Statistics 2: Computer Practical

Jake Ireland (1908320) Isaac Rawcliffe (1603871)

1 Multinomial distribution

Let Y be a random variable taking values in a finite set $\{1, \dots, m\}$. Then, we can model Y as

$$Y \sim \text{Categorical}(\mathbf{p}), \quad \mathbf{p} = (p_1, \dots, p_m),$$

where $p_k = \mathbb{P}(Y = k)$ and $\sum_{k=1}^{m} p_k = 1$.

Note that Y can be used to model m non-numerical outcomes (categories) by labeling them $1, 2, \dots, m$.

Now assume we observed n i.i.d. Categorical($\mathbf{p}$) random variables, $Y_1, \dots, Y_n$. Their counts are

$$N_k := \sum_{i=1}^{n} \mathbb{I}(Y_i = k),$$

We say that

$$\mathbf{X} = (N_1, \dots, N_m) \sim \text{Multinomial}(n, \mathbf{p}),$$

where $\sum_{i=1}^{m} N_i = n$ and $\mathbf{x} = (n_1, \dots, n_m)$. The PMF of $\mathbf{X}$ is given by

$$f_n(\mathbf{x}; \mathbf{p}) = \frac{n!}{\prod_{i=1}^{m} n_i!} \prod_{i=1}^{m} p_i^{n_i}, \quad \text{with } \sum_{i=1}^{m} n_i = n.$$

The maximum likelihood estimate for $\mathbf{p}$ is given by

$$\hat{\mathbf{p}} = (\hat{p_1}, \dots, \hat{p_m}) = \left(\frac{n_1}{n}, \dots, \frac{n_m}{n} \right).$$

(This result is proved in Section 22 of the notes. For this computer practical you may use it without proof.)

2 Contingency tables

Let X and Y be two categorical random variables so that $X \in \{1, 2, \dots r\}$ and $Y \in \{1, 2, \dots c\}$, i.e X has r possible categories and Y has c possible categories.

Now assume that we have a sample of size n, and each observation in the sample can be classified according to variables X, Y. So we can write the $k - th$ observation in the sample as (x_k, y_k), $k = 1, 2, ..., n$, where x_k and y_k are realizations of X and Y respectively. Hence, we obtain the following arrangement (called a *contingency table*),

		Y				
		1	2	...	c	
	1	n_{11}	n_{12}	...	n_{1c}	$n_{1.}$
	2	n_{21}	n_{22}	...	n_{2c}	$n_{2.}$
X	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	r	n_{r1}	n_{r2}	...	n_{rc}	$n_{r.}$
		$n_{.1}$	$n_{.2}$	...	$n_{.c}$	n

where,

$$n_{ij} = \sum_{k=1}^{n} \mathbb{1}_{\{x_k=i, y_k=j\}}$$

and

$$n_{i.} = \sum_{j=1}^{c} n_{ij}, \qquad n_{.j} = \sum_{i=1}^{r} n_{ij}$$

Moreover, it is clear that

$$\sum_{j=1}^{c} n_{.j} = \sum_{i=1}^{r} n_{i.} = n.$$

Now, let

$$p_{ij} := \mathbb{P}(X = i, Y = j), \quad p_{i.} := \mathbb{P}(X = i), \quad p_{.j} := \mathbb{P}(Y = j)$$

so that

$$\sum_{i=1}^{r}\sum_{j=1}^{c} p_{ij} = \sum_{i=1}^{r} p_{i.} = \sum_{j=1}^{c} p_{.j} = 1.$$

We want to test whether X and Y are independent with regard to the population from which we collected the sample.

Question 1. [1 mark] Explain why a suitable test for the independence of X, and Y is

$$H_0 : p_{ij} = p_{i.} \cdot p_{.j} \qquad \text{Vs} \qquad H_1 : H_0 \text{ not true} \tag{1}$$

— **Solution** —

As we're testing for independence, H_0, the null hypthesis is that X and Y are independent. So H_0 is

$$p_{ij} = P(X = i, Y = j) = P(X = i)P(Y = j) = p_{i.}p_{.j}.$$

Then as the only two cases are X and Y are independent or not independent, the alternative hypothesis is the negation of the null hypothesis.

So the proposed test is a suitable test for independence.

— **End of Solution** —

Question 2. [2 marks] Let $\mathbf{z}$ denote the vector of observed frequencies, $\mathbf{z} := (n_{11}, n_{12}, ...n_{rc}) \in \mathbb{N}^{r\times c}$, and let $\mathbf{p} = (p_{11}, p_{12}, ..., p_{rc})$ be the vector of probabilities. The generalized likelihood ratio is

$$\Lambda_n(\mathbf{z}) := \frac{\sup_{\mathbf{p}\in\Theta_0} f_n(\mathbf{z}; \mathbf{p})}{\sup_{\mathbf{p}\in\Theta} f_n(\mathbf{z}; \mathbf{p})}$$

where $f_n(\cdot; \mathbf{p})$ is the density of a Multinomial distribution with $r \times c$ categories and Θ_0, Θ are the parameter spaces under H_0 and without any restrictions respectively.

Let $G := -2\log\Lambda_n(\mathbf{z})$. Show that G is given by

$$G = -2\sum_{i=1}^{r}\sum_{j=1}^{c} n_{ij}\log\left(\frac{n_{i.}\cdot n_{.j}}{n_{ij}\cdot n}\right).$$

You may use the maximum likelihood estimates for the Multinomial model without proof, but you should explain clearly all the steps and adjustments made in the notation.

— **Solution** —

Let $\mathbf{p}_0 = (p_{1.}p_{.1}, p_{1.}p_{.2}, ..., p_{r.}p_{.c})$. Now $\Theta_0 = \{\mathbf{p}_0\}$ so $\sup_{\mathbf{p}\in\Theta_0} f_n(\mathbf{z};\mathbf{p}) = f_n(\mathbf{z};\mathbf{p}_0)$ and $\sup_{\mathbf{p}\in\Theta} f_n(\mathbf{z};\mathbf{p}) = f_n(\mathbf{z};\hat{\mathbf{p}})$ where $\hat{\mathbf{p}}$ is the ML estimate for $\mathbf{p}$. So

$$\Lambda_n = \frac{f_n(\mathbf{z};\mathbf{p}_0)}{f_n(\mathbf{z};\hat{\mathbf{p}})} = \frac{\Pi_{k=1}^{m}(\mathbf{p}_0)_k^{n_k}}{\Pi_{k=1}^{m}(\hat{\mathbf{p}})_k^{n_k}} = \frac{\Pi_{i=1}^{r}\Pi_{j=1}^{c}(\frac{n_{i.}n_{.j}}{n^2})^{n_{ij}}}{\Pi_{i=1}^{r}\Pi_{j=1}^{c}(\frac{n_{ij}}{n})^{n_{ij}}} = \Pi_{i=1}^{r}\Pi_{j=1}^{c}(\frac{n_{i.}n_{.j}}{n_{ij}n})^{n_{ij}}.$$

This is because the kth element is formed from the ith row and jth column as we iterate through each column for each row so

$$(\mathbf{p}_0)_k^{n_k} = (p_{i.}p_{.j})^{n_{ij}} = (\frac{n_{i.}}{n}\cdot\frac{n_{.j}}{n})^{n_{ij}} = (\frac{n_{i.}n_{.j}}{n^2})^{n_{ij}}$$

and

$$(\hat{\mathbf{p}})_k^{n_k} = (\frac{n_{ij}}{n})^{n_{ij}}.$$

Therefore

$$G = -2\log\Lambda_n(\mathbf{z}) = -2\log(\Pi_{i=1}^{r}\Pi_{j=1}^{c}(\frac{n_{i.}n_{.j}}{n_{ij}n})^{n_{ij}}) = -2\Sigma_{i=1}^{r}\Sigma_{j=1}^{c}\log((\frac{n_{i.}n_{.j}}{n_{ij}n})^{n_{ij}}) = -2\Sigma_{i=1}^{r}\Sigma_{j=1}^{c}n_{ij}\log(\frac{n_{i.}n_{.j}}{n_{ij}n}).$$

— **End of Solution** —

Question 3. [1 mark] We know that under the null hypothesis, G can be approximated by a chi-squared distribution. Explain why its degrees of freedom are $(r-1)(c-1)$.

— **Solution** —

We have that $\dim(\Theta) = rc$ since there are rc free variables in the contingency table. Then $\Theta_0 = \{\mathbf{p}_0\}$ so $\dim(\Theta_0) = r + c - 1$ since the free variables are the row and column totals that form the values in $\mathbf{p}_0$. Hence

$$DoF = \dim(\Theta) - \dim(\Theta_0) = rc - (r + c - 1) = rc - r - c + 1 = (r-1)(c-1)$$

— **End of Solution** —

3 Verifying the asymptotic distribution of G.

Question 4. [3 marks] For this question let $r = 3$, $c = 2$, $n = 100$ and use that $p_{i.} = 1/r,\ i = 1, ..., r$ and $p_{.j} = 1/c,\ j = 1, ..., c$. Under the null hypothesis, simulate (i.e. generate by repeated experiments) $N = 1000$ values of the G test statistic. You may use the following code by replacing the 0's in the indicated parts:

```
r<-3
c<-2
n<-100
N=1000
probx<-rep(1/r,r)
proby<-rep(1/c,c)
G<-rep(NA,N)
for(m in 1:N){
M<-rmultinom(1,size=n,prob=rep(1/(r*c),r*c))
obs<-matrix(M,nrow=r)
Gij<-matrix(NA,nrow=r,ncol=c) #initiates a matrix to keep (i,j)
                              #component of the sum in G
for(i in 1:r){
  for(j in 1:c){
Gij[i,j]<-0 # replace 0 - compute the (i,j) entry
  }
}
G[m]<-0 #replace 0 -this is the m-th simulated value of G
}
```

Verify that the asymptotic distribution of G is chi-squared with 2 degrees of freedom ($\mathcal{X}_2^2$), by:

(a) Plotting the density of the N simulated values of G against the density of $\mathcal{X}_2^2$.

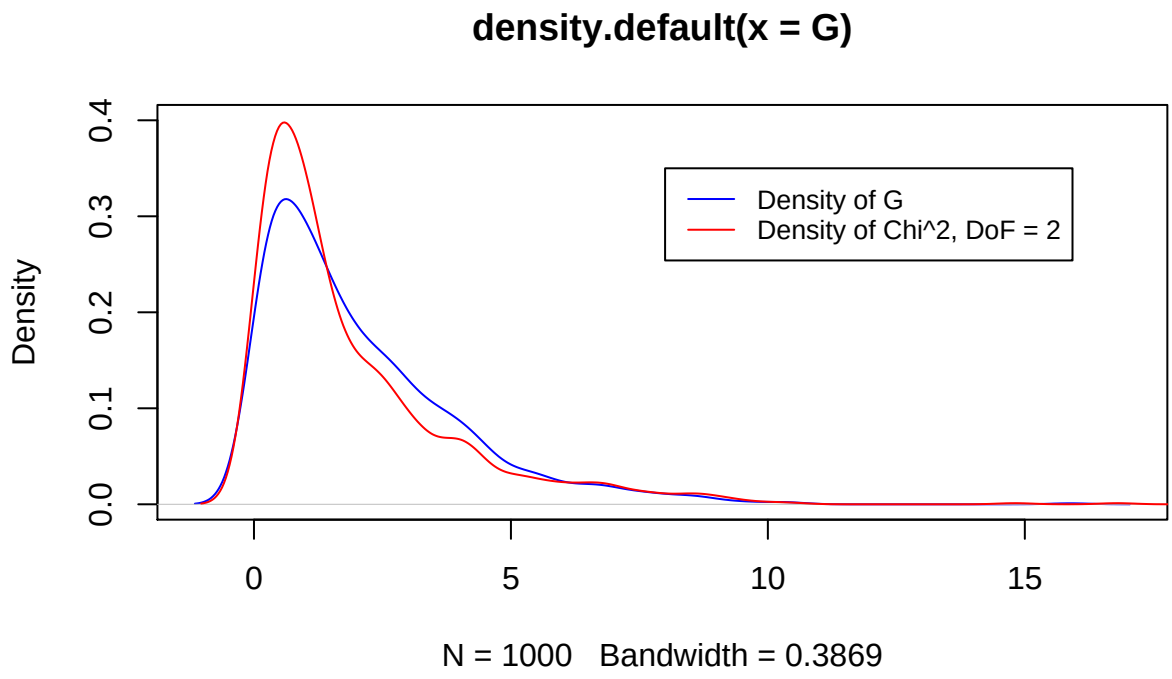
(b) Perform the Pearson goodness-of-fit test for continuous distributions, .i.e. show that the $\mathcal{X}_2^2$ can not be rejected as a good fit for the simulated values of G.

— **Solution** —

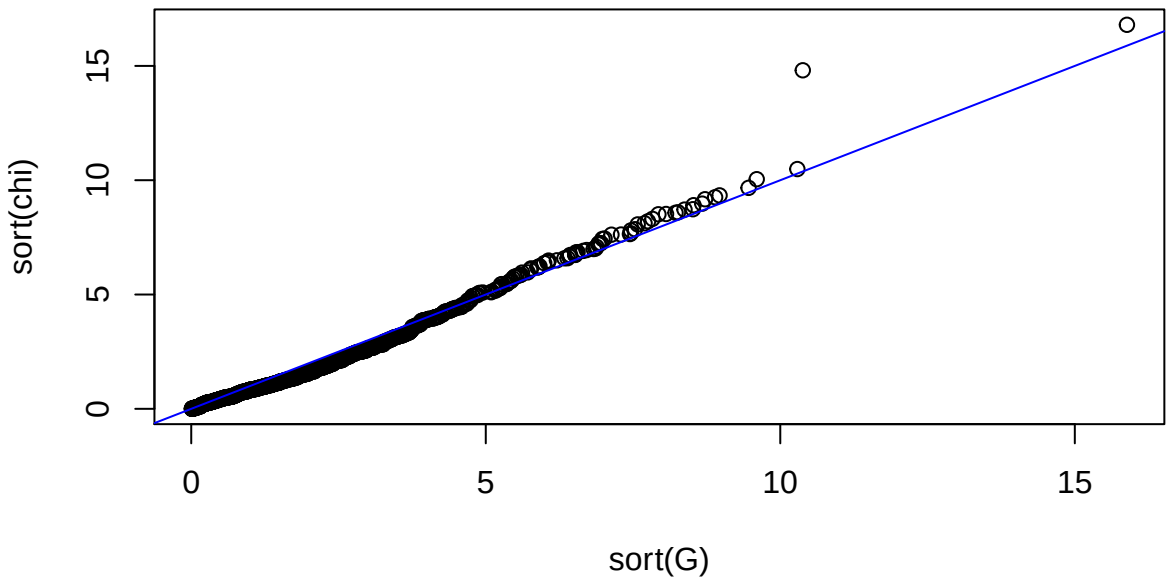
First we will substitute the derived test statistic into the simulation code. Then we will produce two plots. The first will show the densities of the simulated test statistic against the density of the χ_2^2 distribution, the second will plot those values against each other. Both can be used to determine the quality of fit, the second by checking the points fall close to the line $y = x$. Finally we will compute the Pearson goodness-of-fit test.

```
r<-3
c<-2
n<-100
N=1000

probx<-rep(1/r,r)
proby<-rep(1/c,c)

G<-rep(NA,N)

for(m in 1:N){
   M<-rmultinom(1,size=n,prob=rep(1/(r*c),r*c))
   obs<-matrix(M,nrow=r)
   Gij<-matrix(NA,nrow=r,ncol=c) #initiates a matrix to keep (i,j)
                                 #component of the sum in G

   # calculate the column and row totals
   cs <- colSums(obs)
   rs <- rowSums(obs)
```

```
   for(i in 1:r){
      for(j in 1:c){
         # append the value of the test statistic for the ij'th cell
         Gij[i,j] <- obs[i, j] * log((rs[i]*cs[j])/(obs[i, j]*n))
      }
   }

   # Sum values of the test statistic and times by -2
   G[m] <- -2*sum(Gij)
}

# simulate N random numbers from the Chi^2 distribution so we can plot
# its density
chi <- rchisq(N, 2)

# Plot 1: Overlay densities
plot(density(G), ylim=c(0,0.4), col="blue")
lines(density(chi), col="red")
legend(8, 0.35, legend=c("Density of G", "Density of Chi^2, DoF = 2"),
       col=c("blue", "red"), lty=1, cex=0.8)
```




```
# Plot 2: Plot against each other (close to y=x => good fit)
plot(sort(G), sort(chi))
abline(a=0, b=1, col="blue")
```



For the Pearson goodness of fit test we will use the Pearson test statistic

$$T_{\text{Pearson}}(\mathbf{x}) = \sum_{i=1}^{m} \frac{(O_i - E_i)^2}{E_i}$$

which tends in distribution to χ^2_{m-1}. We will quantize the data (ensuring all $E_i > 5$), then we can compute expected values using the χ^2_2 distribution before testing using T_{Pearson} and the χ^2_{m-1} distribution where m is the number of bins which is one less than the number of breaks.

```
# create bins
breaks <- c(seq(0,7,by=0.5),Inf)

# split observed values into bins
Obs <- table(cut(G,breaks))

# compute expected values
prob <- pchisq(breaks[-1],2)-pchisq(breaks[-length(breaks)], 2)
Exp <- N*prob

# display contingency table
round(cbind(Obs,Exp),1)
```

```
##         Obs   Exp
## (0,0.5] 202 221.2
## (0.5,1] 162 172.3
## (1,1.5] 136 134.2
## (1.5,2]  95 104.5
## (2,2.5]  81  81.4
```

```
## (2.5,3]  81  63.4
## (3,3.5]  51  49.4
## (3.5,4]  50  38.4
## (4,4.5]  39  29.9
## (4.5,5]  22  23.3
## (5,5.5]  19  18.2
## (5.5,6]  13  14.1
## (6,6.5]   8  11.0
## (6.5,7]  14   8.6
## (7,Inf]  27  30.2
```

```r
# compute the Pearson test statistic
test.statistic <- sum((Obs-Exp)^2/Exp)

# compute p-value
1-pchisq(test.statistic, df=length(breaks)-2)
```

```
## [1] 0.1595801
```

The p-value suggests little evidence to Reject H_0 so we will retain χ^2_2 as a model for the distribution of the test statistic on a 3×2 contingency table.

— **End of Solution** —

4 Data application

For the remaining questions, we will use data which record the number of homicides in England and Wales, with associated deaths occurring within years 1993-2017. The numbers are recorded separately for males and females and also are recorded by year and age group. In case you are interested, you can access the original data from here https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/deaths/adhocs/009286numberofhomicidesinenglandandwales1993to2017

For this computer practical, we use a modified version as to have only 4 age groups:

- Age_1 for ages < 20
- Age_2 for ages $[20, 34]$
- Age_3 for ages $[35, 59]$
- Age_4 for ages $>= 60$,

and we create one file for males (homicides_males.csv) and one for females (homicides_females.csv). Download the two files and save them in the same working directory as your Rmd file. You can then load the data in RStudio as follows:

```r
females<-read.csv("homicides_females.csv", header=T)
males<-read.csv("homicides_males.csv", header=T)
#View(females)
#View(males)
```

Question 5. [1 mark] Perform hypothesis test (1) to check whether age and gender are independent for the recorded homicides in years 2015 and 2016. The first step is to form the appropriate contingency table by extracting the appropriate information from the two datasets:

```
homicides_age_gender<-matrix(NA,nrow=2,ncol=4)
colnames(homicides_age_gender)<-c("<20","20-34","35-59",">=60")
rownames(homicides_age_gender)<-c("Males", "Females")
for(j in 1:4){
    homicides_age_gender[1,j]<- males[males$Year==2015,j+1]+
   males[males$Year==2016,j+1] #j+1 as data have extra column for year
}
for(j in 1:4){
    homicides_age_gender[2,j]<- females[females$Year==2015,j+1]+
   females[females$Year==2016,j+1]
}
homicides_age_gender
```

```
##         <20 20-34 35-59 >=60
## Males    37   278   286  113
## Females  33    87   132   69
```

— **Solution** —

For this question we reused the appropriate code from question 4 using a chi-squared distribution with $3 = (2-1)(4-1)$ degrees of freedom.

```
obs <- homicides_age_gender
r <- 2
c <- 4
n <- sum(obs)

# code from question 4
Gij<-matrix(NA,nrow=r,ncol=c)

cs <- colSums(obs)
rs <- rowSums(obs)

for(i in 1:r){
   for(j in 1:c){
      Gij[i,j] <- obs[i, j] * log((rs[i]*cs[j])/(obs[i, j]*n))
   }
}

G <- -2*sum(Gij)

G
```

```
## [1] 21.12914
```

```
# chi-squared with 3 DoF
1-pchisq(G, df=(r-1)*(c-1))
```

```
## [1] 9.897092e-05
```

The small $p-value$ suggests we should reject H_0, so for years 2015-2016, age is not independent of gender.

— **End of Solution** —

Question 6. [2 marks] For this question you will need to use the records for years 1995-1996 (representing the past) and years 2015-2016 (representing the recent years). Using the categorical variables *gender* (females, males) and *time* (1995-96, 2015-16), form the appropriate contingency table and perform the independence test for the variables age and time. State your conclusion.

You can initiate your contingency table as follows:

```
homicides_gender_time<-matrix(NA,nrow=2,ncol=2)
colnames(homicides_gender_time)<-c("1995-1996", "2015-2016")
rownames(homicides_gender_time)<-c("Males","Females")
homicides_gender_time
```

```
##         1995-1996 2015-2016
## Males          NA        NA
## Females        NA        NA
```

— **Solution** —

First we need get the appropriate values from the DataFrame which we can do, making sure to remove the Year column before we sum.

```
homicides_gender_time[1,1] <- sum(subset(males[males$Year==1995 | males$Year==1996,], select = -c(Year))
homicides_gender_time[1,2] <- sum(subset(males[males$Year==2015 | males$Year==2016,], select = -c(Year))
homicides_gender_time[2,1] <- sum(subset(females[females$Year==1995 | females$Year==1996,], select = -c
homicides_gender_time[2,2] <- sum(subset(females[females$Year==2015 | females$Year==2016,], select = -c
homicides_gender_time
```

```
##         1995-1996 2015-2016
## Males        1037       714
## Females       509       321
```

Then we can use the same code as before with different values for "obs", "r" and "c".

```
obs <- homicides_gender_time
r <- 2
c <- 2
n <- sum(obs)

Gij<-matrix(NA,nrow=r,ncol=c)

cs <- colSums(obs)
rs <- rowSums(obs)

for(i in 1:r){
   for(j in 1:c){
      Gij[i,j] <- obs[i, j] * log((rs[i]*cs[j])/(obs[i, j]*n))
   }
}

G <- -2*sum(Gij)

G
```

```
## [1] 1.038045
```

```r
1-pchisq(G, df=(r-1)*(c-1))
```

```
## [1] 0.3082767
```

The reasonably large p-value suggests we should retain H_0, implying gender was independent from homicides in 1995-1996 and 2015-2016.

— **End of Solution** —

5 Epilogue

In this computer practical we have covered a new application for the generalized likelihood ratio test that seems a bit different from what we covered in the lectures notes. We studied how we can test the independence of two categorical variables using contingency tables. For the same test, we could have used the Pearson test statistic,

$$T_{Pearson} = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}},$$

where $O_{ij} = n_{ij}$ and $E_{ij} = \frac{n_{\cdot j} n_{i\cdot}}{n}$. For sufficiently large samples it can be approximated by the same distribution as G.

The G test statistic can be performed in R easily using the GTest function from the DescTools package

```r
#install.packages("DescTools")
library(DescTools)
#GTest(table) #table is the contingency table

GTest(homicides_age_gender)
GTest(homicides_gender_time)
```

You can use this **only** to confirm your results in questions 5 and 6.