Dissertation statistics support

Importance of statistical literacy in society

"Good statistics are like a telescope for an astronomer, a microscope for a bacteriologist, or an X-ray for a radiologist. If we are willing to let them, good statistics help us see things about the world around us and about ourselves – both large and small – that we would not be able to see in any other way" (Hartford 2020)

[1]Harford, T. 2020. How to make the world add up: ten rules for thinking differently about numbers. Hachette UK.

Importance of statistical literacy as professionals

Journal of Animal Ecology

Bentlage et al. (2025). **Results**:

"At the population level, total rainfall from February to August had a significant quadratic effect on the annual divorce rate, which increased in years with low and high rainfall (GLM, estimate = 0.335 ± 0.091, p-value = 0.003; Figure S1b). Rainfall effects explained 46.7% of the annual divorce rate's variance ($r^2 = 0.467$)."

[1]Bentlage, AA, et al. 2025. Rainfall is associated with divorce in the socially monogamous Seychelles warbler. Journal of Animal Ecology. 94:85-98.

Importance of statistical literacy as professionals

Marine Biology

Liu et al. (2025). **Materials and Methods**:

"Experimental data were presented as the mean ± standard deviation of results obtained from three parallel samples [...]. A one-way analysis of variance (ANOVA) combined with the least significant difference (LSD) method was employed to assess the variance and significance of nutrient content in algal detritus under different environmental conditions (Table 1) ($p < 0.05$)."

[1] Liu, Z, et al. 2025. The impact of dissolved oxygen and sediments on the decomposition of *Sargassum thunbergii*. Marine Biology, 172:28.

Importance of statistical literacy as professionals

Journal of Environmental Sciences

Tava et al. (2025). **Results**:

"Soil microbial biomass was highly significantly ($P < 0.01$) and positively correlated with alkaline phosphomonoesterase activity in untreated soils (D0, $r_{T03} = 0.92$ and $r_{T14} = 0.77$) and at D01 rate ($r_{T03} = 0.86$ and $r_{T14} = 0.71$) at both time points T03 and T14."

[1]Tava, A, et al. 2025. Saponins in soil, their degradation and effect on soil enzymatic activities. Journal of Environmental Sciences, 154:378-389.

Importance of statistical literacy as professionals

Cell

Nishijima et al. (2025). **Methods**:

"To investigate correlations between the microbiome profile (i.e. species-level taxonomic and functional compositions) and the experimentally measured microbial load, Pearson correlation coefficients were calculated between the log_{10} transformed relative abundance of each microbial species/functions and the microbial load in each cohort separately."

[1] Nishijima, S, et al. 2025. Fecal microbial load is a major determinant of gut microbiome variation and a confounder for disease associations. Cell. 188:222-236

Importance of statistical literacy as professionals

Journal of Environmental Geography

Ziti et al. (2025). **Results**:

"The Mann-Kendall trend test shows an increasing, statistically significant ($p = 0.031$, $\alpha = 0.05$) trend in annual mean temperatures between 1969 and 2021 in the sub-catchment, a condition that may contribute to the proliferation of dry conditions in the catchment."

[1] Ziti, C et al. 2025. Climate Change Response Strategies and Implications on Sustainable Development Goals in Mutirikwi River Sub-Catchment of Zimbabwe. Journal of Environmental Geography, 17:1-14.

Importance of statistical literacy as professionals

Frontiers in Education

Naim et al. (2025). **Methods**:

"The study uses multiple linear regression analysis to examine the relationship between independent variables (socioeconomic status, funding per student, teacher-student ratio, availability of advanced coursework, tuition fees, scholarship availability, and diversity of the student body) and dependent variables (test scores, graduation rates, enrolment statistics, and post-graduation employment rates)."

[1]Naim, A. 2025. Equity Across the Educational Spectrum: Innovations in Educational Access Crosswise all Levels. Frontiers in Education 9:1499642

Overview

1. Data organisation and data types
2. Introduction to jamovi
3. Data visualisation
4. Summary statistics
5. Hypothesis testing
6. T-tests and ANOVA
7. Chi-square test
8. Correlation and regression

1. Data Organisation and Data Types

- Chapter 2: Data organisation
- Chapter 3: *Practical:* Preparing data
- Chapter 5: Types of variables

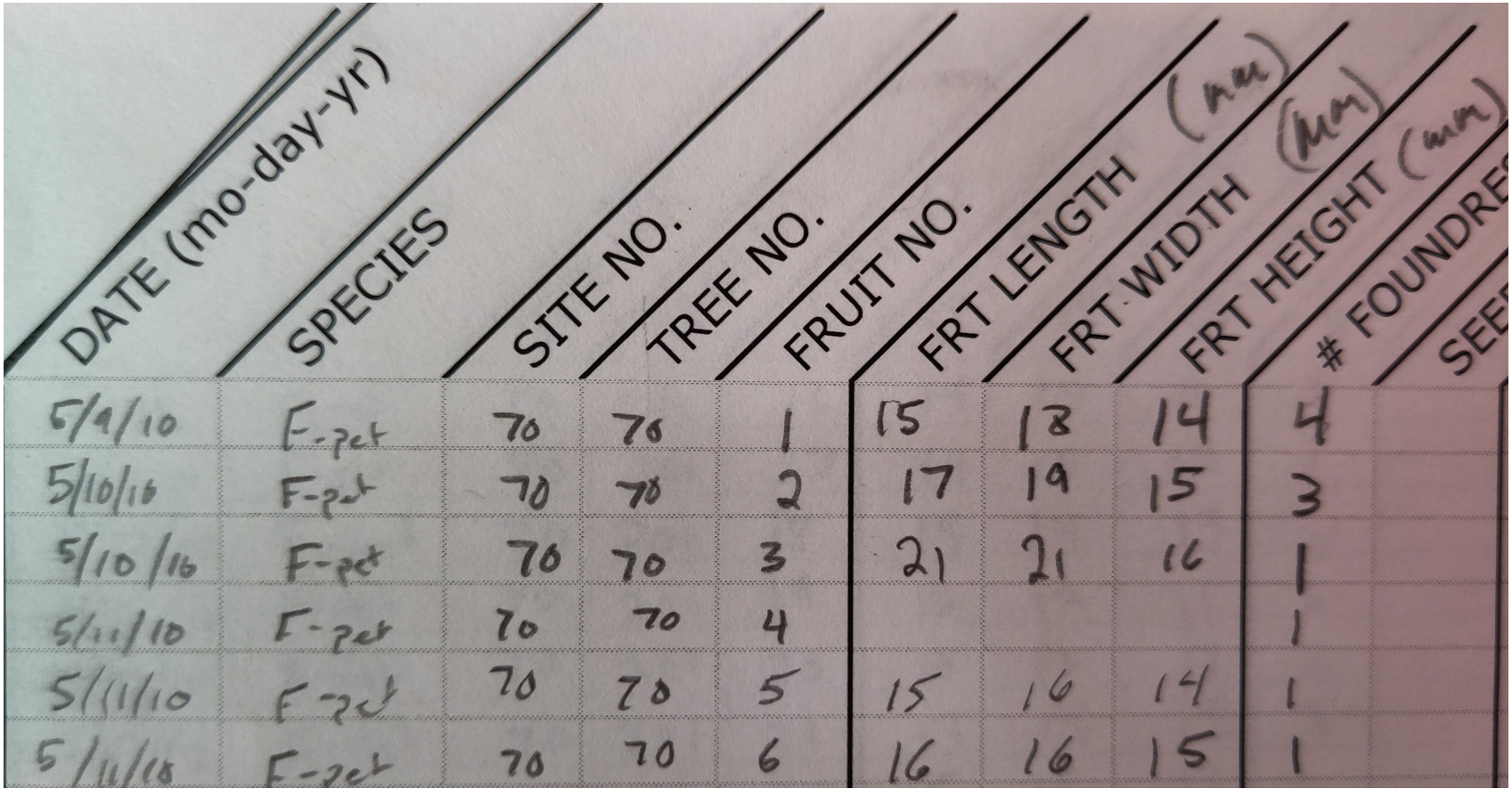
Data collection is messy

Figure 1: Dr Becky Boulton collects data from nest boxes in the field (A), then processes nest material in the lab (B).

Data collection is messy

Figure 2: Sonoran Desert Rock Fig in the desert of Baja, Mexico.

Data collection is messy



DATE (mo-day-yr)	SPECIES	SITE NO.	TREE NO.	FRUIT NO.	FRT LENGTH (mm)	FRT WIDTH (mm)	FRT HEIGHT (mm)	# FOUNDRE	SEE
5/9/10	F-pet	70	70	1	15	18	14	4	
5/10/10	F-pet	70	70	2	17	19	15	3	
5/10/10	F-pet	70	70	3	21	21	16	1	
5/11/10	F-pet	70	70	4				1	
5/11/10	F-pet	70	70	5	15	16	14	1	
5/11/10	F-pet	70	70	6	16	16	15	1	

Figure 3: A portion of a lab notebook used to record measurements of fig fruits from different trees in 2010.

Observations and variables

- **Observation:** Units of sample
- **Variable:** Unit measurement

Observations and variables

- **Observation:** Fig trees
- **Variable:** Tree height, location

Three characteristics of tidy data

Summarised by Wickham (2014)[1]:

1. Each **variable** gets its own column.
2. Each **observation** gets its own row.
3. Different units of observation require different data files.

[1] Wickham, H. 2014. *J. Stat. Softw.* 59:1-23.

Three characteristics of tidy data

Tree	Species	Height (m)	Leaf length (cm)
1	Oak	20.3	8.1
2	Oak	25.4	9.4
3	Maple	18.2	12.5
4	Maple	16.7	11.3

Data types

- Categorical: Discrete types
  - Nominal: No order (Oak, Maple, Elm)
  - Ordinal: Order (Low, Medium, High)

Data types

- Categorical: Discrete types
  - Nominal: No order (Oak, Maple, Elm)
  - Ordinal: Order (Low, Medium, High)
- Quantitative: Numbers that reflect magnitude
  - Discrete: Specific values (e.g., counts)
  - Continous: Any decimal value allowed

Predicting variables

- **Independent variable** (X): Variable to predict another variable
- **Dependent variable** (Y): Variable that we want to predict

2. Introduction to jamovi

- Chapter 3: *Practical:* Preparing data
- Chapter 5: *Practical* Introduction to jamovi

Spreadsheets and jamovi

- Spreadsheets (any)
  - MS Excel
  - LibreOffice Calc
  - Google Sheets

Spreadsheets and jamovi

- Spreadsheets (any)
  - MS Excel
  - LibreOffice Calc
  - Google Sheets
- jamovi
  - User-friendly
  - Free and open-source
  - Windows, Mac, Linux, Chrome

Ways to use jamovi

1. Download: https://www.jamovi.org
2. Jamovi cloud
3. AppsAnywhere (see portal)

[1]The jamovi project. 2024. "Jamovi (Version 2.5)." Sydney, Australia. https://www.jamovi.org.

3. Data visualisation

- Chapter 10: Graphs
- Chapter 30.1: Scatterplots

Sonoran desert rock fig

Figure 4: Sonoran Desert Rock Fig in the desert of Baja, Mexico.

Sonoran desert rock fig fruit

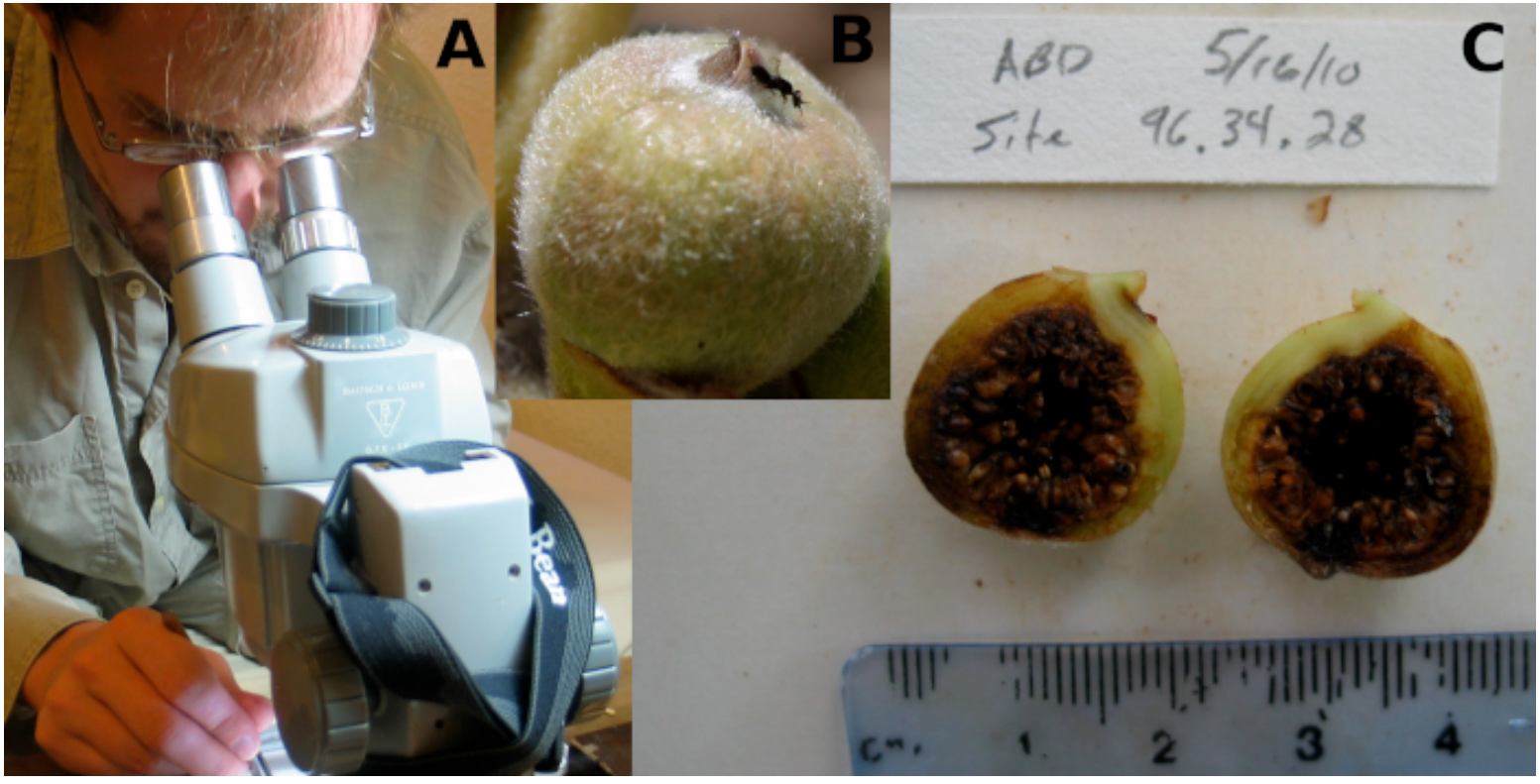


Figure 5: Three images showing the process of collecting data for the dimensions of figs from trees of the Sonoran Desert Rock Fig in Baja.

Histograms

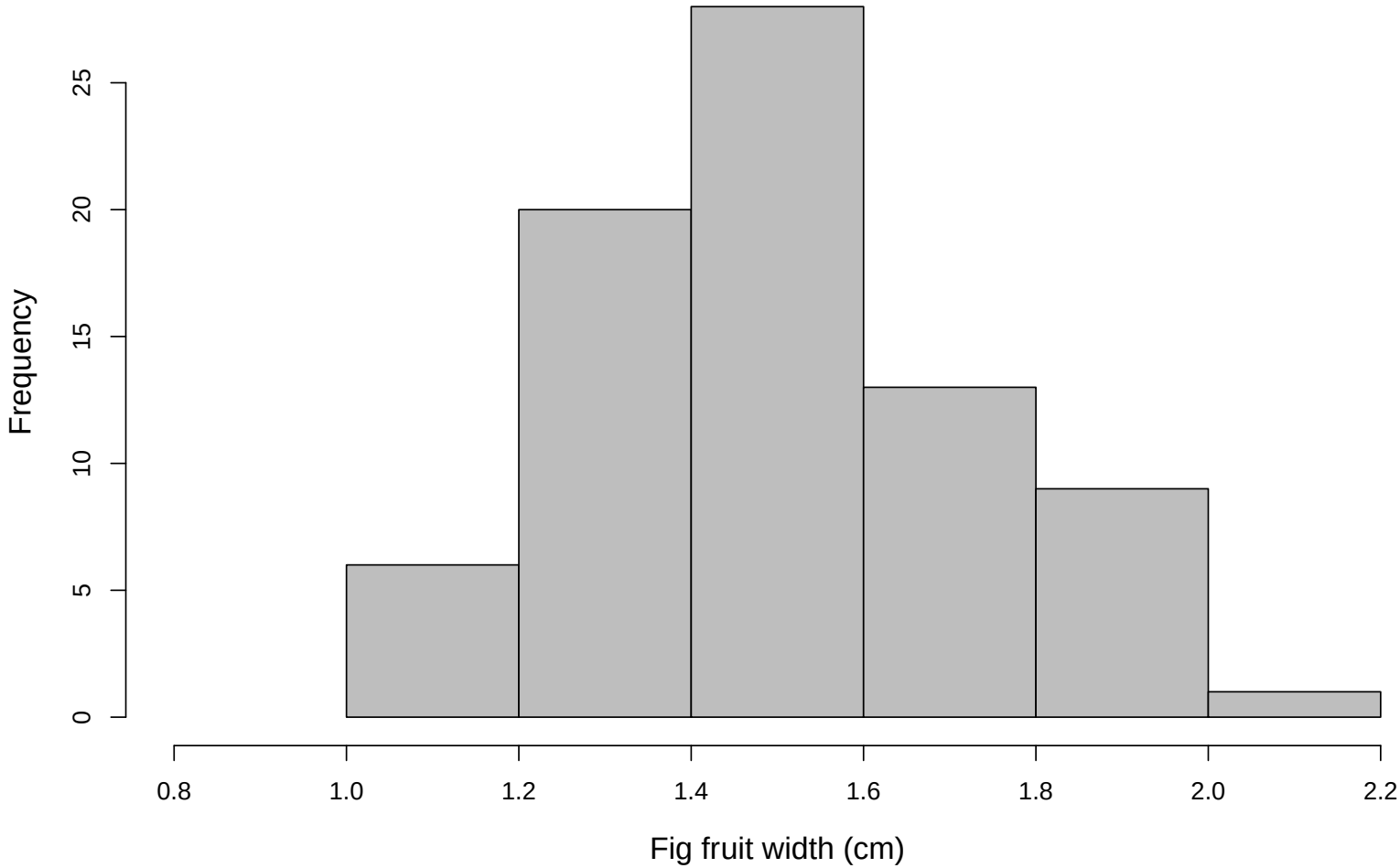


Figure 6: Example histogram fig fruit width (cm) using data from 78 fig fruits collected in 2010 from Baja, Mexico.

[1] https://bradduthie.github.io/stats/app/build_histogram/

Barplots

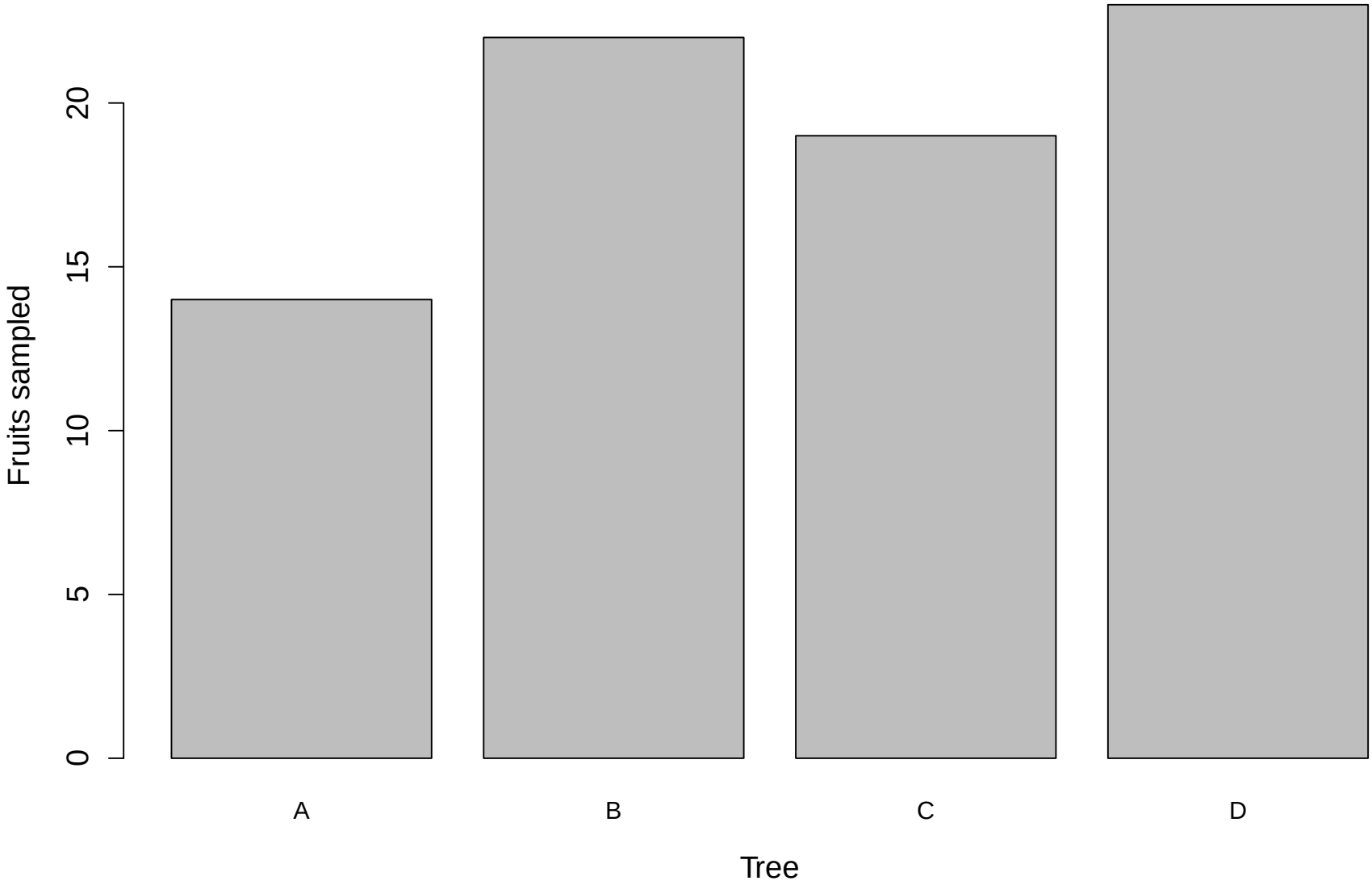


Figure 7: Example barplot showing how many fruits were collected from each of four trees (78 collected in total) in 2010 from Baja, Mexico.

Boxplots

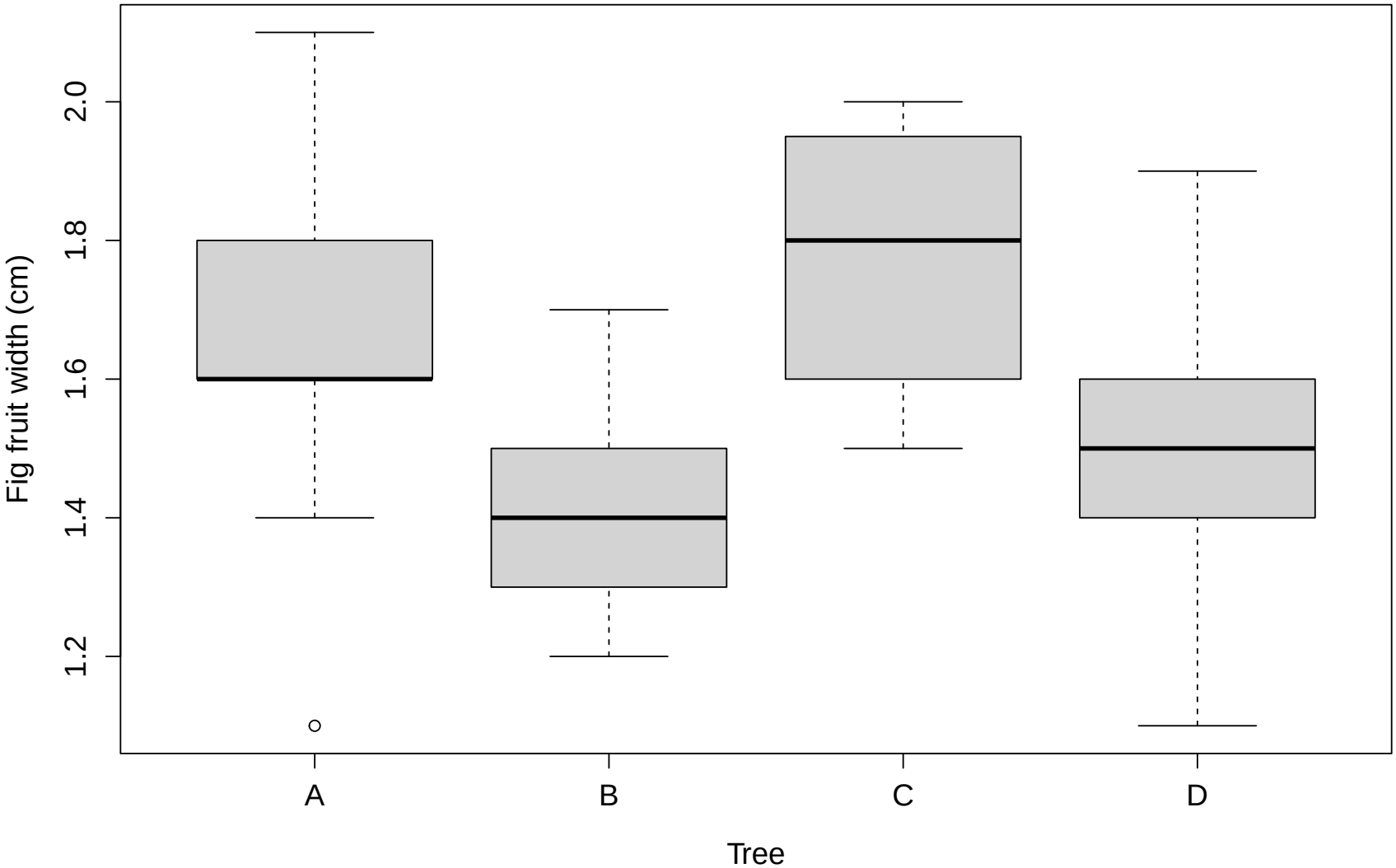


Figure 8: Boxplot of fig fruit widths (cm) collected from four separate trees sampled in 2010 from Baja, Mexico.

Fig wasps

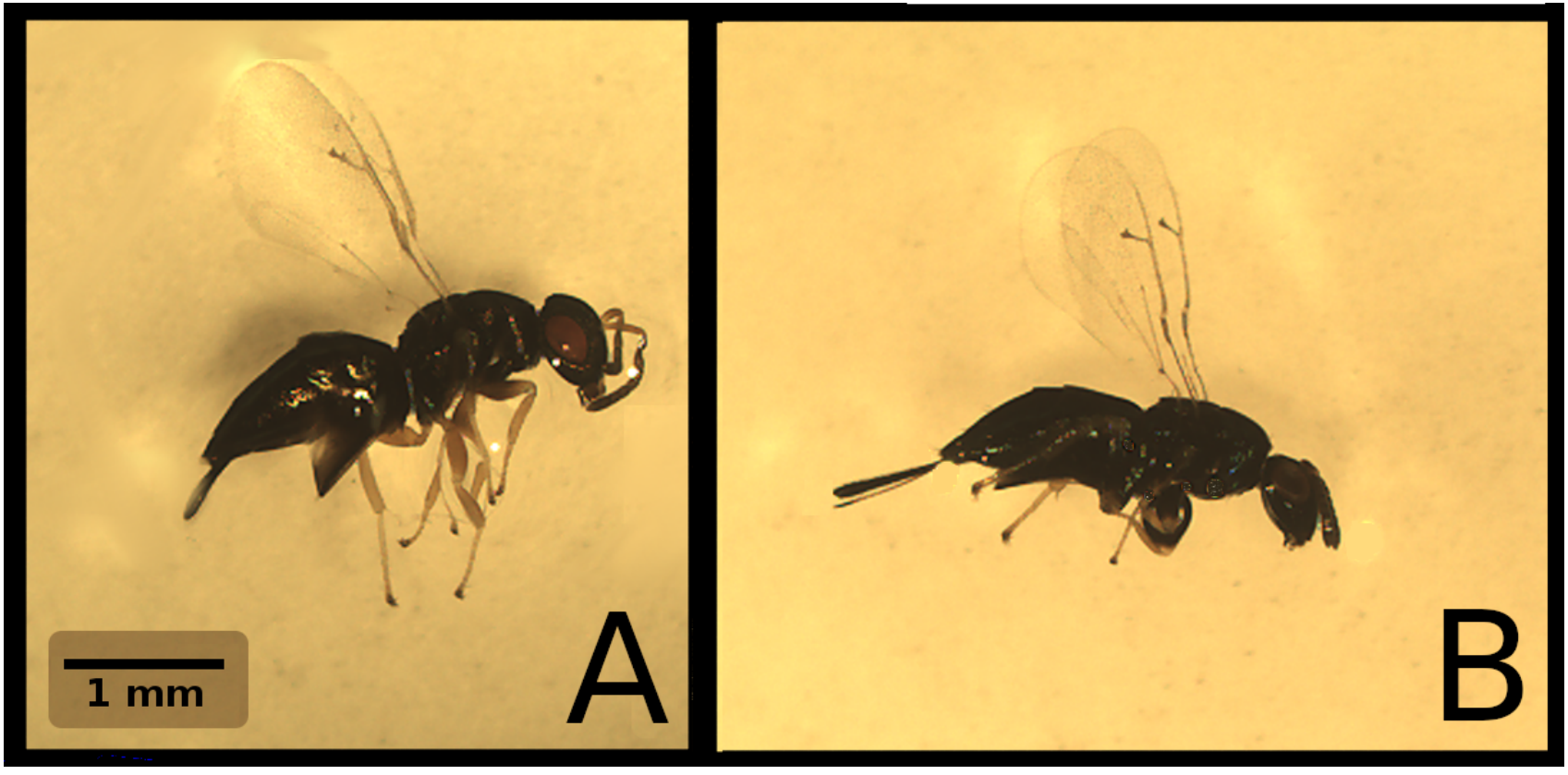


Figure 9: Fig wasps from two different species, (A) Het1 and (B) Het2. Wasps were collected from Baja, Mexico. Image modified from Duthie, Abbott, and Nason (2015).

Scatterplots

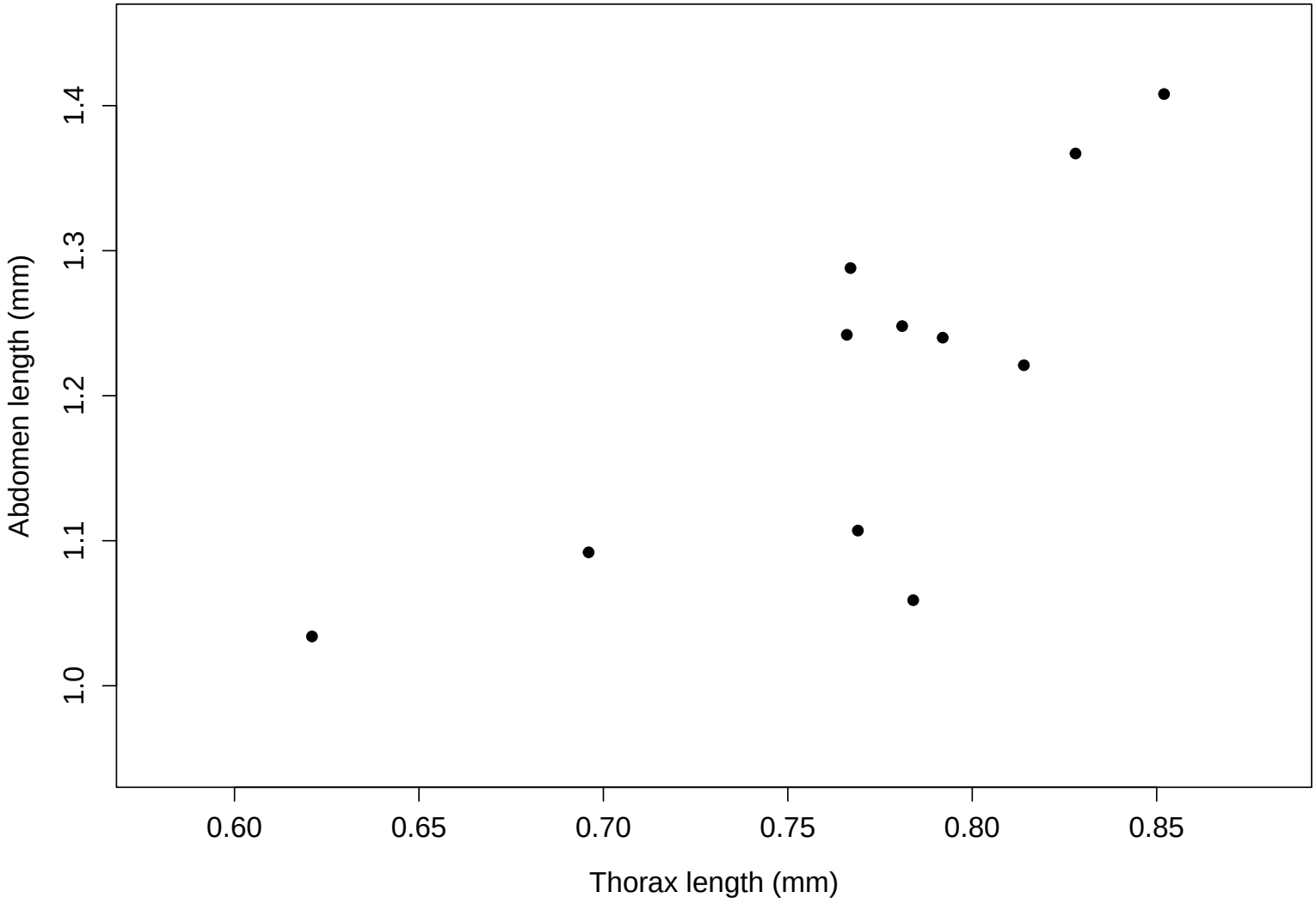


Figure 10: Example of a scatterplot in which fig wasp thorax length (x-axis) is plotted against fig wasp abdomen length (y-axis). Each point is a different fig wasp. Wasps were collected in 2010 in Baja, Mexico.

4. Summary statistics

- Chapter 11: Measures of central tendency
- Chapter 12: Measures of spread
- Chapter 13: Confidence intervals

The mean on a histogram

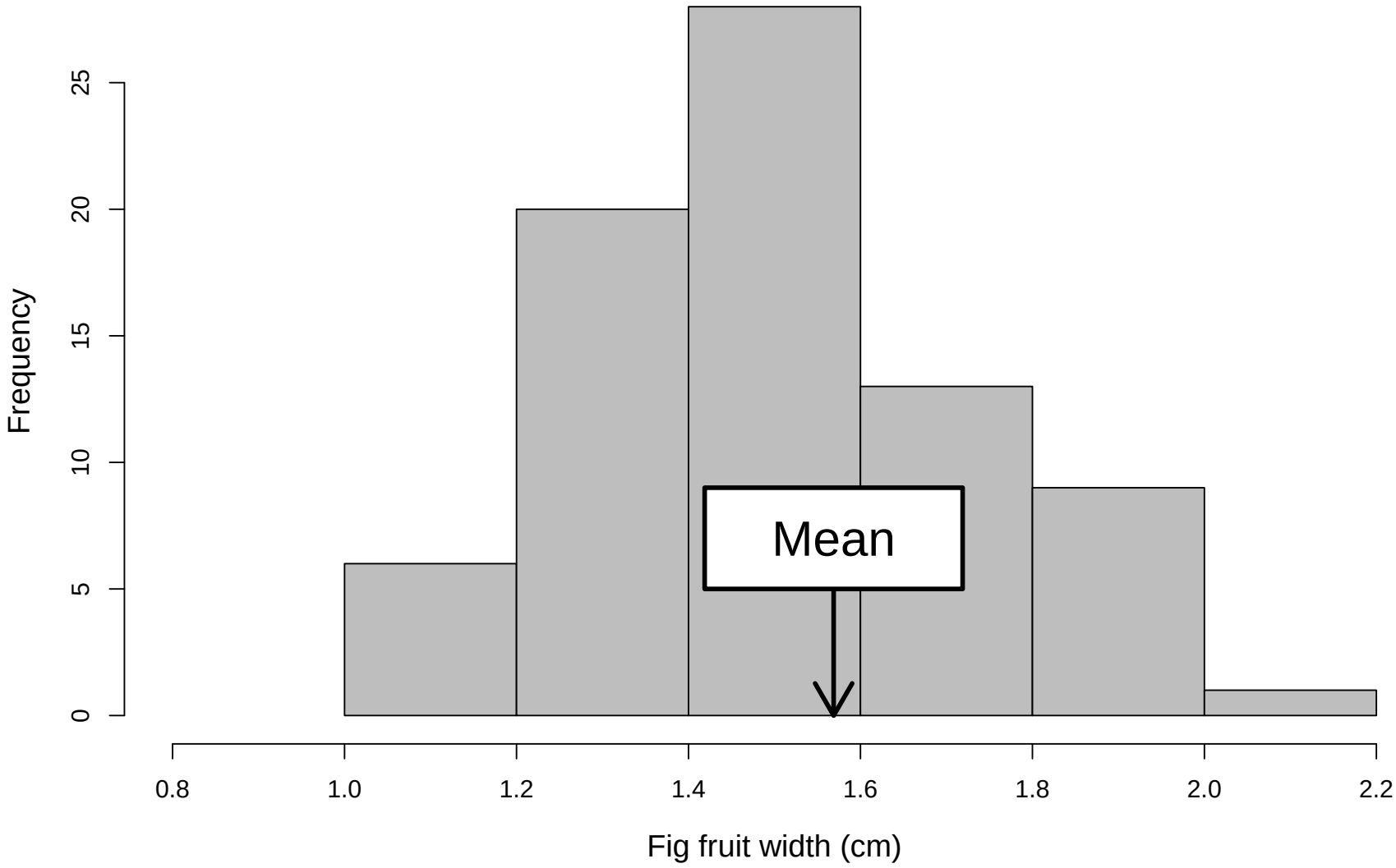


Figure 11: Example histogram fig fruit width (cm) using data from 78 fig fruits collected in 2010 from Baja, Mexico.

Measures of central tendency

- **Mean**: Add up values and divide by total
- **Median**: Middle vaue of a dataset
- **Mode**: Value that occurs most often

Note: Jamovi will calculate these for you in 'Exploration' and 'Descriptives'

Measures of spread

- **Range**: Highest value minus lowest
- **Variance**: Mean squared deviation from mean
- **Standard deviation**: Square root of variance

Note: Jamovi will also calculate these for you

[1] https://bradduthie.github.io/stats/app/forest/
[2] https://bradduthie.github.io/stats/app/normal_pos_neg/

Confidence intervals

- **Standard error**: Accuracy of your estimate of the mean
- **Confidence interval**: Percentage around the mean

Note: Jamovi will calculate, but you need to interpret

[1] https://bradduthie.github.io/stats/app/CI_hist_app/

5. Hypothesis testing

- Chapter 21: What is hypothesis testing?

How ridiculous is our hypothesis?

- Flip a fair coin 100 times

How ridiculous is our hypothesis?

- Flip a fair coin 100 times
- How many times expect heads?

How ridiculous is our hypothesis?

- Flip a fair coin 100 times
- How many times expect heads?
- Variance around expectation?

Distribution of fair coin flips

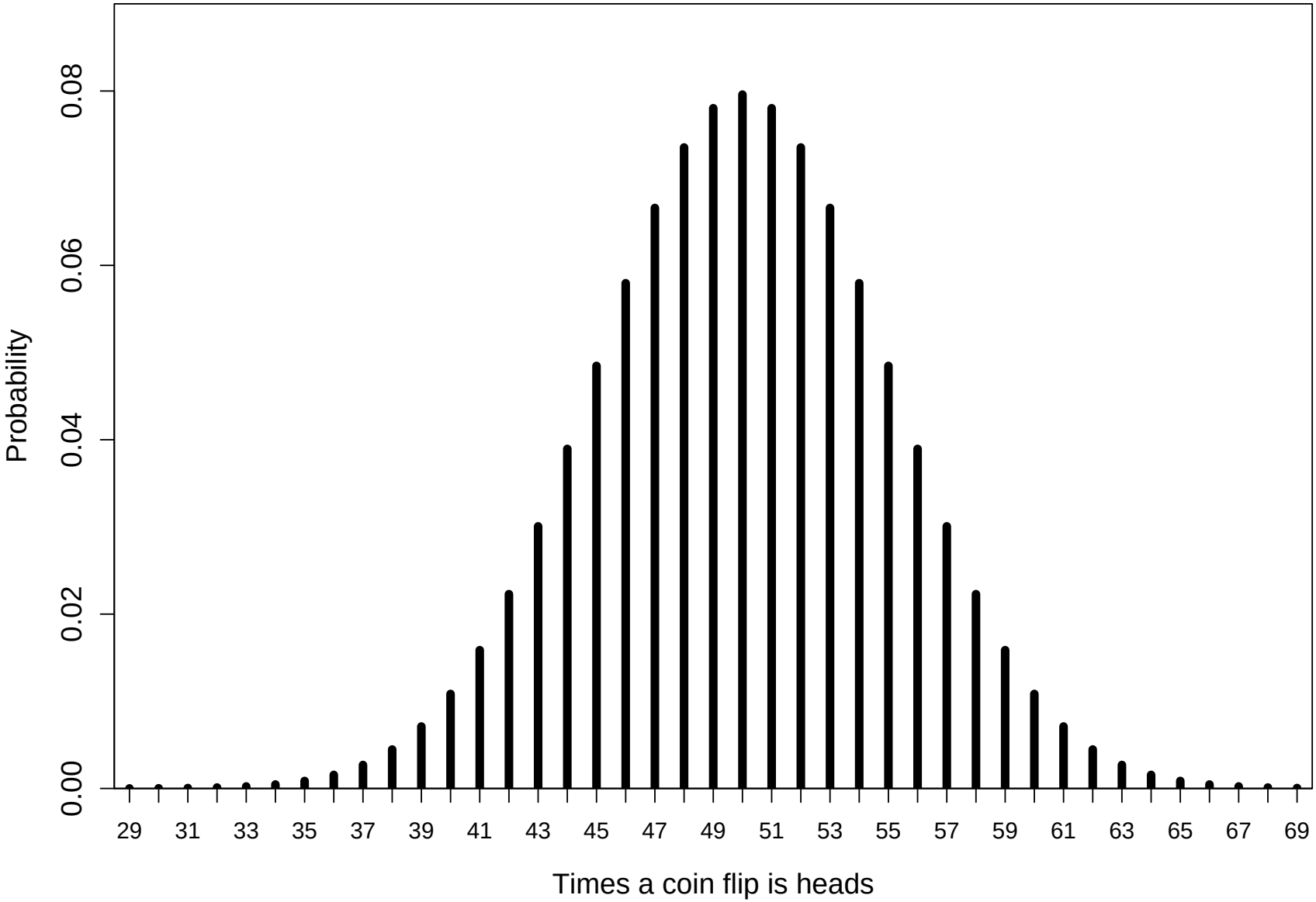


Figure 12: Probability distribution for the number of times that a flipped coin lands on heads in 100 trials.

Statistical hypothesis testing

Null Hypothesis H_0	Alternative Hypothesis H_A
There is no difference between juvenile and adult sparrow mortality	Mortality differs between juvenile and adult sparrows
Amphibian body size does not change with increasing latitude	Amphibian body size increases with latitude
Soil nitrogen concentration does not differ between agricultural and non-agricultural fields	Soil nitrogen concentration is lower in non-agricultural fields

What is a p-value?

A p-value is the probability of getting a result as or more extreme than the one observed assuming H_0 is true.

What is a p-value?

A p-value is the probability of getting a result as or more extreme than the one observed assuming H_0 is true.

Traditionally, we reject the null hypothesis if the p-value is less than 0.05 ($P < 0.05$).

Summary of hypothesis testing outcomes

	Do Not Reject H_0	Reject H_0
H_0 is true	Correct decision	Type I error
H_0 is false	Type II error	Correct decision

6. T-tests and ANOVA

- Chapter 22: The t-test
- Chapter 24: Analysis of variance
- Chapter 28: ANOVA and associated tests

6. T-tests and ANOVA

- Chapter 22: The t-test
- Chapter 24: Analysis of variance
- Chapter 28: ANOVA and associated tests

- **Independent variable(s)**: Categorical
- **Dependent variable**: Continuous

The t-test

A t-test can be used to test if two groups have significantly different means

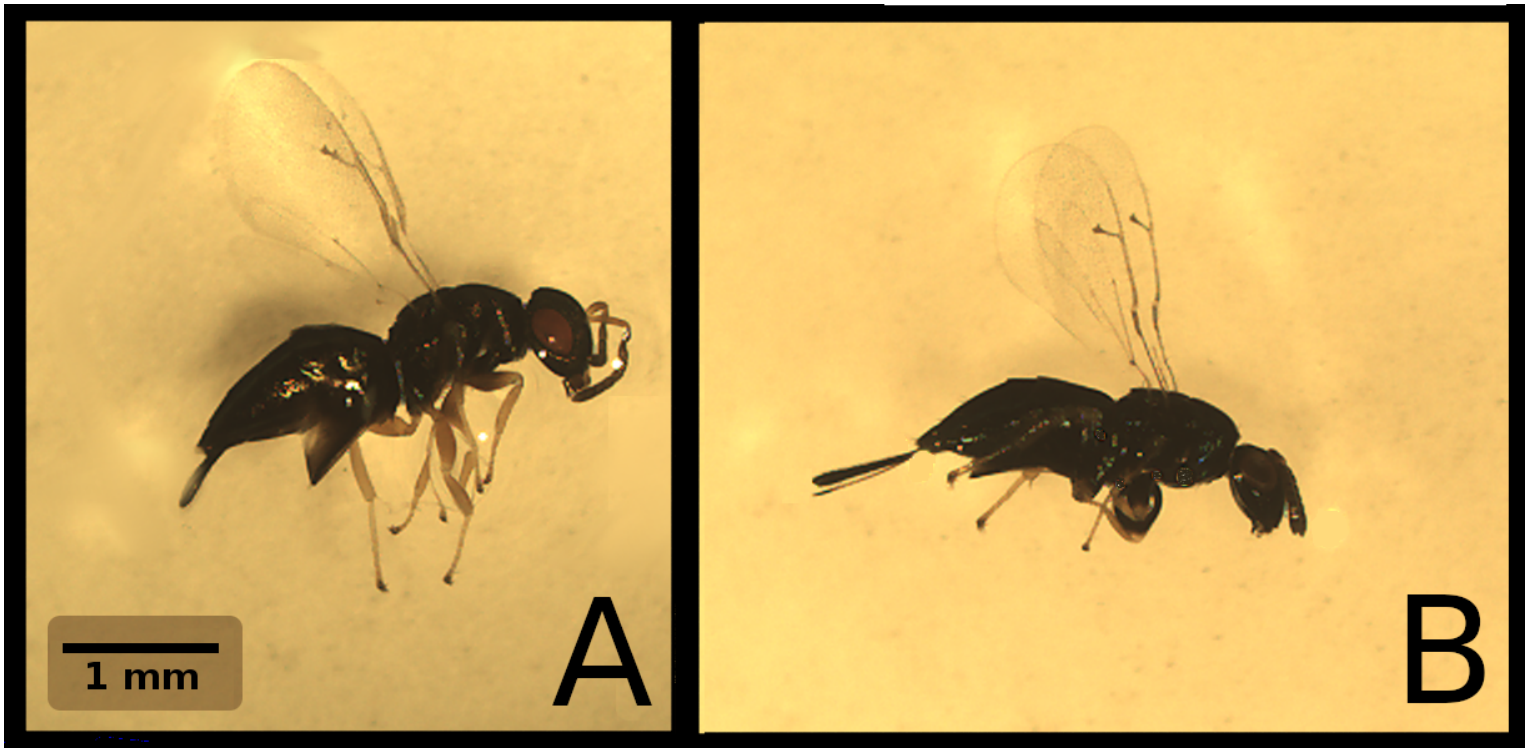


Figure 13: Fig wasps from two different species, (A) Het1 and (B) Het2.

Do two species have the same mean ovipositor length?

[1] https://bradduthie.github.io/petiolaris_data/wasp_morphology.csv

Visual explanation of a t-test

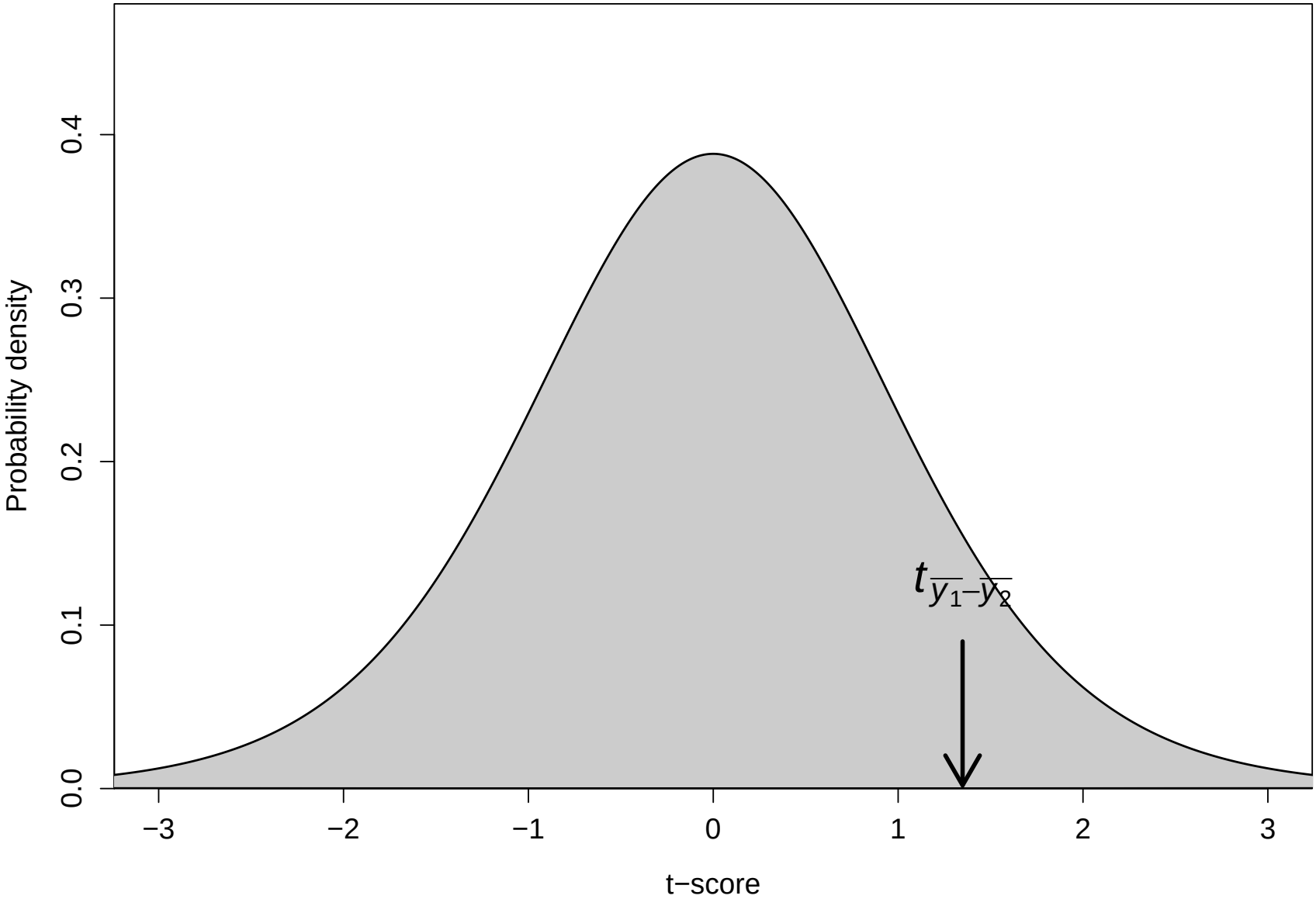


Figure 14: A t-distribution is shown with a calculated t-statistic indicated with a downward arrow.

t-test in jamovi

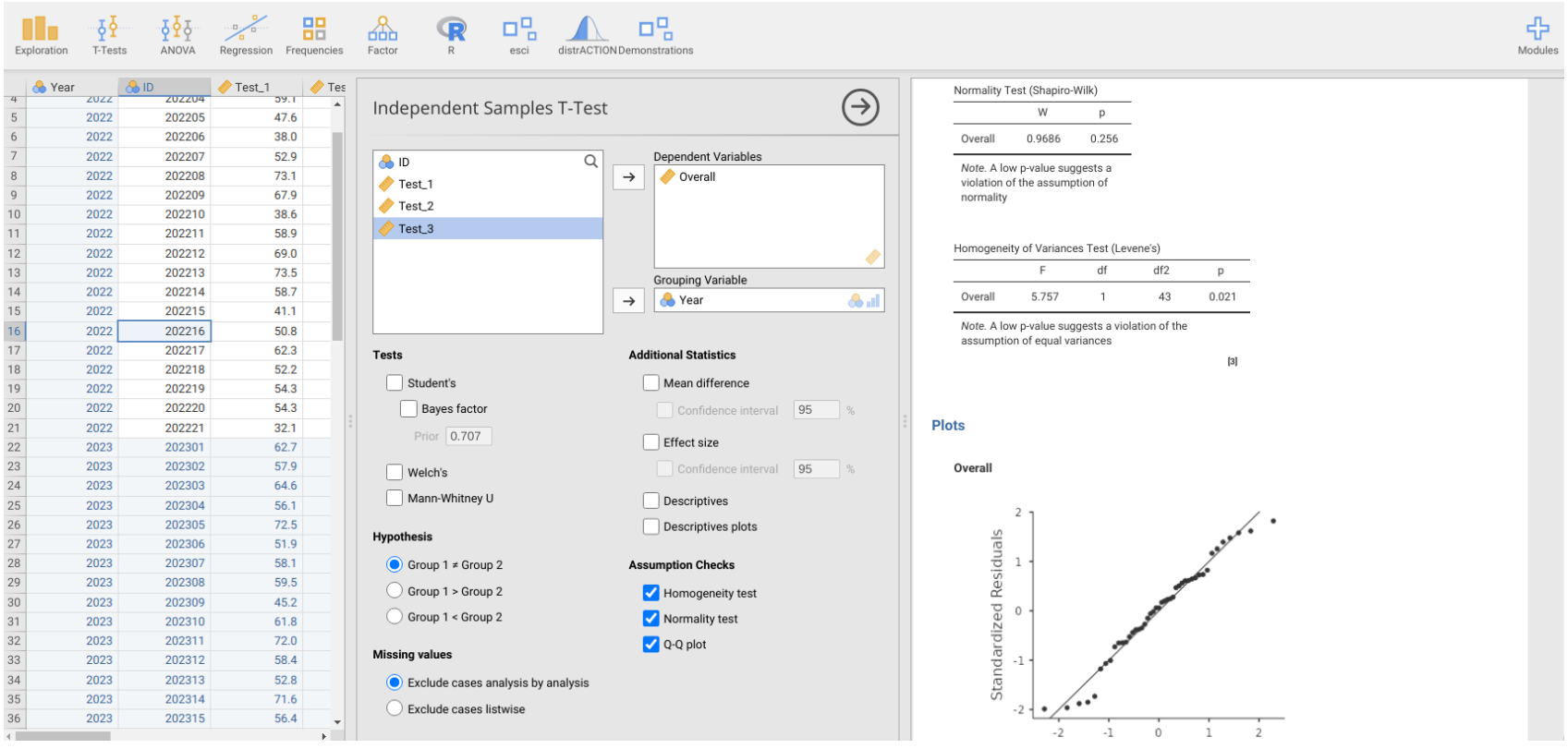


Figure 15: Jamovi interface for running the assumptions of an Independent Samples T-Test.

Assumptions of a t-test

- Data are continuous (i.e., not count or categorical data)
- Sample observations are a random sample from the population
- Sample means are normally distributed around the true mean

Check with a Normality test (Shapiro Wilk test)

If assumptions not met, we might consider a non-parametric alternative such as a Mann-Whitney U test.

Different types of t-test

Groups have equal variances:

- **Student's t-test**

Groups have inequal variances:

- **Welch's t-test**

Check for equal variances using a Homogeneity of Varianes Test (Levene's test)

Analysis of Variance (ANOVA)

A t-test can be used to test if two or more groups have significantly different means

Table 4: Wing lengths (mm) measured for five unnamed species of non-pollinating fig wasps at Site 96.

Het1	Het2	LO1	SO1	SO2
2.122	1.810	1.869	1.557	1.635
1.938	1.821	1.957	1.493	1.700
1.765	1.653	1.589	1.470	1.407
1.700	1.547	1.430	1.541	1.378

Do **all** species have the same mean ovipositor length?

[1] https://bradduthie.github.io/petiolaris_data/wasp_morphology.csv

ANOVA in jamovi

One-Way ANOVA (Fisher's)

	F	df1	df2	p
wing_length	3.24440	4	15	0.04176

Figure 16: Jamovi output for a one-way ANOVA of wing length measurements in five species of fig wasps collected in 2010 near La Paz in Baja, Mexico.

[1] https://bradduthie.github.io/stats/app/f_distribution/

ANOVA assumptions

Assumptions of ANOVA include the following:

1. Observations are sampled randomly
2. Observations are independent of one another
3. Groups have the same variance
4. Errors are normally distributed

Note, ANOVA is quite robust to these assumptions

Post hoc comparisons: which groups are different?

Tab below anova options in jamovi for multiple comparisons.

Post Hoc Comparisons - Species

Comparison			Mean Difference	SE	df	t	p	p_{tukey}	$p_{bonferroni}$
Species		Species							
Het1	-	Het2	0.17350	0.11862	15.00000	1.46270	0.164	0.600	1.000
	-	LO1	0.17000	0.11862	15.00000	1.43319	0.172	0.617	1.000
	-	SO1	0.36600	0.11862	15.00000	3.08557	0.008	0.050	0.075
	-	SO2	0.35125	0.11862	15.00000	2.96122	0.010	0.063	0.097
Het2	-	LO1	−0.00350	0.11862	15.00000	−0.02951	0.977	1.000	1.000
	-	SO1	0.19250	0.11862	15.00000	1.62288	0.125	0.506	1.000
	-	SO2	0.17775	0.11862	15.00000	1.49853	0.155	0.579	1.000
LO1	-	SO1	0.19600	0.11862	15.00000	1.65238	0.119	0.489	1.000
	-	SO2	0.18125	0.11862	15.00000	1.52803	0.147	0.561	1.000
SO1	-	SO2	−0.01475	0.11862	15.00000	−0.12435	0.903	1.000	1.000

Note. Comparisons are based on estimated marginal means

Figure 17: Jamovi output showing a table of 10 post hoc comparisons between species mean wing lengths for five different species of fig wasps.

Two-way ANOVA

We might have two different categorical independent variables.

Species	Tree	Wing Length
Het1	A	2.122
Het1	A	1.938
Het1	B	1.765
Het1	B	1.700
SO2	A	1.635
SO2	A	1.700
SO2	B	1.407
SO2	B	1.378

Just need to add our second category in jamovi

Testing three hypothesies

Testing for significance of (1) Species, (2) Tree, (3) Interaction

ANOVA - Wing length (mm)

	Sum of Squares	df	Mean Square	F	p
Species	0.24675	1	0.24675	45.75115	0.00249
Tree	0.16388	1	0.16388	30.38508	0.00529
Species ✻ Tree	0.00025	1	0.00025	0.04693	0.83909
Residuals	0.02157	4	0.00539		

Figure 18: Jamovi output table for a two-way ANOVA.

Because $P < 0.05$, Species and Tree are significant, but interaction is not.

Visualising a lack of interaction

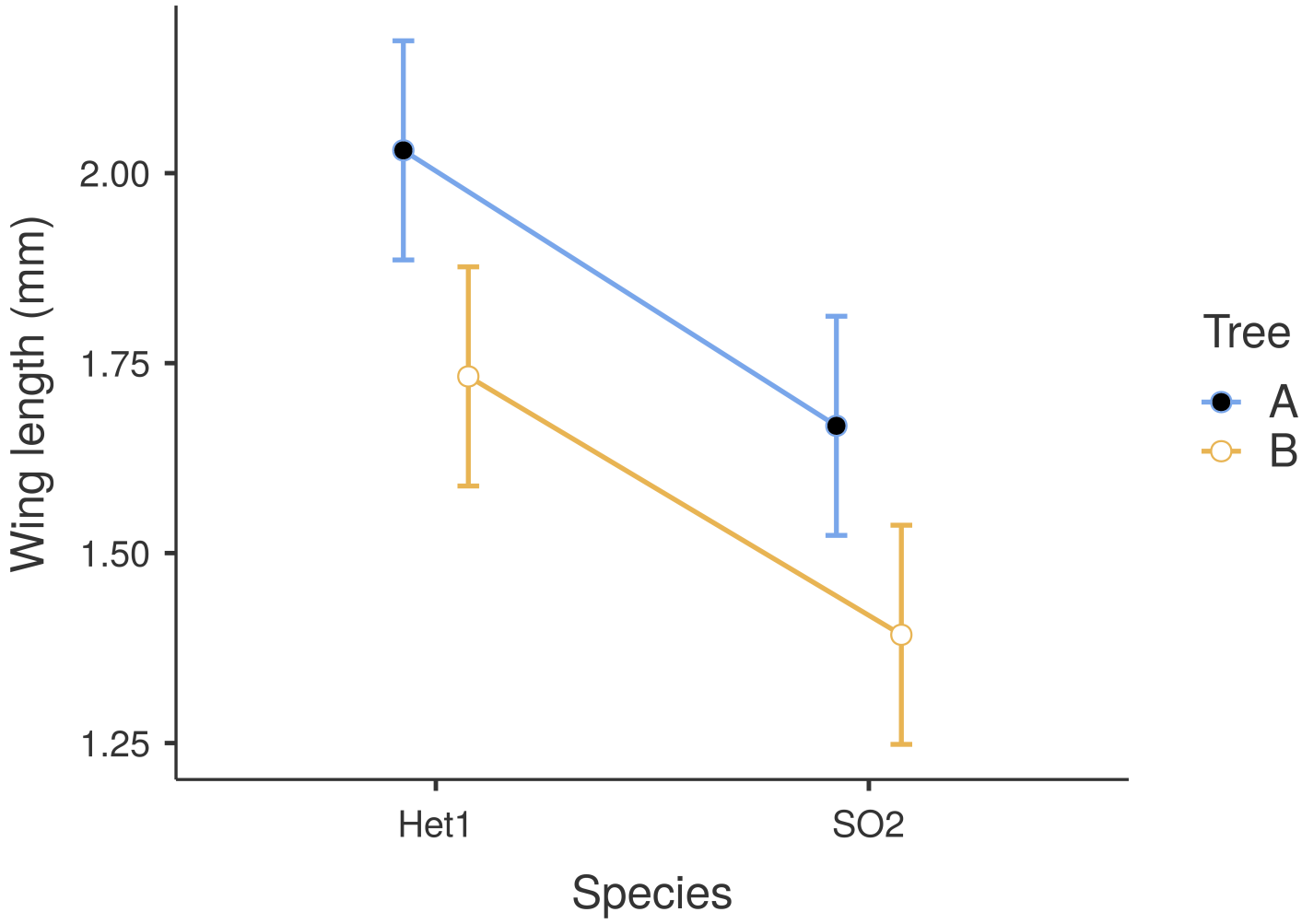


Figure 19: Interaction plot for a two-way ANOVA.

What if there was an interaction?

Now looking at two different species

Table 6: Wing lengths (mm) measured for two unnamed species of non-pollinating fig wasps.

Species	Tree	Wing Length
SO1	A	1.557
SO1	A	1.493
SO1	B	1.470
SO1	B	1.541
SO2	A	1.635
SO2	A	1.700
SO2	B	1.407
SO2	B	1.378

Testing a two-way ANOVA in jamovi again

ANOVA - Wing length (mm)

	Sum of Squares	df	Mean Square	F	p
Species	0.00044	1	0.00044	0.24509	0.64652
Tree	0.04337	1	0.04337	24.42590	0.00780
Species ✻ Tree	0.03264	1	0.03264	18.38492	0.01277
Residuals	0.00710	4	0.00178		

Figure 20: Jamovi output table for a two-way ANOVA.

Interaction plot when interaction is significant

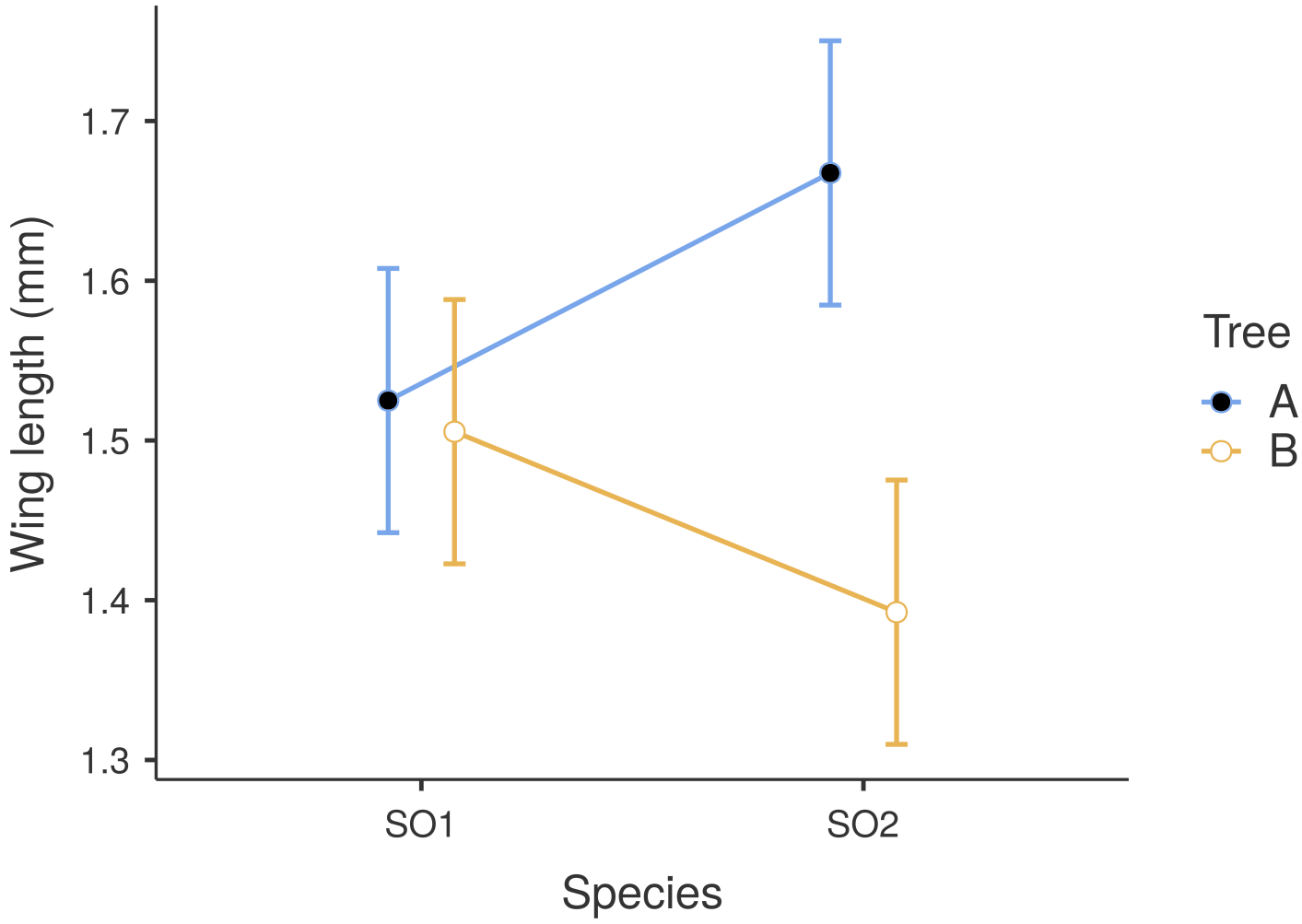


Figure 21: An interaction plot for a two-way ANOVA.

7. Chi-square test

- Chapter 29: Frequency and count data
- Chapter 31: *Practical*. Analysis of counts and correlation

7. Chi-square test

- Chapter 29: Frequency and count data
- Chapter 31: *Practical*. Analysis of counts and correlation

- **Independent variable(s)**: Categorical
- **Dependent variable**: Categorical

Hypothesis tests

- **Goodness of fit**: Test if each category is equally probable to sample
- **Test of association**: Test if different category types are associated

Table of counts (not tidy)

Table 7: Counts ($N = 60$) from a mobile game called 'Power Up!'.

	Small	Medium	Large
Android	8	16	6
macOS	10	8	12

Chi-square Goodness of Fit test in jamovi

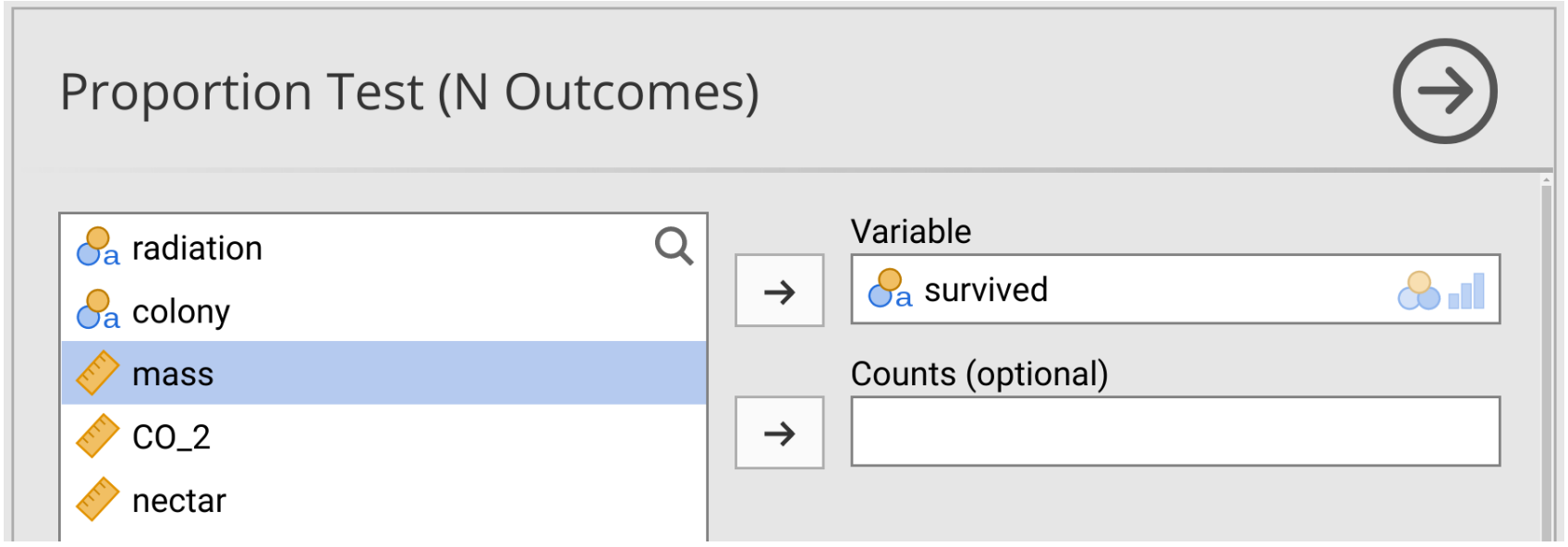


Figure 22: Jamovi interface for running a Chi-square goodness of fit.

Chi-square Test of Association in jamovi

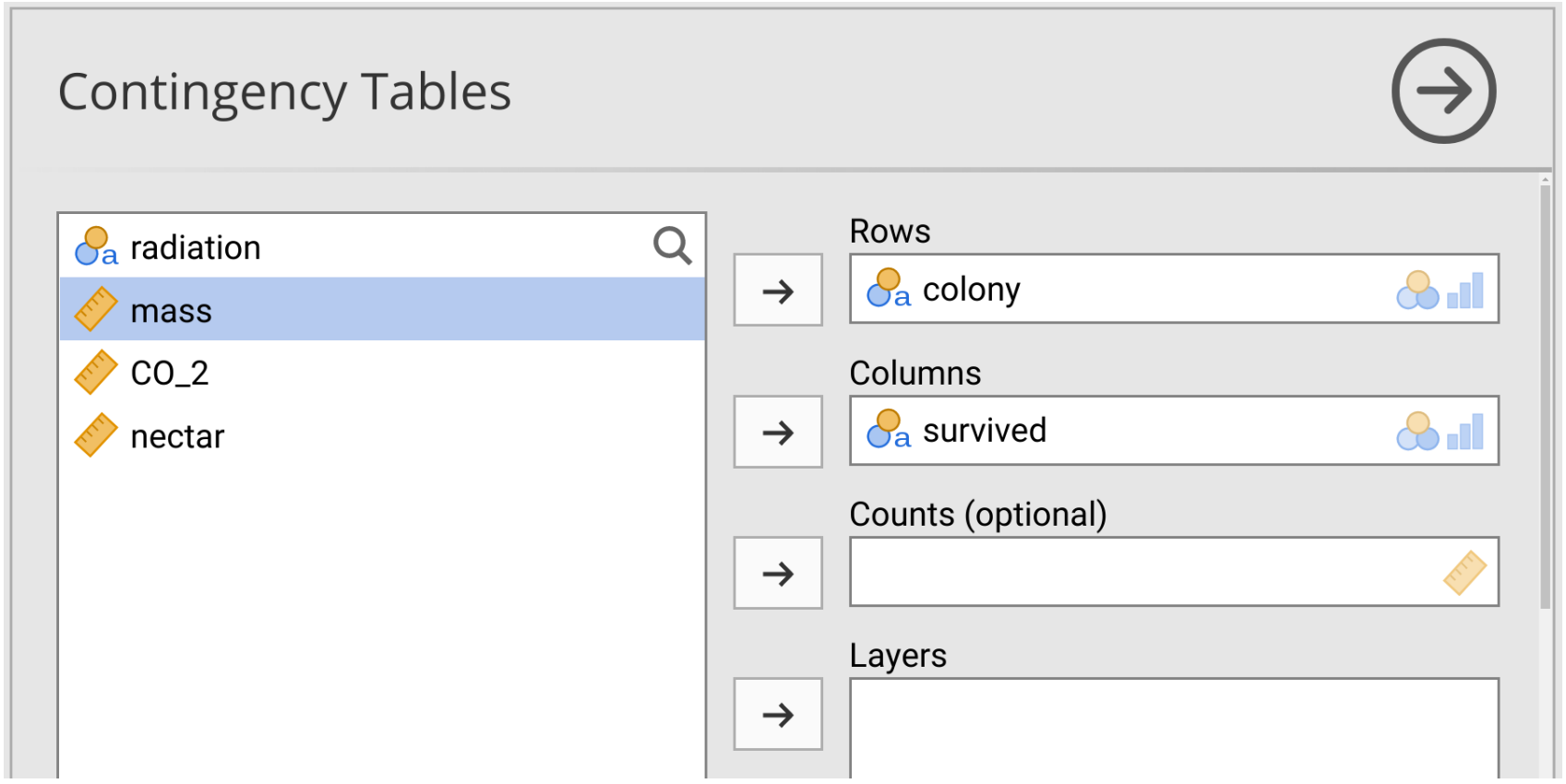


Figure 23: Jamovi interface for running a Chi-square test of association.

8. Correlation and regression

- Chapter 30: Correlation
- Chapter 31: *Practical*. Analysis of counts and correlation
- Chapter 32: Simple linear regression
- Chapter 33: Multiple linear regression
- Chapter 34: *Practical*. Using regression

8. Correlation and regression

- Chapter 30: Correlation
- Chapter 31: *Practical.* Analysis of counts and correlation
- Chapter 32: Simple linear regression
- Chapter 33: Multiple linear regression
- Chapter 34: *Practical.* Using regression

- **Independent variable(s)**: Quantitative
- **Dependent variable**: Quantitative

Introduction to correlation

We often want to investigate the relationship between pairs of variables.

- Vegetation height and mean annual temperature
- Animal body size and metabolic rate
- Number of automobiles in a location and carbon emissions

The **correlation** between pairs of variables, such as those listed above, describes how the variation of each variable is related to the other variable.

Visualising the correlation between two variables

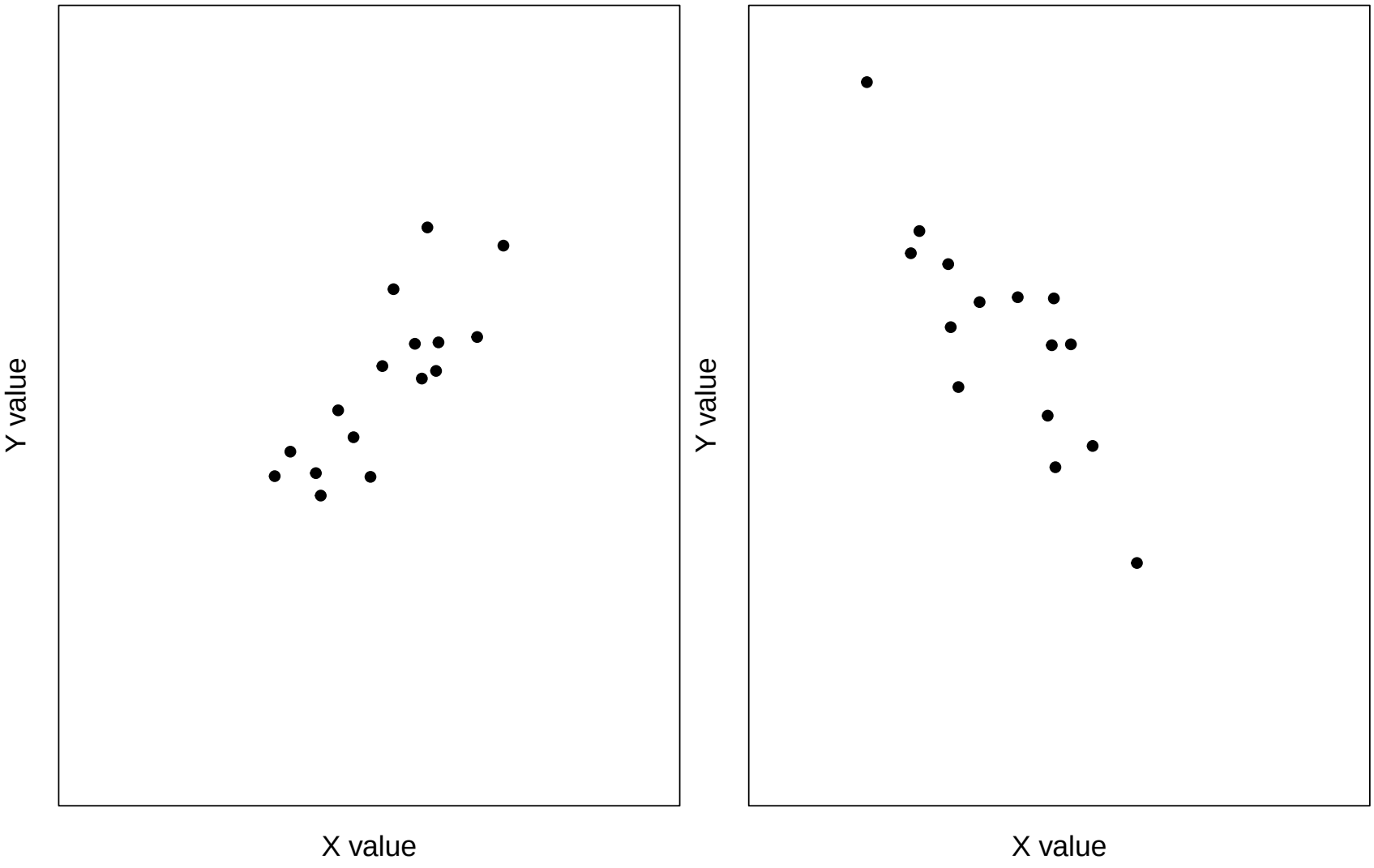


Figure 24: Two plots of hypothetical variables illustrating a positive (left) and negative correlation

Visualising two variables that are not correlated

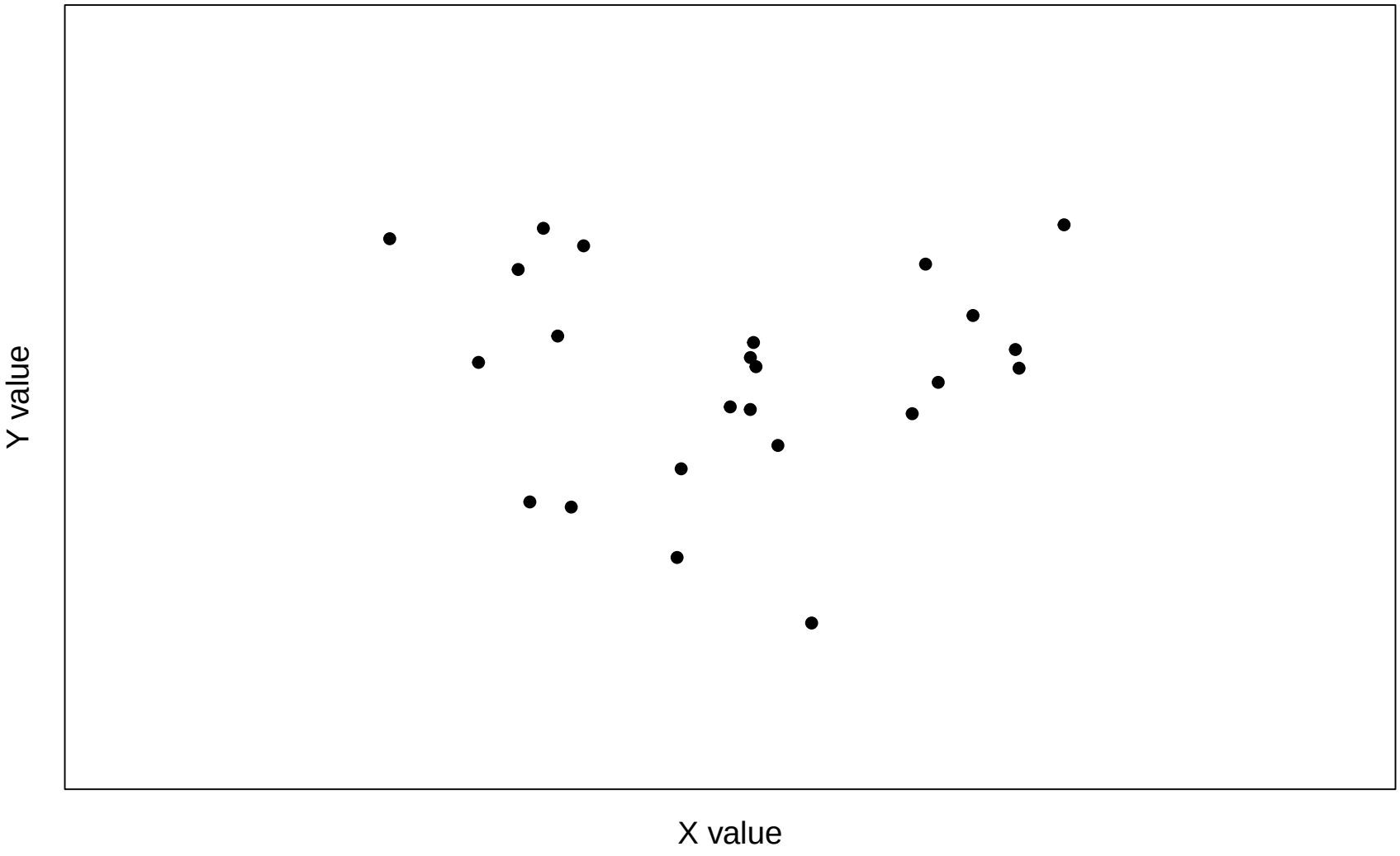


Figure 25: A plot of two hypothetical variables that are not correlated.

Getting a more intuitive sense of correlation

Formalised with the **correlation coefficient** (r)

- Provides a statistical measure of strength and direction of correlation
- Only describes association between variables (**not** cause and effect)

Getting a more intuitive sense of correlation

Formalised with the **correlation coefficient** (r)

- Provides a statistical measure of strength and direction of correlation
- Only describes association between variables (**not** cause and effect)

The value r ranges between -1 and 1

- Negative numbers indicate a negative correlation
- Positive numbers indicate a positive correlation
- Avalue of zero indicates no correlation

We can get a more intuitive understanding of the correlation coefficient with **[this application]**.

Testing if or not a correlation is significant

We often want to test whether or not the correlation between two variables is significant

- Test of Pearson product moment correlation assumes variables are normally distributed
- Test of Spearman's rank correlation coefficient (i.e., correlation of ranks) does not assume normality

To test whether or not two variables are correlated, we first must test the null hypothesis that the two variables are normally distributed.

Testing if soil depth and root density are correlated

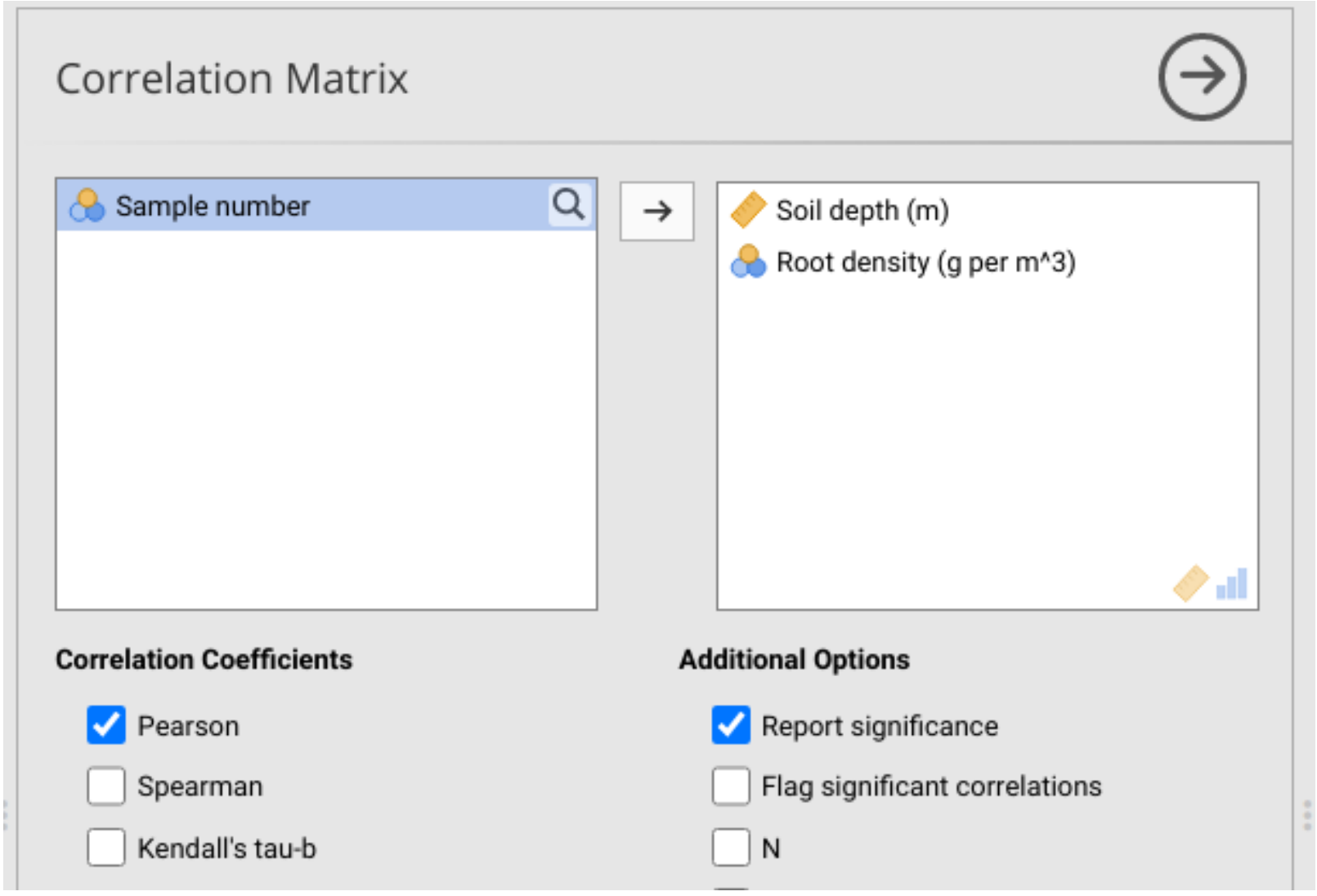


Figure 26: Jamovi box showing how to run a test of the correlation coefficient.

Testing whether soil depth and root density are correlated

A table of output that looks like the one below.

Correlation Matrix

Correlation Matrix

		Soil depth (m)	Root density (g per m^3)
Soil depth (m)	Pearson's r	—	
	p-value	—	
Root density (g per m^3)	Pearson's r	−0.84257	—
	p-value	0.00221	—

Figure 27: Jamovi table showing output of a parameteric test of the significance of a correlation coefficient.

Linear regression

Suppose we want to predict fruit volume from latitude

Table 8: Volumes (mm^3) of fig fruits collected from different latitudes.

Latitude	23.7	24.0	27.6	27.2	29.3	28.2	28.3
Volume	2399.0	2941.7	2167.2	2051.3	1686.2	937.3	1328.2

Visualising regression

Suppose we want to predict fruit volume from latitude

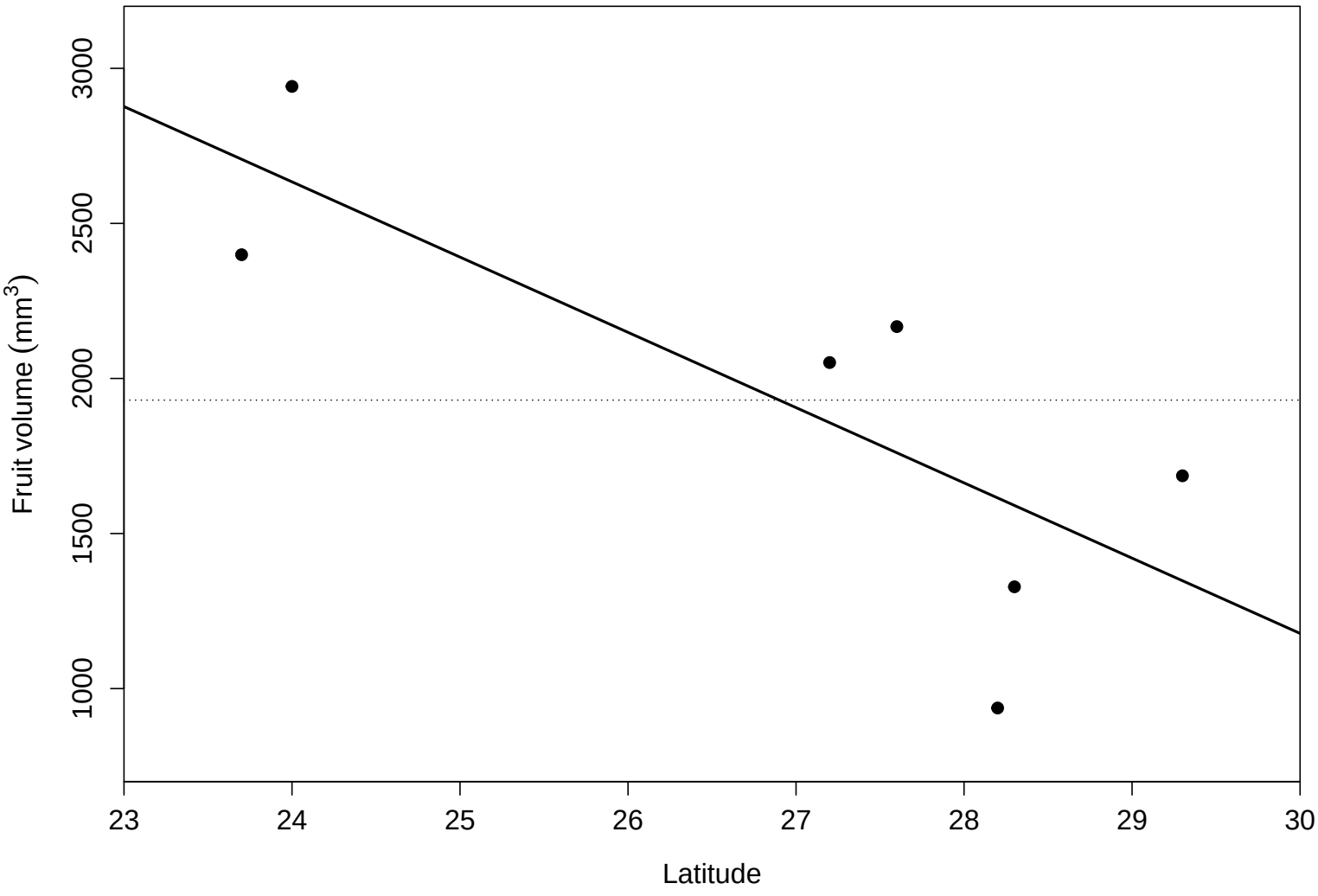


Figure 28: Latitude and fruit volume for seven fig fruits.

Intercepts, slopes, and residuals

Suppose we want to predict fruit volume from latitude

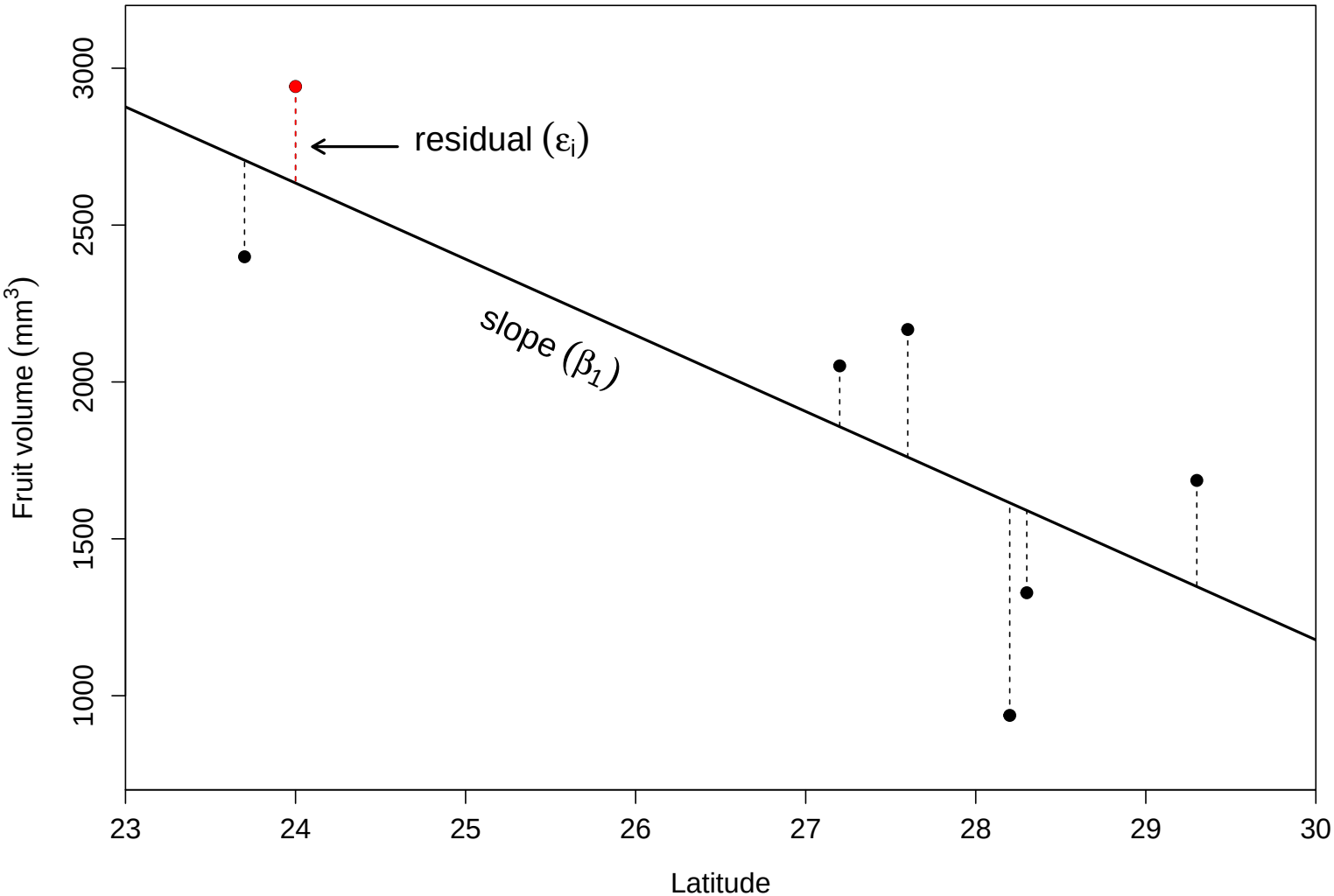


Figure 29: Latitude and fruit volume for seven fig fruits.

[1] https://bradduthie.github.io/stats/app/regr_click/

Regression coefficients

Simple linear regression predicts the dependent variable (y) from the independent variable (x) using the intercept (b_0) and the slope (b_1),

$$y = b_0 + b_1 x.$$

Regression coefficients

Simple linear regression predicts the dependent variable (y) from the independent variable (x) using the intercept (b_0) and the slope (b_1),

$$y = b_0 + b_1 x.$$

For any specific value of x_i, the corresponding y_i can be described more generally,

$$y_i = b_0 + b_1 x_i + \epsilon_i.$$

Regression line calculation

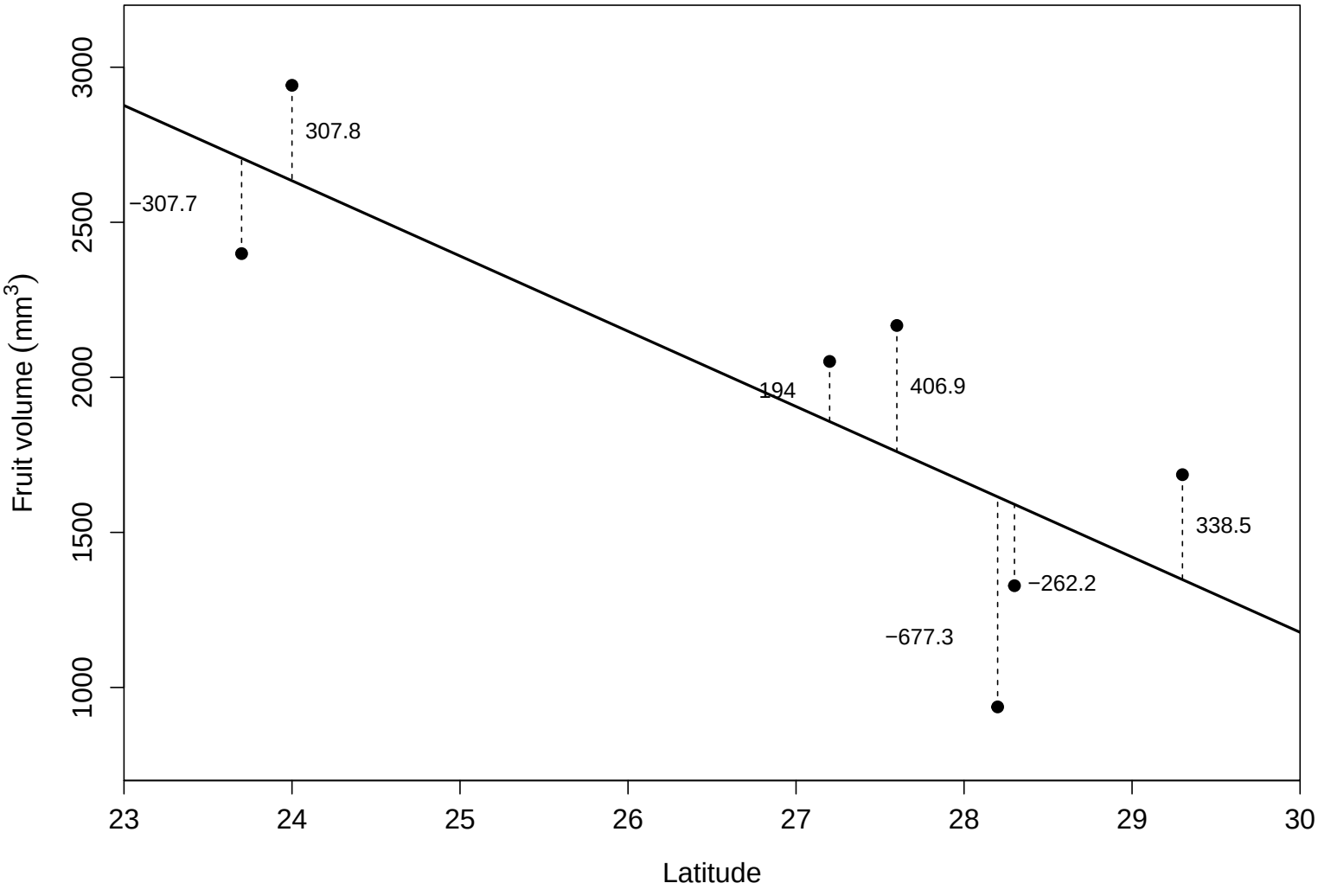


Figure 30: Latitude and fruit volume for seven fig fruits.

Regression assumptions

1. Measurement of the independent variable (x) is completely accurate.
2. The relationship between the independent and dependent variables is linear.
3. For any value of x_i, y_i values are independent and normally distributed.
4. For all values of x, the variance of residuals is identical.

Running a linear regression in jamovi

Test if slope and intercept different from zero

Model Coefficients - Volume

Predictor	Estimate	SE	t	p
Intercept	8458.30961	2293.12354	3.68855	0.01417
Latitude	−242.68331	85.00625	−2.85489	0.03562

Figure 31: Jamovi output table for a simple linear regression showing model coefficients and their statistical significance.

Coefficient of determination

We often want to know how well a regression line fits to the data.

- **Coefficient of determination** (R^2)
- R^2 tells us **how much of the variation in y is explained by the regression equation**
- E.g., if $R^2 = 0.83$, then 83% of the variation in y is accounted for by the fitted regression line
- Visually, how tightly the data points in a scatterplot fit to the regression line

Variation in y explained by x

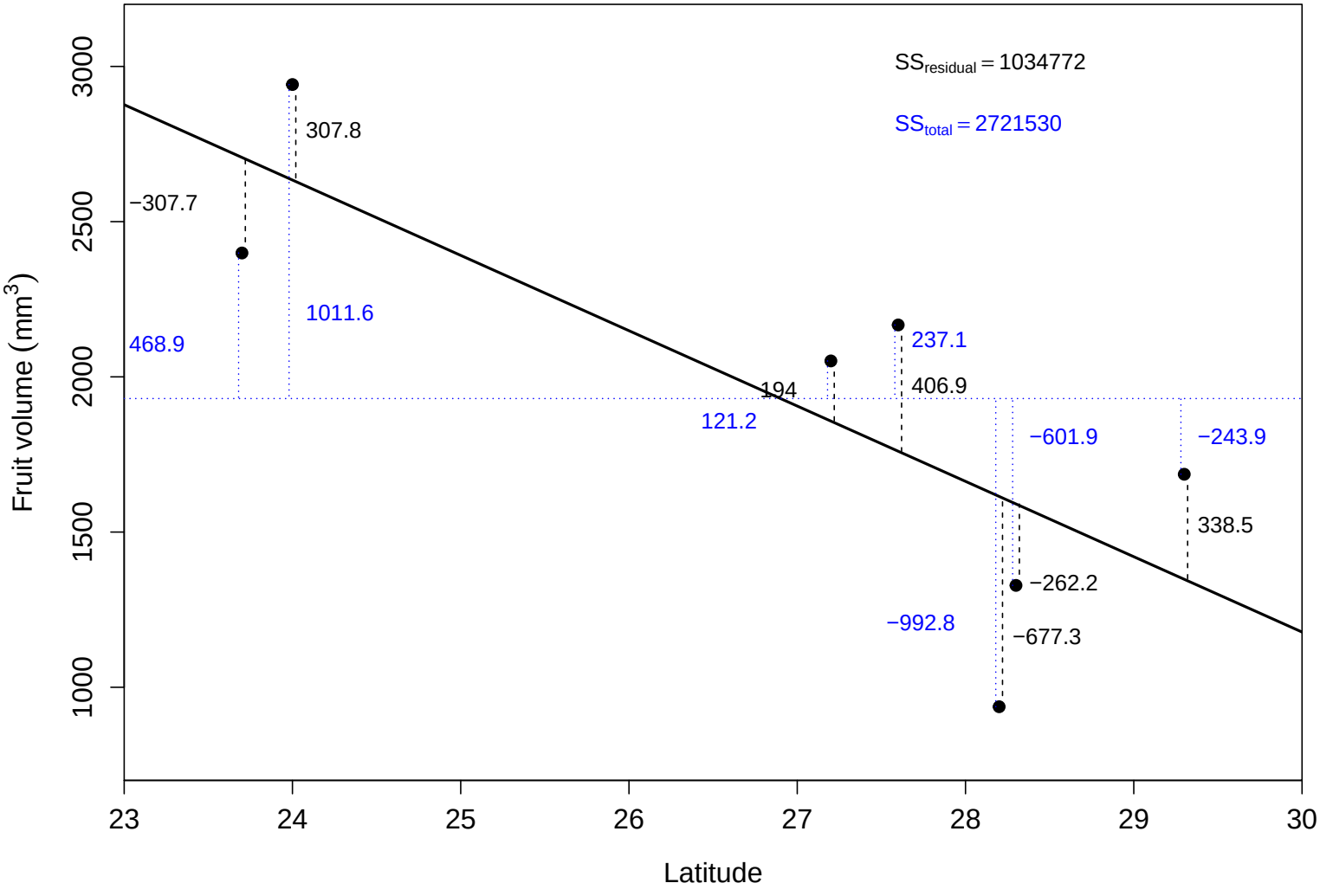


Figure 32: Latitude and fruit volume for seven fig fruits.

Model fit in jamovi

Model Fit Measures

			Overall Model Test			
Model	R	R^2	F	df1	df2	p
1	0.78726	0.61978	8.15039	1	5	0.03562

Figure 33: Jamovi output table for a simple linear regression in which latitude is an independent variable and fig fruit volume is a dependent variable.

Summary of tests: Very rough overview

Independent	Dependent	Test
Categorical	Quantitative	t-test, ANOVA
Categorical	Categorical	Chi-square
Quantitative	Quantitative	Regression

There are a lot more tests to consider; this is just the beginning!

Resources with this lecture

Book, audiobook, datasets:

https://bradduthie.github.io/books.html

Email:

alexander.duthie@stir.ac.uk

Advanced Skills:

https://stirlingcodingclub.github.io/linear_modelling/

https://stirlingcodingclub.github.io/mixed_modelling/