MATH20811 - Practical Statistics

Coursework 2 Submission

1. Probability model:

   The data can be regarded as a single random sample of 27207 Road Casualties cross-classified by two categorical variables according to three Casualty Severity (CS) and five Mode of Transport (MT). A suitable probability model is a single Multinomial distribution given by:

   $$\mathrm{MN}(\mathrm{N} = 27207,\ \mathrm{p} = (\mathrm{p}_{11},\mathrm{p}_{12},\mathrm{p}_{13},\mathrm{p}_{21},...,\mathrm{p}_{53})$$

   where $\mathrm{p}_{jk} = \mathrm{P}(\ \mathrm{MT} = \mathrm{j},\ \mathrm{CS} = \mathrm{k})$, j=1,2,3,4,5; k=1,2,3.

   Assumptions:

   1. The total number of recorded road casualties N = 27207 is treated as fixed.
   2. The sampling design that led to the collection of the data (bivariate discrete random vector)$\mathrm{X}_1,...,\mathrm{X}_N$ on (Transport Mode, Injury Severity) fixed the row totals for each transport mode in advance.
   3. The counts in each row to be independently distributed as $\mathrm{MN}(\mathrm{n}_j,\mathrm{p}_j = (\mathrm{p}_{j,Fatal},\mathrm{p}_{j,Serious},\mathrm{p}_{j,Slight}))$, for j=1,...,5 distributions. Note: n is the row total.
   4. The probabilities p_{jk} are assumed to be constant in this instance, which means that every casualty has the same probability of falling into each of the 15 cells.
   5. The model does not assume any particular pattern on the combination of CS and MT in advance, it only gives probabilities to each of 15 possible outcomes.

   How it treats the structure of data:

My multinomial distribution treats structure by putting 27207 data in total into 15 (MT,CS)categories, with a vector of 15 cell probabilities, corresponding to MTxCS combinations The total sample size of N=27207 is fixed, while the row and column total are random. And X_1,...,X_N is independent from each other with specified probability p_{jk},and the row and column totals arise as the sums of these cells, as mentioned above.

2. Proportions and Graph:

Table 1: Proportion for each mode

Mode	Fatal	Serious	Slight
Pedestrian	0.0090	0.2621	0.7289
Pedal Cycle	0.0014	0.2004	0.7983
Motorcycle	0.0034	0.1419	0.8547
Car	0.0028	0.0557	0.9416
Other	0.0033	0.1125	0.8841

Table 2: Proportion among all modes

Mode	Fatal	Serious	Slight
Pedestrian	0.0015	0.0439	0.1220
Pedal Cycle	0.0003	0.0375	0.1494
Motorcycle	0.0008	0.0321	0.1932
Car	0.0009	0.0184	0.3115
Other	0.0003	0.0100	0.0783

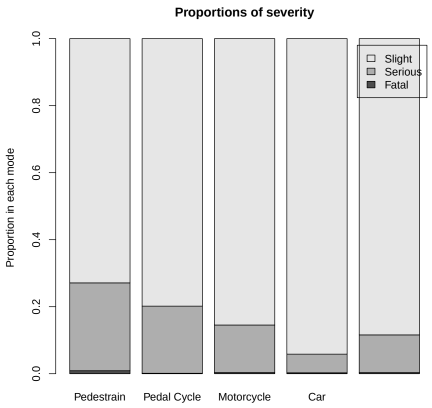

Figure 1: Figure 1: Barplot of proportions for each mode

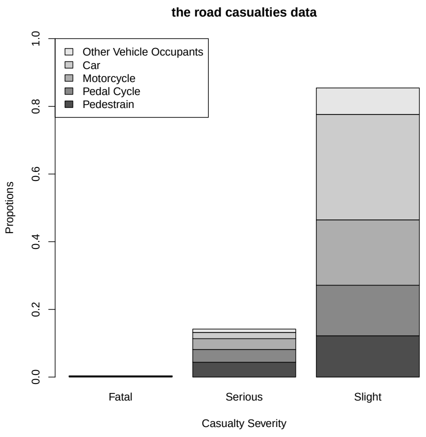

Figure 2: Figure 2: Barplot of proportions for all modes

Figure 3: Summary of proportions and barplots

```
matrix <- matrix(c(41,1194,3320,7,1020,4064,21,873,5257,25,
      +501,8476,8,271,2129), nrow=5, byrow=T)

rownames(matrix)<-c("Pedestrain","Pedal Cycle","Motorcycle",
     "Car","Other Vehicle Occupants")

colnames(matrix)<-c("Fatal","Serious","Slight")

names(dimnames(matrix))<-c("Mode of Transport","Casualty Severity")

mdp <- matrix / rowSums(matrix)

barplot(t(mdp),legend.text = TRUE, beside = FALSE,
       + ylab = "Proportion in each mode",
       + main = "Proportions of severity")

pt<-prop.table(matrix)

ns=0
R=5

for(j in 1:R){
 ns[j]<-sum(matrix[j,])
}
mdp <- matrix / ns

barplot(t(mdp),
        legend.text = TRUE, beside = FALSE,
        ylab = "Proportion in each mode",
        main = "Proportions of severity")

barplot(pt,beside=FALSE,legend.text=TRUE,args.legend
        + = list(x = "topleft"),ylim=c(0,1),ylab="Propotions",
        + xlab ="Casualty Severity",main="the road
        + casualties data")
```

The proportion suggests that Pedestrian has the highest proportion(=0.2621 greater than all the other modes) to be serious or fatal severity, while pedal cycle has the second proportion. Car has the lowest, due to its hard shell outside, however it has 94.16% slight share which is extremely high. Among these 5 modes, Car mode has the highest risk to happen road casualties (any severity). Moreover, it can be seen from table that the fatal casualties are rare.

Interpretation of the results: For Public health and Transport policy: Suggesting for more protective measures for Pedestrian, such as redesigning pedestrian across. Improving helmet policy for cyclist and motorcyclist, and paving more cycle lane. Establish more laws for car speed limit and giving ways policy, but also improve vehicle safety inside car to reduce slight casualties.

3. Null and alternative hypotheses

(i)

H_0(Null hypothesis): Mode of transport(MT) and Casualty Severity(CS) are independent.

vs H_1(Alternative hypothesis): MT and CS are not independent.

Expected frequencies calculation:

Assuming that H_0 is true,

$E_{j,k}$ = n_j* Y_{+k}/N

where n_j is the jth row's sum, Y_{+k} is the kth column's sum.

Mode	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

R code for the frequencies are as follows:

```
chisq.test(matrix)$expected
```

(ii) Using the Chi-square test of independence, The null distribution is

$$X^2 \sim \chi^2_{(5-1)(3-1)} = \chi^2_8$$

R code is

```
chisq.test(matrix, correct=FALSE)
```

Giving that p-value < 2.2e-16(<<0.05)

Conclusion: Based on the chi-square test of independence, with test statistic

$$\chi^2$$

of (5-1)*(3-1)=8 degrees of freedom, the resulting p-value is far below 0.05. Hence, we can reject the null hypothesis: MT and CS are independent. This means that there is evidence that an association between MT and CS exists.

4. Pearson test

Table 3: Pearson residuals

Mode	Fatal	Serious	Slight
Pedestrain	5.7891436	21.556606	-9.1664933
Pedal Cycle	-2.7665176	11.085932	-4.3335914
Motorcycle	-0.4290433	0.018668	0.0208141
Car	-1.5059826	-21.712013	8.9460922
Other	-0.3420334	-3.817279	1.5779672

Interpretion:

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

it measures how far between the observed value and expected value.

Table 3: Pearson standardized residuals

Mode	Fatal	Serious	Slight
Pedestrain	6.3564845	25.5025092	-26.328625
Pedal Cycle	-3.0742281	13.2731720	-12.597168
Motorcycle	-0.4886176	0.0229069	0.062008
Car	-1.8445081	-28.6523708	28.662645
Other	-0.3589280	-4.3161110	4.331708

Interpretion:

$$r_{jk} = \frac{O_{jk} - E_{jk}}{\sqrt{E_{jk}\left[1 - \frac{n_j}{N}\right]\left[1 - \frac{Y_{+k}}{N}\right]}}$$

comparing to the residual it eliminate the influence of differences in sample size and marginal effects, it gives contributions to reject H_0.

The Pedestrian-Serious(with stdres = +25.5), Car-Slight(+28.7) and Car-Serious(-28.7) contribute to rejecting H_0 most, because the first two have the largest and positive residuals and stdres in the table, which indicates that there's much more accident happened than expected under independence. For example, Pedestrian-Serious injuries with 1194 observation, compared with only 646 expectation, which is almost twice of the expectation if MT and CS are unrelated.

On the contrast, Car-Serious cell have the largest negative number in table, which means there's much less accident happened than expected. In this case, the observed count(501) is far below the expected counts(1277).

Implication: Pedestrian are more easily to suffer the serious casualties, while Cars are more likely to suffer the slight injury and far less likely to get serious injuried. In real life, this conclusion indicates that people traveling with unprotected mode of transport face substantially higher severity risk, while car drivers benefited from the vehicle safety protection in reducing serious injury.

R code for Q4 as follows:

```
chisq.test(matrix)$residuals
chisq.test(matrix)$stdres
```

5. Simulating

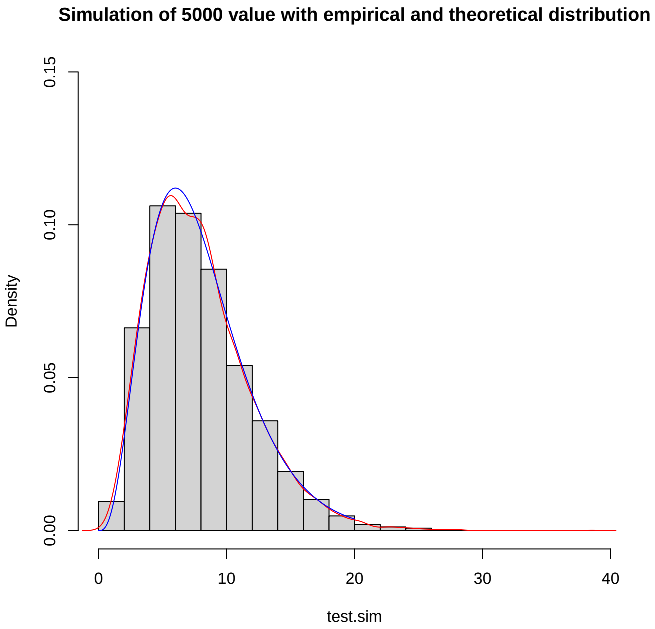

Figure 4: Simulation of 5000 value with empirical and theoretical distribution

R code for the graph:

```
R=5
C=3

N=sum(matrix)

phat=0
for(k in 1:C){
 phat[k]<- sum(matrix[,k])/N
}

B=5000
ysim=matrix(,nrow=R, ncol=C)
```

```
test.sim=0
for(i in 1:B){
 for(k in 1:R){
  ysim[k,]<-rmultinom(n=1, size=ns[k],prob=phat)
 }
 test.sim[i] <- chisq.test(ysim)$statistic
}
hist(test.sim,freq=F,ylim=c(0,0.15),main="Simulation
    + of 5000 value with empirical and
    + theoretical distribution")

lines(density(test.sim),col="red")

xx=seq(from=0, to=35, length.out=600)
dxx=dchisq(xx,df=(R-1)*(C-1))
lines(xx,dxx,col="blue")
```

#Comment:

As shown in the histogram(Figure 4), the empirical distribution of the simulated test statistics (red curve) almost coincides with the theoretical chi-square distribution with 8 degrees of freedom (blue curve). But there is still two small discrepancies need to mention.

First, it can be notice that the empirical distribution(red) has a lower and flatter peak, while the theoretical chi-square one(blue) is with a sharper peak. Second, there's small discrepancies in the upper tail, with the empirical curve sometimes slightly above and sometimes slightly below the theoretical curve, hence this difference is minor.

These discrepancies indicate that although the asymptotic chi-square approximation is almost adequate in this case, it is still not prefect completely. Overall, the main part of the empirical distribution coincides well with the theoretical chi-square curve, indicating that the asymptotic chi-square approximation is adequate for this case. Therefore, using the theoretical chi-square distribution for inference is reliable.

6. **Confidence Interval:**

   (i)Using the R code below, it gives the CI [0.1327275, 0.1566710]

   The large sample of 5091 and 9002 gives a smaller standard error making the interval narrower, leading to higher reliability.

   If the data varies a lot(high variability), the standard error will become larger, and will lead the result to just the opposite (widener interval, lower reliability).

   In this dataset, both groups have large sample sizes, ensuring standard error is small, so the interval is quite reliable.

   (ii)R gives the CI [-0.7251248, -0.7003345]

   Comments: As the result 95% confidence interval is completely within negative interval (entirely below 0), it indicates that the serious injury rate has lower probability than the slight injury rate. Hence, it shows a strong statistical evidence that slight injuries dominates the casualty severity on mode of motorcycle.

```
#6(i)
ser_pedal <- 1020
n_pedal   <- sum(matrix["Pedal Cycle",])

ser_car <- 501
n_car   <- sum(matrix["Car",])

p_pedal <- ser_pedal / n_pedal
p_car   <- ser_car / n_car

diff <- p_pedal - p_car

se <- sqrt( p_pedal*(1-p_pedal)/n_pedal + p_car*(1-p_car)/n_car )

lower <- diff - 1.96*se
upper <- diff + 1.96*se

c(lower, upper)

#6(ii)
ser_mot <- 873
sli_mot <- 5257
```

```
n_mot <- sum(matrix["Motorcycle",])

p_ser <- ser_mot / n_mot
p_sli <- sli_mot / n_mot

diff_mot <- p_ser - p_sli
se_mot <- sqrt( p_ser*(1-p_ser)/n_mot + p_sli*(1-p_sli)/n_mot )

lower2 <- diff_mot - 1.96*se_mot
upper2 <- diff_mot + 1.96*se_mot

c(lower2, upper2)
```