

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

<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
```


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

The summary statistics are this summary()

It seems reasonable to assume $\sigma_1^2 = \sigma_2^2$ and so we estimate the common variance with a two-sided test for two-sample problems (*pooled estimate*).

The test statistics is then:

We do not reject the null hypothesis and conclude that $\mu_{urban} = \mu_{rural}$.

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/>

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

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 states that the sample mean is normally 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 big 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?

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

```
## [1] 508.9586 560.7337
## attr(,"conf.level")
## [1] 0.95
```

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>