Lab 06 – Two-sample *t*-test

ENVX1002 Handbook

Semester 1, 2025

Learning outcomes

- Learn to use R to calculate a 2-sample *t*-test
  - independent samples with constant variance
  - independent samples with unequal variance
  - paired samples
  - data transformations
- Apply the steps for hypothesis testing from lectures
- Learn how to interpret statistical output

Before you begin

Create your Quarto document and save it as `Lab-06.Rmd` or similar. The following data files are required:

1) Barley.csv
2) Plant_growth.csv
3) Turbidity.csv

The following external packages are used in this lab. Install them if you have not done so already.

```
install.packages(c("tidyverse", "car"),
  repo = "https://cloud.r-project.org")
```

Finally, try to complete today's lab exercises in pairs and try out pair programming, where one person writes the code and the other person reviews each line as it is written. You can swap roles every 10 minutes or so. This is a great way to learn from each other and to improve your coding skills.

Exercise 1: barley (walk-through)

Background

An experiment was designed to compare two varieties of spring barley. Thirty four plots were used, seventeen being randomly allocated to variety A and seventeen to variety B. Unfortunately

five plots were destroyed. The yields (t/ha) from the remaining plots were as they appear in the file `Barley.csv`.

Instructions

First, quickly explore the data; then, utilise the **HATPC** process and test the hypothesis that the two varieties give equal yields, assuming that the samples are independent.

HATPC:

- Hypothesis
- Assumptions
- Test (statistic)
- P-value
- Conclusion

> **i** Level of significance
>
> **The level of significance is usually set at 0.05.** This value is generally accepted in the scientific community and is also linked to Type 2 errors, where choosing a lower significance increases the likelihood of failing to reject the null hypothesis when it is false.

Data exploration

First we load the data and inspect its structure to see if it needs to be cleaned or transformed. The `glimpse()` function is a tidy version of `str()` that provides a quick overview of the data that focuses on the variables, ignoring data attributes.

Try to compare `str()` and `glimpse()` to see what the differences are.

```
barley <- readr::read_csv("data/Barley.csv") # packagename:: before a function lets you access a function without having to load the whole library first
```

```
Rows: 29 Columns: 2
── Column specification ────────────────────────────────────────────────────────
Delimiter: ","
chr (1): Variety
dbl (1): Yield

ℹ Use `spec()` to retrieve the full column specification for this data.
ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
dplyr::glimpse(barley)
```

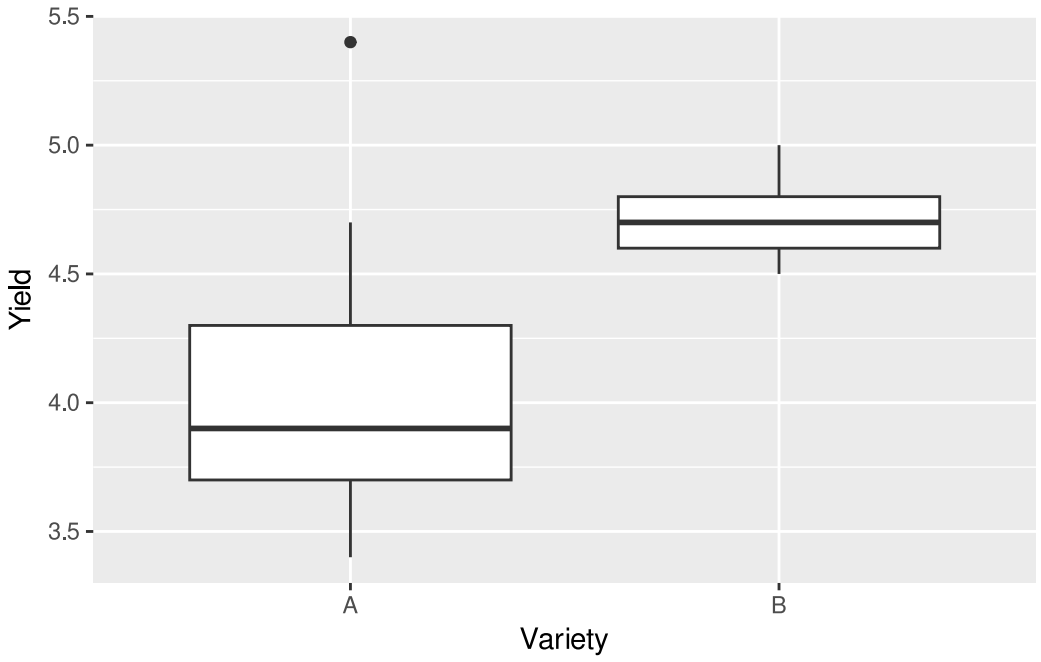
```
Rows: 29
Columns: 2
$ Variety <chr> "A", "A", "A", "A", "A", "A", "A", "A", "A", "A", "A", "A", "A…
$ Yield   <dbl> 4.6, 4.3, 3.8, 3.4, 3.9, 3.9, 3.9, 4.4, 3.6, 3.6, 4.7, 3.9, 3.…
```

The `Variety` column is a factor with two levels, `A` and `B`, but it is defined as a character. We can convert it to a factor using the `mutate()` function from the `dplyr` package, but it is not necessary for the *t*-test since R will automatically convert it to a factor.

```
library(tidyverse)
barley <- mutate(barley, Variety = as.factor(Variety))
```

Quickly preview the data as a plot to see if there are any trends or unusual observations.

```
barley %>%
  ggplot(aes(x = Variety, y = Yield)) +
  geom_boxplot()
```



A trained eye will anticipate that the data may not meet the assumption of equal variance; however, we will test this assumption later. Otherwise, there appear to be no unusual observations in the data.

Hypothesis

What are the null and alternative hypotheses? We can use the following notation:

$$H_0 : \mu_A = \mu_B$$

$$H_1 : \mu_A \neq \mu_B$$

where μ_A and μ_B are the population means of varieties A and B, respectively.

It is **important** that when using mathematical symbols to denote the null and alternative hypotheses, you should **always** define what the symbols mean. Otherwise, the reader may not understand what you are referring to.

The equations above are written in LaTeX, a typesetting system that is commonly used in scientific writing. You can learn more about LaTeX here. The raw syntax used to write the equations are shown below:

```
$$H_0: \mu_A = \mu_B$$
$$H_1: \mu_A \neq \mu_B$$
```

Why do we always define the null and alternative hypotheses? In complex research projects or when working in a team, it is important to ensure that everyone is on the same page. By defining the hypotheses, you can avoid misunderstandings and ensure that everyone is working towards the same goal as the mathematical notation is clear and unambiguous.

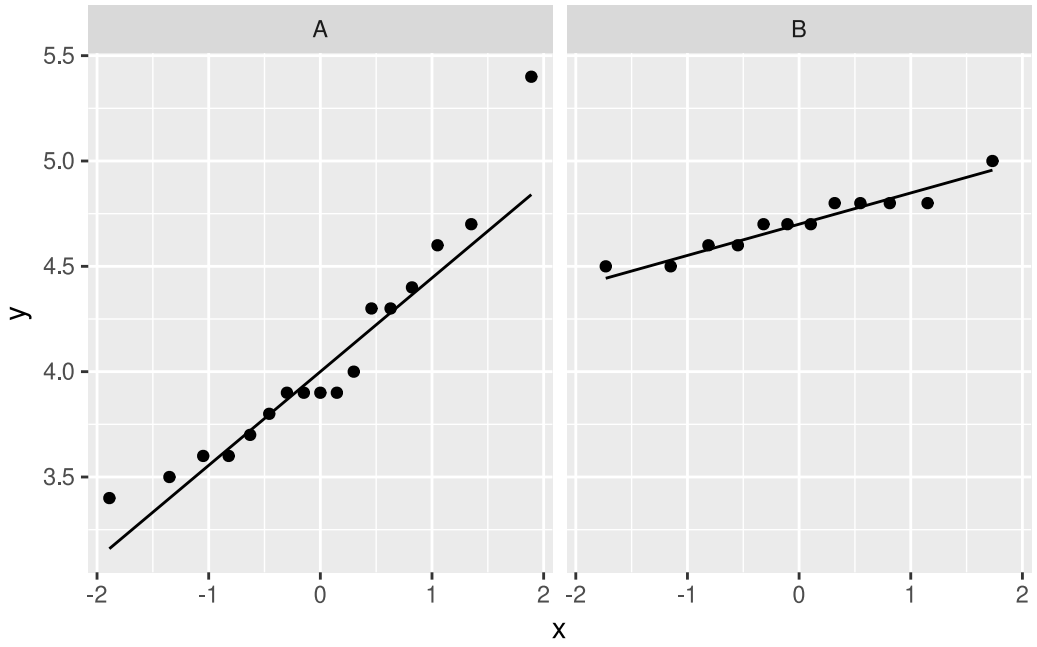
Assumptions

Normality

There are many ways to check for normality. Here we will use the QQ-plot. Use of `ggplot2` is preferred (as a means of practice) but since we are just exploring data, base R functions are not a problem to use.

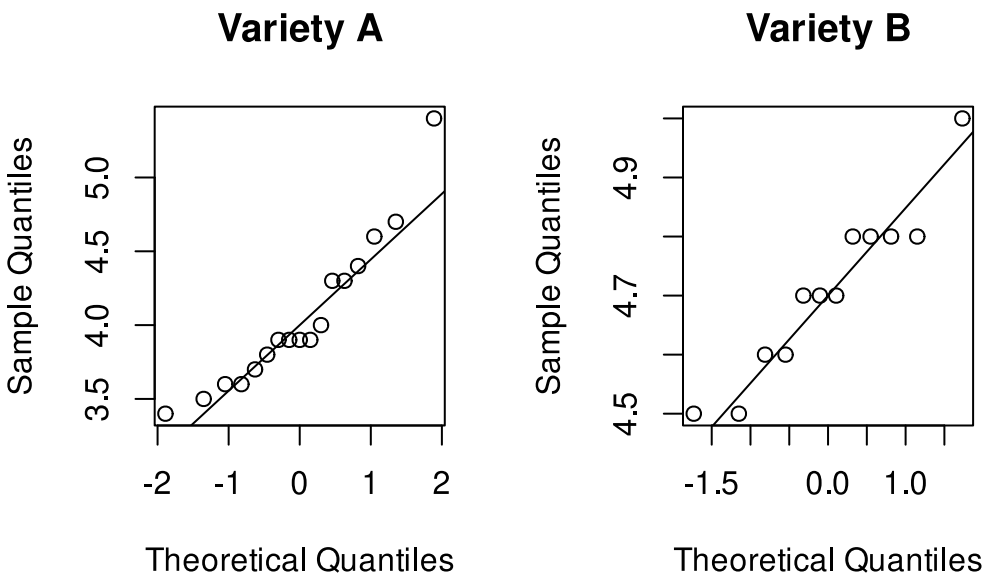
Using `ggplot2`

```
ggplot(barley, aes(sample = Yield)) +
  stat_qq() +
  stat_qq_line() +
   facet_wrap(~Variety) #facet_wrap ensures there are separate plots for each
variety rather than one plot with all the data in Yield
```



Using base R

```
par(mfrow = c(1, 2))
qqnorm(barley$Yield[barley$Variety == "A"], main = "Variety A") # square brackets
to subset the data by variety
qqline(barley$Yield[barley$Variety == "A"])
qqnorm(barley$Yield[barley$Variety == "B"], main = "Variety B")
qqline(barley$Yield[barley$Variety == "B"])
```



Question: Do the plots indicate the data are normally distributed?

Answer: Yes, the data appear to be normally distributed as the QQ-plot shows that the data points are close to the line.

Homogeneity of variance

From the boxplot, we can see that there is some indication that the variances are not equal. We can test this assumption using Bartlett's test or Levene's test; here we will just use Bartlett's test.

```
bartlett.test(Yield ~ Variety, data = barley)
```

```
    Bartlett test of homogeneity of variances

data:  Yield by Variety
Bartlett's K-squared = 14.616, df = 1, p-value = 0.0001318
```

Question: Does the Bartlett's test indicate the two groups have equal variances? What effect will that have on the analysis?

Answer: The two groups have unequal variance (Bartlett's test: $X^2 = 14.6, p < 0.01$). This means that we will need to use the Welch's t-test, which does not assume equal variances.

Test statistic

We can now calculate the test statistic using the `t.test()` function in R. Since the variances are unequal, we do not have to specify the `var.equal` argument – the default test for `t.test()` is the Welch's *t*-test which does not assume equal variances.

```
t.test(Yield ~ Variety, data = barley)
```

```
    Welch Two Sample t-test

data:  Yield by Variety
t = -4.9994, df = 19.441, p-value = 7.458e-05
alternative hypothesis: true difference in means between group A and group B is
not equal to 0
95 percent confidence interval:
 -0.9293569 -0.3814274
sample estimates:
mean in group A mean in group B
       4.052941        4.708333
```

P-value

Since the p-value is < 0.05, we can reject the null hypothesis that the mean yield of both varieties is equal.

Conclusion

The conclusion needs to be brought into the context of the study. In a scientific report or paper, you would write something like this:

> The mean yield of barley variety A was significantly different from that of variety B ($t = -5.0$, $df = 19.4$, $p < 0.01$).

Exercise 2: plant growth

Background

In a test of a particular treatment aimed at inducing growth, 20 plants were grouped into ten pairs so that the two members of each pair were as similar as possible. One plant of each pair was chosen randomly and treated; the other was left as a control. The increases in height (in centimetres) of plants over a two-week period are given in the file Two week plant heights. We wish to compare whether the treatment is actually inducing improved growth, as compared to the control.

Instructions

Here, we have two samples, and the samples are paired as it is a before-after experiment. So we'd like to conduct a paired *t*-test.

For paired t-tests the analysis is performed as a 1-sample t-test on the difference between each pair so the only assumption is the normality assumption.

Copy the structure below and perform your analysis in your document.

```
## Exercise 2: plant growth
### Data exploration
### Hypothesis
### Assumptions
#### Normality
#### Homogeneity of variance
### Test statistic
### P-value
### Conclusion
```

Note that the data is *not* tidy. The code below will convert the data to the long format and assign it to `tidy_plant`.

```
plant_growth <- readr::read_csv("data/Plant_growth.csv")
```

```
Rows: 10 Columns: 3
── Column specification ────────────────────────────────────────────────────────
Delimiter: ","
dbl (3): Pair, Treated, Control

ℹ Use `spec()` to retrieve the full column specification for this data.
ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
tidy_plant <- plant_growth %>%
  pivot_longer(cols = c(Treated, Control), names_to = "Treatment", values_to
= "Height")
```

You may also need to perform a Shapiro-Wilk test to check for normality. To do this for each group, you can use the `tapply()` function.

```
tapply(tidy_plant$Height, tidy_plant$Treatment, shapiro.test)
```

```
$Control

    Shapiro-Wilk normality test

data:  X[[i]]
W = 0.89588, p-value = 0.1973
```

```
$Treated

    Shapiro-Wilk normality test

data:  X[[i]]
W = 0.93139, p-value = 0.4617
```

Exercise 3: turbidity

Background

A new filtering process was installed at a dam which provided drinking water for a nearby town. To check on its success, a number of water samples were taken at random times and locations in the weeks before and after the process was installed. The following are the turbidity values (units = NTU) of the water samples.

Instructions

Now we consider further examples of a two-sample *t*-test, but where the assumption of equal variance and normality may not be met for the raw data. Sometimes after applying a data transformation the analysis can proceed assuming equal variances – but always check after a data transformation.

The data can be read with the code below:

```
turbidity <-read_csv("data/Turbidity.csv")
```

```
Rows: 19 Columns: 2
── Column specification ────────────────────────────────────────────────────────
Delimiter: ","
chr (1): Time
dbl (1): Turbidity

ℹ Use `spec()` to retrieve the full column specification for this data.
ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

For data transformation, you may need to create a new variable in your dataset to store the transformed data. For example, to create a new variable `TurbLog10` that stores the $\log_{10}$ transformed turbidity values, you can use the following code:

```
turbidity$TurbLog10 <- log10(turbidity$Turbidity)
```

To interpret the results for your conclusions, you may need to back-transform the mean and/or confidence interval values. To back transform $\log_{10}$ data you use:

$$10^{\text{mean or CI}}$$

To back-transform natural log, $\log_e$, you use:

$$e^{\text{mean or CI}}$$

Bonus take home exercises

Use the HATPC framework to answer the following questions

1. Hypothesis
2. Assumptions
3. Test (statistic)
4. P-value
5. Conclusion

Exercise 1: Tooth growth

Tooth GRowth is an inbuilt dataset that shows the effect of vitamin c in Guinea pig tooth growth. It has three variables:

- len = tooth length
- supp = type of supplement (Orange juice or ascorbic acid)
- dose = mg/day gieven to the guinea pigs

```
head(ToothGrowth)
```

```
   len supp dose
1  4.2   VC  0.5
2 11.5   VC  0.5
3  7.3   VC  0.5
4  5.8   VC  0.5
5  6.4   VC  0.5
6 10.0   VC  0.5
```

```
str(ToothGrowth)
```

```
'data.frame':   60 obs. of  3 variables:
 $ len : num  4.2 11.5 7.3 5.8 6.4 10 11.2 11.2 5.2 7 ...
 $ supp: Factor w/ 2 levels "OJ","VC": 2 2 2 2 2 2 2 2 2 2 ...
 $ dose: num  0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 ...
```

Using the HATPC framework, test whether the type of supplent affects tooth length.

Exercise 2: Adelie penguin bill length

For this exercise, we will be using a subset of the palmer penguins dataset.

```r
library(palmerpenguins)
library(tidyverse)

adelie <-penguins%>%
  filter(species == "Adelie")%>%
  na.omit()%>%
  droplevels()

head(adelie)
```

```
# A tibble: 6 × 8
  species island    bill_length_mm bill_depth_mm flipper_length_mm body_mass_g
  <fct>   <fct>              <dbl>         <dbl>             <int>       <int>
1 Adelie  Torgersen           39.1          18.7               181        3750
2 Adelie  Torgersen           39.5          17.4               186        3800
3 Adelie  Torgersen           40.3          18                 195        3250
4 Adelie  Torgersen           36.7          19.3               193        3450
5 Adelie  Torgersen           39.3          20.6               190        3650
6 Adelie  Torgersen           38.9          17.8               181        3625
# ℹ 2 more variables: sex <fct>, year <int>
```

Using the HATPC framework, test whether male and female penguins have the same length bill

Exercise 3: Penguin body mass

For this exercise, we will use a subset of the palmer penguins data set again. This time, we will be comparing two different penguin species.

```r
library(palmerpenguins)

penguins2<- penguins%>%
  filter(species != "Adelie")%>%
  na.omit()%>%
  droplevels()
```

Using the HATPC framework, test whether chinstrap and gentoo penguins have different body masses.