Topic 7 – Non-parametric tests

ENVX1002 Introduction to Statistical Methods

Floris van Ogtrop
The University of Sydney

Apr 2025

THE UNIVERSITY OF SYDNEY

Evaluation task (Week 8)

- Testing materials in weeks 2 - 6 (No non-parametric test).
- **No GenAI**: we will check.
- Practice before the test using the practice test found on Canvas

Learning Outcomes

At the end of this week, you will be able to:

1. **Explain the rationale** for using non-parametric tests, including when to choose them over parametric tests due to violations of assumptions (e.g., non-normality, ordinal data).
2. Differentiate between:
    - The **Wilcoxon signed-rank tests** (for one-sample or paired data),
    - The **Mann-Whitney U test** (AKA Wilcoxon rank-sum test) for two independent samples, and
    - The **Chi-squared test**, including tests for proportions (goodness-of-fit) and association (independence).
3. **Apply and interpret** Wilcoxon and Mann-Whitney tests in R, for one-sample paired samples or two independent samples
4. **Conduct and interpret** Chi-squared tests in R, including:
    - Chi-squared test of independence to examine association between two categorical variables
    - Chi-squared goodness-of-fit test to compare observed proportions with expected proportions
    - Check assumptions for the Chi-squared test (e.g., expected cell counts) and understand potential limitations and alternatives (e.g., Fisher's Exact Test).
5. **Present and visualise results** from all non-parametric tests using RStudio output, with appropriate interpretation and reporting.

Parametric vs non-parametric methods

Overview

Parametric methods

Depends on the assumption that the data is normally distributed with mean μ and standard deviation σ ,e.g. t-test, ANOVA, linear regression.

Non-parametric methods

Do **not** make *any* assumptions about the distribution of the data.

Uses other properties e.g. ranking of the data, e.g. Wilcoxon signed-rank test, Mann-Whitney U test, Kruskal-Wallis test.

Rank-based tests

General idea

- Rank the data e.g., from smallest to largest.
- Replace the data with their ranks.
- Perform the test on the ranks.

It's *kind of* like a transformation…

For the **Wilcoxon signed-rank test** suppose we have the following data:

sample:	12	10	8	6	4	10	8	6	10

We arrange the data in ascending order (*similar values are given the same colour for illustration*):

ordered:	4	6	6	8	8	10	10	10	12

Then, we rank the data:

ordered ranks:	1	2	3	4	5	6	7	8	9

Finally, ranks that are *tied* are given the average rank:

final rank:	1	2.5	2.5	4.5	4.5	7	7	7	9

These ranks are then used to perform the test, instead of the original data.

Use case

Two-sample t-test

Consider two sets of **identical** data that compares between a group **A** and **B**, where one contains an outlier.

Data:

▶ Code

A	B
1	7
2	8
3	9
4	10
5	11
6	12
7	13
8	14
9	15
10	16

Data *with* outlier:

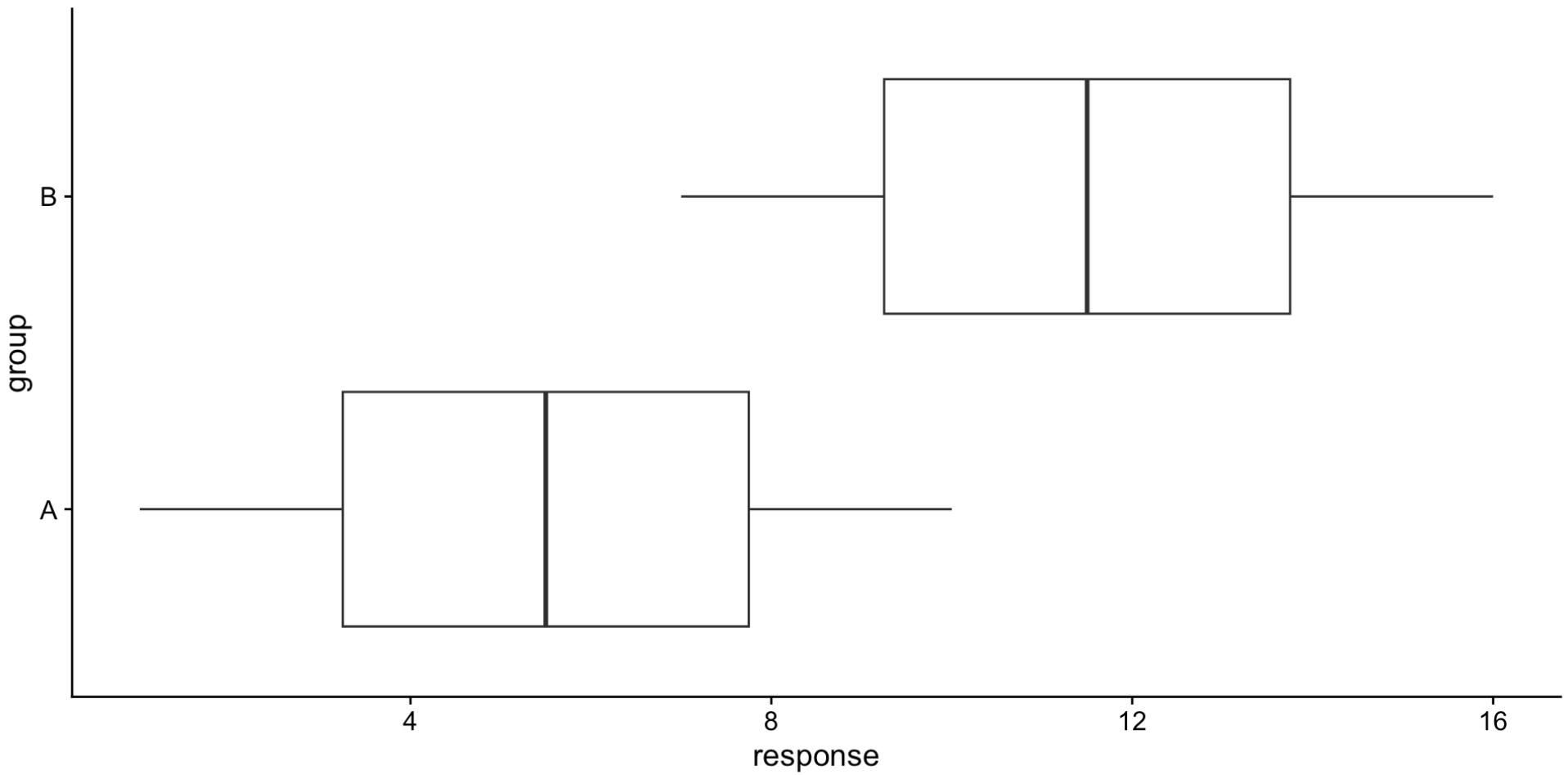
▶ Code

A	B
1	7
2	8
3	9
4	10
5	11
6	12
7	13
8	14
9	15
10	200

Should there be a difference?

Without the outlier, the data would have been normally distributed.

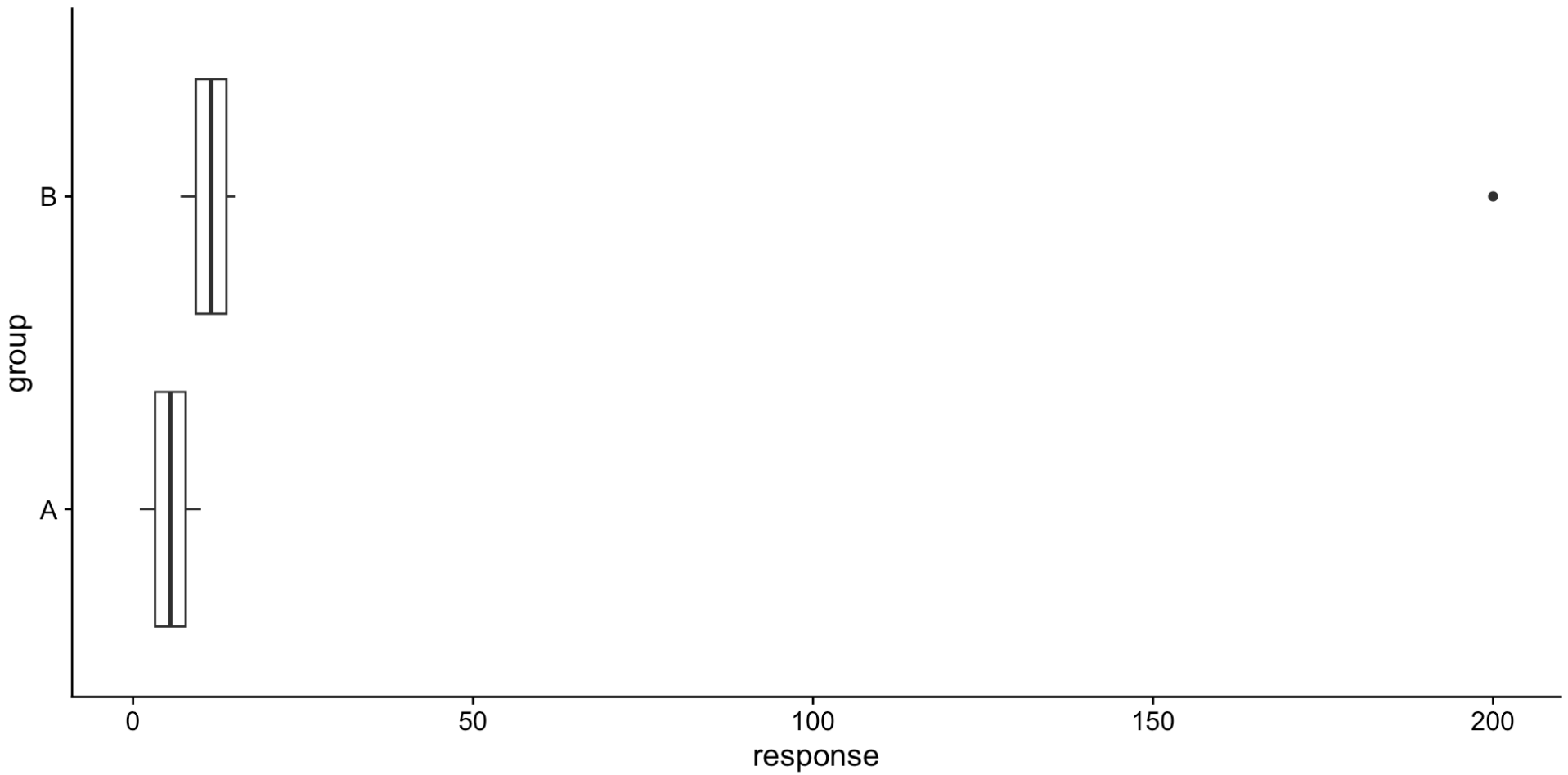
▶ Code



Outlier

The same data, but with a single outlier in group **B**:

▶ Code



Analysis

If we perform t-tests on both data sets, we get the following results:

▶ Code

```
    Welch Two Sample t-test

data:  response by group
t = -4.4313, df = 18, p-value = 0.0003224
alternative hypothesis: true difference in means
between group A and group B is not equal to 0
95 percent confidence interval:
 -8.844662 -3.155338
sample estimates:
mean in group A mean in group B
            5.5            11.5
```

Results indicate that there is a statistically significant difference between the two groups (t_{18} = -4.4, $p < 0.05$).

▶ Code

```
    Welch Two Sample t-test

data:  response by group
t = -1.2882, df = 9.0461, p-value = 0.2297
alternative hypothesis: true difference in means
between group A and group B is not equal to 0
95 percent confidence interval:
 -67.21615  18.41615
sample estimates:
mean in group A mean in group B
            5.5            29.9
```

Results indicate that the two groups are **not** significantly different (t_{18} = -2.1, $p = 0.23$).

The real difference between the two groups is *obscured* by the outlier. **Type II error** (false negative)?

Non-parametric alternatives

When to use

1. If assumptions are met (normality, homogeneity of variance), use parametric tests as they are more powerful and efficient than non-parametric tests.
2. If the normality assumption is violated, transform the data and check for normality again (*optional*).
3. **Non-parametric tests are a good way to deal with circumstances in which parametric tests perform "poorly".**

What to use

▶ Code

Parametric tests	Non-parametric counterpart
One-sample t-test	Wilcoxon signed-rank test
Two-sample t-test	Mann-Whitney U test
ANOVA	Kruskal-Wallis test
Pearson's correlation	Spearman's rank correlation

All of the non-parametric techniques above convert the data into **ranks** before performing the test.

> **Note**
>
> We will focus on the **Wilcoxon signed-rank test** and the **Mann-Whitney U test**.

Wilcoxon signed-rank test

Altenrative to the one-sample t-test and the paired t-test.

Overview

The Wilcoxon signed-rank test is a non-parametric test used to compare two related samples, matched pairs, or repeated measures on a single sample.

Is an alternative to:

- One-sample t-test
- Paired t-test

Assumptions

- Data comes from the same population
- Data are randomly and independently sampled

Basically, used in same situations as the one-sample or paired t-test, but when the data is not normally distributed but **still symmetric**.

Calculating ranks

If comparing two groups, the ranks are calculated as follows:

1. Calculate the difference D between the two groups.
2. Rank the absolute values of the differences in ascending order.
3. Assign the sign of the difference to the rank.
4. Sum the ranks for each group – **zero differences are ignored**.

> **Note**
>
> See Slide 5 to recall how ranks are calculated, but we will show another example in the next slide.

Example: Paired data

Weight gain

We measured weight gain in chickens before and after a diet.

▶ Code

chicken	weight	weight_after
1	2.5	4.0
2	3.5	5.0
3	3.5	5.0
4	3.4	4.6

Is there a significant increase in weight gain after the diet?

Rank values

▶ Code

chicken	weight	weight_after	D	Sign	rank	Signed rank
1	2.5	4.0	1.5	+	3	3
2	3.5	5.0	1.5	+	3	3
3	3.5	5.0	1.5	+	3	3
4	3.4	4.6	1.2	+	1	1

Note

The order of the ranks is based on the **absolute** values of the differences; the signs are assigned afterward.

Hypothesis

> Is there a significant increase in weight gain after the diet?

$$H_0 : \mu_{\text{before}} = \mu_{\text{after}}$$

$$H_1 : \mu_{\text{before}} < \mu_{\text{after}}$$

In words:

- H_0: There is no difference in weight gain before and after the diet.
- H_1: There is an increase in weight gain after the diet.

Alternatively, since the data is paired, we may also consider hypotheses based on the differences between the two groups:

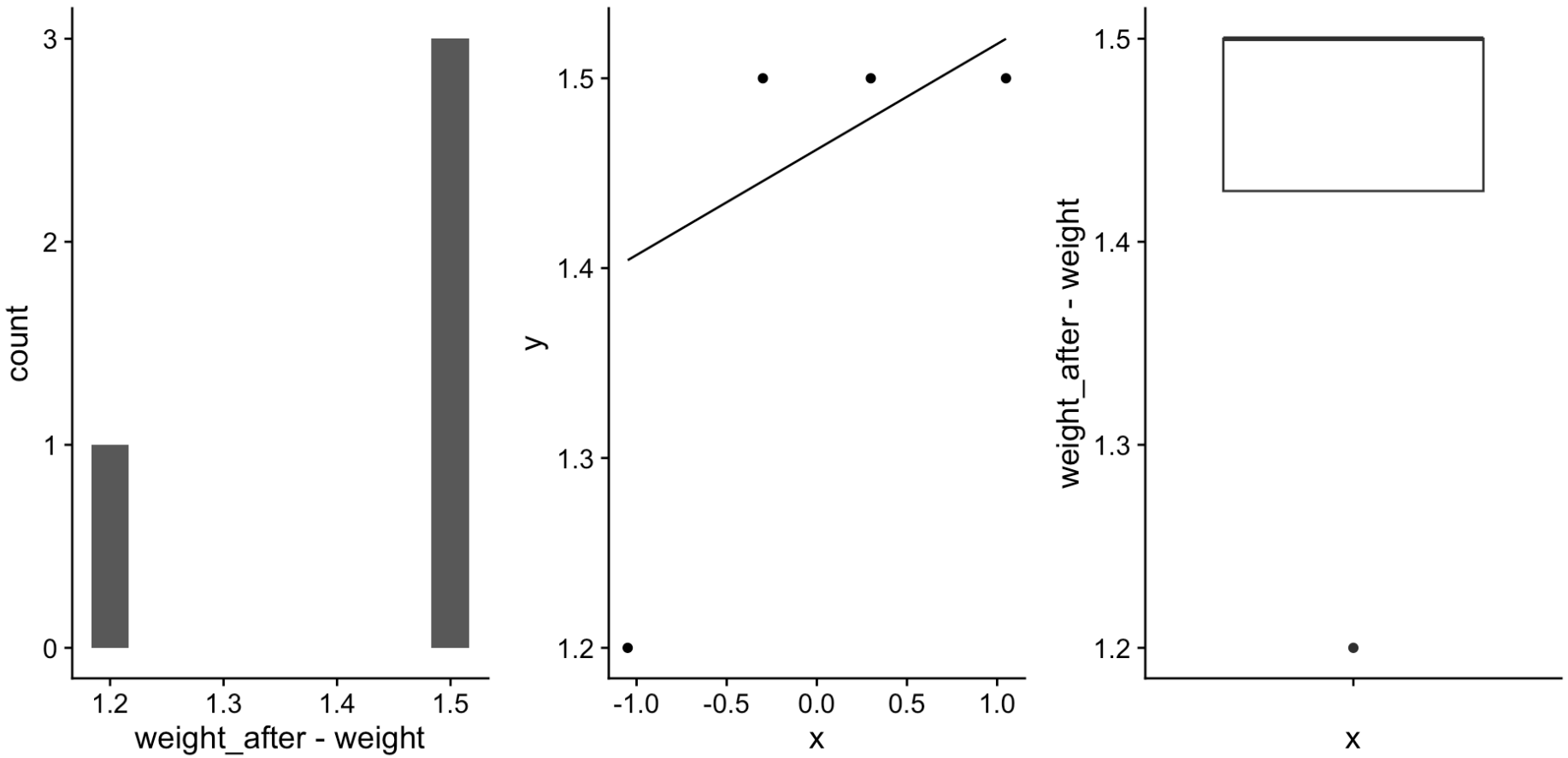
$$H_0 : \mu_D = 0$$

$$H_1 : \mu_D > 0$$

where D is the difference between the two groups.

Assumptions

▶ Code



With so few data points, we may want to use a formal test to check for normality.

Assumptions

```r
shapiro.test(df$weight_after - df$weight)
```

```
    Shapiro-Wilk normality test

data:  df$weight_after - df$weight
W = 0.62978, p-value = 0.001241
```

Results indicate that the data significantly deviates from normality ($W = 0.63$, $p < 0.05$). We will use the Wilcoxon signed-rank test.

Performing the test

In R

```r
wilcox.test(x = df$weight_after, y = df$weight,  # data must be x - y
  alternative = "greater",  # because we are testing for an increase
  paired = TRUE)            # because the data is paired
```

```
    Wilcoxon signed rank test with continuity correction

data:  df$weight_after and df$weight
V = 10, p-value = 0.04449
alternative hypothesis: true location shift is greater than 0
```

where `V` is the sum of the signed ranks.

The results indicate that there is a **significant increase in weight gain** after the diet ($V = 10$, $p < 0.05$).

Example: One-sample data

Beetle consumption in lizards

Researchers investigated differences in beetle consumption between two size classes of eastern horned lizard (*Phrynosoma douglassi brevirostre*)

- **Larger class**: adult females.
- **Smaller class**: adult males, yearling females.

Focusing on just the smaller size class (for now) – it was hypothesised that this size class would eat a minimum of 100 beetles per day.

Hypothesis

Does the average smaller size class lizard eat about 100 beetles per day?

$$H_0 : \mu = 100$$

$$H_1 : \mu \neq 100$$

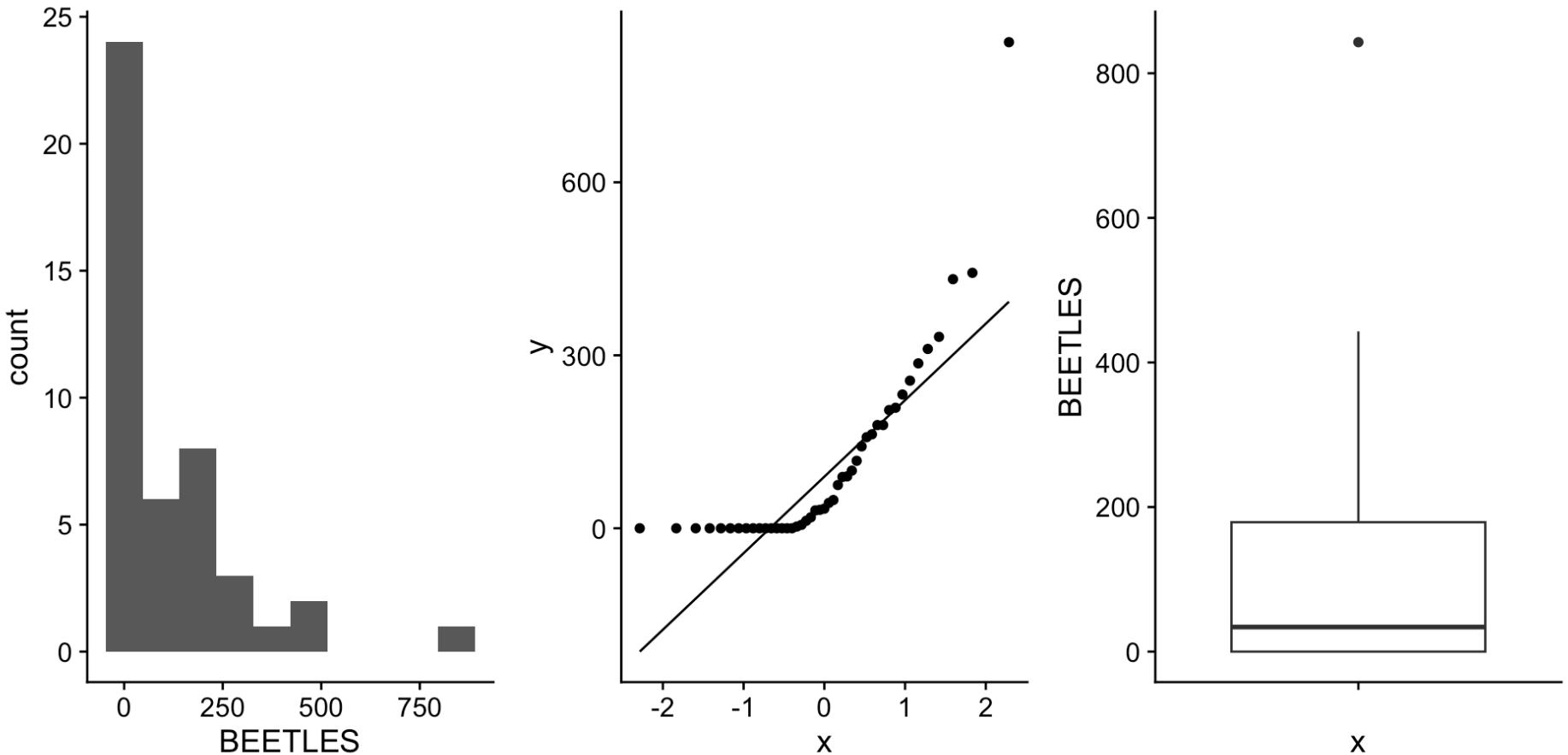
Dataset Download

▶ Code

```
Rows: 45
Columns: 2
$ SIZE    <chr> "small", "small", "small", "small", "small", "small", "small",…
$ BEETLES <dbl> 256, 209, 0, 0, 0, 44, 49, 117, 6, 0, 0, 75, 34, 13, 0, 90, 0,…
```

First, check assumptions

▶ Code



Is it normally distributed?

Run the test

The Wilcoxon signed-rank test for one sample can be performed as follows:

```
beetle %>%
  filter(SIZE == "small") %>%  # filter only the smaller size class
  pull(BEETLES) %>%            # convert to a vector using pull()
  wilcox.test(mu = 100)        # wilcox.text(x, mu)
```

```
    Wilcoxon signed rank test with continuity correction

data:  .
V = 92, p-value = 0.09755
alternative hypothesis: true location is not equal to 100
```

Results indicate that the average number of beetles consumed by the smaller size class lizard is **not significantly different from 100** (V = 92, p = 0.1).

> **Important**
>
> We are unable to make a conclusion about effect size from non-parametric tests as the information is lost when the data is transformed into ranks.

Mann-Whitney U test

Alternative to the **two-sample *t*-test**.

*Also called the Mann–Whitney–Wilcoxon (MWW/MWU), **Wilcoxon rank-sum test (what R calls it)**, or Wilcoxon–Mann–Whitney test.*

About

- A **non-parametric test** used to compare two independent samples similar to the two-sample t-test.
- Like the Wilcoxon signed-rank test, it uses **ranks** to perform the test and does not assume normality.
- It is also more *relaxed* in that it does not assume symmetry in the distribution of the data – instead, it assumes that the two groups have the same shape/distribution.

Example: Back to the lizards

Beetle consumption in lizards

Researchers investigated differences in beetle consumption between two size classes of eastern horned lizard (*Phrynosoma douglassi brevirostre*)

- **Larger class**: adult females.
- **Smaller class**: adult males, yearling females.

We will now compare the number of beetles consumed by the **larger** and **smaller** size classes of lizards.

Hypotheses

> Are the number of beetles consumed by the larger and smaller size classes of lizards different?

Loosely speaking, because we are not assuming symmetry, the most appropriate summary statistic to use when comparing the two groups is the **median**.

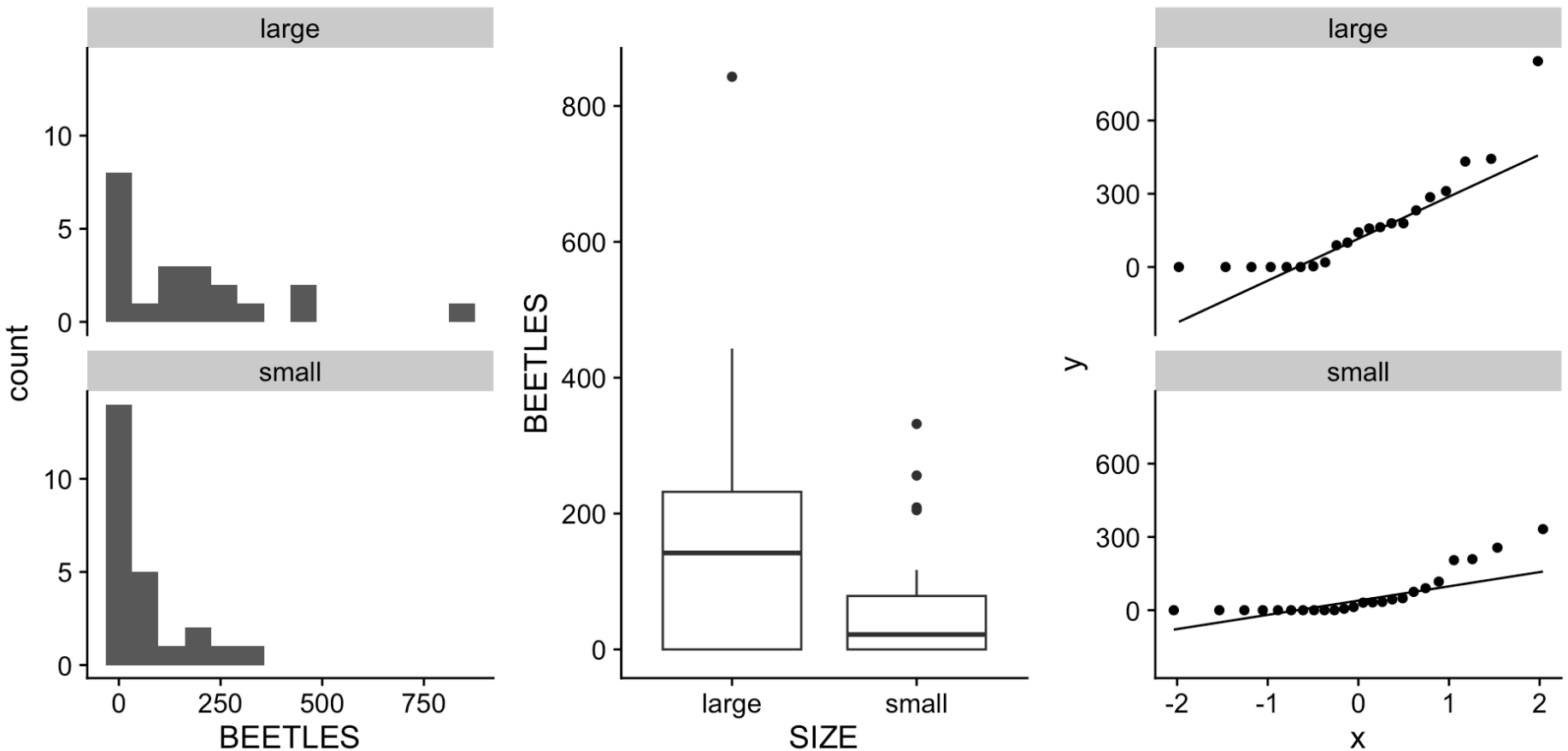
$$H_0 : median_{\text{larger}} = median_{\text{smaller}}$$

$$H_1 : median_{\text{larger}} \neq median_{\text{smaller}}$$

More accurately, we are testing for a difference in the *distribution* of the two groups.

Assumptions

▶ Code



Data does not meet the normality assumption.

Test statistic

The same function `wilcox.test()` can be used to perform the Mann-Whitney U test.

```
1 wilcox.test(BEETLES ~ SIZE, data = beetle)
```

```
    Wilcoxon rank sum test with continuity correction

data:  BEETLES by SIZE
W = 329, p-value = 0.07494
alternative hypothesis: true location shift is not equal to 0
```

- `W` is the sum of the ranks of the *smaller* group.
- The "true location shift" is the median of the larger group minus the median of the smaller group.

The results indicate that the number of beetles consumed by the larger and smaller size classes of lizards is **not significantly different** (W = 329, p = 0.07).

What about transformations?

Transform, or non-parametric?

- As usual, there is ongoing debate on whether to transform the data or use non-parametric tests, but the general consensus is to always prefer parametric tests and transformations when assumptions are met using those techniques.
    - ➠ e.g. Parametric analysis of transformed data is more powerful than non-parametric analysis
- Some argue that non-parametric tests **must** be decided during experimental design and not when the data fails to meet the normality assumption: **as the decision to rank data has implications on the interpretation of the results.**
    - ➠ e.g. Graphpad Advice: Don’t automate the decision to use a nonparametric test
- The conventional wisdom is to **transform the data** and check for normality if the assumption is not met. If the data is *still* not normal, then use non-parametric tests (after considering the implications on interpretation).
    - ➠ Or, use bootstrapping (next week!).

Summary

- **Wilcoxon signed-rank test**: alternative to the one-sample t-test and paired t-test.
- **Mann-Whitney U test**: alternative to the two-sample t-test.
- **Advantages**: **Robust** to outliers, skewness, and non-normality.
- **Drawbacks**: Less powerful than parametric tests when assumptions are met, provide no insight into the size of the effect.

Questions?

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:

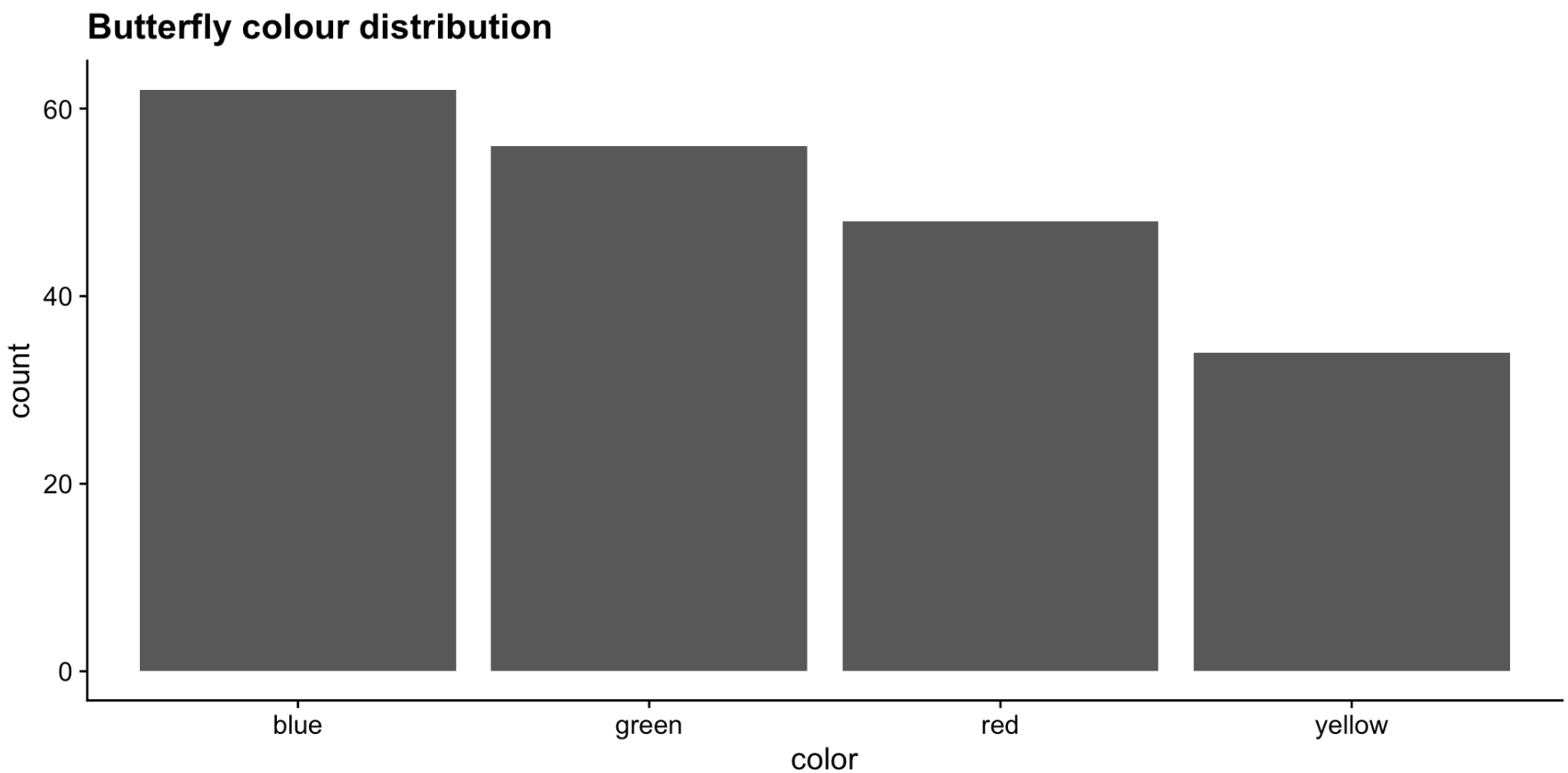
> A biologist claims that when sampling the Australian Botanical Gardens for butterflies, the ratio of the most dominant colours (red, blue, green, and yellow) is equal. How would you determine if the biologist's claim is true?

> 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).

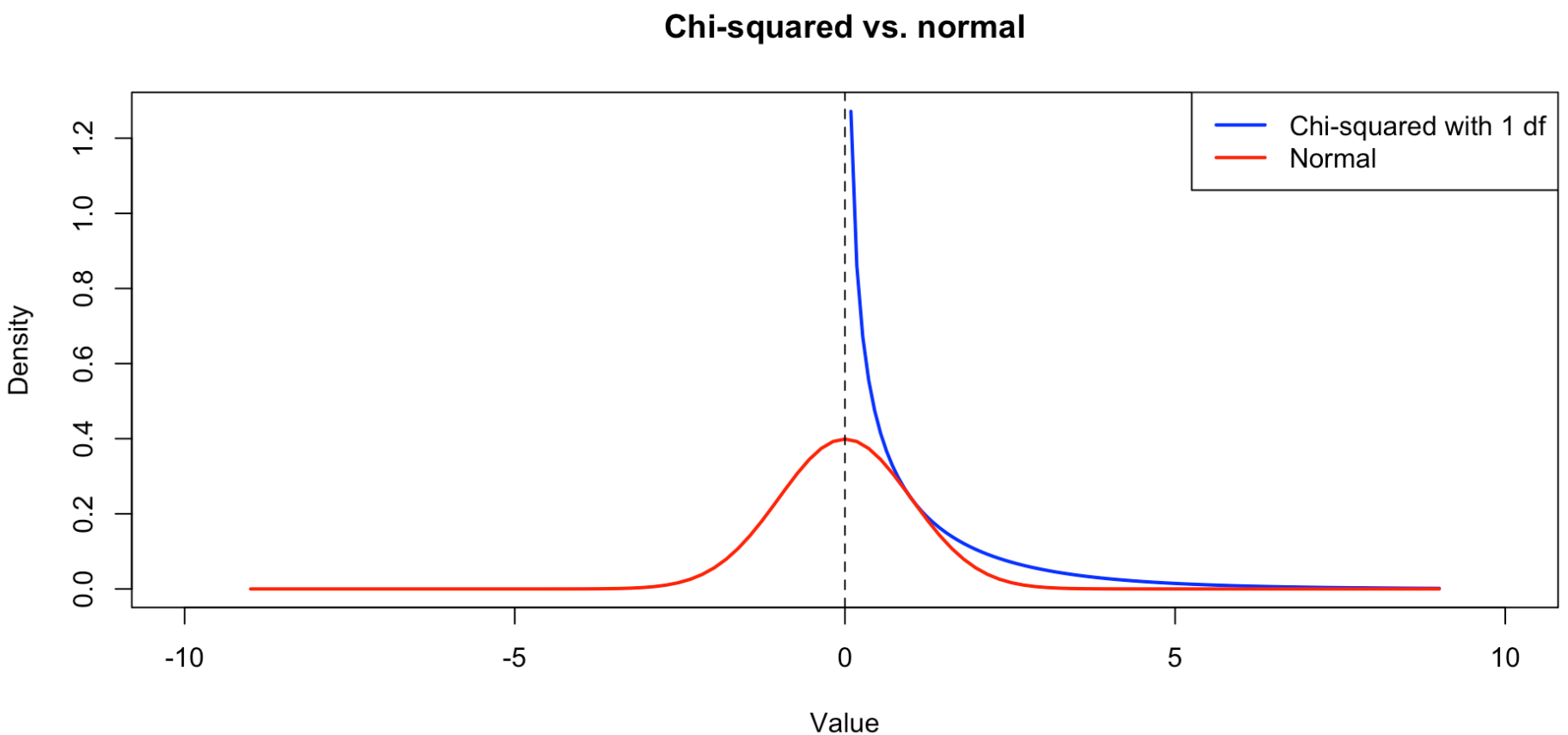
Chi-squared distribution

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:

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

```
[1] 8.8
```

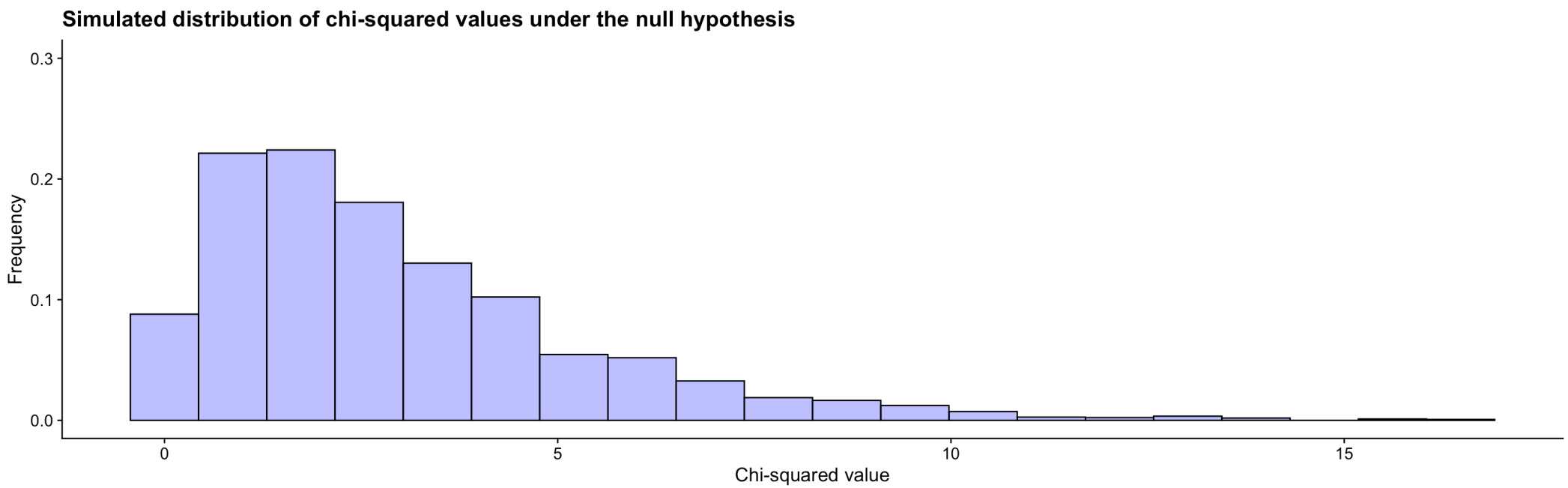
This is the test statistic for one sample. How do we interpret this value?

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.

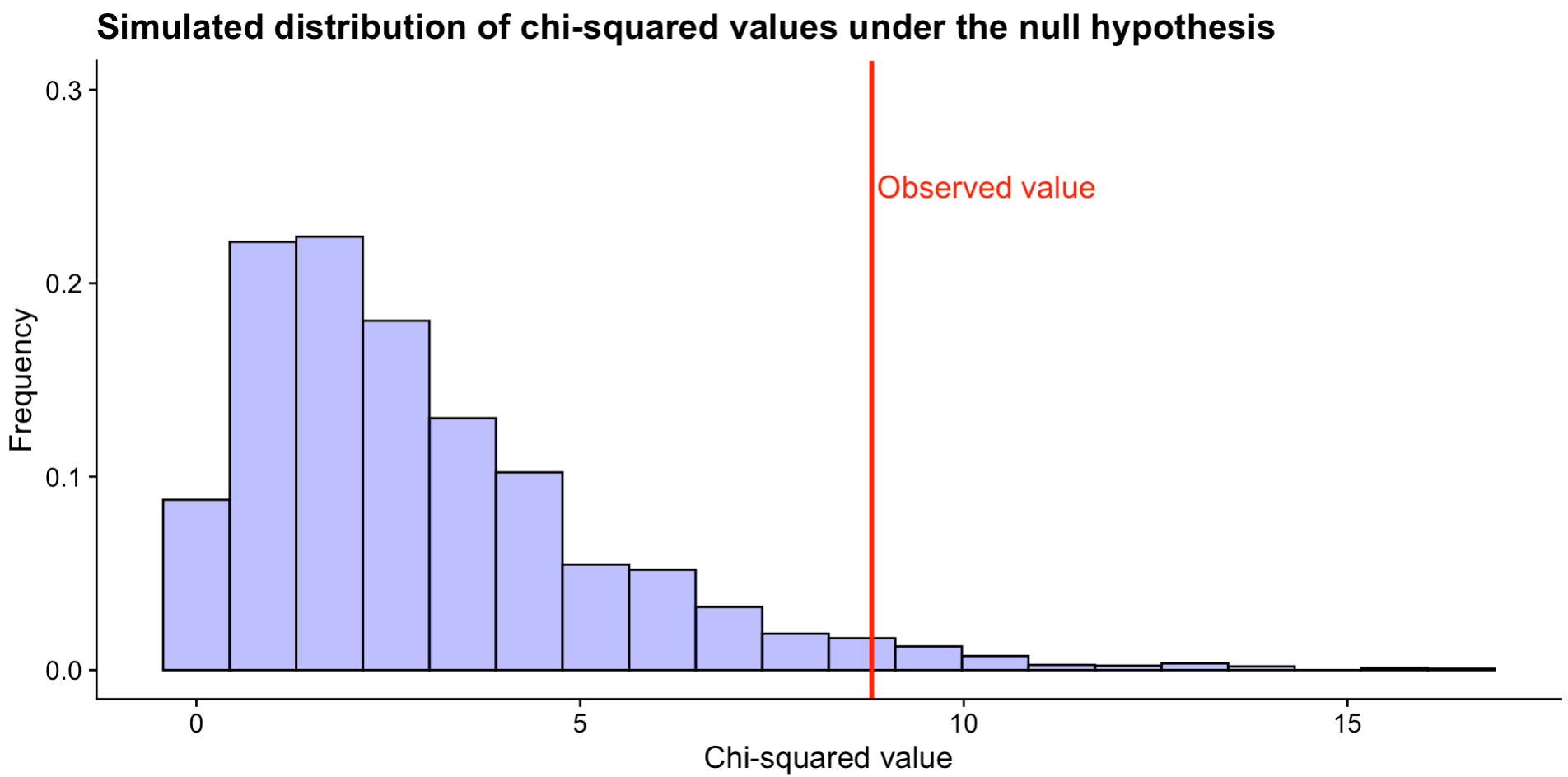
Suppose we repeat the sampling process many times, **assuming the null hypothesis is true**, each time calculating the test statistic. What would the distribution of test statistics look like?

▶ Code



What does our test statistic tell us?

▶ Code



▶ Code

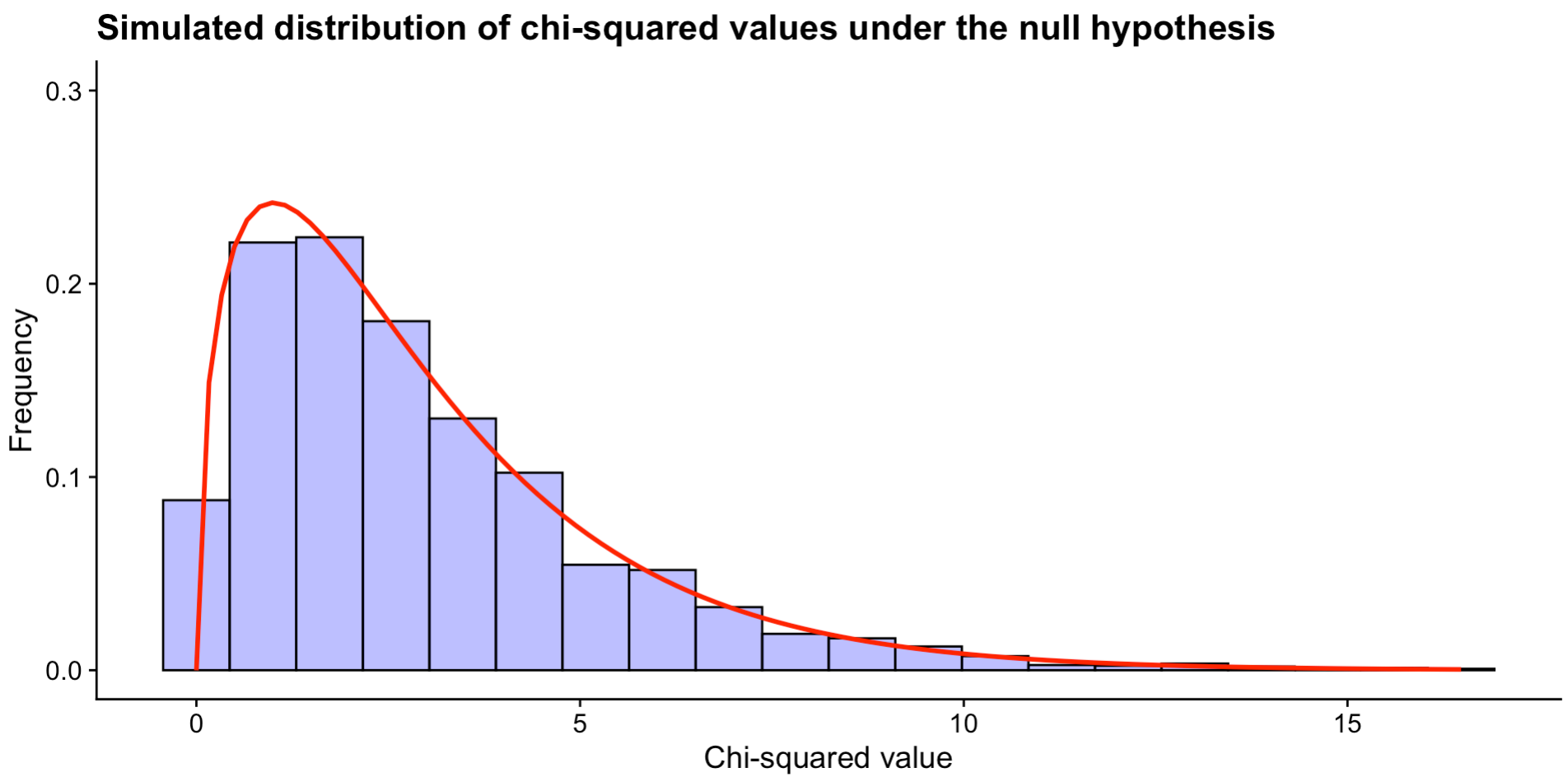
```
[1] 0.034
```

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



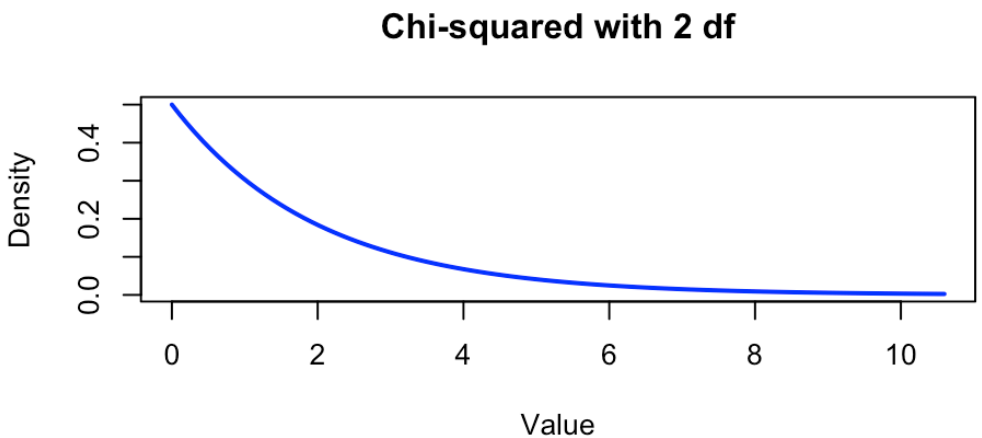
Conclusion?

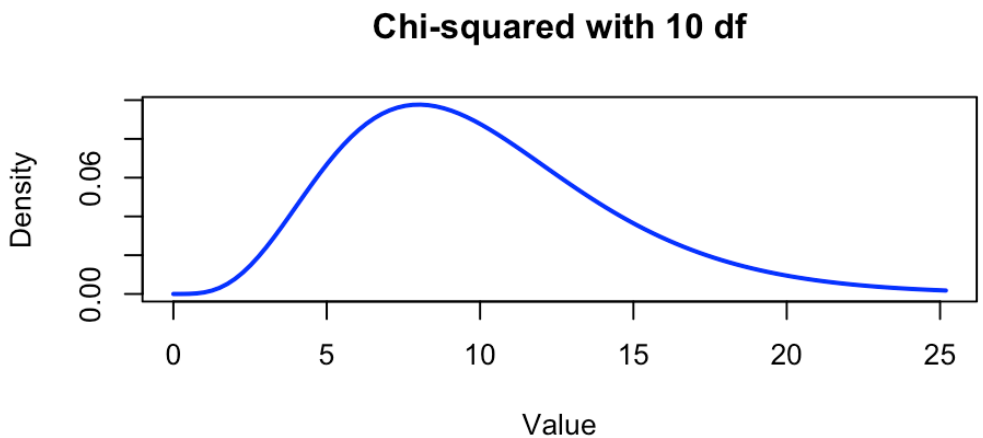
The results of the simulation suggest that the observed frequencies of butterfly colours are **significantly different** from the expected frequencies, and we can **reject** the biologist's claim.

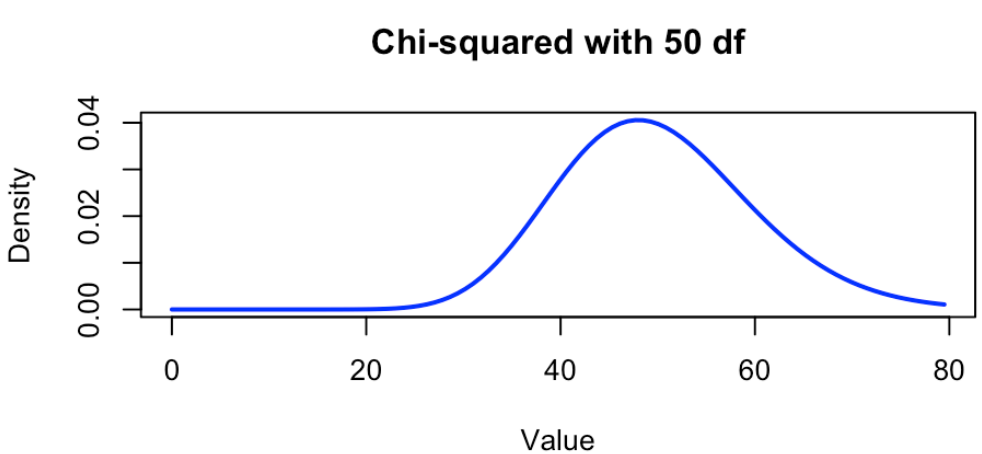
More on the chi-squared distribution

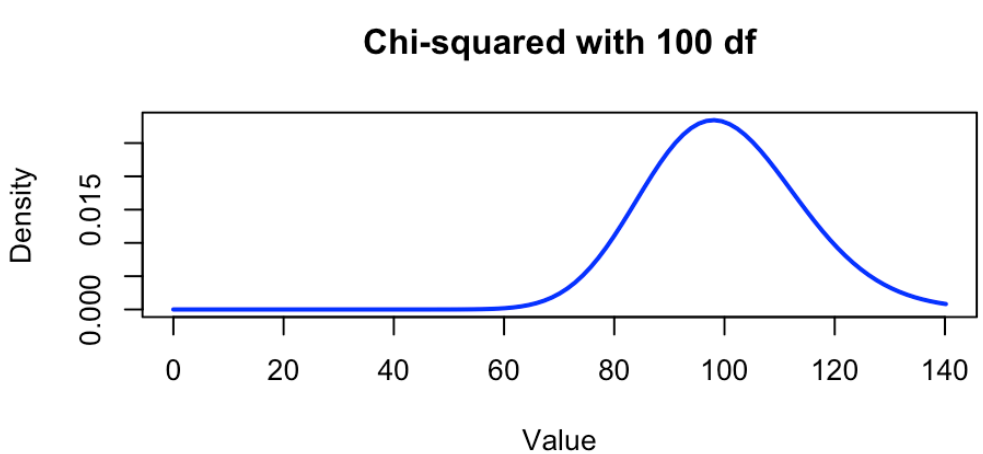
- The chi-squared distribution is **non-symmetric** and **right-skewed**.
- The shape of the distribution is determined by the **degrees of freedom**, calculated as the number of categories minus 1.
- As the degrees of freedom increase, the chi-squared distribution approaches a normal distribution due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

▶ Code

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

Test statistic

▶ Code

```

    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.

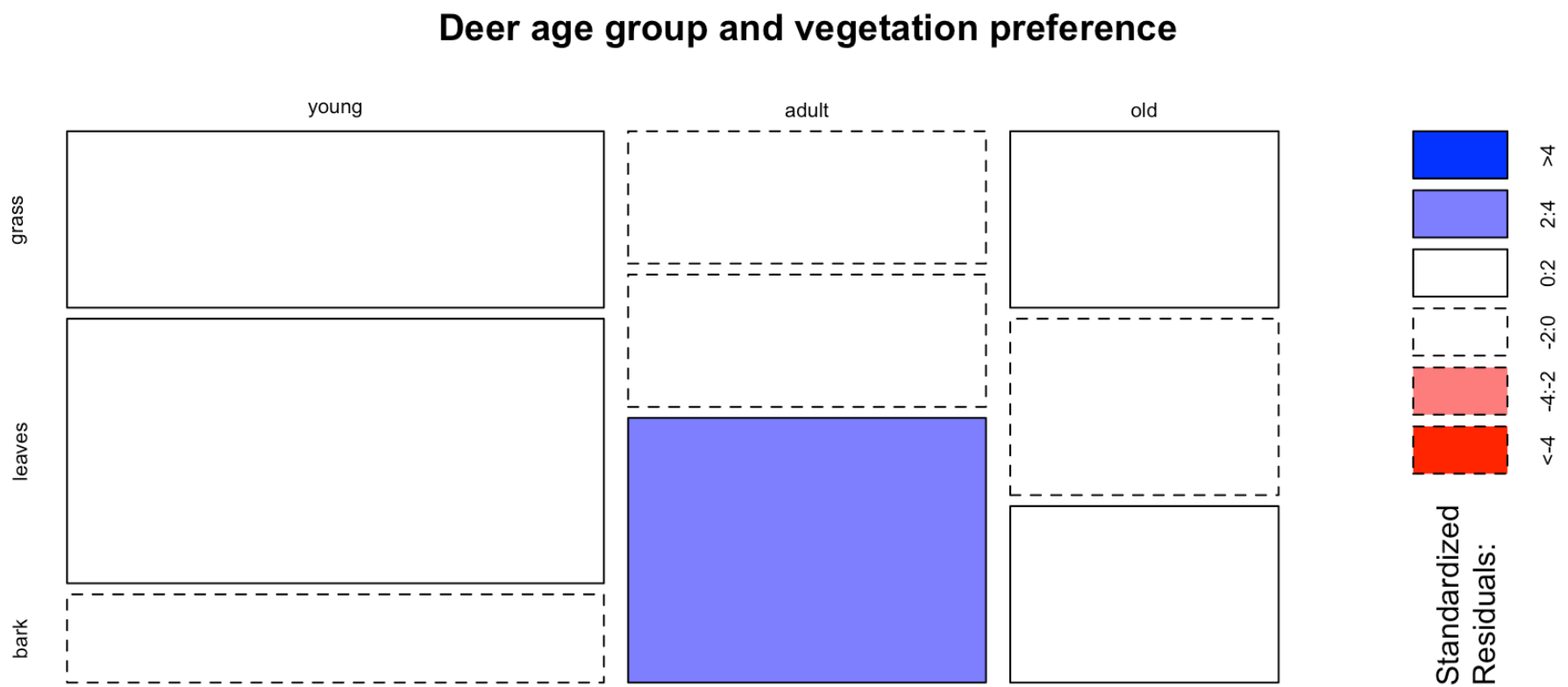
Conclusion

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

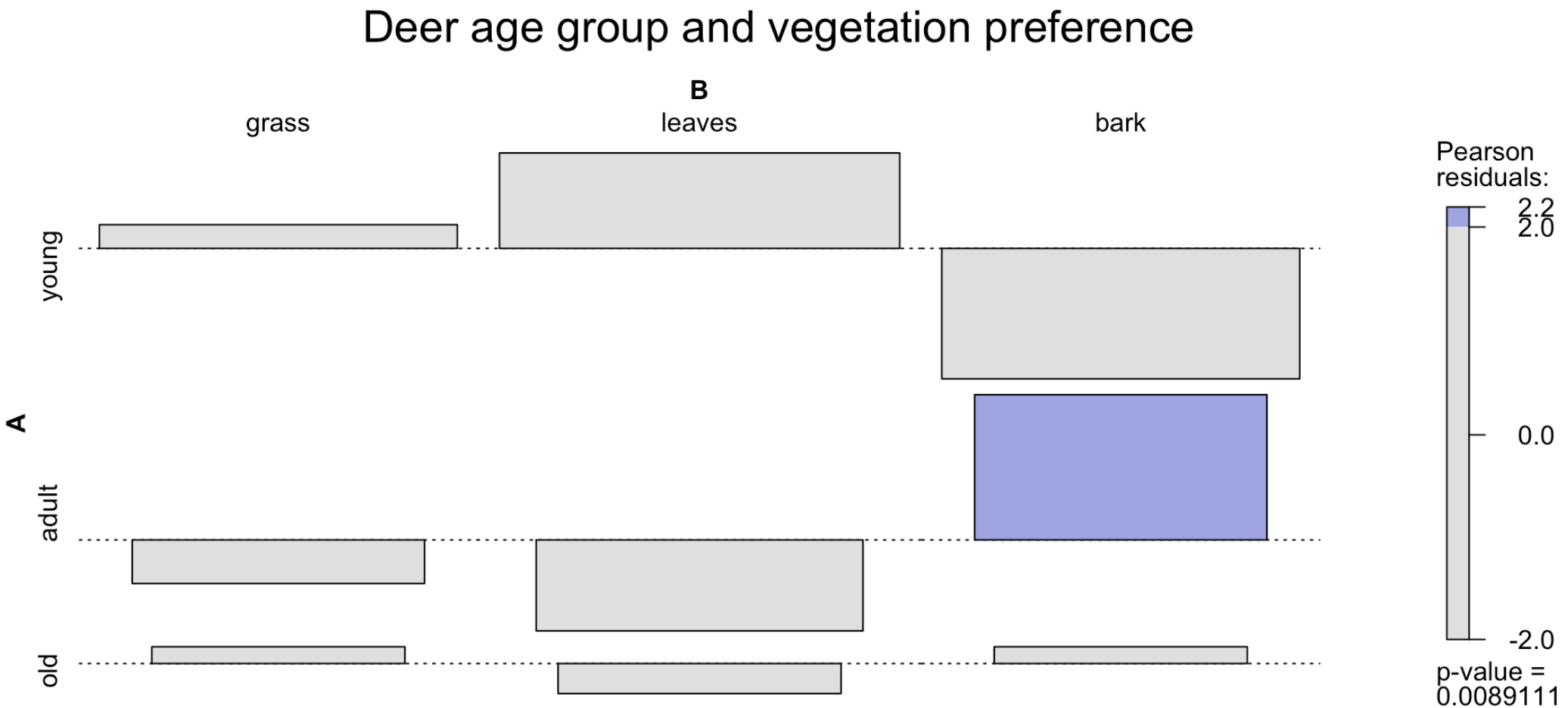


Interpretation

- 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.

Association plots

▶ Code



Interpretation

- Size of cells indicate the number of observations in that category.
- The shadings are made based on the residuals of the chi-squared test (see legend), highlighting which cells contribute most to the chi-squared statistic.
- Colour of the shadings indicate whether they are more or less frequent than expected (again, see legend).

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.

Differences are subtle

- In the test of independence, observational units are collected at random from **a single population** and two (or more) categorical variables are observed for each unit.
- For the deer example, the experimental design would involve randomly sampling deer and recording their age group and vegetation preference.

> Is age group independent of vegetation preference?

- In the test of homogeneity, the data are collected by randomly sampling from **two or more subgroups**, and the same categorical variable is observed for each unit.
- For the deer example, the experimental design would have to be modified to sample the vegetation preference of deer from young, adult, and old populations.

> Is the distribution of vegetation preference the same if we compare young, adult, and old deer?

Summary

When to use a chi-squared test?

- The chi-squared test is not an “alternative” to a parametric test, but is **better suited** for certain types of data and requires delibarate experimental design that collects data in a certain way.
- If we have **categorical data** and we want to determine whether the observed frequencies differ from the expected frequencies, then we can use a **chi-squared test**.

Thanks!