Chi-squared tests

ENVX1002 Introduction to Statistical Methods

Januar Harianto
The University of Sydney

Apr 2024

CRICOS 00026A TEQSA PRV12057

Outline

- Recap
- Categorical data
- Chi-squared distribution
- The Chi-squared test
- Example: Goodness of fit
- Example: Test of independence
- How do we visualise the differences in a contingency table?
- What about the test of homogeneity?
- Summary
- Thanks!

Recap

Parametric and non-parametric alternatives

- So far, **all** of our techniques have been aimed at comparing **means/medians** of continuous variables.
- The *assumption of normality* **underpins** these techniques – if the data is not normally distributed, we have alternatives like *transforming* the data or using *non-parametric tests*.
- **Does this apply to all data?**

A rational assumption?

- **Are all randomly sampled data normally distributed?**
- Recall probability distributions (Week 3) – **normal distribution** is just *one of several* possible distributions of data.
- It turns out that there are non-parametric techniques that are not just *alternatives* of parametric tests, but **better suited** for certain types of data.

Categorical data

Some data are not measured on a continuous scale, but rather as **categories**.

What are categorical variables?

Consider the following questions:

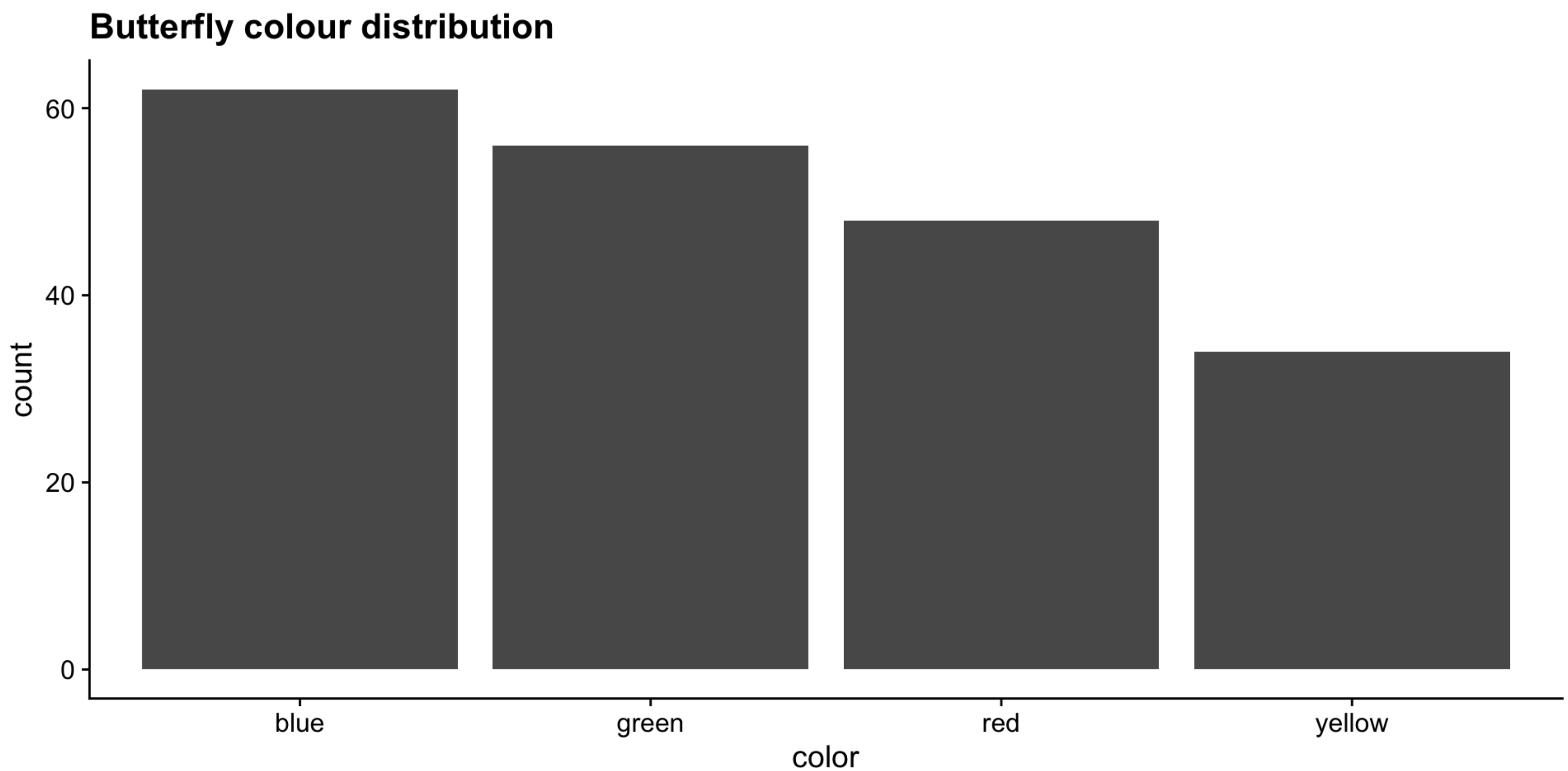
> A biologist claims that when sampling the Australian Botanical Gardens for butterflies, the ratio of the most dominant colours (red, blue, green, and yellow) is equal. How would you determine if the biologist's claim is true?

> A study was conducted on a population of deer to see if there is a relationship between their age group (young, adult, old) and their preferred type of vegetation (grass, leaves, bark). Is age group of the deer independent of their vegetation preference?

How would you **measure** these variables, and what sort of summary statistics can you use?

Visualising categorical variables

▶ Code



We can only **count** the number of times a particular category occurs, or the **proportion** of the total that each category represents.

Types of categorical data

- Rather than measuring a continuous variable, we are interested in **counting** the number of times a particular category occurs, or the **proportion** of the total that each category represents.
- These are known as **categorical variables**.
- Generally 3 types of categorical data:
    - ➟ **Nominal**: Categories have no inherent order (e.g. colours, breeds of dogs).
    - ➟ **Ordinal**: Categories have an inherent order (e.g. Likert scales, grades).
    - ➟ **Binary**: Only two mutually exclusive categories (e.g rain or no rain).

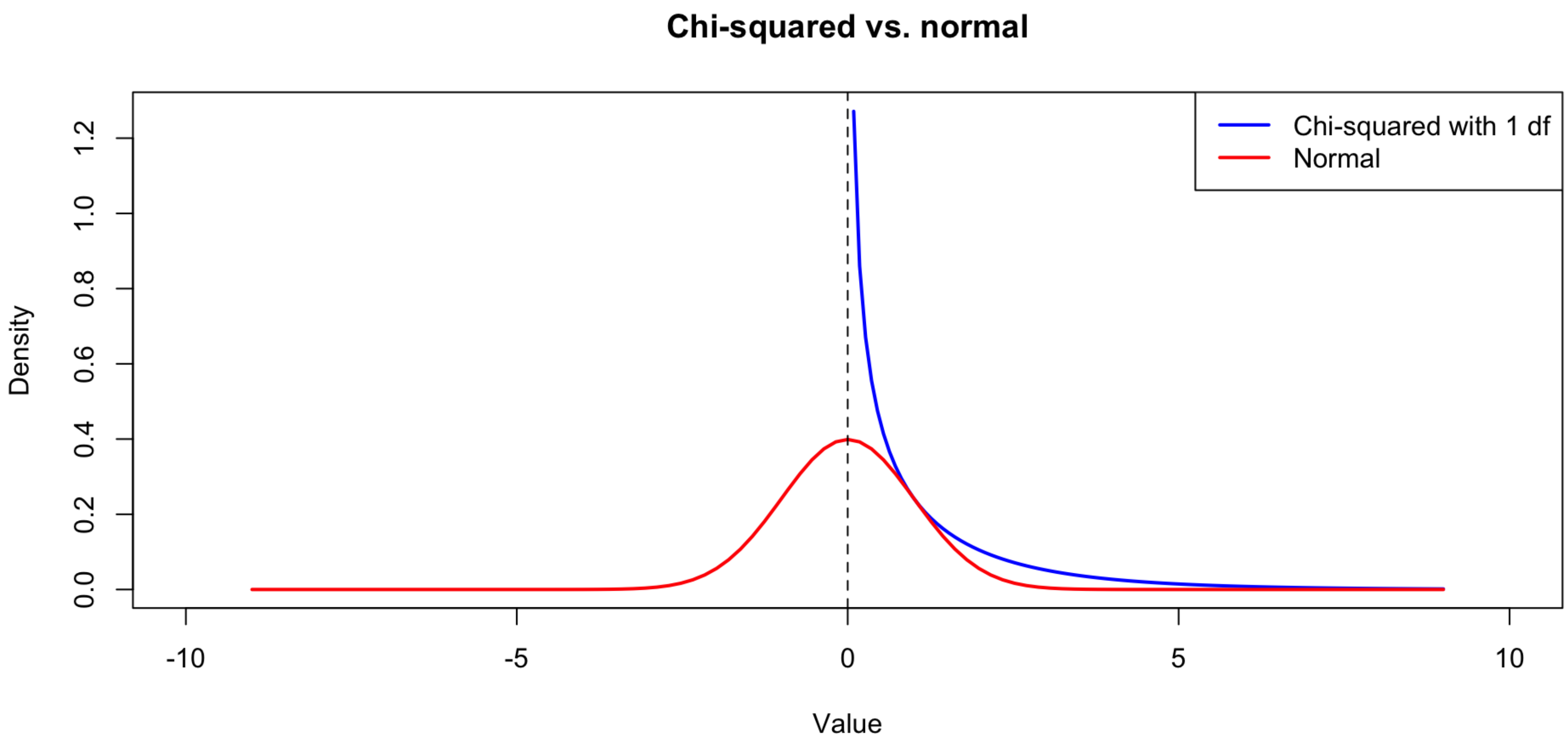
Chi-squared distribution

The chi-squared test

- The chi-squared test is perhaps one of the most prominent examples of non-parametric tests.
- Developed by Karl Pearson in 1900, pronounced “ki” as in “kite”, uses the Greek letter χ.
- Actually derived from the normal distribution: a chi-squared distribution is the sum of squared standard normal deviates – essentially a **folded-over** and **stretched out** normal.

Chi-squared distribution vs normal distribution

▶ Code



How is the chi-squared distribution used in hypothesis testing?

Butterflies data

> A biologist claims that when sampling the Australian Botanical Gardens for butterflies, the ratio of the most dominant colours (red, blue, green, and yellow) is equal. How would you determine if the biologist's claim is true?

Suppose we have the following data on the colours of butterflies after randomly sampling 200 of them:

▶ Code

color	count
red	48
blue	62
green	56
yellow	34

Testing the claim

- If the biologist's claim is true, we would expect the number of butterflies of each colour to be equal.
- If 200 butterflies were sampled, we would expect 50 of each colour, as the expected frequency of each colour is $200 \times 0.25 = 50$.

Therefore:

▶ Code

color	count	expected
red	48	50
blue	62	50
green	56	50
yellow	34	50

Test statistic

The **test statistic** for the chi-squared test is calculated as:

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

where O is the observed frequency and E is the expected frequency.

So for the butterfly data:

```r
chi_squared <- sum((df$count - df$expected)^2 / df$expected)
chi_squared
```

```
[1] 8.8
```

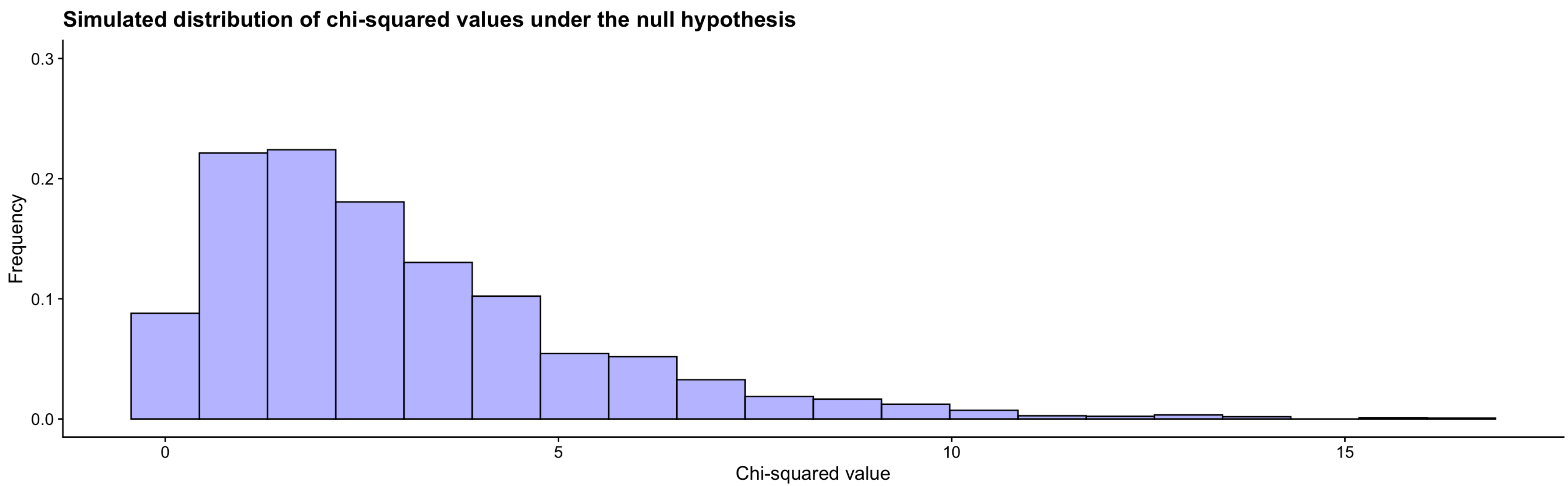
This is the test statistic for one sample. How do we interpret this value?

Simulate the null distribution

Under the null hypothesis, the observed frequencies are equal to the expected frequencies i.e. the biologist's claim is true.

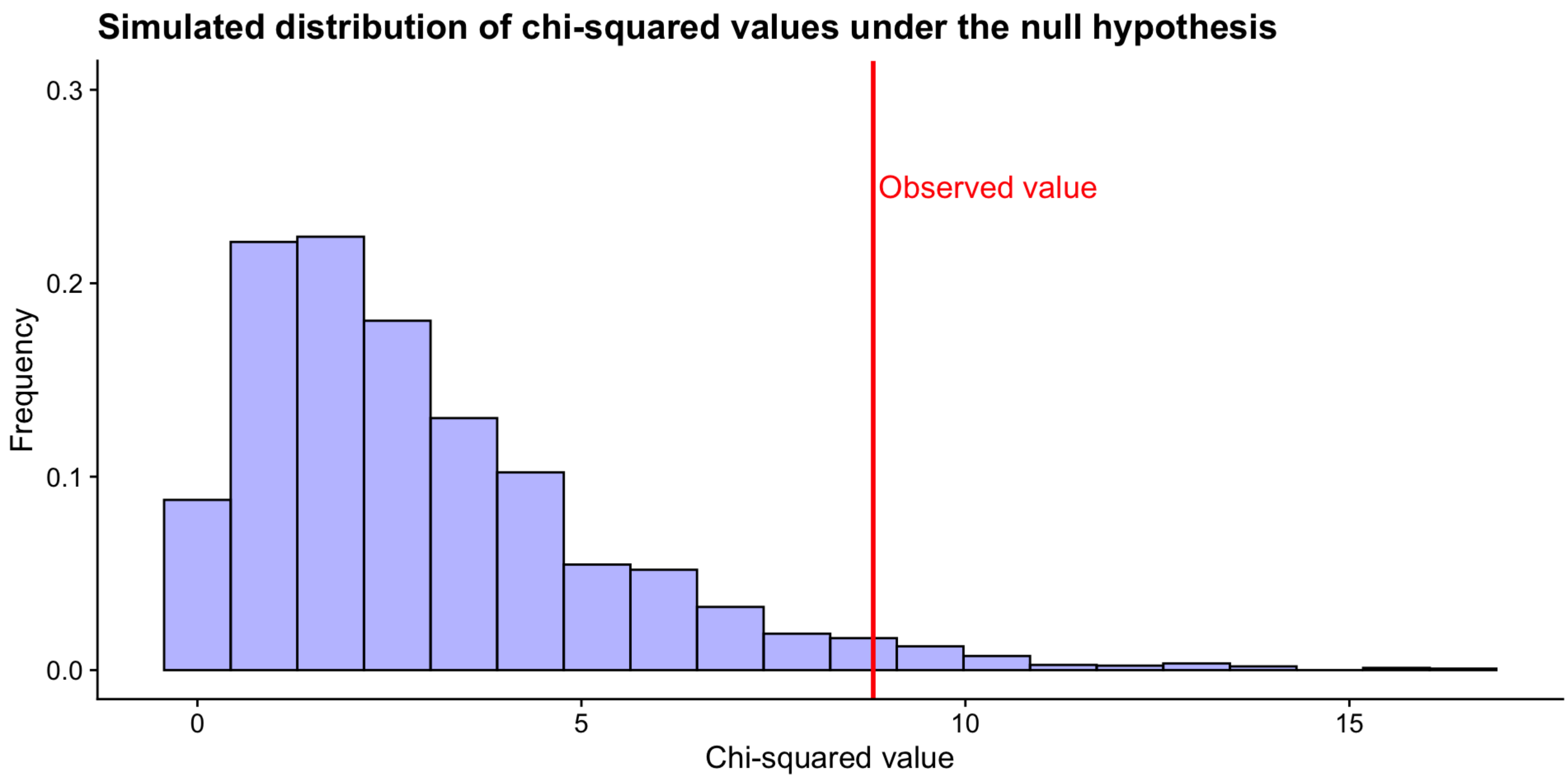
Suppose we repeat the sampling process many times, **assuming the null hypothesis is true**, each time calculating the test statistic. What would the distribution of test statistics look like?

▶ Code



What does our test statistic tell us?

▶ Code




```
mean(test_statistic >= chi_squared)
```

```
[1] 0.034
```

Comparing our test statistic to the simulated distribution, we can see that the 0.03% of the simulated values are greater than our test statistic. **What does this tell us?**

A χ^2 test

A chi-squared distribution allows us to perform the same hypothesis test without the need for simulation.

▶ Code

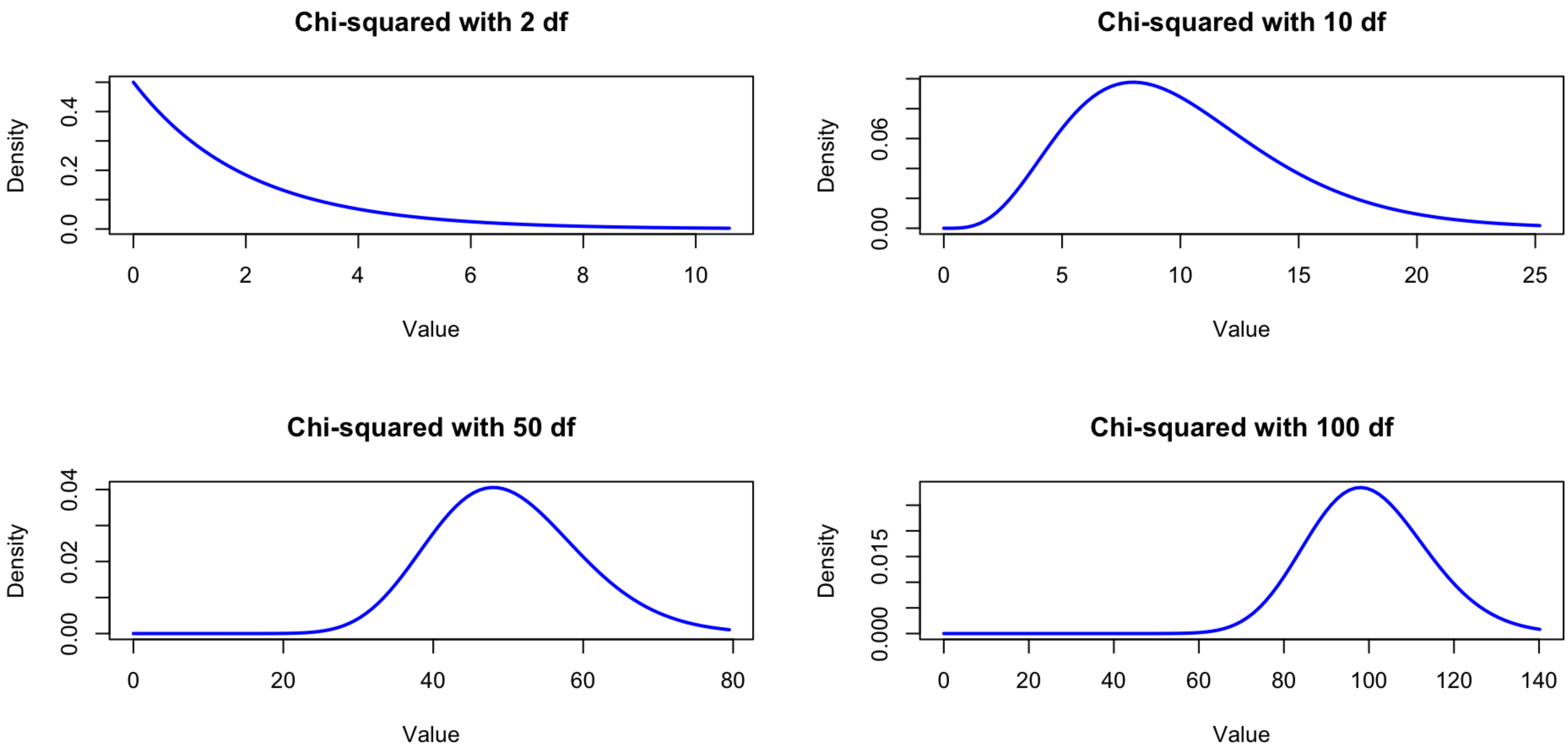
Conclusion?

The results of the simulation suggest that the observed frequencies of butterfly colours are **significantly different** from the expected frequencies, and we can **reject** the biologist's claim.

More on the chi-squared distribution

- The chi-squared distribution is **non-symmetric** and **right-skewed**.
- The shape of the distribution is determined by the **degrees of freedom**, calculated as the number of categories minus 1.
- As the degrees of freedom increase, the chi-squared distribution approaches a normal distribution due to the **central limit theorem**.

▶ Code


Chi-squared with 2 df
Density
0.0
0.2
0.4
0
2
4
6
8
10
Value
Chi-squared with 10 df
Density
0.00
0.06
0
5
10
15
20
25
Value
Chi-squared with 50 df
Density
0.00
0.02
0.04
0
20
40
60
80
Value
Chi-squared with 100 df
Density
0.000
0.015
0
20
40
60
80
100
120
140
Value

The Chi-squared test

Definitions

- **Chi-squared distribution**: a distribution derived from the normal distribution that allows us to determine whether the observed frequencies of a categorical variable differ from the expected frequencies.
- **Contingency table**: a table that displays the frequency of observations for two or more categorical variables.
- **Expected frequency**: the frequency that we would expect to observe if the null hypothesis is true.
- **Observed frequency**: the frequency that we actually observe.
- **Test statistic**: a **measure** of how much the observed frequencies differ from the expected frequencies, standardised by the expected frequencies.

Types of chi-squared tests

- **Goodness-of-fit test**: used to determine whether the observed frequencies of a categorical variable differ from the expected frequencies.
- **Test of independence**: used to determine whether there is a relationship between two or more categorical variables.
- **Test of homogeneity**: used to determine whether the distribution of a categorical variable is the same across different groups.

Assumptions

- The chi-squared test is a **non-parametric** test, so it does not rely on the assumption of normality. However, it does have some assumptions:
  - ➠ **Independence**: the observations are independent.
  - ➠ **Sample size**: the expected frequency of each category is at least 5, and no more than 20% of the expected frequencies are less than 5.

The sample size assumption ensures that the chi-squared distribution is a good approximation of the normal distribution.

Example: Goodness of fit

> A biologist claims that when sampling the Australian Botanical Gardens for butterflies, the ratio of the most dominant colours (red, blue, green, and yellow) is equal. How would you determine if the biologist's claim is true?

Hypothesis

- **Null hypothesis**: the observed proportion of butterfly colours are equal to the expected proportions of 0.25 each.
- **Alternative hypothesis**: the observed proportions are not equal.

$$H_0 : p_1 = p_2 = p_3 = p_4 = 0.25$$

$$H_1 : \text{at least one } p_i \neq 0.25$$

Test statistic and check assumptions (in R)

```r
# chi-squared test for goodness of fit
fit <- chisq.test(df$count, p = rep(0.25, 4))
```

Assumptions

By performing the chi-squared test, we can check the assumptions of the test by looking at the calculated frequences in the output:

```r
fit$observed
```

```
[1] 48 62 56 34
```

Test statistic

```r
fit
```

```

	Chi-squared test for given probabilities

data:  df$count
X-squared = 8.8, df = 3, p-value = 0.03207
```

Conclusion

The results of the chi-squared test suggest that the observed frequencies of butterfly colours are **significantly different** from the expected frequencies ($\chi^2 = 8.8$, $df = 3$, $p < 0.001$). We can reject the null hypothesis and conclude that the biologist's claim is not true.

> **Note**
>
> If you're interested, compare this result to the simulation we performed earlier.

Example: Test of independence

> A study was conducted on a population of deer to see if there is a relationship between their age group (young, adult, old) and their preferred type of vegetation (grass, leaves, bark). Is age group of the deer independent of their vegetation preference?

Hypothesis

- **Null hypothesis**: the age group of the deer is independent of their vegetation preference.
- **Alternative hypothesis**: the age group of the deer is not independent of their vegetation preference.

$$H_0 : \text{Age group is independent of vegetation preference}$$

> No relationship between the two variables

$$H_1 : \text{Age group is not independent of vegetation preference}$$

> There is a relationship between the two variables

Data

Suppose we have the following data on the age group and vegetation preference of 100 deer:

▶ Code

```
       grass leaves bark
young     20     30   10
adult     10     10   20
old       10     10   10
```

Test statistic and check assumptions (in R)

Assumptions are met as we can see the contingency table in the previous slide.

Test statistic

```r
# chi-squared test for independence
fit <- chisq.test(deer_data) # exclude the age group column
fit
```

```

	Pearson's Chi-squared test

data:  deer_data
X-squared = 13.542, df = 4, p-value = 0.008911
```

We reject the null hypothesis since the p-value is less than 0.05.

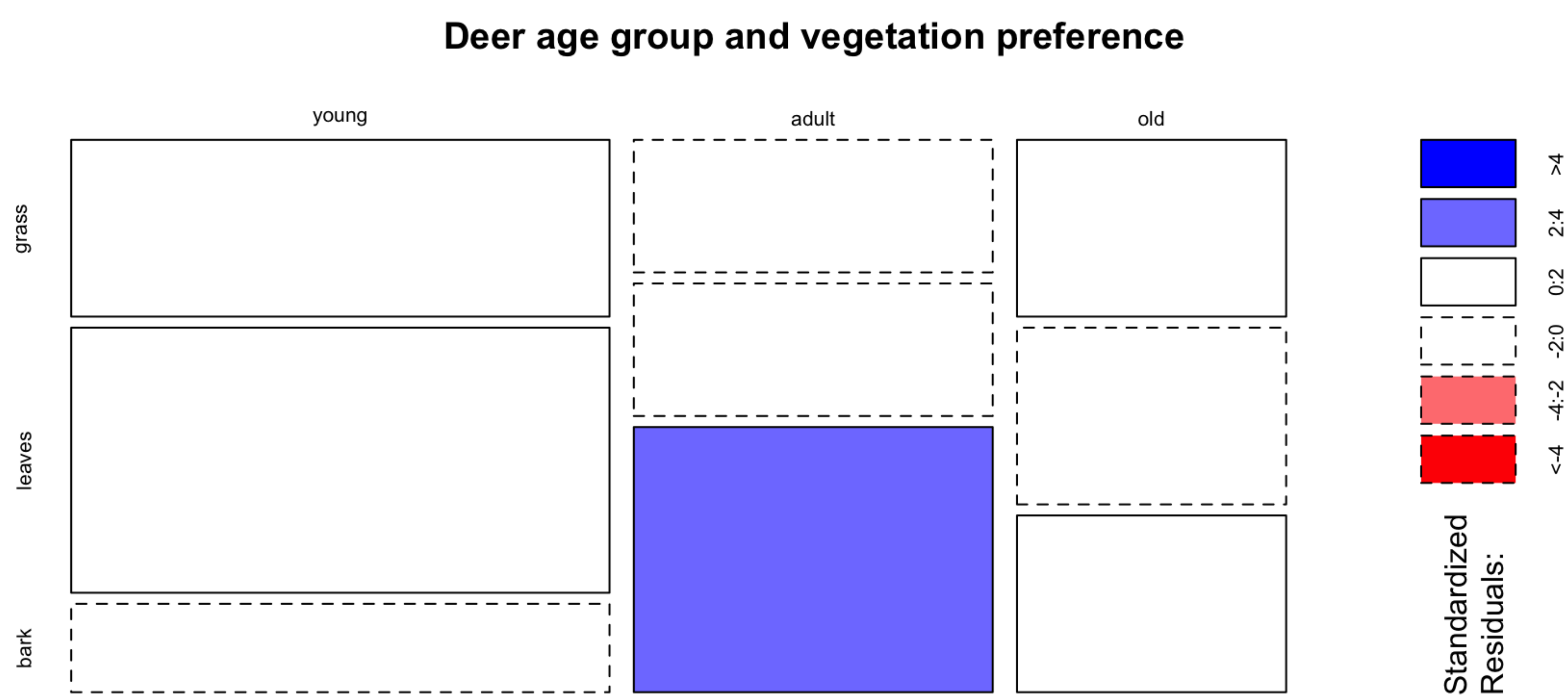
Conclusion

The results of the chi-squared test suggest that the age group of the deer is **not independent** of their vegetation preference ($\chi^2 = 12.4$, $df = 4$, $p < 0.001$). We can reject the null hypothesis and conclude that there is a relationship between the age group of the deer and their vegetation preference.

How do we visualise the differences in a contingency table?

Mosaic plots

▶ Code



Interpretation

- The area of each rectangle is proportional to the number of observations in that category.
- The **shading** of each rectangle indicates the **expected** frequency of observations in that category.
- The **darker** the shading, the **greater** the difference between the observed and expected frequencies.
- Dotted lines indicate **independence** between the two variables.
- Solid lines indicate **dependence** between the two variables.

What about the test of homogeneity?

Test of homogeneity vs. test of independence

- The **test of homogeneity** is similar to the **test of independence**, but is used when we have **two or more groups** and we want to determine whether the distribution of a categorical variable is the same across different groups.
- In general, this means that the null hypothesis is stated differently, and the test statistic is calculated in a slightly different way with different degrees of freedom.
- Homogeneity
  - ➠ H_0: The distribution of the categorical variable is the same across different groups.
  - ➠ H_1: The distribution of the categorical variable is not the same across different groups.
- Independence
  - ➠ H_0: The variables of interest are independent.
  - ➠ H_1: The variables of interest are not independent.

Summary

When to use a chi-squared test?

- If we have **categorical data** and we want to determine whether the observed frequencies differ from the expected frequencies, we can use a **chi-squared test**.
- We're asking whether the observed incidence of outcomes for each level of the predictor variable could have happened by chance.

Thanks!