**Dr Isabella Gollini**

**Probability and Statistics**

# Lab - Test for independence of two categorical variables

- Section 1 contains a summary of the test for independence in two-way contingency tables.
- Section 2 [page 3] contains a guided example
- Section 3 [page 7] contains other three exercises. [Try to do them on your own, but do not hesitate to ask any questions!]

Remember to write your code into an `R` script, in order to be able to save and reuse it.

## 1.1 Test for Independence in Two-way Contingency tables

1. Each observation is classified according to two different criteria. The first variable has $r$ categories and the second has $s$ categories.

   The data may be summarized in a $r \times s$ contingency table of observed frequencies:

$$\begin{array}{cccc|c} O_{11} & O_{12} & \dots & O_{1s} & R_1 \\ O_{21} & O_{22} & \dots & O_{2s} & R_2 \\ \vdots & & \ddots & \vdots & \vdots \\ O_{r1} & O_{r2} & \dots & O_{rs} & R_r \\ \hline C_1 & C_2 & \dots & C_s & n \end{array}$$

   where the general entry $O_{ij}$ is the observed count of sample members who are classified into category $i$ according to the first variable and into category $j$ according to the second variable. $R_i$ is the $i$th row total, and $C_j$ is the $j$th column total.

2. $H_0$ **and** $H_1$

   $H_0$ : There is no association between the two variables, i.e., they are independent.

   $H_1$ : There is an association between the two variables, i.e., they are not independent.

3. **Test Statistic**

   To test the null hypothesis $H_0$ we use the test statistic:

$$X^2 = \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \quad \sim \chi^2_{(r-1)\times(s-1)}$$

   where $E_{ij} = \frac{R_i C_j}{n}$ is the expected frequency for the $(i, j)$th cell under the null hypothesis.

   Requirement: The sample size $n$ should be large enough such that $E_{ij} < 5$ in no more than 20% of the cells, and there is no expected frequency $E_{ij} < 1$.

4. **Rejection region Significance-level-$\alpha$ test**

   Reject $H_0$ if $X^2 \geq \chi^2_{(r-1)\times(s-1),\alpha}$. Do not reject $H_0$ otherwise.

5. $p$**-value**

$$p-\text{value} = 1 - F(X^2)$$

   where $F(X^2)$ is the c.d.f. of a $\chi^2$-distribution with $(r-1) \times (s-1)$ degrees of freedom.

   Reject $H_0$ if $p-\text{value} \leq \alpha$. Do not reject $H_0$ otherwise.

**`R` function**

`Tab` is a $r \times s$ matrix containing the observed frequencies $O_{ij}$.

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

# 2 Guided Examples

## 2.1 Hypertension

A doctor wants to analyse if overweight is associated with higher prevalence of hypertension. For this reason he collects data from 324 patients:

| | Overweight | NOT Overweight |
|---|---|---|
| Hypertension | 121 | 63 |
| NO Hypertension | 52 | 88 |

The aim of this exercise is to carry out a chi-square test to investigate whether there is an association between overweight and hypertension.

Firstly we create a $2 \times 2$ matrix that contains the data:

```
tab <- matrix(c(121, 63, 52, 88),
  2, 2, byrow = TRUE)

rownames(tab) <- c("Hypertension", "NO Hypertension")
colnames(tab) <- c("Overweight", "NOT Overweight")

tab
```

We can visualise the row and the column sums with the function `addmargins`:

```
addmargins(tab)
```

The `chisq.test` function allows us to perform the chi-square test of independence in `R`. If the argument specified is a matrix, it automatically performs the test of independence, and calculates the degrees of freedom correctly:

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

The output of the test includes the matrix `test$expected` containing the expected frequencies under the null hypothesis of independence. We can visualise this table with the row sums and column sums by using the `addmargins` function, so we can notice that the row and column sums are the same observed in the matrix `tab`.

```
addmargins(test$expected)
```

Draw conclusions and interpret your conclusion in the context of the application:

Knowing the value of the test statistic, try to get to the conclusions by using the statistical tables:

You can not use the exact value of the $p$-value, but you can use the statistical tables to get an upper bound of the $p$-value.

The value of the test statistics $X^2 = 26.167$ is higher then 12.12 that is the percentage point for the 0.05 percentage point (that is the lowest you can get from the Table 8 of *Lindley and Scott*). So, you can conclude that the $p$-value ¡ 0.0005.

An alternative approach is to use Table 7 of *Lindley and Scott*, and notice that the maximum value you can get for a chi-square distribution with 1 degree of freedom is $F_1(8) = 0.9953$, since the test statistic is $X^2 = 26.167$, we can conclude that the $p$-value is less then $1 - 0.9953 = 0.0047$.

### 2.1.1 Test between two proportions and Chi-Square test of independence in $2 \times 2$ contingency tables

The chi-square test of independence in $2 \times 2$ contingency tables, when used for comparing two proportions, is equivalent to the test for comparing two proportions in Section 11.4.

We can use a $2 \times 2$ contingency table to display the frequency of occurrence of successes and failures for two groups:

| | Success | Failure | |
|---|---|---|---|
| Group 1 | $n_1\hat{p}_1$ | $n_1(1-\hat{p}_1)$ | $n_1$ |
| Group 2 | $n_2\hat{p}_2$ | $n_2(1-\hat{p}_2)$ | $n_2$ |
| | $n\hat{p}$ | $n(1-\hat{p})$ | $n$ |

where:

- $n_1$ and $n_2$ are the number of observations from Group 1 and Group 2 respectively.
- $\hat{p}_1$ and $\hat{p}_2$ are the observed proportions of successes in Group 1 and Group 2 respectively.
- $n = n_1 + n_2$ is the total number of observations.
- $\hat{p} = \dfrac{n_1\hat{p}_1 + n_2\hat{p}_2}{n}$.

Using this notation we get that $|O_{ij} - E_{ij}| = \dfrac{|\hat{p}_1 - \hat{p}_2|}{1/n_1 + 1/n_2}$ for all the entries of the table.

After some simplifications the test statistic for $2 \times 2$ contingency tables can be written as:

$$X^2 = \frac{(\hat{p}_1 - \hat{p}_2)^2}{\hat{p}\hat{q}\,(1/n_1 + 1/n_2)} \quad \sim \chi^2_1$$

where $\hat{q} = (1 - \hat{p})$. You can notice that the square of the $z$-statistic used in Section 11.4 is equal to the chi-square statistic used here.

It follows from the fact that if $Z \sim N(0, 1)$ then $Z^2 \sim \chi^2_1$. If a two-tail test is used with the $z$-statistic then the $p$-values for both statistics are identical.

Now we verify numerically from the R outputs that the chi-square test of independence in $2 \times 2$ contingency tables gives results that are exactly equivalent to those found using the two-tail test for comparing two proportions.

We can do two different tests of proportions:

- In the first one we fix our populations of interest to be the people that suffer from hypertension, and the people who do not suffer from hypertension, and test if the difference in the proportions of overweight people within the two groups is significantly different from 0.

  The point estimate of the proportion of people with hypertension who are overweight is:

```
pH_O <- 121 / 184
pH_O
```

  The point estimate of the proportion of people without hypertension who are overweight is:

```
pNH_O <- 52 / 140
pNH_O
```

  The test to compare the proportion of overweight people in the two groups formed by people with hypertension or not is:

```
prop.test(c(121, 52), c(184, 140),
  correct = FALSE)
```

Draw conclusions in the context of the test, and compare this output with the one obtained with the chi-square test:

- In the second one we fix our populations of interest to be the overweight/not overweight people, and test if the proportions of people with hypertension in the two groups is significantly different from 0.

  The point estimate of the proportion of overweight people who have hypertension is:

```
pO_H <- 121 / 173
pO_H
```

and the point estimate of the proportion of people that are not overweight who have hypertension is::

```
pNO_H <- 63 / 151
pNO_H
```

The test to compare the proportion of people with hypertension in the two groups formed by overweight and not overweight people is:

```
prop.test(c(121, 63), c(173, 151),
  correct = FALSE)
```

Draw conclusions in the context of the test, and compare this output with the outputs above:

# 3 Extra Examples

## 3.1 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.

## 3.2 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.

## 3.3 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 $2 \times 2$ 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.