Assessment

On the VLE Module Assessment link

(ii) A report of the summer term Experimental Design Practical including an Appendix with R script.

- logical folder structure
- script, any accessory functions and the data itself.
- script should be well-commented, well-organised and follow good practice in the use of spacing, indentation, and variable naming.
- include all the code required to reproduce data import and formatting as well as the summary information, analyses, and figures in your report.
- Examples of well-formatted scripts are given at the end of most workshops.

17C Laboratory & Professional Skills: Data Analysis

Laboratory & Professional skills for Bioscientists
Term 2: Data Analysis in R

Two sample tests: two-sample *t*-test, two-sample Wilcoxon

Practical: one and two sample tests examples, different data formats ggplot.

Summary

- Tests for one-, two-, and paired-samples (*t*-tests and non-parametric equivalents)
- Today
  - two-sample *t*-test for independent samples
  - Wilcoxon for one- and two-samples when *t*-test assumptions are not met
- Choosing appropriate tests: type of question, type of data
- Second of two lectures; single workshop

Overview of topics

Week	Topic	
2	Introduction. Logic of hypothesis testing	Foundation
3	Hypothesis testing, variable types	
4	Chi-squared tests	Hypothesis testing
5	The normal distribution, summary statistics and CI	Estimation
6 and 7	**One- and two-sample tests (2 lectures)**	
8	One-way ANOVA and Kruskal-Wallis	Hypothesis testing
9	Two-way ANOVA incl understanding the interaction	
10	Correlation and regression	

Lecture 2

1. Estimation
   – what is the mean of the population?
2. Hypotheses testing

   e.g., is there a difference between 2 means (*t*-test)

   e.g., is the expected number of observations what we expect (chi-squared test)

Also in previous lecture

Learning objectives

By actively following the lecture and practical and carrying out the independent study the successful student will be able to:

- Explain dependent and independent samples (MLO 2)
- Select, appropriately, *t*-tests and their non-parametric equivalents (MLO 2)
- Apply, interpret and evaluate the legitimacy of the tests in R (MLO 3 and 4)
- Summarise and illustrate with appropriate figures test results scientifically (MLO 3 and 4)

Also in previous lecture

Types of *t*-test

1. One-sample
   - Compares the mean of sample to a particular value (compares the response to a reference)
     - Includes paired-sample test for dependent samples (i.e., two linked measures)
2. Two-sample
   - Compares two (independent) means to each other

Also in previous lecture

Paired-sample *t*-tests example

Is there a difference between the maths and stats marks of 10 students?

The one sample is the difference between the pairs of values

Same student

	subject	mark
1	maths	97
2	maths	58
3	maths	65
4	maths	65
5	maths	80
6	maths	48
7	maths	85
8	maths	63
9	maths	58
10	maths	98
11	stats	89
12	stats	49
13	stats	68
14	stats	70
15	stats	74
16	stats	30
17	stats	78
18	stats	69
19	stats	40
20	stats	85

NEW!

Two-sample *t*-tests

- Is there a difference between two independent means
  - Independent – values in one group not related to values in the other group
- Example: is there a significant difference between the masses of male and female chaffinches?

Fringilla coelebs

filecopy.R* × Untitled1* ×

Filter

	mass	sex
1	18.3	females
2	22.1	females
3	22.4	females
4	18.5	females
5	22.2	females
6	19.3	females
7	17.8	females
8	20.2	females
9	22.1	females
10	16.6	females
11	20.7	females
12	18.7	females
13	22.6	females
14	21.5	females
15	21.7	females
16	19.9	females
17	23.1	females
18	17.8	females
19	19.5	females
20	24.6	females
21	22.7	males
22	20.6	males
23	25.4	males

Showing 1 to 23 of 40 entries

NOT LINKED

t-tests

Also in previous lecture

- Standard formula for all t-tests

$$t = \frac{statistic - hypothesised\ value}{s.e.of\ statistic}$$

NEW!

- $$t_{[d.f]} = \frac{(\bar{x}_1 - \bar{x}_2) - 0}{s.e.of(\bar{x}_1 - \bar{x}_2)}$$

To aid understanding, not for remembering

- d.f.= $n_1 + n_2 - 2$

Also in previous lecture

t-tests in general

Assumptions

All *t*-tests assume the “residuals” are normally distributed and have homogeneity of variance

A residual is the difference between the predicted and observed value

Predicted value is the mean / group mean

Also in previous lecture

Checking Assumptions

- Common sense
  - response should be continuous
  - No/few repeats
- Plot the residuals
- Using a test in R

Also in previous lecture

When data are not normally distributed

- Transform (not really covered)
  - E.g. Log to remove skew, arcsin squareroot on proportions
- Use a non-parametric test (covered)
  - Fewer assumptions
  - Generally less powerful

NEW!

Two-sample *t*-test example

- Example: is there a significant difference between the masses of male and female chaffinches?

Fringilla coelebs

NEW!

Two-sample *t*-test example

filecopy.R* Untitled1*

Filter

	mass	sex
1	18.3	females
2	22.1	females
3	22.4	females
4	18.5	females
5	22.2	females
6	19.3	females
7	17.8	females
8	20.2	females
9	22.1	females
10	16.6	females
11	20.7	females
12	18.7	females
13	22.6	females
14	21.5	females
15	21.7	females
16	19.9	females
17	23.1	females
18	17.8	females
19	19.5	females
20	24.6	females
21	22.7	males
22	20.6	males
23	25.4	males

Showing 1 to 23 of 40 entries

```
chaff <- read.table("../data/chaff.txt", header = T)
```

Note: these data are ‘tidy’

All the responses in one column with other variables indicating the group

Organise your data this way

Extension of previous lect

Tidy data

- Each variable should be in one column.
- Each different observation of that variable should be in a different row.
- There should be one table for each "kind" of variable.
- If you have multiple tables, they should include a column in the table that allows them to be linked.

Independent study: Wickham, H (2013). Tidy Data. Journal of Statistical Software.

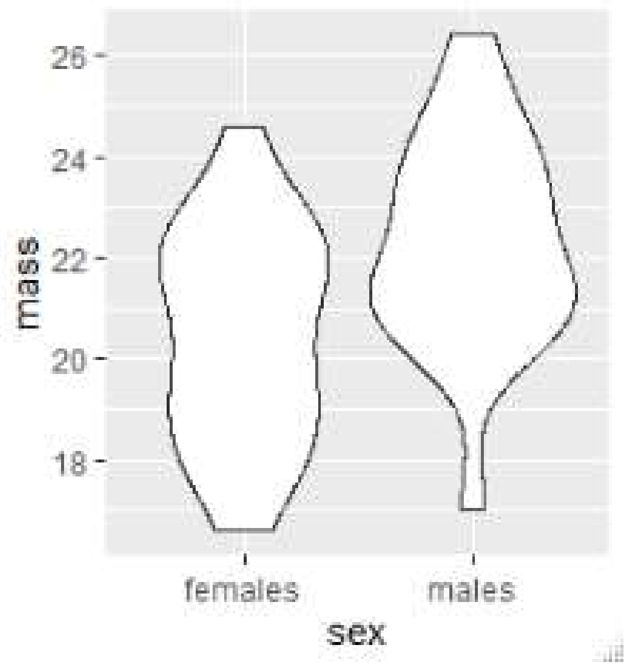
https://www.jstatsoft.org/article/view/v059i10

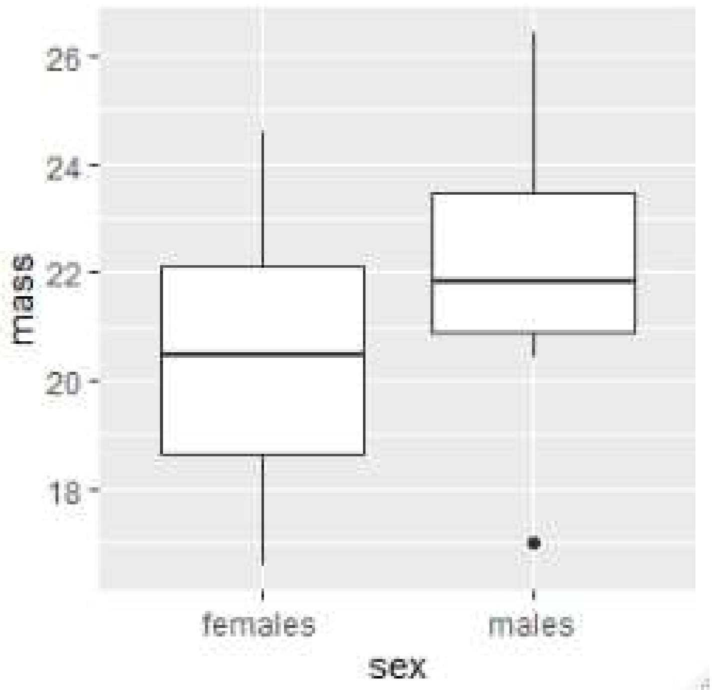
Two-sample *t*-test example

Plot your data: roughly – perhaps one of these…

```
ggplot(data = chaff,
       aes(x = sex, y = mass)) +
  geom_violin()
```

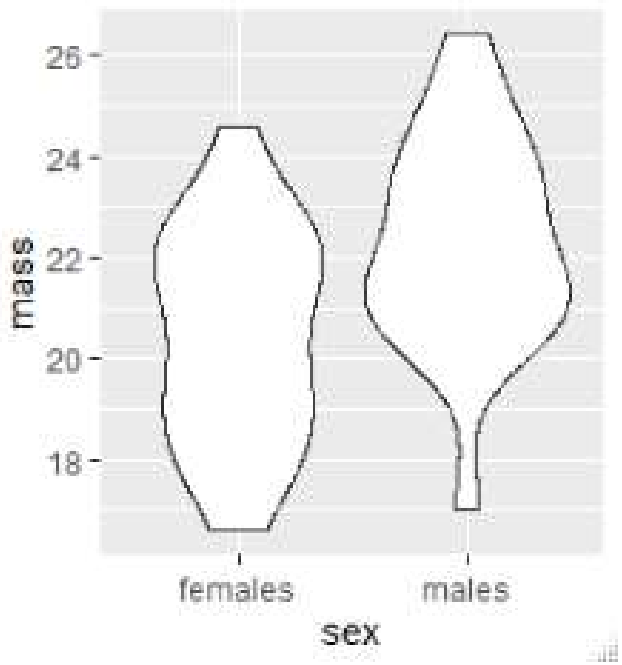
```
ggplot(data = chaff,
       aes(x = sex, y = mass)) +
  geom_boxplot()
```

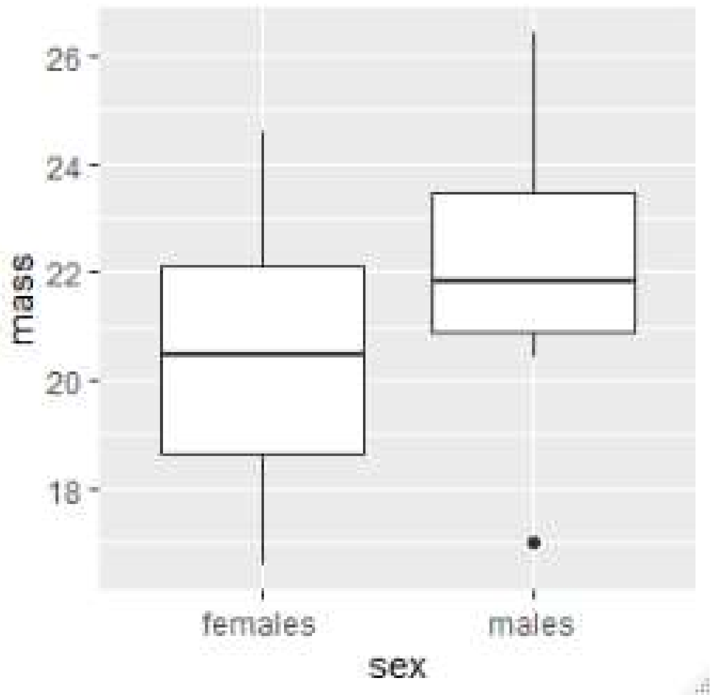




Two-sample *t*-test example

Plot your data: don't overthink. Just gives you idea of what to expect and helps identify issues (missing data, outliers etc)





Two-sample *t*-test example

Summarise the data:

```
chaffsum <- chaff %>%
  group_by(sex) %>%
  summarise(mean = mean(mass),
            std = sd(mass),
            n = length(mass),
            se = std/sqrt(n))
```

```
chaffsum
# A tibble: 2 x 5
  sex       mean   std     n    se
  <fct>    <dbl> <dbl> <int> <dbl>
1 females   20.5  2.14    20 0.478
2 males     22.3  2.15    20 0.481
```

Two-sample *t*-test example

Run the *t*-test

```
t.test(data = chaff,
       mass ~ sex,
       paired = F,
       var.equal = T)
```

Name of the dataframe

The ‘model’ explain mass by sex

The data are not paired, they are independent

We are assuming homogeneity of variance

Two-sample *t*-test example

Run the *t*-test

```
t.test(data = chaff,
       mass ~ sex,
       paired = F,
       var.equal = T)
```

```
data:  mass by sex
t = -2.6471, df = 38, p-value = 0.01175
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -3.167734 -0.422266
sample estimates:
mean in group females   mean in group males
               20.480                22.275
```

Two-sample *t*-test example

Checking the assumptions: calculate the residuals – the difference between predicted and observed (i.e., group mean and value)

```
# add the group means to the data
chaff <- merge(chaff, chaffsum[,1:2], by = "sex")

# add the residuals
chaff <- chaff %>%
  mutate(residual = mass - mean)
```

Two-sample *t*-test example

Checking the assumptions: normally and homogenously distributed residuals

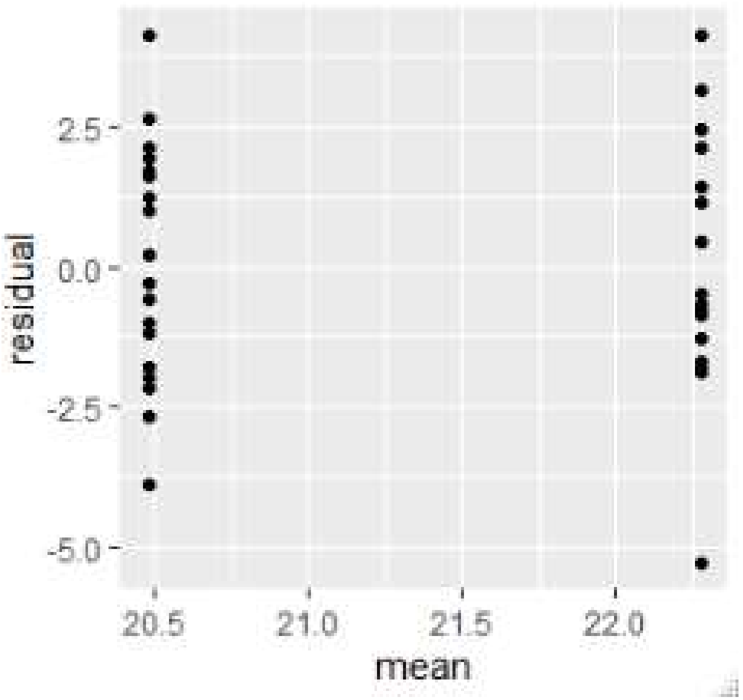
```
shapiro.test(chaff$residual)

	Shapiro-Wilk normality test

data:  chaff$residual
W = 0.98046, p-value = 0.7067
```

Two-sample *t*-test example

Checking the assumptions: normally and <u>homogenously</u> distributed residuals

```
ggplot(data = chaff,
       aes(x = mean, y = residual)) +
  geom_point()
```

Variance is about the same for all values of x



Extension of previous lect

Two-sample *t*-test example

Reporting the result: "significance of effect, direction of effect, magnitude of effect"

Male chaffinches ($\bar{x} \pm s.e$: 22.48 ± 0.48) are significantly heavier than females (20.28 ± 0.48) (t = 2.65; *d.f.* = 38; p = 0.012). See figure 1.

NEW!

Two-sample *t*-test: figures

Supports your claim:

– Show the data (all if possible)
– Show the 'model' (the predicted values i.e., means and error bars)
– Say what kind of error bars
– Full but concise figure legends

NEW!

Two-sample *t*-test: figures

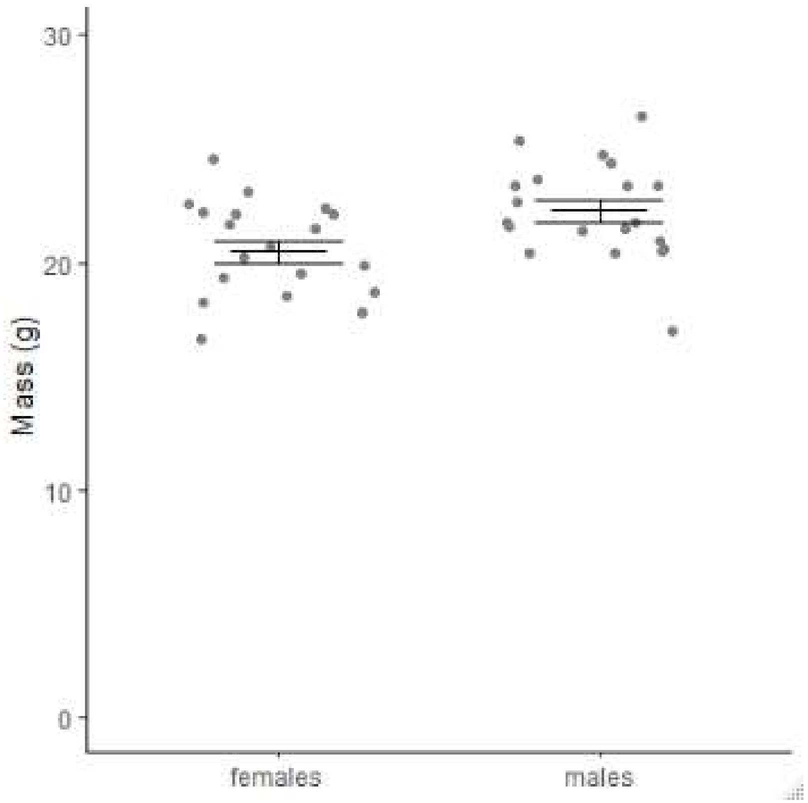


Figure 1. Mean mass of male and female chaffinches. Error bars are means +/- one standard error.

NEW!

Figure 1. Graph to show mass of male and female chaffinches.
Error bars are means +/- one standard error.

When the *t*-test assumptions are not met: non- parametric tests

- Non-parametric tests make fewer assumptions

- Based on the ranks rather than the actual data

- Null hypotheses are about the *mean rank* (not the mean)

Non-parametric tests

t-test equivalents

i,.e., the type of question is the same but the response variable is not normally distributed

- one – sample *t*-test and paired-sample *t*-test: the one-sample Wilcoxon
- Two-sample *t*-test: two-sample Wilcoxon aka Mann-Whitney

Non-parametric tests

two-sample Wilcoxon (Mann-Whitney)

Example: comparing the number of leaves on 8 mutant and 7 wild type plants (small samples, counts)

	leaves	type
1	3	mutant
2	5	mutant
3	6	mutant
4	7	mutant
5	3	mutant
6	4	mutant
7	5	mutant
8	8	mutant
9	8	wild
10	9	wild
11	6	wild
12	7	wild
13	8	wild
14	9	wild
15	10	wild

Non-parametric tests

two-sample Wilcoxon (M-W): example

Carrying out the test two-sample Wilcoxon

```
wilcox.test(data = plants, leaves  ~ type, paired = FALSE)

       Wilcoxon rank sum test with continuity correction

data:  leaves by type
W = 5, p-value = 0.008664
alternative hypothesis: true location shift is not equal to 0

Warning message:
In wilcox.test.default(x = c(3, 5, 6, 7, 3, 4, 5, 8), y = c(8, 9,  :
  cannot compute exact p-value with ties
```

No need to worry!

Non-parametric tests

two-sample Wilcoxon (M-W): example

Reporting the result: "significance of effect, direction of effect, magnitude of effect"

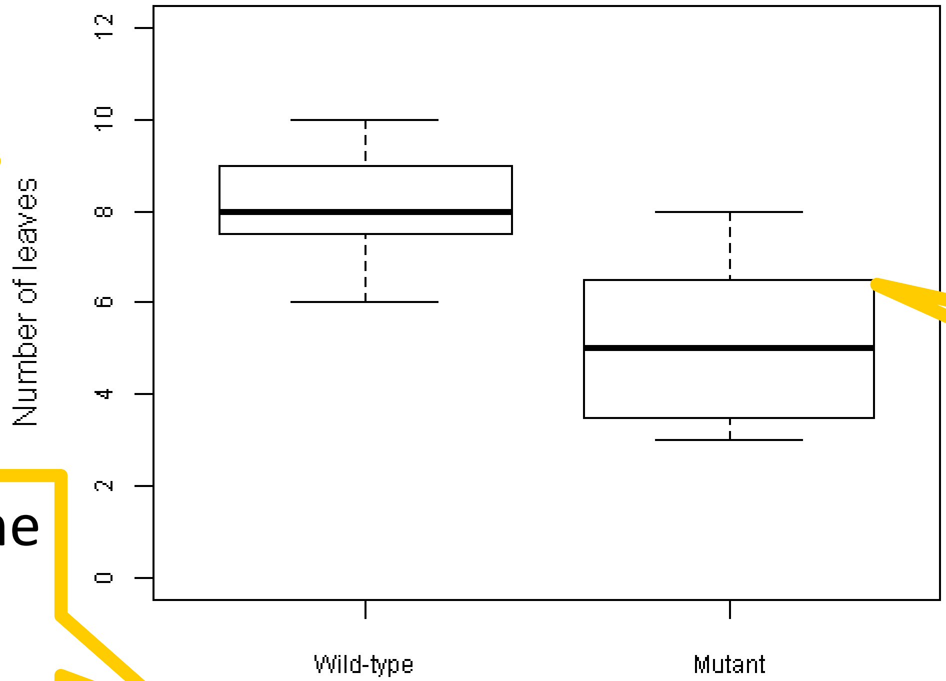
There are significantly more leaves on wild-type (median = 8) than mutant (median = 5) plants (Mann-Whitney: $W=5$, $n_1=7$, $n_2=8$, $p = 0.009$)

Non-parametric tests

two-sample Wilcoxon (M-W): example

Label with units

Non-parametric tests: use median+IQR



Measure of dispersion - IQR

What is the figure?

Figure 2. Median (heavy lines) number of leaves on Wild-type and mutant plants. Boxes indicate the inter-quartile range.

Always refer to figure in the text

Learning objectives for the week

By attending the lectures and practical the successful student will be able to

- Explain dependent and independent samples (MLO 2)
- Select, appropriately, *t*-tests and their non-parametric equivalents (MLO 2)
- Apply, interpret and evaluate the legitimacy of the tests in R (MLO 3 and 4)
- Summarise and illustrate with appropriate R figures test results scientifically (MLO 3 and 4)