Data Analysis and Statistical Inference 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 from 2005 collected by 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/Data_Analysis_and_Statistical_Inference_using_R.

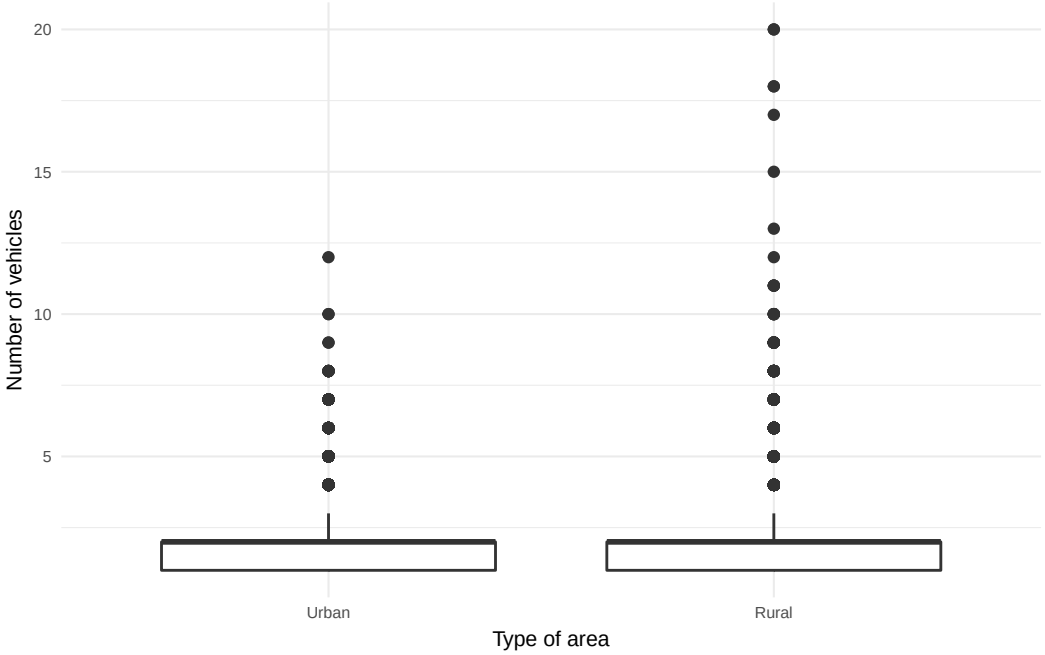
2 Results

2.1 Question 1

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

For this, we first prepare the data removing "Unallocated" values from the `Urban_or_Rural Area` variable. Then, we plot the graph.

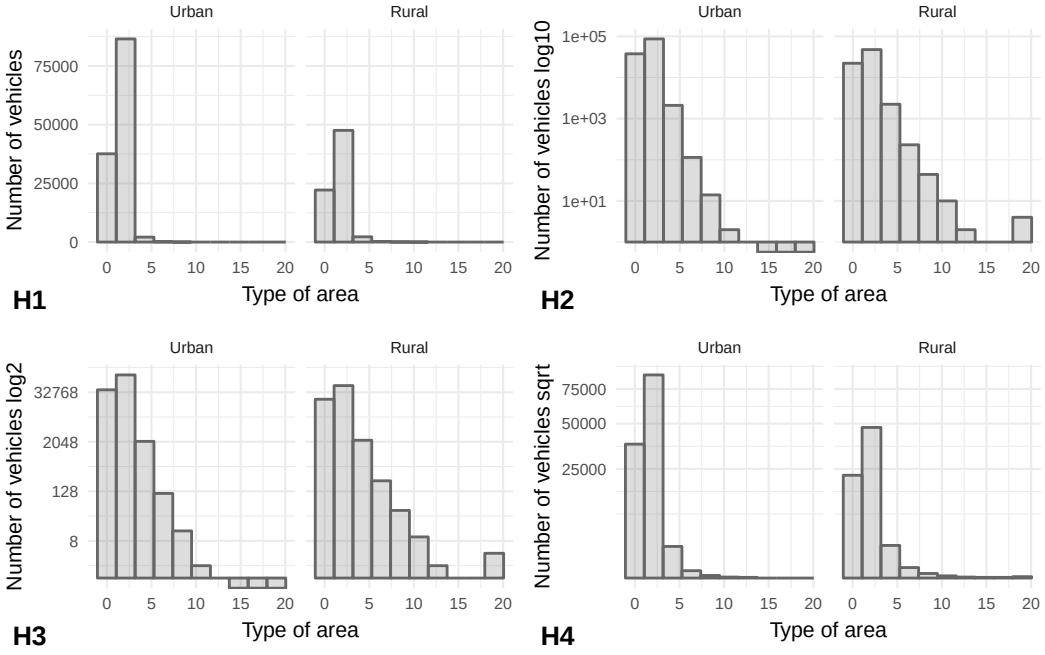
Figure 1: Number of vehicles involved in accidents grouped by type of area



Apart from the fact that rural areas have more outliers than urban areas, in the boxplot we cannot appreciate the differences between their quantiles. Both boxes look identical and the median and upper quartile seem to be coincident.

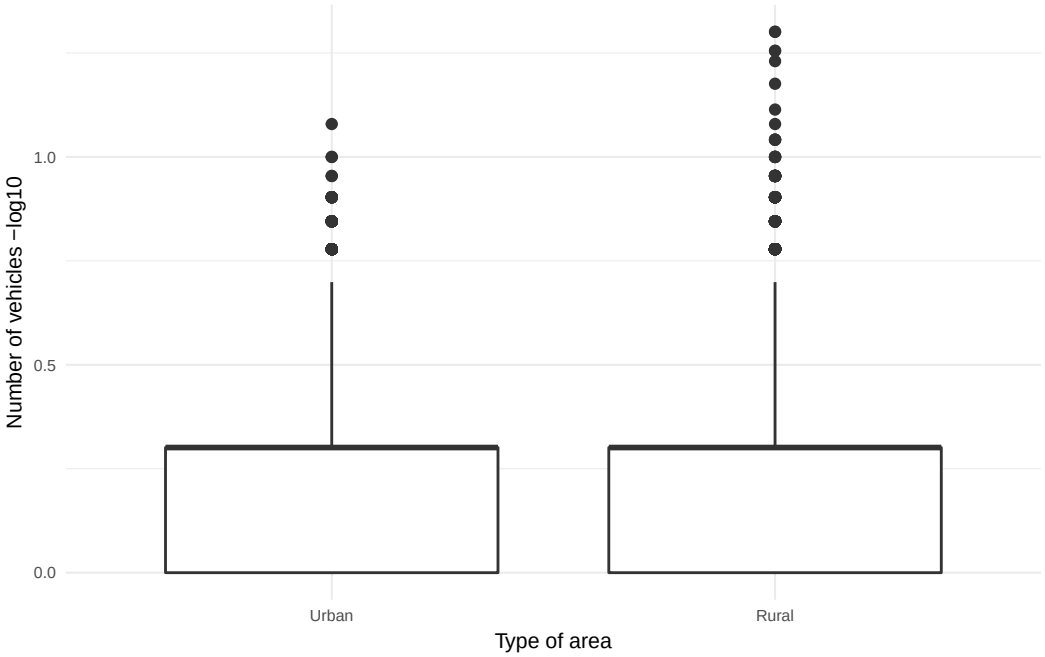
This is because the data is not symmetrical. As we can see in histogram H1 (Figure 2), the data is very skewed to the right. To normalise it, we transform the `Number_of_Vehicles` variable in three different ways: taking the log10, log2 and using the square root (see histograms, H2, H3 and H4 in Figure 2). In these new histograms, the distribution is not entirely symmetric, but they have improved, particularly those that take log10 and log2.

Figure 2: Data histogram (H1) and transformed data histograms (H2, H3, H4)



We choose the log10 transformation, which looks closer to normal distribution, and draw a second boxpot.

Figure 3: Number of vehicles (log10) involved in accidents grouped by type of area



This time the appearance of the boxplot is better, with bigger boxes. However, the interpretation is still hard, and we cannot be 100% sure whether the average number of vehicles involved in accidents differs per type of area.

To investigate this, we carry out a statistical test, which is the second requirement of the question. The null hypothesis is that the mean of vehicles involved in both types of areas is equal. The alternative hypothesis is that they differ. Denoting the rural areas by subscript r and urban by subscript u, we have:

$$H_0 : \mu_u = \mu_r \quad vs. \quad H_1 : \mu_u \neq \mu_r$$

We will use a critical region approach with $\alpha = 0.01$.

The summary statistics are:

$$n_u = 126378 \;\; \bar{x_u} = 0.2305898 \;\; s_u^2 = 0.1622405 \;\; n_r = 72267 \;\; \bar{x_r} = 0.2389048 \;\; s_r^2 = 0.1775904$$

It seems reasonable to assume $\sigma_u^2 = \sigma_r^2$. Consequently, we apply a **pooled standard deviation** for our estimates[1].

$$s_p^2 = \frac{(n_u - 1)s_u^2 + (n_r - 1)s_r^2}{n_u + n_r - 2} = \frac{126377 * 0.02632197984 + 72266 * 0.03153835017}{126378 + 72267 - 2} = 0.0282197$$

The test statistic is then:

$$\frac{\bar{x}_u - \bar{x}_r - 0}{s_p\sqrt{\frac{1}{n_u} + \frac{1}{n_r}}} = \frac{0.2389048 - 0.2305898}{0.0001316089} = -0.001374182$$

We compare this to the critical point of t-distribution with ν = 198,643 degrees of freedom[2], which is $t_{198643}(0.005)$=2.575854. Since $|0.001374182| < 2.575854$, we do not reject the null hypothesis and conclude that the mean of vehicles involved in both types of areas is equal ($\mu_u = \mu_r$).

2.2 Question 2

In this question, we have to investigate whether the frequency of accidents varies by day of the week using a suitable statistical hypothesis test. **Chi-square test** can be used to test whether observed data differ significantly from theoretical expectations (Lane 2018). So, this is the test we apply.

The null hypothesis is that the frequency of accidents is evenly distributed per days of the week (i.e. the probability of accidents occurring per each day is 1/7). The alternative hypothesis is that their frequency differs (i.e. the probability of accidents occurring per each day is not 1/7).

$$H_0 : p = 1/7 \quad vs. \quad H_1 : p \neq 1/7$$

[1] We can assume equal variances when the ratio of max/min is less than 3 or less than 4 for small samples (Taylor 2017, 69).
[2] With this number of degrees of freedom we could have also apply z-statistic and the result would have been nearly the same.

To carry out this test, first, we prepare the data, aggregating it by `Day_of_Week`. Secondly, we create a table with the observed values, the expected values and other necessary contributions for the test per day of week.

Table 1: Observed, expected and contributions to X^2. Mon-Sun

week days	observed	expected	oi - ei	(oi - ei)^2/ei
Monday	27812	28390.71	578.7143	11.79647
Tuesday	29219	28390.71	-828.2857	24.16485
Wednesday	30373	28390.71	-1982.2857	138.40640
Thursday	29738	28390.71	-1347.2857	63.93565
Friday	32738	28390.71	-4347.2857	665.67163
Saturday	26945	28390.71	1445.7143	73.61878
Sunday	21910	28390.71	6480.7143	1479.34487

The value of $\chi^2 = 2456.93865$. This can be compared to the χ^2 distribution with 7 - 1 = 6 degrees of freedom, giving a p-value of 2.2e-16. This p-value represents the probability that we are wrong in the assumption they are basically not equally distributed. So, we reject the null hypothesis and affirm that the frequency of accidents is not evenly distributed per days of the week ($p \neq 1/7$).

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

This time the null hypothesis is that the frequency of accidents is equally distributed per week days (i.e. the probability of accidents per each week day is 1/5). The alternative hypothesis is that their frequency differs (i.e. the probability of accidents per each week day is not 1/5).

$$H_0 : p = 1/5 \quad vs. \quad H_1 : p \neq 1/5$$

First, we prepare the data, aggregating it by `Day_of_Week` and removing Saturday and Sundays. Then, we create a new table with the summaries from Monday to Friday.

Table 2: Observed, expected and contributions to X^2. Mon-Fri

week days	observed	expected	oi - ei	(oi - ei)^2/ei
Monday	27812	29976	2164	156.221510
Tuesday	29219	29976	757	19.116927
Wednesday	30373	29976	-397	5.257840
Thursday	29738	29976	238	1.889645
Friday	32738	29976	-2762	254.491727

The value of $\chi^2 = 436.9776$. This is compared to the χ^2 distribution with 5-1=4 degree of freedom, giving a p-value again of 2.2e-16. So, again we reject the null hypothesis and state that the frequency of accidents in week days is not equally distributed ($p \neq 1/5$).

2.3 Question 3

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

To prepare the data, we filter the accidents occurred on Mondays and group them by date.

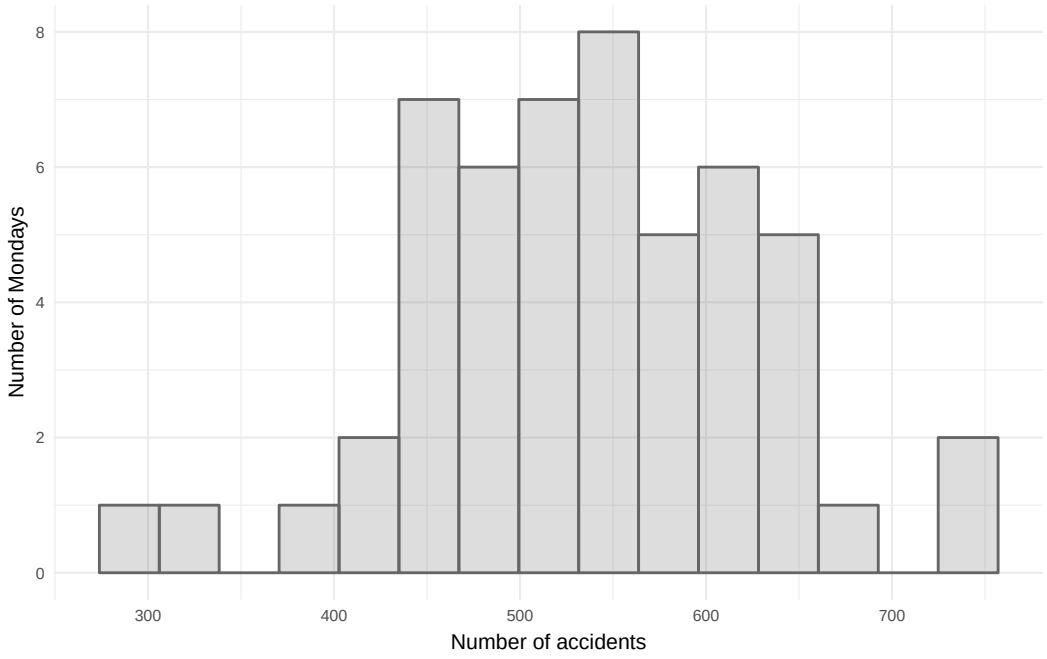
In total we get 52 observations ($n = 52$). The sample mean and variance are: $\bar{x}$= 534.8462 and s^2= 92.98627 respectively. Since we desire a 95% interval, our α= 0.05. We then find that $t_{51}(0.025)$= 2.007584.

Substituting all these quantities into the form of the confidence interval, we have the 95% confidence interval for the expected number of accidents on a Monday.

$$\left(\bar{x} - t_{n-1}(\alpha/2)\frac{s}{\sqrt{n}}, \bar{x} + t_{n-1}(\alpha/2)\frac{s}{\sqrt{n}}\right) = (534.8462 - 25.88754,\ 534.8462 + 25.88754) = 508.9586, 560.7337$$

Computing this interval, we state the assumption that the data are normally distributed. An informal approach to check that this assumption is reasonable, is to compare a histogram (or another kind of graph) of the sample data to a normal probability curve, as we did in question 1.

Figure 4: Histogram number of accidents which occur on a Monday



The histogram does not show perfect symmetry, but its shape is close to normal distribution.

However, to be more certain, there are various formal hypothesis tests to check normality that can be used. The one that we will perform here is the **Shapiro-Wilk test**, which takes account of the expected values, but also the correlations between the order statistics (Taylor 2017, 85).

These are the hypothesis:

$$H_0 : \textit{data come from a normal distribution} \quad \textit{vs.} \quad H_1 : \textit{data do not come from a normal distribution}$$

We perform Shapiro-Wilk test of normality using the command `shapiro.test(x)` in **R**.

The results are W = 0.98537 and p-value = 0.7681.

The p-values gives evidence against the null hypothesis. Since the p-value = 0.7681 is large (i.e. greater than 0.05), we accept that the data come from a normally distributed population.

2.3.1 Appendix

```
knitr::opts_chunk$set(fig.pos = 'H')
# Activate libraries
library(tidyverse)
library(cowplot)
library(knitr)
# Read csv in R
xx=read.csv("DfTaccidents.csv", header=T)
# Have a look at the variables we got
names(xx)
# Drop all those we do not need
xx <- xx[, c(8, 10, 11, 30)]
# Create a variable count with value 1 (for future aggregates)
xx$count <- 1
# Rename labels of Urban_or_Rural_Area
xx$Urban_or_Rural_Area[xx$Urban_or_Rural_Area == "1"] <- "Urban"
xx$Urban_or_Rural_Area[xx$Urban_or_Rural_Area == "2"] <- "Rural"
xx$Urban_or_Rural_Area[xx$Urban_or_Rural_Area == "3"] <- "Unallocated"

# Rename labels of day of the week
xx$Day_of_Week[xx$Day_of_Week == "1"] <- "Sunday"
xx$Day_of_Week[xx$Day_of_Week == "2"] <- "Monday"
xx$Day_of_Week[xx$Day_of_Week == "3"] <- "Tuesday"
xx$Day_of_Week[xx$Day_of_Week == "4"] <- "Wednesday"
xx$Day_of_Week[xx$Day_of_Week == "5"] <- "Thursday"
xx$Day_of_Week[xx$Day_of_Week == "6"] <- "Friday"
xx$Day_of_Week[xx$Day_of_Week == "7"] <- "Saturday"

# Give order the weekdays names
xx$Day_of_Week<- factor(xx$Day_of_Week, levels=c("Monday", "Tuesday", "Wednesday", "Thursday", "Friday"
xx<-xx[order(xx$Day_of_Week),]
# Data ready for analysis - show first 6 rows
head(xx)
# Remove Unallocated recods
xx_a = xx[!xx$Urban_or_Rural_Area == "Unallocated",]
# Draw boxplot
xx_a$Urban_or_Rural_Area<-factor(xx_a$Urban_or_Rural_Area, levels=c("Urban", "Rural"))
p <- ggplot(xx_a, aes(x=Urban_or_Rural_Area, y=Number_of_Vehicles))+geom_boxplot()
p +  xlab("Type of area") + ylab("Number of vehicles") +  theme_minimal(base_size = 8)
# Create variables transformations
xx_a$log2 <- log2(xx_a$Number_of_Vehicles)
xx_a$sqrt <- sqrt(xx_a$Number_of_Vehicles)
xx_a$log10 <- log10(xx_a$Number_of_Vehicles)
# Plot histograms
p2 <- ggplot(xx_a, aes(x=Number_of_Vehicles)) +
    geom_histogram(position="identity", colour="grey40", alpha=0.2, bins = 10) +
    facet_grid(. ~ Urban_or_Rural_Area) +  xlab("Type of area") + ylab("Number of vehicles")+  theme_mi
p3<-p2 + scale_y_continuous(trans = "log10") +   xlab("Type of area") + ylab("Number of vehicles log10")
p4<-p2+ scale_y_continuous(trans = "log2") +  xlab("Type of area") + ylab("Number of vehicles log2")
p5<-p2+ scale_y_continuous(trans = "sqrt") + xlab("Type of area") + ylab("Number of vehicles sqrt")
p6<-plot_grid(p2, p3, p4, p5, labels = c("H1","H2","H3","H4"), label_size = 10, label_x = 0, label_y =
p6
```

```r
# Plot boxplot with data transformed
p7 <- ggplot(xx_a, aes(x=Urban_or_Rural_Area, y=log10))+geom_boxplot()
p7 +  xlab("Type of area") + ylab("Number of vehicles -log10") +  theme_minimal(base_size = 8)
# Prepare the data
xx_u <- xx_a%>%
  filter(Urban_or_Rural_Area=="Urban")
xx_r <- xx_a%>%
  filter(Urban_or_Rural_Area=="Rural")
# Calculate pooled variance
nu=length(xx_u$Number_of_Vehicles)
nr=length(xx_r$Number_of_Vehicles)
su= sd(xx_u$log10)
sr= sd(xx_r$log10)
sp = (((nr-1)*sr^2+(nu-1)*su^2))/((nr+nu)-2)
sp
# Compute test
xbaru= mean(xx_u$log10)
xbarr= mean(xx_r$log10)
t = (xbaru-xbarr-0)/sp*sqrt(1/nu+1/nr)
t
# Critical value
df=(nu+nr-2)
df
abs(qt(0.01/2, df))
# Prepare data
xx_group_alldays<- xx%>%
  select(Day_of_Week, count)%>%
  group_by(Day_of_Week)%>%
  summarise(observed = sum(count))
# Table
xx_group_alldays$expected<-198735*(1/7)
xx_group_alldays$o_e<-xx_group_alldays$expected-xx_group_alldays$observed
xx_group_alldays$x2<-((xx_group_alldays$o_e)^2)/xx_group_alldays$expected
colnames(xx_group_alldays) <- c("week days", "observed", "expected", "oi - ei", "(oi - ei)^2/ei")
kable(xx_group_alldays, align=rep('r'), booktabs = T, caption = "Observed, expected and contributions t
xsup2 <-(11.79647 + 24.16485 + 138.40640 + 63.93565 + 665.67163 + 73.61878  + 1479.34487)
# Compute test
chisq.test(xx_group_alldays$observed)
# Prepare data
xx_group_weekdays<- xx%>%
  select(Day_of_Week, count)%>%
  group_by(Day_of_Week)%>%
  summarise(observed = sum(count))%>%
  filter(Day_of_Week != "Saturday" & Day_of_Week != "Sunday")
# Table
xx_group_weekdays$expected<-149880*(1/5)
xx_group_weekdays$o_e<-xx_group_weekdays$expected-xx_group_weekdays$observed
xx_group_weekdays$x2<-((xx_group_weekdays$o_e)^2)/xx_group_weekdays$expected
colnames(xx_group_weekdays) <- c("week days", "observed", "expected", "oi - ei", "(oi - ei)^2/ei")
kable((xx_group_weekdays), align=rep('r'), booktabs = T, caption = "Observed, expected and contributions
xsup2 <-(156.221511 + 19.116927 + 5.257840 + 1.889645 + 254.491727)
# Compute test
chisq.test(xx_group_weekdays$observed)
```

```r
# Prepare data
xx_mon<- xx%>%
  select(Day_of_Week, Date, count)%>%
  filter(Day_of_Week=="Monday")%>%
  group_by(Date)%>%
  summarise(n = sum(count))
# Compute the 95% confidence interval
xbar= mean(xx_mon$n)
xbar
n=length(xx_mon$n)
n
conf.level<- 0.95
z<-qt((1+conf.level)/2,df=n-1)
z
se<-sd(xx_mon$n)/sqrt(n)
sd(xx_mon$n)
ci<-z*se
ci
xbar-ci
xbar+ci
# Other way of computing the interval
x<- xx_mon$n
t.test(x, conf.level = 0.95)$conf.int
# Plot histogram
p8 <- ggplot(xx_mon, aes(x=n)) +
    geom_histogram(position="identity", colour="grey40", alpha=0.2, bins = 15)+ ylab("Number of Mondays
p8
# Shapiro-Wilk normality test
shapiro.test(xx_mon$n)
```

References

Lane, David M. 2018. "Online Statistics Education: An Interactive Multimedia Course of Study." http://onlinestatbook.com/.

Taylor, Charles. 2017. "MATH5741M: Statistical Theory and Methods." School of Mathematics - University of Leeds.