

Introduction to t -tests – a special case of the GLM

BIOL2022 – Biology Experimental Design and Analysis (BEDA)

Januar Harianto

The University of Sydney

Semester 2, 2025



THE UNIVERSITY OF
SYDNEY

Learning objectives

You will:

1. Learn the concept of the t -test and its different forms: one-sample, two-sample, and paired t -tests.
2. Be able to model the different forms of the t -test using the general linear model (GLM) framework.
3. Understand that only certain assumptions need to be checked depending on the type of t -test used, and how to check them.
4. Be able to interpret the results of the t -test using the GLM framework or the traditional approach.

Comparing means



Are male Gentoo penguins heavier than female penguins?

Most baby foods fail health test, make dodgy claims

By [Alex Mitchell](#)

Updated August 13 2024 - 3:31am, first published 3:30am



Most baby and toddler foods in Australia fail to meet nutritional guidelines, according to a study. Photo: Bianca De Marchi/AAP PHOTOS

Most Australian baby and toddler foods fail to meet international nutritional guidelines while featuring dodgy health claims in their marketing.

Source

Is the sugar content of this particular baby food product different from that advertised?

TL;DR

- Often, we see patterns that clearly show differences from expectations, or between groups.
- The **t-test** allows us to quantify these differences and determine if they are **statistically significant**.
- Specifically, the test compares **two things**:
 - ➡ a group to a known value,
 - ➡ **two** groups to each other,
 - ➡ **two related** groups to each other...
- Answering different questions:
 1. Is the mean of a sample different from the *mean* of another that has unknown variance? **One-sample t-test**
 2. Are the means of two groups different from each other? **Two-sample t-test**
 3. Are two *related* measurements different from each other when we look at their *mean difference*? **Paired t-test**

But really, these are general linear models (GLMs) but with either no predictor or binary predictors.

History and context



William Sealy Gosset (1876-1937)

Fisher would have discovered it anyway...

– *on his invention of the t -statistic and the t -distribution*

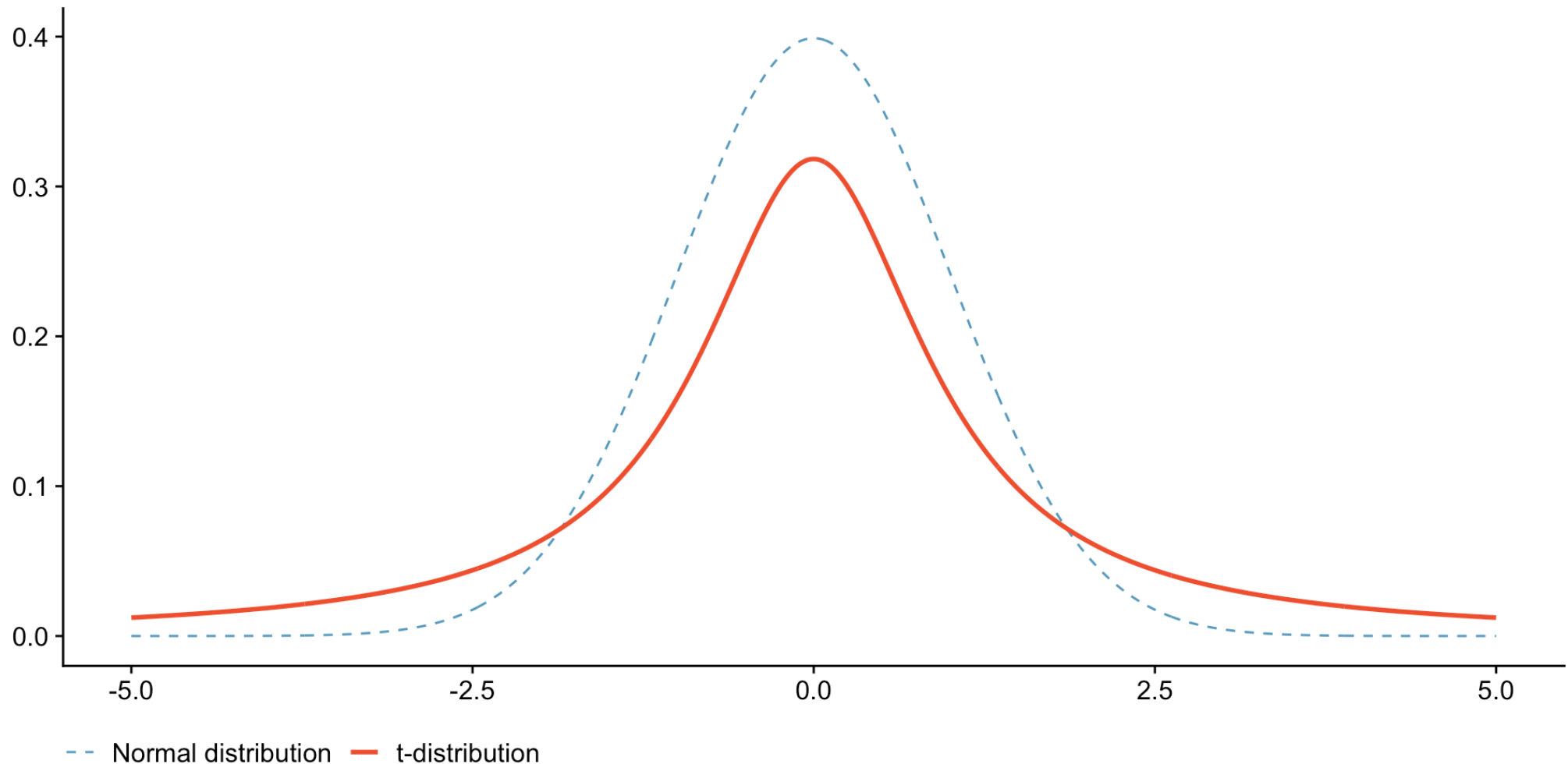
- A statistician at Guinness Brewery.
- Developed the t -test to monitor the quality of stout (beer) – **by comparing barley yields.**
- Friends with both Karl Pearson and Roland Fisher (you know [the story](#))

Speed-run: the t -distribution

What is the t -distribution?

A [probability distribution](#) that is similar to the normal distribution, but with **heavier tails**.

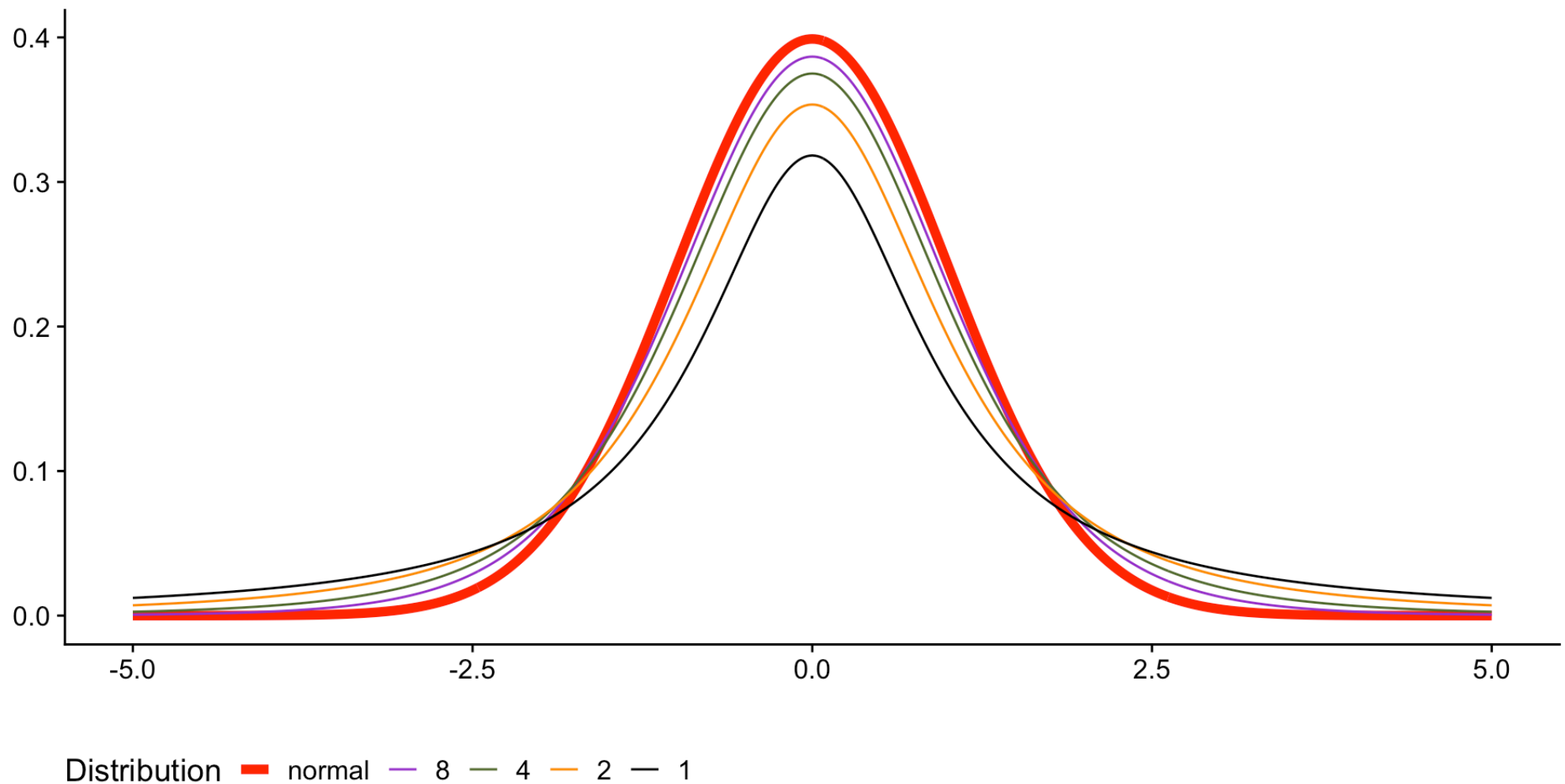
► Code



Sample size and the t -distribution

With increasing sample size, the t -distribution approaches the normal distribution.

► Code



How do we use the t -distribution?

The t -statistic: A signal-to-noise ratio

The test calculates a single number called a **t -statistic**. It's best to think of this as a “signal-to-noise” ratio.

- The signal: the difference between the average results of your two groups. **A bigger difference means a stronger signal**
- The noise: the amount of variation within each group

To put this into an equation, it looks like this:

$$t = \frac{\text{signal}}{\text{noise}}$$

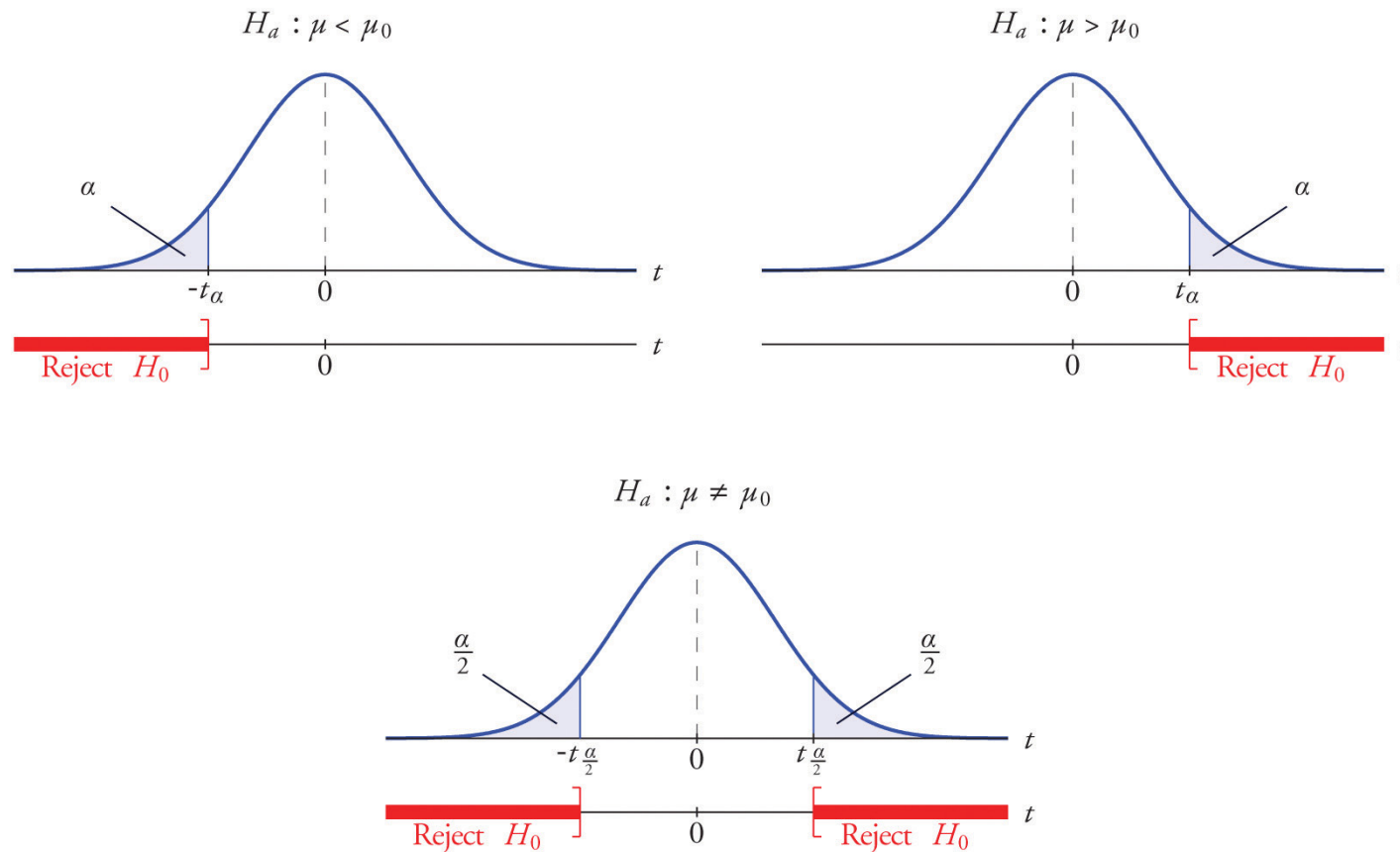
Mathematically, the t -statistic could be calculated as (for two independent samples):

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

A *large* t -statistic tells you that the signal is strong and clear, standing out above the background noise (the natural variation).

The traditional t -test

1. First, check if the data meets the assumptions of the test
2. Calculate a test statistic, t
3. Compare the t value to the t -distribution
4. If the value falls within critical (shaded) regions, conclude that there is a significant difference



In practice...

It's more simple than that. Simply identify the two groups of interest and most modern software will do the rest.

One-sample *t*-test

Are penguin bill lengths different from 40 mm?

► Code

One Sample t-test

```
data:  penguins$bill_length_mm
t = 13.285, df = 341, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 40
95 percent confidence interval:
 43.34125 44.50261
sample estimates:
mean of x
 43.92193
```

Two-sample *t*-test

Is bill length different between Male and Female penguins?

► Code

Two Sample t-test

```
data:  bill_length_mm by sex
t = -6.667, df = 331, p-value = 1.094e-10
alternative hypothesis: true difference in means
between group female and group male is not equal to 0
95 percent confidence interval:
 -4.866557 -2.649027
sample estimates:
mean in group female    mean in group male
      42.09697           45.85476
```

Modelling the t -test

Model-centric approach

The t -test is a special case of the GLM because it can be considered a linear model, but the predictors are **binary** (two categories in a two-sample t -test), or **absent** (one-sample t -test or paired t -test) rather than continuous.

So, given that

$$\text{response} \sim \text{predictor}$$

then for the **two-sample t -test**, the model is

$$\text{response} = \beta_0 + \beta_1 \cdot \text{predictor} + \epsilon$$

and for a **one-sample or paired t -test**, the model is

$$\text{response} = \beta_0 + \epsilon$$

Essentially, we need to understand that β_1 is the difference between the two groups, and β_0 is the mean of the first group.

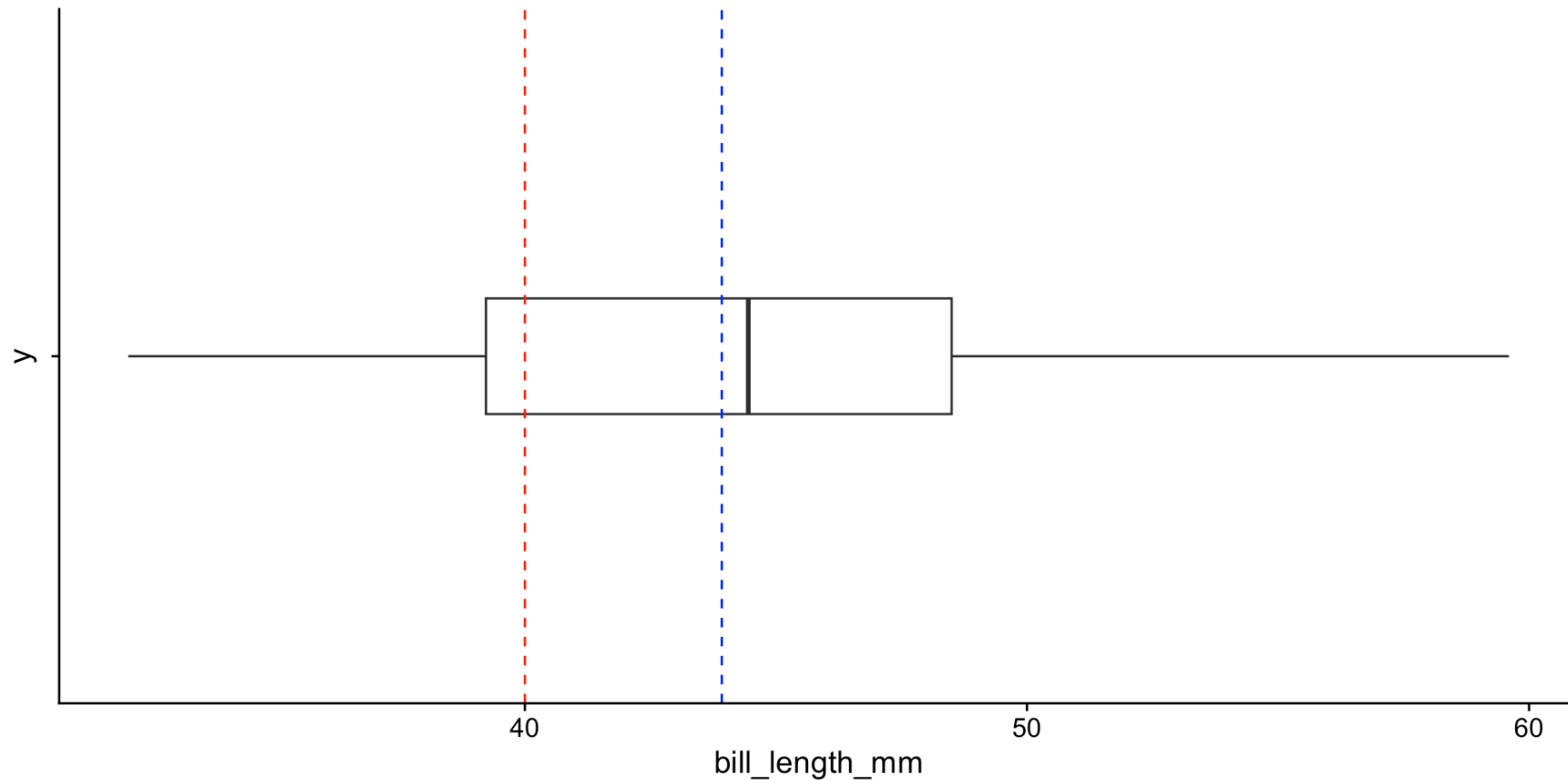
One-sample t -test

The question

Are mean penguin bill lengths in my sample (blue line) different from 40 mm (red line)?

One-sample *t*-test

► Code



Modelling the one-sample t -test

- **Response:** bill length
- **Predictor:** none, instead we test whether the response is equal to a known value (intercept).

The model

Bill length is defined by a mean of 40 mm:

$$\text{bill length} \sim 40$$

Modelling the one-sample *t*-test

We generally need to subtract the known value from the response variable such that:

$$\text{bill_length_mm} - 40 \sim 0$$

In R

► Code

```
Call:
lm(formula = bill_length_mm - 40 ~ 1, data = penguins)

Residuals:
    Min       1Q   Median       3Q      Max
-11.8219  -4.6969   0.5281   4.5781  15.6781

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.9219     0.2952   13.29  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.46 on 341 degrees of freedom
(2 observations deleted due to missingness)
```


Modelling the one-sample t -test

We generally need to subtract the known value from the response variable such that:

$$\text{bill_length_mm} - 40 \sim 0$$

In Jamovi

You will need to double-click on the variable name and transform it to `$source-40`, then use the new variable in the analysis.



+ Add recode condition

f_x = \$source-40

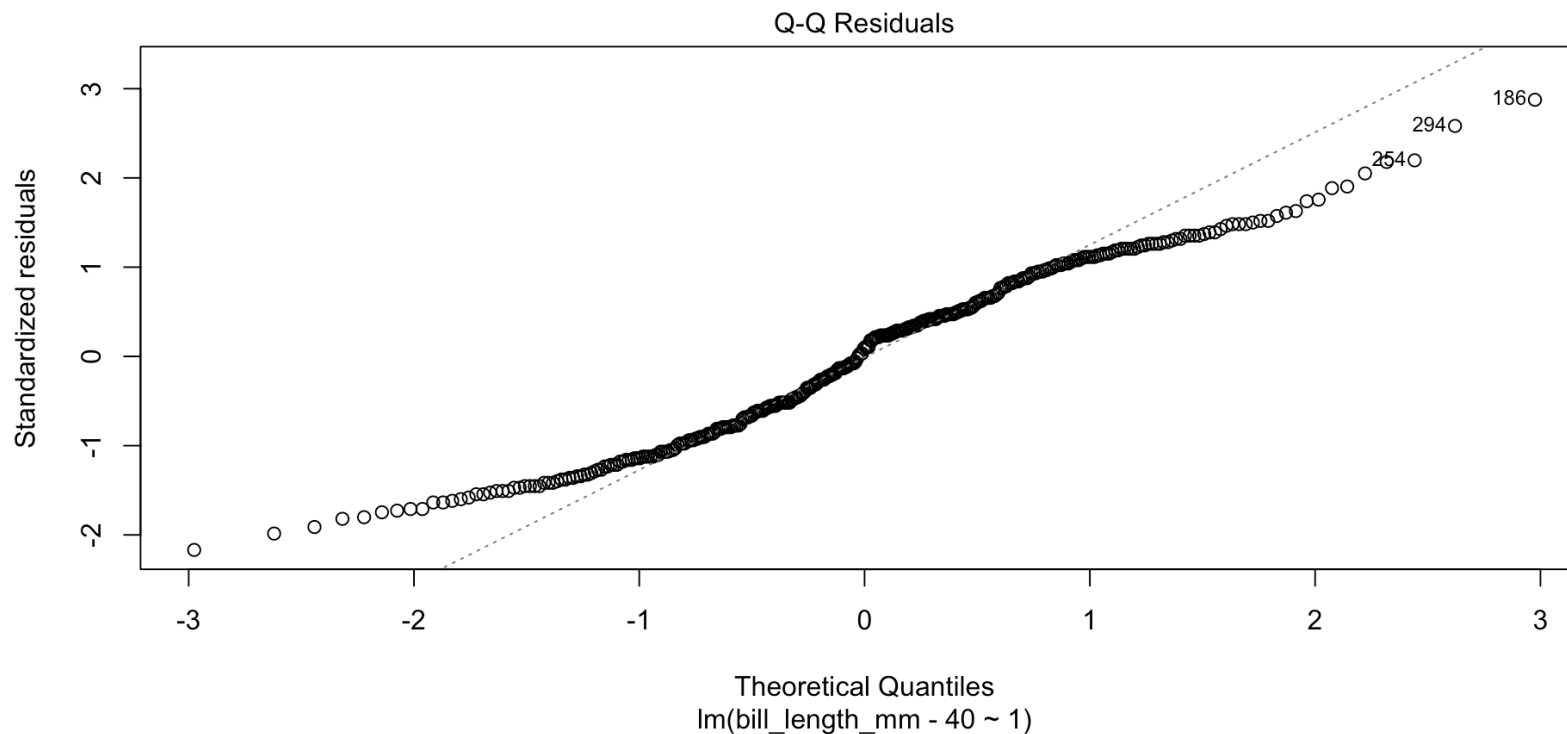
! Important

The one-sample t -test is also one of the few cases where it is probably easier to use the traditional approach. So think of this exercise as a way to understand the GLM approach (and proof that it applies).

Assumptions of the one-sample t -test

For the one-sample t -test, the only assumption is that the residuals are normally distributed. Under the GLM framework, we can check this by looking at the QQ-plot of the residuals.

► Code

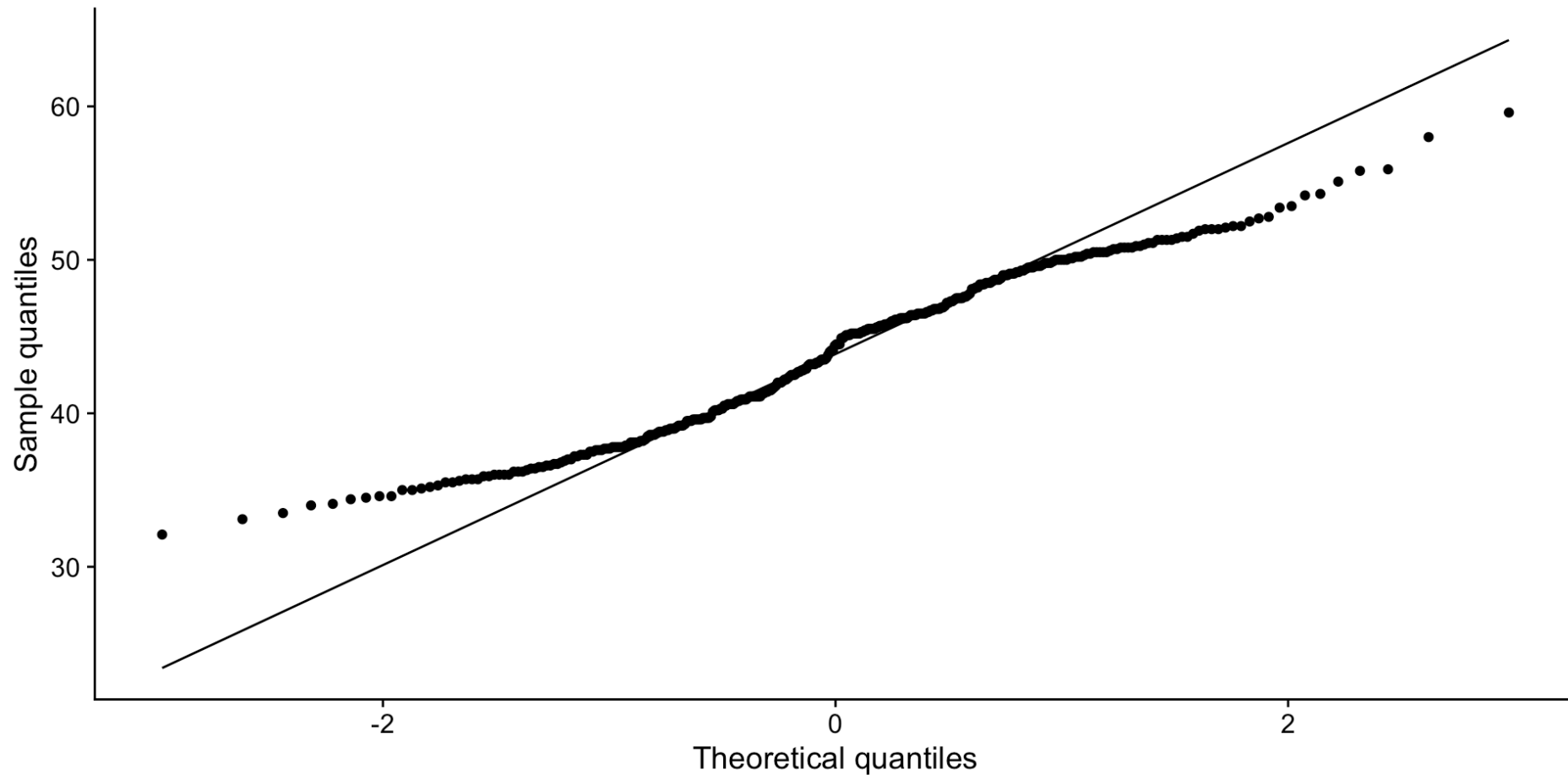


The residuals show that the data does not meet the assumption of normality, but the model is robust to this violation as long as the sample size is large and the samples are representative of the population.

Assumptions of the one-sample t -test

We can also just plot the raw data in the traditional way to check for normality, since there is only one group of data to plot.

► Code



The patterns are identical since there are no predictors in the model.

Interpretation

► Code

```
One Sample t-test

data:  penguins$bill_length_mm
t = 13.285, df = 341, p-value < 2.2e-16
alternative hypothesis: true mean is not
equal to 40
95 percent confidence interval:
 43.34125 44.50261
sample estimates:
mean of x
 43.92193
```

► Code

```
Call:
lm(formula = bill_length_mm ~ 40, data = penguins)

Residuals:
    Min       1Q   Median       3Q      Max
-11.8219  -4.6969   0.5281   4.5781  15.6781

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.9219     0.2952   13.29  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.46 on 341 degrees of freedom
(2 observations deleted due to missingness)
```

Notice the similarities and differences between the two approaches.

Reporting

Methods

A **one-sample t-test** was conducted to determine whether the mean bill length of penguins significantly differed from a hypothesised mean of 40 mm. The analysis was conducted using R version 4.4.0 (R Core Team, 2024) with the `t.test()` function. The assumption of normality was checked visually by examining the QQ-plot of the raw data. This assumption was not met, but the *t*-test was considered robust due to the large sample size and representativeness of the data and was therefore reported.

Results

The analysis revealed a statistically significant difference, with the mean bill length (43.9 mm) being significantly greater than 40 mm ($t = 13.29$, $df = 341$, $p < 0.01$). The 95% confidence interval for the mean bill length was [43.34, 44.50] mm.

Reporting

Methods

A **general linear model (GLM)** ~~equivalent to a one-sample t-test~~ was performed to determine whether the mean bill length of penguins differed from a hypothesised value of 40 mm. The analysis was conducted using R version 4.4.0 (R Core Team, 2024), with the model specified as:

$$\text{bill_length_mm} - 40 \sim 1$$

where the intercept represents the difference between the mean bill length and 40 mm. The assumption of normality was checked visually by examining the QQ-plot of the residuals. This assumption was not met, but the model was considered robust due to the large sample size and representativeness of the data.

Results

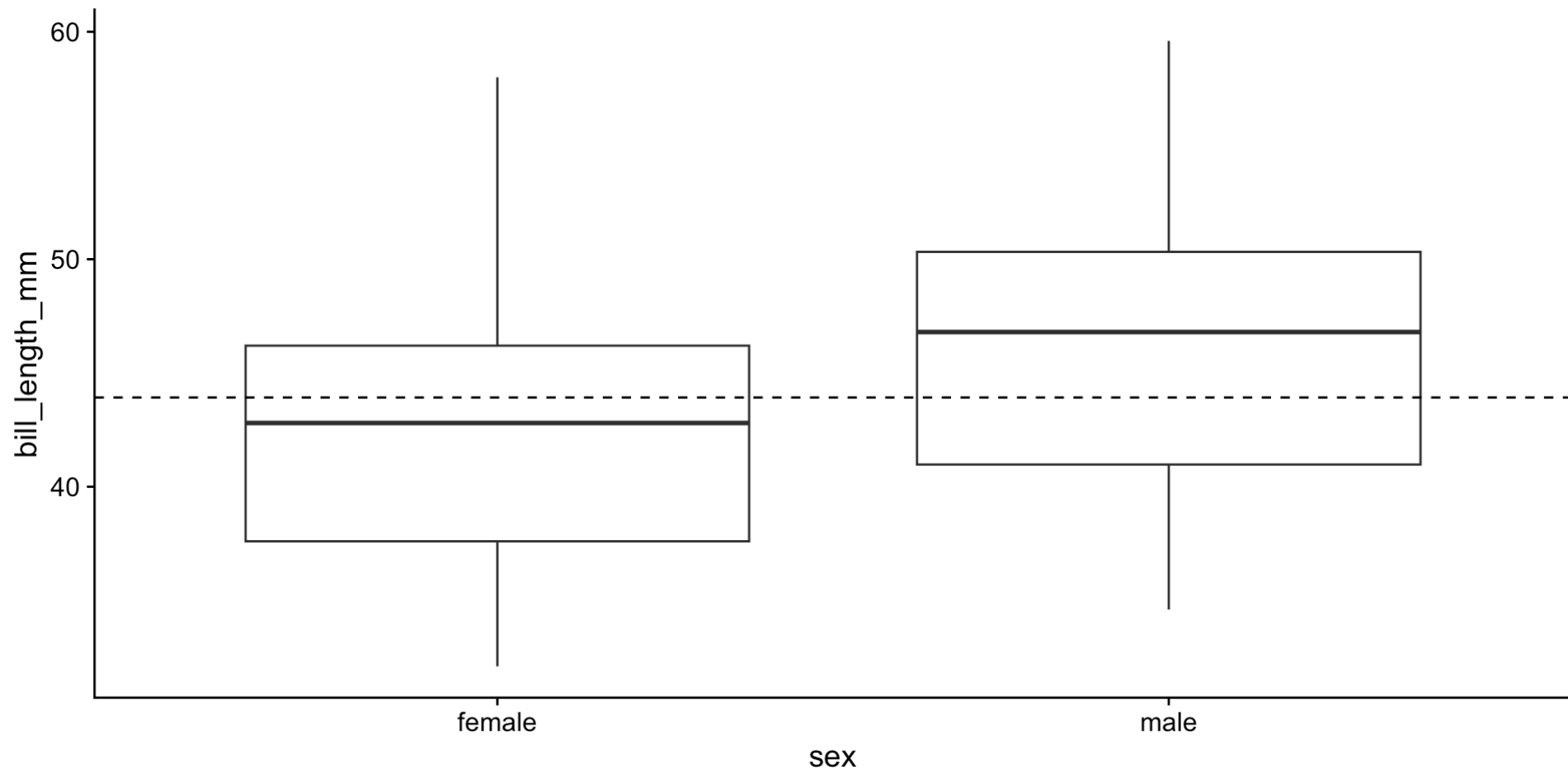
The results indicated that the mean difference in bill length from 40 mm was statistically significant. The mean bill length was estimated to be 3.92 mm greater than 40 mm, with a standard error of 0.30 mm (GLM, $t = 13.29$, $df = 341$, $p < 0.001$).

Two-sample t -test

The question

Are the mean bill lengths between male and female penguins different? *If they are not, then both groups should have the same mean (dotted line).*

► Code



Modelling the two-sample t -test

- **Response:** bill length
- **Predictor:** sex

The model

Bill length is a function of sex:

$$\text{bill length} \sim \text{sex}$$

Modelling the two-sample *t*-test

The model specification is more straightforward since a predictor is present:

$$\text{bill_length_mm} \sim \text{sex}$$

► Code

```
Call:
lm(formula = bill_length_mm ~ sex, data = penguins)

Residuals:
    Min       1Q   Median       3Q      Max
-11.2548  -4.7548   0.8452   4.3030  15.9030

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  42.0970     0.4003 105.152  < 2e-16 ***
sexmale       3.7578     0.5636   6.667 1.09e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

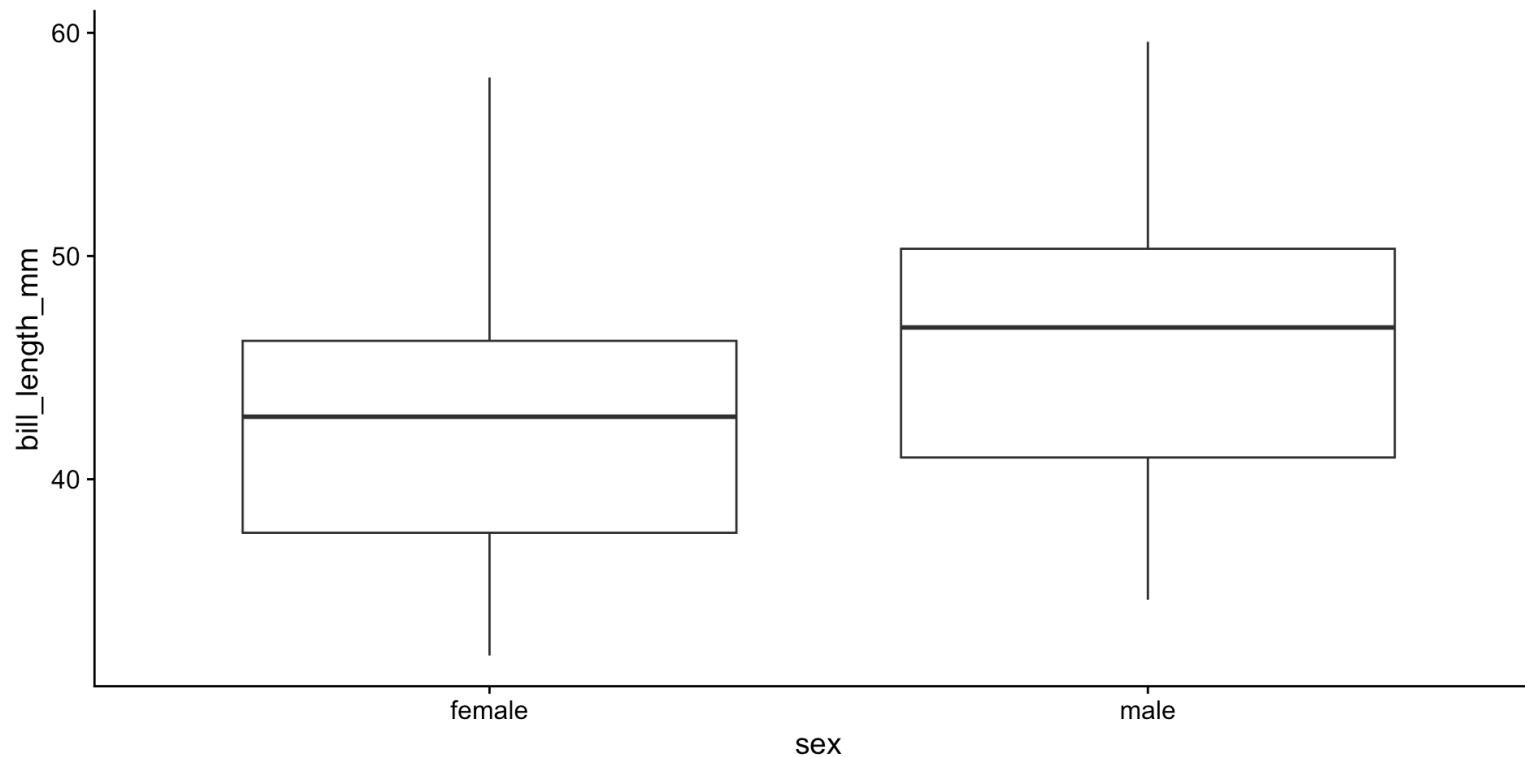
Residual standard error: 5.143 on 331 degrees of freedom
(11 observations deleted due to missingness)
Multiple R-squared:  0.1184,    Adjusted R-squared:  0.1157
F-statistic: 44.45 on 1 and 331 DF,  p-value: 1.094e-10
```

Assumptions of the two-sample t -test: LINE

If we were to use the traditional t -test approach

- Divide the data into the two groups – male and female
- Check the normality and equal variances of the residuals for each group

► Code

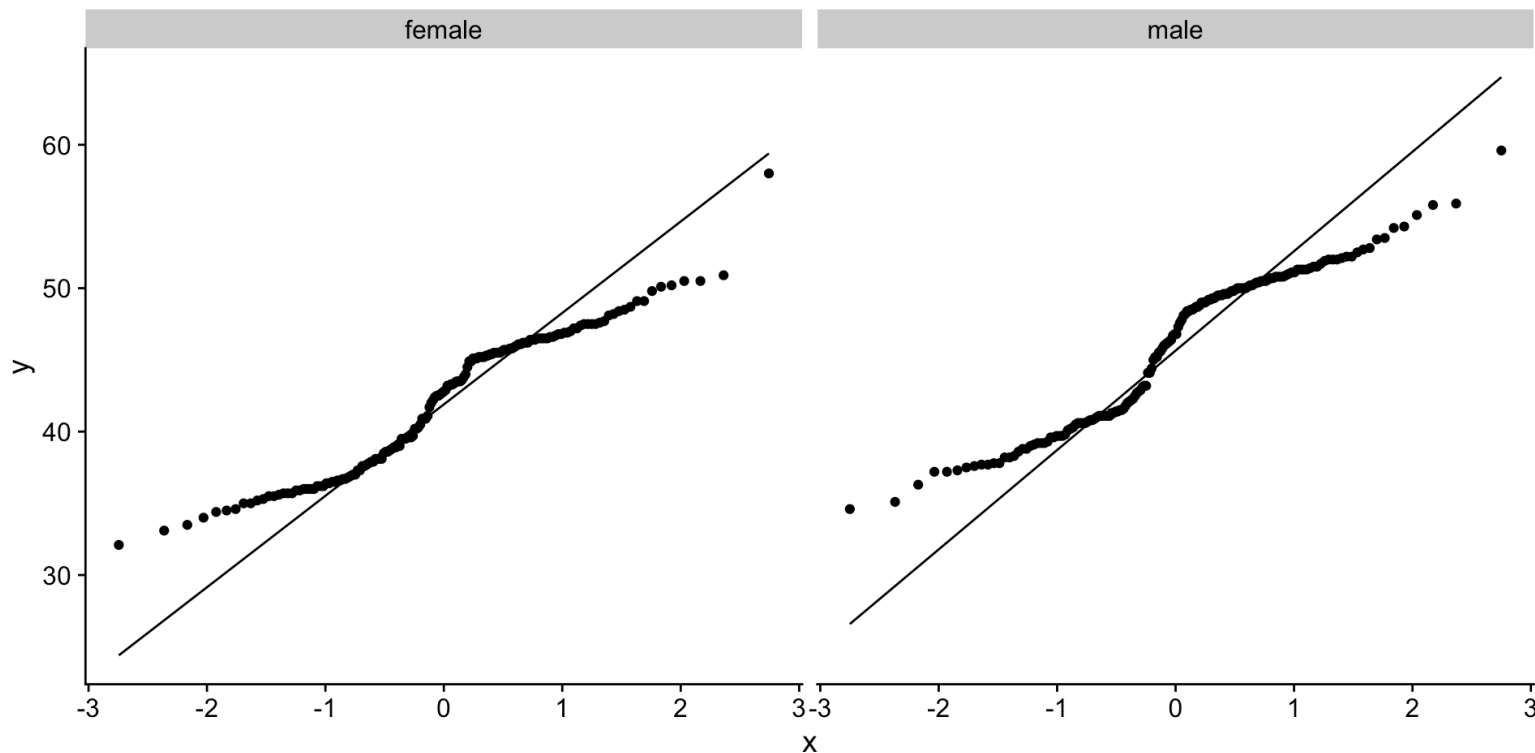


Assumptions of the two-sample t -test: LINE

If we were to use the traditional t -test approach

- Divide the data into the two groups – male and female
- Check the normality and equal variances of the residuals for each group

► Code



Assumptions of the two-sample t -test: LINE

GLM approach

View the residuals of the model – no need to divide the data

► [Code](#)

Interpretation

► Code

```
Two Sample t-test

data:  bill_length_mm by sex
t = -6.667, df = 331, p-value = 1.094e-10
alternative hypothesis: true difference in
means between group female and group male
is not equal to 0
95 percent confidence interval:
 -4.866557 -2.649027
sample estimates:
mean in group female    mean in group male
      42.09697           45.85476
```

► Code

```
Call:
lm(formula = bill_length_mm ~ sex, data = penguins)

Residuals:
      Min       1Q   Median       3Q      Max
-11.2548  -4.7548   0.8452   4.3030  15.9030

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  42.0970     0.4003  105.152  < 2e-16 ***
sexmale       3.7578     0.5636   6.667 1.09e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.143 on 331 degrees of freedom
(11 observations deleted due to missingness)
Multiple R-squared:  0.1184,    Adjusted R-squared:  0.1157
F-statistic: 44.45 on 1 and 331 DF,  p-value: 1.094e-10
```

Suggestion: paste the results into ChatGPT and ask for a summary.

What if the assumptions are violated?

1. **Normality:** the t -test is robust to violations of normality, especially with large sample sizes. However we can consider transforming the data or using a non-parametric test such as the Wilcoxon rank-sum test if the data is highly skewed.
2. **Equal variances:** transforming the data can help, but we can consider Welch's t -test, which does not assume equal variances, as a viable alternative. Note: there is a [GLM approach](#) to this but it is complex, so we will not cover it here.

Paired t -test

First, about related (or paired) samples

- **Related samples** are those where the measurements are taken from the same individual or object at different times or under different conditions:
 - ➡ before and after treatment,
 - ➡ left and right eyes,
 - ➡ two different methods of measurement of the same subject, etc.
- This causes the measurements to be **dependent** on each other.
- There is correlation or covariance between the measurements, therefore we use the difference between the measurements as the response variable rather than the measurements separately.

! Important

This is one of the few cases where the **independence** assumption matters and can result in using a paired t -test instead of a two-sample t -test.

The question

Is sleep duration different before and after treatment? [Student's sleep data](#) – the same 10 patients were observed for two weeks, with the first week being a control week and the second week being a treatment week.

Modelling the question

- **Response:** sleep duration
- **Predictor:** week

The model

Sleep duration is a function of the week:

sleep duration \sim week

Modelling the paired *t*-test

We need to subtract the sleep duration in the control week from the sleep duration in the treatment week such that:

$$\text{sleep duration (treatment)} - \text{sleep duration (control)} \sim 0$$

► Code

```
Call:
lm(formula = treatment - control ~ 1)

Residuals:
    Min       1Q   Median       3Q      Max
-1.58  -0.53  -0.28   0.12   3.02

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    1.580      0.389    4.062  0.00283 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.23 on 9 degrees of freedom
```

Assumptions of the paired t -test

Because the data is paired, variances should be equal and do not need to be checked, although unequal variances highlight possible issues with the data.

Otherwise, only the normality assumption needs to be checked.

We will not check the assumptions here, since the process remains the same.

Further thinking

- The GLM approach means that the model *just so happens* to match a *particular* traditional test.
- Interpretation of the model can change as soon as the **structure** of any of the variables change.

Example: *t*-test vs. ANOVA

The two-sample *t*-test can be extended to a **one-way ANOVA** if we have more than two groups to compare. The model remains the same, but the interpretation changes!

$$\text{response} = \beta_0 + \beta_1 \cdot \text{predictor} + \epsilon$$

We will cover this in great detail next week.

Questions to consider

- What are the assumptions of the one-sample t -test, two-sample t -test, and paired t -test? Are they the same?
- How do you interpret the results of a GLM compared to a traditional t -test?
- What are the advantages and disadvantages of using the GLM approach over the traditional approach?

Thanks

This presentation is based on the [SOLES Quarto reveal.js template](#) and is licensed under a [Creative Commons Attribution 4.0 International License](#). A pdf version of this document can be found [here](#).