Two-sample *t*-test in R

Cheatsheet

2024-08-13

ℹ License

ℹ Assumed knowledge

- You know how to install and load packages in R.
- You know how to import data into R.
- You recognise data frames and vectors.

! Data structure

The data should be in a **long format** (also known as tidy data), where each row is an observation and each column is a variable (Figure 1). If your data is not already structured this way, reshape it manually in a spreadsheet program or in R using the `pivot_longer()` function from the `tidyr` package.

Sex	BW
F	2.15
M	2.55
F	2.95
F	2.70
M	2.20
F	1.85
M	2.55
M	2.60

F	M
2.15	2.55
2.95	2.20
2.70	2.55
1.85	2.60

Figure 1: Data should be in long format (left) where each row is an observation and each column is a variable. This is the preferred format for most statistical software. Wide format (right) is also common, but may require additional steps to analyse or visualise in some instances.

! Data

For this cheatsheet we will use data from the penguins dataset from the `palmerpenguins` package. You may need to install this package:

```r
install.packages("palmerpenguins")
```

About

The two-sample *t*-test is used to determine whether the means of two independent groups are significantly different from each other. **Examples**:

- **Comparing plant growth:** Is the mean height of plants in a shaded area significantly different from those in a sunlit area?
- **Species abundance:** Are the mean numbers of a specific bird species found in two different forest types significantly different?
- **Soil nutrient levels:** Is the mean nitrogen concentration in soil samples from an agricultural field significantly different from that in a nearby natural forest?

Modelling

The model used in this cheatsheet is based on the `penguins` dataset.

> Is the bill length of penguins significantly different between male and female penguins?

The **simplified model** for the mathematically-adverse individual is

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

which translates to "the bill length of penguins is influenced by sex". The **statistical model** is

$$\text{bill length} = \beta_0 + \beta_1 \cdot sex + \epsilon$$

where β_0 is the mean of the first group in the `sex` variable (in this case, female), β_1 is the difference in means between the two groups, and ϵ is the error term.

Preparing the data

Extract **only** the variables of interest from the dataset using `select()` from the `dplyr` package. This makes it easier to work with the data, especially if cleaning is required. In this case, there are `NA` values in the `sex` column, so we remove them using `drop_na(sex)`.

```
library(palmerpenguins)
library(dplyr)
data(penguins)
df <- penguins |>
  select(bill_length_mm, sex) |>
  drop_na(sex)
```

Your own data should be in a similar format.

Analytical approaches

The traditional approach to the two-sample *t*-test is to use the `t.test()` function in R, while the modern approach is to use a general linear model (GLM) with the `lm()` or `glm()` functions.

`t.test()` function

Methods reporting

> Assumption checks for normality and homogeneity of variance were performed performed using the [insert check(s)] and [insert check(s)], respectively. Both groups

> were found to be normally distributed and have equal variance, so a two-sample *t*-test was used to determine whether the mean body weight of possums was significantly different from 3.5 kg. This was computed using the `t.test()` function in R version 4.4.0 (R Core Team, 2024).

Check assumption(s)

In the traditional approach, assumptions of normality and homogeneity of variance are checked using the raw data and can be done **before** performing the *t*-test.

> ℹ Note
>
> Assumption checks are not required to be reported in the results, but it is good practice to include them in the methods section and indicate how they were checked.

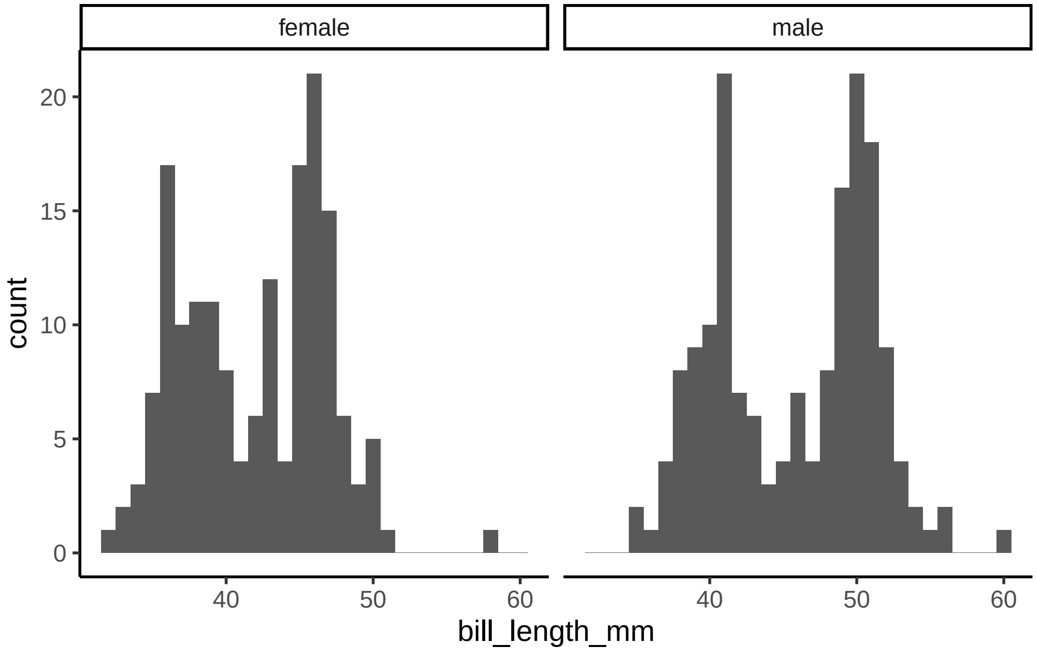
Normality

Any of the following methods can be used to check for normality. It is important to note that using raw data to check for assumptions requires each group (i.e. sex) to be checked **separately**.

Histogram

The histogram shows that the data is not normally distributed as it is multimodal.

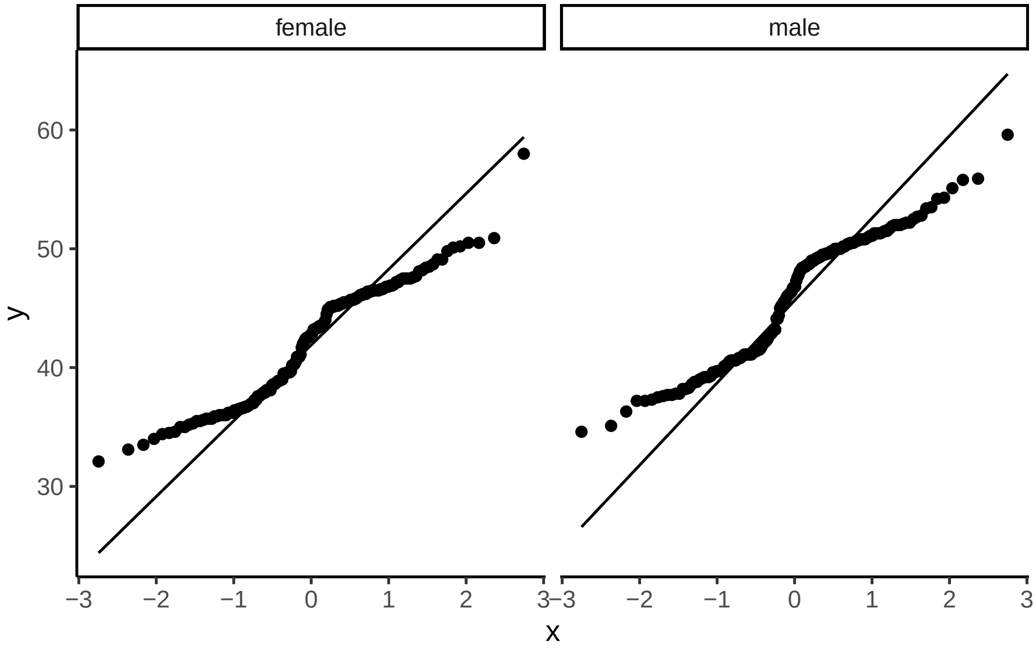
```r
library(ggplot2)
ggplot(df, aes(x = bill_length_mm)) +
  geom_histogram(binwidth = 1) +
  facet_wrap( ~ sex) +
  theme_classic()
```



Q-Q plot

The Q-Q plot shows that the data is not normally distributed as many points do not fall on the line. The S-shaped curvature indicates underdispersion.

```
ggplot(df, aes(sample = bill_length_mm)) +
  stat_qq() +
  stat_qq_line() +
  facet_wrap( ~ sex) +
  theme_classic()
```



Shapiro-Wilk test

This test is not recommended for large sample sizes (> 50) as it is sensitive to even the smallest deviations from normality. Nevertheless, if you choose to use it, the results show that the data is not normally distributed in either group since the p-value is less than 0.05.

```
df |>
  filter(sex == "male") |>
  pull(bill_length_mm) |>
  shapiro.test()

df |>
  filter(sex == "female") |>
  pull(bill_length_mm) |>
  shapiro.test()
```

Equal variance

Eyeballing a boxplot of the data can be used to check for homogeneity of variance. The `var.test()` and `leveneTest()` functions can also be used to check for homogeneity of variance, with the latter being more robust. **In general, it is preferable to not rely on the results of these formal tests if the samples are not normally distributed.**

```
#
var.test(bill_length_mm ~ sex, df)

# Levene's test
library(car)
leveneTest(bill_length_mm ~ sex, df)
```

Perform the analysis

These days, irregardless of whether the data meets the assumptions of equal variance, or not, the Welch's *t*-test is recommended since it will provide almost identical results to the traditional *t*-test if equal variance is met anyway.

```
t.test(bill_length_mm ~ sex, data = df)
```

How to report results

> The mean bill length of penquins was significantly different between males and females (Welch's t_{329} = -6.67, $p < 0.001$). Mean male bill length (45.9 mm) is higher than mean female bill length (42.1 mm).

`lm()` function

Methods reporting

> A general linear model was used to determine whether the bill length of penguins is significantly different between the two sexes by fitting the model `bill length ~ sex`. This was computed using the `lm()` function in R version 4.4.0 (R Core Team, 2024). The assumptions of normality and homogeneity of variance were checked visually using the residuals of the model. Normality was not met but the analysis was still performed using a GLM as the sample size was large enough to assume that the sampling distribution of the mean is approximately normal.

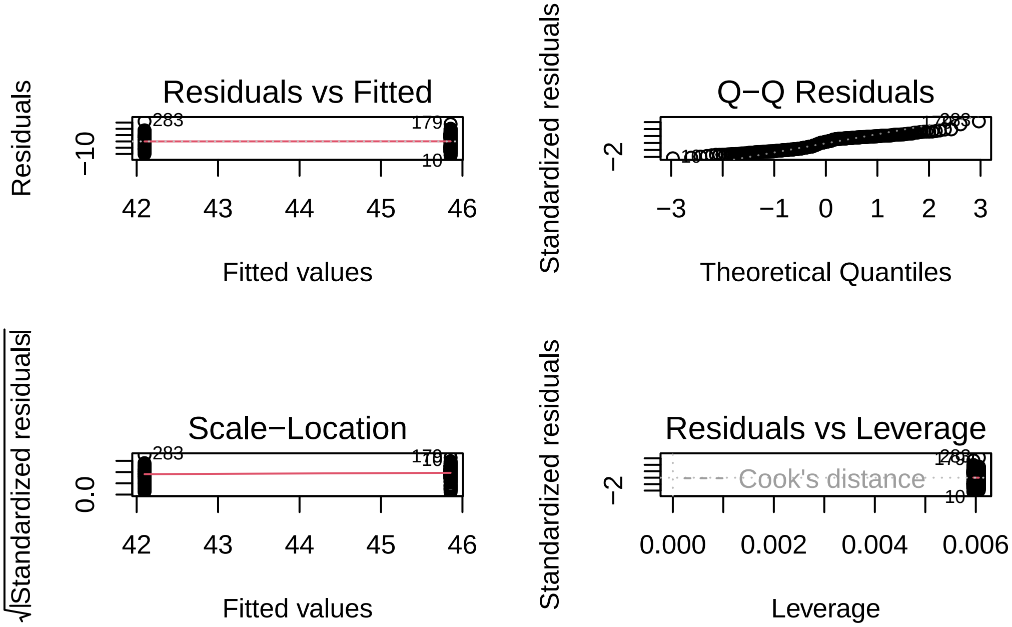
Perform the analysis

```
fit <- lm(bill_length_mm ~ sex, data = df)
summary(fit)
```

Check assumption(s)

With a GLM, assumptions of normality and homogeneity of variance are checked using the residuals of the model. A `plot()` function can be used to check these assumptions.

```
par(mfrow = c(2, 2))
plot(fit)
```



How to report results

> There is evidence to suggest that the mean bill length of penguins is significantly different between sexes (GLM, F~1, 331~ = 44.5, p < 0.001). Male penguins had a higher mean bill length than females (mean difference = 3.76 mm).