17C Laboratory & Professional Skills: Data Analysis

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

Week 3: Hypothesis testing, data types, reading data in to R and saving figures in reports

- Last week…
  - Why we need statistics: Are these results, or ones as probable or less probable, so unlikely that we suspect an effect?
- This week…
  - More on the logic and what governs the type of test we use?
  - Data types

Slide from last week:

The logic of ‘hypothesis’ testing

- Have a ‘null’ hypothesis’
- Calculate probability of getting your data if that null hypothesis is true
- If the probability is less than 0.05 reject the null hypothesis

- Frequentist/classical statistics

Summary of this week

- We will consider how we can classify variables in terms of the type of values they can take and their role in analysis and the impact these have on the tests that we conduct.
- In RStudio we will cover reading in data files of various formats, data types, summarising and plotting data. We also cover saving figures and laying out a report in word.

Learning objectives for the week

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

- to able to explain what response and explanatory variables are, distinguish between data types and describe how these impact choice of test (MLO 1 and 2)
- demonstrate the process of hypothesis testing with an example and evaluate potential inferences (MLO 1 and 2)
- read in data in to RStudio, create simple summaries and plots using manual pages where necessary (MLO 3)
- create neat reports in Word which include text and figures (MLO 4)

Science – generalisation

population

Impossible to measure

sample

Possible to measure

We draw inferences about the population(s) from the sample(s) based on statistics

Uses of statistics

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)

Uses of statistics

1. Estimation
   - what is the mean of the population?

L04 and W04: Describing normal distributions and Confidence Intervals

1. 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)

L03 and W03; L05 and W05 to L08 and W08

Regardless, the choice of statistic depends on ….

1. Type of data

The type of values a variable can take: Discrete or continuous?

2. Their role in the analysis

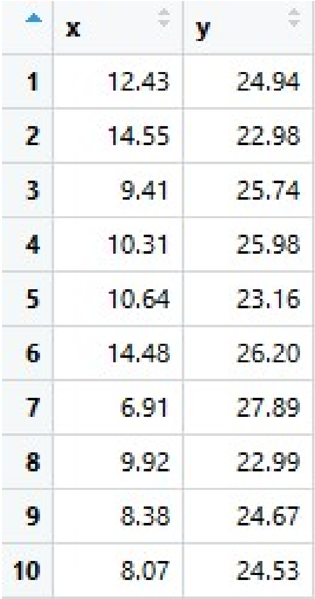
Which is the response and which is/are explanatory?

Overview

- ‘Experiments’

Some things we control, choose or set

Independent variables
Explanatory variables
The ‘x’ s



	x	y
1	12.43	24.94
2	14.55	22.98
3	9.41	25.74
4	10.31	25.98
5	10.64	23.16
6	14.48	26.20
7	6.91	27.89
8	9.92	22.99
9	8.38	24.67
10	8.07	24.53

Something we measure

Dependent variables
Response variables
The ‘y’ s

Which variable is the response? (2)
Which variables are explanatory? (2)

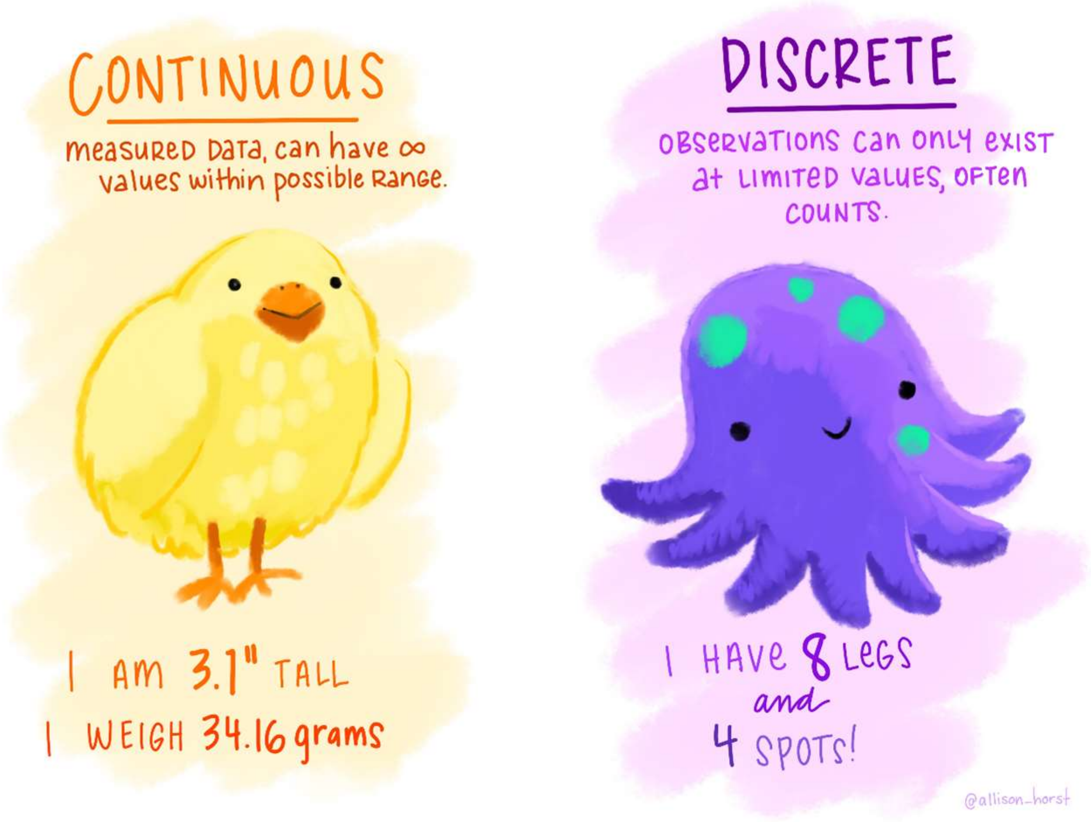
What kind of values can they take? (1)

The choice of statistic depends on:

Type of data

Two main types

- discrete
- continuous


CONTINUOUS
measured data, can have ∞ values within possible range.
I AM 3.1" TALL
I WEIGH 34.16 grams
DISCRETE
observations can only exist at limited values, often counts.
I HAVE 8 LEGS
and
4 SPOTS!
@allison_horst

The choice of statistic depends on:

Type of data - discrete

Discrete

- Categories (not quantitative)
- Counts (quantitative but discrete)

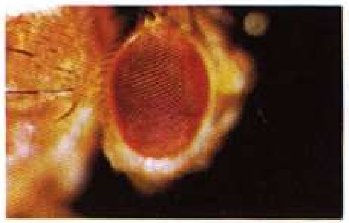
The choice of statistic depends on:

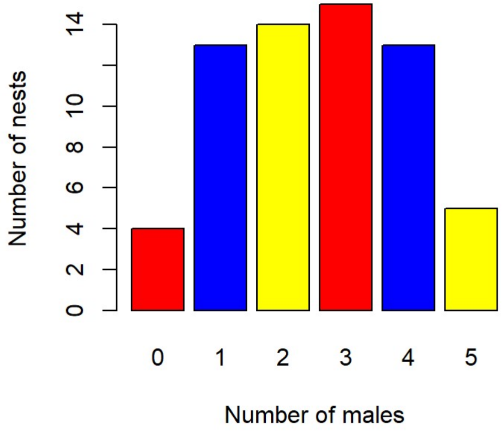
Type of data - discrete

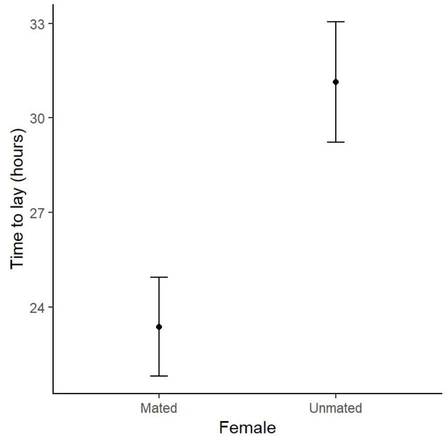
- Categories

No scale e.g., colour, species

Often an 'explanatory' variable







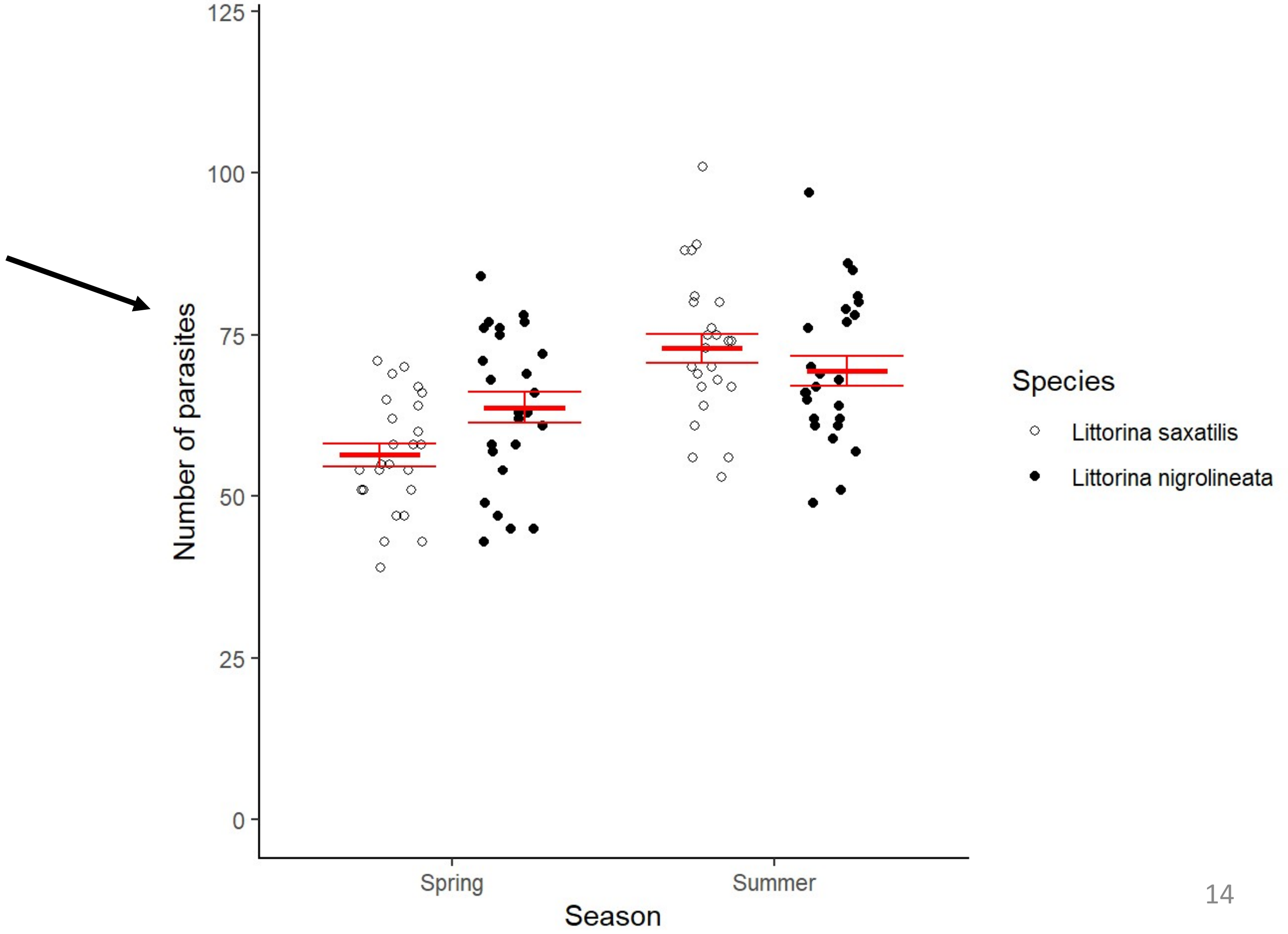
Category

The choice of statistic depends on:

Type of data - discrete

- Counts

Normally a 'response' variable



The choice of statistic depends on:

Type of data - continuous

- e.g., length, height, concentration
- Infinite number of possible values
- Can be a response or an explanatory

The choice of statistic depends on:

Type of data

- Theory vs practice
- Limit of measurement

Numbers of hairs on head: discrete but can be treated as continuous

Height to nearest metre: continuous but discretised by measurement

The choice of statistic depends on ….

1. Type of data

   What kind of values? Discrete or continuous? ✓

2. Their role in the analysis

   Which is the response and which are the explanatory

   What is the relationship between them?

Rest of the module!

Hypothesis Testing: deeper

- Set up H_0 “no effect”
- Test generates a test statistic from data (a summary)
- Converted to a probability (*p*-value) = prob of data if H_0 is true
- $p \leq 0.05$ reject H_0; $p > 0.05$ do not reject H_0

Hypothesis Testing – relationship to L1 example

- Set up H_0

There is no effect of maternal poverty on birthweight

Hypothesis Testing – relationship to L1 example

- Test generates a test statistic from the data
- ‘Converted’ to a probability (*p*-value) = Probability of getting a test statistic of that size or as extreme or more extreme if H_0 is true

Hypothesis Testing – relationship to L1 example

Compare our *p*-value to 0.05

$p \leq 0.05$ reject H_0

$p > 0.05$ do not reject H_0

In that example, our *p*-value was 0.096
Thus: We do not reject the null hypothesis.

Our sample is consistent with poverty having no effect.

Hypothesis Testing:

The *p*-value

- Probability of result if null hypothesis true

  if we calculate $p = 0.45$ we can expect results as extreme or more extreme as those we observe 45% of the time.

- 0.05 is the crucial level
- If $p \leq 0.05$. We reject the null hypothesis
- And conclude there is a significant difference between our sample and what we would expect if there was no effect

Hypothesis Testing:

Type 1 and type 2 errors

Inherent in the approach - not ‘mistakes’ you can prevent

Decision after testing	(unknown) True state of H_0	
	True	False
Reject (evidence it is false)	Type 1 error	Correct
Do not reject (no evidence it is false)	Correct	Type 2 error

Hypothesis Testing:

Type 1 and type 2 errors

For our birthweight example…..p > 0.05 (0.096)

Decision after testing	(unknown) True state of H_0	
	True	False
Reject (evidence it is false)	Type 1 error	Correct
Do not reject (no evidence it is false)	Correct	Type 2 error

Learning objectives for the week

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

- to able to explain what response and explanatory variables are, distinguish between data types and describe how these impact choice of test (MLO 1 and 2)
- demonstrate the process of hypothesis testing with an example and evaluate potential inferences (MLO 1 and 2)
- read in data in to RStudio, create simple summaries and plots using manual pages where necessary (MLO 3)
- create neat reports in Word which include text and figures (MLO 4)