Topic 6 – Two-sample *t*-tests

ENVX1002 Introduction to Statistical Methods

Floris van Ogtrop
The University of Sydney

Apr 2025

THE UNIVERSITY OF SYDNEY

Testing differences in means

Recap: we have one sample

One-sample *t*-test: compare the sample of data to a *fixed* value of interest (e.g. a hypothesised value, or a population mean).

Examples

- *Is the mean height of students in ENVX different from the population mean of 170 cm?*
- *Is the mean heart rate of students in ENVX different from the population mean of 70 bpm*

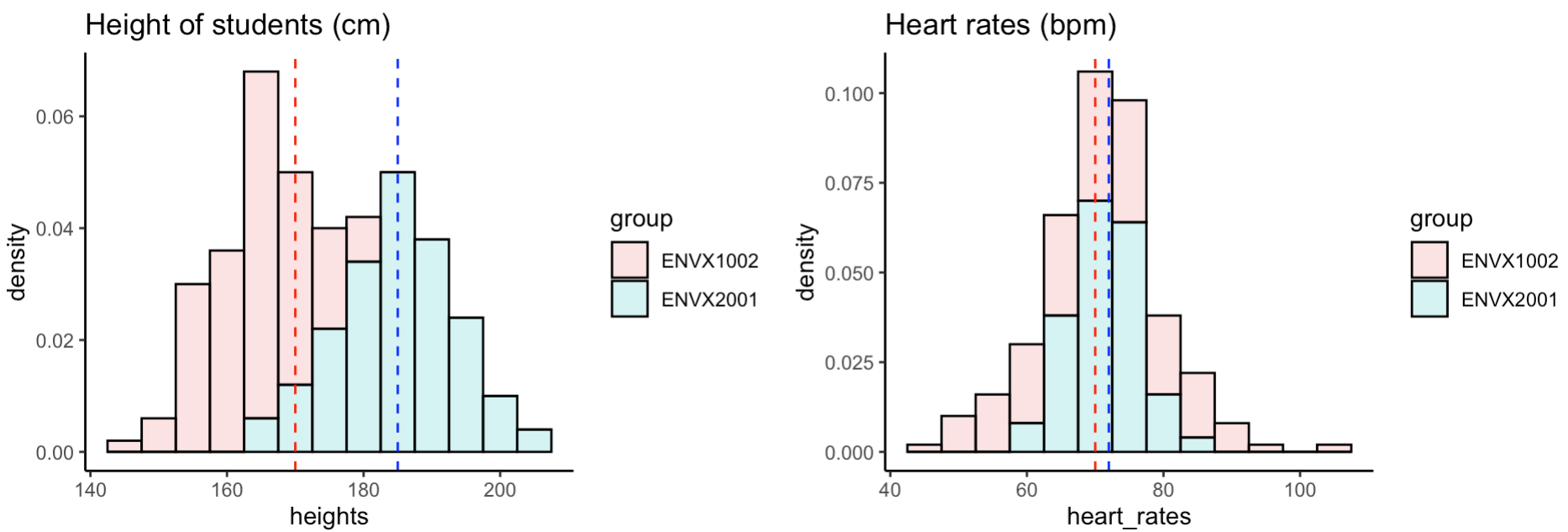
▶ Code

What if we want to compare a sample of data to *another* sample?

Examples

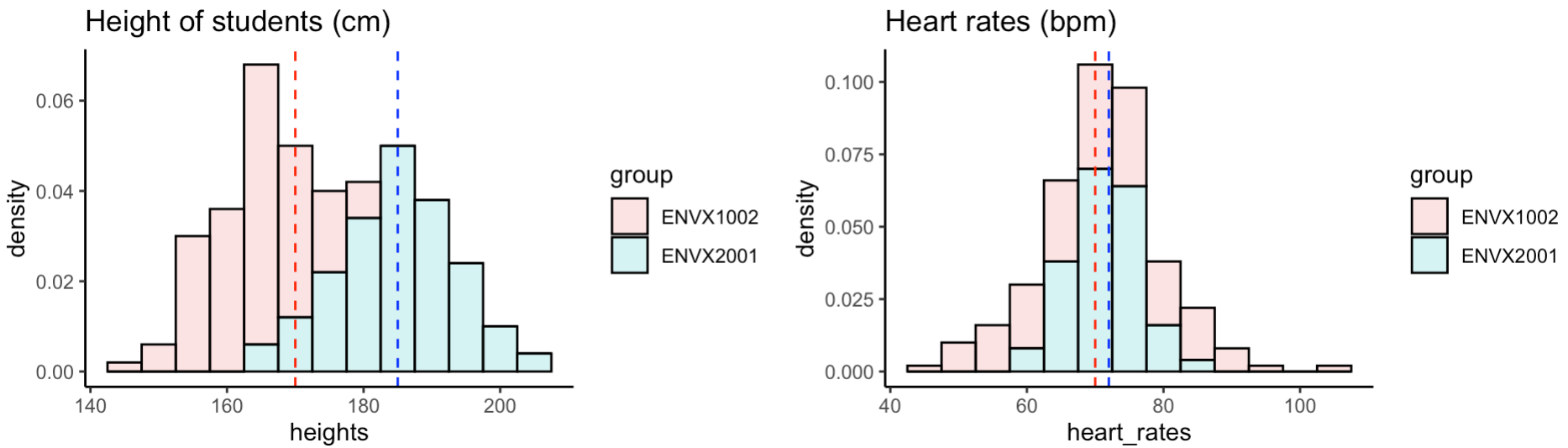
- *Is the mean height of students in ENVX1002 different from ENVX2001?*
- *Is the mean heart rate of students in ENVX1002 different from ENVX2001?*

▶ Code

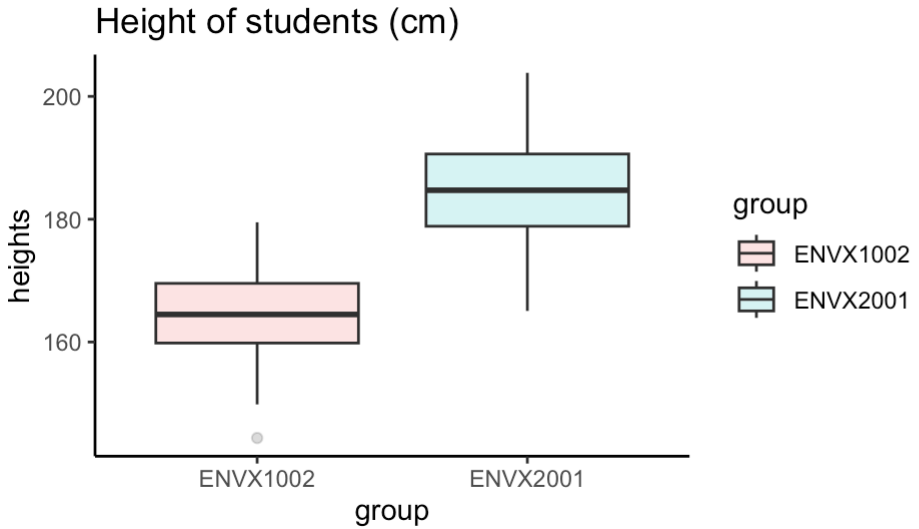


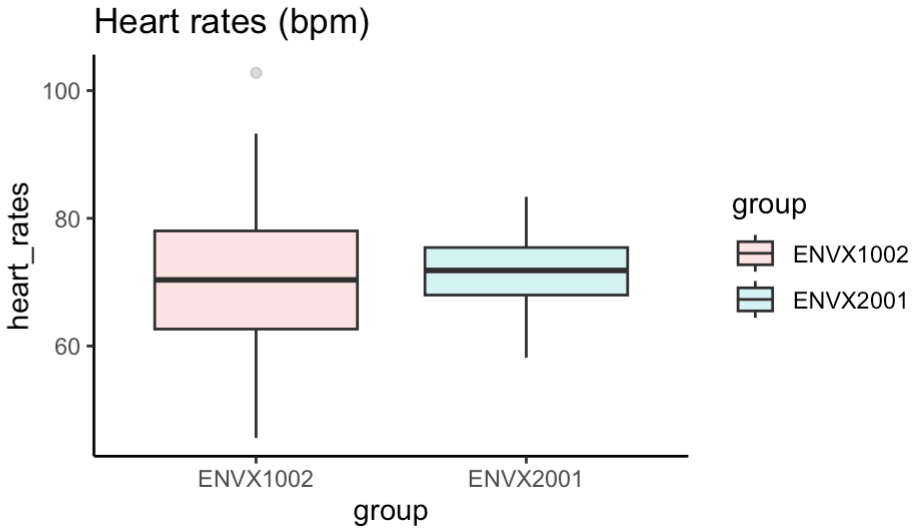
Two-sample *t*-test

Comparing two samples: visualisation

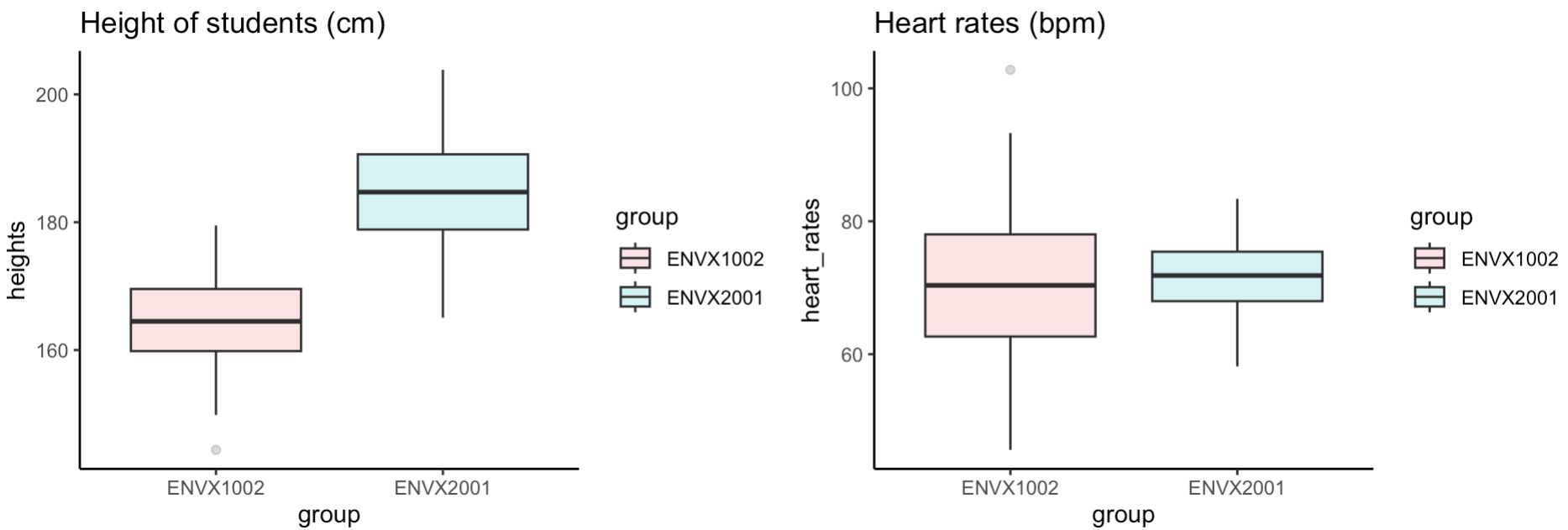


▶ Code


Height of students (cm)
heights
200
180
160
ENVX1002
ENVX2001
group
group
ENVX1002
ENVX2001


Heart rates (bpm)
heart_rates
100
80
60
ENVX1002
ENVX2001
group
group
ENVX1002
ENVX2001

Some considerations: the boxplot



- Trade-off between being able to see the distribution of the data and being able to compare between groups.
- The *recommended* approach when comparing two or more groups of data in most cases.

Drinking two cans of Red Bull ‘increases risk of cardiac arrest by a fifth’ for people with an underlying heart condition

The popular drink could put people with long QT syndrome in serious danger

Sarah Young • Tuesday 18 July 2017 12:46 BST • Comments

Does Red Bull increase the heart rate of students?

Data

A simulated example (data is not real):

▶ Code

Experimental design: two groups of students selected at random, *without* replacement, from the ENVX1002 cohort.

- `redbull` group: students who consumed 250 ml of Red Bull.
- `control` group: students who consumed 250 ml of water (control group).

Heart rate in beats per minute (bpm) was measured *20 minutes after consumption*.

Structure of data

▶ Code

```
'data.frame':   24 obs. of  2 variables:
 $ group     : chr  "redbull" "redbull" "redbull" "redbull" ...
 $ heart_rate: num  72 88 72 88 76 75 84 80 60 96 ...
```

HATPC

Hypothesis | Assumptions | Test | P-value | Conclusion

Hypothesis

For a two-sample *t*-test, the null hypothesis is that the means of the two groups are equal, and the alternative hypothesis is that the means are different.

$$H_0 : \mu_{\text{redbull}} = \mu_{\text{control}}$$

$$H_1 : \mu_{\text{redbull}} \neq \mu_{\text{control}}$$

Compare this to the one-sample t*-test, where the null hypothesis is that the sample mean is equal to a fixed value:*

$$H_0 : \mu = \mu_0$$

$$H_1 : \mu \neq \mu_0$$

Assumptions

The assumptions of the two-sample *t*-test include:

1. **Normality**: the data are normally distributed.
2. **Homogeneity of variance**: the variances of the two groups are equal.

Why are these assumptions important?

- Since the *t*-test compares the means of two groups, normality ensures that the means are the *best estimate* of the population means.
- Equal variances indicates that the two groups have similar "noise" influencing their means, except for the "treatment" effect.
    - ➠ In the Red Bull example, this means that the range of heart rate values in students for both groups is similar, except for the effect of consuming Red Bull.

Assumption: normality

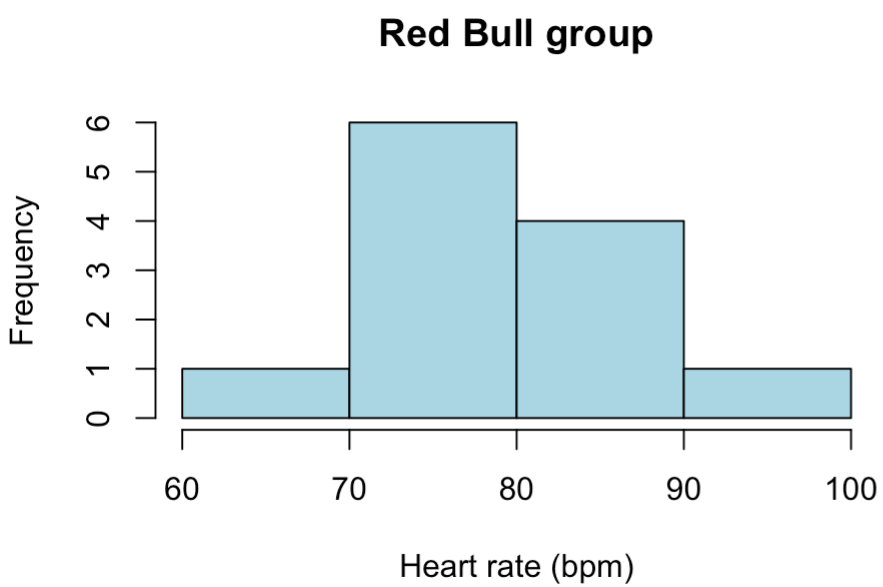
Histogram | QQ-plot | Shapiro-Wilk test

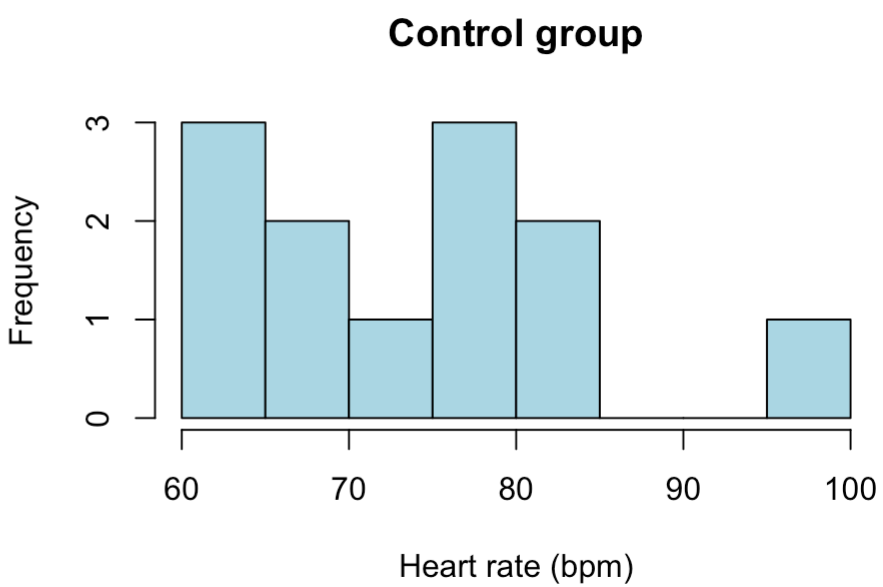
Normality: histogram

- We visually inspect the distribution of the data using histograms, generally for **each group**.
- Look out for: symmetry, skewness, and multimodality.
- Hard to visualise when n (sample size is small)

R base graphics ggplot2 ggpubr

▶ Code



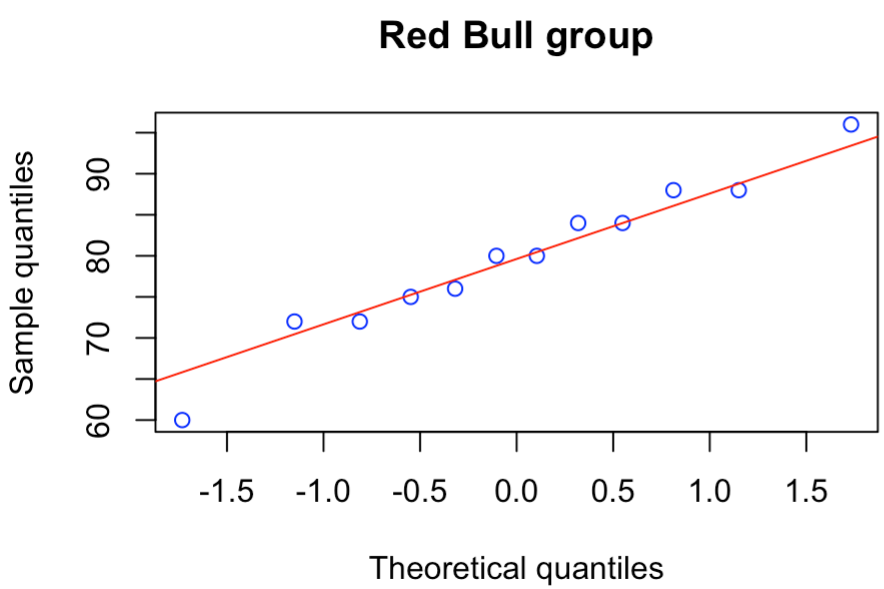


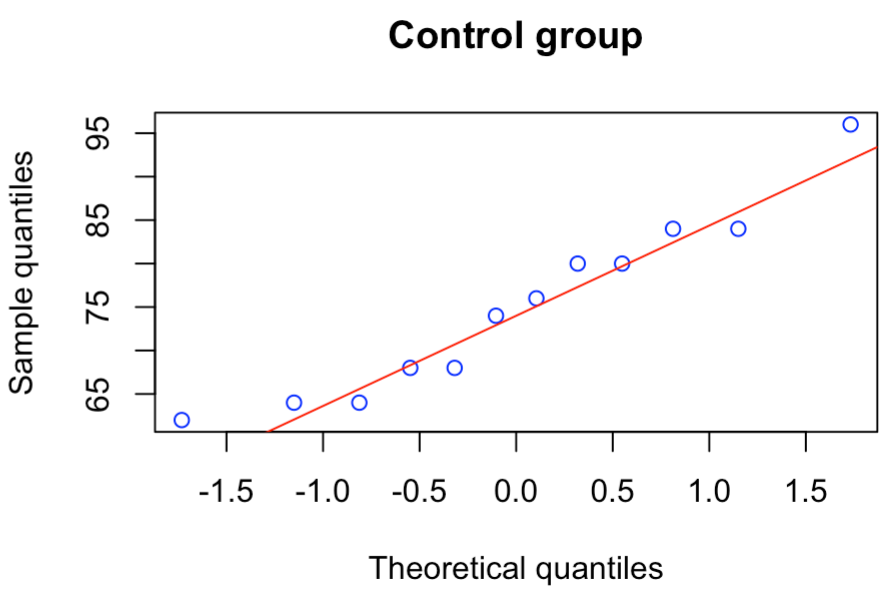
Normality: QQ-plot

- The qq-plot is a graphical method to *specifically* assess the normality of the data. Again, we look at the data for **each group**.
- Look out for: deviations from the straight line.

R base graphics ggplot2 ggpubr

▶ Code





Normality: formal test

- Use the Shapiro-Wilk test which tests the null hypothesis that the data are normally distributed.
- This test is sensitive to deviations from normality in the tails of the distribution, and is suitable for small sample sizes (about 5 to 50 observations).

▶ Code

```
    Shapiro-Wilk normality test

data:  redbull$heart_rate[redbull$group == "redbull"]
W = 0.97459, p-value = 0.9524
```

▶ Code

```
    Shapiro-Wilk normality test

data:  redbull$heart_rate[redbull$group == "control"]
W = 0.93733, p-value = 0.4643
```

Conclusion: *p*-values are greater than 0.05, so we do not reject the null hypothesis of normality. The data are normally distributed.

What if the normality assumption is violated?

- The *t*-test is robust to deviations from normality, especially for large sample sizes due to the Central Limit Theorem.
- If the sample size is small, consider using a non-parametric test (e.g. the Wilcoxon rank-sum test): **next week**
- Alternatively, transform the data: **later**

Assumption: homogeneity of variance

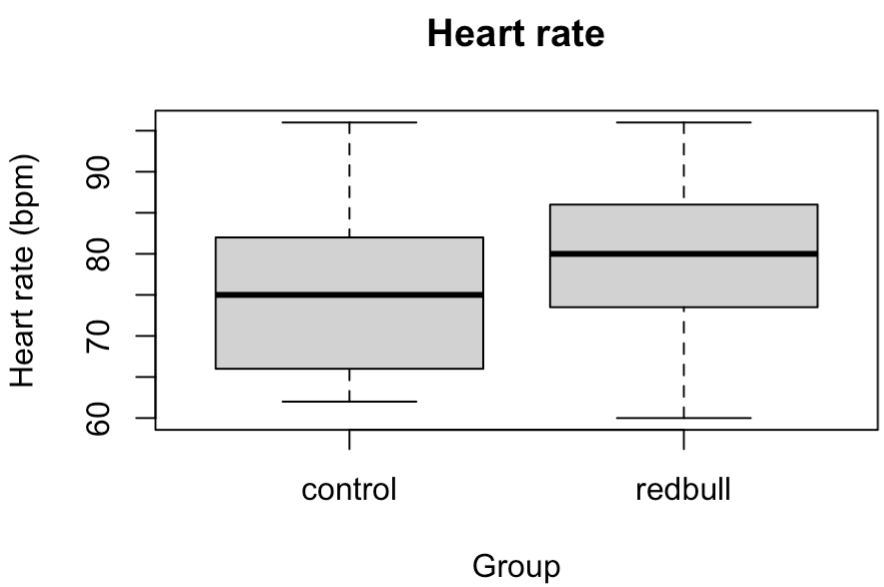
Boxplots | Formal tests

Equal variances: boxplot

- We visually inspect the spread of the data using boxplots, generally for **each group**.
- Look out for: differences in spread, outliers, and symmetry.

R base graphics ggplot2 ggpubr

▶ Code



Conclusion: *The spread of the data appears to be similar between the two groups.*

Equal variances: formal tests

- Bartlett's and Levene's tests may be used to test the null hypothesis that the variances of the groups are equal.
- These tests are sensitive to deviations from normality (Levene's test is less so compared to Bartlett's), and are suitable for small sample sizes.

Levene's test Bartlett's test

▶ Code

```
Levene's Test for Homogeneity of Variance (center = median)
      Df F value Pr(>F)
group  1   0.289 0.5962
      22
```

Conclusion: *p*-values are greater than 0.05, so we do not reject the null hypothesis of equal variances. The variances of the two groups are equal.

What if the equal variance assumption is violated?

Some debate exists on what to do, but choices include:

- Use the **Welch's *t*-test**, which is robust to unequal variances: *coming up next*
- **Transform** the data to stabilise the variances: **later**
- Perform a non-parametric test: **next week**

The Welch's *t*-test

- The Welch's *t*-test is a modification of the two-sample *t*-test that **does not assume equal variances**.
- Also applicable when the **sample sizes are unequal**.

Why not use the Welch's *t*-test all the time?

- Ongoing debate on whether to use the Welch's *t*-test or the Student's *t*-test when the variances are equal.
- The Welch's *t*-test is generally considered more robust, and **is the default in R's `t.test()` function**.
- You can still use the Student's *t*-test by setting `var.equal = TRUE` in the `t.test()` function.

Are the assumptions of normality and homogeneity of variance met?

When reporting in journals, it is common to simply state that the assumptions were met and what tests were used to confirm them, without showing the exact results of the tests!

Example 1

> *The assumptions of normality and homogeneity of variance were met for the data (Shapiro-Wilk test, $p > 0.05$; Levene's test, $p > 0.05$). Thus, we performed a two-sample* t*-test...*

Example 2

> *Visual inspection of the histograms, QQ-plots, and boxplots showed that the data met the assumptions of both normality and homogeneity of variance. Thus, we performed a two-sample* t*-test..*

For your lab reports, you should show the results of the tests (because we want to check your work!)

P-value and Conclusion

Performing the *t*-test

```r
t.test(heart_rate ~ group, data = redbull, var.equal = TRUE)
```

```
    Two Sample t-test

data:  heart_rate by group
t = -1.1365, df = 22, p-value = 0.268
alternative hypothesis: true difference in means between group control and group redbull is not equal to 0
95 percent confidence interval:
 -12.947117   3.780451
sample estimates:
mean in group control mean in group redbull
             75.00000              79.58333
```

Results indicate that the means of the two groups are **not** significantly different (p = 0.27).

Compare with the Welch's *t*-test

```r
t.test(heart_rate ~ group, data = redbull)
```

```
    Welch Two Sample t-test

data:  heart_rate by group
t = -1.1365, df = 21.845, p-value = 0.2681
alternative hypothesis: true difference in means between group control and group redbull is not equal to 0
95 percent confidence interval:
 -12.950568   3.783902
sample estimates:
```

Conclusion

Differences in heart rate we not statistically significant between the Red Bull and control groups (t_{22} = -1.1, p = 0.27) indicating that Red Bull did not significantly increase the heart rate of students sampled from ENVX1002.

The paired *t*-test

Are the two sample independent?

When testing if two samples are different from each other, we need to consider two possible scenarios:

- **Independent samples**: The samples are drawn from two different populations, **or** the samples are not related to each other – **independent groups**.
- **Related samples**: The samples are drawn from the same population, **and/or** the samples are related to each other – **repeated measures** or **matched pairs**.

If the samples are related, a **paired *t*-test** is more appropriate than a two-sample *t*-test as it accounts for the relationship between the samples that could confound the results.

Paired *t*-test

Experimental design (what if?)

Before

> Two groups of students selected at random, *without* replacement, from the ENVX1002 cohort.

Paired design

The same student was used in a before/after experiment, where the heart rate was measured before and after consuming 250ml of Red Bull. Twelve (12) students were selected at random from the ENVX1002 cohort.

- Data is no longer independent, as the same student is measured twice.
- The student now confounds the results, as the heart rate of the same student is likely to be **correlated** even without consuming Red Bull.
- Total number of students is now 12, not 24.
- **Let's assume the data collected are exactly the same.**

Hypothesis

For a paired *t*-test, the null hypothesis is that the mean difference between the two groups is zero, and the alternative hypothesis is that the mean difference is different from zero.

$$H_0 : \mu_{\text{diff}} = 0$$

$$H_1 : \mu_{\text{diff}} \neq 0$$

Compare this to the two-sample t*-test, where the null hypothesis is that the means of the two groups are equal:*

$$H_0 : \mu_{\text{redbull}} = \mu_{\text{control}}$$

$$H_1 : \mu_{\text{redbull}} \neq \mu_{\text{control}}$$

Assumptions of the paired *t*-test

- The assumption of the paired *t*-test is that the differences between the two groups are normally distributed.
- There is no assumption of equal variances, as the paired *t*-test is a one-sample *t*-test on the differences.
    - ➠ Another way to think about it is that since the data are paired, the variance of the differences is the same for both groups.

Performing the paired *t*-test

There are two ways.

Method 1: Calculate the differences, then perform a one-sample *t*-test using `t.test()`

```
diff ← redbull$heart_rate[redbull$group == "redbull"] -
  redbull$heart_rate[redbull$group == "control"]
t.test(diff)
```

```
    One Sample t-test

data:  diff
t = 1.6578, df = 11, p-value = 0.1256
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -1.501628 10.668294
sample estimates:
mean of x
 4.583333
```

Performing the paired *t*-test

There are two ways.

Method 2: Use the `t.test()` function with the x and y and `paired = TRUE` argument

```r
# t.text(group1, group2, paired = TRUE)
t.test(redbull$heart_rate[redbull$group == "redbull"],
  redbull$heart_rate[redbull$group == "control"],
  paired = TRUE)
```

```
    Paired t-test

data:  redbull$heart_rate[redbull$group == "redbull"] and redbull$heart_rate[redbull$group == "control"]
t = 1.6578, df = 11, p-value = 0.1256
alternative hypothesis: true mean difference is not equal to 0
95 percent confidence interval:
 -1.501628 10.668294
sample estimates:
mean difference
       4.583333
```

The results for both methods are identical; the mean difference is not significantly different from zero (p = 0.13).

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

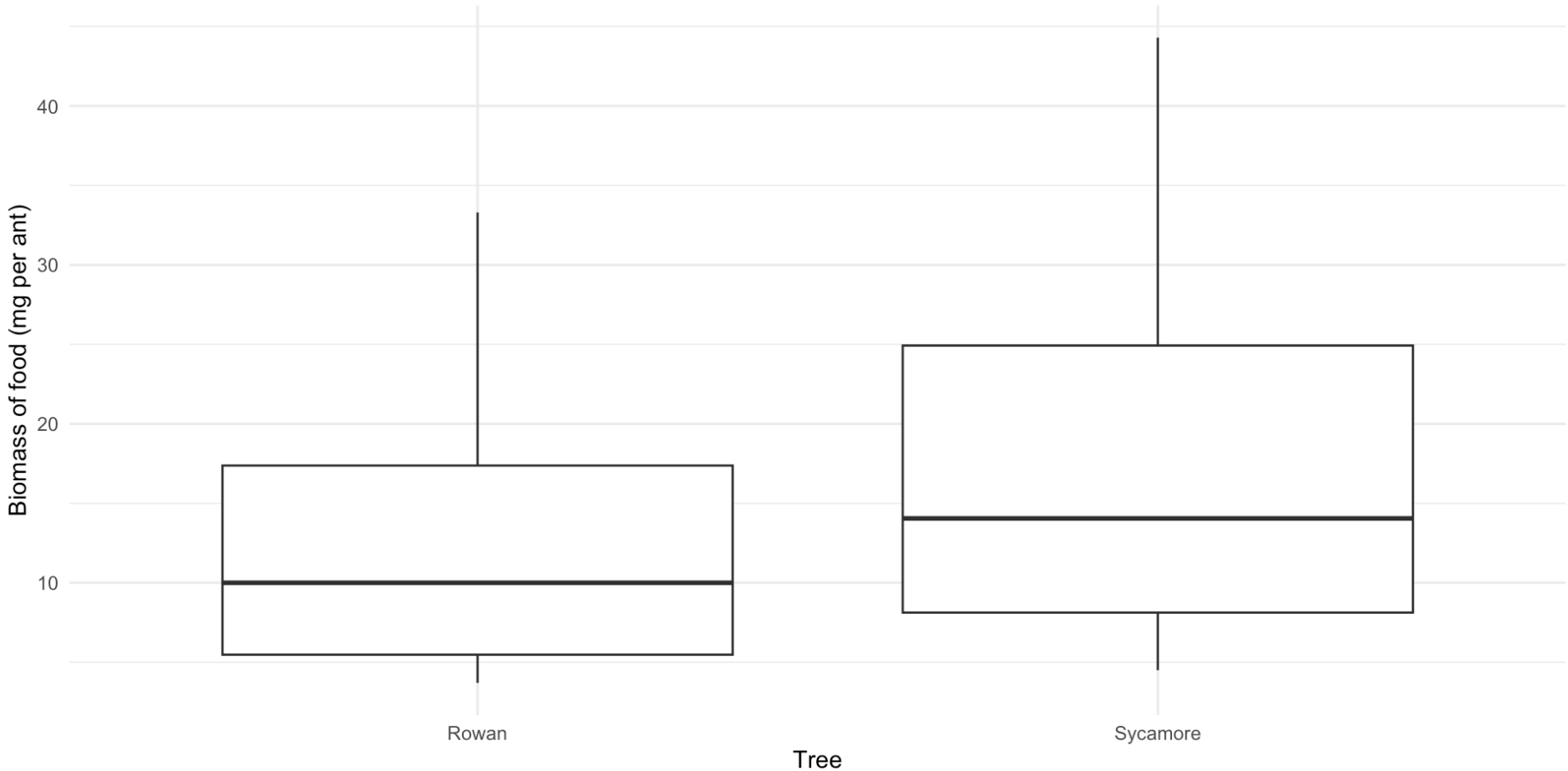
▶ Code

```
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

▶ Code

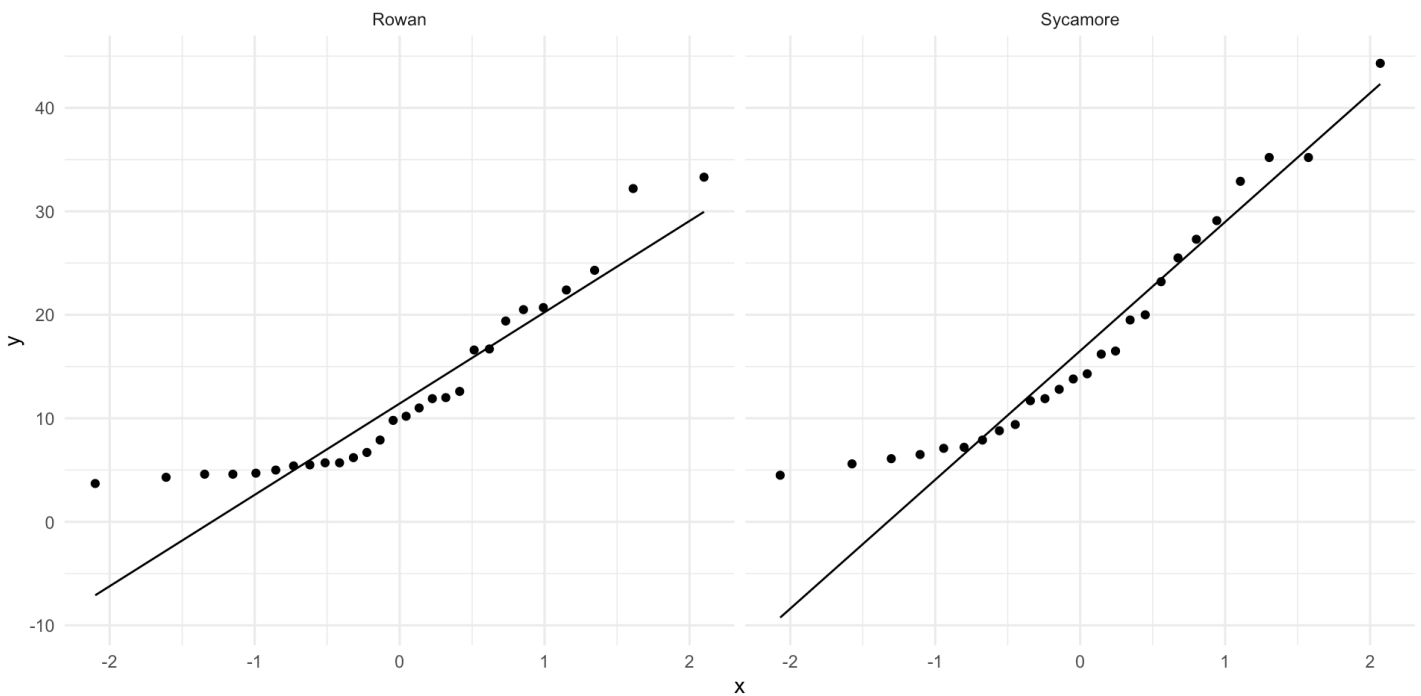


Does this data meet the assumptions of the two-sample *t*-test?

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.

▶ Code



- 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

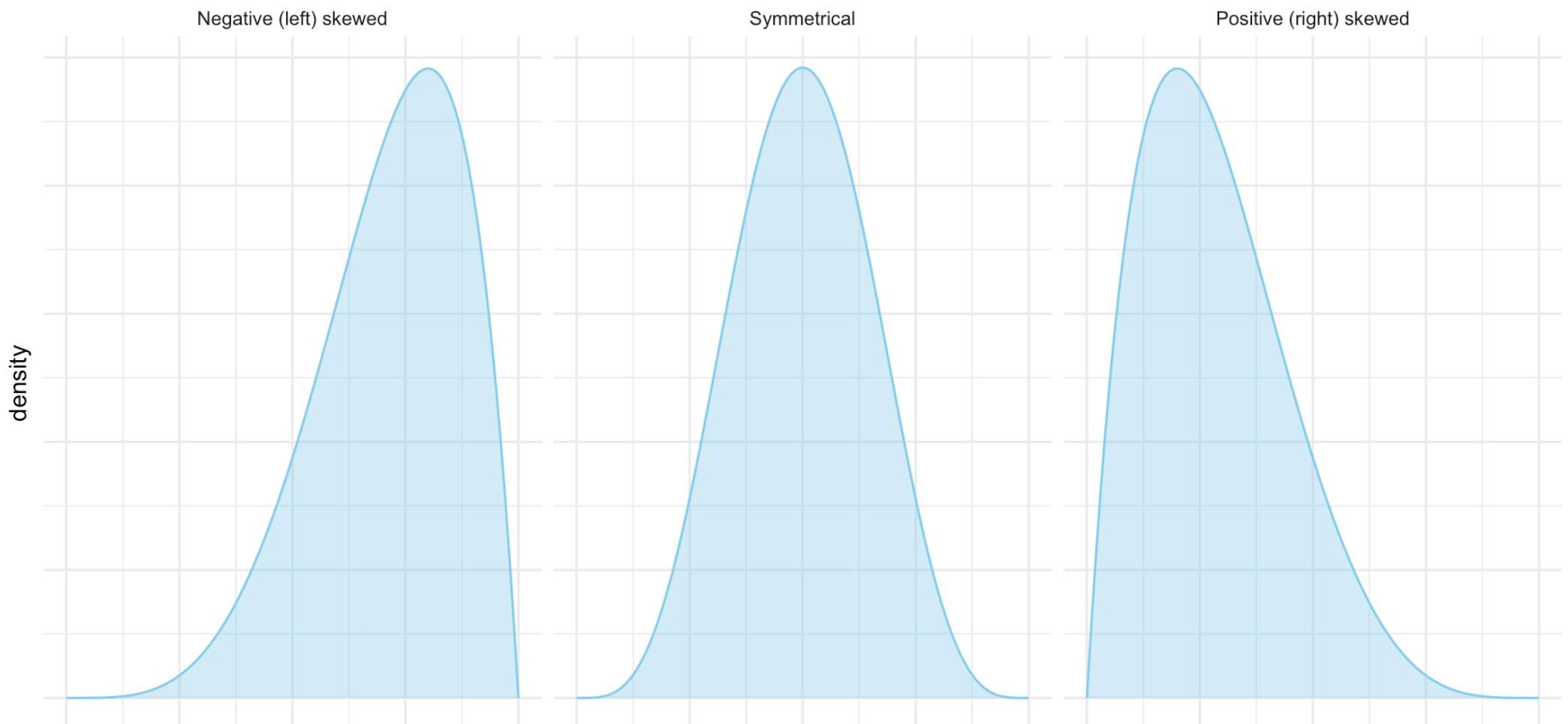
Skewness

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

▶ Code



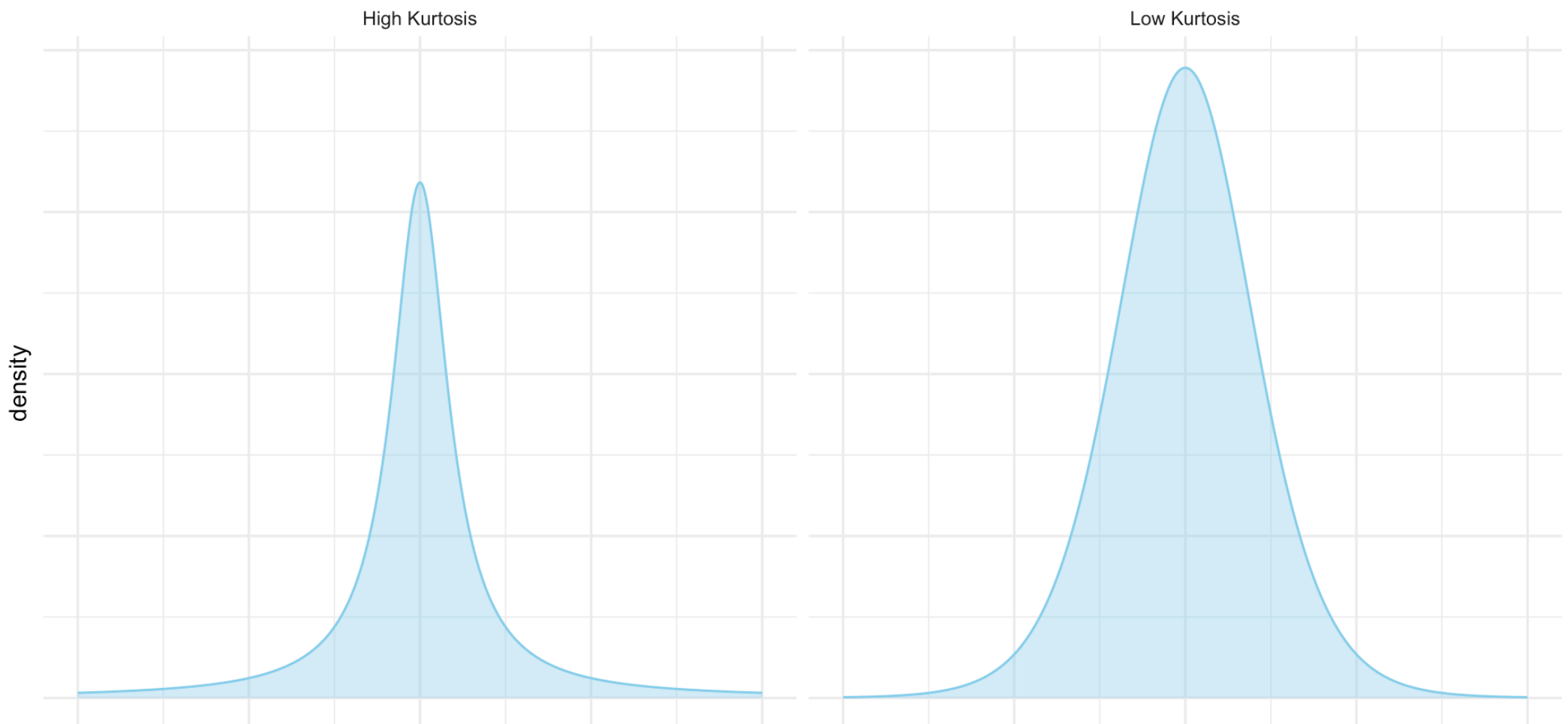
Kurtosis

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

▶ Code

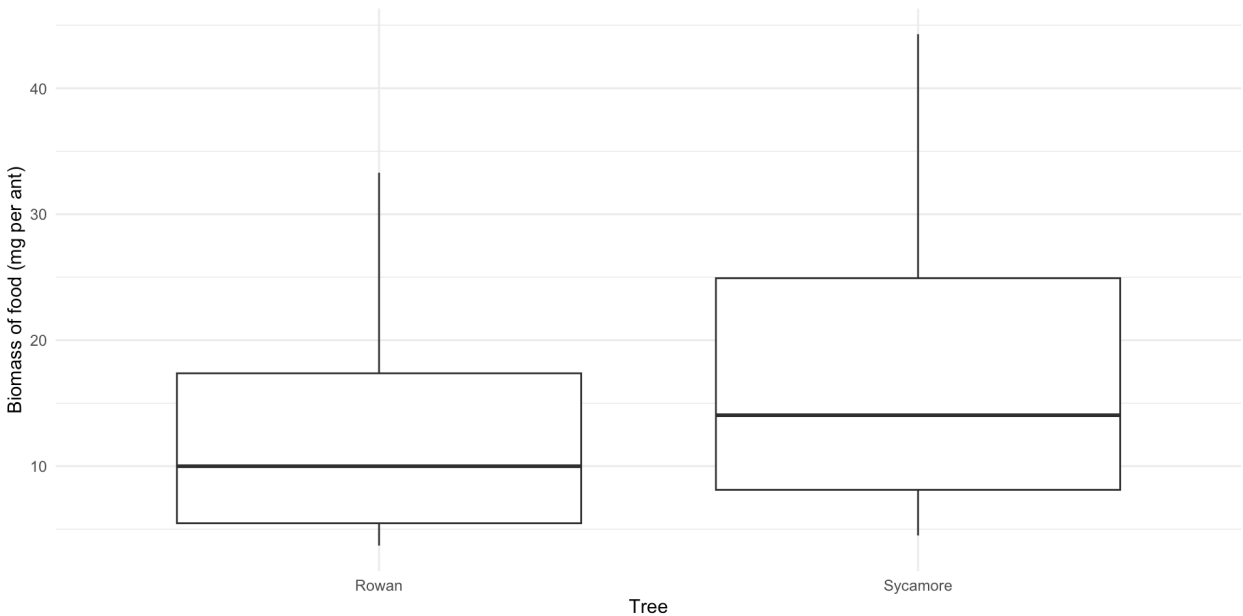


Skewness and kurtosis in the ants data

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

▶ Code

```
# 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 ALL 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,

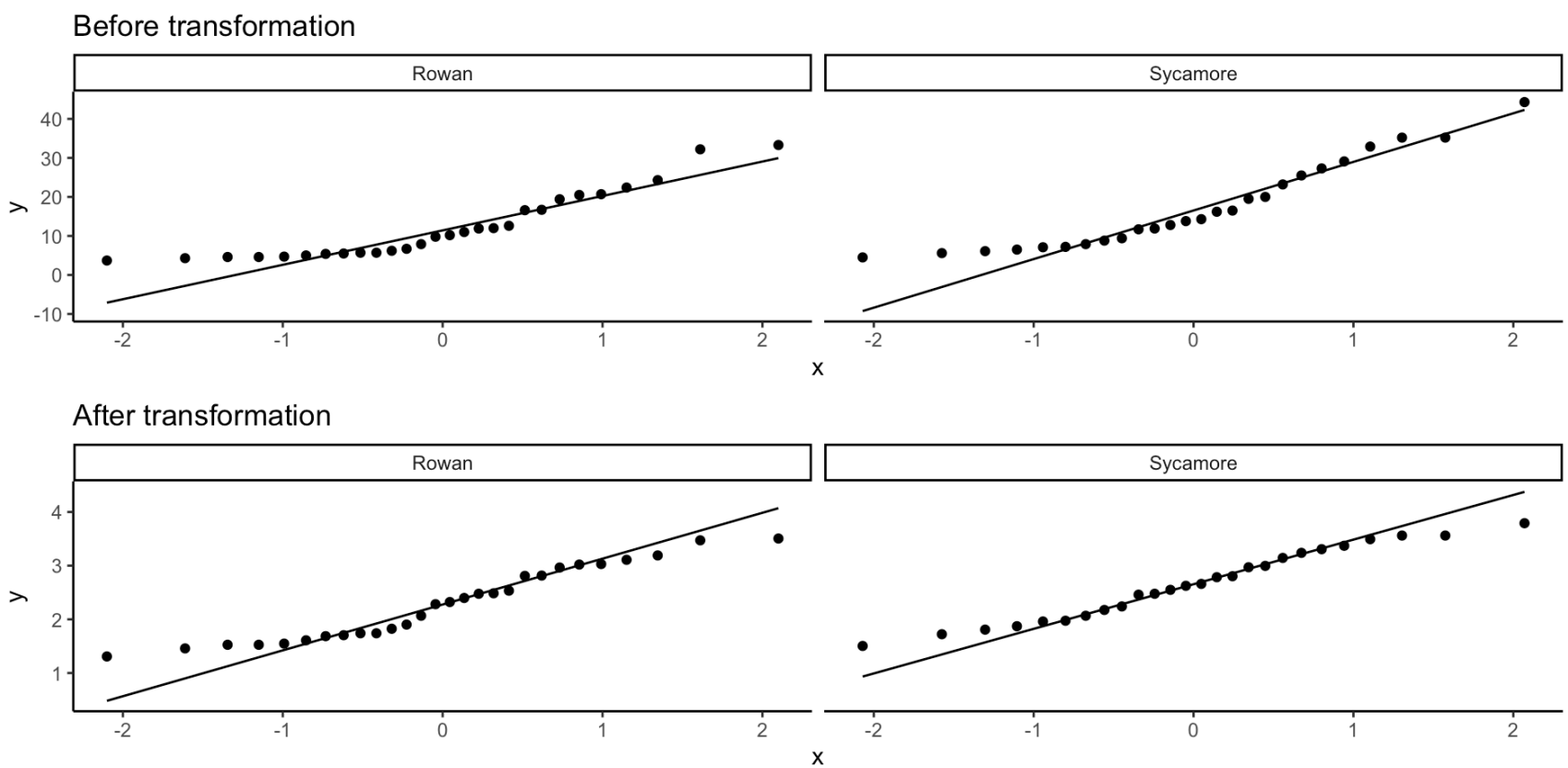
but you can read more about it **here** (the simple R version) or **here** (more detailed mathematical explanation).

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.

▶ Code

▶ Code



Checking skewness and kurtosis after transformation

▶ Code

```
# 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

▶ Code

```
    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

▶ Code

```
[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

▶ Code

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

Comparing to a test without transformation

▶ Code

```
    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!

References

- Quinn G. P. & Keough M. J. (2002) Experimental design and data analysis for biologists. Cambridge University Press, Cambridge, UK.
- Logan, M. (2010). Biostatistical design and analysis using R a practical guide. Hoboken, N.J., Wiley-Blackwell.