Dr Isabella Gollini

Probability and Statistics

Solution Extra Examples – Lab – Test for Independence of two categorical variables

1 Survey

The data frame `survey` in the `MASS` package contains the responses of 237 Statistics I students at the University of Adelaide to a number of questions (type `?survey` for further details). We are interested in the following three categorical variables:

`Sex` The sex of the student. (`Male` and `Female`.)

`W.Hnd` The writing hand of student. (`Left` and `Right`.)

`Exer` How often the student exercises. (`Freq` (frequently), `Some`, `None`.)

The commands to load and visualise the data frame are:

```
library(MASS)
data(survey)
attach(survey)
```

```
head(survey)
```

(a) Carry out a chi-square test to investigate whether there is an association between the sex and writing hand of student, and draw conclusions.

[HINT: To create the table use the command: `tabSWh <- table(Sex, W.Hnd)`]

(b) Carry out a chi-square test to investigate whether there is an association between the sex of the student and how often the student exercises, and draw conclusions.

Solution

(a)
```
tabSWh <- table(Sex, W.Hnd)
tabSWh
```

```
##         W.Hnd
## Sex      Left Right
##   Female    7   110
##   Male     10   108
```

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

```
##
##  Pearson's Chi-squared test
##
## data:  tabSWh
## X-squared = 0.54351, df = 1, p-value = 0.461
```

The p-value > 0.05, thus at 5% significance level we do not reject the null hypothesis of independence between the variables `Sex` and `W.Hnd`. There is no strong evidence of any association between the sex and writing hand of student.

(b)
```
tabSS <- table(Sex, Exer)
tabSS
```

```
##         Exer
## Sex      Freq None Some
##   Female   49   11   58
##   Male     65   13   40
```

```
##
##  Pearson's Chi-squared test
##
## data:  tabSS
## X-squared = 5.7184, df = 2, p-value = 0.05731
```

The p-value > 0.05, thus at 5% significance level we do not reject the null hypothesis of independence between the variables `Sex` and `Exer`. There is no strong evidence of any association between the sex of the student and how often the student exercises.

2 Voting intentions

In the example shown in class regarding stated voting intentions in two neighbouring constituencies, the data was restricted to whether the respondents stated that they intended to vote Conservative or not. In the table below more detailed data are provided including information on whether they stated that they intended to vote Liberal or Labour.

	Conservative	Liberal	Labour	Other/ Undecided
Constituency 1	73	23	54	50
Constituency 2	43	22	12	23

Again it is to be investigated whether there is any association between constituency and stated voting intentions.

(i) State the null and alternative hypotheses to be tested.

(ii) Under the null hypothesis, calculate a table of expected frequencies.

(iii) Carry out the relevant chi-square test and draw conclusions.

(iv) In particular, state what is the nature of any apparent association.

Solution

Load the data in R:

```
tab <- matrix(c(73, 23, 54, 50,
  43, 22, 12, 23),
  2, 4,
  byrow = TRUE)
colnames(tab) <- c("Conservative", "Liberal", "Labour", "Other")
rownames(tab) <- c("Constituency 1", "Constituency 2")
addmargins(tab)

##                Conservative Liberal Labour Other Sum
## Constituency 1           73      23     54    50 200
## Constituency 2           43      22     12    23 100
## Sum                     116      45     66    73 300
```

(i) The null hypothesis is that there is no association between constituency and stated voting intention. The alternative hypothesis is that there is an association between constituency and stated voting intention.

(ii) The table of expected frequencies under the null hypothesis is given below. First the column totals C_j from the table in the question have to be entered. Using also the row totals R_i, the expected frequencies are given by $E_{ij} = R_i C_j/300$, and calculations can be reduced by using the fact that the row and column totals of the expected frequencies must match those of the observed frequencies.

	Conservative	Liberal	Labour	Other/ Undecided	total
Constituency 1	77.33	30.00	44.00	48.67	200
Constituency 2	38.67	15.00	22.00	24.33	100
total	116	45	66	73	300

(iii)

$$X^2 = \sum_{i=1}^{2}\sum_{j=1}^{4} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = 12.554$$

The individual terms that make up the sum are given in the table below.

	Conservative	Liberal	Labour	Other/ Undecided
Constituency 1	0.242	1.633	2.273	0.036
Constituency 2	0.485	3.267	4.545	0.073

The associated degrees of freedom are (2 - 1)(4 - 1) = 3. From Table 8 of *Lindley and Scott*, $\chi^2_3(1) = 11.34$. Our test statistic is significant at the 1% level. There is very strong evidence of an association between constituency and stated voting intention.

R code:

```
test <- chisq.test(tab,
  correct = FALSE)
test
```

```
##
##  Pearson's Chi-squared test
##
## data:  tab
## X-squared = 12.556, df = 3, p-value = 0.005702
```

This is the table to answer to point (ii)

```
round(addmargins(test$expected), 2)
```

```
##                Conservative Liberal Labour Other Sum
## Constituency 1        77.33      30     44 48.67 200
## Constituency 2        38.67      15     22 24.33 100
## Sum                  116.00      45     66 73.00 300
```

(iv)
```
round((tab - test$expected)^2 / (test$expected), 3)
```

```
##                Conservative Liberal Labour Other
## Constituency 1        0.243   1.633  2.273 0.037
## Constituency 2        0.486   3.267  4.545 0.073
```

The biggest components of the chi-square statistic are the ones for Liberal and Labour. Labour is particularly badly supported and Liberals well supported in Constituency 2. In Constituency 1 Labour appears to be the main rival to the Conservatives, but in Constituency 2 it is the Liberals.

3 Treatment for Influenza

Question 2 of Examples 6 states:

> "In a pandemic of a virulent new strain of influenza, a randomly selected 150 of influenza patients admitted to a hospital in a given week were given treatment A and the remaining 300 were given treatment B. Of the 150 patients given treatment A, 15 died, and of the 300 patients given treatment B, 45 died"

(i) Recast the data of Question 2 of Examples 6 in the form of a 2×2 contingency table.

(ii) Carry out a chi-square test to investigate whether there is an association between which treatment is given and whether death occurs, and draw conclusions.

(iii) Check that this method of analysis gives results that are exactly equivalent to those found using the test for comparing two proportions.

Solution

(i) The observed frequencies are given in the following table.

	Died	Survived	total
Treatment A	15	135	150
Treatment B	45	255	300
total	60	390	450

R code:

```
tab <- matrix(c(15, 135, 45, 255),
  2, 2,
  byrow = TRUE)
colnames(tab) <- c("Died", "Survived")
rownames(tab) <- c("Treatment A", "Treatment B")
addmargins(tab)
```

```
##             Died Survived Sum
## Treatment A   15      135 150
## Treatment B   45      255 300
## Sum           60      390 450
```

(ii) The corresponding expected frequencies ($E_{ij} = R_i C_j / 450$) under the null hypothesis of no association are given in the following table.

	Died	Survived	total
Treatment A	20	130	150
Treatment B	40	260	300
total	60	390	450

$$X^2 = 5^2(1/20 + 1/130 + 1/40 + 1/260) = 2.163 \ ,$$

with 1 degree of freedom. From Table 8 of *Lindley and Scott*, $\chi^2_1(10) = 2.706$. Our test statistic is not significant even at the 10% level. There is no strong evidence of any association between which treatment is given and whether death occurs.

R code:

```
test <- chisq.test(tab,
  correct = FALSE)
test


##
##  Pearson's Chi-squared test
##
## data:  tab
## X-squared = 2.1635, df = 1, p-value = 0.1413

round(addmargins(test$expected), 2)

##             Died Survived Sum
## Treatment A   20      130 150
## Treatment B   40      260 300
## Sum           60      390 450
```

(iii) In Question 2 of Examples 6, $z = -1.47$. Here $X^2 = 2.163 = (-1.47)^2$, given some rounding error. In both cases, $p = 0.141$.

R code:

```
prop.test(c(15, 45), c(150, 300),
correct = FALSE)


##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  c(15, 45) out of c(150, 300)
## X-squared = 2.1635, df = 1, p-value = 0.1413
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.11274946  0.01274946
## sample estimates:
## prop 1 prop 2
##   0.10   0.15
```