# An Analysis of Road Traffic Accident Data

Philip Nouris, Andy Pearce, Gunnar Peters, Tomos Power, Liwan Sun
Cardiff University Mathematics, MAT002

## INTRODUCTION

The purpose of this analysis is to investigate the factors surrounding Road Traffic Accidents (RTAs), in an attempt to determine whether there are any meaningful relationships of interest, or areas that would warrant further investigation. The data are based on personal injury RTAs on public roads during 2016 which were reported to the police and subsequently recorded using the STATS19 accident reporting form [1]. There are, however, some caveats surrounding this data set:

- It is not a full population of RTAs, for example it does not include those which are not reported to the police or do not incur personal injury. These would form a much larger set.
- Other RTAs involving personal injury will have occurred that are not reported, for example through criminal behaviour or uninsured drivers, although this would be expected to be smaller.
- The data is collated through manual form-filling, giving rise to potential for human error.
- Several variables of interest are not available (Driver experience, overall age distribution of drivers, journey frequency / duration / distance for example)
- Some of the analyses only consider data where there has been a significant accident, thus there is an inherent risk that any conclusions will be skewed due to not considering traffic density

Each section below considers a various slice of the data. Within each, assumptions are made based on existing preconceptions, which are then tested and any appropriate conclusions drawn.

## ROAD TYPE

Looking at how the road feature may impact the number of casualties and number of vehicles involved in an accident our main hypotheses are:

- The two target variables (number of vehicles and number of casualties) are correlated with each other:

Neither the number of casualties nor the number of vehicles are normally distributed which is as we would expect. We created a heat map (Fig. 1) showing a clear trend that the fewer vehicles are involved in an accident the fewer casualties there will be too. Performing a Spearman test confirms our visual conclusion with relatively high value of 0.238 and a p value less than 0.001. Thus there is enough evidence to suggest that there is a positive correlation between number of casualties and number of vehicles involved in an accident. We therefore accept our first hypothesis.

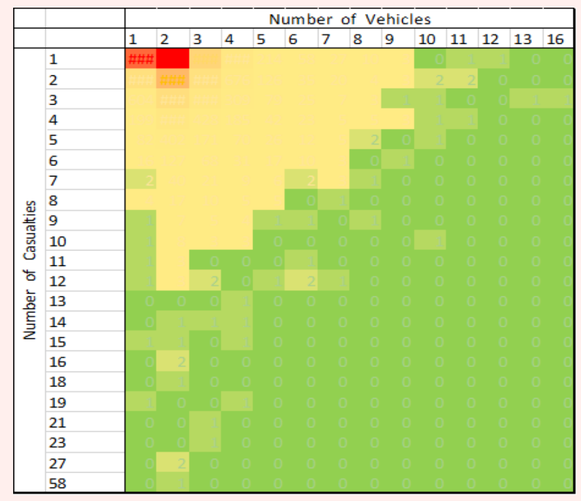


*Fig 1: Num. of vehicles/Num. casualties heat map*

- The mean number of casualties and number of vehicles will differ by speed limit:

We can see a general upward trend saying that the higher the speed, the more casualties and vehicles involved, with the exception of 60mph where number of vehicles has a dip while the number of casualties continues to go up (Table 1). This makes sense as 60mph is the default speed for rural roads thus the likelihood of involving another car is limited. This discrepancy in passengers may be due to the type of journey taken on rural roads. It is sadly impossible to gain clearer insights into this from the current data set.

Investigating if there was a correlation between speed limit and number of casualties and vehicles, a Spearman test indicates a correlation between speed limit and the target variables with values of 0.1 for number of vehicles and 0.18 for number of casualties with p values of <0.001 for both and thus we can comfortably say that they are in fact correlated.

**Report**
Mean

| Speed_limit | Number_of_Vehicles | Number_of_Casualties |
|---|---|---|
| 20 | 1.69 | 1.15 |
| 30 | 1.80 | 1.25 |
| 40 | 1.96 | 1.42 |
| 50 | 2.03 | 1.50 |
| 60 | 1.83 | 1.52 |
| 70 | 2.17 | 1.58 |
| Total | 1.85 | 1.33 |

*Table 1: Distribution of accidents*

- There will be a correlation between the target variables and the type of road (i.e. the bigger the road the more casualties and vehicles are likely to be involved)

Doing a Spearman test for road type shows us that there is a strong correlation (0.14 for number of vehicles and 0.88 for number of casualties with p<0.001) between road type and number of casualties as well as vehicles involved in an accident. This means that the larger the road, the more casualties and vehicles involved there will be.

We continue using non-parametric tests and the Kruskal-Wallis method to check if there are significant differences in the medians between the Speed Limit groups and Road types when looking at median number of casualties and median number of vehicles. Unsurprising the Kruskal-Wallis test has come back rejecting the hypothesis that the number of casualties and vehicles are evenly distributed across all speed limits and all road types with a significance of <0.001.

## AGE OF DRIVER

For this analysis, any data where the age of driver was unknown was excluded from the sample set, removing approximately 11.65% of the data. Prior to beginning, assumptions of the data were made along stereotypical lines to investigate whether there is a relationship between the age of a driver and the propensity to be involved in an RTA.


*Fig 2: Scatter plot of accidents by age*

From a scatter plot of accidents by age (Fig.2), we can see an initial peak in the number of RTAs for drivers between the ages of 20 and 30. This appears to tie in with our original assumption that younger drivers are more likely to be involved in these RTAs. There is also a clear second hump in the data between the ages of 40 and 58.

Similar charts were produced for each subset of accident severity – fatal, serious or slight – and these were seen to conform in overall shape to that of the total sample.

There are noticeable spikes in number of accidents where the age of driver is a multiple of five and between the ages of 25 and 50, highlighted for clarity in red in the chart. While this may indeed be accurate, these values happen to be coincident with typical age bands used for reporting, so suggest a level of approximation in the recorded figures. This could also lead to a degree of bias within the sample data, but this is mitigated by the large sample size.

A more specific hypothesis was considered, that "younger drivers are involved in more accidents after pub closing time". To test this hypothesis the data was categorised into "Late Evening" (22:00 to 00:29) and "Other time of day" (00:30 to 21:59). Independence was assumed for this test to be valid; it may well be that a given individual has been involved in numerous recorded accidents within the course of the year, but this was again mitigated by the very large sample size.

Using a non-parametric (due to non-normality) independent samples test with the null and alternative hypotheses: $H_0 : \eta_{Late} = \eta_{Other}$ and $H_1 : \eta_{Late} <> \eta_{Other,}$ we found that, at a significance level of $\alpha$=0.01, all tests returned a rejection of the null hypothesis, leading to the conclusion that there is a difference in the median age of a driver depending on the time of day (late evening or otherwise) of an accident.

A second non-parametric test was also carried out to consider if there was variability in the median age depending on accident severity, regardless of time of day. For this test, again at a significance level of 0.01, we retained the null hypothesis, from which we infer that there is no relationship between the age of driver and the severity of an accident.

Returning to look more closely at the second hump of the scatter plot, a hypothesis was posited that there is no correlation between whether a driver falls into this age group, and their gender. To test this, the age and gender variable (again excluding unknown values) were categorised into binary variables – Male or Female / "Between 40 and 58" or "Other Age Group" - and a $\chi^2$-test of association for categorical variables was performed. The cross-tabulated results are shown in the table below:

| | | Gender | | |
|---|---|---|---|---|
| | | Female | Male | Total |
| Age_Group | Age 40 to 58 | 22238 | 50565 | 72803 |
| | Other Age | 45216 | 103422 | 148638 |
| Total | | 67454 | 153987 | 221441 |

*Table 2: Gender/Age cross tabulation results*

These results led to a correlation $\phi$-coefficient of 0.001 and a test statistic of 0.362, well beneath the critical value (at significance level $\alpha$=0.01) of 6.635, and so we did not reject our hypothesis. Our inference from this test is that there must be factor(s) other than gender that drive the apparent increase in accidents within this age group.

Despite warranting further investigation, multiple correlation analyses were avoided to ward against the enormous inflation in the probability of false positive results.

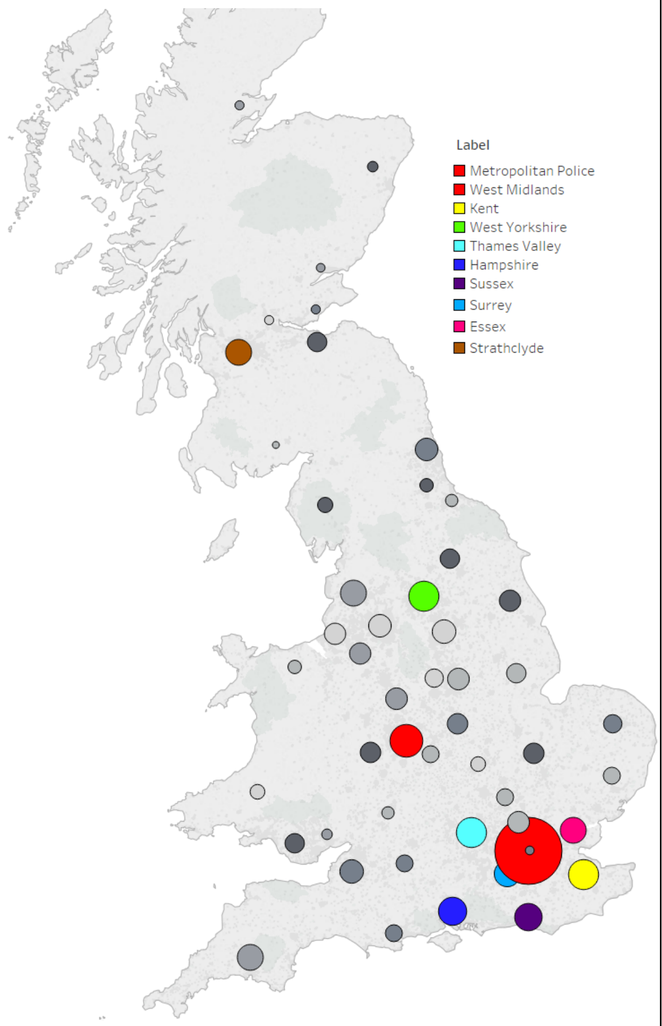


*Fig 4: Illustration of number of accidents split by police force*

## TIME AND WEATHER

The time of an accident is recorded within the data set, and visual inspection of the histogram below (Fig. 3) shows that more accidents seem to happen during peak rush hours (7:00-9:00 and 17:00-19:00). It is therefore assumed that the number of casualties is different during rush hour and other times of day.

To test if there is such a relationship between the time of day and number of casualties, two random samples of 100 accidents were selected and since the sample size is reasonably large, the Central Limit Theorem can be applied, meaning that the sample means, and difference in means, will be normally distributed. Our null hypothesis is that the means between the samples will be equal, and we calculated a 95% confidence interval for this difference. The resultant interval of (-0.2163,-0.0037) does not contain zero, therefore we can reject our null hypothesis and conclude that there is a difference in the mean number of accidents depending on time of day. This clearly aligns with our expectation as more people are on the road.

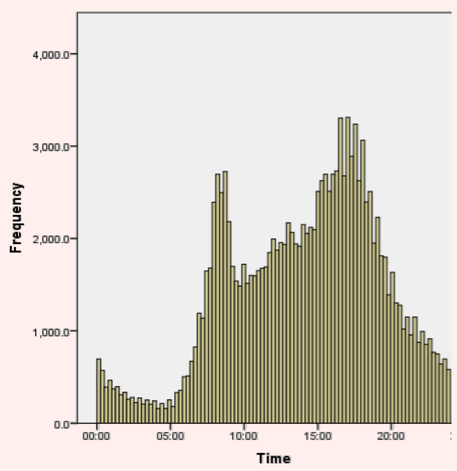


*Fig 3: Time/Num. of accidents histogram*

Weather is another factor that may affect the number of casualties in an accident. The original data set classified weather into nine types and ranked them from 1 to 9. Our assumption is that bad weather will cause more casualties in an accident, which means weather and number of casualties have a negative linear correlation. This correlation between two variables can be estimated by the sample correlation, and 100 samples were selected, calculating the Pearson product-moment correlation coefficient to get r, indicating the correlation between the two variables. This analysis shows that weather and the number of casualties do not have significant correlations because r is estimated to -0.098.

## POLICE DISTRICT

The first step of this analysis was to plot the number of accidents on a map of Britain, split by police force to visually inspect the distribution of incidents as illustrated in the diagram above (Fig. 4). As seems logical, the top 10 Police Forces by number of accidents concur with major population centres.

Looking at the number of casualties, and using the Spearman's rank-order correlation to evaluate the monotonic relationship between two continuous/ordinal variables, we discovered that there is a statistically significant, albeit weak, positive correlation (approximately 0.045) between Police Force and Number of Casualties. We then conducted a Kruskal-Wallis test (due to non-normality from the Kolmogorov-Smirnov test) with the null hypothesis: The distribution of Number of Casualties is the same across Police Forces, applying a p-value of 0.05. The result of this test is statistically significant enough to strongly reject the null hypothesis and infer that there are differences across the Police Forces (Table 3).

Returning to the apparent concurrency of accidents with major population centres, we enriched the data set to include the estimated population metrics for each Police Force [3] (This data is from mid-2010 but it is assumed for the purposes of this test that any population change since then was uniform and proportional across the data set. It is also only applicable for England and Wales), and performed a further correlation analysis between the number of accidents and population by Police Force. This produced a very strong, significant positive correlation of 0.943 between the two metrics, from which we logically infer that the population size is a key factor in the number of accidents per police force.

**Tests of Normality**

| | Kolmogorov-Smirnov[a] | | |
|---|---|---|---|
| | Statistic | df | Sig. |
| Police_Force | .127 | 136621 | .000 |
| Number_of_Casualties | .443 | 136621 | .000 |

*Table 3: Testing for normality of variables*

**Test Statistics[a,b]**

| | Number_of_Casualties |
|---|---|
| Chi-Square | 1275.560 |
| df | 50 |
| Asymp. Sig. | .000 |

*Table 4: Kruskal Wallis test with grouping variable Police Force*

## VEHICLE TYPE

Prior to analysis it was proposed that number of casualties is positively correlated with the power of vehicle involved and number of people in the vehicle. We treated number of casualties as a continuous variable in this case to enable sensible comparisons of results.

Firstly, descriptive analysis was performed revealing that the mean number of casualties in accidents involving minibuses was noticeably higher than all other vehicles (Fig. 5). This is somewhat intuitive as there are more passengers in minibuses than other vehicles. However, it is counterintuitive that minibus (8-16 passengers) accidents on average had higher casualty rates than accidents with buses + coaches (17+). A Mann Whitney U Test, with $\alpha$=0.01 and Normal approx., as both groups are large enough (>450), was used to test the significance of the difference. We rejected the null hypothesis that the means were the same. Possible explanations are that bus/coach passengers tend to be higher up than minibus passengers, bus drivers are more qualified and buses have stricter speed limit restrictions than minibuses.

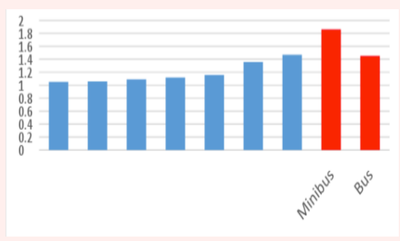


*Fig. 5: Bar chart of mean casualties by vehicle type*

| | Number of Casualties | | |
|---|---|---|---|
| | Mean | N | St. Dev. |
| <1600cc | 1.43 | 110935 | 0.873 |
| >= 1600cc | 1.49 | 82785 | 1.004 |
| Total | 1.45 | 193720 | 0.932 |

*Table 5: Num. of casualties by engine size*

We then looked at Engine Size. To test correlation between Engine Size and Number of Casualties we carried out a Spearman's rank test for correlation (after failure of the Kolmogorov Smirnov test). We found correlation at the 0.01 level, however the magnitude of the correlation is low. To delve deeper into the magnitude of the difference, we found where the approximate median of the data was (1600), split the data and applied a Mann Whitney U test at the 0.1 level. The result were significant but it can be seen that the mean of the two groups differs by a very small amount (Table 5).

In conclusion we can infer that minibuses have higher casualty rates than all other transport, including buses, and that the size of the engine of the vehicle involved has a small, but measurable, effect on the casualty rate.

## CONCLUSION

Although most of our initial assumptions were confirmed - and any anomalies explained by logical arguments - whilst analysing this data it was clear that there are no simple factors determining relationships in road safety, with many confounding elements that should be considered. Indeed, the Department for Transport Road Use Statistics 2016 [2] considers some of these elements and provides areas for further investigation, such as:

- People in higher household income groups travel more than people in the lowest
- The average distance travelled (in miles per person) is greatest in the 40-49 age band.
- Men drive much further than women on average
- Overall distance driven is growing for women
- There has been a reduction in numbers gaining a driving licence

Further targeted analysis would be warranted to explore these areas in more detail, utilising richer and more diverse data sets.

REFERENCES:
[1] https://data.gov.uk/dataset/road-accidents-safety-data
[2] https://www.gov.uk/government/statistics/road-use-statistics-2016
[3] http://www.homeoffice.gov.uk/publications/science-research-statistics/research-statistics/crime-research/population-estimates/pfa-la-pop-house-nos-xls