MATH20811 - Practical Statistics

Coursework 2 Submission

This report aims to analyse the data provided of the numbers of road casualties in Greater London during 2013 and study whether the five modes of transport differ in their respective probabilities of different casualty severity. The collection of the data involves collecting fixed samples of casualties in each group of modes of transport. They are then separated into three categories of severity; ”fatal”, ”serious” or ”slight”, summarised in the contingency table, Table 1 below.

		Casualty Severity			
		Fatal	Serious	Slight	Sum
	Pedestrian	65	773	4343	5181
	Pedal Cycle	14	475	4134	4623
Mode of Transport	Powered 2 Wheeler	22	488	3992	4502
	Car	25	310	9850	10185
	Other	6	146	2556	2708
	Sum	132	2192	24875	27199

Table 1: The numbers of road casualties in Greater London during 2013

1. Based on the way that the data was collected, the counts, Y_{jk} in each jth row are independently distributed as Multinomial distributions with the following parameters;

   - Pedestrian: $n_1 = 5181$ and $\underline{p}_1$
   - Pedal Cycle: $n_2 = 4623$ and $\underline{p}_2$
   - Powered 2 Wheeler: $n_3 = 4502$ and $\underline{p}_3$
   - Car: $n_4 = 10185$ and $\underline{p}_4$
   - Other: $n_5 = 2708$ and $\underline{p}_5$

   where each $\underline{p}_j$ are vectors of probabilities of length 3 with $j = 1, 2, 3, 4, 5$.

2. We read the data into object `roadcas` in R using the `matrix` command including the following commands to label the dimensions and add an extra row and column for the sums;

```
> roadcas <- matrix(c(65, 14, 22, 25, 6, 773, 475, 488, 310,
    146, 4343, 4134, 3992, 9850, 2556), nrow = 5, byrow = F)
> dimnames(roadcas) <- list(c("pedestrian", "pedal cycle",
    "powered 2 wheeler","car", "other"), c("fatal", "serious",
    "slight"))
> names(dimnames(roadcas)) <- c("mode of transport", "casualty
severity")
```

```
> addmargins(roadcas)
                     casualty severity
mode of transport   fatal serious slight   Sum
  pedestrian           65     773   4343  5181
  pedal cycle          14     475   4134  4623
  powered 2 wheeler    22     488   3992  4502
  car                  25     310   9850 10185
  other                 6     146   2556  2708
  Sum                 132    2192  24875 27199
```

We can make some informal deductions about whether $\underset{-}{p}_j$ differs with each j by looking at the proportions, summarised in the following table and obtained by the `prop.table` command in R.

```
> prop.table(roadcas, 1)
```

		Casualty Severity		
		Fatal	Serious	Slight
Mode of Transport	Pedestrian	0.012545841	0.14919900	0.8382552
	Pedal Cycle	0.003028337	0.10274713	0.8942245
	Powered 2 Wheeler	0.004886717	0.10839627	0.8867170
	Car	0.002454590	0.03043692	0.9671085
	Other	0.002215657	0.05391433	0.9438700

Table 2: Proportions of casualty counts to its corresponding fixed sample value

Based on the shown proportions in Table 2, we can see that in each severity category, there are slight differences between the values. In the ”Fatal” category, the value that stands out the most is the proportion for pedestrians as it is very large compared to the rest. Since the other values are very small it seems like the difference between them are not very large, however, the only two proportions that are actually fairly similar are those for car and other mode of transport. Looking at the next category, the mode of transport that has the smallest proportion for the casualty severity to be ”Serious” is by Car and the largest being for Pedestrians. The difference between these two values are 0.1187621 which is too large of a proportion range for us to say that their probabilities do not differ. Lastly, in the ”Slight” category, at a glance it seems that all the proportion values are very high and may be similar to each other, but similar to the ”Serious” category, the range that the proportions are valued is quite large, between 0.9671085 and 0.8382552 with a difference of 0.1288533.

3. To better picture the comparison between values presented in Table 2, a bar chart can be plotted, using the following R code lines, and inserted in Figure 1 below it.

```
>barplot(prop.table(roadcas, 1), beside = T, legend.text = F,
    ylim = c(0, 1), ylab = "Proportions", xlab = "Casual Severity",
    main = "", col = c("green", "purple", "red", "pink", "yellow"))
>legend("topleft", c("pedestrian", "pedal cycle", "powered 2
    wheeler", "car", "other"), fill = c("green", "purple",
    "red", "pink", "yellow"))
```

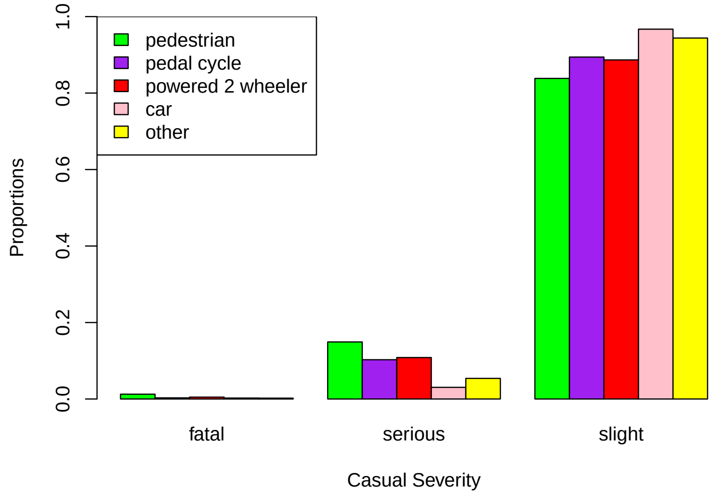

Figure 1: Proportions of casualty counts to its corresponding fixed sample value grouped by casualty severity

With a graphical figure, it is easier to comment on the differences between the proportions. In this case, we can take a look at the heights of the bars in each category. First when we look at the "Fatal" category, we can mainly note that there is one green bar that seems to be taller than the rest, which refers to the Pedestrian proportion. In the "Serious" category of casualty severity, we see that the heights of the bars are quite significantly different from each other, and the same can be said for the "Slight" category. These bars confirm the comments said in part 2. Based on the sample, it does seem to support that the five modes of transport *do* differ in their respective probabilities of different casualty severity.

4. Keeping in mind the question at hand, we can carry out a hypothesis test to help us answer it. The test at 5% significance level on the true vectors of probability of casualty severity in each mode of transport, $\underline{p}_j$, $j = 1, 2, 3, 4, 5$ is to be conducted, letting H_0 and H_A denote the null hypothesis and alternate hypothesis respectively.

$$H_0 : \underline{p}_1 = \underline{p}_2 = \underline{p}_3 = \underline{p}_4 = \underline{p}_5$$

$$\text{vs}$$

$$H_A : \text{at least two of the probabilities are not as stated under } H_0$$

The values presented in Table 1 are regarded as observed frequencies, O_{jk} however, calculating the X^2 statistic involves the expected frequencies, E_{jk}, (k being the kth column) in place of each cell of the table as well. Below are listed some formulas to calculate both values.

$$\hat{p}_k = \frac{Y_{+k}}{N} \tag{1}$$

$$E_{jk} = n_j \hat{p}_k \tag{2}$$

$$X^2 = \sum_{k=1}^{3} \sum_{j=1}^{5} \frac{(E_{jk} - O_{jk})^2}{E_{jk}} \tag{3}$$

where

- Y_{+k} are the column totals
- n_j are the fixed row totals
- $\hat{p}_k$ are the estimated probability vectors
- N is the total number of samples, 27199.

In R, it is very easy to extract these values using the few commands listed below with the expected frequencies summarised in Table 3:

```
> test <- chisq.test(roadcas); test

Pearson's Chi-squared test

data:  roadcas
X-squared = 870.25, df = 8, p-value < 2.2e-16
> names(test) #shows the values available in chisq.test result#
[1] "statistic" "parameter" "p.value"   "method"    "data.name"
    "observed"  "expected"  "residuals"
[9] "stdres"
> test$expected
```

		Casualty Severity		
		Fatal	Serious	Slight
	Pedestrian	25.14401	417.5430	4738.313
	Pedal Cycle	22.43597	372.5731	4227.991
Mode of Transport	Powered 2 Wheeler	21.84874	362.8216	4117.330
	Car	49.42902	820.8214	9314.750
	Other	13.14225	218.2410	2476.617

Table 3: Expected frequencies calculated under the assumption that H_0 is true

With these values, we are able calculate the X^2 statistic manually. However, with the `chisq.test` command, we immediately are provided with it. To conduct the test, it is important to note that the statistic follows the following distribution:

$$X^2 \sim \chi^2(8)$$

For a 5% significance level test, we will be comparing the X^2 statistic with the value of the 95th quantile in the stated distribution.

```
> qchisq(0.95, df = 8)
[1] 15.50731
```

From there, it is very clear that the statistic, $X^2 = 870.25 > 15.50731$. Therefore, there is sufficient evidence for us to reject the null hypothesis in favour of the alternate one. To answer the question of interest, the three probabilities in the five Multinomial distributions are *not* identical.

5. In light of the conclusions made previously, we should take a look at the Pearson residuals, r_{jk} for a more thorough explanation of why the conclusion was made. The formula used to calculate the residuals for each jth row and kth column are as follows;

$$r_{jk} = \frac{O_{jk} - E_{jk}}{\sqrt{E_{jk}}}$$

The main use of this set of residuals is to compare between observed and expected frequencies and look at how far off from each other they are. Using the command listed below, we get an output summarised in Table 4.

```
> test$residuals
```

		Casualty Severity		
		Fatal	Serious	Slight
Mode of Transport	Pedestrian	7.94833712	17.395482	-5.742873
	Pedal Cycle	-1.78099492	5.306501	-1.445503
	Powered 2 Wheeler	0.03235922	6.571779	-1.953197
	Car	-3.47468218	-17.829728	5.545892
	Other	-1.97015362	-4.890074	1.595142

Table 4: Residuals calculated under the assumption that H_0 is true

The positive or negative signs for each value shows that the observed frequency, O_{jk} is greater or lesser than expected under H_0. The value that stands out the most among all of the residuals is the one for a powered 2 wheeler rider being fatally injured due to it being the smallest, 0.03, which means that the observed frequency was not far off from the expected frequency calculated. The other differences range from 1.44 up to 17.83 with the largest being that the observed counts for a car driver to be seriously injured is much lesser than expected, by approximately 18. The observed count exceeds the expected the greatest when a pedestrian was seriously injured, by about 17 counts.

Another set of residuals that are useful to further dissect the X^2 statistic value would be the squared residuals. They provide us an insight to how much each cell in the contingency table contributes to its value. We include the sums of rows and columns to show they add up to the test statistic.

```
> addmargins(test$residuals^2)
```

This command gives us the output summarised in Table 5 below. As we can see, the sum at the most bottom right of the table matches with the X^2 statistic obtained by the `chisq.test` command in part 4. Now we look at the values in each cell to see which ones have contributed to the very large value of test statistic. The casualty severity category that contributed the most to it is "Serious" since it contains the two largest values of squared residuals. Some values that contribute the least by less than 3, include powered 2 wheeler drivers injured in category "Fatal", and pedal cyclists and other mode of transport in category "Slight". It is also important to note that this set of residuals range from 0 up to 317.9 which is very large. So it makes sense that the test statistic becomes much larger than its critical value in the χ^2 distribution.

		Casualty Severity			
		Fatal	Serious	Slight	Sum
Mode of Transport	Pedestrian	63.176063030	302.60280	32.980590	398.75945
	Pedal Cycle	3.171942912	28.15895	2.089478	33.42037
	Powered 2 Wheeler	0.001047119	43.18828	3.814980	47.00430
	Car	12.073416247	317.89920	30.756915	360.72954
	Other	3.881505296	23.91282	2.544477	30.33880
	Sum	82.303974604	715.76205	72.186440	870.25247

Table 5: Squared residuals contributing to the value of X^2 statistic

The final set of useful residuals is the standardised residuals, rs_{jk}, calculated using the stated formula below, essentially dividing the Pearson residuals, r_{jk} with its estimated standard deviation, making them approximately follow the standard normal distribution. With these values, we can test the following null hypothesis on r_{jk}, treating them as individual test statistics for the test.

$$rs_{jk} = \frac{O_{jk} - E_{jk}}{\sqrt{E_{jk}(1 - n_j/N)(1 - Y_{+k}/N)}}$$

$$H_0 : r_{jk} = 0 \text{ vs } H_A : r_{jk} \neq 0$$

We extract these values in R, from command `chisq.test` stored in `test` earlier in part 4, giving the output in Table 6.

```
> test$stdres
```

To test the H_0 at 5% significance that the Pearson residuals are zero, we compare each value in the table with the critical value obtained from the $N(0, 1)$ distribution,

```
> qnorm(0.975)
[1] 1.959964
```

In order to accept the null hypothesis, the standardised residuals must fall between -1.959964 and 1.959964. There are only two values that pass this condition, pedal cycle and powered 2 wheeler in the "Fatal" category of casualty severity. For the rest of the values, there is sufficient evidence to reject the H_0 that the Pearson residuals, r_{jk} are equal to zero, in favour of the alternate hypothesis.

		Casualty Severity		
		Fatal	Serious	Slight
Mode of Transport	Pedestrian	8.85564546	20.163674	-21.836104
	Pedal Cycle	-1.95962088	6.074449	-5.427880
	Powered 2 Wheeler	0.03550968	7.502756	-7.314704
	Car	-4.40397402	-23.510596	23.988498
	Other	-2.08127615	-5.374453	5.750835

Table 6: Standardised residuals calculated under the assumption that H_0 in part 4 is true

6. In part 4, it was previously mentioned that the X^2 statistic testing the homogeneity of the probability vectors in each row of our table, follows a particular distribution under the assumption that H_0 is true,

$$X^2 \sim \chi^2(8)$$

with the degree of freedom obtained from $(r-1)(c-1)$ where r is the number of rows and c is the number of columns. In this section of the report we will simulate $B = 5000$ values of test statistic from random observed frequencies obtained from a Multinomial distribution. All while keeping the row totals and $\hat{p}_k$ from the `roadcas` data inserted in R. The function is run in R as follows;

```
> R = 5 #number of rows#
> C = 3 #number of columns#
>
> ns = 0
> for (j in 1:R){
+   ns[j]<-sum(roadcas[j,])
+ }
>
> ns #row totals#
[1]  5181  4623  4502 10185  2708
>
> N = sum(roadcas)
> N
[1] 27199
>
> p.hat = 0
> for(k in 1:C){
+   p.hat[k]<- sum(roadcas[,k])/N
+ }
>
> p.hat #estimated probabilities#
[1] 0.00485312 0.08059120 0.91455568
> sum(p.hat)
[1] 1
>
> B=5000
> ysim=matrix(, nrow = R, ncol = C) #random observed frequencies#
> test.sim=0
> for (i in 1:B){
+ for (k in 1:R){
+ ysim[k,]<- rmultinom(n=1, size=ns[k], prob=p.hat)
+ }
+ test.sim[i]<- chisq.test(ysim, p=p.hat)$statistic
+ }
>
> hist(test.sim, freq=F, ylim=c(0, 0.15), main = "",
    xlab = "x")
> lines(density(test.sim), col="red")
>
> xx=seq(from=0, to=20, length.out=600)
> dxx=dchisq(xx, df=8)
> lines(xx, dxx, col="blue")
> legend('topright', c('simulated test statistic',
    'chisq pdf'), lty=1,col=c('red','blue'), bty='n',
    cex= 1.0)
```

Every time the function is run in R, different values of test statistics are produced since they depend on the random observed frequencies. So the histogram as well as its density changes each time. It is also important to note that R will give warnings to the user since the some of the test statistic values could be quite small. An example of the output is presented in Figure 2.

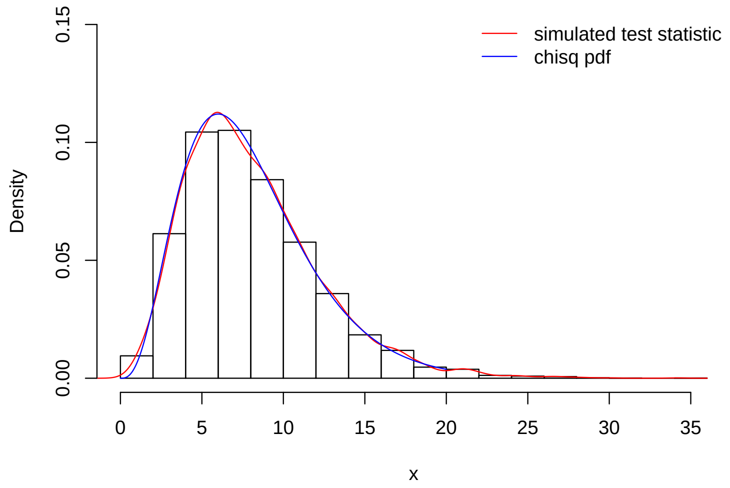

Figure 2: Histogram of simulated test statistic superimposed by its density and the density of the chi-squared distribution

It is very clear, by looking at the density of the simulated test statistic and comparing it with the χ^2 probability density function on 8 degrees of freedom that they are remarkably close.

Therefore, it is safe to say that the X^2 statistic approximately follows the aforementioned distribution under H_0.

7. In this final section of the report we will be looking into 95% confidence intervals for the differences of certain probabilities. Before we get into the calculations, we introduce a property of variance that will be useful to obtain the standard errors below;

$$Var(X - Y) = Var(X) + Var(Y) - 2Cov(X, Y)$$

where

- X and Y are random variables
- $Cov(X, Y)$ are the covariance between X and Y

In order to find these values, we make use of the distribution stated in part 1. We know that the counts, Y follow a Multinomial distribution, while the marginal distribution of any of its subset follows;

$$Y_{jk} \sim Bi(N, \hat{p}_{jk})$$

so the variance and covariance between them can be calculated as follows;

$$\begin{align} Var(Y_{jk}) &= -N\hat{p}_{jk}(1 - \hat{p}_{jk}) \tag{4} \\ Cov(Y_{jh}, Y_{jk}) &= -N\hat{p}_{jh}\hat{p}_{jk} \tag{5} \\ Cov(Y_{gh}, Y_{jk}) &= 0 \tag{6} \end{align}$$

where $g, j = 1, 2, 3, 4, 5$ and $g \neq j$, and $h, k = 1, 2, 3$. In simpler terms, when the difference of probabilities compared are in the same row, we have to note the value of their covariance when calculating their difference's variance due to their dependency.

Then, we'll need the two-sample confidence interval formula for the population proportion,

$$\bar{p} \pm z_{1-\frac{\alpha}{2}} \sqrt{Var(\bar{p})}$$

where, in this case, $\bar{p}$ will be the difference between the two $\hat{p}_{jk}$ chosen. Combining all these information, we can calculate the confidence interval for,

(a) The difference between the probability that a pedestrian is seriously injured and the probability that a car driver is seriously injured.

We expect that the confidence interval will not contain zero, as they should agree with the conclusions of the hypothesis made in part 4- that the probabilities *do* differ.

```
> dest1 <- (773/5181) - (310/10185); dest1
[1] 0.1187621
> ese.dest1 <- sqrt(((773/5181)*(1-773/5181)/5181)
    +((310/10185)*(1-310/10185)/10185))
>
```

```
> ci1a <- dest1 - qnorm(0.975)*ese.dest1
> ci2a <- dest1 + qnorm(0.975)*ese.dest1
>
> ci1a ; ci2a
[1] 0.108503
[1] 0.1290212
```

Based on the positive sign of the upper and lower bound, this suggests that the probability that a pedestrian is seriously injured is larger than the probability of a car driver to experience serious casualty. They also do not contain zero as we expected, and so the confidence interval does match with the test conducted in part 4.

(b) The difference between the probabilities of a serious injury and a slight injury for cyclists

Note that these probabilities are both in the same row, so they are dependent of each other. The calculation of the upper and lower bound of the confidence intervals are as follows;

```
> dest2 <-(475/4623) - (4134/4623); dest2
[1] -0.7914774
> cov2 <- -(475/4623)*(4134/4623) ; cov2
[1] -0.09187901
> ese.dest2 <- sqrt(((475/4623)*(1-475/4623)/4623)+
    ((4134/4623)*(1-4134/4623)/4623)-(2*cov2/4623))
>
> ci1b <- dest2 - qnorm(0.975)*ese.dest2
> ci2b <- dest2 + qnorm(0.975)*ese.dest2
>
> ci1b ; ci2b
[1] -0.8090243
[1] -0.7739305
```

The negative sign and quite the high value in the interval tells us that the probability for a cyclist to get seriously injured is a lot less than for them to get only slightly injured. Both these confidence intervals gives us a range of values that the population proportion of the difference might be in 95% of the time, which gives us very useful information on which mode of transport would be less risky to take.