

# Lab 07 – Chi-squared test

## ENVX1002 Handbook

Semester 1, 2025

### Learning Outcomes

- Learn to use R to calculate a chi-squared test for:
  - Test of proportions
  - Test of independence
- Learn how to interpret statistical output.

### Before we begin

Create your Quarto document and save it as Lab-07.qmd or similar. There are no data files to download for this lab.

### Quick introduction

The chi-square test is used to compare the *observed* distribution to an *expected* distribution, in a situation where we have **two or more categories** in a discrete data. In other words, it compares multiple observed proportions to expected probabilities.

The formula is:

$$\chi^2 = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i}$$

where  $O_i$  is the observed frequency,  $E_i$  is the expected frequency for each category and  $k$  is the number of categories.

For more information about the technique, consult your lecture slides and tutorial 7.

### Exercise 1: wild tulips (walk-through)

#### Background

Suppose we collected wild tulips and found that 81 were red, 50 were yellow and 27 were white. Are these colours equally common?

**If these colours were equally distributed, the expected proportion would be 1/3 for each of the colour.** Therefore, we want to test if the observed proportions are significantly different from the expected proportions.

The data is below.

```
tulip <- c(81, 50, 27)
```

### Instructions

Utilise the **HATPC** process and test the hypothesis that the proportion of flower colours of tulips are equally common, assuming that the samples are independent. We can explore the data as we check the assumptions of the test.

HATPC:

- Hypothesis
- Assumptions
- Test (statistic)
- P-value
- Conclusion

#### **i** Level of significance

**The level of significance is usually set at 0.05.** This value is generally accepted in the scientific community and is also linked to Type 2 errors, where choosing a lower significance increases the likelihood of failing to reject the null hypothesis when it is false.

### Hypotheses

What are the null hypothesis and alternative hypotheses?

[Click here to view answer](#)

- $H_0$ : There is no significant difference between the observed and the expected proportions of flower colours.
- $H_1$ : There is a significant difference between the observed and the expected proportions of flower colours.

### Assumptions

Recall that the assumptions of the  $\chi^2$  test are:

1. No cell has expected frequencies less than 1
2. No more than 20% of cells have expected frequencies less than 5

#### **i** Note

In the case that the above assumptions are violated then the probability of a type 1 error occurring (rejecting the null hypothesis when it is true, i.e. false positive) increases.

To calculate expected frequencies, we first calculate the total number of tulips and then divide by the number of categories.

```
expected <- rep(sum(tulip) * 1 / 3, 3) #rep function replicated the value we're
calculating inside the brackets
expected
```

```
[1] 52.66667 52.66667 52.66667
```

Does the data satisfy the assumptions of a  $\chi^2$  test?

[Click here to view answer](#) Yes, as expected frequencies > 5.

### Test statistic

The `chisq.test()` function in R is used to calculate the chi-squared test.

```
res <- chisq.test(tulip, p = c(1 / 3, 1 / 3, 1 / 3))
res
```

Chi-squared test for given probabilities

```
data: tulip
X-squared = 27.886, df = 2, p-value = 8.803e-07
```

Note that we could check our assumptions post-analysis by checking the expected frequencies stored in the `expected` object of the output:

```
res$expected
```

```
[1] 52.66667 52.66667 52.66667
```

### P-value

Write down how you should report the critical value, p-value and df in a scientific paper?

[Click here to view answer](#)  $\chi^2 = 27.9, d.f. = 2, p < 0.01$

### Conclusions

Based on the p-value, do we accept or reject the null hypothesis?

[Click here to view answer](#) We reject the null hypothesis as the p-value is less than 0.05.

Now write a scientific (biological) conclusion based on the outcome.

[Click here to view answer](#) There is a significant difference in the proportion of flower colours of tulips ( $\chi^2 = 27.9, d.f. = 2, p < 0.01$ ).

## Exercise 2: hermit crabs

## Background

In a study of hermit crab behaviour at Point Lookout, North Stradbroke Island, a random sample of 3 types of gastropod shells was collected. Each shell was then scored as being either occupied by a hermit crab or empty. Do hermit crabs prefer a certain shell?

| Shell species  | Occupied | Empty |
|----------------|----------|-------|
| Austrochochlea | 47       | 42    |
| Bembicium      | 10       | 41    |
| Cirithid       | 125      | 49    |

The data is stored in a `table` object in R below. Note that it is different from a `data.frame` object. You can verify this by using the `str()` or `class()` functions.

```
crabs <- as.table( #Make a table with values for each row
  rbind(
    Aus = c(47, 42),
    Bem = c(10, 41),
    Cir = c(125, 49)
  )
)

colnames(crabs) <- c("Occupied", "Empty") #Add column names to table
str(crabs)
```

```
'table' num [1:3, 1:2] 47 10 125 42 41 49
- attr(*, "dimnames")=List of 2
..$ : chr [1:3] "Aus" "Bem" "Cir"
..$ : chr [1:2] "Occupied" "Empty"
```

crabs

|     | Occupied | Empty |
|-----|----------|-------|
| Aus | 47       | 42    |
| Bem | 10       | 41    |
| Cir | 125      | 49    |

## Data exploration

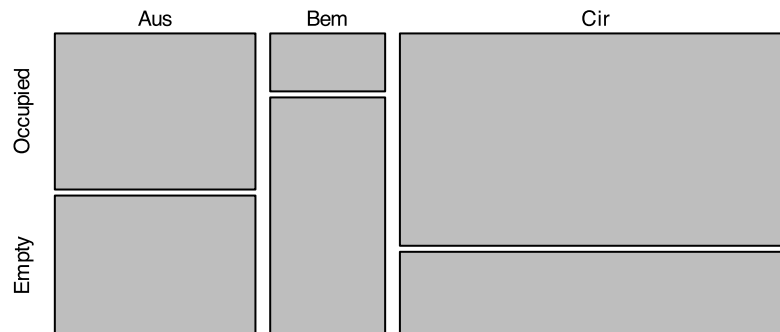
Since we have a multi-dimensional dataset, we can try to plot the data to visualise it.

A mosaic plot is a graphical representation of the data in a two-way contingency table. It is a way of visualising the relationship between two categorical variables.

Try the code below. Can you interpret the plot?

```
# mosaic plot of crabs
mosaicplot(crabs, main = "Hermit crabs and shell species")
```

## Hermit crabs and shell species



Click [here](#) to view interpretation The plot shows that the distribution of hermit crabs in the different shell species is not equal. The majority of hermit crabs are found in the *Cirithid* shell species, followed by *Austrochochlea* and *Bembicium*. This can be observed by the **width** of the boxes in the plot.

There are also differences in the number of empty shells in the different shell species. The *Bembicium* shell species has the highest number of empty shells, followed by *Austrochochlea* and *Cirithid*. This can be observed by the **height** of the boxes in the plot.

### HATPC

Now it is your turn to test the hypothesis that the three shell species are equally preferred by hermit crabs. Follow the HATPC process with the following questions in mind (but you don't have to answer them individually):

1. What are the null hypothesis and alternative hypotheses?
2. Does the data satisfy the assumptions of a  $\chi^2$  test?
3. How should you report the critical value, p-value and df in a scientific paper?
4. Based on the p-value, do we accept or reject the null hypothesis?
5. Write a scientific (biological) conclusion based on the outcome.

Use the markdown structure below to answer the questions (if you wish):

```
## Exercise 2: hermit crabs
### Hypothesis
### Assumptions
### Test statistic
### P-value
### Conclusions
```

Take your time, and when you are ready, check your answers with your demonstrators.

## Done!

This is the end of the lab. Remember to save your work, render your document and ask your demonstrators for feedback if you are unsure about your answers.

## Bonus take home exercises

### Exercise 1: National RSPCA statistics

The RSPCA releases statistics on the number of animals they receive, reclaim and rehome every year. In the 2023-24 financial year, the RSPCA received 17468 dogs, 26704 cats, and 37497 other animals. The “other” category includes horses, small animals, livestock and wildlife.

Using the HATPC framework, test whether these animals were received in equal proportions.

```
received <- c(17468, 26704, 37497)
```

### Exercise 2: UC Berkeley Admissions

For the two exercises we’re going to use simplified versions of the inbuilt dataset ‘UCBAdmissions’, which has data on student admissions for Berkeley. The dataset shows how many students were rejected and admitted to the university by both department and gender.

#### 2.1 Admissions by department

Did every department at UC Berkeley admit students in equal proportions?

```
dept_admissions <- c(601, 370, 322, 269, 147, 46)
```

#### 2.2

Are male and female students admitted and rejected in the same proportion?

```
gender <- as.table( #Make a table with values for each row
  rbind(
    admitted = c(1198, 557),
    rejected = c(1493, 1278)
  )
)
```

```
colnames(gender) <- c("Male", "Female") #Add column names to table
```