Emma Rand
Data Analysis in R

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

Summary of this week

- We will consider how we can classify variables in terms of the type of values they can take and the logic of hypothesis testing with an example.
- In RStudio we will cover reading in data files, 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 material and carrying out the independent study the successful student will be able to:

- distinguish between data types (MLO 2)
- demonstrate the process of hypothesis testing with an example (MLO 1)
- Explain type 1 and type 2 errors (MLO 4)
- 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)

Choosing data analysis methods

In Data Analysis in R you will learn several methods for statistically analysing data.

Here we start to consider how we make appropriate choices.

It’s a journey!

The choice of test 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 test depends on:

Type of data

Two main types

- discrete
- continuous

The choice of test depends on:

Type of data - discrete

Discrete

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

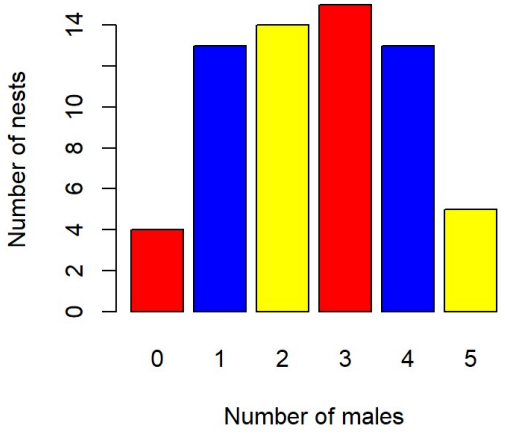
The choice of test depends on:

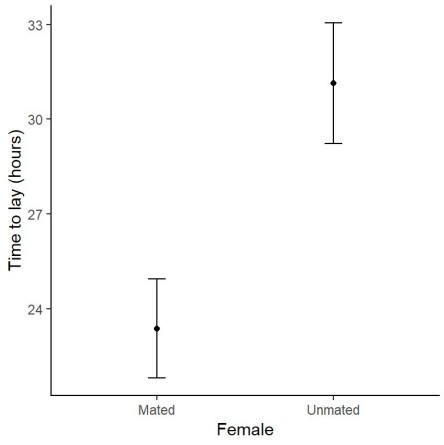
Type of data - discrete

- Categories

  No scale e.g., colour, species

  Often an ‘explanatory’ variable





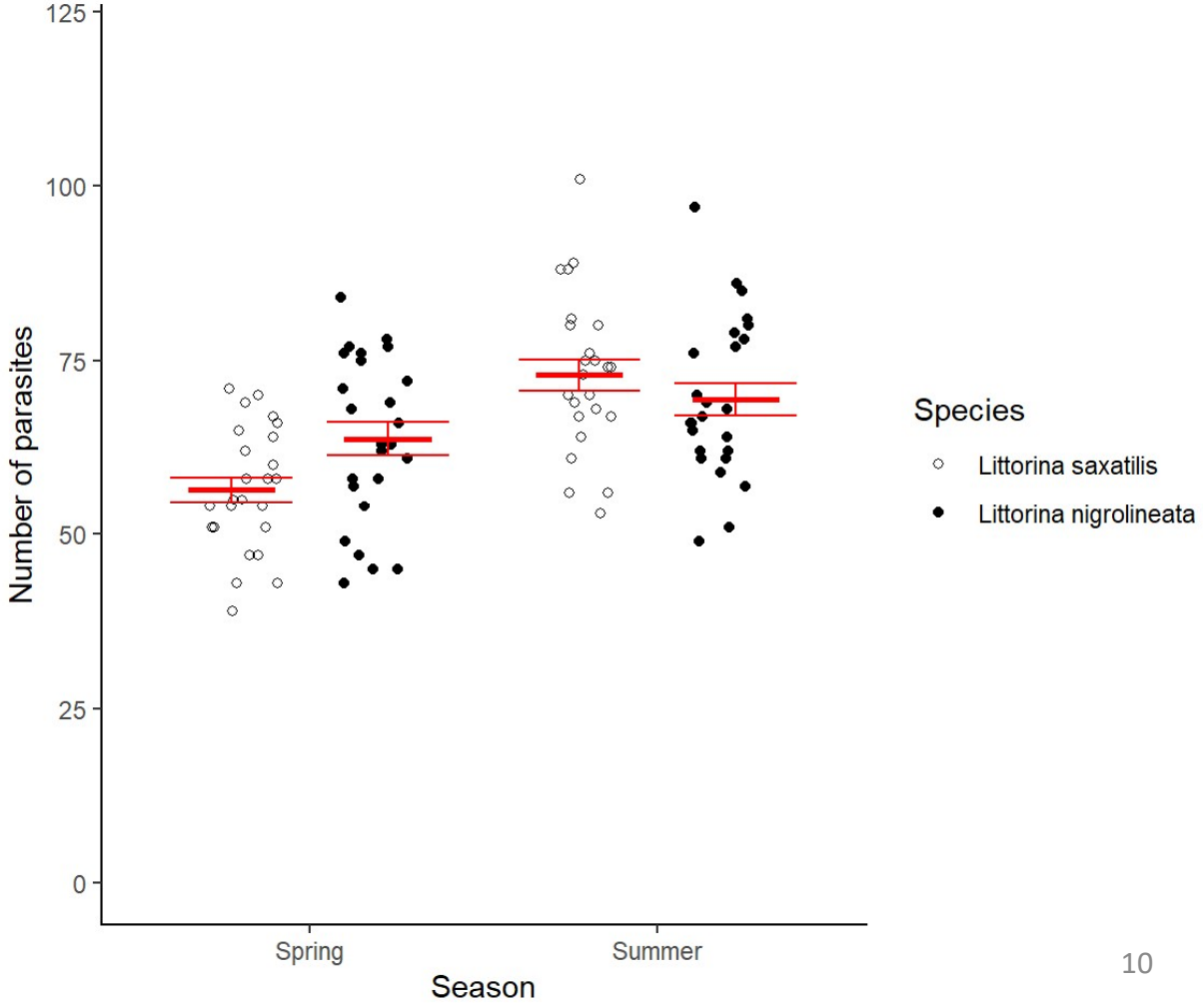
Category

The choice of test depends on:

Type of data - discrete

- Counts

  Normally a ‘response’ variable



The choice of test 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 test depends on:

Type of data

- Theory vs practice
- Limit of measurement

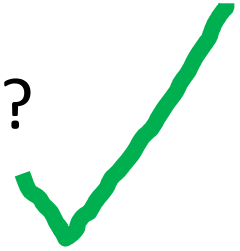
Numbers of hairs on head: discrete but can be treated as continuous

Height to nearest metre: continuous but discretised by measurement

The choice of test 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!

R 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

Hypothesis Testing: steps

1. Set up H_0 “no effect”
2. Collect data
3. Determine the probability of our data if H_0 is true
4. $p \leq 0.05$ reject H_0; $p > 0.05$ do not reject H_0

Hypothesis Testing example

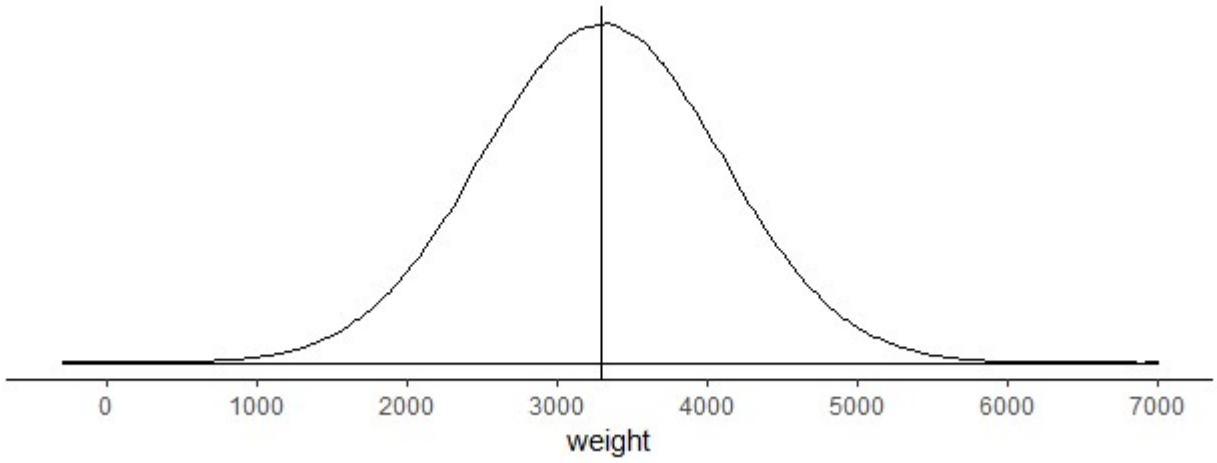
Question: National average birthweight is 3300 grams with an s.d. = 900 grams. Does maternal poverty influence birthweight?

1. Set up $H_{0:}$ There is no effect of maternal poverty on birthweight

The null hypothesis. H_0

What you expect to happen if nothing interesting biologically is occurring.

We would expect a mean of 3300 if poverty has no effect.



Hypothesis Testing example

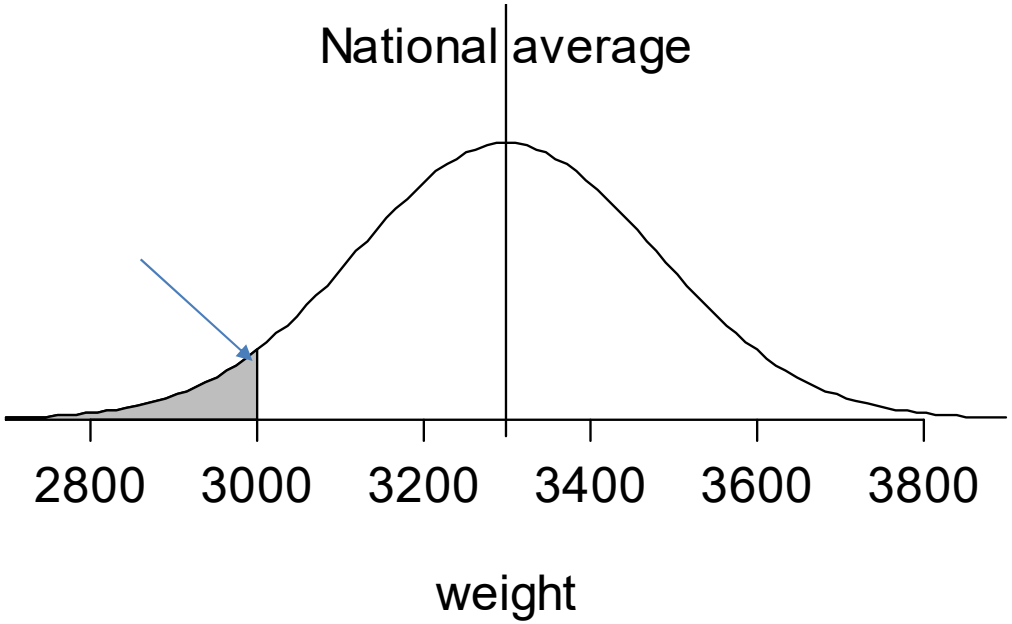
2. Collect data:

We take a sample of 25 women who live in poverty and determine the birthweight of their baby.

The mean, $\bar{x}$=3000 grams

This is lower than the national average but is that enough?

How far is too far? Distributions



Hypothesis Testing example

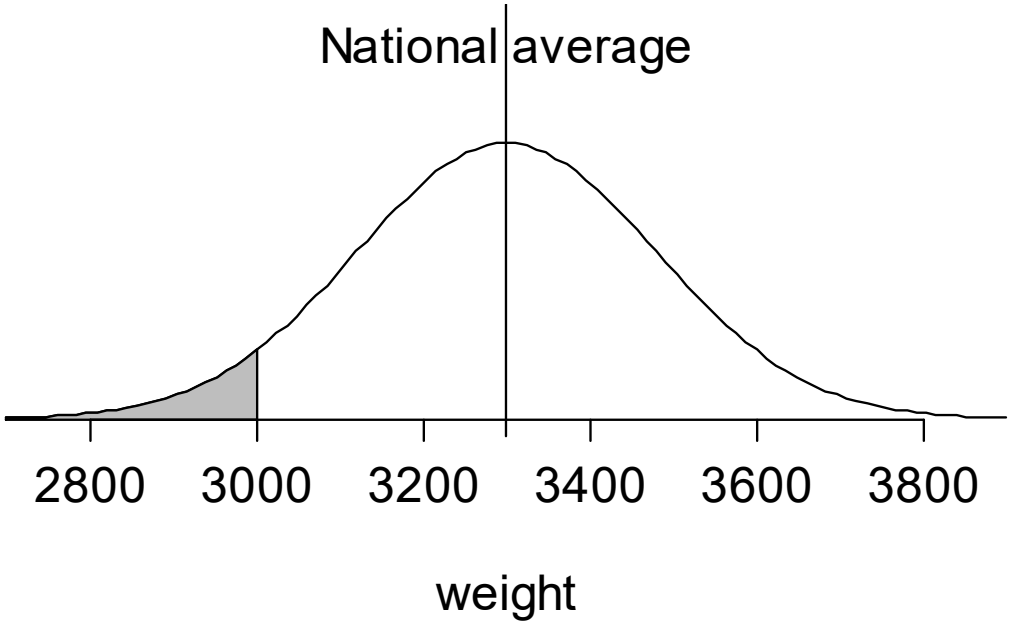
3. Determine the probability of our data if H_0 is true

How far is too far? Distributions

We calculate the probability of 3000 g if we expect 3300 g on average

What is P(3000) *or lower* from a distribution with mean 3300

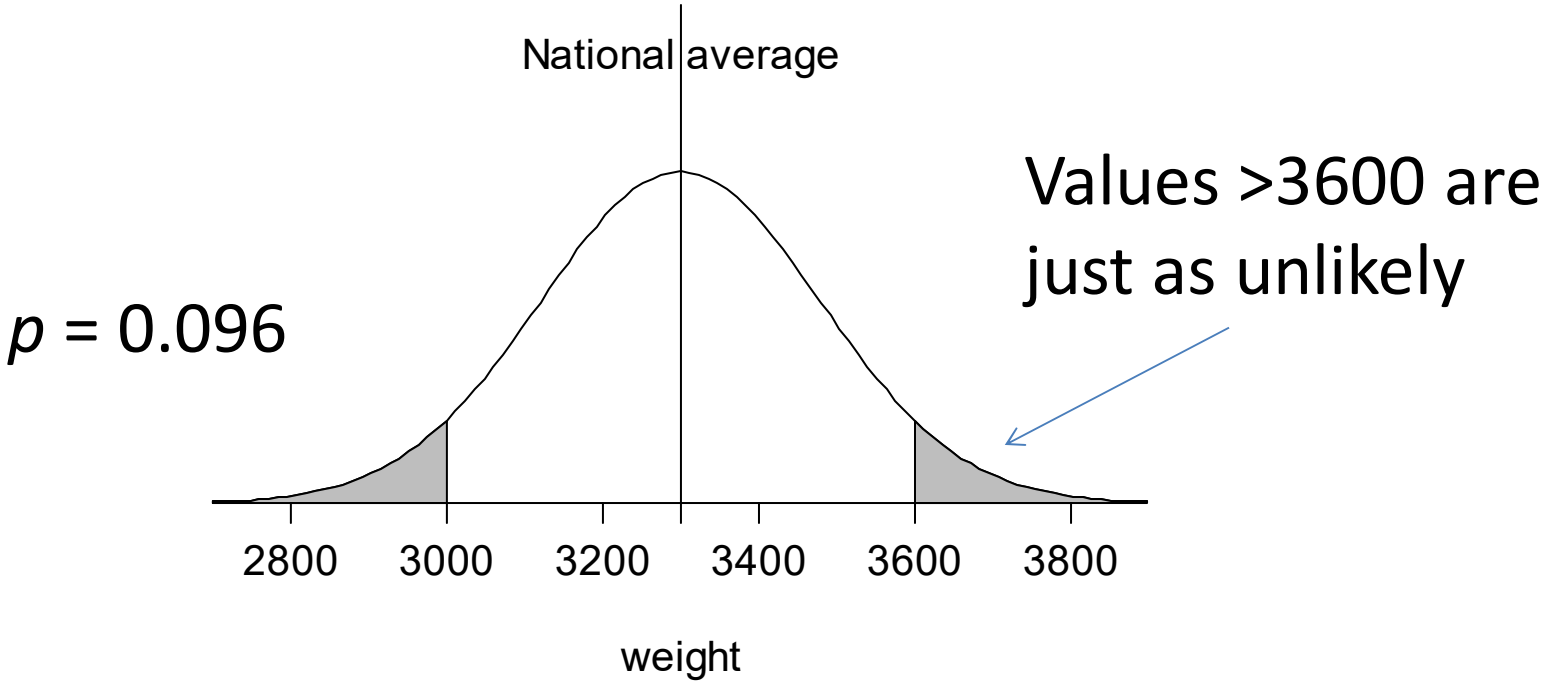
$p = 0.048$



Don’t worry how p was calculated

But more appropriately

What is P(3000) or a mean *as unlikely or more unlikely* from a distribution with mean 3300?



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

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 example

4. $p \leq 0.05$ reject H_0; $p > 0.05$ do not reject H_0

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 data if null hypothesis true

  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