewed to the right, we take the log10 and see that it substantially improves (histogram 2).

```
# show table of values
table(xx_area$Number_of_Vehicles)
```

```
##
##      1      2      3      4      5      6      7      8      9     10
##  59770 117114  17002   3516    819    262     83     40     18      8
##     11     12     13     15     17     18     20
##      4      2      1      1      1      2      2
```

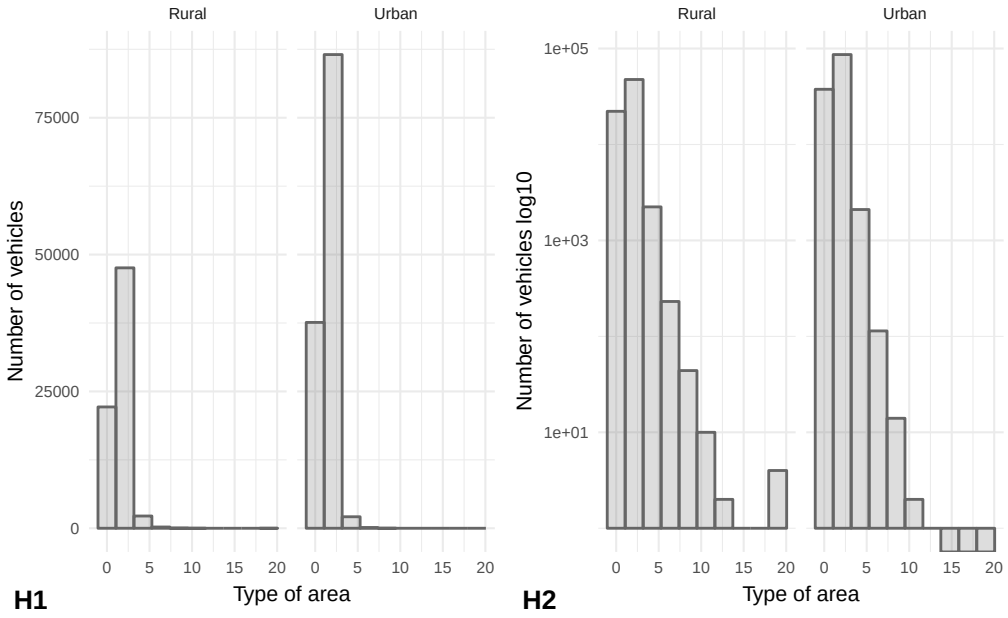


Figure 1: Histograms

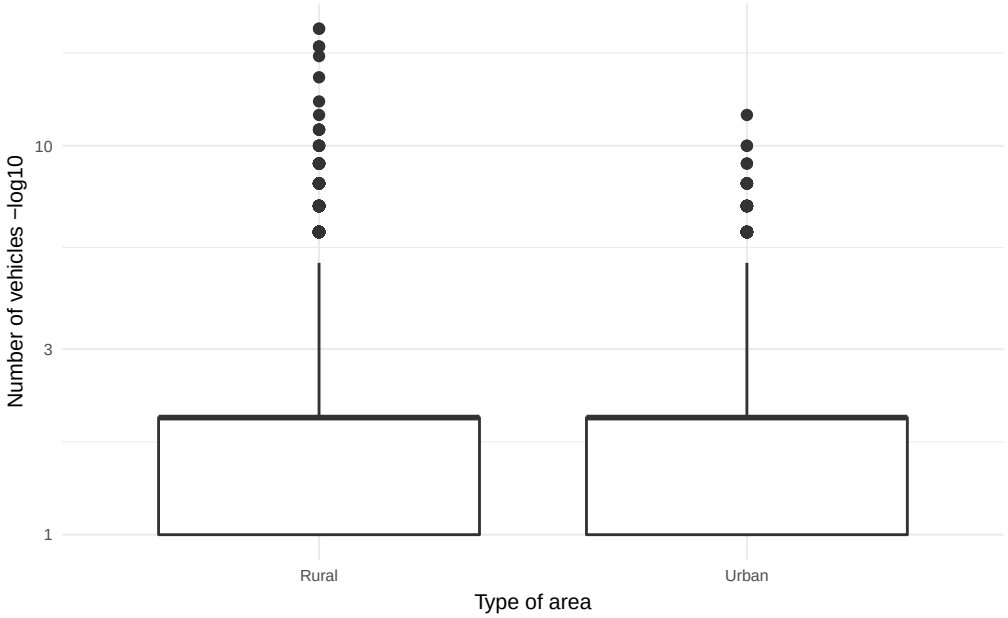


Figure 2: Number of vehicles involved grouped by type of area

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**. We assume that $\sigma_1^2 = \sigma_2^2 = \sigma^2$

http://www.jbstatistics.com/inference-for-two-means-introduction/

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

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

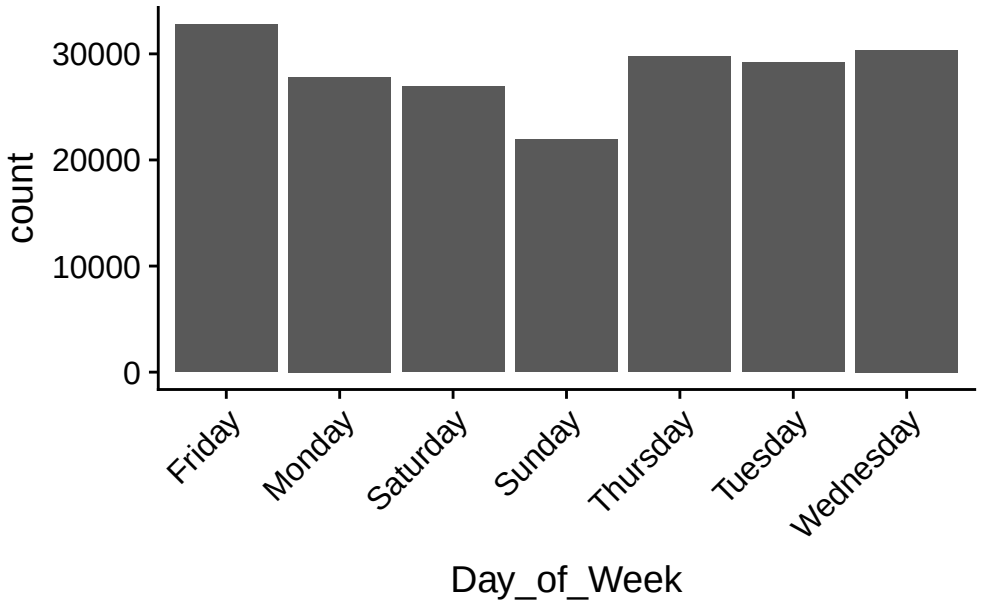
2.2 Question 2

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**.

http://www.jbstatistics.com/chi-square-tests-for-one-way-tables/

```
# 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).

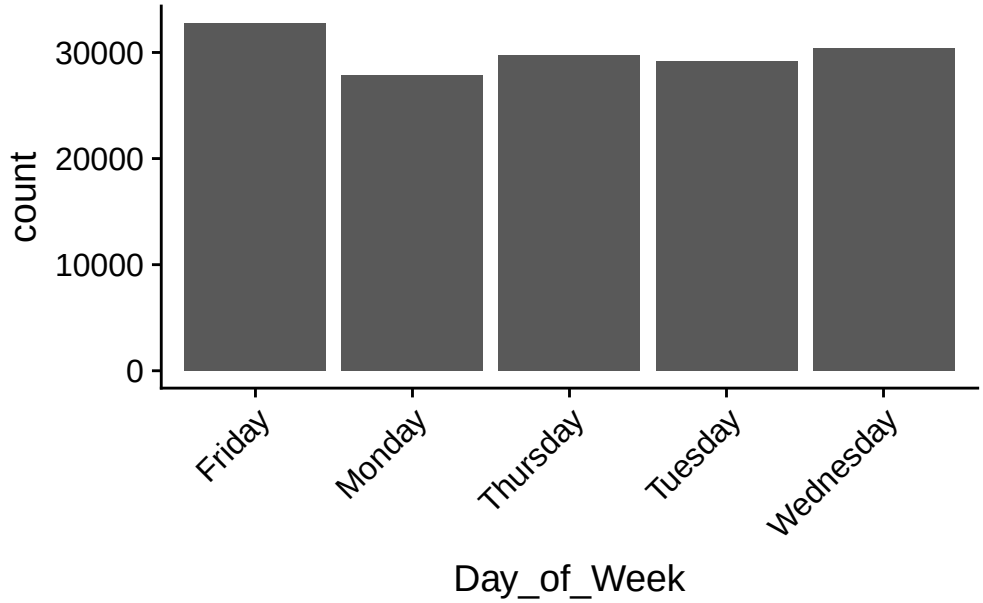
```
# Select data using only week-days
week_days <- xx%>%
  select(Day_of_Week)%>%
  filter(!Day_of_Week %in% c("Saturday", "Sunday"))
```

```
## 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
```

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




```r
# 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.

Look at example in p.62 to answer this:

- The data is clearly not normal. The distribution is discrete and very skewed to the right. The table of values is:
- The sample is huge (n= 198,735) so the central limit theorem stats that the sample mean in normallly distributed (even when the data are not).
- However, we need to think - for which population are we estimating a mean for?.

Is this a simple t-distribution confidence interval? WE could say that because the sample is that bing could be just a normal confidence interval?

I need to create calculate the mean of accidents on Mondays, so per each Monday which is the number of accidents expected. How to do that?

```r
# Select data using only week-days
Mondays<- xx%>%
  select(count, Day_of_Week)%>%
  group_by(Day_of_Week)%>%
  summarise(sum = sum(count))
```

```r
#t.test(Mondays, conf.level = 0.95)$conf.int
```

3 Bibliography

The resources used to carry out this project were:

- 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