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 three questions through data analysis and statistical inference 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.

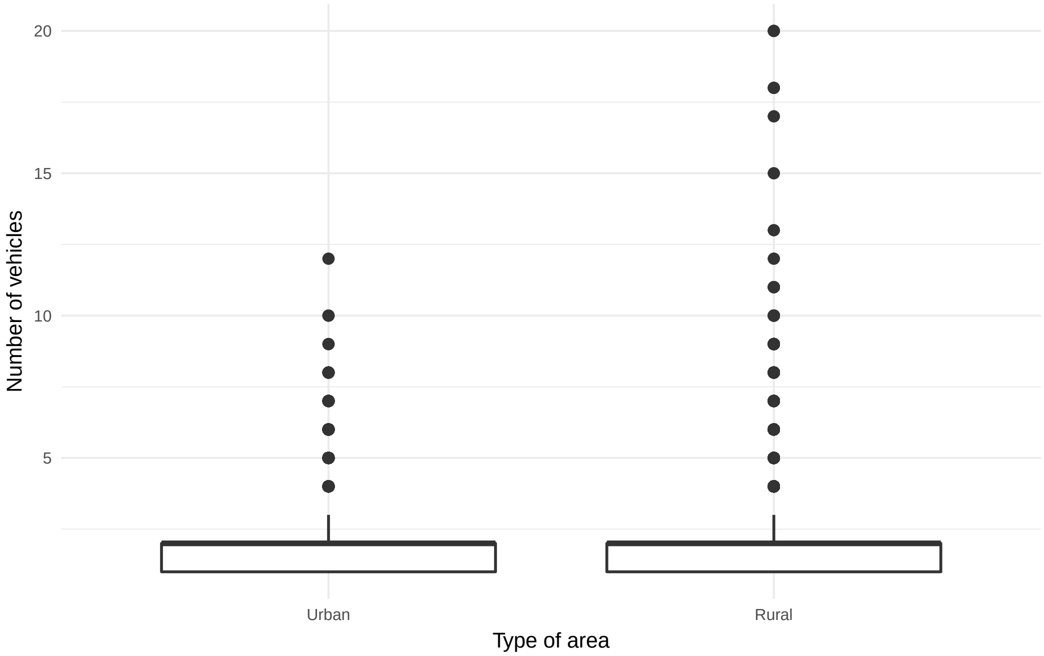
2 Results

2.1 Question 1

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

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

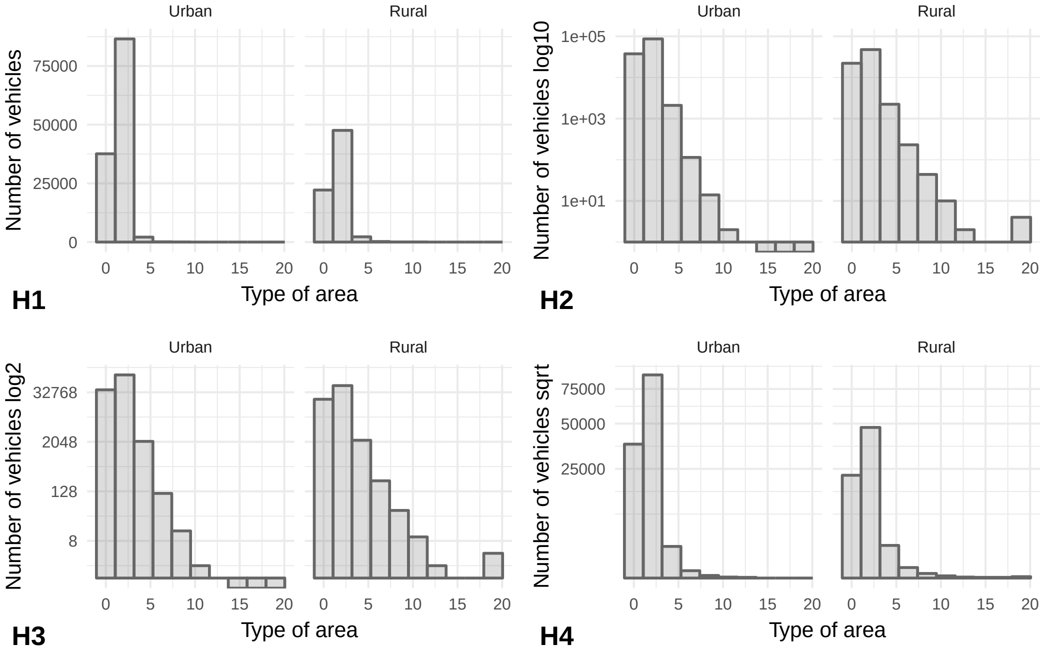
Figure 1: Number of vehicles involved in accidents grouped by type of area



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

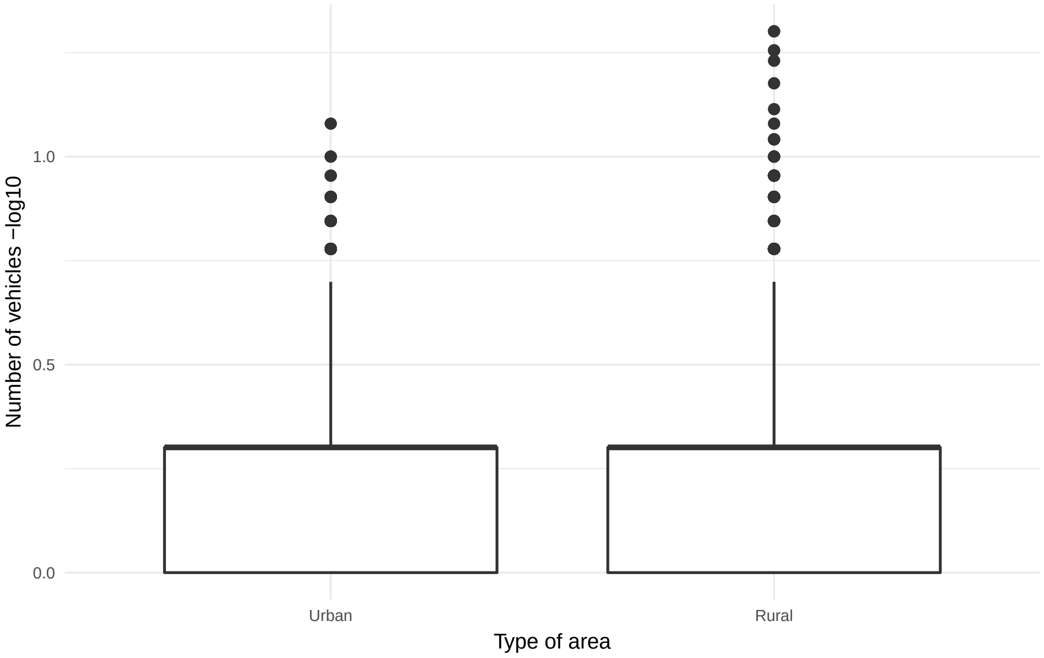
This is because the data is not symmetrical. As we can see in histogram H1 (Figure 2), the data is very skewed to the right. To normalise it and improve the interpretability or appearance of the boxplot, we transform the `Number_of_Vehicles` variable in three different ways: taking the log10, log2 and 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 and draw a second boxpot.

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



This time the size of the boxes is bigger. However, the interpretation of the graph is still difficult, 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 this question. The null hypothesis is that 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 $\sigma_u^2 = \sigma_r^2$.[1]. Consequently, we apply a **Pooled estimate**.

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

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 $\nu = 198{,}643$ 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 $\mu_u = \mu_r$.

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. For this, we apply a **Chi-square test** which can be used to test whether observed data differ significantly from theoretical expectations (Lane 2018).

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 differ (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.

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

Table 1: Observed, expected and contributions to X^2

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

The value of χ^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 can reject the null hypothesis and say that the frequency of accidents is not evenly distributed per days of the week.

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 that their frequency differ (i.e. the probability of accidents per each week day is not 1/5).

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

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

week days	Monday	Tuesday	Wednesday	Thursday	Friday
observed	27812	29219	30373	29738	32738
expected	29976	29976	29976	29976	29976
oi - ei	2164	757	-397	238	-2762
(oi - ei)^2/ei	156.221511	19.116927	5.257840	1.889645	254.491727

The value of χ^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 also unequally distributed.

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.

First, we prepare the data, filtering the accidents occurred on Mondays and grouping them by date.

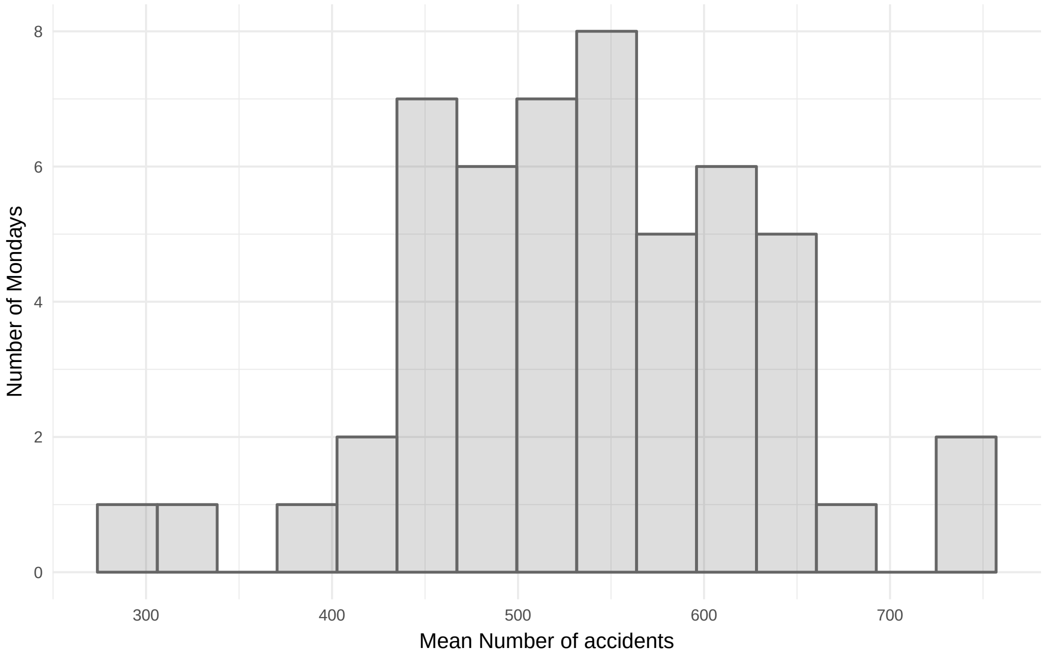
Then, we compute the sample mean and variance to be: $\bar{x}$= 534.8462 and s^2= 92.98627 respectively. Since we desire a 95% interval, our α= 0.05. n = 52. We then find that $t_{51}(0.025)$= 2.007584.

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. To check that this assumption is reasonable, we can draw a histogram.

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



The graph does not show perfect symmetry, but it is much closer to normal distribution than the data analysed in the first question.

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,

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

We can 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.

W is the the value of the Shapiro-Wilk statistic. 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.

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.