MATH20811 - Practical Statistics

Coursework 2 Submission

This report aims to analyse whether the five modes of transport: pedestrian, pedal cycle, motorcycle, car or other vehicle occupants are related to the severity of road casualties happened in Greater London during 2022.

The casualties are categorised as:

- Fatal: where the accident causes the injured to die in less than 30 days
- Serious: when the injured is detained in hospital as a patient, or the injury causes death 30 or more days after the accident, e.g. concussion, fractures etc.
- Slight: an injury that is neither fatal nor serious, e.g. bruise, sprain etc.

Introduction of the Data

	Casualty Severity			
Mode of Transport	Fatal	Serious	Slight	Sum
Pedestrian	41	1194	3320	4555
Pedal Cycle	7	1020	4064	5091
Motorcycle	21	873	5257	6151
Car	25	501	8476	9002
Other Vehicle Occupants	8	271	2129	2408
Sum	102	3859	23246	27207

Table 1: Data of Casualty Severity in respect to the Mode of Transport

A suitable probability model for this data is the product of five independent Multinomial Distribution, i.e. $MN(N_1 = 4555, \mathbf{p}_1) \times MN(N_2 = 5091, \mathbf{p}_2) \times MN(N_3 = 6151, \mathbf{p}_3) \times MN(N_4 = 9002, \mathbf{p}_4) \times MN(N_5 = 2408, \mathbf{p}_5)$, where $\mathbf{p}_n = (p_{n1}, p_{n2}, p_{n3})$, n = 1, 2, 3, 4, 5 are the probabilities of each casualty severity in respect to the mode of transports.

The hypothesis interest for this report would be a Pearson's Chi-squared Test to test on

$$H_0\text{: } \mathbf{p}_1 = \mathbf{p}_2 = \mathbf{p}_3 = \mathbf{p}_4 = \mathbf{p}_5 \text{ vs } H_1\text{: } \mathbf{p}_1 \neq \mathbf{p}_2 \neq \mathbf{p}_3 \neq \mathbf{p}_4 \neq \mathbf{p}_5$$

The R code below was used to construct the data above:

```
#creating the data and its proportion table
> road <- matrix(c(41, 1194, 3320, 7, 1020, 4064, 21, 873, 5257, 25, 501, 8476, 8,
271, 2129), nrow = 5, ncol = 3, byrow = T)

> dimnames(road) <- list(c("Pedestrian", "Pedal Cycle", "Motorcycle", "Car",
"Other Vehicle Occupants"), c("Fatal", "Serious", "Slight"))
> names(dimnames(road)) <- c("Mode of Transport", "Casualty Severity")

#adding outer dimnames into the table (xtable strips them out by default)
> addtorow <- list()
> addtorow$pos <- list(0, 0)
> addtorow$command <- c("& \\multicolumn{4}{c}{Casualty Severity} \\\\\n",
                        "Mode of Transport & Fatal & Serious & Slight & Sum \\\\\n")

> library(xtable)
> print(xtable(addmargins(road), digits = c(0, 0, 0, 0, 0)), add.to.row = addtorow,
include.colnames = FALSE)
```

Proportions of the Data

The data is converted into a proportion table using R to assist the analysis of the relation between the mode of transport its respective severity of injury.

	Casualty Severity		
Mode of Transport	Fatal	Serious	Slight
Pedestrian	0.00900	0.26213	0.72887
Pedal Cycle	0.00137	0.20035	0.79827
Motorcycle	0.00341	0.14193	0.85466
Car	0.00278	0.05565	0.94157
Other Vehicle Occupants	0.00332	0.11254	0.88414

Table 2: Proportions of the Casualty Severity with the Mode of Transports

The values in the proportion table is the number of mode of transports in each of the severity categories divided by the total of accidents with the mode of transports involved i.e $N_1 = 4555$, $N_2 = 5091$, $N_3 = 6151$, $N_4 = 9002$ and $N_5 = 2408$.

The differences between motorcyclists and other vehicle occupants are almost similar for all three categories, they only differ by 0.00009, 0.02939 and -0.02948.

The biggest differences between these samples of the mode of transports would be pedestrians and cyclists on fatality with pedestrians proportion being approximately 224% larger than the cyclists proportion; along with pedestrians and drivers with serious and slight injuries with pedestrians proportion being 371% greater than drivers whereas drivers proportion is 29% bigger than pedestrians.

A bar plot is then constructed according to the table above for the better view of the relevance of mode of transport and casualty severity.

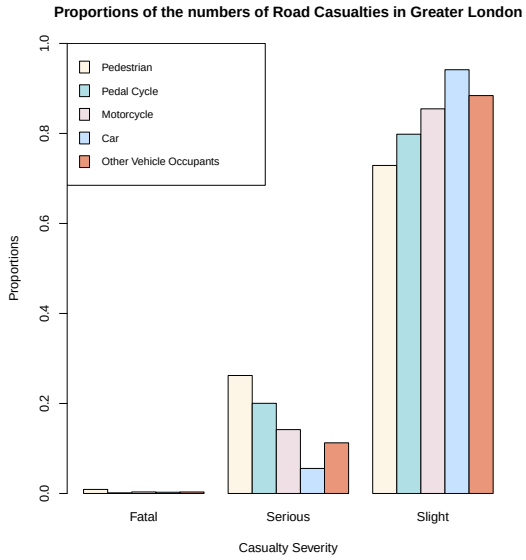

Figure 1: Bar Plot regarding Road Casualties in Greater London

This plot has confirmed the calculations of the proportions from the table above: for all three levels of severity, the proportions of motorcyclists and other vehicle occupants in the samples are almost similar, whereas pedestrians and cyclists, pedestrians and drivers, differ greatly for the mentioned severity above.

The R codes used for the table and bar plot were as follows:

```
#adding outer dimnames into the table (xtable strips them out by default)
> addtorow1 <- list()
> addtorow1$pos <- list(0, 0)
> addtorow1$command <- c("& \\multicolumn{3}{c}{Casualty Severity} \\\\\n",
                         "Mode of Transport & Fatal & Serious & Slight \\\\\n")

> print(xtable(prop.table(road, 1), digits = c(0, 5, 5, 5)), add.to.row = addtorow1,
include.colnames = FALSE )

#constructing a bar plot
> barplot(prop.table(road, 1), beside = T, ylim = c(0,1), ylab = "Proportions",
xlab = "Casualty Severity", main = "Proportions of the numbers of Road Casualties
in Greater London", col = c("oldlace", "powderblue", "lavenderblush2", "slategray1",
"darksalmon"))
> legend("topleft", cex = 0.8, c("Pedestrian", "Pedal Cycle", "Motorcycle", "Car",
"Other Vehicle Occupants"), fill = c("oldlace", "powderblue", "lavenderblush2",
"slategray1", "darksalmon"))
```

Hypothesis Testing

Under the assumption that H_0 is true, the expected frequencies, $E_{j,k}$, j = 1,...,5, k = 1,...,3 are calculated by: $E_{j,k} = n_j \hat{p}_k$, where $\hat{p}_k$ is the estimated common probabilities: $\hat{p} = \frac{Y_k}{N}$ for k = 1, 2, 3 and N is the sum of the data. The expected frequencies can be computed in R and they are as shown:

	Casualty Severity		
Mode of Transport	Fatal	Serious	Slight
Pedestrian	17.08	646.07	3891.85
Pedal Cycle	19.09	722.10	4349.81
Motorcycle	23.06	872.45	5255.49
Car	33.75	1276.83	7691.42
Other Vehicle Occupants	9.03	341.55	2057.43

Table 3: Expected Frequencies of the Data

For casualties that led to fatality, we observed more pedestrians, but less cyclists, drivers and other vehicle occupants than expected. Similarly, the observed values for serious injuries had more pedestrians and cyclists than expected while more drivers and other vehicle occupants were expected. However, there were more observed drivers and other vehicle occupants that suffered from slight injuries but less pedestrians and cyclists. For motorcycles that were caught in casualties, on the other hand, have similar observed and expected values for all three levels of casualty severity.

We will now compute the Pearson's Chi-squared test with a significance level $\alpha = 0.05$ will now be computed using R. We want to test for:

$$H_0\text{: } \mathbf{p}_1 = \mathbf{p}_2 = \mathbf{p}_3 = \mathbf{p}_4 = \mathbf{p}_5 \text{ vs } H_1\text{: } \mathbf{p}_1 \neq \mathbf{p}_2 \neq \mathbf{p}_3 \neq \mathbf{p}_4 \neq \mathbf{p}_5$$

using the *chisq.test* function in R and the results is as shown:

	X-squared	df	p-value
Data	1302.63	8	6.339507e-276

Table 4: Pearson's Chi-squared Test with significant level $\alpha = 0.05$

Since the p-value is less than α i.e. 6.339507e-276 < 0.05, and the value of X-squared greater than the critical value 1302.6 $>$ 15.50731 (computed from a χ^2 distribution with (5-1)(3-1) = 8 degrees of freedom), there is sufficient evidence for H_0 to be rejected i.e. we conclude that $\mathbf{p}_1 \neq \mathbf{p}_2 \neq \mathbf{p}_3 \neq \mathbf{p}_4 \neq \mathbf{p}_5$.

The test results were generated with the R codes:

```
#solving for expected frequencies
> test <- chisq.test(road)
> test$expected

> xtable(as.table(test$expected))

#hypothesis testing using the chi-squared test
> test
> qchisq(0.95, df = 8) ##critical value

> testdf <- data.frame(
   X2 = test$statistic,
   df = test$parameter,
   pval = format(test$p.value, scientific = TRUE) ##because of a very small p-value
 )

> colnames(testdf) <- c("X-squared", "df", "p-value")
> rownames(testdf) <- c("Data")

> xtable(testdf)
```

Residuals

The residuals, squared residuals and standardised residuals are also calculated to have an idea on how well the expected values fits the observed data.

- Residuals: constructed to analyse the differences between observed and expected values, can be calculated in R or manually by the formula : $r_{ij} = \frac{O_{ij}-E_{ij}}{\sqrt{E_{ij}}}$

	Fatal	Serious	Slight
Pedestrian	5.79	21.56	-9.17
Pedal Cycle	-2.77	11.09	-4.33
Motorcycle	-0.43	0.02	0.02
Car	-1.51	-21.71	8.95
Other Vehicle Occupants	-0.34	-3.82	1.58

Table 5: Residuals of Person's Chi-squared Test

- Squared Residuals: similar to the residual values, but the squared values provide smaller errors with the formula as such : $r_{ij}^2 = \frac{(O_{ij}-E_{ij})^2}{E_{ij}}$

	Fatal	Serious	Slight
Pedestrian	33.5142	464.6872	84.0246
Pedal Cycle	7.6536	122.8979	18.7800
Motorcycle	0.1841	0.0003	0.0004
Car	2.2680	471.4115	80.0326
Other Vehicle Occupants	0.1170	14.5716	2.4900

Table 6: Squared Residuals of Pearson's Chi-squared Test

- Standardised Residuals: we can assume that it has an asymptotic $N(0,1)$ distribution as: $rs_{jk} = \frac{O_{jk}-E_{jk}}{\sqrt{E_{jk}(1-\frac{n_j}{N})(1-\frac{Y_{+k}}{N})}}$

	Fatal	Serious	Slight
Pedestrian	6.36	25.50	-26.33
Pedal Cycle	-3.07	13.27	-12.60
Motorcycle	-0.49	0.02	0.06
Car	-1.84	-28.65	28.66
Other Vehicle Occupants	-0.36	-4.32	4.33

Table 7: Standardised Residuals of Person's Chi-squared Test

with the R code as follows:

```
> xtable(test$residuals)
> xtable(test$residuals^2, digits = c(0, 4, 4, 4))
> xtable(test$stdres)
```

From the table of residual values, it has proven the above expected frequencies to be true, i.e. positive residuals indicate larger observed value than expected vice versa. For example, it was previously said that more pedestrians were observed to result in fatality in casualties than it was expected. This was shown as a positive value of 5.79 in the residuals table. The squared residuals table shows us which data contributed most to the X^2 value, where it is equal to

$$\sum_{j=1}^{r}\sum_{k=1}^{c}\frac{(O_{ij}-E_{ij})^2}{E_{ij}}$$

with r being the number of rows and c the number of columns. We can see from the table that pedestrians and drivers that were seriously injured had a squared residual of 464.69 and 471.41 respectively. They are relatively larger than the rest of the values in the table, hence we can conclude that both samples have contributed most to the X-squared value. The largest values of standardised residuals are pedestrians, cyclists and drivers with serious and slight injuries, whereas motorcyclists with serious and slight injuries have all residuals (at 5% level) close to zero.

Simulated Data and its Goodness-of-fit

Using the information we have from before, we can generate 5000 random vectors from the MN distributions. Each of the data set will be tested using the Pearson's χ^2 test and its values will be stored in a vector. This was tested in R:

```
> B = 5000
> R = 5
> C = 3
> N = sum(road)

> sumT = 0
> for (i in 1:R){
   sumT[i] = sum(road[i,])
 }

> phat = 0
> for (j in 1:C){
   phat[j] = sum(road[,j])/N
 }

> ysim=matrix(, nrow = R, ncol = C)
> test.sim=0
> for (k in 1:B){
   for (m in 1:R){
     ysim[m,] <- rmultinom(n = 1, size = sumT[m], prob = phat)
   }
   test.sim[k] <- chisq.test(ysim, p = phat)$statistic
 }
```

The code above allows a histogram of the null empirical sampling distribution of χ^2 to be plotted as well as the asymptotic null distribution to show the goodness-of-fit of these simulated values graphically:

```
> hist(test.sim, freq = F, ylim = c(0,0.13), xlim = c(0, 30),
main = "Histogram of simulated X^2 values (B = 5000) with df = 8 pdf", xlab =
"Simulated Test", col = "mistyrose4")
> lines(density(test.sim), col = "rosybrown1")

#plotting asymptotic null distribution
> xx = seq(from = 0, to = 30, length.out = 600)
> dxx = dchisq(xx, df = 8)
> lines(xx, dxx, col = "navy")
```

and the result is as shown:

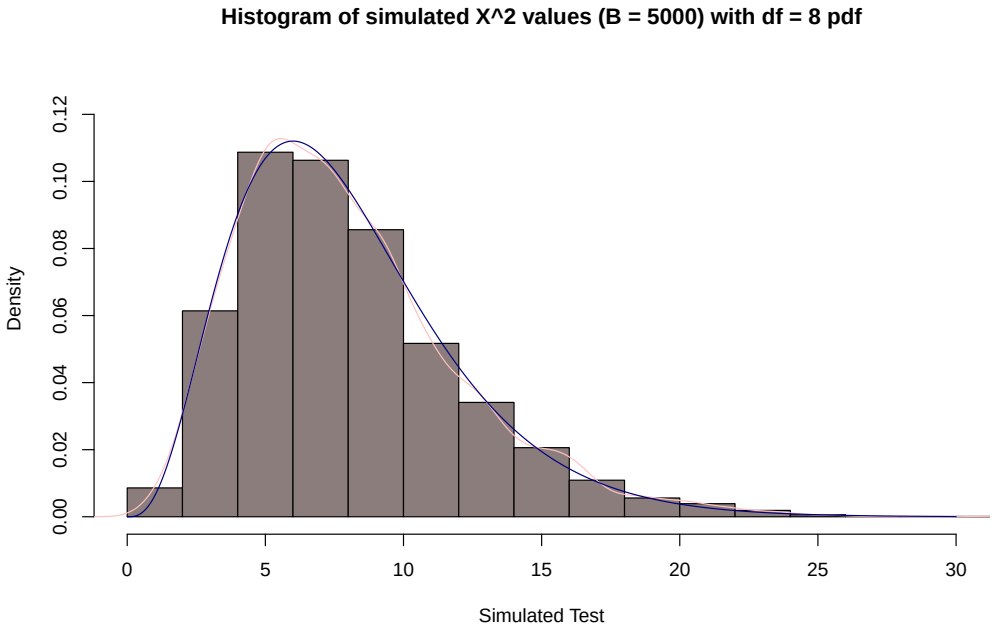

Figure 2: Histogram of the Simulated Values

It can be seen from the graph that the simulated values appears to be a good fit in comparison to the actual pdf of the χ^2 distribution with df $= 8$.

Confidence Intervals

The following R functions have been written to calculate the 95% confidence intervals of:

- Difference between the probabilities of seriously injured cyclists and car drivers, answer: [0.1327278, 0.1566708]

```
> propc_s <- prop.table(road,1)[2,2]
> propd_s <- prop.table(road,1)[4,2]

> diff.est <- abs(propc_s - propd_s)
> ese.diff <- sqrt((propc_s * (1- propc_s))/addmargins(road)[2,4] +
(propd_s * (1 - propd_s))/addmargins(road)[4,4])

> ci.lower <- diff.est - qnorm(0.975)*ese.diff
> ci.upper <- diff.est + qnorm(0.975)*ese.diff

> ci.lower;ci.upper
[1] 0.1327278
[1] 0.1566708
```

- Difference between the probabilities of serious and slight injuries of motorcyclists, answer: [0.7003347, 0.7251246]

```
> propserious <- prop.table(road, 1)[3,2]
> propslight <- prop.table(road, 1)[3,3]

> diff.est1 <- abs(propserious - propslight)
> ese.diff1 <- sqrt(((propserious * (1 - propserious))/addmargins(road)[3,4]) +
((propslight * (1 - propslight))/addmargins(road)[3,4]))

> cil <- diff.est1 - qnorm(0.975)*ese.diff1
> ciu <- diff.est1 + qnorm(0.975)*ese.diff1

> cil;ciu
[1] 0.7003347
[1] 0.7251246
```

References

[1] Reported Road Casualties Scotland 2014 Appendix D Definitions used in road accident statistics, and some other points to note.