Topic 6 – Two-sample *t*-tests – Part II

ENVX1002 Introduction to Statistical Methods

Januar Harianto
The University of Sydney

Dec 2024

THE UNIVERSITY OF SYDNEY

CRICOS 00026A TEQSA PRV12057

When assumptions of the *t*-test are violated

Recap: Assumptions of the two-sample *t*-test

With independent samples:

- **Normality**: the data are normally distributed
- **Homogeneity of variance** (equal variances): the variances of the two groups are equal

With paired samples:

- **Normality**: the differences between the paired samples are normally distributed
- Equal variances is implied

If we analyse the data anyway…

The *t*-test:

- may provide incorrect results as **mean and variance calculations depend on normally distributed data**.
- may be **less powerful** (i.e., less likely to detect a true difference).
- may be **biased** (i.e., systematically over- or under-estimating the true difference).

Don’t throw the data away…

What can we do?

The *t*-test is quite robust to violations of normality, especially when the sample size is large. However, the assumption of equal variances is more critical – we cannot simply depend on large sample sizes to "fix" the problem.

Options include:

- **Transform** the data to normalise the data and/or scale the variance
- Use a **Welch's *t*-test** or a **Welch's ANOVA** (limited cases)
- Use a **non-parametric test**, such as the **Mann-Whitney U test** or **Wilcoxon signed-rank test** (paired samples) – however, these tests have *less power* than the *t*-test i.e. less likely to detect a true difference.

Ants - a foraging biomass study

{fig-align="left"}

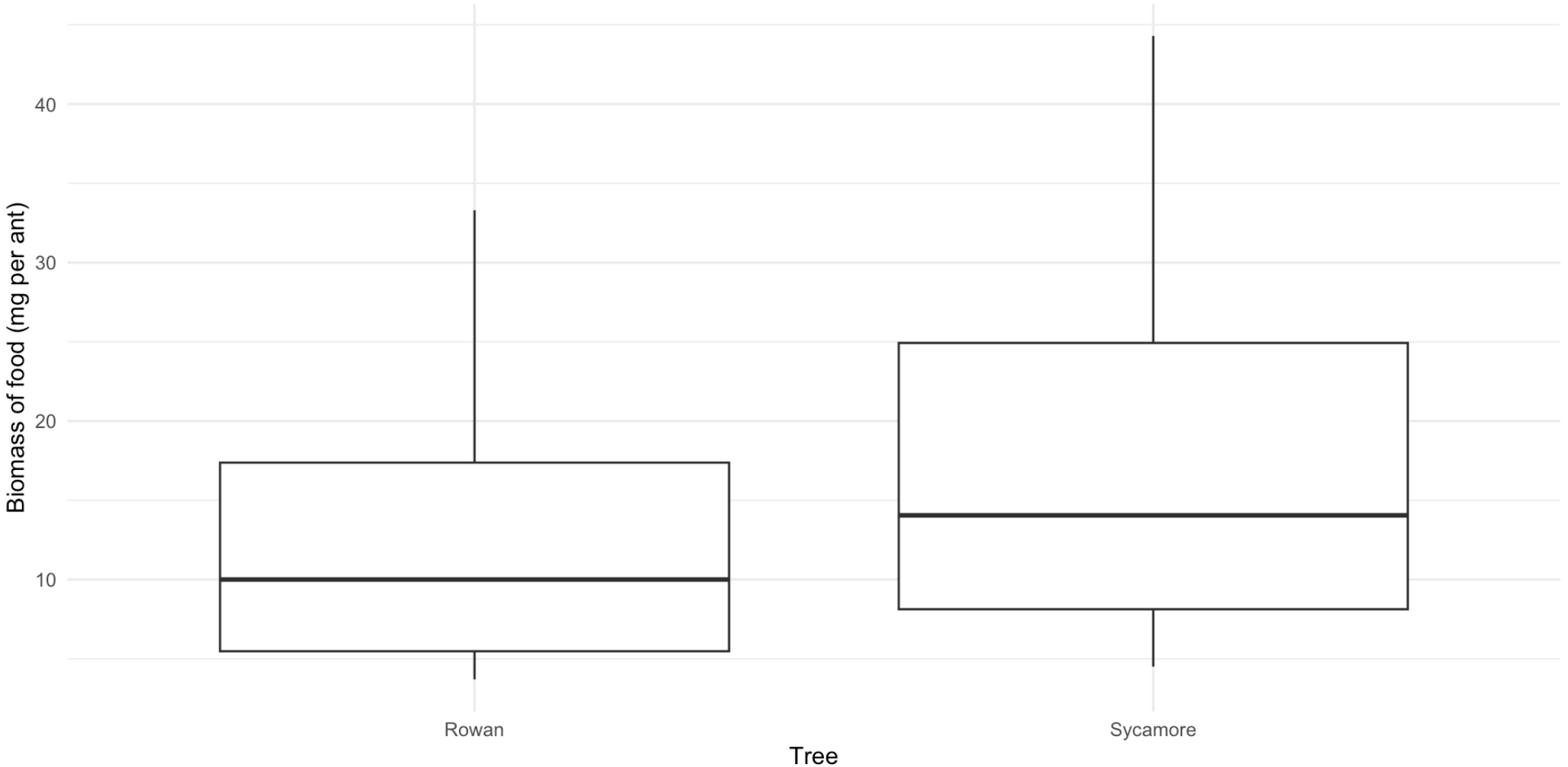
Is the food collected by ants different between two sites?

Data structure

```
Rows: 54
Columns: 2
$ Food <dbl> 11.9, 33.3, 4.6, 5.5, 6.2, 11.0, 24.3, 20.7, 5.7, 12.6, 10.2, 4.7…
$ Tree <fct> Rowan, Rowan, Rowan, Rowan, Rowan, Rowan, Rowan, Rowan, Rowan, Ro…
```

We want to compare the mean biomass of food, collected by ants between the two sites in **dry weight (mg) of prey, divided by the total number of ants leaving the tree in 30 minutes**.

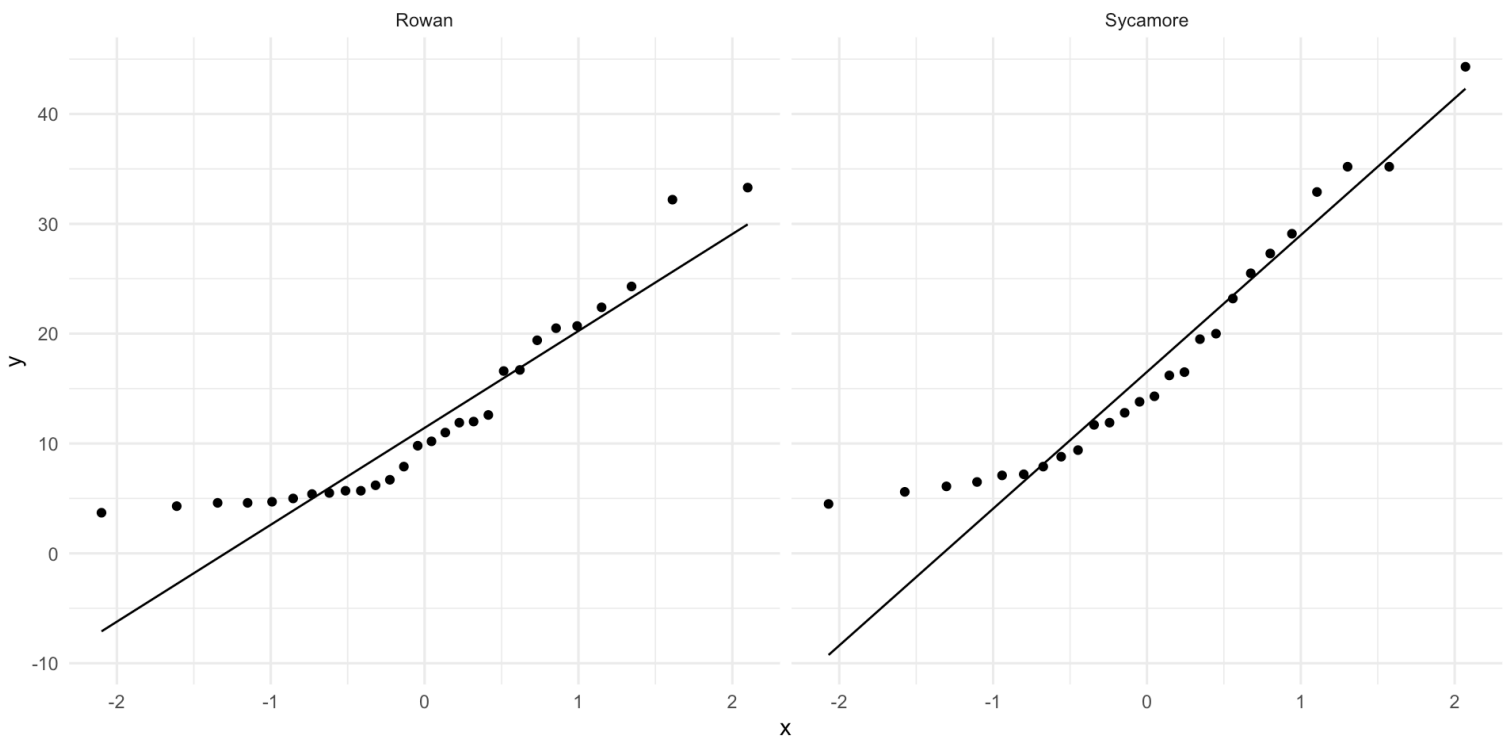
Visualising the data



Does this data meet the assumptions of the two-sample *t*-test?

Checking assumptions

We have some idea that the data may not be normally distributed, but are not quite sure. So let's check using the Q-Q plot.



- Curvature of the data points away from the line indicates non-normality.
- Boxplots (previous slide) suggest equal variances.
- **Let's transform the data.**

Picking a transformation

We need to consider the **type of data** and the **shape of its distribution** when choosing a transformation. These can be assessed using:

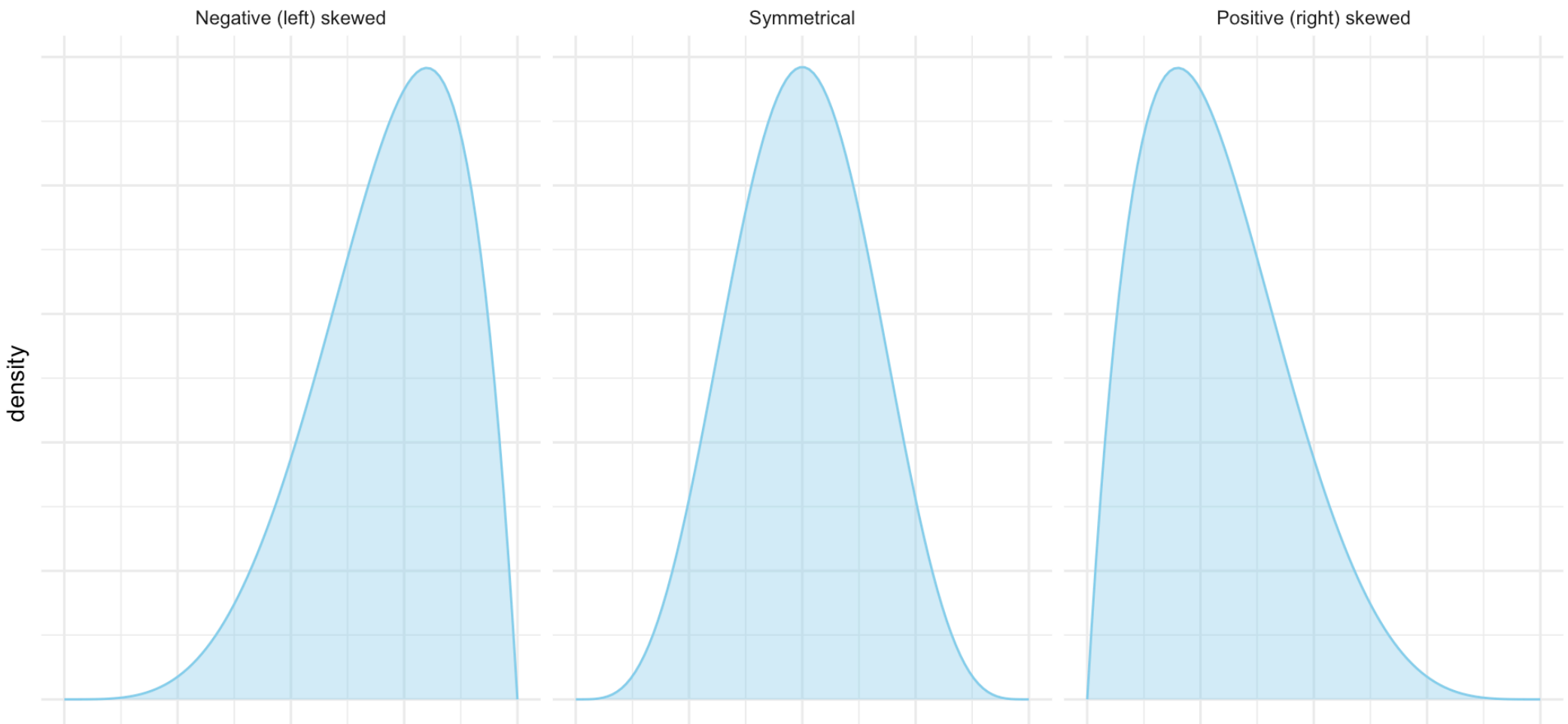
- **Histograms** and **Q-Q plots** to assess normality - DONE
- **Box plots** to assess homogeneity of variance - DONE
- **Skewness** and **kurtosis** to assess the shape of the distribution - NEXT

Skewness

The degree of asymmetry in the data distribution when compared to a normal distribution.

- Represented by the **skewness coefficient** (γ_1) and can be positive, negative, or zero.
- Skewness values between **-0.5 and 0.5** are considered acceptable (fairly symmetrical).
- **Negative** skewness indicates a *left*-skewed distribution, while **positive** skewness indicates a *right*-skewed distribution.
- Above 1 or below -1, the distribution is considered ***highly* skewed**.

Example: skewness

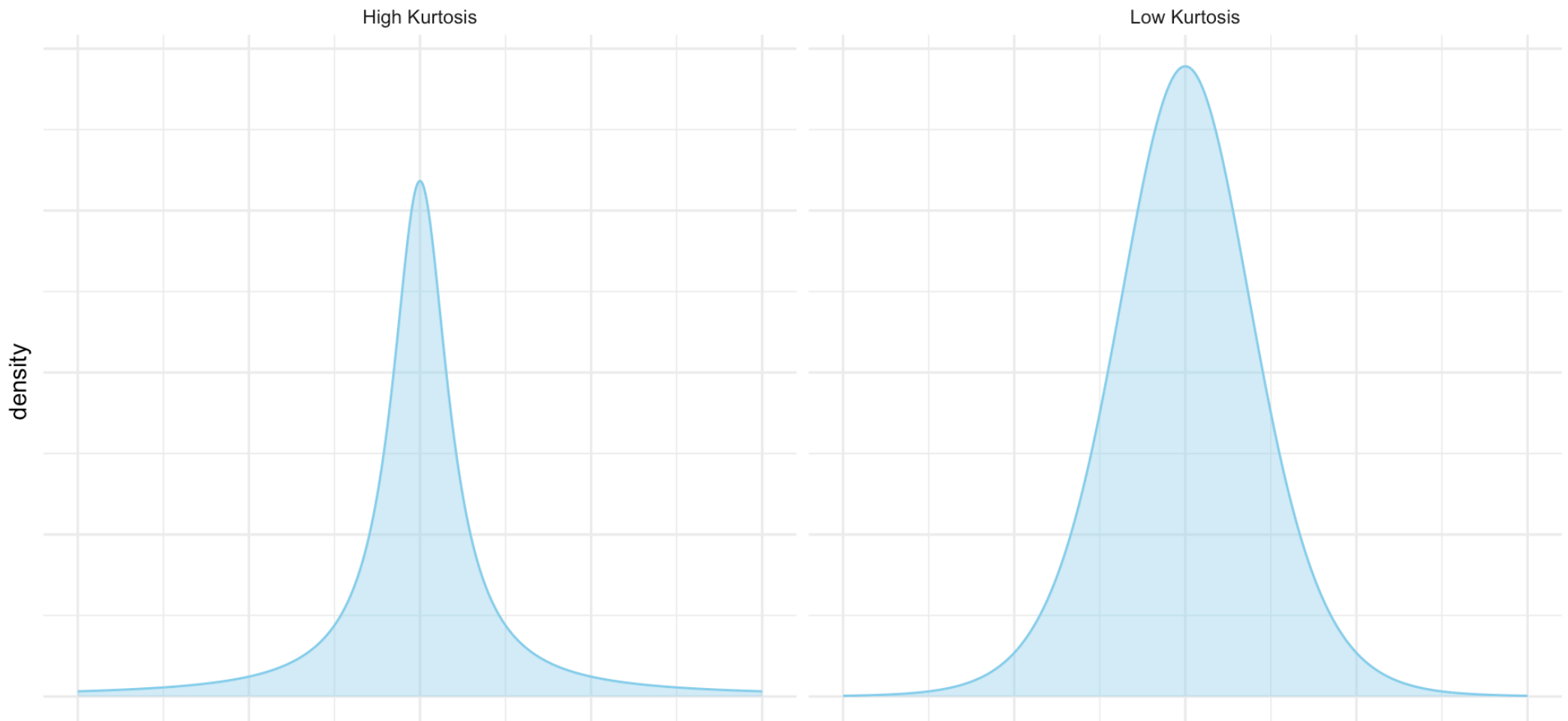


Kurtosis

Used to describe the extreme values (outliers) in the distribution versus the tails.

- **High kurtosis (>3)** indicates a distribution with **heavy tails** and a **peaked centre**. When this happens, we should investigate the data for outliers.
- **Low kurtosis (<3)** indicates a distribution with **light tails** and a **flat centre**. There are fewer to no outliers in the data.

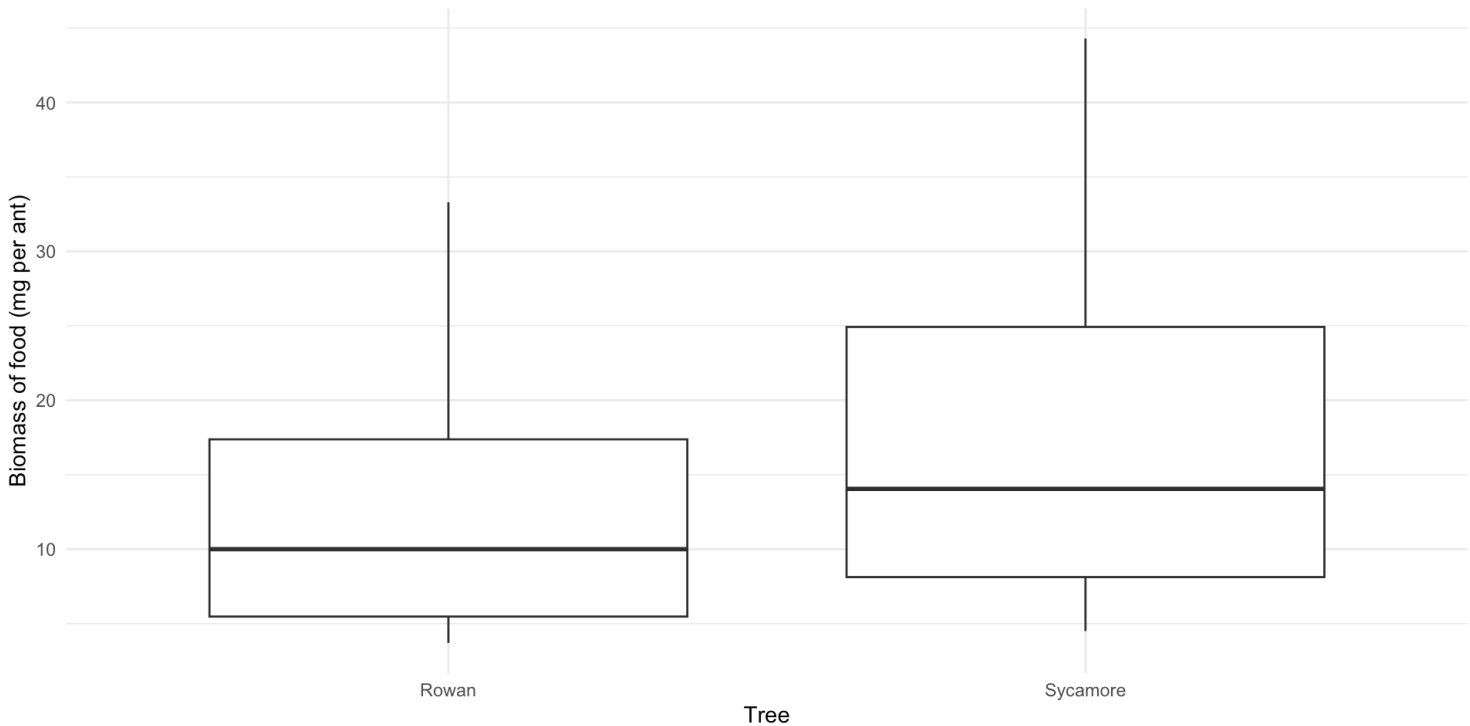
Example: kurtosis



Skewness and kurtosis in the ants data

With experience we can “eyeball” the data, but we can also calculate the skewness and kurtosis.

```
# A tibble: 2 × 3
  Tree     skewness kurtosis
  <fct>       <dbl>    <dbl>
1 Rowan        1.04     3.15
2 Sycamore     0.807    2.63
```



From the results we can see that both sites have a **positive skewness**. Site `Rowan` has high kurtosis.

Data transformation

Workflow

1. Check the data for normality and homogeneity of variance (i.e. **test assumptions**).
2. If the assumptions are violated, consider **transforming the data**.
3. **Repeat** checks on assumptions. If assumptions are **met**, proceed with the *t*-test on the transformed scale. *Otherwise, use a different transformation or consider using a non-parametric test.*
4. Interpret the statistical results and **back-transform the results** to the original scale (optional but recommended) to aid interpretation.

Picking a transformation

For positive skewness

- **Square root** transformation: $\sqrt{x}$ for skewness between 0.5 and 1 and kurtosis < 3.
- **Logarithmic** transformation: $\log(x)$ for skewness > 1 and kurtosis < 3.
- **Reciprocal** transformation: $\frac{1}{x}$ for skewness > 1 and kurtosis > 3 (quite extreme).

For negative skewness

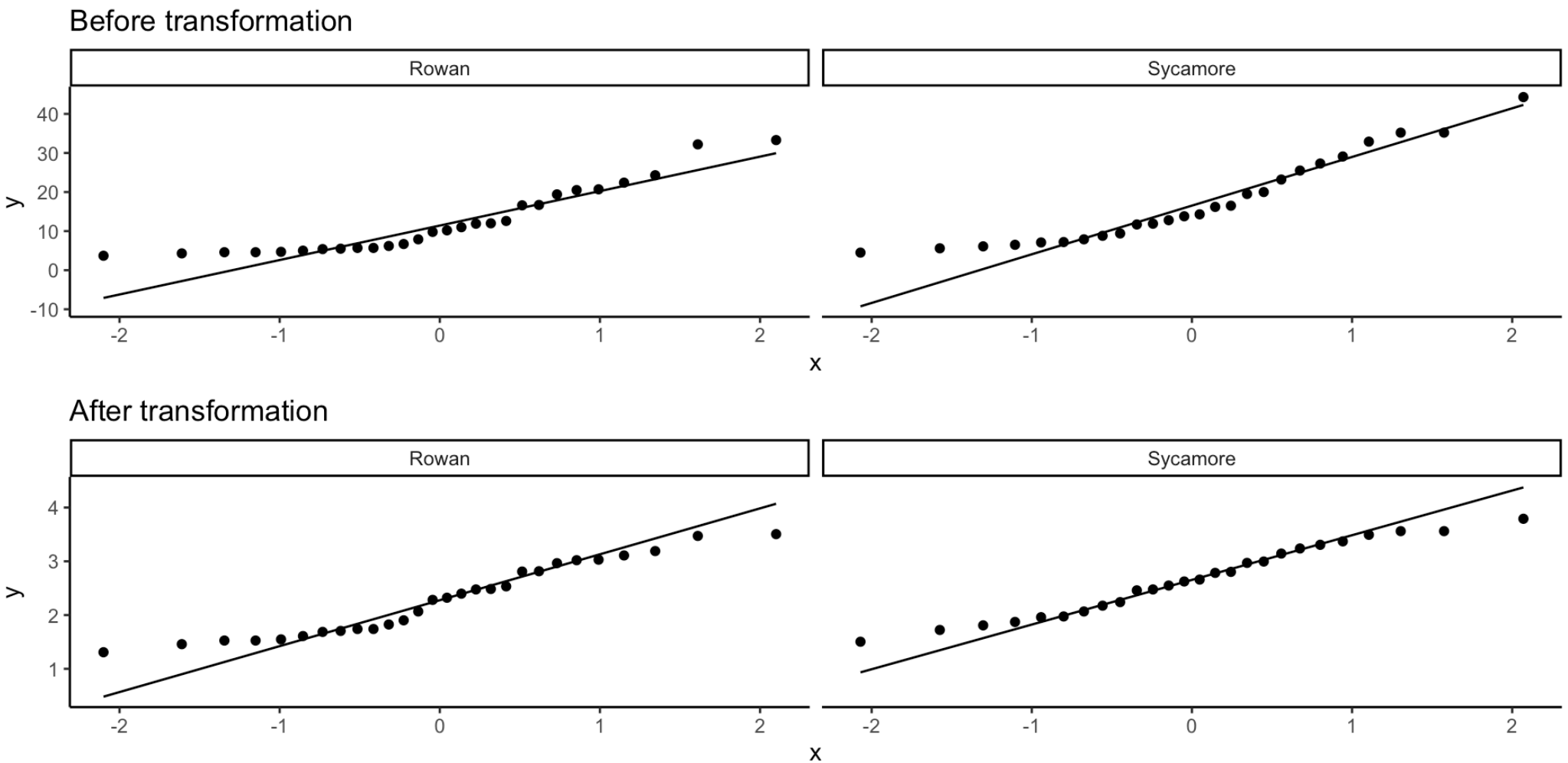
- This is rare as most biological data are positively skewed. However, you can try the **square** x^2 or **cube** x^3 transformation.
- If negatively skewed data contains zeros, consider using the log transform and adding a constant to the data before transformation e.g. $\log(x+1)$.

> **Note**
>
> There is also the **Box-Cox transformation** which informs us of the best transformation to apply to the data without the need to check skewness and kurtosis. This method is not covered in this unit,

but you can read more about it here (the simple R version) or here (more detailed mathematical explanation).

How do we check if the transformation worked?

We need to apply the transformation to the entire dataset and check the Q-Q plot again.



Checking skewness and kurtosis after transformation

```
# A tibble: 2 × 3
  Tree      skewness kurtosis
  <fct>        <dbl>    <dbl>
1 Rowan    0.271         1.77
2 Sycamore 0.0000457     1.86
```

Performing the *t*-test

```
    Two Sample t-test

data:  Food_log by Tree
t = -2.0521, df = 52, p-value = 0.04521
alternative hypothesis: true difference in means between group Rowan and group Sycamore is not equal to 0
95 percent confidence interval:
 -0.732858080 -0.008203447
sample estimates:
   mean in group Rowan mean in group Sycamore
              2.287756                2.658287
```

How do we interpret the results?

Evidence suggests that the log-transformed mean biomass of food collected by ants from the Rowan site is significantly different from the log-transformed mean biomass of food collected by ants from the Sycamore site (t = -2.05, df = 52, p = 0.045).

Back-transforming the results

- For power transformations, we can back-transform the results to the original scale using the inverse function.
- Log transformations are a bit tricky as the inverse function is the exponential function.
  - For the natural log transformation which is `log()` in R, the inverse function is the exponential function: e^x.
  - For the base 10 log transformation which is `log10()` in R, the inverse function is 10^x.

Interpretation

Back-transforming mean values

```
[1] 1.448503
```

> Evidence suggests that the log-transformed mean biomass of food collected by ants from the Rowan site is significantly different from the log-transformed mean biomass of food collected by ants from the Sycamore site (t = -2.05, df = 52, p = 0.045).

The mean biomass of food collected by ants from the Sycamore site (14.3 mg) is 1.4 times greater than the mean biomass of food collected by ants from the Rowan site (9.9 mg).

Back-transforming confidence intervals

```
[1] 0.4805336 0.9918301
attr(,"conf.level")
[1] 0.95
```

Comparing to a test without transformation

```
    Two Sample t-test

data:  Food by Tree
t = -1.9217, df = 52, p-value = 0.06013
alternative hypothesis: true difference in means between group Rowan and group Sycamore is not equal to 0
95 percent confidence interval:
 -10.4916030   0.2267678
sample estimates:
   mean in group Rowan mean in group Sycamore
              12.27143               17.40385
```

- Original mean values:
  - Rowan = 12.3 mg
  - Sycamore = 17.4 mg
- Log-transformed mean values:
  - Rowan = 2.3 lg(mg)
  - Sycamore = 2.7 lg(mg)
- Back-transformed mean values:
  - **Rowan = 9.9 mg**
  - **Sycamore = 14.3 mg**

The original mean values are based on the arithmetic mean, while the log-transformed mean values are based on the geometric mean. The geometric mean is more appropriate for skewed data.

- Original 95% confidence interval:
  - -10.5 to 0.2 mg
- Log-transformed 95% confidence interval:
  - 0.5 to 1 lg(mg)
- Back-transformed 95% confidence interval:
  - **1.6 to 2.7 mg**

The influence of kurtosis on the 95% confidence interval is evident when comparing the original and back-transformed confidence intervals, as the log transform reduces the effect of outliers on the data.

Thanks!

This presentation is based on the SOLES Quarto reveal.js template and is licensed under a Creative Commons Attribution 4.0 International License.