

STAT 110: Week 1

University of Otago

What is statistics?

- Learning from data
- What do statisticians do?
 - ▶ Examples
- Wikipedia:
 - ▶ "Statistics is the discipline that concerns the collection, organization, analysis, interpretation, and presentation of data" ¹
 - ▶ It's all about data.
 - ▶ Learning from data involves all of those concepts

¹I got a similar answer when I asked ChatGPT.

Data

- Data is all around us
 - ▶ It informs us about the natural world, business, society, ...
- In the past, data sets tended to be small
 - ▶ Data was expensive to collect (it often still is!)
 - ▶ Much was done with pen and paper
- It is now common to have large data sets
 - ▶ Computing is an essential part of modern statistics

Then and now

- To illustrate the differences, compare two data visualisations (both electoral)
 - ▶ One from 1975
 - Three parties
 - Points inside the triangle relate to probabilities of various parties finishing 1st
 - Contour plot (like a topo map)
 - Very confusing – limited by technology
 - ▶ One more recent
 - County level: which presidential candidate got most votes (2020 presidential election)
 - Republican (red), or democrat (blue)
 - One plot based on population
 - One plot based on (land) area

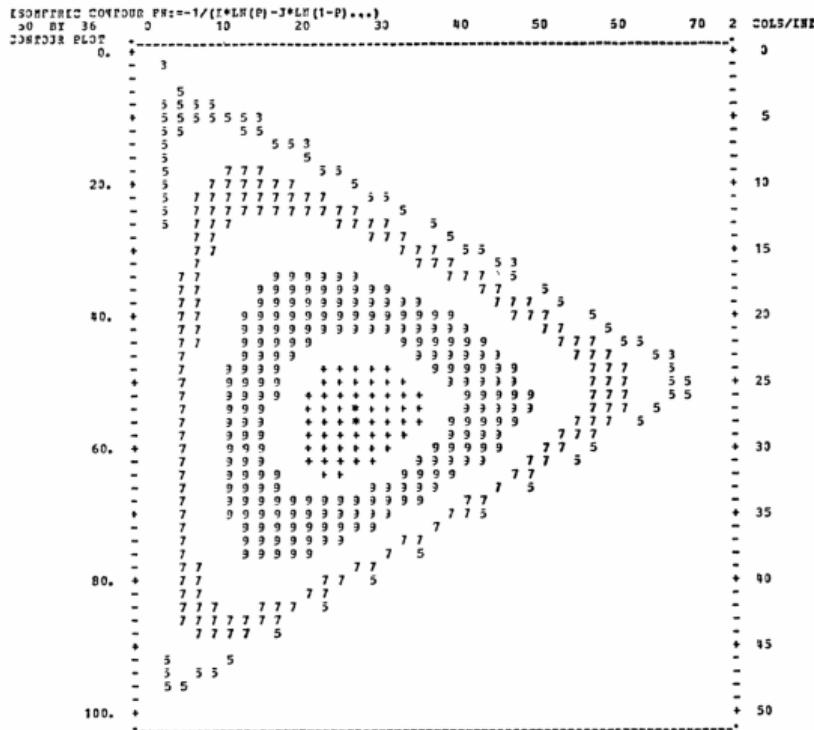
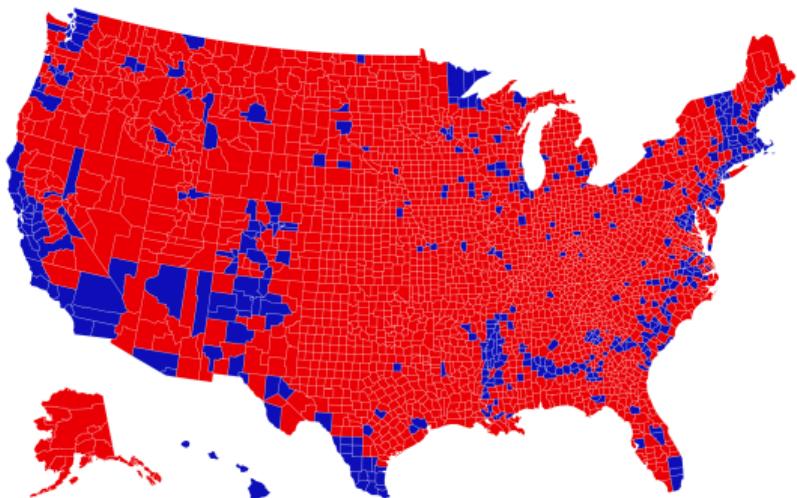
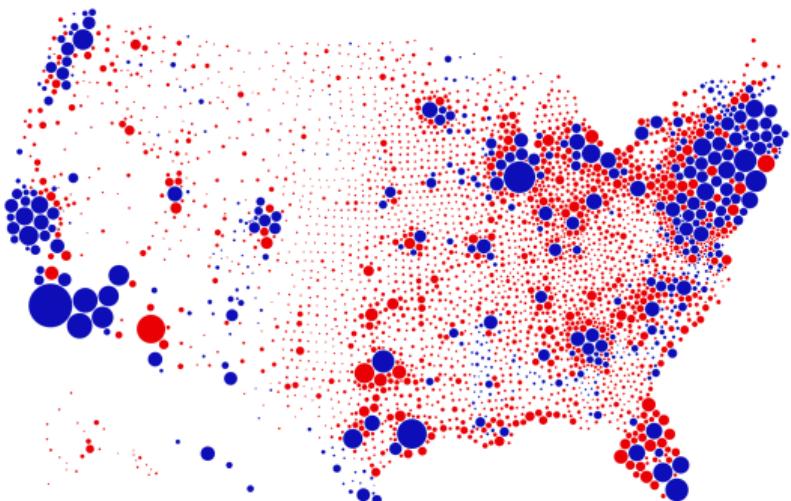


FIG. 1. Isometric contour plot for election data.

The ratio $(f - f_{\min}) / (f_{\max} - f_{\min})$ is indicated by contour zone codes as follows.

3	4	5	6	7	8	9	+	*		
25%	35%	45%	55%	65%	75%	85%	95%	97.5%	99.9%	100%

¹From: Plackett (1975); JRSS C (24); p. 193-202



¹From [@karim_douieb](#)

Data

- We will see a lot of data in this course
 - ▶ Lectures
 - ▶ Tutorial exercises
 - ▶ Assignments
- Data comes from a variety of sources
 - ▶ Variety of subject areas
 - ▶ You will hopefully see examples that are from your area of interest
 - ▶ See examples from other scientific areas

Statistics is about ... variation and uncertainty?

- In science (and life!), we can rarely be certain
- Statistics: trying to describe and predict scientific process
 - ▶ Using data
 - ▶ In the presence of uncertainty
- We will be talking about variability and uncertainty a lot
 - ▶ Understand (or describe) variability
 - ▶ Control sources of variability
 - ▶ Quantify uncertainty where possible
 - ▶ Make use of probability
 - (Mathematical) language of uncertain events
 - ▶ Look at an example

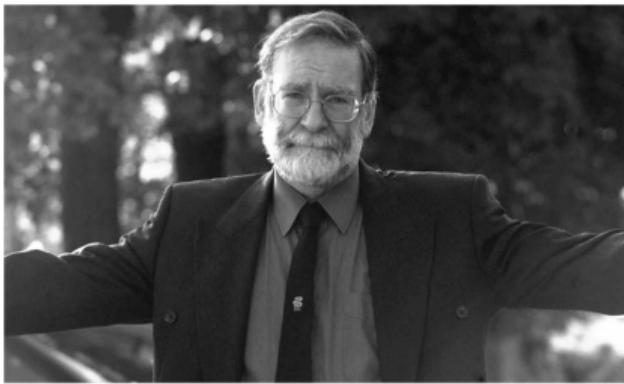
Example 1: Harold Shipman

- Harold Shipman was a notorious serial killer
- He had 215 confirmed kills
 - ▶ A further 45 suspected kills
- He was a British GP
- His victims were predominantly:
 - ▶ Older
 - ▶ Female
- What does this have to do with statistics?

Shipman's statistical legacy

Harold Shipman, who in January committed suicide in prison, has become notorious the world over as one of the most prolific serial killers of all time. His case has also seriously dented public confidence in doctors. **David Spiegelhalter** and **Nicky Best** explain how industrial quality control techniques could be adapted to signal when death rates among a doctor's patients are surprisingly high, and the tricky issues that would arise in implementing such a monitoring system.

Dr Harold Shipman arrives at Ashton-under-Lyne police station (photograph copyright Chris Glanvill, REX syndication)



Variability

- Could statistics have detected Harold Shipman's offending earlier?
- A patient dying is not unusual
 - ▶ Older patients tend to be more likely to die
- The number of patient deaths varies
- The timing of patient deaths varies
- The death rate varies by age, sex, ...
- Expect variation in the number of death certificates signed by different doctors
 - ▶ Some would sign more, some less

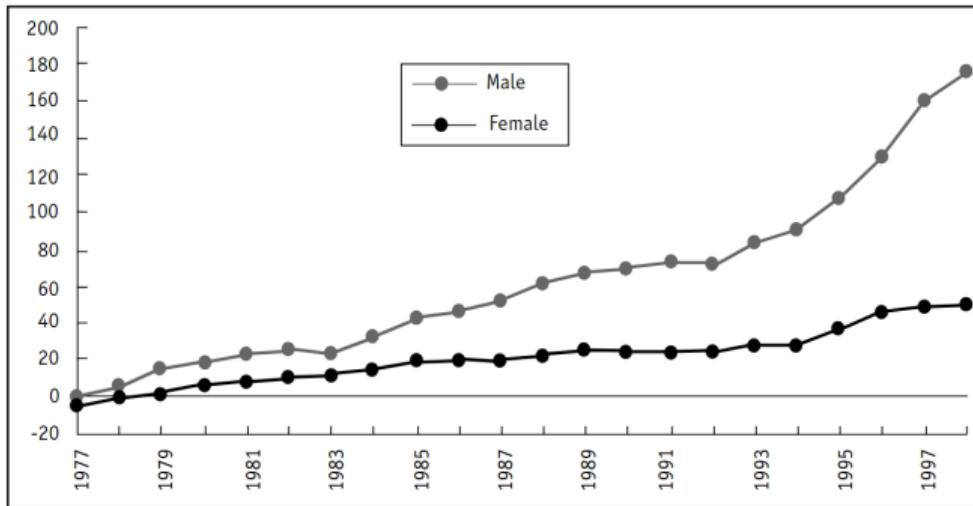
Variability

- A doctor signs one death certificate in one afternoon
 - ▶ Not unusual
- A doctor signs a million death certificates in one afternoon
 - ▶ Terrifying
- Idea: there is some range of values that are 'expected' or 'normal'
- Calculate excess deaths compared to an average doctor
 - ▶ Based on probability
 - ▶ Account for factors like patient age

Excess deaths: Harold Shipman

- Excess deaths in 1998: 175 women and 49 men
 - ▶ Close to number of confirmed kills

Figure 2. Cumulative excess death certificates signed by Shipman, for people older than 64 and who died at home or in his practice



Role of statistics

- The authors conclude it may be feasible to monitor doctors using statistics
- Highlight potential problems
 - ▶ Data availability
 - ▶ Privacy concerns
 - ▶ False positives: unusually high numbers of deaths
 - By chance
 - Case mix (e.g. predominantly work in rest homes)
 - Data quality

Example 1b: Lucy Letby

- Lucy Letby is a convicted UK serial killer
 - ▶ Neonatal nurse convicted of killing seven babies (convicted August 2023)
 - ▶ Prosecution case relied heavily on statistical evidence
- There have been concerns raised about the statistical evidence
 - ▶ Jury shown a chart listing 25 deaths and collapses
 - Lucy Letby was on shift for all of them
 - Other nurses were only on shift for a few of them
 - ▶ Another six deaths in the period were omitted from the table

Statistics and crime: another perspective?

- In September 2022 the Royal Statistical Society (RSS) published a report
 - ▶ Healthcare Serial Killer or Coincidence?
 - ▶ Prompted by concerns with cases in Italy and Netherlands due to association between shift patterns and deaths
- With regard to Letby²:
 - ▶ John O'Quigley (UCL London): “... *all the shift chart shows is that when Letby was on duty, Letby was on duty.*”
 - ▶ Richard Gill (Leiden University): “*The police investigation and crown prosecution made all the mistakes the RSS warned about. Nobody studied the statistics in a professional way.*”

²Both quotes taken from Guardian article linked on previous page.

Examples

- In both situations:
 - ▶ There is variability and uncertainty
 - ▶ What is the likely ‘range’ of variability
- Goal in STAT 110: describe the variation and uncertainty mathematically
 - ▶ Statistical model

Roadmap

- Spend the rest of the week with data
 - ▶ Introduction to the software we will use (R / Rstudio)
- Exploring and visualizing data will help motivate
 - ▶ Probability
 - ▶ Statistical models

Summary

- Statistics is learning from data
- Statistics is about describing and quantifying variability

Data

- Data is all around us
 - ▶ But how do we interact with it?
- In the past: pen and paper
- More recently: computers
 - ▶ Software for data and statistics
- Today: start interacting with data

Statistical software

- There are many statistical software packages
 - ▶ R
 - ▶ SAS
 - ▶ Stata
 - ▶ SPSS
 - ▶ JMP
 - ▶ PRISM
- Other software packages are also used
 - ▶ Excel
 - ▶ Python
 - ▶ Julia
 - ▶ ...

R (and excel)

- We are going to focus on one of these: R
 - ▶ R has a learning curve
 - Provide support in lectures, tutorials and assignments
- We will also see excel
 - ▶ Excel is used by many researchers to record data
 - ▶ It is also used by many researchers to analyze data
 - ▶ Excel has many weaknesses for data handling and statistics
 - Data handling: easy to (unintentionally) change/corrupt data
 - Statistical modelling: has basic functionality
 - ▶ Learn how to import data into R

R: NZ on the world stage

- R was developed at the University of Auckland in the early 90s
 - ▶ Ross Ihaka (Ngati Kahungunu, Rangitane)
 - ▶ Robert Gentleman
- It is used around the world
- Advantages:
 - ▶ Freely available
 - ▶ External packages that extend base functionality ^a
 - Contributed by researchers around the world
 - New methodology often readily implemented in R



^aWe may see how to install and use packages later

R: Installation

- We will be using Rstudio
 - ▶ R is the language (command line)
 - ▶ Rstudio is an IDE (integrated development environment) for R
 - Provides a more user-friendly experience
- We need to download and install both R and Rstudio
 - ▶ See video on blackboard information for installation instructions
 - Installing on chromebook or tablet is difficult or impossible
 - See video on blackboard for possible workarounds
 - ▶ Tutorials on Thursday that provides support for installing R and Rstudio

R: hands on

- Move into Rstudio
- Look at some data
- We will mostly see data in csv files
 - ▶ Comma separated file
 - ▶ Tabular (or rectangular) data
 - ▶ Opened by spreadsheet (like excel), but is plain text
 - ▶ See video on blackboard for how to obtain a csv from excel
 - ▶ It is possible to import data directly from excel
 - It requires installing and loading an additional package
 - Not considered further in STAT 110

Rstudio: hands on I

- Four panes³:
 1. LL: Console pane (where R code is run)
 - Start with this today: get things working initially
 2. UL: Editor pane (where we work)
 - Circle back around to how to use editor
 - This is our primary ‘work environment’
 3. UR: Environment (etc) pane (what have we done)
 4. LR: Files (etc) pane (help, plots, packages)

³The hands on lecture slides are a reminder for me of what to show you in Rstudio

Rstudio: hands on II

- Get (import) data
- Option one: use the drop down menu
 - ▶ File > Import Dataset > From text (base)
- Import penguin dataset
 - ▶ Available on blackboard
 - ▶ peng_lect1.csv
 - ▶ Various options
- Data should automatically be viewed
 - ▶ If closed, view again by clicking on object in 'Environment' tab

Data

- The data are a subset from a larger dataset of penguins from the Palmer Archipelago⁴
 - ▶ Group of islands off the northwestern coast off Antarctica
- Measurements of flipper length and bill length for a sample of gentoo penguins⁵
- We will interact with this (and the larger dataset) a few times this semester



⁴Data collected by Dr. Kristen Gorman with Palmer Station LTER.

⁵Photo: Andrew Shiva / Wikipedia

Rstudio: we're stuck

- We can:
 - ▶ Order the values
 - ▶ Look at a ‘spreadsheet’
- To do anything more we have to engage with editor
 - ▶ Command line
 - ▶ Typing commands to R

Rstudio: another look at workflow

- If we exit out of Rstudio
 - ▶ Lose most of what we have done
 - ▶ Start again
 - ▶ Frustrating: assignments and bigger projects
- Solution is to work in the editor
 - ▶ It can be intimidating at first
 - ▶ Rstudio itself helps out
 - ‘History’

Rstudio: hands on (getting started with editor)

- Instructions for importing data onto editor
 - ▶ ‘History’ tab shows the R commands for what we have done
 - Put this in the editor window (for when we come back next time)
 - ▶ Care is needed with file structures
 - I suggest creating a STAT 110 folder
 - Use this as a ‘working directory’

Rstudio: hands on (where are we?)

- The working directory is the folder (on your computer) that R uses
- Change the working directory:
 - ▶ Session > Set Working Directory > Choose Directory
 - ▶ Equivalent command line expression
- Many of the mistakes we see with 100-level students
 - ▶ Asking R to find a file, but you're in the wrong folder
- First ensure in the correct folder
 - ▶ Then import the data

Rstudio hands on (bill length)

- The data has information about two variables
 - ▶ Flipper length
 - ▶ Bill length
- What if we only want to look at one (bill length)?
 - ▶ Use \$: allows us to access specific variables by name
 - ▶ Use [,1] : allows us to access columns of the data frame by number

```
peng_lect1$bill_length_mm
```

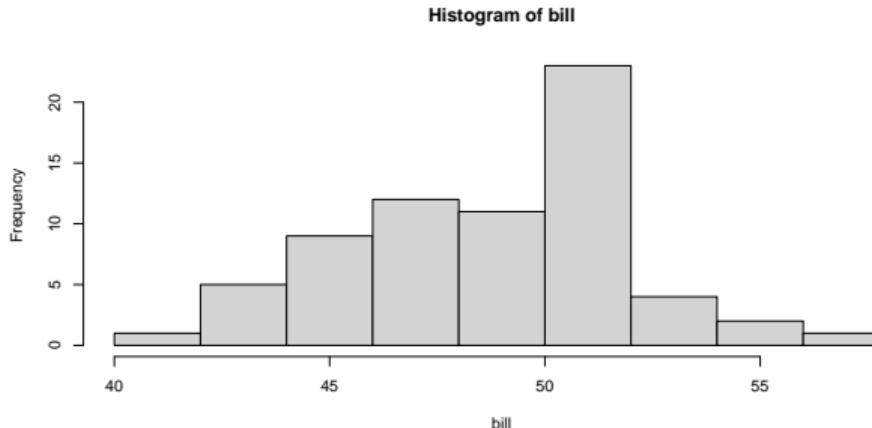
- Assign the new variable to bill
 - ▶ Use = or <-
 - ▶ Use these values later (next slide!)

```
bill = peng_lect1$bill_length_mm
```

Rstudio hands on (bill length)

- We can now look at numeric summaries of bill length, e.g.
 - ▶ mean: `mean(bill)`
 - ▶ median: `median(bill)`
 - ▶ standard deviation: `sd(bill)`
- We can also look at graphical summaries of bill length, e.g. histogram

```
hist(bill)
```



Rstudio: help!

- How would we know that in R:
 - ▶ `mean`: calculate the mean
 - ▶ `hist`: plot a histogram?
- There is internal help: probably not the first place to look
- For you in STAT 110:
 - ▶ Lecture slides
 - ▶ Assignments
 - ▶ Tutorials
 - ▶ Google: e.g. 'Finding an average in R'
 - ▶ AI (e.g. chatgpt)⁶

⁶A word of caution: AI tools are excellent for helping you get started with R. AI tools are not a replacement for thinking, but can be helpful tools for learning.

R code

- The norm for us interacting in Rstudio will not be 'hands on'
- Most of the time R code will be displayed on lecture slides

```
mean(bill)
```

```
## [1] 49
```

- These commands can be copied and pasted
 - ▶ Focus on understanding what the R code is doing
 - ▶ Support for Rstudio in tutorials

Summary

- We will be using R/Rstudio in STAT 110
- Free, powerful, and widely used
- We saw how:
 - ▶ Change our working directory
 - ▶ Import data
 - ▶ Subset one variable (bill length)
 - ▶ Summarize that variable
 - Numerically
 - Graphically

Outline

- Long-term goal: fit, and interpret statistical models to real data
- We need some more background information first:
 - ▶ What is a statistical model?
 - ▶ Introduction to probability and random variables
- Today: look at data summaries
 - ▶ You may have seen these summaries before
 - ▶ Calculate these in R
 - ▶ Introduce 'mathematical notation'
 - ▶ Look at how these summaries point toward statistical modelling
 - Data summaries are the starting point, not the finish line
 - Motivate a better understanding of probability

Data: Palmer penguin data

- What do the data say about flipper length of gentoo penguins?
- Option 1: provide (list) the data
 - ▶ Not practical: $n = 68$ observations ⁷
 - ▶ It might not be possible
 - Privacy concerns
 - Other considerations (ethical or otherwise) which prevent sharing of data
- Option 2: visualize the data
 - ▶ Good idea, but hard to summarize
- Option 3: numerically summarize the data
- Option 4: approaches we are yet to learn

⁷if we considered all 3 penguins species, there are over 300 observations

Intro R

- Step 1: call data into R

- ▶ Import using menu (File Import Dataset)
 - ▶ Use commands

```
peng_lect1 = read.csv('peng_lect1.csv')
```

- ▶ peng_lect1.csv needs to be in the current working directory in Rstudio

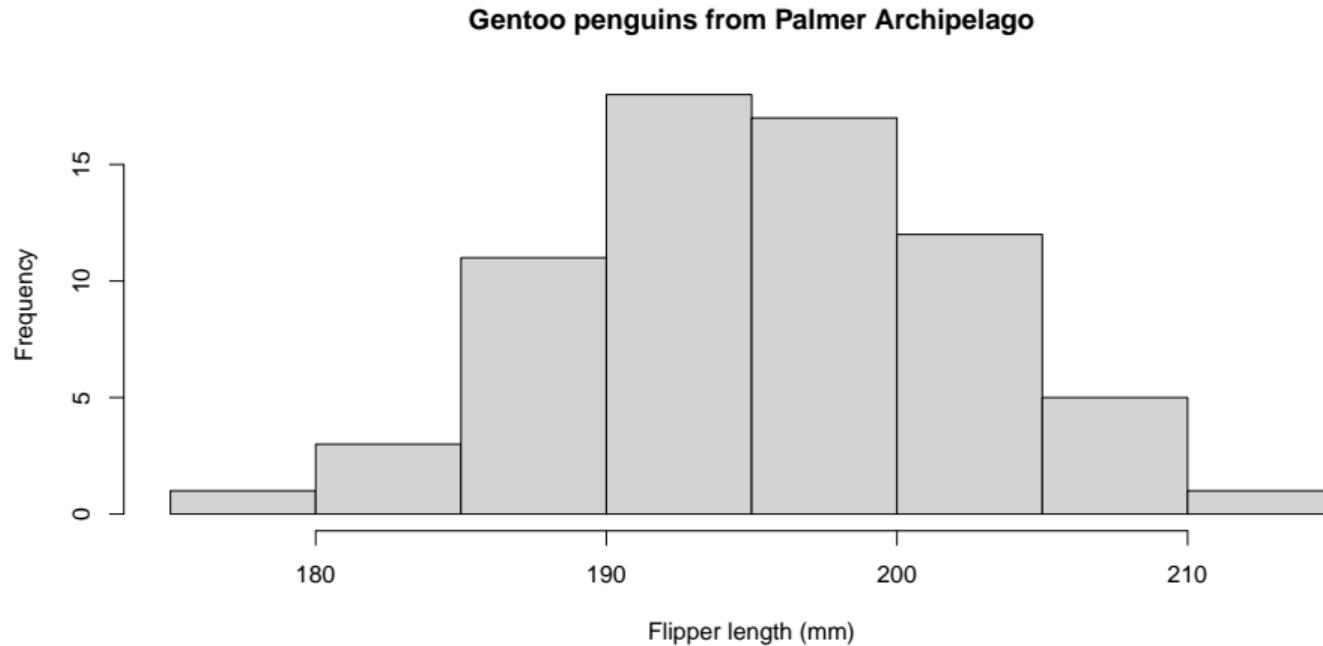
- Step 2: visualize the data

```
hist(peng_lect1$flipper_length_mm, xlab = "Flipper length (mm)",  
      main = "Gentoo penguins from Palmer Archipelago")
```

- Remember: peng_lect1 has two variables

- ▶ peng_lect1\$flipper_length_mm obtains the flipper length variable

Histogram



Summary 1: sample mean

- The mean is a common summary
 - ▶ Often called the average
- The sample mean is the sum of the observed values divided by the number of observations

$$\bar{y} = \frac{y_1 + y_2 + \dots + y_n}{n}$$

- Let's unpack:
 - ▶ What does \bar{y} represent?⁸
 - ▶ What does y_1 represent?
 - ▶ What does y_2 represent?
 - ▶ What does n represent?

⁸ \bar{y} is said: y-bar

Summary 1: sample mean

- The sample mean is given as

$$\bar{y} = \frac{y_1 + y_2 + \dots + y_n}{n}$$

- Commonly we will see this written as

$$\bar{y} = \frac{\sum_{i=1}^n y_i}{n}$$

- Let's unpack:
 - What does y_i represent?
 - What does $\sum_{i=1}^n$ represent?
- The two equations say exactly the same thing

Tutorial: what the \sum ?

- The sample mean is

$$\bar{y} = \frac{y_1 + y_2 + \dots + y_n}{n} = \frac{\sum_{i=1}^n y_i}{n}$$

- \sum is the Greek letter Sigma (capital)
 - ▶ It represents a sum
 - ▶ $\sum_{i=1}^n y_i$ says that we:
 - Set $i = 1$ and find y_i : gives y_1
 - Set $i = 2$ and add y_i : gives $y_1 + y_2$
 - Set $i = 3$ and add y_i : gives $y_1 + y_2 + y_3$
 - Keep going...

Finding the mean

- It is worth knowing how to find a mean ‘the old fashioned way’
 - ▶ What is the mean of 10, 6, 13, 7?
 - ▶ It means you can (in principle) calculate a mean anywhere, anytime
 - In your head (if not exactly, then approximately)
 - On a calculator / phone

Finding the mean

- The majority of the time we use the computer (R or other software)

```
y = c(10, 6, 13, 7) # c() is used to create a vector (or collection) of values  
y  
## [1] 10 6 13 7
```

- Use the R function `mean()` to find the mean

```
mean(y)  
## [1] 9
```

- For the flipper data

```
mean(peng_lect1$flipper_length_mm)  
## [1] 196
```

R: excursion

- You may have noticed that sometimes I have created an R object

```
y = c(10, 6, 13, 7) # c() is used to create a vector (or collection) of values
```

- This has created the object y
 - ▶ This object is then available to ‘use’, e.g. when finding the mean

```
mean(y)
```

```
## [1] 9
```

- In the code above, the mean value is not assigned to an object
 - ▶ It can be – it is then available to ‘use’ later on
 - e.g. if we were to compare it to the mean flipper length for chinstrap penguins

```
ybar = mean(y)
```

```
ybar
```

```
## [1] 9
```

Other summaries

- The (sample) mean tells us a lot
 - ▶ Among our sample of 64 gentoo penguins the average flipper length was 196 mm
 - ▶ A penguin with a flipper length of 205 mm is bigger than average
- There is a lot the mean does not tell us
 - ▶ Is it surprising if we saw a gentoo penguin with a flipper length of 170 mm?
- Another summary that tells us how variable the data are would be useful?
 - ▶ High variability: we commonly see flipper less than 120 mm or more than 270 mm
 - ▶ Low variability: unlikely to see flipper less than 190 mm or more than 200 mm

Summary 2: sample variance and standard deviation

- We will focus on two measure of variation
 - ▶ Variance
 - ▶ Standard deviation
- These are different expressions of the same thing
 - ▶ The variance is $(\text{standard deviation})^2$
 - ▶ The standard deviation is $\sqrt{\text{variance}}$

Summary 2: sample variance

- Sample variance: average squared distance between observations and the mean

$$s^2 = \frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n - 1}$$

- ▶ We divide by $n - 1$ (and not n)
 - There is some mathematical nuance
 - For our purposes: it gives a more reliable answer
- ▶ It is a difficult calculation to do by hand
 - It is worth doing for a small problem to ensure you understand the formula
 - What is the variance of 10, 6, 13, 7?⁹
- We can find it easily in R

```
var(peng_lect1$flipper_length_mm)  
## [1] 51
```

⁹The answer is 10

Summary 2: sample variance

- Sample variance: average squared distance between observations and the mean

$$s^2 = \frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n - 1}$$

- ▶ If an observation y_i is far from \bar{y}
 - $(y_i - \bar{y})^2$ will be large
- ▶ If the observations y_1, \dots, y_n are spread out
 - Many of the values $(y_i - \bar{y})^2$ will be large
 - s^2 will be large
- ▶ If an observation y_i is close to \bar{y}
 - $(y_i - \bar{y})^2$ will be small
- ▶ If the observations y_1, \dots, y_n are close together
 - Most of the values $(y_i - \bar{y})^2$ will be small
 - s^2 will be small

Summary 2: sample standard deviation

- The sample standard deviation

$$s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n - 1}}$$

- It represents the typical deviation of observations from the mean (approximately)
 - ▶ Useful when considering how far the data are distributed from the mean
 - ▶ Easier to interpret than the variance
- We can find it easily in R

```
sd(peng_lect1$flipper_length_mm)
## [1] 7.1
```

- A typical observation is *approximately* 7.1 mm from the sample mean

Standard deviation: rules of thumb

- To better help us understand what the standard deviation represents
 - ▶ Approximately 70% of the data will be within one standard deviation of the mean
 - ▶ Approximately 95% of the data will be within two standard deviations of the mean
- These are only rules of thumb.
 - ▶ e.g. they do not hold if the data are skewed

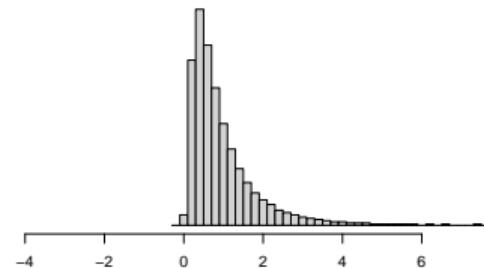
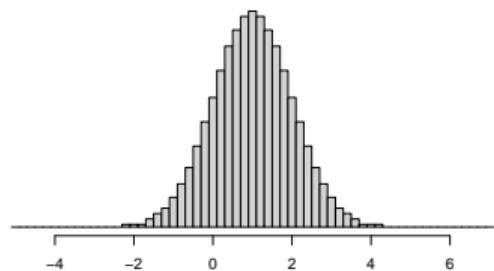
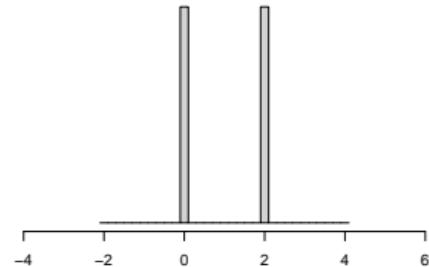
Data summaries: big picture

- On one hand: lost a lot of information
 - ▶ $n = 68$ into two numbers
- On the other hand: created order out of chaos
 - ▶ It is hard for us to get an understanding of $n = 68$ values ¹⁰
 - ▶ Summarized the data to gain an understanding about important features of the data
 - ▶ Suppose we got $n = 205$ observations from penguins in a different location
 - Hard to compare 68 observations to 205 observations by eye
 - Relatively easy to compare the mean from each group
 - ▶ The idea of finding a “simple” description (or model) of complex data will be a theme
- Look into the limitations of data summaries

¹⁰It is even worse if we have $n = 6800$ values!

Limitations of data summaries I

- Data summaries are useful, but...
 - ▶ Lose a lot of information: $n = 68$ into two numbers
 - ▶ Be careful not to over-interpret
- Three histograms: data with the same sample mean ($\bar{y} = 1$) and variance ($s^2 = 1$)



Limitations of data summaries II

- Data summaries are useful, but...
 - ▶ Samples do not give perfect information about the population
 - ▶ If we took a different sample, get a different sample mean (and variance)
- The population is all gentoo penguins in the Palmer archipelago
- The mean flipper length of the population is unlikely to be 196 mm
 - ▶ The value of 196 mm can be thought of as an educated guess (or estimate)
 - ▶ Can we quantify how precise (or uncertain) that estimate is?
- We cannot get this information from data summaries alone
 - ▶ What we will be working toward
 - ▶ Use probability to describe the variation in the data
 - ▶ Statistical models

Limitations of data summaries IIb

- Samples do not give perfect information about the population
- Again: suppose we also have $n = 205$ observations from a different location
- Above: claimed it was relatively easy to compare the mean from each group
 - ▶ Not that simple
 - We can easily compare the sample means
 - We care about the comparison between the population means
- Does flipper length vary by location if sample means are 196 mm vs 302 mm?
- Does flipper length vary by location if sample means are 196 mm vs 197 mm?
- Can we quantify the precision (or uncertainty) of the estimates?
 - ▶ Use probability to describe the variation in the data
 - ▶ Statistical models

Summary

- Calculate basic data summaries in R
- Understand how to calculate data summaries by hand (if we need to)
- Introduce mathematical notation
- Looked at limitations of data summaries

STAT 110: Week 2

University of Otago

Outline

- We've started interacting with data
- Data summaries: sample mean and standard deviation
- Summaries are limited
 - ▶ To go further we need statistical models
 - Use probability to describe the variation in the data
- Over the next few lectures we will look at probability
 - ▶ Start today with foundational knowledge
 - ▶ Much of this knowledge remains important even in complex applications

Probability

- Immediate problem:
 - ▶ We want to describe the data using probability
 - ▶ We need to understand probability

Probability: mathematical language of uncertain events

- What is the probability that:
 - ▶ A randomly sampled penguin in the Palmer archipelago is an Adélie?
 - ▶ The all black kicker is successful with their next kick?
 - ▶ A rat will choose one reward (out of many) when moving through a maze?
 - ▶ A person has a certain genotype?
 - ▶ A female skink is a breeder?
 - ▶ An earthquake of magnitude 5 or larger occurs this year?
 - ▶ A cancer patient will die within 12 months?
 - ▶ The sliced ham you got at the supermarket is safe to consume?

Probability

- Setup
 - ▶ Random process with a number of possible outcomes
 - Roll a die. Possible outcomes: 1, 2, 3, 4, 5, or 6
 - Flip a coin. Possible outcomes: head or tail
 - Observe a penguin at Palmer. Possible outcomes: Adelie, chinstrap, gentoo
 - The set of all possible outcomes is called the sample space
- A probability has to satisfy a number of mathematical principles, including:
 - ▶ Between 0 and 1
 - We can't have a probability of -0.4 or 1.2
 - ▶ Probabilities sum to 1
 - If we observe the random process, we must see one of the possible outcomes
 - If we flip a coin, we must see either a head, or a tail.

Probability

- From here, things get a little murky
 - ▶ There are several definitions (or interpretations) of probability¹
- We will define probability in terms of relative frequency:
 - ▶ The probability of an outcome is the proportion of times the outcome occurs if we were to observe the random process a large (infinite) number of times.
 - Imagine a (bored!) person repeatedly tossing a coin²

¹We may return to this later in the semester.

²John Edmund Kerrich.

Mutually exclusive outcomes

- Two outcomes are mutually exclusive (or disjoint) if they cannot both happen
 - ▶ e.g. a coin flip cannot land on heads and tails
 - ▶ e.g. A penguin we observe cannot be a chinstrap and gentoo
- The probability of mutually exclusive outcomes can be found with addition
 - ▶ For two outcomes A and B that are mutually exclusive:

$$\Pr(A \text{ or } B) = \Pr(A) + \Pr(B)$$

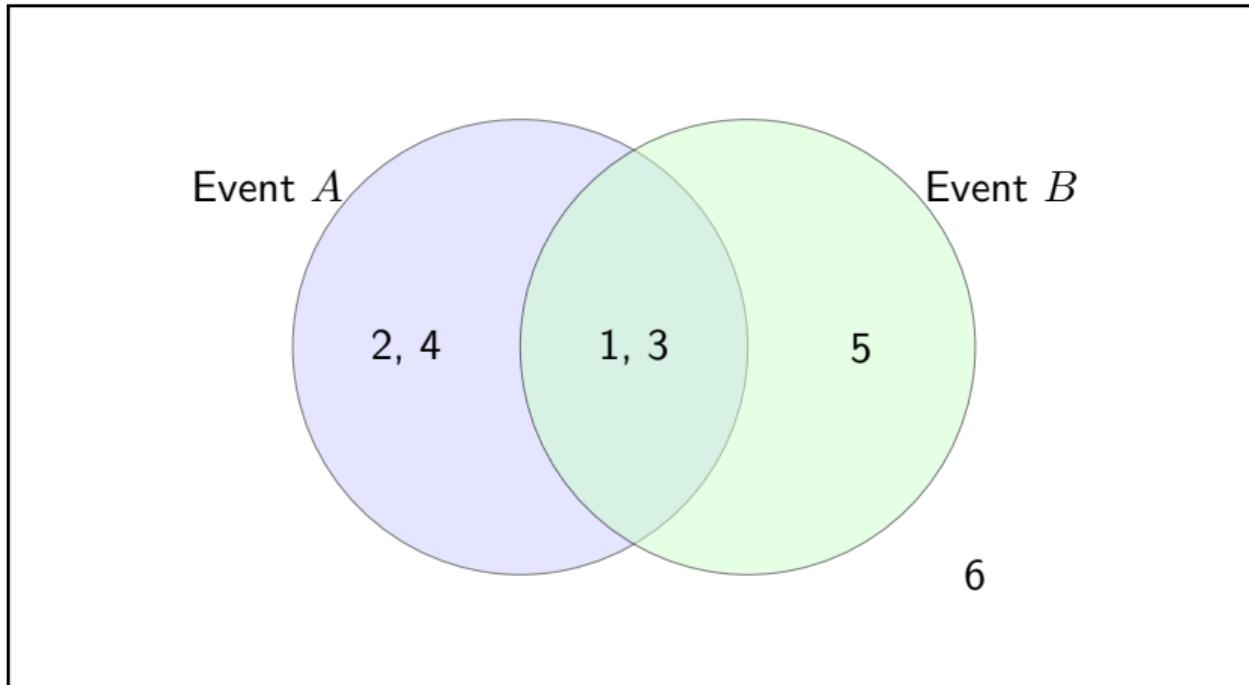
- Die roll: A : roll a 1, B : roll a 6
 - ▶ $\Pr(A \text{ or } B) = \Pr(A) + \Pr(B) = 1/6 + 1/6 = 1/3$
- Penguins: A : observe a chinstrap, B : observe a gentoo
 - ▶ $\Pr(A \text{ or } B) = \Pr(A) + \Pr(B)$

Events

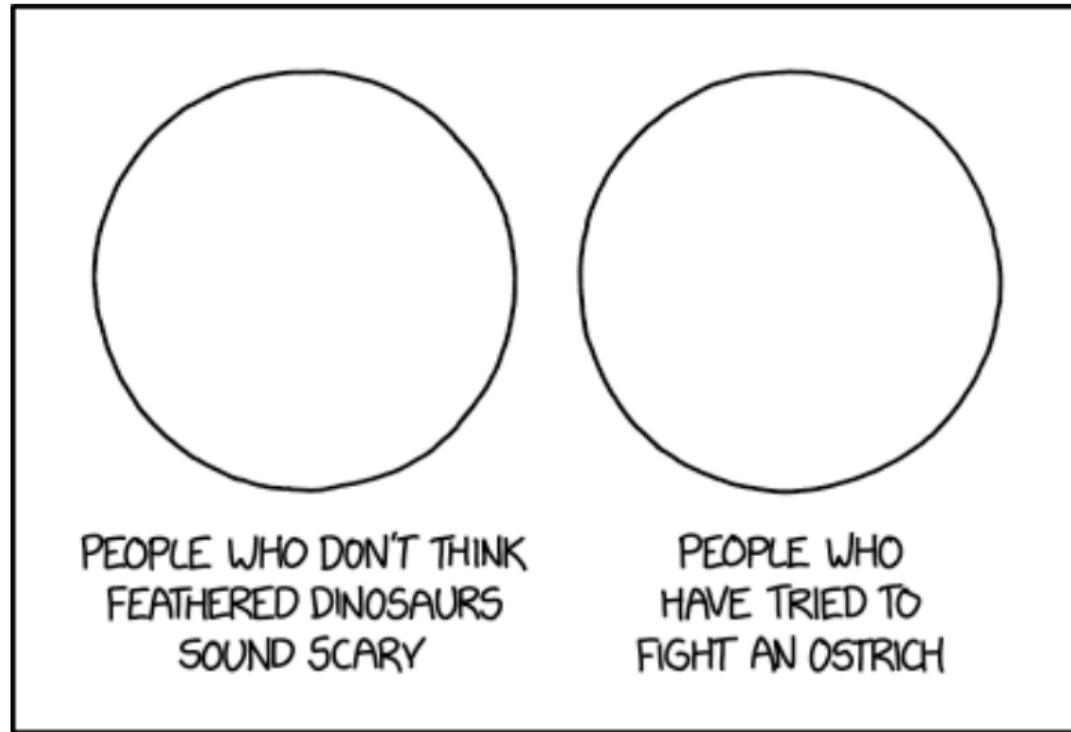
- We often work with collections of outcomes
 - ▶ These are called events
- Examples:
 - ▶ Die roll: event A : roll 1, 2, or 4, event B : roll 5 or 6.
 - ▶ Penguins: event C : Adelie or chinstrap, event D : gentoo or chinstrap
- Events can be mutually exclusive if they have no outcomes in common
 - ▶ Events A and B are mutually exclusive
 - $\Pr(A \text{ or } B) = \Pr(A) + \Pr(B) = 3/6 + 2/6 = 5/6$
 - ▶ Events C and D are not mutually exclusive
 - What is $\Pr(C \text{ or } D)$?
- An event can comprise a single outcome
 - ▶ e.g. the event E : roll a 3

Venn diagram

- Venn diagrams can be used to visualize small sample spaces
- Die roll: event A : roll 4 or less, event B : roll odd number



Venn diagram

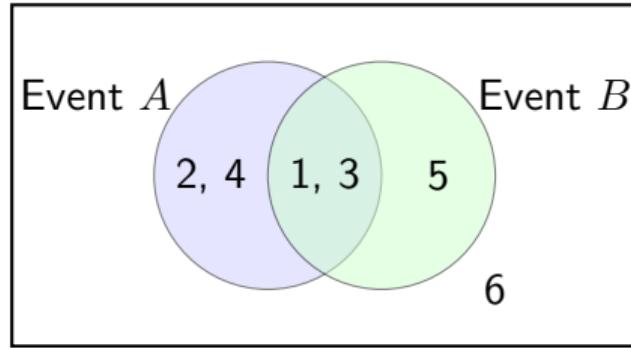


²<https://xkcd.com/2090/>

Venn diagram and sets

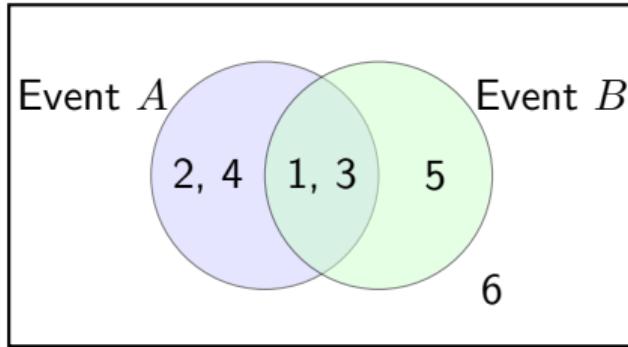
- Venn diagrams are useful when looking at when:
 - ▶ Event A or B occurs
 - This is inclusive, i.e. A or B means that event A , B or both A and B occur.
 - ▶ Event A and event B occurs

Sets of interest



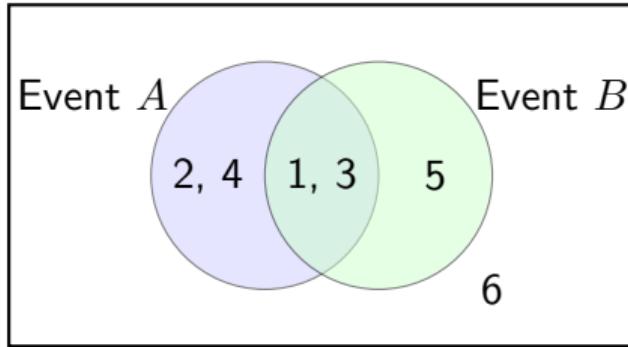
- $A \text{ or } B$: 1, 2, 3, 4, 5
- $A \text{ and } B$: 1, 3

Probabilities



- What is $\Pr(A \text{ or } B)$?
 - ▶ Same question asked a few slides ago (the events were called C and D then)
 - ▶ Events A and B are not mutually exclusive
 - ▶ $\Pr(A) + \Pr(B) = 4/6 + 3/6 = 7/6$
 - Clearly incorrect

Probabilities



- What is $\Pr(A \text{ or } B)$?
 - ▶ Probability of observing a 1, 2, 3, 4, or 5: probability of $5/6$
- Problem with $\Pr(A) + \Pr(B)$ is that it double counts outcomes 1 and 3
 - ▶ Double counting $\Pr(A \text{ and } B)$

General addition rule

- If A and B are any two events, then the probability that at least one of them occurs is

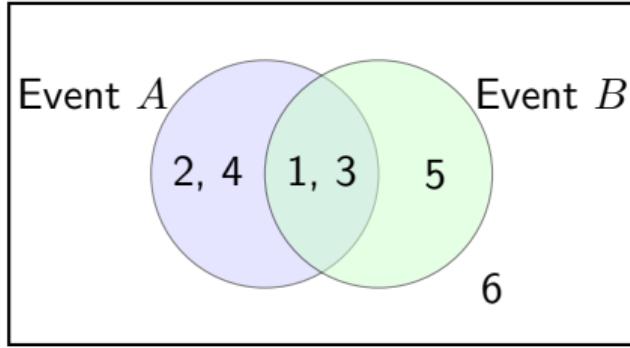
$$\Pr(A \text{ or } B) = \Pr(A) + \Pr(B) - \Pr(A \text{ and } B)$$

- If events A and B are mutually exclusive, then $\Pr(A \text{ and } B) = 0$.
- Example (from previous slide)
 - ▶ $\Pr A + \Pr(B) - \Pr(A \text{ and } B) = 4/6 + 3/6 - 2/6 = 5/6$

Complement

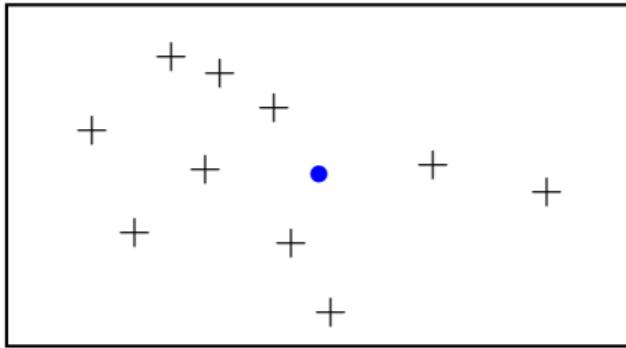
- Compliment: STAT 110 students are amazing!
- Complement of event A : the outcomes in the sample space that are not in A
- Roll a die: sample space is $\{1, 2, 3, 4, 5, 6\}$
 - ▶ The event E is rolling even: $\{2, 4, 6\}$
 - ▶ Its complement E^C is $\{1, 3, 5\}$
- $\Pr(E) + \Pr(E^C) = 1$, or $\Pr(E) = 1 - \Pr(E^C)$
 - ▶ For the example above: $\Pr(E) = 0.5$, $\Pr(E^C) = 0.5$
- Complements seem obvious and simple
 - ▶ I frequently remind 400-level students how useful they can be

Complements



- Complements ‘play nice’ with Venn diagrams
- What is:
 - ▶ $\Pr(A^c)$?
 - ▶ $\Pr(B^c)$?
 - ▶ $\Pr((A \text{ or } B)^c)$?
 - ▶ $\Pr((A \text{ and } B)^c)$?

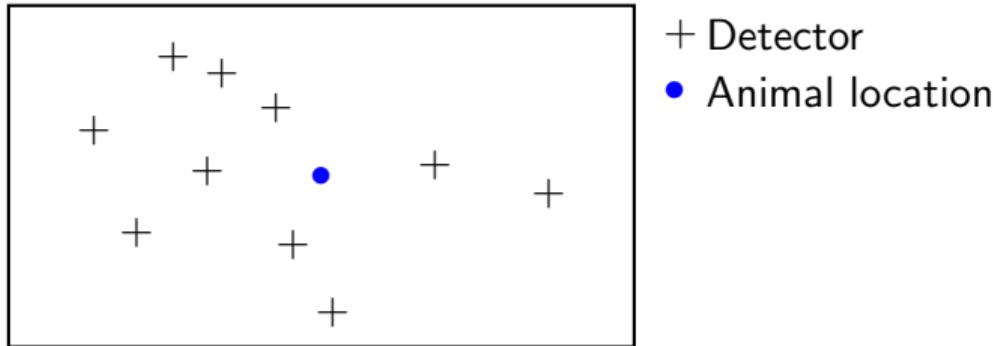
Complement: real example



- + Detector
- Animal location

- The picture represents an array of 'detectors' (e.g. motion activated camera)
 - ▶ Assume we know the probability the animal is detected (in some period of time) for each of the 10 detectors, based on its location
 - p_1, p_2, \dots, p_{10}
- What is the probability it is seen by at least one detector?

Complement: real example



- There are over 1000 possible ways an animal could be detected:
 - ▶ Seen at one detector: detector 1, detector 2, detector 3, ...
 - ▶ Seen at two detectors: detector 1 & 2, detector 1 & 3
 - ▶ etc
- There is only one way an animal cannot be seen
 - ▶ Complement of being seen by at least one detector

Summary

- We are working toward statistical model for data
 - ▶ Use probability to describe the variation in the data
- Foundational knowledge in probability
 - ▶ Outcomes and events
 - ▶ Sample space, sets, and complements
 - ▶ General addition rule
- Relate probability back to examples
- Tomorrow: everyone bring a coin
 - ▶ Explore some (interactive) probability results

Outline

- Continue to build our knowledge of probability
- Today we look at two (or more) random processes
 - ▶ Independence
 - ▶ Conditional probability
 - ▶ Contingency tables
- Begin with an interactive exercise

Probability is hard

- It can be easy to trick ourselves that probability is easy
 - ▶ What is the probability that a die lands on a 4?

Probability is hard

- It can be easy to trick ourselves that probability is easy
 - ▶ What is the probability that a die lands on a 4?
 - ▶ It goes from easy to difficult very quickly
- In each of the next two lectures:
 - ▶ Start with an exercise that *might* be surprising to you
 - ▶ Hopefully broaden understanding of probability
- If I told you that I was flipping coins and saw 7 heads in a row
 - ▶ Would you think I am telling the truth?
 - ▶ How likely is it that I flip a fair coin and get 7 heads in a row?
 - Without getting calculators (or phones!) out
 - Is it closest to: 1 in one million? 1 in ten thousand? 1 in 100?

Exercise

- Everyone stand up
- Flip your coin (when I tell you to)³
 - ▶ Head: remain standing
 - ▶ Tail: sit down
- Those who are still standing, flip again (when I tell you to)
- Repeat until everyone sits down
 - ▶ See how many flips the 'best' person can get
- Is that what you were expecting?
- How would this look if we repeated the experiment at Beaver Stadium?
 - ▶ Penn State football stadium: capacity of over 106 000

³If you have forgotten a coin, google 'flip a coin' and you can do it online

Examples

- We might want to know the probability that:
 - ▶ A try was scored under the posts given the conversion was successful?
 - ▶ A participant undertakes a task left handed given they have a certain genotype?
 - ▶ A person has a certain genotype given they have brown eyes?
 - ▶ A female skink observed without offspring is a breeder?
 - ▶ An aftershock of magnitude 5 or larger occurs within six hours of a earthquake of magnitude 6?
 - ▶ Chicken is safe to consume given it has been heated to 60°C for five minutes?
- Each of these probabilities depends on another variable

Independence

- There are situations where we would expect two random processes to be unrelated
 - ▶ Process 1: rolling a die, process 2: flipping a coin
 - ▶ Process 1: eye colour of a person, process 2: success of rugby kick (conversion)
- We refer to these as independent
 - ▶ Two events A and B are independent if the outcome of one event provides no information about the outcome of the other
 - Knowing that our coin flip landed on heads does not change the probability of rolling a six
- Other processes may not be independent
 - ▶ Process 1: stock price of asset A, process 2: stock price of asset B

Independence

- Consider again the two processes:
 - ▶ Process 1: rolling a die, process 2: flipping a coin
- If A is the event 'roll four or lower', and B is the event 'coin lands head'
 - ▶ What is $\Pr(A \text{ and } B)$?

Independence

- Consider again the two processes:
 - ▶ Process 1: rolling a die, process 2: flipping a coin
- If A is the event 'roll four or lower', and B is the event 'coin lands head'
 - ▶ What is $\Pr(A \text{ and } B)$?
- $\Pr(A) = 4/6$ and $\Pr(B) = 1/2$
- Since the two events are independent we can reason that:
 - ▶ Event A will occur $4/6$ of the time
 - ▶ Event B will subsequently occur $1/2$ of those times
 - ▶ Event A and B occur together $4/6 \times 1/2 = 2/6$ of the time

Multiplication rule: independent processes

- If A and B are independent events, then⁴

$$\Pr(A \text{ and } B) = \Pr(A) \Pr(B)$$

- Example: suppose that 10% of the population are left handed, and 50% are female. If handedness and sex are independent, then what is the probability that a randomly selected person is right-handed and female?

$$\begin{aligned}\Pr(\text{right-handed and female}) &= \Pr(\text{right-handed}) \Pr(\text{female}) \\ &= 0.9 \times 0.5 \\ &= 0.45\end{aligned}$$

⁴This can be extended to more than two events

Diversion: sex and gender

- Often we look at differences to do with sex or gender
- If interest is in biological differences (like the example above)
 - ▶ Sex: XX and XY
 - Recognize that intersex individuals exist
 - Usually not accounted for as prevalence is low
- In other applications (e.g. social science) we may be interested in gender
 - ▶ Often two genders still used
 - ▶ In time, increasingly see wider representation

Conditional probability

- Conditional probability describes the relationship between two events
- The probability of event B given event A has occurred is written $\Pr(B|A)$
 - ▶ Let T be the event that a try was scored under the posts
 - ▶ Let C be the event that the conversion was successful
 - $\Pr(T|C)$ is the probability that a try was scored under the posts given the conversion was successful?
 - ▶ Let B be the event that a female skink is a breeder
 - ▶ Let O be the event that a female skink is observed without offspring
 - $\Pr(B|O)$ is the probability that a female skink observed without offspring is a breeder

Conditional probability

- We can find $\Pr(B|A)$ using

$$\Pr(B|A) = \frac{\Pr(A \text{ and } B)}{\Pr(A)}$$

- $\Pr(A \text{ and } B)$: joint probability of events A and B occurring
- $\Pr(A)$: marginal probability of event A
- Two events A and B are independent if
 - ▶ $\Pr(B|A) = \Pr(B)$
 - ▶ The event A occurring does not change the probability of B occurring.
- Helpful to look at contingency tables

Contingency table: titanic

- Contingency tables allow us to compare two (categorical) variables⁵
- Data from the adult passengers on the titanic. Two variables:
 - ▶ Sex: male or female
 - ▶ Survived: yes or no
- Two tables: the first gives the counts, the second gives proportions

Sex	survived			Total
	yes	no		
male	338	1329	1667	
female	316	109	425	
Total	654	1438	2092	

Sex	survived			Total
	yes	no		
male	0.162	0.635	0.797	
female	0.151	0.052	0.203	
Total	0.313	0.687	1.000	

⁵They can be extended to more than two variables

Contingency table: titanic

- Two tables: the first gives the counts, the second gives proportions
 - ▶ For now, we will treat the proportions as if they are probabilities
 - ▶ See better approaches for estimating probabilities from contingency tables later

		survived		survived	
		yes	no	Total	
Sex	male	338	1329	1667	Sex
	female	316	109	425	
	Total	654	1438	2092	
		yes	no	Total	
	male	0.162	0.635	0.797	
	female	0.151	0.052	0.203	
	Total	0.313	0.687	1.000	

- Proportion are found by dividing entries by total, 2092
 - ▶ e.g. $316/2092 = 0.151$
 - ▶ e.g. $1438/2092 = 0.687$

Contingency table: titanic

- S : randomly selected passenger survived
- M : randomly selected passenger is male
- Marginal probability
 - ▶ $\Pr(M) = 0.797$
 - ▶ $\Pr(S) = 0.313$
 - ▶ Found in margin of contingency table

Sex	survived		Total
	yes	no	
male	338	1329	1667
female	316	109	425
Total	654	1438	2092

Sex	survived		Total
	yes	no	
male	0.162	0.635	0.797
female	0.151	0.052	0.203
Total	0.313	0.687	1.000

Contingency table: titanic

- S : randomly selected passenger survived
- M : randomly selected passenger is male
- Joint probabilities
 - ▶ $\Pr(M \text{ and } S) = 0.162$
 - ▶ $\Pr(M \text{ and } S^C) = 0.635$
 - ▶ $\Pr(M^C \text{ and } S) = 0.151$
 - ▶ $\Pr(M^C \text{ and } S^C) = 0.052$
 - ▶ Found in cells of contingency table

Sex	survived		Total
	yes	no	
male	338	1329	1667
female	316	109	425
Total	654	1438	2092

Sex	survived		Total
	yes	no	
male	0.162	0.635	0.797
female	0.151	0.052	0.203
Total	0.313	0.687	1.000

Contingency table: titanic

- S : randomly selected passenger survived
- M : randomly selected passenger is male
- Conditional probabilities
 - ▶ $\Pr(S|M) = \frac{\Pr(S \text{ and } M)}{\Pr(M)} = \frac{0.162}{0.797} = \frac{338}{1667} = 0.20$
 - Only consider male row
 - Survival and sex are not independent ([why?](#))
 - ▶ $\Pr(S|M^C) = \frac{0.151}{0.203} = \frac{316}{425} = 0.74$
 - ▶ Could also find $\Pr(M|S)$, ...

Sex	survived		Total
	yes	no	
	male	female	
male	338	1329	1667
female	316	109	425
Total	654	1438	2092

Sex	survived		Total
	yes	no	
	male	female	
male	0.162	0.635	0.797
female	0.151	0.052	0.203
Total	0.313	0.687	1.000

Contingency table: smallpox

- Data are 6224 observations from individuals in Boston in 1721 who were exposed to smallpox⁶. There are two variables:
 - ▶ Inoculated: yes or no⁷
 - ▶ Result: lived or died

		inoculated		Total
		yes	no	
result	lived	238	5136	5374
	died	6	844	850
	Total	244	5980	6224

		inoculated		Total
		yes	no	
result	lived	0.038	0.825	0.863
	died	0.001	0.136	0.137
	Total	0.039	0.961	1.000

⁶Fenner F. et al. 1988. Smallpox and Its Eradication (History of International Public Health, No.

6). Geneva: World Health Organization. ISBN 92-4-156110-6, p. 257

⁷Exposing a person to the disease in a controlled form

Contingency table: smallpox

- I : individual exposed to smallpox was inoculated
- L : individual exposed to smallpox lived
 - ▶ Find marginal probabilities:
 - $\Pr(I)$, $\Pr(L)$, $\Pr(I^C)$, $\Pr(L^C)$
 - ▶ Find joint probabilities:
 - $\Pr(I \text{ and } L)$, $\Pr(I^C \text{ and } L)$, $\Pr(I \text{ and } L^C)$,
 $\Pr(I^C \text{ and } L^C)$
 - ▶ Find conditional probabilities:
 - $\Pr(L|I)$, $\Pr(L|I^C)$
 - Are L and I independent?
 - ▶ Find $\Pr(I \text{ or } L)$
 - Is this a meaningful quantity in this example?

result	inoculated		Total
	yes	no	
	lived	died	
lived	238	5136	5374
died	6	844	850
Total	244	5980	6224

result	inoculated		Total
	yes	no	
	lived	died	
lived	0.038	0.825	0.863
died	0.001	0.136	0.137
Total	0.039	0.961	1.000

Summary

- Looked at two random processes
- Introduced:
 - ▶ Independence
 - ▶ Multiplication rule (for independent events)
 - ▶ Conditional probability
 - ▶ Contingency tables
- Still building our knowledge of probability
 - ▶ So that we can apply it for statistical modelling
- We will start the next lecture with another probability exercise

Outline

- Look more at (conditional) probability
- Start with another probability exercise

Challenge

- I will select a small group of you
 - ▶ Most likely a row or two
- Question: does anyone in this group shares a birthday?
 - ▶ Compare our predictions with reality

Recap

- In the last lecture we talked about:
 - ▶ Joint probability: $\Pr(A \text{ and } B)$
 - ▶ Marginal probability: $\Pr(A)$
 - ▶ Conditional probability: $\Pr(B|A) = \frac{\Pr(A \text{ and } B)}{\Pr(A)}$
- Pick up from where we left off

Multiplication rule: general

- Last lecture we saw how to find conditional probability

$$\Pr(B|A) = \frac{\Pr(A \text{ and } B)}{\Pr(A)}$$

- If we rearrange this, we get the general multiplication rule

$$\Pr(A \text{ and } B) = \Pr(A) \Pr(B|A)$$

- We can ‘switch’ A and B so that we also have

$$\Pr(A \text{ and } B) = \Pr(B) \Pr(A|B)$$

Multiplication rule: smallpox example

- Suppose we were told: “96.1% of the residents were not inoculated, and 85.9% of the residents who were not inoculated ended up surviving.”
 - ▶ What is the probability that a resident was not inoculated and lived?

Multiplication rule: smallpox example

- Suppose we were told: “96.1% of the residents were not inoculated, and 85.9% of the residents who were not inoculated ended up surviving.”
 - ▶ What is the probability that a resident was not inoculated and lived?
- $\Pr(I^C) = 0.961$: 96.1% of the residents were not inoculated
- $\Pr(L|I^C) = 0.859$: 85.9% of the residents who were not inoculated ended up surviving
- $\Pr(I^C \text{ and } L)$: probability that a resident was not inoculated and lived

$$\begin{aligned}\Pr(I^C \text{ and } L) &= \Pr(I^C) \Pr(L|I^C) \\ &= 0.961 \times 0.859 \\ &= 0.825\end{aligned}$$

Joint and conditional probability (smallpox example)

- Order doesn't matter for the joint probability
 - ▶ $\Pr(A \text{ and } B) = \Pr(B \text{ and } A)$
 - ▶ Probability both A and B occur
- Order does matter for the conditional probability
 - ▶ $\Pr(A|B)$ and $\Pr(B|A)$ are two different quantities
- Smallpox: compare $\Pr(L|I)$ and $\Pr(I|L)$.
 - ▶ $\Pr(L|I) = \frac{0.038}{0.039} = \frac{238}{244} = 0.975$
 - Probability of a resident living given inoculation
 - ▶ $\Pr(I|L) = \frac{0.038}{0.863} = \frac{238}{5374} = 0.044$
 - Probability of a resident being inoculated given lived

result	inoculated		Total
	yes	no	
	lived	5136	
died	6	844	850
Total	244	5980	6224

result	inoculated		Total
	yes	no	
	lived	0.825	
died	0.136	0.001	0.137
Total	0.961	0.039	1.000

Marginal probability: law of total probability

- Previously found “intuitively” from contingency table
- To find $\Pr(B)$
 - ▶ Sum over possible outcomes that could co-occur with the event B
- If there are two outcomes: A_1 and A_2
 - ▶ $\Pr(B) = \Pr(A_1 \text{ and } B) + \Pr(A_2 \text{ and } B)$
- Smallpox: find $\Pr(L)$ and $\Pr(I)$

result	inoculated		
	yes	no	Total
	lived	5136	5374
died	6	844	850
Total	244	5980	6224

result	inoculated		
	yes	no	Total
	lived	0.825	0.863
died	0.136	0.137	
Total	0.961	1.000	

Marginal probability: law of total probability

- Previously found “intuitively” from contingency table
- To find $\Pr(B)$
 - ▶ Sum over possible outcomes that could co-occur with the event B
- If there are two outcomes: A_1 and A_2
 - ▶ $\Pr(B) = \Pr(A_1 \text{ and } B) + \Pr(A_2 \text{ and } B)$
- Smallpox: find $\Pr(L)$ and $\Pr(I)$
 - ▶ $\Pr(L) = \Pr(I \text{ and } L) + \Pr(I^C \text{ and } L) = 0.038 + 0.825 = 0.863$

result	inoculated		
	yes	no	Total
	lived	5136	5374
died	6	844	850
Total	244	5980	6224

result	inoculated		
	yes	no	Total
	lived	0.825	0.863
died	0.136	0.137	
Total	0.961	0.039	1.000

Marginal probability: law of total probability

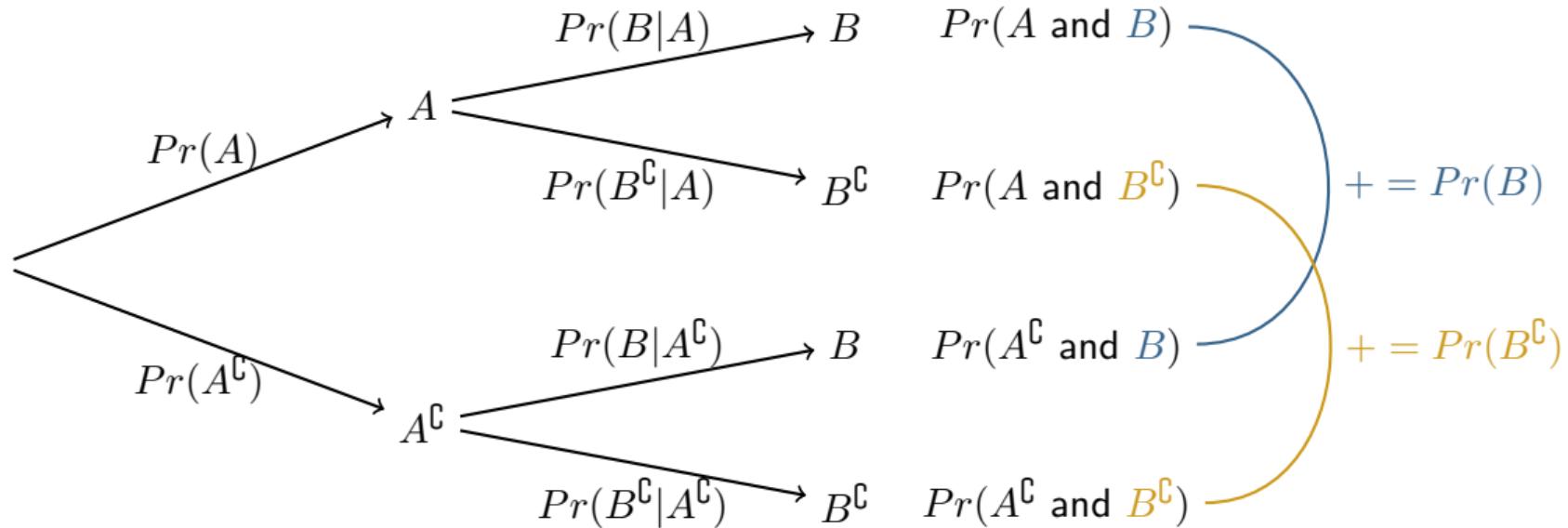
- Previously found “intuitively” from contingency table
- To find $\Pr(B)$
 - ▶ Sum over possible outcomes that could co-occur with the event B
- If there are two outcomes: A_1 and A_2
 - ▶ $\Pr(B) = \Pr(A_1 \text{ and } B) + \Pr(A_2 \text{ and } B)$
- Smallpox: find $\Pr(L)$ and $\Pr(I)$
 - ▶ $\Pr(I) = \Pr(I \text{ and } L) + \Pr(I \text{ and } L^C) = 0.038 + 0.001 = 0.039$

result	inoculated		Total
	yes	no	
	lived	5136	
died	6	844	850
Total	244	5980	6224

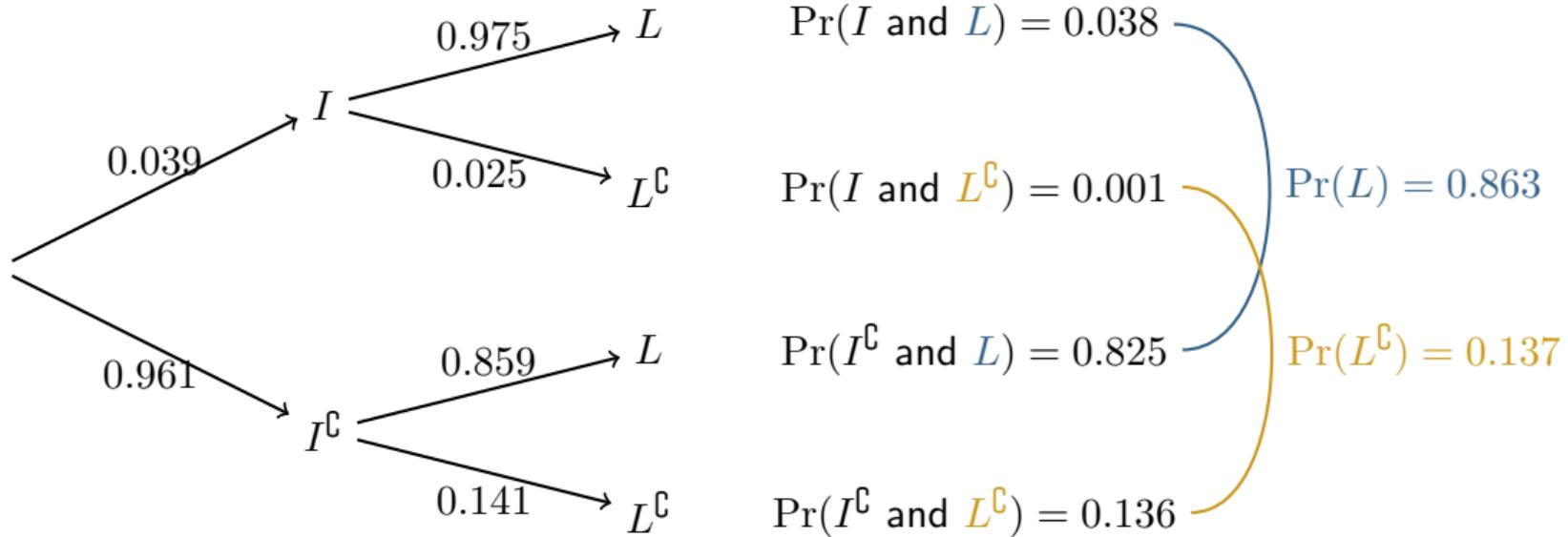
result	inoculated		Total
	yes	no	
	lived	0.825	
died	0.136	0.137	
Total	0.961	1.000	

Tree diagrams

- Tree diagrams are an alternate way to visualize outcomes and probabilities
- General form:

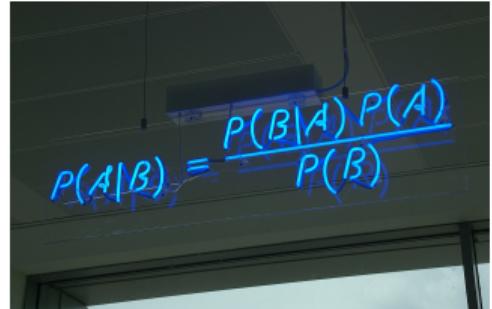


Tree diagrams: smallpox example



Bayes' theorem

- An important result in probability theory
 - ▶ Underpins a lot of modern statistics/data science/AI
 - Hopefully return to this at the end of the semester


$$P(A|B) = \frac{P(B|A) P(A)}{P(B)}$$

- Idea: find $\Pr(A|B)$ from $\Pr(B|A)$

$$\Pr(A|B) = \frac{\Pr(B|A) \Pr(A)}{\Pr(B)} = \frac{\Pr(B|A) \Pr(A)}{\Pr(B|A) \Pr(A) + \Pr(B|A^c) \Pr(A^c)}$$

Bayes' theorem

- Example: sleep apnea
- A : patient has sleep apnea
- S : patient snores
- Question: we know our patient snores. What is the prob they sleep apnea?
 - ▶ 90% of patients with sleep apnea snore: $\Pr(S|A) = 0.9$
 - ▶ 50% of patients without sleep apnea snore: $\Pr(S|A^C) = 0.5$
 - ▶ 5% of the population have sleep apnea: $\Pr(A) = 0.05$

Bayes' theorem

- Example: sleep apnea
- A : patient has sleep apnea
- S : patient snores
- Question: we know our patient snores. What is the prob they sleep apnea?
 - ▶ 90% of patients with sleep apnea snore: $\Pr(S|A) = 0.9$
 - ▶ 50% of patients without sleep apnea snore: $\Pr(S|A^C) = 0.5$
 - ▶ 5% of the population have sleep apnea: $\Pr(A) = 0.05$

$$\begin{aligned}\Pr(A|S) &= \frac{\Pr(S|A) \Pr(A)}{\Pr(S|A) \Pr(A) + \Pr(S|A^C) \Pr(A^C)} \\ &= \frac{0.9 \times 0.05}{0.9 \times 0.05 + 0.5 \times 0.95} = 0.087\end{aligned}$$

Bayes' theorem: example

- We had: $\Pr(S|A) = 0.9$
 - ▶ Tempting to think that $\Pr(A|S)$ will also be high
 - ▶ $\Pr(A|S) \approx 0.09$ seems surprisingly low
 - ▶ Confusing conditional probabilities is a common mistake
- Another perspective:
 - ▶ How does the probability of patient having sleep apnea change
- In the general population, we have
 - ▶ $\Pr(A) = 0.05$
- After learning that patient snores, the probability increases to
 - ▶ $\Pr(A|S) \approx 0.09$

Bayes' theorem: understanding

- It can be difficult to understand why $\Pr(A|S)$ is low when $\Pr(S|A)$ is high
- Construct a hypothetical ('expected') contingency table
 - ▶ Pretend there are 100 000 patients
 - ▶ Recall: $\Pr(A) = 0.05$

		snores (S)		Total
		yes	no	
sleep apnea (A)	yes			5000
	no			95 000
Total				100 000

Bayes' theorem: understanding

- It can be difficult to understand why $\Pr(A|S)$ is low when $\Pr(S|A)$ is high
- Construct a hypothetical ('expected') contingency table
 - ▶ Pretend there are 100 000 patients
 - ▶ Recall: $\Pr(S|A) = 0.9$

		snores (S)		Total
		yes	no	
sleep apnea (A)	yes	4500	500	5000
	no			95 000
Total				100 000

Bayes' theorem: understanding

- It can be difficult to understand why $\Pr(A|S)$ is low when $\Pr(S|A)$ is high
- Construct a hypothetical ('expected') contingency table
 - ▶ Pretend there are 100 000 patients
 - ▶ Recall: $\Pr(S|A^C) = 0.5$

		snores (S)		Total
		yes	no	
sleep apnea (A)	yes	4500	500	5000
	no	47 500	47 500	95 000
Total				100 000

Bayes' theorem: understanding

- It can be difficult to understand why $\Pr(A|S)$ is low when $\Pr(S|A)$ is high
- Construct a hypothetical ('expected') contingency table
 - ▶ Pretend there are 100 000 patients

		snores (S)		Total
		yes	no	
sleep apnea (A)	yes	4500	500	5000
	no	47 500	47 500	95 000
Total		52 000	48 000	100 000

- Most of those who snore do not have sleep apnea!
- $\Pr(A|S) = \frac{4500}{52\,000} = 0.087$

Summary

- Looked in more detail at conditional probability
- Generalised the multiplication rule
- Tree diagrams
- Bayes' theorem
 - ▶ Using formula
 - ▶ Constructing an expected contingency table
- Next week: begin exploring how to use probability to model data

STAT 110: Week 3

University of Otago

Outline

- Data summaries: sample mean and standard deviation
- Summaries are limited
 - ▶ To go further we needed statistical models
 - Use probability to describe the variation in the data
- Had an introduction to probability
- Today we will introduce idea of a random variable
 - ▶ Useful in helping us use probability to describe data

Example: bovine leptospirosis

- An inspector visits cattle & dairy farms for signs of bovine leptospirosis
- If they visit three farms, the sample space has eight possible outcomes
 - ▶ LLL, LLC, LCL, LCC, CLL, CLC, CCL, CCC
 - L: evidence of leptospira at farm
 - C: farm is clear
 - ▶ Each outcome has an associated probability
- If the inspector visits 30 farms, there are 1 073 741 824 possible outcomes
- The way the problem is expressed makes it difficult to answer questions:
 - ▶ How many farms would we expect to have evidence of leptospira?
 - ▶ How likely is it that 24 or more farms will have evidence of leptospira?
- We need a better way of writing/expressing things

Random variable

- A random variable assigns a numerical value to each outcome in sample space
- For our purposes, we can use a simpler definition:
 - ▶ A random variable is a (random) process with a numerical outcome
- Common to represent a random variable with capital letter
 - ▶ e.g. X or Y or Z
- The possible values are given with lowercase letters
 - ▶ e.g. x, y, z

Random variables: leptospirosis example

- Y represents the number of farms with evidence of leptospira
- Visit three farms
 - ▶ Four possible values: $y_1 = 0, y_2 = 1, y_3 = 2, y_4 = 3$
- Visit 30 farms
 - ▶ 31 possible values: $y_1 = 0, y_2 = 1, \dots, y_{31} = 30$.
- We may use i (or j) as an index of possible values
 - ▶ e.g. $i = 2$ is the second possible value; $y_i = y_2 = 1$
- We use the k to represent the number of possible values
 - ▶ $k = 4$ if we visit three farms
 - ▶ $k = 31$ if we visit 30 farms

Probability distribution

- A random variable has an associated probability distribution
- For the leptospirosis example

i	1	2	3	4	Total
y_i	0	1	2	3	
$\Pr(Y = y_i)$	0.25	0.15	0.4	0.2	1

- $\Pr(Y = y_i)$: the probability that (the random variable) Y takes the value y_i
 - ▶ e.g. for $i = 3$: $\Pr(Y = 2) = 0.4$, the probability that Y takes the value 2

Probability distribution: example

- Suppose we open an online store that sells two products
- A given online visitor may:
 - ▶ With probability 0.4 buy nothing: we receive \$0
 - ▶ With probability 0.3 buy item A: we receive \$20
 - ▶ With probability 0.2 buy item B: we receive \$35
 - ▶ With probability 0.1 buy item A and B: we receive \$50
- If Y represents the money we receive from an online visitor

i	1	2	3	4	Total
y_i	0	20	35	50	
$\Pr(Y = y_i)$	0.4	0.3	0.2	0.1	1

Using probability distributions

- With these definitions we can start to ask useful questions
 - ▶ How likely is it that 2 or more farms will have evidence of leptospira?
 - ▶ How likely is it that we will receive \$20 or below from an online visitor?

Using probability distributions

- With these definitions we can start to ask useful questions
 - ▶ How likely is it that 2 or more farms will have evidence of leptospira?
 - ▶ How likely is it that we will receive \$20 or below from an online visitor?
- We use results from last week to answer those questions
- Using the online store as an example
 - ▶ Think of the y values as events: $y_1 = 0$, $y_2 = 20$, $y_3 = 35$, $y_4 = 50$
 - ▶ The events are mutually exclusive
 - ▶ $\Pr(Y \leq 20) = \Pr(Y = 0 \text{ or } Y = 20) = \Pr(Y = 0) + \Pr(Y = 20) = 0.4 + 0.3 = 0.7$

Expectation

- We can't yet answer the other question from earlier
 - ▶ How many farms would we expect to have evidence of leptospira? or
 - ▶ How much money do we expect to receive from an online visitor?
- We want to find $E[Y]$, the expected value of the random variable Y
 - ▶ The expected value is the same as the mean and is often represented by μ
- To find this, we weight each possible value by its corresponding probability

$$E[Y] = \sum_{i=1}^k y_i \Pr(Y = y_i)$$

- k is the number of possible values (in both our examples $k = 4$)
 - ▶ $E[Y] = y_1 \Pr(Y = y_1) + y_2 \Pr(Y = y_2) + y_3 \Pr(Y = y_3) + y_4 \Pr(Y = y_4)$

Expectation: leptospirosis example

- How many farms would we expect to have evidence of leptospira?

i	1	2	3	4	Total
y_i	0	1	2	3	
$\Pr(Y = y_i)$	0.25	0.15	0.4	0.2	1

$$E[Y] = \underbrace{0 \times 0.25}_0 + \underbrace{1 \times 0.15}_{0.15} + \underbrace{2 \times 0.4}_{0.8} + \underbrace{3 \times 0.2}_{0.6} = 1.55$$

- We expect to find 1.55 farms with evidence of leptospira infection

Expectation: online store

- How much money do we expect to receive from an online visitor?

i	1	2	3	4	Total
y_i	0	20	35	50	
$\Pr(Y = y_i)$	0.4	0.3	0.2	0.1	1

$$E[Y] = \underbrace{0 \times 0.4}_0 + \underbrace{20 \times 0.3}_6 + \underbrace{35 \times 0.2}_7 + \underbrace{50 \times 0.1}_5 \\ = 18$$

- We expect to receive \$18 from an online visitor

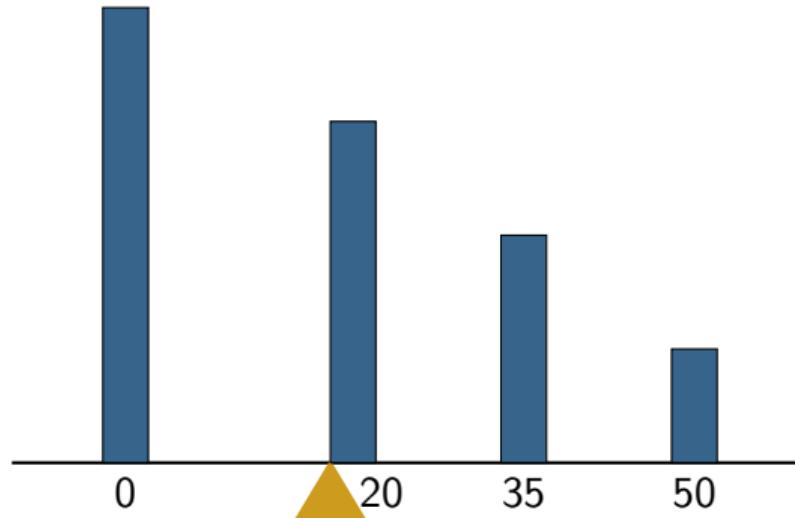
Expectation: intuition

i	1	2	3	4	Total
y_i	0	20	35	50	
$\Pr(Y = y_i)$	0.4	0.3	0.2	0.1	1

- If we saw 100 online visitors
 - ▶ We would expect 40 of them to spend nothing: receive \$0
 - ▶ We would expect 30 of them to spend \$20: receive \$600
 - ▶ We would expect 20 of them to spend \$35: receive \$700
 - ▶ We would expect 10 of them to spend \$50: receive \$500
- We would expect to receive \$1800 per 100 visitors = \$18 per visitor
- Multiplying y_i by $\Pr(Y = y_i)$ is taking a 'direct route' to this answer

Expectation: intuition

- Another way we can look at expectation is by thinking of the probability distribution as a old-fashioned scale
- The expected value balances the probability distribution (gold triangle)



Variance

- We could also ask questions that relate to variability
 - ▶ How much would we expect income from our store to vary from one day to the next?
- For small problems (like those we have been looking at)
 - ▶ Probably preferable to base this off the probability distribution
- For larger problems (which we are moving toward)
 - ▶ We need a measure of variability
 - ▶ Typically use variance / standard deviation

Variance

- The variance of the random variable Y is $\text{Var}(Y)$
 - ▶ Find the average of squared deviations from the mean
 - ▶ Weight the squared deviations by their probability

$$\text{Var}(Y) = \sum_{i=1}^k (y_i - E[Y])^2 \Pr(Y = y_i)$$

- For $k = 4$
 - ▶ $\text{Var}(Y) = (y_1 - E[Y])^2 \Pr(Y = y_1) + (y_2 - E[Y])^2 \Pr(Y = y_2) + (y_3 - E[Y])^2 \Pr(Y = y_3) + (y_4 - E[Y])^2 \Pr(Y = y_4)$

Variance: leptospirosis example

- What is the variance in the number of farms that have evidence of leptospira?
 - We know $E[Y] = 1.55$

i	1	2	3	4	Total
y_i	0	1	2	3	
$\Pr(Y = y_i)$	0.25	0.15	0.4	0.2	1

$$\begin{aligned}\text{Var}(Y) &= \underbrace{(0 - 1.55)^2 \times 0.25}_{2.4025 \times 0.25} + \underbrace{(1 - 1.55)^2 \times 0.15}_{0.3025 \times 0.15} + \underbrace{(2 - 1.55)^2 \times 0.4}_{0.2025 \times 0.4} + \underbrace{(3 - 1.55)^2 \times 0.2}_{2.1025 \times 0.2} \\ &= 1.1475\end{aligned}$$

Standard deviation

- The standard deviation is the square root of variance
 - ▶ $\text{sd}(Y) = \sqrt{\text{Var}(Y)}$
- For the leptospirosis example
 - ▶ $\text{sd}(Y) = \sqrt{1.1475} = 1.07$
- The standard deviation is (approximately) the average deviation from the mean
- Often the variance will be represented by σ^2
 - ▶ The standard deviation as σ

Example: online visitors

- What is the variance in the amount we receive from an online visitor?
 - ▶ We know $E[Y] = 18$

i	1	2	3	4	Total
y_i	0	20	35	50	
$\Pr(Y = y_i)$	0.4	0.3	0.2	0.1	1

$$\begin{aligned}\text{Var}(Y) &= \underbrace{(0 - 18)^2 \times 0.4}_{324 \times 0.4} + \underbrace{(20 - 18)^2 \times 0.3}_{4 \times 0.3} + \underbrace{(35 - 18)^2 \times 0.2}_{289 \times 0.2} + \underbrace{(50 - 18)^2 \times 0.1}_{1024 \times 0.1} \\ &= 291\end{aligned}$$

$$\text{sd}(Y) = 17.1$$

We've seen this before

- We saw expectation (mean), standard deviation, and variance in Week 1
 - ▶ Sample mean, sample variance, sample standard deviation
 - ▶ These are summaries of a particular data set (a sample)
- Today we've found these quantities for a distribution
 - ▶ Summaries of a random variable
 - ▶ Tells us something about what realizations from the distribution should look like

Summary

- Introduced random variables
- Probability distribution of random variable
- Saw several summaries of random variables
 - ▶ Mean
 - ▶ Variance
 - ▶ Standard deviation

Outline

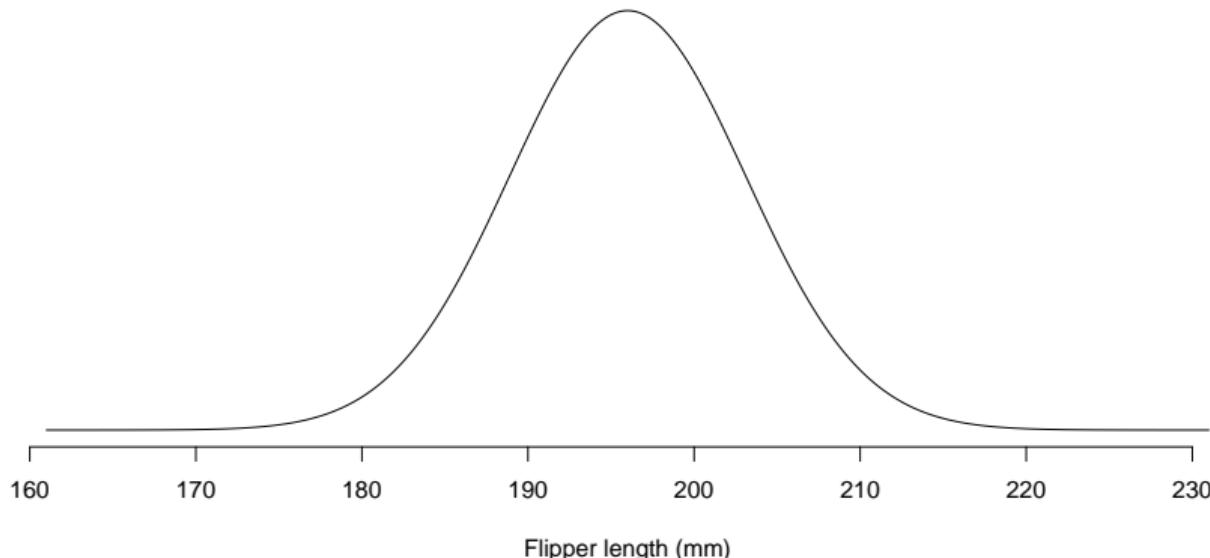
- We saw random variables in the last lecture
 - ▶ Probability distribution
 - ▶ Expectation
 - ▶ Variance
- Continue learning about random variables today
 - ▶ Can we have continuous random variables?
 - ▶ What happens when we combine random variables?

Discrete vs continuous

- The random variables we looked at in the last lecture were all discrete
 - ▶ Countable number of distinct values
- Discrete random variables are useful in a range of problems, e.g.
 - ▶ Number of eggs in a nest
 - ▶ Number of tasks completed in fixed time
 - ▶ Number of bugs in a piece of computer code
 - ▶ Number of voters who prefer candidate X
- There are other situations where things aren't discrete, e.g.
 - ▶ The flipper length of a gentoo penguin
 - ▶ The time taken in reflex test
 - ▶ The pH of seawater
- These can take continuous values

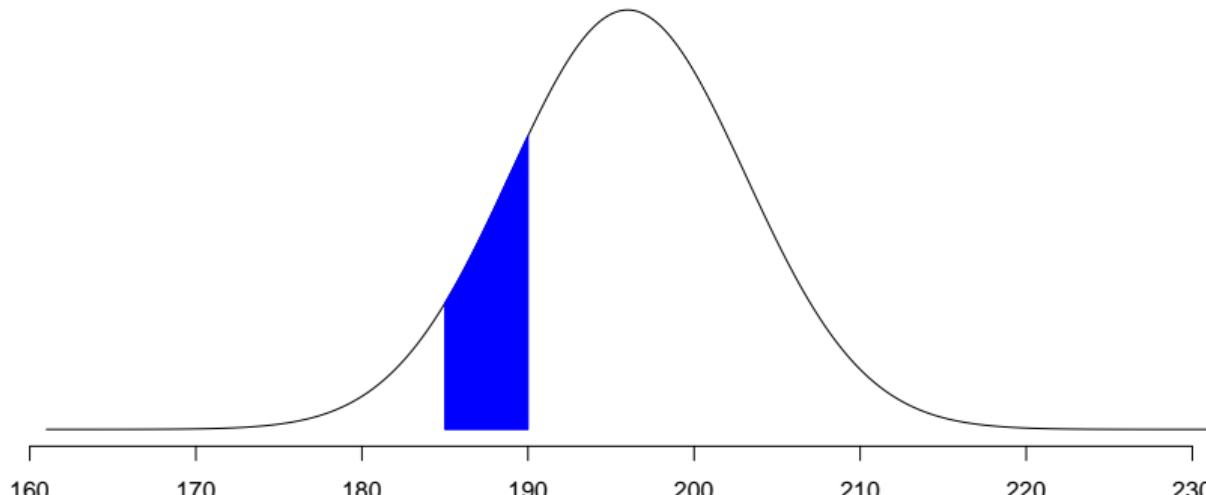
Continuous random variables

- An infinite (and uncountable) number of possible values
- Each value has a probability density
 - ▶ Best seen graphically (e.g. for flipper length)



Probability density

- This curve is called a probability density function (pdf)
- Probability is given by the area under the curve (pdf)
 - ▶ The total area under the curve (pdf) is 1
- The probability of flipper length between 185 and 190 mm is given by:



Continuous vs discrete

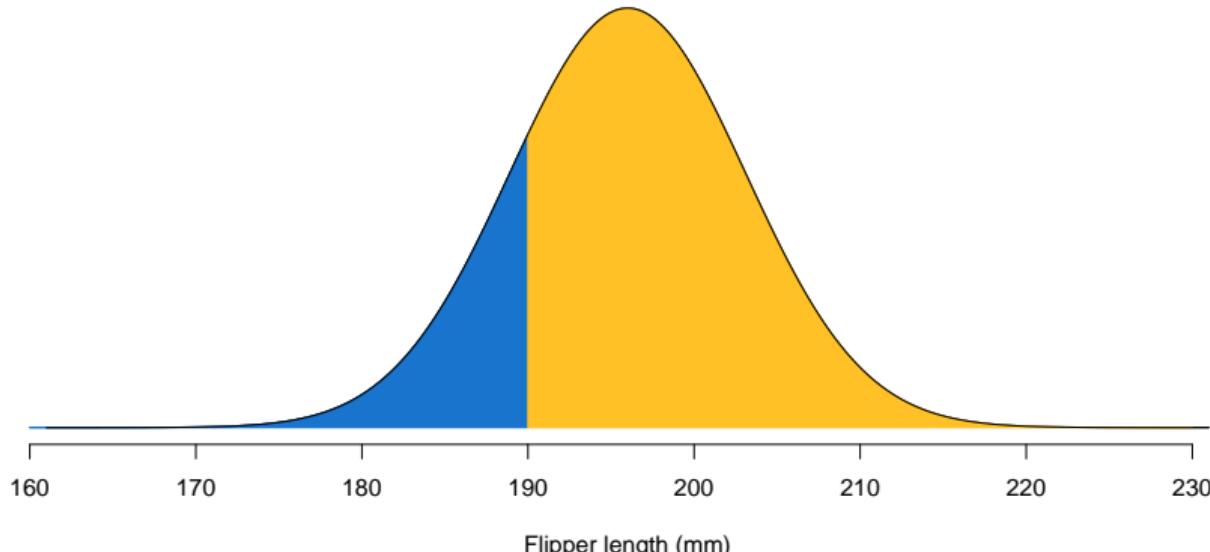
- Much of what we have already learned applies to continuous random variables
 - ▶ We can find expectation, variance, standard deviation
 - ▶ The calculations are more complex (sums are replaced by integrals)
 - ▶ Explore in more detail in more advanced courses (e.g. STAT 270)

Continuous

- Much of what we have already learned applies to continuous random variables
 - ▶ We can find expectation, variance, standard deviation
 - ▶ The calculations are more complex (sums are replaced by integrals)
 - ▶ Explore in more detail in more advanced courses (e.g. STAT 270)
- Look at examples on the next two slides

Complement

- Suppose we know the probability that flipper length is less than 190 mm (blue)
 - ▶ $\Pr(\text{flipper length} < 190) = 0.2$
- What is $\Pr(\text{flipper length} > 190)$? (gold)
 - ▶ It is a complement!



Combinations of random variables

- We may be interested in the combination of several random variables
 - ▶ Adélie penguins: feeding trip time
 - Random variables: time spent (i) feeding, (ii) resting, (iii) transit, in a trip
 - Combination: total trip time
 - ▶ Genetic linkage (crossover¹)
 - Random variables: number of crossovers in each chromosome
 - Combination: total number of crossovers
 - ▶ Cricket: runs scored
 - Random variables: number of singles, twos, threes, fours, sixes in an innings.
 - Combination: total score
 - ▶ Finance: portfolio value
 - Random variables: share prices for spark (SPK) and port of Tauranga (POT)
 - Combination: portfolio value (e.g. portfolio: 5 SPK, 10 POT)

¹ segments of DNA from one parent's chromosome swap with corresponding segments on the other parent's chromosome during meiosis

Combination of random variables

- Suppose we have random variables X and Y
 - ▶ To guide the development, we will think about
 - X : value of one SPK share in one months time
 - Y : value of one POT share in one months time
- We may be interested in a linear combination of X and Y
 - ▶ $aX + bY$
- What is the expected value of $aX + bY$?
- What is the variance of $aX + bY$

Expected value of combination

- If we owned shares: 5 SPK and 10 POT
 - ▶ Linear combination represents the value of our portfolio in one months time
 - ▶ $5X + 10Y$
 - Here, a is the number of SPK shares: 5
 - Here, b is the number of POT shares: 10
- How do we find the expected value of the linear combination?

$$E[aX + bY] = aE[X] + bE[Y]$$

- If $E[X] = 3$ and $E[Y] = 6.3$ then, the expected portfolio value is

$$\begin{aligned} E[5X + 10Y] &= 5E[X] + 10E[Y] \\ &= 5 \times 3 + 10 \times 6.3 \\ &= 78 \end{aligned}$$

Expected value of combination

- Ice cream is sold from 16 L containers in NZ
 - ▶ Expect that there is 16 L when opened
 - ▶ Can vary: let's say a standard deviation of 0.1 L (variance 0.01)
 - ▶ Let X be the amount of ice cream in a container: $E[X] = 16$, $Var(X) = 0.01$
- A new container of goldrush icecream is opened for the person ahead of us in line.
- They get a scoop of gold rush
 - ▶ Expect each scoop to get 0.1 L of ice cream
 - ▶ Standard deviation of 0.01 L (variance 0.0001).
 - ▶ Let Y be the amount in a scoop of ice cream: $E[Y] = 0.1$, $Var(Y) = 0.0001$
- The amount of goldrush icecream when we come to order is $X - Y$
 - ▶ What is $E[X - Y]$?

Variance of combination

- Can also be important to have a measure of variability for the combination of random variables
 - ▶ Trip time for Adélie penguins
 - ▶ Number of crossovers
 - ▶ Runs in cricket innings
 - ▶ Value of portfolio
- If X and Y are independent, then

$$\text{Var}(aX + bY) = a^2\text{Var}(X) + b^2\text{Var}(Y)$$

- If X and Y are not independent
 - ▶ The variance is more complicated (additional term needed)
 - ▶ Considered in higher level courses

Variance of combination

- What is $Var(X - Y)$ for ice cream example?
 - ▶ $a = 1$
 - ▶ $b = -1$

$$\begin{aligned}Var(X - Y) &= 1^2 Var(X) + (-1)^2 Var(Y) \\&= Var(X) + Var(Y) \\&= 0.01 + 0.0001 \\&= 0.0101\end{aligned}$$

- Portfolio: what is $Var(5X + 10Y)$?
 - ▶ Assume that share prices are independent (unlikely to be the case in reality)

Variance of combination

- We saw that $Var(X - Y) = Var(X) + Var(Y)$
 - ▶ We are subtracting Y from X . Why do the variances add?
- A server with low variability
 - ▶ Each scoop has is consistent in terms of the amount of ice cream
- A server with high variability
 - ▶ Each scoop can vary greatly (small or large or anywhere in between)
- If server is highly variable, will amount left in container be highly variable?

Variance of combination

- We saw that $\text{Var}(X - Y) = \text{Var}(X) + \text{Var}(Y)$
 - ▶ We are subtracting Y from X . Why do the variances add?
- A server with low variability
 - ▶ Each scoop has is consistent in terms of the amount of ice cream
- A server with high variability
 - ▶ Each scoop can vary greatly (small or large or anywhere in between)
- If server is highly variable, will amount left in container be highly variable?
- The variability in the amount of ice cream is the same if:
 - ▶ Add a scoop of ice cream to the container, or
 - ▶ Took a scoop of ice cream away

Abstract example

- Look at another example: somewhat abstract
 - ▶ Provide some useful results that we will use in coming weeks
- Let Y_1 and Y_2 be independent observations from a distribution
 - ▶ Mean μ
 - ▶ Standard deviation σ
- What is the mean and variance of $\frac{Y_1+Y_2}{2}$?
 - ▶ Sample mean of two values from a distribution

Abstract example: expected value

- The expected value of the sample mean is

$$\begin{aligned} E\left[\frac{Y_1 + Y_2}{2}\right] &= \frac{1}{2}E[Y_1] + \frac{1}{2}E[Y_2] \\ &= \frac{1}{2}\mu + \frac{1}{2}\mu \\ &= \mu \end{aligned}$$

- The variance of the sample mean is

$$\begin{aligned} Var\left(\frac{Y_1 + Y_2}{2}\right) &= \frac{1}{4}Var(Y_1) + \frac{1}{4}Var(Y_2) \\ &= \frac{1}{4}\sigma^2 + \frac{1}{4}\sigma^2 \\ &= \frac{\sigma^2}{2} \end{aligned}$$

Abstract example: extension

- This can be extended to when we have n independent observations: Y_1, Y_2, \dots, Y_n
- The expected value of the sample mean is

$$\begin{aligned} E\left[\frac{Y_1 + Y_2 + \dots + Y_n}{n}\right] &= \frac{1}{n}E[Y_1] + \frac{1}{n}E[Y_2] + \dots + \frac{1}{n}E[Y_n] \\ &= \mu \end{aligned}$$

- The variance of the sample mean is

$$\begin{aligned} Var\left(\frac{Y_1 + Y_2 + \dots + Y_n}{n}\right) &= \frac{1}{n^2}Var(Y_1) + \frac{1}{n^2}Var(Y_2) + \dots + \frac{1}{n^2}Var(Y_n) \\ &= \frac{\sigma^2}{n} \end{aligned}$$

Summary

- Looked at continuous random variables
 - ▶ There are differences, but much remains the same
- Looked at combination of random variables
 - ▶ Expectation
 - ▶ Variance
- Next lecture: start developing models for data

Outline

- Introduction to statistical modeling
 - ▶ Populations and parameters
 - ▶ Samples and statistics
 - ▶ Estimation of parameters
 - ▶ Introduce the normal distribution

Big picture

- We may be interested in flipper lengths of gentoo penguins in the Palmer archipelago
 - ▶ e.g. what is the mean flipper length?
- How could we find the mean flipper length of gentoos on Palmer?

Big picture

- We may be interested in flipper lengths of gentoo penguins in the Palmer archipelago
 - ▶ e.g. what is the mean flipper length?
- How could we find the mean flipper length of gentoos on Palmer?
- Problem: question refers to population (of gentoos in the archipelago)
 - ▶ We cannot answer it unless we measure every individual in the population
 - ▶ Likely impossible
- Formulate a statistical model
 - ▶ Use a sample to tell us about the population

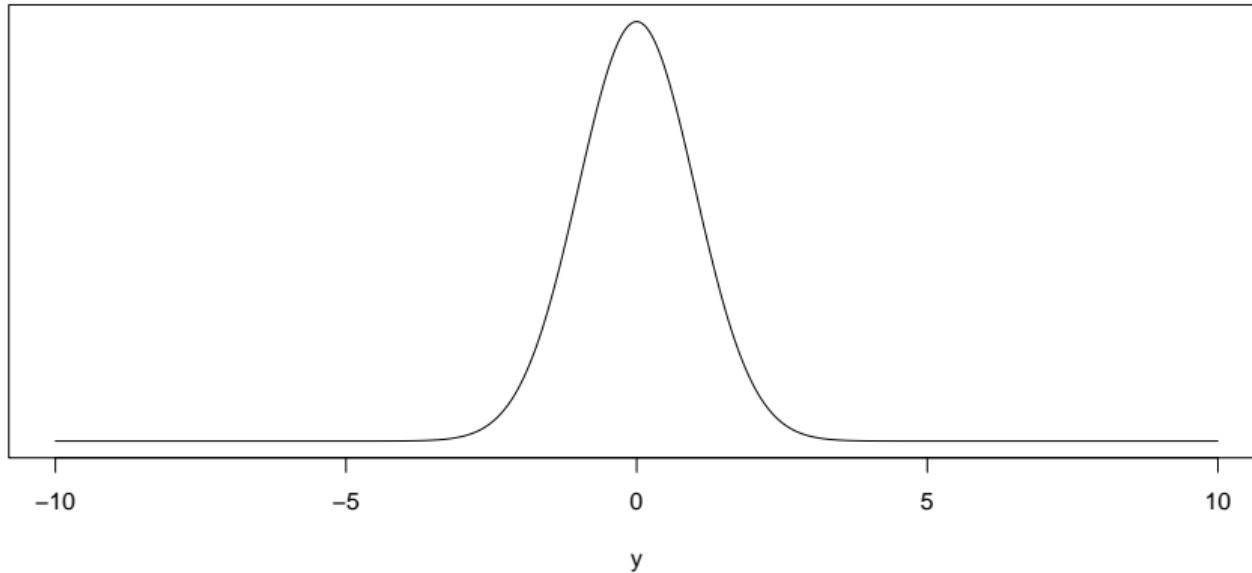
Statistical model

- Idea: assume the population values follow some distribution
 - ▶ The distribution tell us how the values vary in the population
- The distribution has unknown parameters
 - ▶ Parameter: any quantity that describes a population
- It is the parameter(s) that are of interest
 - ▶ Tell us about the population
- Abstract concepts
 - ▶ Introduce an example to make the idea more concrete
- We have seen probability distributions in simple ‘generic’ cases
 - ▶ Introduce specific case: normal distribution

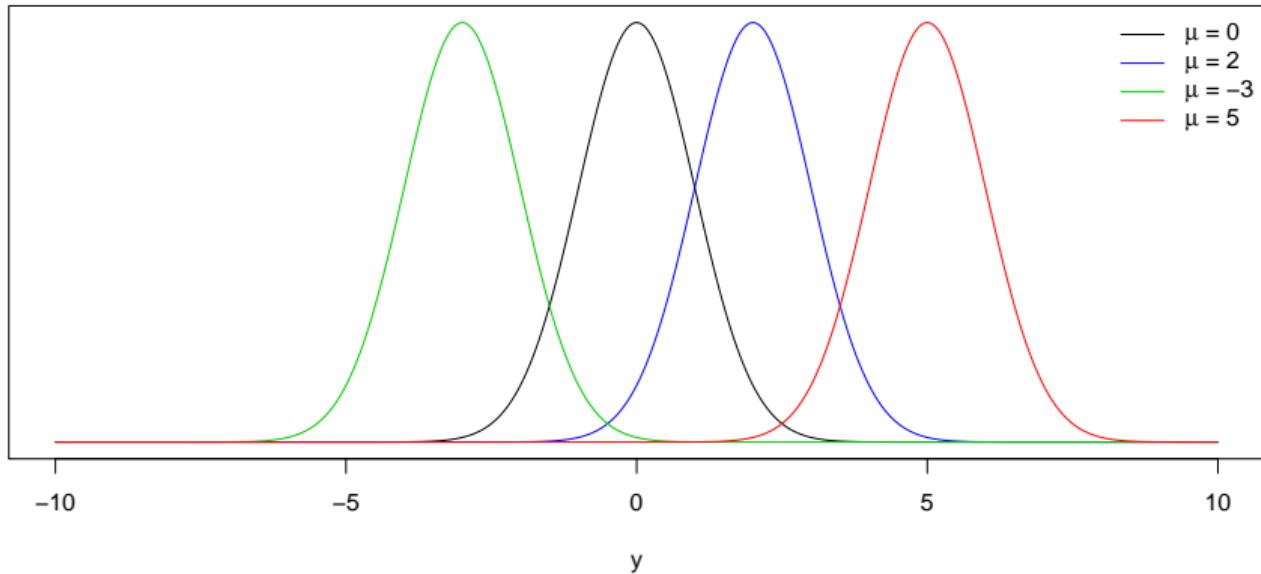
Normal model

- Assume that the data are normally distributed
 - ▶ We might also say we are using a normal model
- The normal distribution is sometimes called:
 - ▶ Bell-shaped curve
 - ▶ Gaussian model
- Described by two parameters:
 - ▶ Mean μ (Greek letter mu)
 - ▶ Standard deviation σ (Greek letter sigma)
 - Often refer to the variance σ^2 instead of the standard deviation
- We will spend some time familiarizing ourselves with the normal distribution

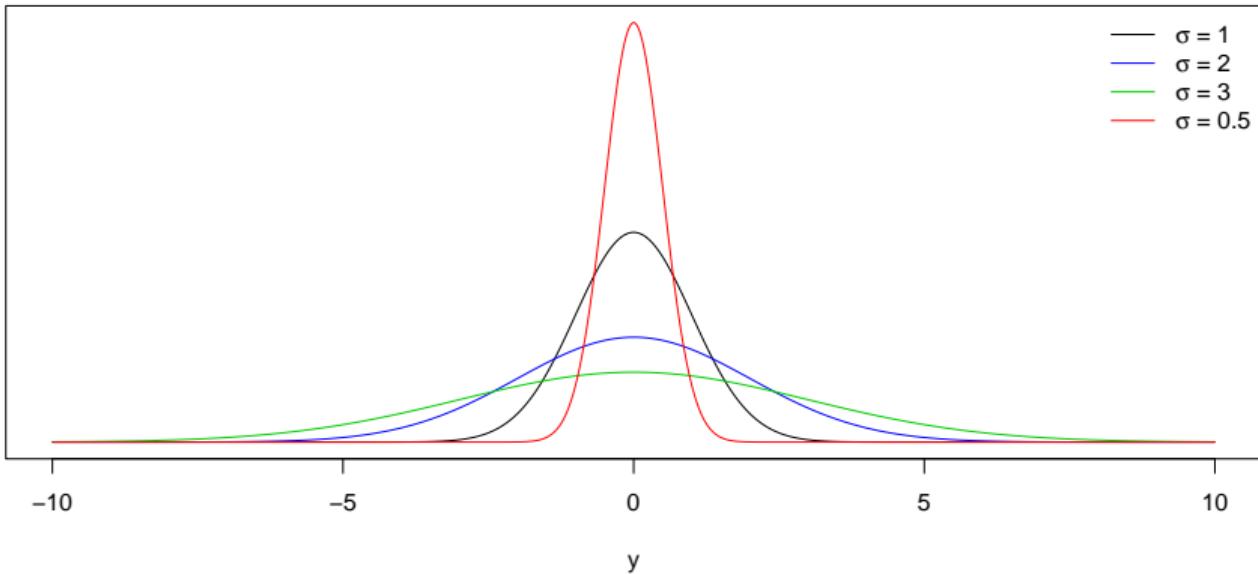
Probability density function (pdf) of normal distribution: $\mu = 0$, $\sigma = 1$



Pdf of normal distribution: different μ



Pdf of normal distribution: different σ



Model for gentoo

- We assume flipper lengths follow a normal distribution
 - ▶ This is an assumption about the population of gentoo penguins in Palmer archipelago
 - ▶ Parameters μ and σ are unknown
 - μ : mean flipper length (population level)
 - σ : standard deviation of flipper lengths (population level)
- Typically use greek letters for parameters
 - ▶ Here we are using μ and σ

Populations and samples

- Big idea: use a sample (and statistics) to estimate parameters
 - ▶ The estimate is an educated guess at the parameter value
- We have flipper length measurements from 68 gentoo penguins (cf. week 1)
- How could we use this sample to estimate μ ?

Populations and samples

- Big idea: use a sample (and statistics) to estimate parameters
 - ▶ The estimate is an educated guess at the parameter value
- We have flipper length measurements from 68 gentoo penguins (cf. week 1)
- How could we use this sample to estimate μ ?
- The sample mean \bar{y} could be used to estimate the population mean μ
 - ▶ The sample mean \bar{y} is an example of a statistic
 - ▶ Statistic: any quantity computed from values in a sample

That's easy ... are we done?

- Our example: finding a 'suitable' statistic is straightforward
 - ▶ Not always the case
- Let's imagine a more extensive penguin study:
 - ▶ Interested in understanding how feeding patterns, spatial structure (within a colony and between colonies), time of year, (and other factors) might influence penguin condition
 - What statistic(s) should we use for that?
- Later in semester we will (hopefully) think more about general strategies for finding suitable statistics (estimators)

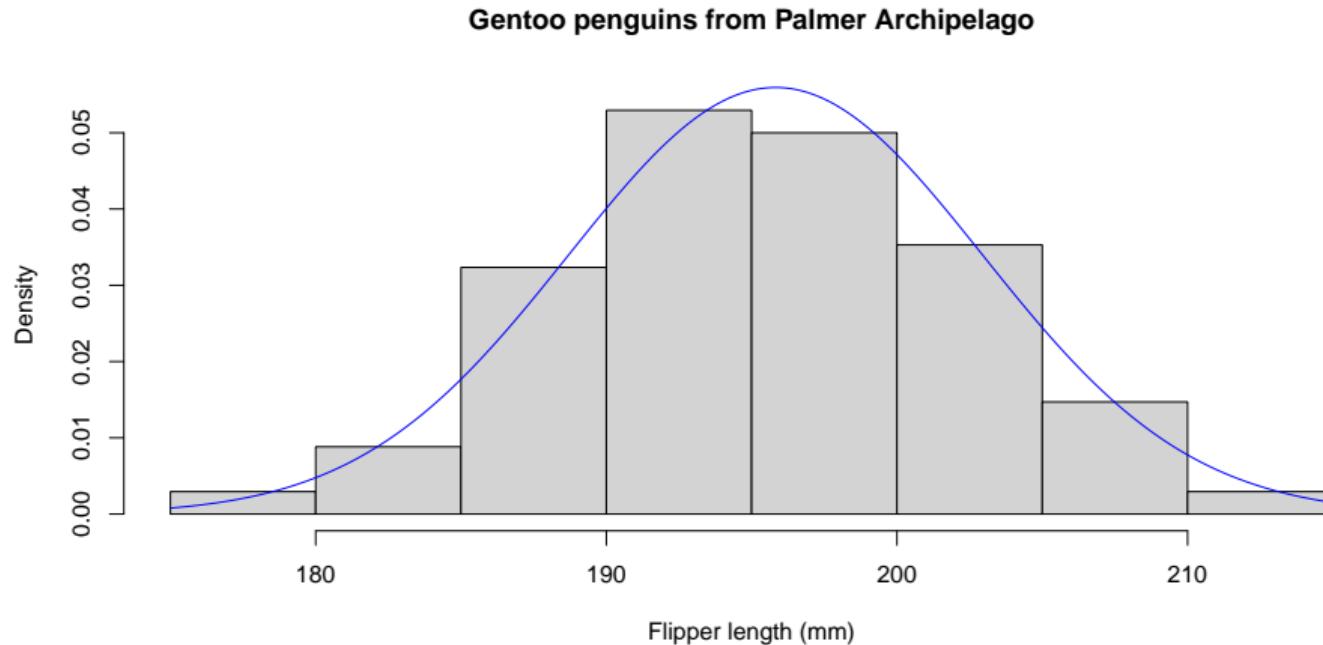
Model fitting

- For our model, we have
 - ▶ $\hat{\mu} = \bar{y}$
 - ▶ $\hat{\sigma} = s$
 - ▶ The population std deviation (σ) is being estimated by the sample std deviation (s)
- We have used the hat symbol $\hat{}$ to represent that we are estimating a parameter
 - ▶ $\hat{\mu}$ is said “mu-hat”
 - ▶ $\hat{\mu} = \bar{y}$: the parameter μ is being estimated by \bar{y} (a statistic)

Fitted model

- Look at the fitted model (graphically)
 - ▶ (Normal) model at the estimated parameter values
- Compare the fitted model to the data
 - ▶ Load the gentoo penguin data into R (for the next few slides)
 - ▶ Revise material from week 1

Fitted model

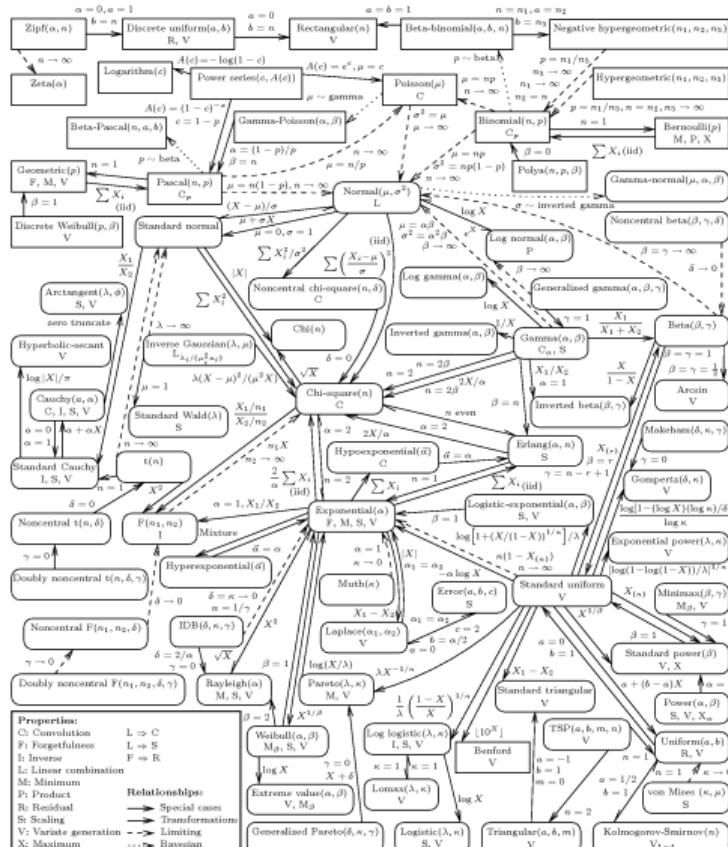


Statistical models

- Common mistakes:
 - ▶ Believing that μ is 195.8 mm (the sample mean)
 - This is the estimate of μ
 - It is impossible to know the value of μ
 - ▶ Believing that the model is true
 - Hope the model chosen is a reasonable approximation to reality
 - We should check this
- Checking the model fit
 - ▶ Looking to see if the model and the data are 'out of sync'
 - Plot: normal appears to describe the data reasonably well
 - ▶ Think a lot more about model fit in a couple of weeks (regression)

Statistical models

- In this example we have used a normal model for the data
- Should be reasonable for what we believe about flipper length:
 - ▶ Continuous values (flipper lengths can take any (positive) value)
 - ▶ Reasonably close to symmetric
 - An adult gentoo is unlikely to have a flipper four times the size of another adult gentoo
 - Cf. income
- Not all data looks like this!
 - ▶ Different types of data (yes/no, count, categories, time, space, ...)
 - ▶ Different characteristics (e.g. income)
 - ▶ Different complexity
- Many probability distributions with different characteristics
 - ▶ Next figure is for illustration, we don't need to learn it!



Looking forward I

- We will be working with a normal model for a few weeks
 - ▶ Look more at the normal distribution
 - Use it to describe (and model) data
 - Want to understand it as much as possible
- Explore a strategy for estimating parameters
 - ▶ Barely scratch the surface
 - ▶ Cover in more depth in higher level courses STAT 270, 370, 371
- Explore ‘extensions’ to normal model (e.g. regression)
- Explore models for other types of data: yes/no data

Looking forward II

- What does our estimate tell us about the parameter?
 - ▶ We have an estimate of μ from the sample of size 68
 - ▶ How ‘close’ to the true value of μ is it likely to be?
- Is the estimate likely to be better / worse if it were from:
 - ▶ A sample of size 6?
 - ▶ A sample of size 600?
- Explore how to determine how precise/uncertain the estimate is
- Also important is how were the data were collected?
 - ▶ e.g. does our sample consist of only adults?
 - ▶ We'll come back to this later in the semester
- Use the model for prediction (regression)

Summary

- Introduction to statistical modeling
 - ▶ Fit a normal model
 - ▶ Estimated the parameter μ with the statistic \bar{y}
- Next: get a better understanding of the normal distribution

STAT 110: Week 4

University of Otago

Outline

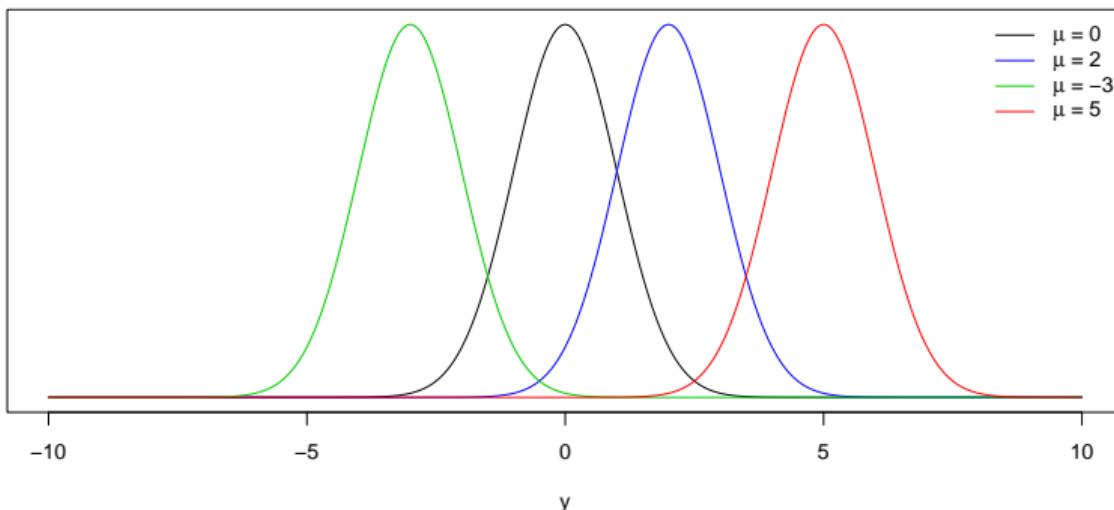
- Previous lectures:
 - ▶ Introduction to probability, random variables
 - ▶ First example of a statistical model
 - Normal model
- Today: learn more about the normal distribution

Normal distribution

- We used a normal model to describe flipper length (gentoo penguins in Palmer archipelago)
- Is the normal model appropriate
 - ▶ Does it make sense scientifically
 - Understand ‘properties’ of a normal distribution
 - Looked at some aspects in last lecture
 - Understand more about the normal distribution today
 - ▶ After estimation: check model fit
 - Looked briefly at this in last lecture
 - Consider it further in future lectures

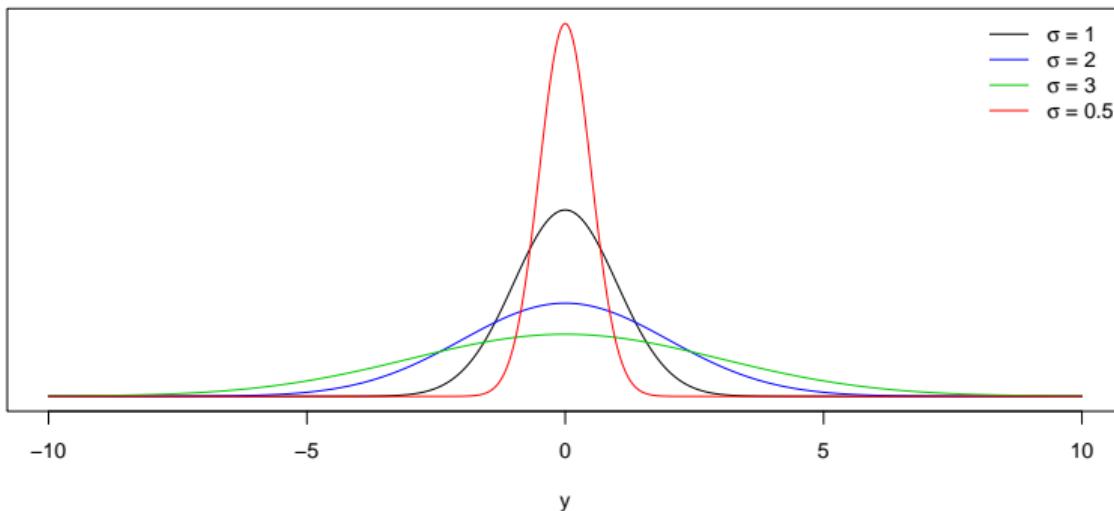
Recap: normal distribution

- Described by two parameters
 - ▶ Mean μ
 - ▶ Standard deviation σ
- Changing μ shifts the pdf side to side



Recap: normal distribution

- Described by two parameters
 - ▶ Mean μ
 - ▶ Standard deviation σ
- Changing σ compresses or expands the pdf



IQ scores

- IQ tests are designed so that scores are (approximately) normally distributed
 - ▶ $\mu = 100$
 - ▶ $\sigma = 15$
- We may be interested in knowing things like:
 - ▶ What is the probability of a randomly chosen individual scoring less than 85?
 - ▶ What is the probability of a randomly chosen individual scoring between 85 and 115?
 - ▶ For membership Mensa require a score at or above the 98th percentile on certain standardized IQ tests. For an IQ test (as above) what score would you need?
- All of these require us to be able to find probabilities from the normal distribution

Probabilities

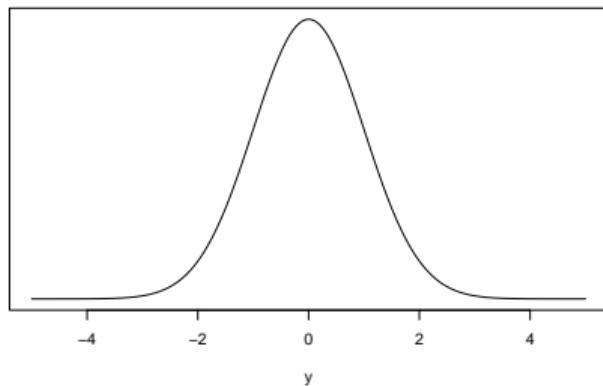
- Recall: we find probabilities by finding the area under pdf
- The normal pdf is a mathematical function: $f(y) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y - \mu)^2\right)$
 - ▶ Not expected (or required) to remember or understand this
 - ▶ Mathematical representation of the pdfs we saw in earlier slides
- Theory: to find probabilities we can use calculus and integrate $f(y)$ ¹
 - ▶ Problem: can't integrate $f(y)$ by hand
- Historical solution: tables of values we could refer to
 - ▶ Problem: lots of possible values of μ and σ
 - ▶ Solution: find them for a single standardized version of the distribution

¹Integration can be thought of as (mathematically) finding the area under curve

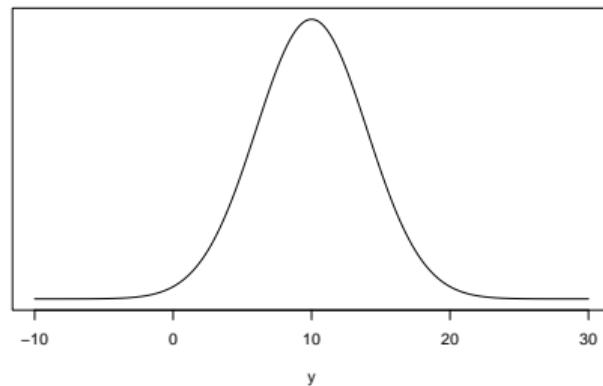
Standard normal distribution

- Normal pdfs have the same shape
 - ▶ Irrespective of the value of μ, σ
 - Hard to see on the previous plots
 - More clear if change the scale of the axes for different values of μ, σ

normal pdf: $\mu = 0, \sigma = 1$



normal pdf: $\mu = 10, \sigma = 4$



- Idea: work with a standard normal distribution: $\mu = 0, \sigma = 1$

Standardizing

- Idea: define a standard normal distribution
 - ▶ $\mu = 0, \sigma = 1$
- Find probabilities, etc, for this standard distribution
- Convert a value (y) to a z -score
 - ▶ y -value from distribution with mean μ and standard deviation σ
 - ▶ z -score from distribution with mean 0 and standard deviation 1
 - ▶ Going from y to z is often called standardizing
- The z -score tells us how many standard deviations above the mean a value is
 - ▶ $z = 1$: value is 1 standard deviation above the mean
 - ▶ $z = -1.5$: value is 1.5 standard deviations below the mean

Standardizing

- We can find a z -score from y

$$z = \frac{\text{value} - \text{mean}}{\text{standard deviation}} = \frac{y - \mu}{\sigma}$$

- IQ test of $y = 115$:

$$z = \frac{y - \mu}{\sigma} = \frac{115 - 100}{15} = 1$$

- ▶ An IQ test of 115 is one standard deviation above the mean
- We can also find y from a z -score

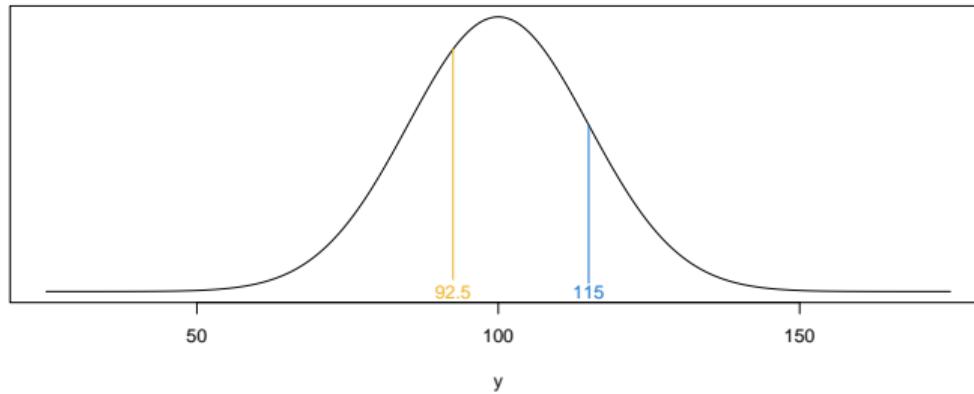
$$y = \mu + z\sigma$$

- A z -score of 1 for IQ corresponds to a score of:

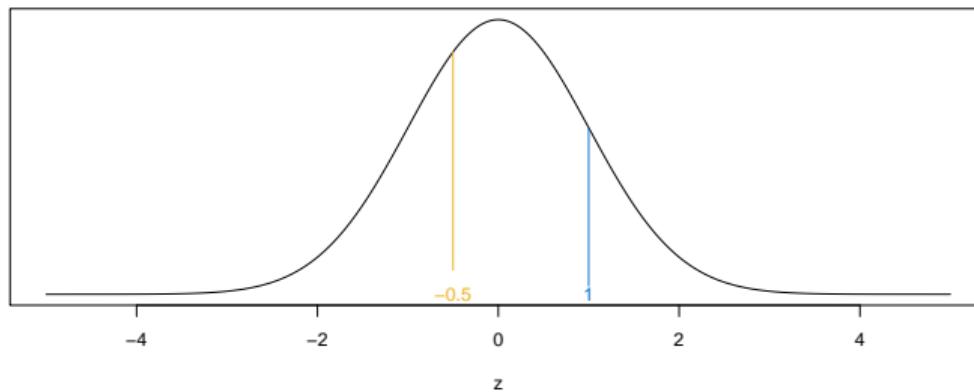
$$y = 100 + 1 \times 15 = 115$$

Graphical representation

normal pdf: $\mu = 100, \sigma = 15$

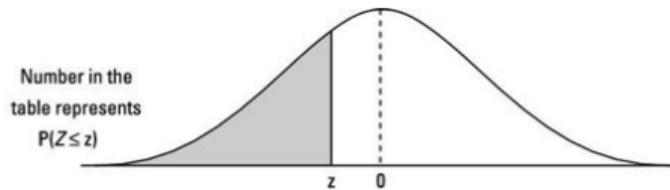


normal pdf: $\mu = 0, \sigma = 1$



Finding probabilities: deep dark past

- We used to find probabilities from tables



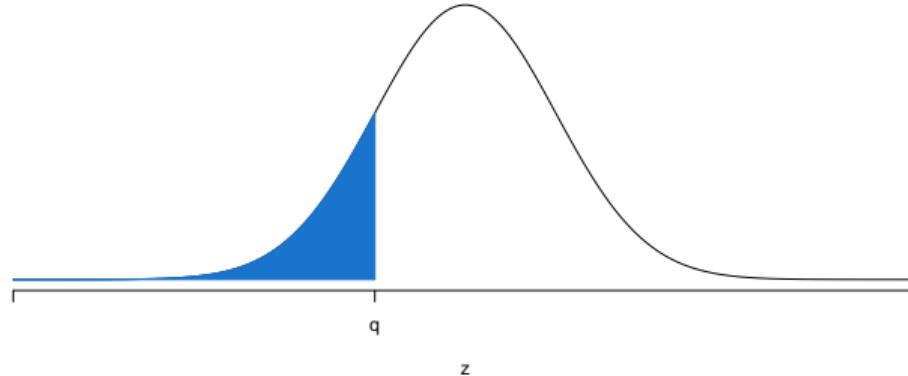
z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-3.6	.0002	.0002	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0001
-3.5	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0002
-3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
-3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
-3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
-3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
-3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
-2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
-2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
-2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
-2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
-2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
-2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
-2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
-2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
-2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
-2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
-1.9	.0279	.0274	.0269	.0263	.0258	.0252	.0244	.0239	.0233	.0228

Finding probabilities: computing age

- We can find them using a graphical calculator or computer
- We will use R
- R has four functions for the normal distribution
 - ▶ `dnorm`: density function
 - ▶ `pnorm`: probability function
 - ▶ `qnorm`: quantile function
 - ▶ `rnorm`: generate random values
- In STAT 110, most our interest is in `pnorm` and `qnorm`
 - ▶ Look at each in turn

Probability function

- This is best seen graphically
- The blue area is given by `pnorm(q)`
 - ▶ $\Pr(Z < q)$



- Look at three examples

Example 1

- What is the probability that IQ is less than 85?
- Find z -score:

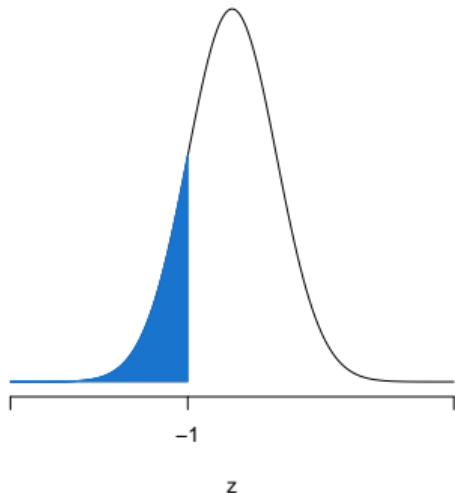
$$z = \frac{y - \mu}{\sigma} = \frac{85 - 100}{15} = -1$$

- Find $\Pr(Z < -1)$

```
mu = 100; sigma = 15 # the mean and sd for IQ
z = (85 - mu)/sigma # finding the z-score
pnorm(z)

## [1] 0.159

pnorm(-1) # for those who want to check
## [1] 0.159
```



Example 2

- Probability that IQ is more than 120?

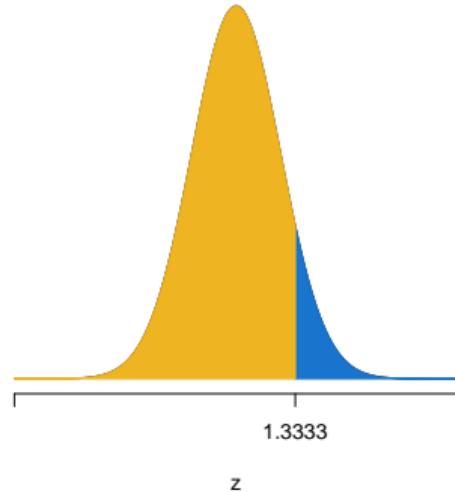
$$z = \frac{y - \mu}{\sigma} = \frac{120 - 100}{15} = 1.3333$$

- Use `pnorm` to find $\Pr(Z < 1.3333)$ (gold area)

```
z = (120 - mu)/sigma # finding the z-score  
pnorm(z)  
## [1] 0.909
```

- $\Pr(Z > 1.3333)$ (blue area) is the complement
 - $\Pr(Z > z) = 1 - \Pr(Z < z)$

```
1-pnorm(z)  
## [1] 0.0912
```



Example 3

- Probability that IQ is between 110 and 130?

$$z_{110} = \frac{y - \mu}{\sigma} = \frac{110 - 100}{15} = 0.6667$$

$$z_{130} = \frac{y - \mu}{\sigma} = \frac{130 - 100}{15} = 2$$

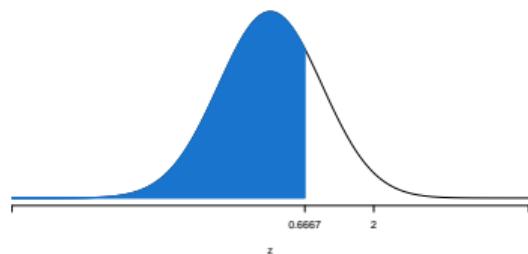
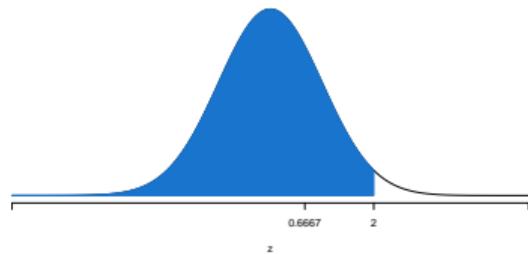
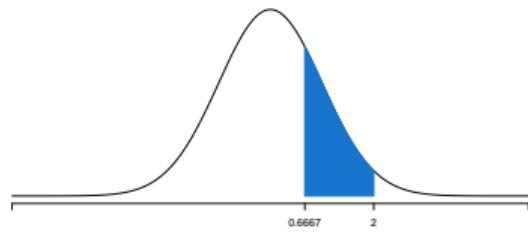
```
z110 = (110 - mu)/sigma # finding the z-score
```

```
z130 = (130 - mu)/sigma
```

- $\Pr(z_{110} < Z < z_{130}) = \Pr(Z < z_{130}) - \Pr(Z < z_{110})$
 - Best seen graphically on RHS

```
pnorm(z130)-pnorm(z110)
```

```
## [1] 0.23
```

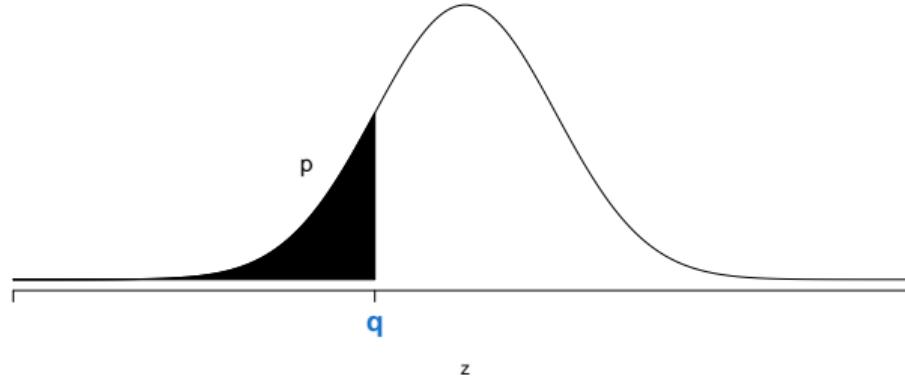


Important properties

- We can use this to learn some important characteristics of a normal distribution
- $\Pr(-1 < Z < 1) = 0.683$
 - ▶ Approximately 68% of values should be within 1 sd of the mean
- $\Pr(-2 < Z < 2) = 0.955$
 - ▶ Approximately 95% of values should be within 2 sd of the mean
- $\Pr(-3 < Z < 3) = 0.997$
 - ▶ More than 99% of values should be within 3 sd of the mean
- Challenge: confirm these numbers using `pnorm` in R before next class

Quantile function

- Basically the same graphic as before: interest is switched
- The value q is given by `qnorm(p)`
 - ▶ The value of p is the black area (known)



- Look at an example

Example

- What score is required for Mensa membership
 - ▶ At or above the 98th percentile
 - In the top 2%
- Find the z -score corresponding to $p = 0.98$

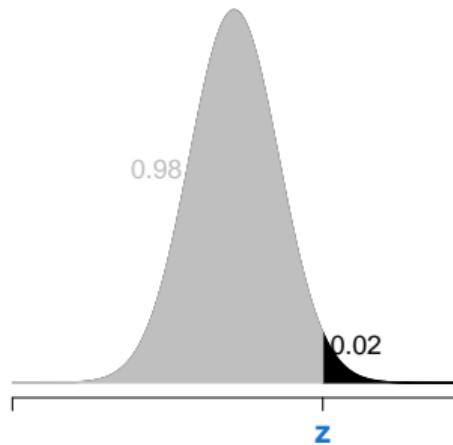
```
z = qnorm(0.98)
```

- Find the y -value

$$y = \mu + z\sigma$$

```
mu + z * sigma  
## [1] 131
```

- Need an IQ score of 131 or higher



z or y ?

- Throughout we have done calculations using standard normal
 - ▶ Standardized to find z
- With R it is comparatively easy to find using y
 - ▶ `pnorm` has optional arguments for the mean and `sd`
- First example: $\Pr(IQ < 85)$

```
pnorm(q = 85, mean = 100, sd = 15)  
## [1] 0.159
```

- Rstudio guides you as to the arguments (in R)
- Important to know about z / standardization
 - ▶ Required knowledge in the scientific world
 - ▶ Need it to understand how confidence interval and t-tests work

Summary

- Looked in some detail at normal distribution
 - ▶ Standardization and z -scores
 - ▶ Finding probabilities from z -scores
 - ▶ Finding z -scores from probabilities
- Next class: sampling distributions
 - ▶ If we took another sample, how much variation would we expect in the sample mean \bar{y} ?

Outline

- Previous:
 - ▶ Introduction to statistical modelling
 - ▶ Looked into the normal distribution
- Today:
 - ▶ Look at sampling distribution
 - ▶ Explore: how precise is the estimate \bar{y} ?

Example

- Previously we have been exploring flipper length of gentoo penguins
- Today we will use a different example
- Data from urine tests of $n = 314$ children (aged 0 – 17 years)
 - ▶ (log) GAG concentration²
 - ▶ GAG: glycosaminoglycan
 - Test is used to diagnose disorders of glycosaminoglycan metabolism
 - Glycosaminoglycans are important in cell signalling
- Data were collected to help paediatricians assess normal level of GAG concentration
- Today we'll consider a simpler problem
 - ▶ What is the expected (or mean) GAG concentration?

²We will refer to this as the concentration from here on

Data

- The data are in `lect4GAG.csv`
- Import the data into R ³

```
lect4GAG = read.csv('lect4GAG.csv')
```

- The function `head` shows us the first few lines of data

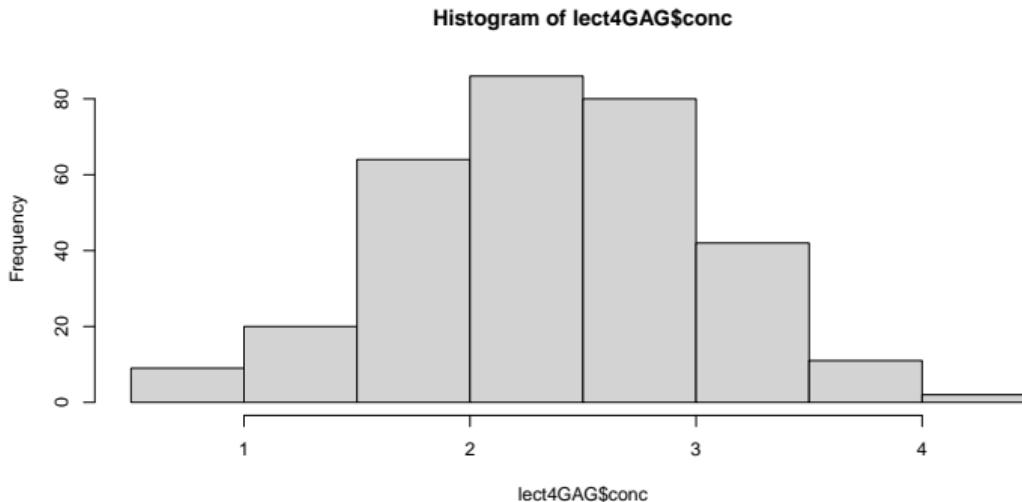
```
head(lect4GAG)  
  
##      age  conc  
## 1 0.00 3.14  
## 2 0.00 3.17  
## 3 0.00 2.83  
## 4 0.00 2.92  
## 5 0.01 2.88  
## 6 0.01 3.25
```

³Recall there are several ways to do this: see week 1 of lectures

Data

- Look at a histogram

```
hist(lect4GAG$conc) # dollar sign: selects the appropriate variable (conc)
```



- We can 'adapt' this plot to change axes labels, title, etc.
 - ▶ Keep it simple, getting an idea of the data

Recap: normal model

- We model the data as from a normal distribution
 - ▶ Modelling GAG concentration as being normally distributed
- Two parameters μ and σ
- Parameters are unknown
 - ▶ μ : mean GAG concentration
 - ▶ σ : standard deviation of GAG concentrations
- Return to our question: what is the expected (or mean) GAG concentration?
 - ▶ Estimate μ with sample mean
 - ▶ $\hat{\mu} = \bar{y}$

```
ybar_conc = mean(lect4GAG$conc)  
ybar_conc  
## [1] 2.36
```

Critical thinking

- Do we now know the expected GAG concentration?
 - ▶ That we could use (if we were a paediatrician) seeing patients

Critical thinking

- Do we now know the expected GAG concentration?
 - ▶ That we could use (if we were a paediatrician) seeing patients
- No, we don't
 - ▶ Mean GAG concentration is a parameter μ
 - ▶ Estimated it with a statistic: sample mean, \bar{y}
- How precise is the estimate?
 - ▶ If we took another sample of 314 children, how much would the estimate change?
 - ▶ Would you 'trust' the estimate more, less, or the same, if:
 - The estimate was from a sample of 8 children?
 - The estimate was from a sample of 50 000 children?

Thought experiment

- How close to μ is \bar{y} ?

Thought experiment

- How close to μ is \bar{y} ?
- To answer it, let's play god:
 - ▶ Assume that GAG concentration really is normal
 - ▶ Pretend that we know μ and σ
 - $\mu = 2.4$
 - $\sigma = 0.75$
- Take a sample of size $n = 314$ from the population
 - ▶ Observe how close the sample mean \bar{y} is to μ
- Take many (separate) samples of size n
 - ▶ See how much \bar{y} varies from one sample to another

Let's try it

- We saw a function previously for simulating from a normal distribution

```
rnorm(n,mean,sd)
```

- Generates a sample of size `n` from a normal distribution with mean (`mean`) and std deviation (`sd`)

```
n = 314; mu = 2.45; sigma = 0.75
y = rnorm(n = n, mean = mu, sd = sigma)
mean(y)
## [1] 2.52
```

- True mean: $\mu = 2.45$; sample mean: $\bar{y} = 2.52$

What if we took a lot of samples?

- Repeat this m times (using R)
 - ▶ You will not be expected to replicate the R code below

```
m = 10000 # the number of samples
ybar = rep(NA, m) # this 'initializes' a vector to store each
# of the m sample means
for(i in 1:m){ # repeats the code below m times
  y = rnorm(n, mu, sigma) # takes a sample of size n = 314
  ybar[i] = mean(y) # finds the sample mean and stores it in ybar
}
```

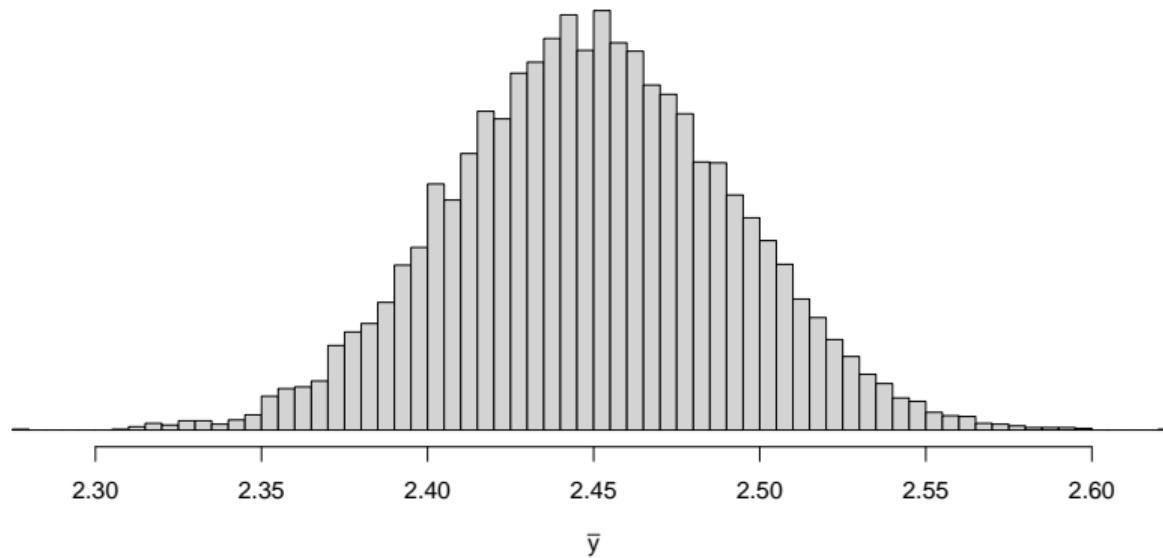
What if we took a lot of samples?

- The first few sample means are:

```
head(ybar)  
## [1] 2.48 2.46 2.45 2.46 2.44 2.50
```

- We could look at a histogram of these
 - ▶ Get an idea of the distribution of sample means
 - ▶ Evaluate how variable \bar{y} is: one sample to another
 - ▶ Assess whether \bar{y} accurately estimates the mean (on average)

What if we took a lot of samples?



Sampling distribution

- This is called the sampling distribution
 - ▶ Sampling distribution of \bar{y}
- Tells us how we would expect our statistic (\bar{y}) to vary from one sample to another
- From the histogram we can see
 - ▶ On average it is 2.45: the value of μ
 - ▶ Sample means less than 2.35 or larger than 2.55 are unlikely

What if?

- We can use this to answer ‘what if’ questions, e.g.
- What is the chance of observing a sample mean as extreme as $\bar{y} = 2.36$
 - ▶ If the $\mu = 2.45$ and $\sigma = 0.75$?
- Look at the histogram again:
 - ▶ Possible, but unlikely
- Could use R to count how many samples (of 10 000) had mean less than 2.36
 - ▶ Estimate the probability
 - ▶ R shown for interest only

```
sum(ybar < ybar_conc) # ybar_conc = 2.36 (from data)
## [1] 218
```

What is extreme?

- We asked 'what is the chance of observing a sample mean as extreme as ...'
 - ▶ Did we answer that correctly?

What is extreme?

- We asked ‘what is the chance of observing a sample mean as extreme as ...’
 - ▶ Did we answer that correctly?
- No: we looked at chance of observing a sample mean less than 2.36
 - ▶ A sample mean higher than 2.54 is just as extreme as one below 2.36
 - ▶ Both are 0.09 units away from the true mean ($\mu = 2.45$)
- An extreme observation could be below or above the mean
 - ▶ Calculating the probability of an extreme value needs to account for both
- This is a principle we will use often

Theory

- It turns out that when we have a normal model for y
 - ▶ The sampling distribution (distribution of sample means \bar{y}) is also normally distributed
- What are the mean and variance?
 - ▶ The mean of the sampling distribution is μ
 - ▶ The variance of the sampling distribution is $\frac{\sigma^2}{n}$
 - ▶ The standard deviation of the sampling distribution is $\frac{\sigma}{\sqrt{n}}$

Theory

- Where do these results come from?
 - ▶ We worked these out a few lectures ago! (lecture 8; copied below)
 - The expected value of the sample mean is

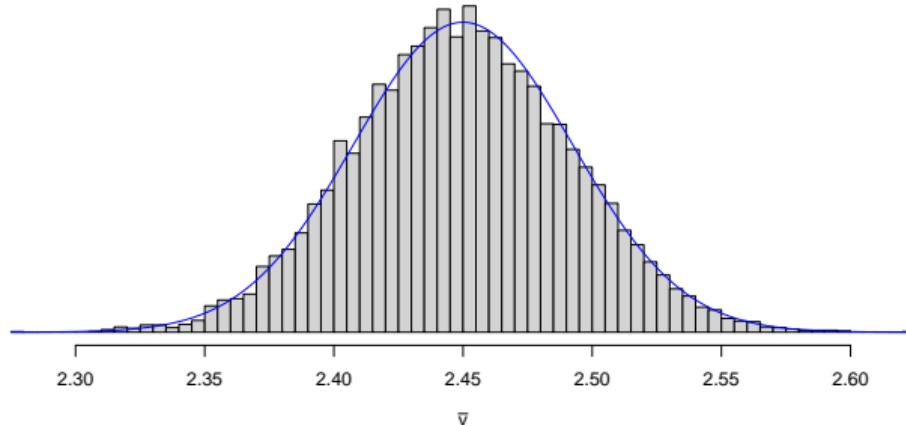
$$\begin{aligned} E\left[\frac{Y_1 + Y_2 + \dots + Y_n}{n}\right] &= \frac{1}{n}E[Y_1] + \frac{1}{n}E[Y_2] + \dots + \frac{1}{n}E[Y_n] \\ &= \mu \end{aligned}$$

- The variance of the sample mean is

$$\begin{aligned} Var\left(\frac{Y_1 + Y_2 + \dots + Y_n}{n}\right) &= \frac{1}{n^2}Var(Y_1) + \frac{1}{n^2}Var(Y_2) + \dots + \frac{1}{n^2}Var(Y_n) \\ &= \frac{\sigma^2}{n} \end{aligned}$$

Sampling distribution

- When using a normal model for y , the sampling distribution for \bar{y}
 - ▶ Normally distributed with mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$
- For the example above, the sampling distribution has:
 - ▶ Mean: 2.45, standard deviation $\frac{0.75}{\sqrt{314}}$
- Compare to the sampling distribution found in R



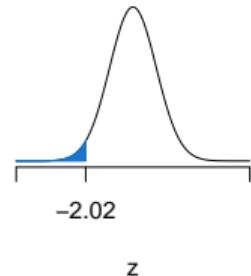
Sampling distribution

- Use our knowledge of the normal distribution to earlier questions
- What is the chance of observing a sample mean as extreme as $\bar{y} = 2.36$?
 - ▶ If the $\mu = 2.45$ and $\sigma = 0.75$?
- Three steps
 1. Find mean and sd of sampling distribution
 2. Convert to z-value
 3. Find the probability

Sampling distribution

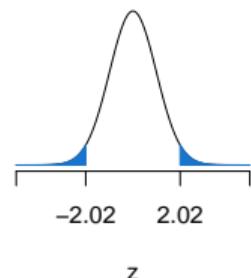
- Mean: $\mu = 2.45$
- Standard deviation: $\frac{\sigma}{\sqrt{n}} = \frac{0.75}{\sqrt{314}}$
- z-value: $z = \frac{\bar{y} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{2.36 - 2.45}{\frac{0.75}{\sqrt{314}}} = -2.021$
- Probability of z-value less than -2.021

```
z = (ybar_conc - mu) / (sigma / sqrt(n)) # z-value  
pnorm(z)  
## [1] 0.0216
```



- Probability of z-value more extreme than -2.021

```
2*pnorm(z) # same area in each tail (see graphic)  
## [1] 0.0432
```



Does this make sense?

- The standard deviation of the sampling distribution $\frac{\sigma}{\sqrt{n}}$
 - ▶ Decreases as n increases
- Makes sense
 - ▶ As the sample size (n) increases, the estimate \bar{y} is increasingly precise
- If n is small ($n = 1$)
 - ▶ Sample mean is the same as an observation: same sd (σ)
- If n is large ($n = 1\,000\,000$)
 - ▶ Standard deviation of the sample mean is 1/1000th the sd of observations
 - ▶ Lots of data: sample mean is a precise estimate of true mean

Summary

- Introduced the concept of sampling distribution
 - ▶ Tells us how much \bar{y} varies from one sample to the next
- Introduced some core principles that we will see again and again
- Standard deviation of sampling distribution is $\frac{\sigma}{\sqrt{n}}$
 - ▶ Use this to evaluate how precise an estimate is
 - ▶ Problem: relies on σ being known
 - ▶ What happens if σ is unknown
 - Always the case in the real world
 - ▶ Explore in the next lecture

Outline

- Previous:
 - ▶ Introduction to (normal) statistical model
 - ▶ Sampling distributions
 - Describe variation in the sample mean \bar{y} (or any other statistic) from one sample to another
 - Relies on us knowing σ
- Today:
 - ▶ Use that to find confidence interval
 - Interval estimate for the parameter value
 - ▶ Look at what happens when σ is unknown

Example

- Continue using the GAG concentration data
 - ▶ Data from urine tests of $n = 314$ children (aged 0 – 17 years)
 - ▶ (log) concentration of glycosaminoglycan (GAG)
- Asking: what is the expected (or mean) GAG concentration?

Sampling distribution

- Recall we have a normal model for the data
 - ▶ Data come from a normal distribution with mean μ and standard deviation σ
- Last lecture we found the sampling distribution for \bar{y}
 - ▶ Distribution that describes how \bar{y} will vary from one sample to another
 - ▶ Sampling distribution is normally distributed (for a normal model)
 - Mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$

Cool result!

- We know about what will happen in repeated samples
 - ▶ Without having to take repeated samples!
- If we know the data distribution (i.e. we know μ and σ):
 - ▶ We know how variable we expect \bar{y} to be without even sampling from the population
- If we know σ (but don't know μ):
 - ▶ Can we use a single sample to tell us about a range of plausible values of μ ?

Cool result!

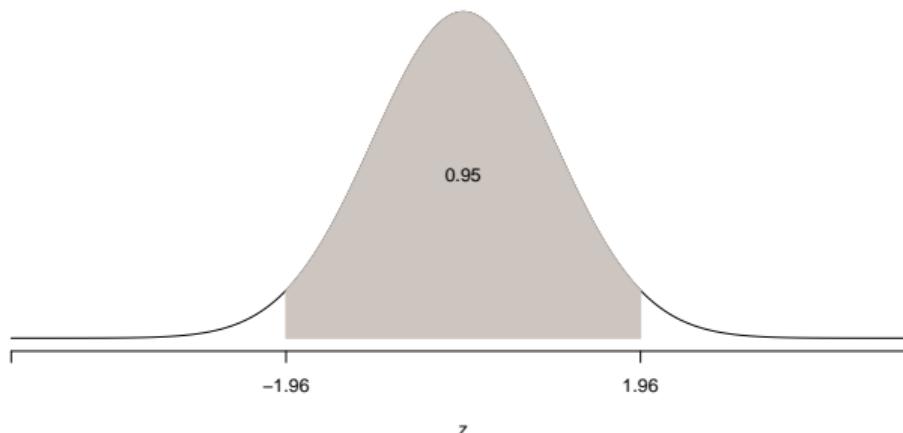
- We know about what will happen in repeated samples
 - ▶ Without having to take repeated samples!
- If we know the data distribution (i.e. we know μ and σ):
 - ▶ We know how variable we expect \bar{y} to be without even sampling from the population
- If we know σ (but don't know μ):
 - ▶ Can we use a single sample to tell us about a range of plausible values of μ ?
- Yes!

Excursion: standard error

- Confusing notation to discuss
- Over the past few lectures, we have seen:
 - ▶ Population standard deviation σ
 - ▶ Sample standard deviation s
 - ▶ Standard deviation of sampling distribution of \bar{y}
 - It is $\frac{\sigma}{\sqrt{n}}$
 - Has a special name: standard error
 - Can be represented with notation $\sigma_{\bar{y}}$
 - ▶ Estimate of the standard deviation of the sampling distribution of \bar{y}
 - It is $\frac{s}{\sqrt{n}}$
 - It is often also called the standard error
 - Can be represented with notation $s_{\bar{y}}$
- Very confusing

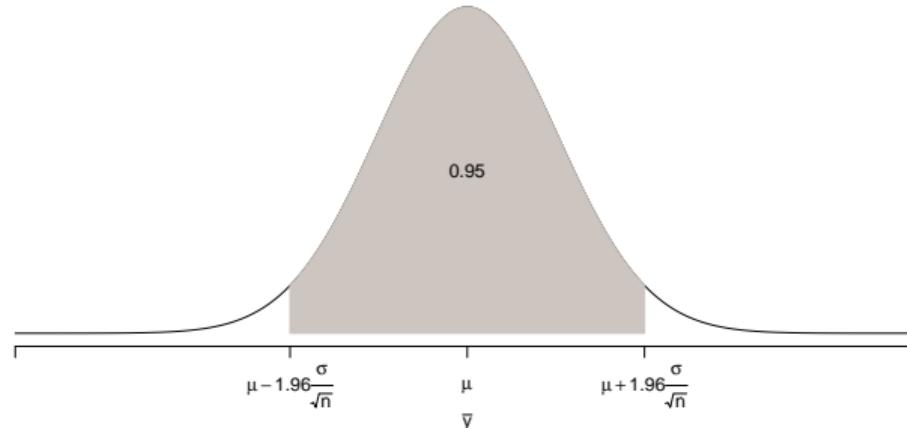
Previous knowledge

- Want to determine an interval estimate of μ from \bar{y}
- From our knowledge of normal distribution:
 - ▶ 95% of observations will fall within (approx) ± 2 standard deviations of mean
 - More precisely it is ± 1.96
 - In R: `qnorm(0.025)` and `qnorm(0.975)`
 - ▶ $\Pr(-1.96 < Z < 1.96) = 0.95$



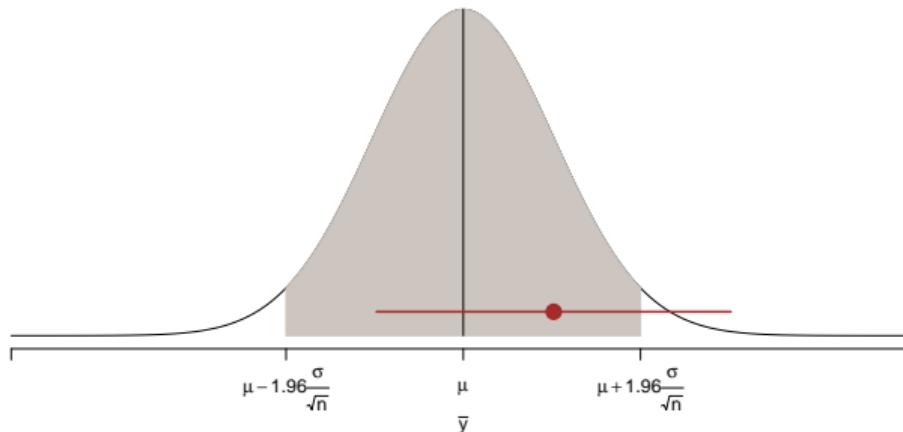
Sampling distribution

- Applying this to the sampling distribution we have:
 - ▶ 95% of sample means (\bar{y}) are between ± 1.96 standard errors ($\frac{\sigma}{\sqrt{n}}$) of the mean
- 95% of samples we collect will have sample means in the grey area
 - ▶ Given by $\mu \pm 1.96 \frac{\sigma}{\sqrt{n}}$



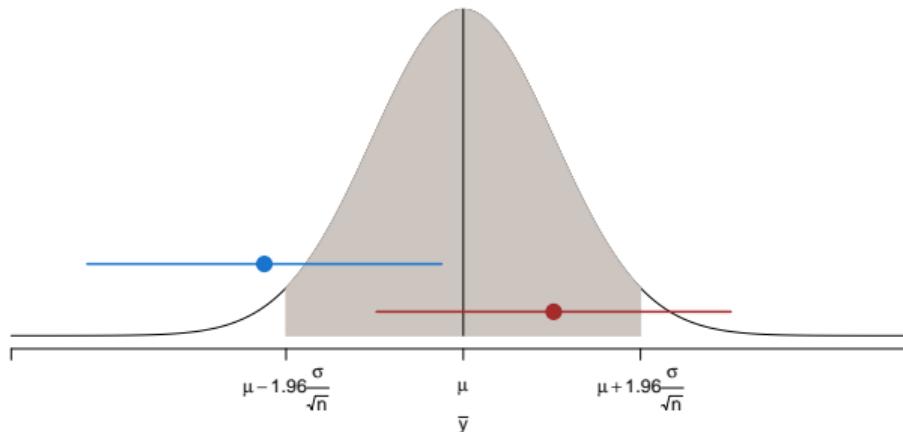
Flipping things I

- Consider any sample mean that is **inside** the shaded grey area
 - ▶ We've plotted one in brown on plot below
- Here's the magic:
 - ▶ If \bar{y} is inside the grey area ($\mu \pm 1.96 \frac{\sigma}{\sqrt{n}}$) (brown point)
 - ▶ Then μ (vertical black line) is inside the interval $\bar{y} \pm 1.96 \frac{\sigma}{\sqrt{n}}$ (brown interval)



Flipping things II

- Consider any sample mean that is **outside** the shaded grey area
 - ▶ We've plotted one in blue on plot below
- Here's the magic:
 - ▶ If \bar{y} is outside the grey area $(\mu \pm 1.96 \frac{\sigma}{\sqrt{n}})$ (blue point)
 - ▶ Then μ (vertical black line) is outside the interval $\bar{y} \pm 1.96 \frac{\sigma}{\sqrt{n}}$ (blue interval)



Confidence interval

$$\bar{y} \pm 1.96 \frac{\sigma}{\sqrt{n}}$$

- This is a 95% confidence interval for μ
 - ▶ Interval estimate of μ
 - ▶ Quantifies how precise the estimate of μ is
- On average, 95% of sample means will lie in shaded grey area (established above)
 - ▶ That means that our confidence interval should contain the true μ in 95% of samples
 - ▶ Gives us confidence in the procedure (hence the name)
 - Care is needed: we cannot say that there is a probability of 0.95 that μ is in the interval

A few notes on confidence intervals

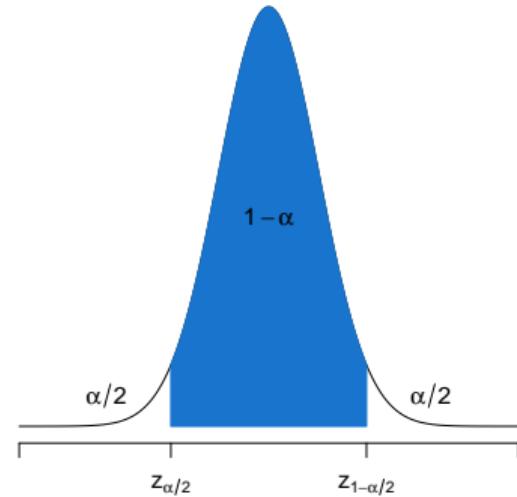
- The confidence interval is in a general form:

$$\text{estimate} \pm \text{multiplier} \times \text{standard error}$$

- estimate: \bar{y}
- multiplier:
 - 1.96 for 95% confidence interval
 - More generally, we write $z_{1-\alpha/2}$
 - More details on next slide
- Standard error: $\frac{\sigma}{\sqrt{n}}$

Multiplier

- Multiplier: $z_{1-\alpha/2}$
 - ▶ Also referred to as the critical value
- α : significance level
 - ▶ significance level = $1 - \text{confidence level}$
 - 95% interval: $\alpha = 1 - 0.95 = 0.05$
 - 90% interval: what is α ?
- $\Pr(Z < z_{1-\alpha/2}) = 1 - \alpha/2$
 - ▶ Find z-value so that tails have probability $\alpha/2$



Multiplier

- For a 95% interval
 - ▶ $\alpha = 0.05$
 - ▶ $1 - \alpha/2 = 0.975$
 - ▶ We want to find $z_{0.975}$

```
qnorm(0.975)
```

```
## [1] 1.96
```

- How do we find the multiplier for a 90% interval?

Multiplier

- For a 95% interval
 - ▶ $\alpha = 0.05$
 - ▶ $1 - \alpha/2 = 0.975$
 - ▶ We want to find $z_{0.975}$

```
qnorm(0.975)
```

```
## [1] 1.96
```

- How do we find the multiplier for a 90% interval?
 - ▶ $\alpha = 0.10$
 - ▶ $1 - \alpha/2 = 0.95$
 - ▶ We want to find $z_{0.95}$

```
qnorm(0.95)
```

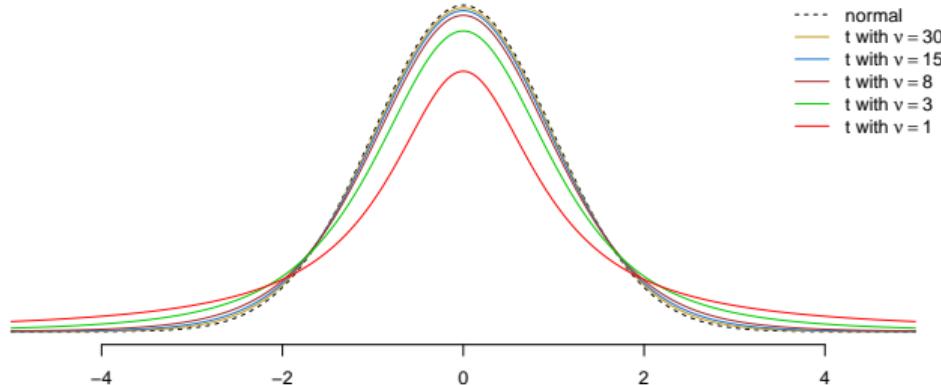
```
## [1] 1.64
```

GAG concentrations

- Let's find an interval estimate for mean GAG concentration!
- We can't... we don't know σ
 - ▶ Population standard deviation
- Can we just replace σ with s ?
 - ▶ No, the sampling distribution is no longer normal
 - All is not lost: most of the reasoning we worked through remains the same
- Replacing σ by s introduces additional noise (variability)
 - ▶ Sampling distribution no longer normally distributed
 - ▶ We need to use a t-distribution instead

t -distribution

- A t -distribution looks a lot like a (standard) normal distribution
 - ▶ Has fatter tails
- Additional parameter $\nu > 0$, called the degrees of freedom
 - ▶ This defines how fat the tails are



Historical excursion: William Gosset (1876 – 1937)

- Head Brewer of Guinness who ‘discovered’ the t -distribution
- Running experiments on yield of barley varieties and did not have statistical tools he needed to analyze the data
 - ▶ Statistical methodology developed due to applications in food science, agriculture
- The t -distribution is commonly known as Student’s t -distribution
 - ▶ Gosset published under the pseudonym ‘Student’
 - ▶ Guinness allowed its scientists to publish research if they did not mention:
 - Beer
 - Guinness
 - Their own surname

Confidence interval: unknown σ

- Replacing σ by s leads to the confidence interval

$$\bar{y} \pm t_{\nu, 1-\alpha/2} \frac{s}{\sqrt{n}}$$

- $t_{\nu, 1-\alpha/2}$: multiplier for the t -distribution
 - Significance level α
 - Degrees of freedom ν
- When finding confidence interval for μ
 - Degrees of freedom $\nu = n - 1$
- Find multiplier in R: for 95% interval when $n = 30$

```
n = 30  
qt(0.975, df = n-1)  
## [1] 2.05
```

GAG concentrations

- We are now ready to find an interval estimate for mean GAG concentration
- We need to get a few bits and pieces together:
 - ▶ Call in the data:

```
lect4GAG = read.csv('lect4GAG.csv')
```

- ▶ Find the sample mean: \bar{y}

```
ybar = mean(lect4GAG$conc)  
ybar  
## [1] 2.36
```

- ▶ Find the sample standard deviation: s

```
s = sd(lect4GAG$conc)  
s  
## [1] 0.668
```

GAG concentrations

- ▶ Find the sample size: n

```
n = length(lect4GAG$conc) # length() tells us the number of values  
n  
## [1] 314
```

- ▶ Find the standard error: $s_{\bar{y}} = \frac{s}{\sqrt{n}}$

```
se = s/sqrt(n)  
se  
## [1] 0.0377
```

- ▶ Find the multiplier: 95% confidence interval

```
alpha = 0.05  
tcrit = qt(1-alpha/2, df = n-1)  
tcrit  
## [1] 1.97
```

GAG concentrations

- ▶ Put it all together

```
lower = ybar - tcrit * se # lower confidence limit  
upper = ybar + tcrit * se # upper confidence limit  
ci = c(lower, upper)  
ci  
## [1] 2.29 2.44
```

- ▶ The 95% confidence interval for μ is (2.29, 2.44)
 - Interval estimate for μ
- Spent some time interpreting the interval in the next lecture

Summary

- Found confidence interval for μ
 - ▶ Interval that quantifies how precise our estimate of μ is
- Found confidence interval if σ is known
 - ▶ Useful for understanding
 - ▶ Not practically useful
- Found confidence interval if σ is unknown
 - ▶ Introduced the t -distribution
- Looking forward:
 - ▶ More about confidence intervals

STAT 110: Week 5

University of Otago

Outline

- Previous lecture:
 - ▶ Confidence interval for population mean μ
- Today: understand more about the confidence interval
 - ▶ How to find the confidence interval
 - ▶ How to interpret the confidence interval
 - ▶ Understanding the properties of the confidence interval
 - ▶ How large of a sample do we need?

Data: GAG concentration

- Call in the data

```
lect4GAG = read.csv('lect4GAG.csv') # it doesn't matter that we are now in week 5
```

- Remember what the data set looks like:

```
head(lect4GAG)

##      age  conc
## 1 0.00 3.14
## 2 0.00 3.17
## 3 0.00 2.83
## 4 0.00 2.92
## 5 0.01 2.88
## 6 0.01 3.25
```

Recall: GAG concentration

- Data from urine tests of $n = 314$ children (aged 0 – 17 years)
 - ▶ Interest in estimating the mean (log) concentration of glycosaminoglycan (GAG)
- In the last lecture we found a confidence interval
 - ▶ Quite an involved process
- Several steps
 1. Call the data into R
 2. Find the sample mean: \bar{y}
 3. Find the sample standard deviation: s
 4. Find the sample size: n
 5. Find the standard error: $s_{\bar{y}} = \frac{s}{\sqrt{n}}$
 6. Find the multiplier: $t_{\nu,1-\alpha/2}$
 7. Find the confidence interval: $\bar{y} \pm t_{\nu,1-\alpha/2} \frac{s}{\sqrt{n}}$

That's a lot of steps!

- That's not how we find a confidence interval in practice
 - ▶ R function that finds it for us: `t.test`
- So why did we go through those steps?
 - ▶ Important for our understanding of what a confidence interval is
 - We will be exploring 'properties' of confidence intervals that use this information
 - ▶ To use any tool well, it helps to know how it works
 - What its limitations are

Finding confidence interval: in practice

- We can find a confidence interval for μ with `t.test`

```
output = t.test(lect4GAG$conc)

output

##

##  One Sample t-test

##

## data: lect4GAG$conc

## t = 63, df = 313, p-value <2e-16

## alternative hypothesis: true mean is not equal to 0

## 95 percent confidence interval:

##  2.29 2.44

## sample estimates:

## mean of x

##      2.36
```

Output of t.test

- We can understand some of the output
 - ▶ df = degrees of freedom for the multiplier
 - ▶ sample mean
 - ▶ 95% confidence interval
 - ▶ We will be learning about the other things soon
- We can isolate the confidence interval

```
output$conf.int  
## [1] 2.29 2.44  
## attr(,"conf.level")  
## [1] 0.95
```

Using `t.test`

- The input to `t.test` is the full data set
 - ▶ No need to summarize data in terms of \bar{y} and s
 - ▶ No need to find the multiplier

Changing the confidence level

- The function `t.test` has optional arguments
 - ▶ These are arguments that have some default, but we can choose to change them
 - ▶ One of these is `conf.level`
 - Defaults to 0.95 (95% confidence interval)
- For a 90% confidence interval:

```
output90 = t.test(lect4GAG$conc, conf.level = 0.9)
output90$conf.int
## [1] 2.30 2.43
## attr("conf.level")
## [1] 0.9
```

- How would we find a 99% interval?

Diversion: R help

- How would you figure out that `conf.level` changes the confidence level?
- Many answers:
 - ▶ In this course: we will show you how to make changes like this
 - ▶ Outside this course: you can consult the R help
 - Surprisingly, not really the recommended first option
 - ▶ This is where chatGPT (or equivalent) can be really helpful
 - e.g. ask “how do I find a 90% confidence interval when using `t.test` in R?”
 - Not always 100% accurate, but it is pretty good
 - ▶ Google can also be very helpful

Interpreting the confidence interval

- What do we do with the confidence interval: (2.29, 2.44)?
 - ▶ We are 95% confident that mean GAG concentration is between 2.29 and 2.44
- What does 95% confident mean?
 - ▶ Recall the definition of a confidence interval
 - ▶ It does not guarantee that the true mean GAG concentration is inside the interval
 - Across many samples, the true mean should be in the interval 95% of the time
 - ▶ Confidence in the procedure: long-term performance

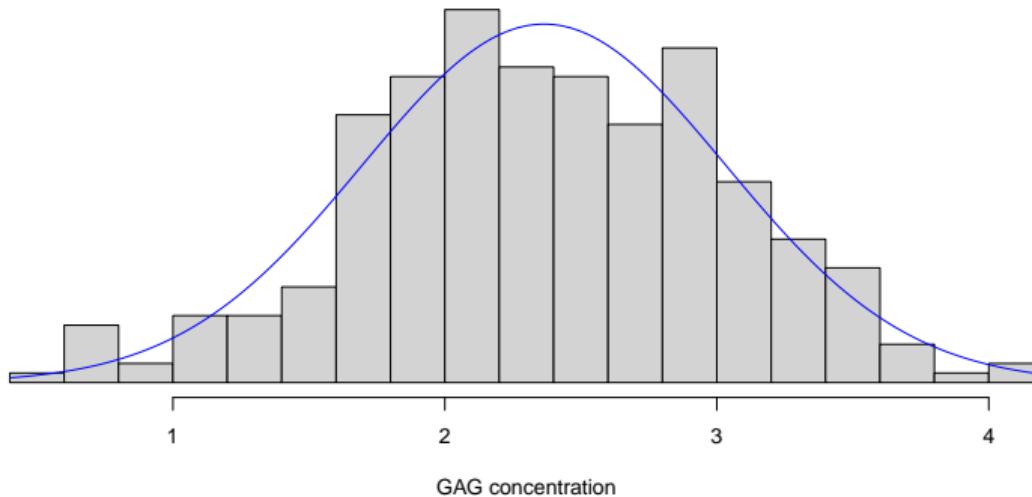
Interpreting the confidence interval

- What do we do with the confidence interval: (2.29, 2.44)?
 - ▶ We are 95% confident that mean GAG concentration is between 2.29 and 2.44
- This is a statement about the population parameter
 - ▶ Average GAG concentration for children aged 0-17
 - ▶ Population isn't well defined
 - Geographical area?
 - It isn't clear how the data were collected
 - Important factors in determining whether the confidence interval tells us anything useful
 - We will be talking more later in the course about the importance of data collection

Checking model assumptions

- Recall: it is important to check model assumptions
- We have assumed the data came from a normal distribution
- STAT 110 approach: check visually
 - ▶ Histogram
 - ▶ Looking for major departures from normality
 - Obvious skew
 - Large outliers
- If the sample size is large enough
 - ▶ Confidence intervals for μ are suitable for non-normal data
 - ▶ $n > 30$ is rule of thumb often used
 - If there are major departures from normality, we may need a much larger n
 - ▶ Discuss more in a few weeks

Model fit: GAG



- No obvious departures from normality
 - ▶ Blue curve: normal density using the sample mean and sd

Width of the confidence interval

- The width of the confidence interval is important
 - ▶ Tells us how precise the estimate is
- The CI we found is (2.29, 2.44)
 - ▶ An example of a wider (less precise) interval: (2.22, 2.51)
 - ▶ An example of a narrower (more precise) interval: (2.34, 2.39)
- The width of a confidence interval is given by upper limit - lower limit
 - ▶ Width: $2.44 - 2.29 = 0.15$
- We often refer to the margin of error: half of the interval width
 - ▶ Recall our confidence interval formula:

$$\bar{y} \pm t_{\nu, 1-\alpha/2} \underbrace{\frac{s}{\sqrt{n}}}_{\text{margin of error}}$$

Changing confidence level

- What happens to interval width if we increase the confidence level, say from 95% to 99%? Why?

Changing confidence level

- What happens to interval width if we increase the confidence level, say from 95% to 99%? Why?
 - ▶ The interval gets wider (margin of error gets larger)
 - Confidence level increases, α decreases
 - Multiplier $t_{\nu,1-\alpha/2}$ increases
 - Can be seen graphically
- This makes sense:
 - ▶ Making the interval wider: increasing the confidence that parameter (μ) is in interval
 - ▶ If we have a wider interval, the true mean will be in the interval a higher percentage of the time
- The opposite also holds:
 - ▶ If we decrease the confidence level: interval gets narrower

Changing confidence level: 95%

```
output95 = t.test(lect4GAG$conc, conf.level = 0.95)
output95

##
## One Sample t-test
##
## data: lect4GAG$conc
## t = 63, df = 313, p-value <2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
## 2.29 2.44
## sample estimates:
## mean of x
## 2.36
```

Changing confidence level: 99%

```
output99 = t.test(lect4GAG$conc, conf.level = 0.99)
output99

##
## One Sample t-test
##
## data: lect4GAG$conc
## t = 63, df = 313, p-value <2e-16
## alternative hypothesis: true mean is not equal to 0
## 99 percent confidence interval:
## 2.27 2.46
## sample estimates:
## mean of x
## 2.36
```

Changing confidence level: 90%

```
output90 = t.test(lect4GAG$conc, conf.level = 0.90)
output90

##
## One Sample t-test
##
## data: lect4GAG$conc
## t = 63, df = 313, p-value <2e-16
## alternative hypothesis: true mean is not equal to 0
## 90 percent confidence interval:
## 2.30 2.43
## sample estimates:
## mean of x
## 2.36
```

Standard error

- The standard error is a critical part of the calculation of a confidence interval:

$$s_{\bar{y}} = \frac{s}{\sqrt{n}}$$

- Recall: tells us how variable the statistic \bar{y} is
 - ▶ Quantifies how much we expect \bar{y} to vary
 - If we took multiple samples of size n from the population
- It has two components

1. s : sample standard deviation

- The larger the variation in the data, the larger the standard error
- The larger the variation in the data, the wider the confidence interval for μ

2. n : sample size

- The larger the sample size, the smaller the standard error
- The larger the sample size, the narrower the confidence interval for μ

Caution

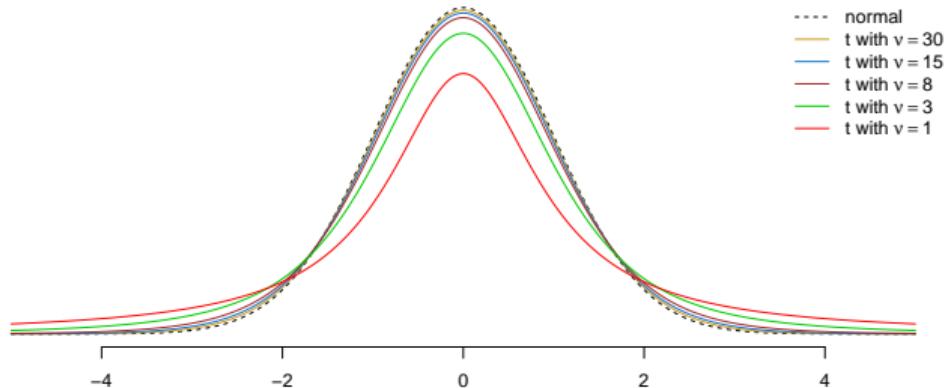
- The statements on the previous slide assume all else is held fixed
 - ▶ e.g. the larger the sample size, the narrower the confidence interval, all else held fixed
- In reality: if we took a different (larger) sample, things would not be held fixed
 - ▶ \bar{y} varies from one sample to the next
 - ▶ s also varies from one sample to the next
 - ▶ On average: \bar{y} from a larger sample will be closer to the true mean
- We cannot (and should not) use the \bar{y} and s we observe and pretend we had a larger sample size to find a narrower confidence interval
 - ▶ Fabricating (or falsifying) data
 - ▶ Unethical
 - ▶ Scientific misconduct

Sample size calculation

- The GAG data appear to be from the UK
- We may choose to replicate the study here in NZ
 - ▶ We want the study to be accurate: margin of error of 0.04
 - ▶ How large of a sample should we take?
- We want to find value n such that the margin of error is 0.04
- This is a common scenario when designing research studies
 - ▶ Too few samples: imprecise estimates of limited value
 - ▶ Too many samples: poor use of precious resources (time and money)

Sample size calculation

- This is an approximate process (we'll see why as we go)
- Recall: the margin of error is $t_{\nu,1-\alpha/2} \frac{s}{\sqrt{n}}$
 - ▶ Find n so that the margin of error has a desired level of accuracy
- This is problematic for two reasons:
 1. The multiplier $t_{\nu,1-\alpha/2}$ depends on n ($\nu = n - 1$)
 - Approximate it with $z_{1-\alpha/2}$



Sample size calculation

- We want to find n so that the margin of error has a desired level of accuracy
- This is problematic for two reasons:
 2. The standard deviation s is an estimate that will change from one sample to the next
 - Take s as our best estimate of σ
- To find n , we use an approximate margin of error $\approx z_{1-\alpha/2} \frac{s}{\sqrt{n}}$
- If the desired level of accuracy (in our case 0.04) is given by the symbol ξ , we want to find the value of n such that

$$z_{1-\alpha/2} \frac{s}{\sqrt{n}} \leq \xi$$

Sample size calculation

- We rearrange the formula to get:

$$n \geq \left(\frac{z_{1-\alpha/2} s}{\xi} \right)^2$$

- In our case

```
alpha = 0.05 # 95% confidence interval
z = qnorm(1-alpha/2) # approximate multiplier: normal distribution
s = sd(lect4GAG$conc) # best guess as to the sigma
xi = 0.04 # desired margin of error
n = ceiling((z * s / xi)^2) # sample size; ceiling rounds up
n
## [1] 1073
```

Sample size calculation

- This is an approximate process
 - ▶ Approximated the multiplier
 - ▶ Used an estimate of standard deviation
- Always ‘round up’ (R command `ceiling` rounds up)
- We tend to be conservative
 - ▶ It’s better to have a few more observations than you need, than too few.
 - Often round up further, to say $n = 1100$ or $n = 1200$ participants, or
 - ▶ In practice, we often find a confidence interval for σ
 - Use the upper limit of the CI in the calculation (in place of s)
 - Outside the scope of STAT 110

Summary

- Looked at more detail into calculation and use of confidence intervals
 - ▶ How to find them in R: `t.test`
 - Changing confidence level
 - ▶ Interpreting the confidence interval
 - ▶ Width and margin of error
 - ▶ Sample size calculation
 - ▶ Tomorrow: hypothesis testing

Outline

- Previous:
 - ▶ Learned how to find and interpret confidence intervals
 - ▶ Interval estimates of parameter
- Today:
 - ▶ Look at hypothesis testing

Hypotheses

- Data are often collected to test a hypothesis
 - ▶ e.g. sleep deprivation affects reaction time
 - ▶ e.g. survival rates of kākāpō are higher today than they were 10 years ago
 - ▶ e.g. calorie values listed on labels of chip packets are not accurate
- Collect data to investigate our hypothesis

Example 1: Shoshone Rectangles

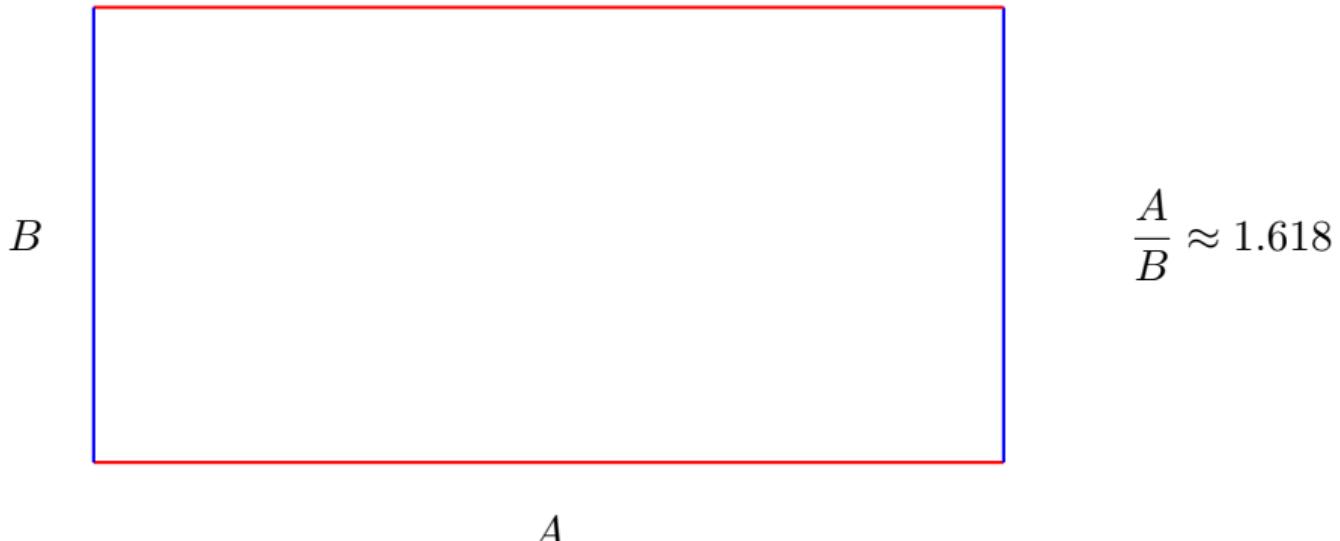
- Shoshone Native Americans used beaded rectangles to decorate their goods



- Native American tribe that originated in the western Great Basin and spread north and east into present-day Idaho and Wyoming
- Anthropologists are interested to know whether there is evidence against the claim that Shoshone Native Americans produced rectangles which conform to the golden ratio
 - ▶ What is the golden ratio?

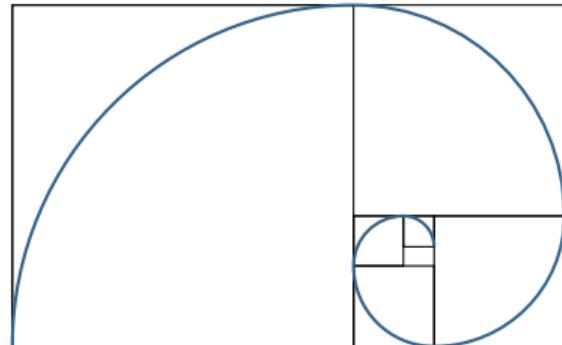
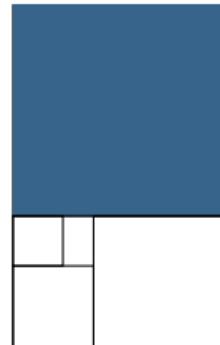
Golden ratio

- The golden ratio is a number that appears frequently in geometry
 - ▶ First studied by the Greeks
 - Euclid called it the ‘extreme and mean ratio’
- A rectangle is ‘golden’ if the ratio of its long to short side is $\frac{1+\sqrt{5}}{2} \approx 1.618$.

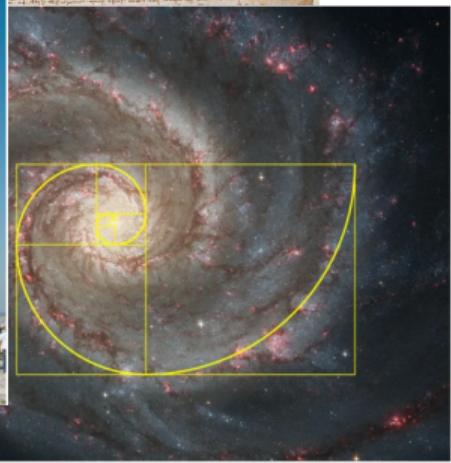
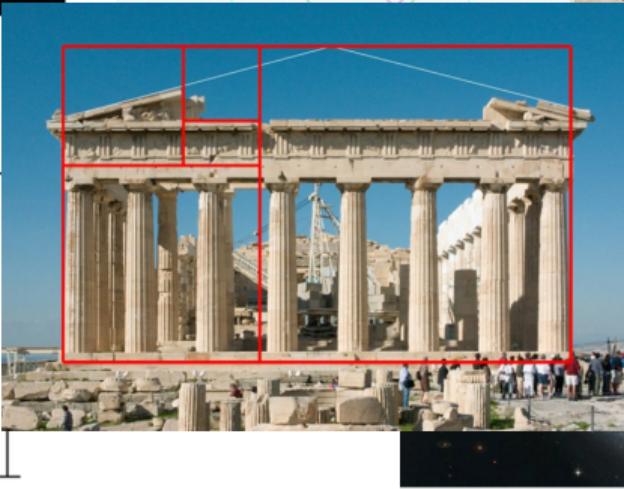
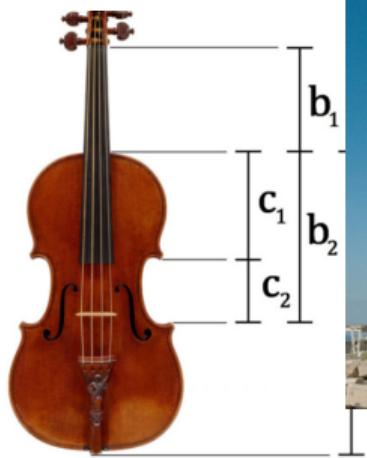
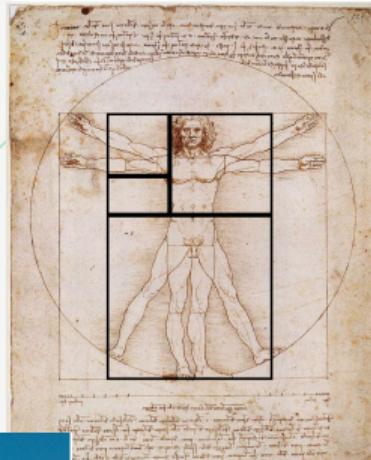
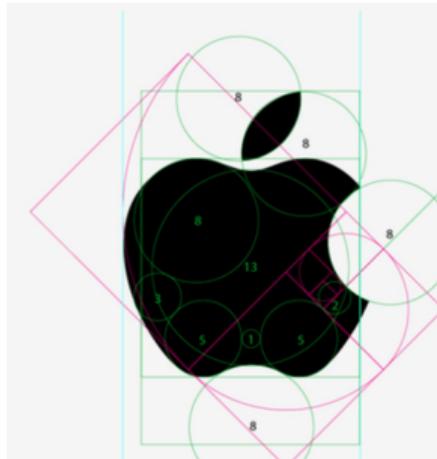
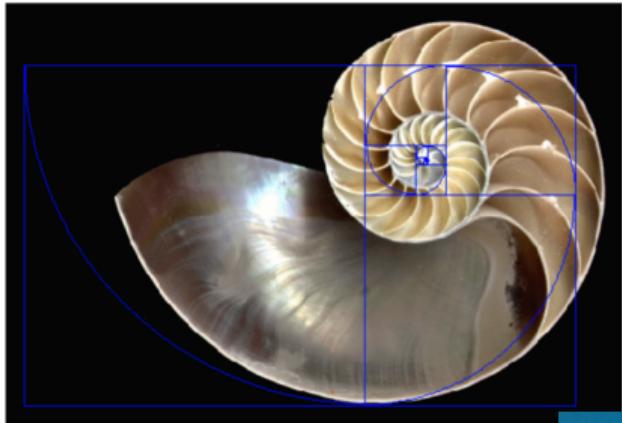


Golden ratio

- Rectangles with the golden ratio have some nice mathematical properties
 - ▶ e.g. if you take away a square (in blue) you get another golden rectangle
 - ▶ Related to a Fibonacci sequence



- The golden ratio is apparent in art, architecture and nature



Example: Shoshone rectangles

- Data of the length-to-width ratios for 20 Shoshone rectangles

```
shoshone = read.csv("shoshone.csv")
```

```
shoshone$ratio
```

```
## [1] 1.44 1.51 1.45 1.65 1.75 1.33 1.49 1.59 1.64 1.19 1.53 1.63 1.50 1.66 1.74  
## [16] 1.49 1.65 1.64 1.81 1.07
```

- How can we investigate how compatible the data are with the golden ratio?

Set up hypotheses

- Two hypotheses: null hypothesis and alternate hypothesis
 - ▶ Null: we compare the data to what we expect under the null hypothesis
 - Assess the compatibility of the data to the null hypothesis
 - Often the claim to be tested, the status quo, or assumption of no difference
 - The hypothesis we find evidence against
 - ▶ Alternate: alternate claim under consideration
 - Alternate ‘state of the world’
 - Hypothesis we want to find evidence in support of

Set up hypotheses: examples I

- Quality control: manufacturing cell phone case
 - ▶ To specifications: say mean length $\mu = 6$ inches
 - ▶ Collect data to ensure quality
 - ▶ $H_0 : \mu = 6$ (status quo)
 - ▶ $H_A : \mu \neq 6$
- Collect data from group with specific disease
 - ▶ Interested in expression of particular gene
 - ▶ Know the expression in the population is 10 TPM
 - ▶ $H_0 : \mu = 10$ (claim to be tested)
 - ▶ $H_A : \mu \neq 10$

Set up hypotheses: examples II

- Collect data to find evidence that the pH may differ from neutral in some environment
 - ▶ $H_0 : \mu = 7$
 - ▶ $H_A : \mu \neq 7$
- Collect data on recovery time of a new surgery (for a particular condition)
 - ▶ The recovery time for the current surgery is known to average 10 days
 - ▶ $H_0 : \mu = 10$
 - ▶ $H_A : \mu \neq 10$

Set up hypotheses

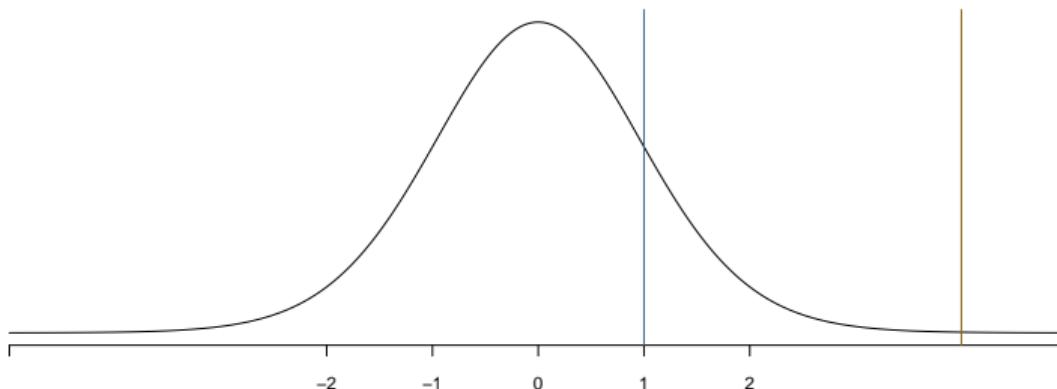
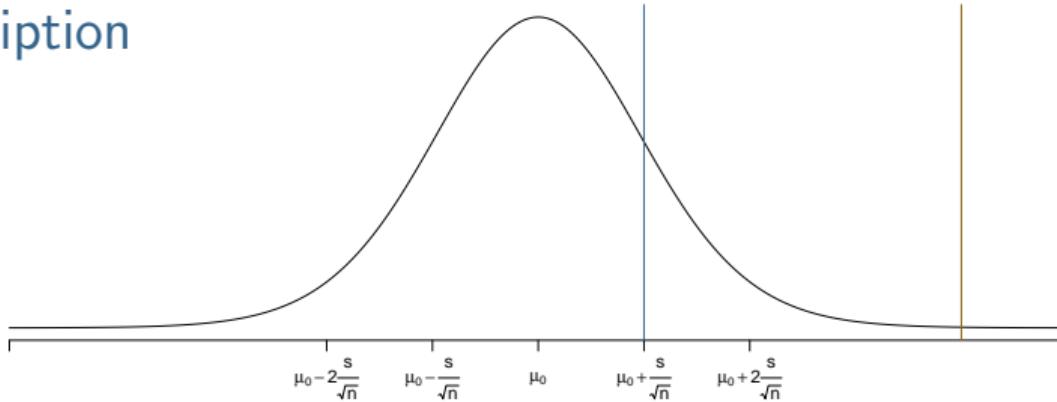
- Hypotheses are statements about parameter values (in our case μ)
- For the Shoshone rectangles we would have:
 - ▶ $H_0: \mu = 1.618$ (the golden ratio)
 - ▶ $H_A: \mu \neq 1.618$
- The null hypothesis is say that the true mean ratio is the golden ratio
 - ▶ Often refer to this as μ_0
 - ▶ An individual garment might have a ratio larger or smaller than the golden ratio
 - ▶ Mean value (in the population) is given by the golden ratio
- The alternative hypothesis¹ says that the true mean ratio is some other value

¹The alternative hypothesis is sometimes referred to as H_1

What if?

- Now we play a ‘what if’ game:
 - ▶ How extreme is the data we observed if the null hypothesis were true?
- Very similar to questions we asked when looking at sampling distributions
 - ▶ Only difference: accounting for not knowing σ
- We calculate a test statistic to help us answer the question
 - ▶ How many standard errors separate the sample mean from null value ($\mu = 1.618$)
 - The standard error is a measure of how variable the sample mean is
 - If the sample mean is 4 standard errors from the null value: unusual
 - If the sample mean is 1 standard error from the null value: not unusual

Graphical description



Test statistic

- Finding how many standard errors separate the sample mean from null value

$$T = \frac{\text{sample mean} - \text{null value}}{\text{standard error}} = \frac{\bar{y} - \mu_0}{\frac{s}{\sqrt{n}}}$$

- Find the relevant quantities for the Shoshone example

```
mu0 = 1.618 # null value
ybar = mean(shoshone$ratio)
ybar # sample mean
## [1] 1.54
n = length(shoshone$ratio) # number of samples (20)
se = sd(shoshone$ratio)/sqrt(n) # standard error
se
## [1] 0.041
```

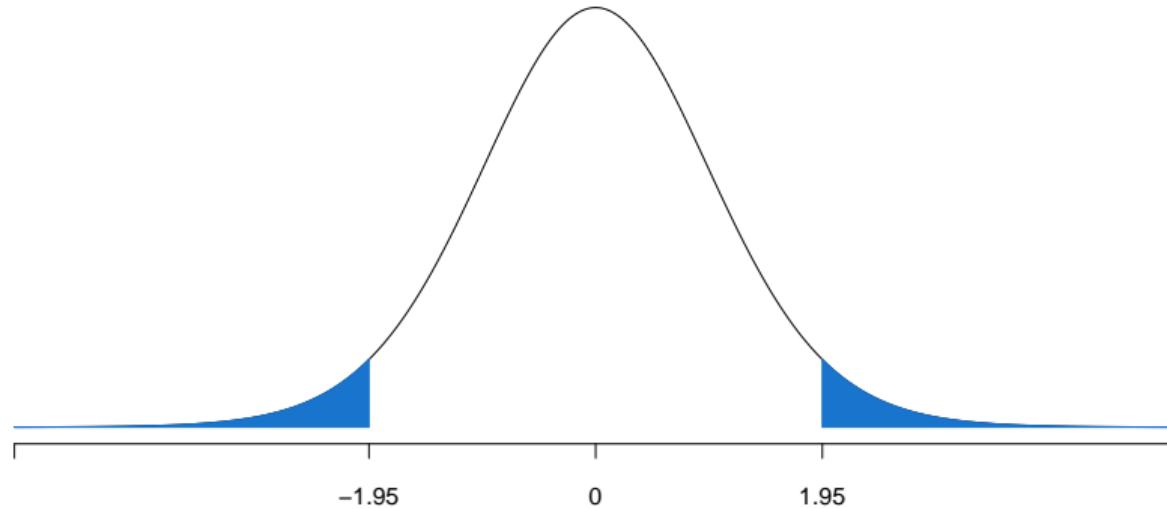
Test statistic

- Find the test statistic

```
Tstat = (ybar - mu0)/se # test statistic  
Tstat  
## [1] -1.95
```

- The sample mean is 1.951 standard errors below the null value
- Is that consistent with the null hypothesis?
 - ▶ Compare it to a t -distribution with $n - 1$ degrees of freedom

Test statistic



p-value

- The tail areas (on the previous slide) give the *p*-value
 - ▶ Find that with `pt` function in R
 - Remember we need to find both tails (or find one and double it)

```
# Find the lower tail: here we have a negative value
tail_lower = pt(Tstat, df = n-1)

# in general, we would use -abs(Tstat) to ensure it is the lower tail
pval = 2*tail_lower

pval
## [1] 0.0659
```

In R: using t.test

- t.test assumes the null value is 0 ($\mu_0 = 0$): change with mu input

```
out = t.test(shoshone, mu = 1.618)

out

##
## One Sample t-test
##
## data: shoshone
## t = -1.95, df = 19, p-value = 0.066
## alternative hypothesis: true mean is not equal to 1.618
## 95 percent confidence interval:
## 1.4523 1.6238
## sample estimates:
## mean of x
## 1.538
```

In R: using `t.test`

- The R output has all of the features we have discussed today:
 - ▶ Test statistic
 - ▶ Degrees of freedom
 - ▶ p -value
 - ▶ Alternative hypothesis (null hypothesis is implicit)

Interpretation

- You would think it should be easy to use and interpret hypothesis tests
 - ▶ It is not
- Hypothesis tests are one of the most heavily used statistical ‘concepts’
 - ▶ Most articles in the (applied science) literature use hypothesis testing in some way
- They are probably the most abused, misunderstood, and misinterpreted concept
 - ▶ Controversial: one psychology journal has banned the use of p-values
 - ▶ American Statistical Association has published articles on their use
 - ▶ We will try to offer a balanced view
 - Further discuss many of the issues later in the semester

What is a *p*-value?

- The *p*-value is the probability of observing data as or more extreme than that observed given the null hypothesis is true
- It provides a measure of incompatibility with statistical model
 - ▶ Model given by null hypothesis
- The smaller the *p*-value, the greater the incompatibility between the data and the null hypothesis
 - ▶ Often expressed as evidence against the null hypothesis
- A *p*-value is **not**:
 - ▶ The probability the null hypothesis is true
 - ▶ The probability that random chance produced the observed data
 - Both of these ‘flip’ a conditional probability

Hypothesis testing in this course

- If the study / example was (likely) confirmatory
 - ▶ Collected data to confirm (or test) a specific hypothesis
 - ▶ We will use formal hypothesis testing
 - Compare p -value to α and make a decision
- If the study / example was (likely) exploratory
 - ▶ Collect data to try and explore and understand scientific phenomena
 - ▶ Use the data to generate hypotheses
 - ▶ We will use p -value to assess the incompatibility of the data to null hypothesis
 - Use α as a guide (more details on next slide)
 - Try not to make a decision between competing hypotheses
 - ▶ Often prefer confidence intervals

Formal test

- If the object of the analysis was to test a particular hypothesis
 - ▶ Use p -value to help make a ‘decision’ between H_0 and H_A
- We have a threshold α (significance level) specified in advance
 - ▶ Often $\alpha = 0.05$ (or 0.01, etc)
- If the p -value $< \alpha$: reject H_0
 - ▶ Evidence in support of H_A
 - ▶ Sometimes called ‘statistically significant’
- If the p -value $> \alpha$: fail to reject H_0
 - ▶ Not enough evidence to reject H_0
 - Not the same as support (or evidence) for H_0
 - Absence of evidence is not evidence of absence
 - ▶ Sometimes referred to as ‘not statistically significant’

Formal test

- It is very easy to abuse a formal hypothesis testing approach
 - ▶ e.g. one of the ASA principles: ‘Scientific conclusions and business or policy decisions should not be based only on whether a p -value passes a specific threshold.’
- What often happens in [practice](#):
 - ▶ Collect data with no clear hypotheses in mind
 - ▶ Explore every possible situation trying to find $p\text{-value} < \alpha$
 - ▶ ‘Torture the data until it confesses’
- Return to a discussion about the use of p -values later in the course

Interpretation of exploratory model

- Use α as a guide for incompatibility between the data and null hypothesis
 - ▶ If $p\text{-value} < \alpha$
 - There is evidence of incompatibility between the data and null hypothesis (relative to α)
 - Investigate further: e.g. look at designing confirmatory study
 - ▶ If we obtain a $p\text{-value} > \alpha$
 - There is no evidence of incompatibility between the data and null hypothesis (relative to α)
 - The degree of incompatibility between the data and null hypothesis (as quantified by the $p\text{-value}$) is similar to what we would expect if the data came from a model under H_0
 - This is not evidence in support of H_0
- Assessing the incompatibility of the data and the null hypothesis
 - ▶ Not making a decision about which hypothesis to adopt based solely on the $p\text{-value}$

Shoshone rectangles

- The Shoshone rectangles: specific hypothesis in mind
 - ▶ Confirmatory: use formal hypothesis testing
- Significance value is $\alpha = 0.05$
- p -value is 0.06593
 - ▶ No obvious incompatibility with the null hypothesis
 - ▶ Formal statement: no evidence to reject H_0
- Recall: $H_0 : \mu = 1.618$
 - ▶ There is no evidence that rectangles used by Shoshone do not follow golden ratio

Summary

- Introduced hypothesis testing
- Two hypothesis:
 - ▶ Null hypothesis
 - ▶ Alternative hypothesis
- Introduced p -value: measure of incompatibility between data and null hypothesis
- Formal hypothesis test
 - ▶ Confirmatory study
- Care is needed in interpretation

Outline

- Previous:
 - ▶ Confidence interval for μ
 - ▶ Hypothesis test
- Today:
 - ▶ Explore more of the properties around the hypothesis test
 - ▶ Type I and Type II Errors
 - ▶ Power of a Test
 - ▶ Trade-offs Between Errors and Power

Height of STAT 110 students

- In previous years there was a questionnaire (optional) for STAT 110 students
 - ▶ Questions about age, height, sex, ...
- Exploratory study
 - ▶ Explore the height of females in STAT 110 relative to national average
 - Average height for NZ female aged 15-24 is 164.7 cm ([figure.nz](#))²
 - Restrict ourselves to female STAT 110 students aged 15-24

```
STAT110 = read.csv('./data/STAT110_height_f.csv')
head(STAT110$height)

## [1] 167 153 171 177 161 173
```

- Heights from $n = 451$ female students aged 15-24

²Data from New Zealand Health Survey, 2023

Hypothesis test

- Write down the null and alternate hypothesis
 - ▶ $H_0 : \mu = 164.7$
 - ▶ $H_A : \mu \neq 164.7$
- Use $\alpha = 0.05$
- We can conduct the test in R

```
h_test = t.test(STAT110$height, mu = 164.7)
h_test
##
##  One Sample t-test
##
## data: STAT110$height
## t = 8.073, df = 450, p-value = 6.32e-15
## alternative hypothesis: true mean is not equal to 164.7
## 95 percent confidence interval:
##  166.891 168.301
## sample estimates:
## mean of x
## 167.596
```

Interpretation

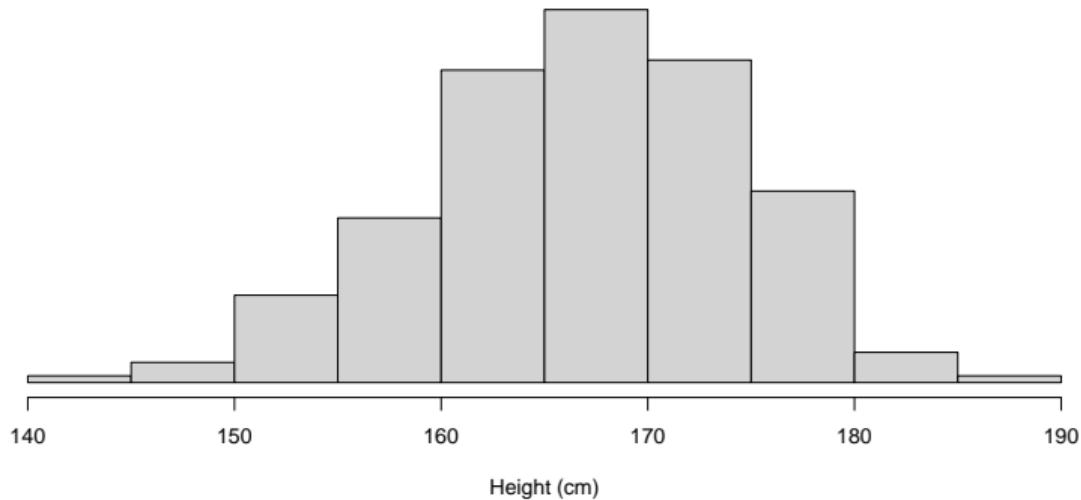
- Exploratory study: interpret p -values (no formal test)
- There is evidence that the data are incompatible with null hypothesis
 - ▶ p -value is approximately 1 in a quadrillion³
- Evidence that the (mean) height of female STAT 110 students is incompatible with national average
- Do we trust it? It pays to be cautious
 - ▶ Students in STAT 110 are not a random selection of 15 – 24 year olds in NZ
 - ▶ STAT 110 data are voluntary and heights are self-reported
 - There are also very different rates of left-handedness from national averages
- If this question were of interest
 - ▶ There is ‘enough’ to look into designing a (confirmatory) study

³The progression is million (10^6), billion (10^9), trillion (10^{12}), quadrillion (10^{15}), ...

Assumptions

- We have made an assumption that our data are normally distributed
 - ▶ Just as we did with confidence intervals
- To check this assumption: looking for serious departures from normality
 - ▶ We check visually (histogram)
- As with confidence intervals: if the sample size is large enough
 - ▶ p -values are reasonable for non-normal data
 - ▶ Discuss more in a few weeks

Histogram



- No obvious departures from normality
- Large sample (~ 450)

Setup

- We want to better understand how hypothesis testing works
- We do this in the context of formal hypothesis test
 - ▶ If $p\text{-value} < \alpha$ we reject H_0
 - ▶ If $p\text{-value} > \alpha$ we fail to reject H_0
- There are four possibilities:

		Decision	
		Do not reject H_0	Reject H_0
H_0 true	H_0 true	✓	Type I error
	H_0 not true	Type II error	✓

Setup

- Consider a specific gene: GENE-X
 - ▶ Reference expression value of 5.0 TPM (transcripts per million) in healthy individuals
- Design a confirmatory study to test if GENE-X is expressed differently in a sample of people with a specific disease
 - ▶ $H_0 : \mu = 5$ (the mean expression for the diseased group is the same as the reference)
 - ▶ $H_A : \mu \neq 5$
- In this study:
 - ▶ We want to find evidence against the null
 - ▶ We want to find evidence that gene expression differs in the diseased group
- In the rest of the lecture an effect is defined as:
 - ▶ Effect: difference between the mean for the disease group and $\mu_0 = 5$

A tale of two errors

- Type I Error (α): Rejecting H_0 when it is true.
 - ▶ Concluding the expression of GENE-X is different for the diseased group, when it isn't
- Type II Error (β): Failing to reject H_0 when H_a is true.
 - ▶ Concluding that there is no evidence that expression of GENE-X differs for diseased group, when there is a non-zero effect

Type I error

- Type I error rate is given by α , the significance level
 - ▶ Decreasing α from 0.05 to 0.01 will reduce the number of type I errors we make
 - Recall: α is the threshold for incompatibility with null
 - A lower α is applying a higher threshold for incompatibility

Type II error

- The type II error rate is represented as β
- We often refer to the power = $1 - \beta$
- Power: the probability of rejecting the null hypothesis, given it is incorrect
 - ▶ i.e. it is the probability of detecting an effect, given there is one
- All else equal, we want a powerful test
 - ▶ More likely to correctly reject H_0
 - ▶ More likely to correctly conclude that gene expression differs in diseased group
- We will look at four factors that change the type II error / power

Type I error rate α

- Trade off between type I error rate and power
 - ▶ If we decrease α (lower type I error rate)
 - Increase type II error rate β
 - Decrease power
 - If we increase α (higher type I error rate)
 - ▶ Decrease type II error rate β
 - ▶ Increase power

Effect size

- Recall: $\mu_0 = 5$ TPM (transcripts per million)
- Consider two scenarios:
 1. The true mean of the diseased population is $\mu_A = 5.1$ TPM
 2. The true mean of the diseased population is $\mu_A = 12$ TPM
- In which scenario will power be higher (all else equal)?

⁴ $|x|$ is the absolute value of x

Effect size

- Recall: $\mu_0 = 5$ TPM (transcripts per million)
- Consider two scenarios:
 1. The true mean of the diseased population is $\mu_A = 5.1$ TPM
 2. The true mean of the diseased population is $\mu_A = 12$ TPM
- In which scenario will power be higher (all else equal)?
- The larger⁴ the effect $|\mu_A - \mu_0|$
 - ▶ The more powerful the test, all else equal
- The size of the effect is not something we can typically control

⁴ $|x|$ is the absolute value of x

Sample size

- For a fixed α and effect size, consider these two scenarios:
 1. The sample size (of diseased participants) is $n = 20$
 2. The sample size (of diseased participants) is $n = 200$
- In which scenario will power be higher?

Sample size

- For a fixed α and effect size, consider these two scenarios:
 1. The sample size (of diseased participants) is $n = 20$
 2. The sample size (of diseased participants) is $n = 200$
- In which scenario will power be higher?
- The larger the sample size
 - ▶ The more powerful the test, all else equal
- Scientific research (grant) funding in ecology, food science, global health, etc
 - ▶ Typically have to justify your research design
 - ▶ Power calculation: determining sample size needed to achieve a certain power

Population standard deviation

- For a fixed n , α , and effect size, consider these two scenarios:
 1. The population standard deviation (of gene expression in the disease group) is $\sigma = 0.1$
 2. The population standard deviation (of gene expression in the disease group) is $\sigma = 1$
- In which scenario will power be higher?

Population standard deviation

- For a fixed n , α , and effect size, consider these two scenarios:
 1. The population standard deviation (of gene expression in the disease group) is $\sigma = 0.1$
 2. The population standard deviation (of gene expression in the disease group) is $\sigma = 1$
- In which scenario will power be higher?
- The smaller the population standard deviation
 - ▶ The smaller the standard error
 - ▶ The more precise \bar{y} is
 - ▶ The more powerful the test, all else equal
- The value of σ is not something we can typically control

p-value

- ASA principle: “A *p*-value, or statistical significance, does not measure the size of an effect or the importance of a result”
- Suppose we have $p = 0.0000001$. This could be because:
 - ▶ This could be because the effect size is large
 - ▶ It could occur when the effect size is small (but non-zero) and sample size is large
- Care is needed that we don’t confuse a small *p*-value, with an important result

Relationship with confidence intervals

- If we are testing the hypothesis:
 - ▶ $H_0 : \mu = \mu_0$
 - ▶ $H_A : \mu \neq \mu_0$
- There is an equivalence between p -value and confidence interval
 - ▶ p -value $< \alpha$ is equivalent to μ_0 outside the $(1 - \alpha)100\%$ confidence interval
 - e.g. if p -value < 0.05 , then μ_0 is outside 95% confidence interval
 - e.g. if p -value > 0.01 , then μ_0 is inside 99% confidence interval

Quiz

- It's quiz time!
- Three possible answers for the questions below:
 - ▶ (1) increase; (2) decrease; (3) can't tell
- What is the effect on (i) type I error rate, and (ii) power if we:
 - ▶ Increase the sample size?
 - ▶ Decrease α ?
 - ▶ Decrease the sample size and increased α ?
 - ▶ Changed the research design so that the type II error rate β decreased?
 - ▶ Collected a sample twice the size for a different gene (GENE-Y) that has a smaller effect and larger σ ?

Summary

- Checking assumptions
- Looked more at the properties of hypothesis testing
 - ▶ Type I error
 - ▶ Type II error
 - ▶ Power
- Looked at the effect of
 - ▶ Sample size
 - ▶ Effect size
 - ▶ α
 - ▶ σ

STAT 110: Week 6

University of Otago

Outline

- Previous lectures:
 - ▶ Explored statistical models for normally distributed data
 - ▶ Data are modelled as normal with mean μ and variance σ^2
 - ▶ Found confidence interval for μ
 - ▶ Hypothesis test for μ
- Today: begin to look at relationships between variables
 - ▶ Relationship between a continuous variable and a categorical variable
 - ▶ Continuous variable: can take any value
 - e.g. height, weight, time to run 100 m
 - It could be limited a range (e.g. height must be positive)
 - ▶ Categorical variable: represents categories or groups
 - e.g. sex, country of birth, blood type, etc.

Motivation

- What is the effect of sensory deprivation?¹
 - ▶ Study designed to explore this question, where all participants were prisoners
- Twenty participants were selected
 - ▶ 82 inmates initially volunteered
 - Removed: medically unfit, low IQ, history of behaviour or psychiatric problems in prison
- The 20 participants were randomly allocated into two groups
 - ▶ Solitary confinement
 - ▶ Control (ordinary prison life)
- EEG² frequencies were obtained on day 7
 - ▶ Is there a difference in arousal levels? (as measured by EEG frequency)

¹ From Journal of Abnormal Psychology, 1972, 79, 54–59

² EEG (Electroencephalogram) measures the frequency of brain waves

Data: EEG frequencies

- Import the data

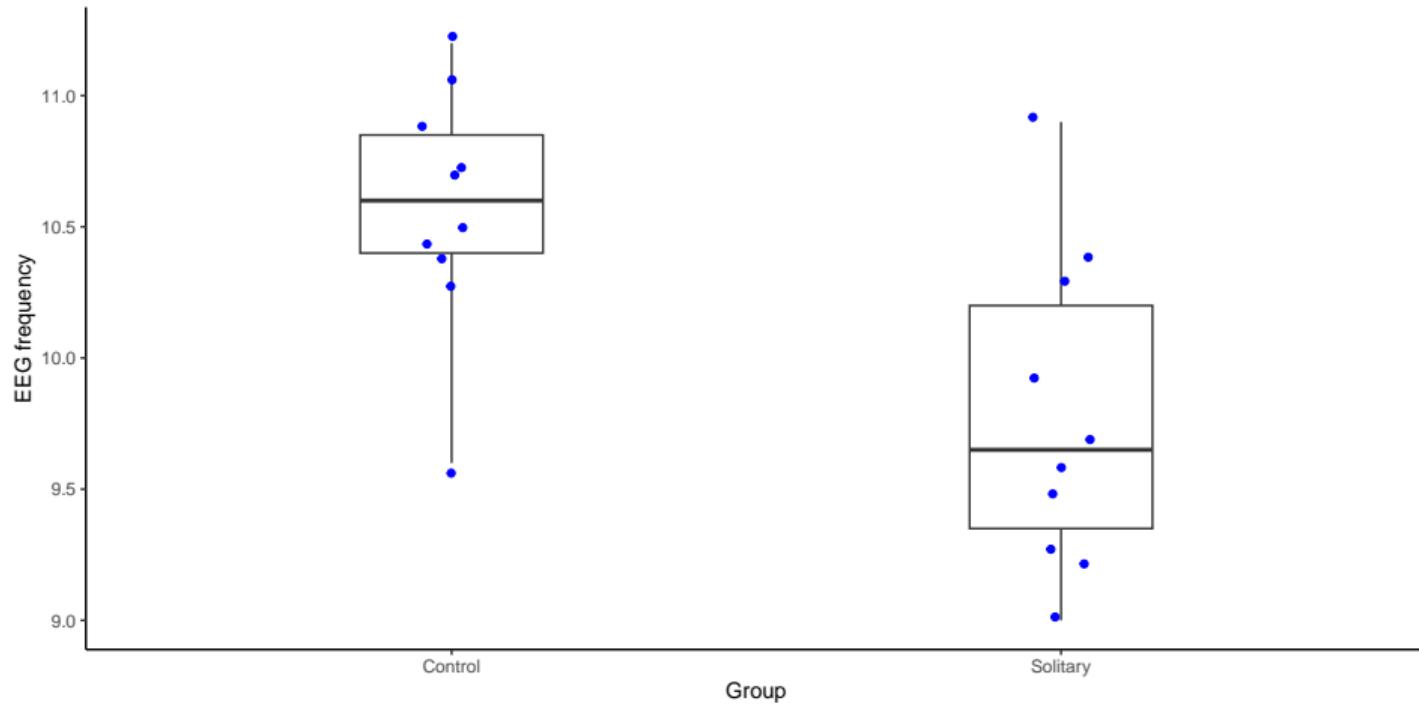
```
EEG = read.csv('EEG.csv')
```

- Have a look at the data:

```
head(EEG)

##      Group Freq
## 1 Control 10.7
## 2 Control 10.7
## 3 Control 10.4
## 4 Control 10.9
## 5 Control 10.5
## 6 Control 10.3
```

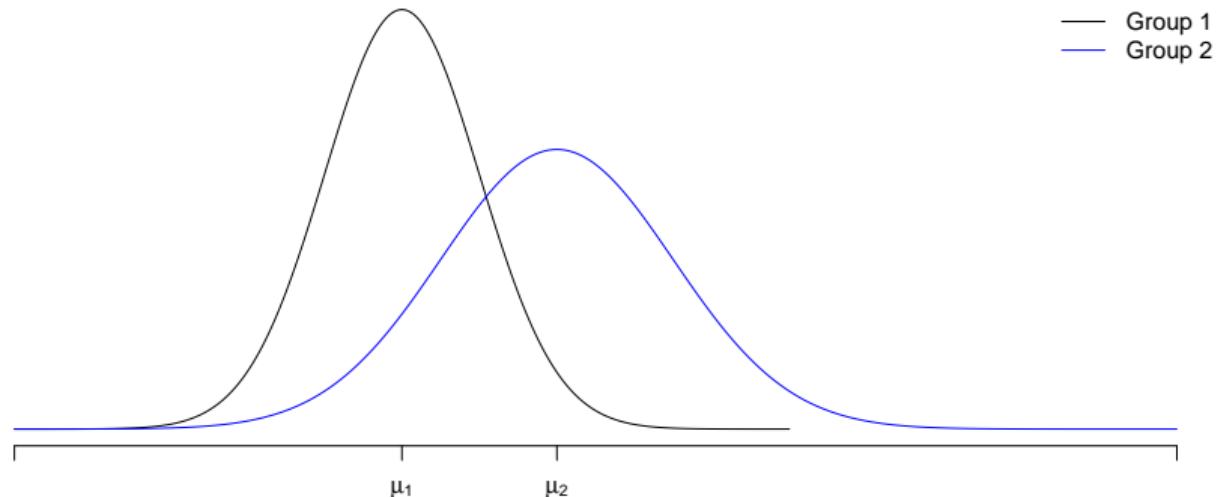
Visualise the data



Problem

- We have looked at models:
 - ▶ Data are normally distributed with mean μ and variance σ^2
 - ▶ Focus has been on the estimation of a (single) mean μ
- We need to extend our model to allow for two groups of data
 - ▶ Group 1 (experimental): normally distributed with mean μ_1 and variance σ_1^2
 - ▶ Group 2 (control): normally distributed with mean μ_2 and variance σ_2^2
- Interest is in the difference in means between the two groups
 - ▶ $\mu_1 - \mu_2$ (or $\mu_2 - \mu_1$)
- Difference in the mean arousal level between the deprived and the controls

Model (graphical representation)



Other examples

- There are other applications we could have used to motivate:
 - ▶ Cuckoos are avian brood parasites: they lay their eggs in the nest of other birds
 - Compare the length of cuckoo eggs in wren and robin nests
 - ▶ Explore differences in chemical composition of wine or olives
 - Different cultivars (wine)
 - Different regions (olives)
 - ▶ Comparing athletic performance
 - Comparing resistance training and traditional training for athletes in some sport
 - ▶ Survival time for breast cancer patients
 - Comparing candidate drug and placebo
 - ▶ Gene expression in a section of the brain
 - Comparing diseased, with healthy controls
 - ▶ You will see some of these in the Assignment

How to find a confidence interval

- Much of what we have learned previously ‘carries over’
- Use statistics (from sample) to estimate parameters (from population)
 - ▶ Parameter: $\mu_1 - \mu_2$
 - ▶ Statistic: $\bar{y}_1 - \bar{y}_2$
- Standard error for $\bar{y}_1 - \bar{y}_2$
 - ▶ Tells us about the variation in $\bar{y}_1 - \bar{y}_2$ in repeated samples
- Estimated standard error: $s_{\bar{y}_1 - \bar{y}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$
- The confidence interval is given as

$$\underbrace{\bar{y}_1 - \bar{y}_2}_{\text{statistic}} \pm \underbrace{t_{\nu, 1-\alpha/2}}_{\text{multiplier}} \underbrace{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}_{\text{standard error}}$$

Standard error

- The standard error is different from before, but similar
 - ▶ Follows from variance rules (week 3; ice cream)
 - ▶ Observations in the two groups are independent

$$\begin{aligned} \text{Var}(\bar{y}_1 - \bar{y}_2) &= \text{Var}(\bar{y}_1) + \text{Var}(\bar{y}_2) \\ &= \frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} \end{aligned}$$

Multiplier

- The multiplier is again given by the t -distribution
 - ▶ The use of the t -distribution relies on an approximation
 - Approximation is accurate provided we have more than a handful of observations ($n_1 > 5, n_2 > 5$)
- The degrees of freedom, ν , we use is given by a complicated formula
 - ▶ You have no need to know or learn this

$$\nu = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} \right)^2}{\frac{s_1^4}{n_1^2(n_1-1)} + \frac{s_2^4}{n_2^2(n_2-1)}}.$$

- If software isn't available, simpler approximations for ν are sometimes used
 - ▶ e.g. using smaller of $n_1 - 1$ and $n_2 - 1$
 - ▶ Conservative

Calculating the confidence interval

- We could calculate the confidence interval by hand:
 - ▶ Find the sample mean in each group: \bar{y}_1, \bar{y}_2
 - ▶ Find the sample variance in each group: s_1^2, s_2^2
 - ▶ Find the standard error
 - ▶ Calculate the degrees of freedom
 - ▶ Find the t -multiplier
 - ▶ Construct the confidence interval
- Tedious task
 - ▶ Important to know how the interval is constructed
 - You may be asked to do various aspects of it for assignment/test/exam
 - ▶ Easier to use R to calculate the interval

In R

- We use the same function as before: `t.test`
 - ▶ This requires us to have the data for each group separately
 - ▶ Currently our data are in a single data frame

```
head(EEG)

##      Group Freq
## 1 Control 10.7
## 2 Control 10.7
## 3 Control 10.4
## 4 Control 10.9
## 5 Control 10.5
## 6 Control 10.3
```

- The variable `Group` distinguishes which group the observation is from
 - ▶ Either `Control` or `Solitary`

In R

- There are several ways in R we could separate into two groups
 - ▶ We will use `subset`
 - Subsets the data based on a specified criteria
 - ▶ Only cover 'basic' data handling in STAT 110
 - See STAT 260

```
control = subset(EEG, Group == "Control")
solitary = subset(EEG, Group == "Solitary")
```

- We use two equal signs (==) to *check* equality
 - ▶ `Group == "Solitary"` is checking which observations are Solitary

In R

- Check each of these objects

control

```
##      Group Freq
## 1 Control 10.7
## 2 Control 10.7
## 3 Control 10.4
## 4 Control 10.9
## 5 Control 10.5
## 6 Control 10.3
## 7 Control  9.6
## 8 Control 11.1
## 9 Control 11.2
## 10 Control 10.4
```

solitary

```
##      Group Freq
## 11 Solitary  9.6
## 12 Solitary 10.4
## 13 Solitary  9.7
## 14 Solitary 10.3
## 15 Solitary  9.2
## 16 Solitary  9.3
## 17 Solitary  9.9
## 18 Solitary  9.5
## 19 Solitary  9.0
## 20 Solitary 10.9
```

In R

- Each of the groups is a separate argument in `t.test`

```
out = t.test(control$Freq, solitary$Freq)
out

##
##  Welch Two Sample t-test
##
## data: control$Freq and solitary$Freq
## t = 3, df = 17, p-value = 0.004
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.297 1.303
## sample estimates:
## mean of x mean of y
##      10.58      9.78
```

R output

- R calculates the degrees of freedom for us: $\nu = 16.875$
- R gives us the means

```
out$estimate # gives the samples means of the two groups  
## mean of x mean of y  
##      10.58      9.78  
  
out$estimate[1] - out$estimate[2] # find the diff in sample means  
## mean of x  
##      0.8
```

- When interpreting, we must be careful to not confuse the order
 - ▶ Mean of x corresponds to the first argument: controls
 - ▶ Mean of y corresponds to the second argument: solitary
 - ▶ Confidence interval is for $\mu_x - \mu_y$, or $\mu_{\text{control}} - \mu_{\text{solitary}}$

Confidence interval

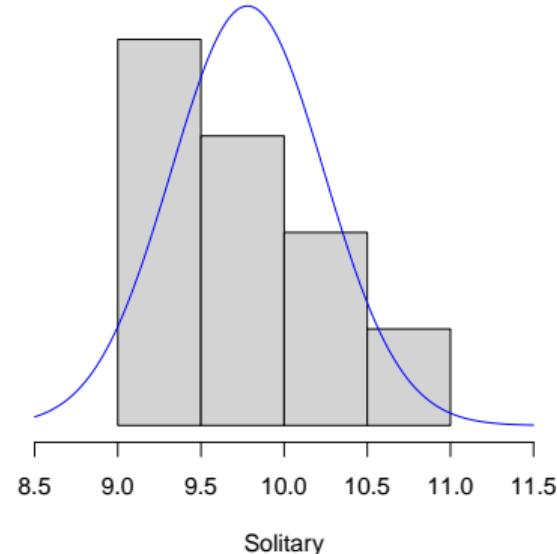
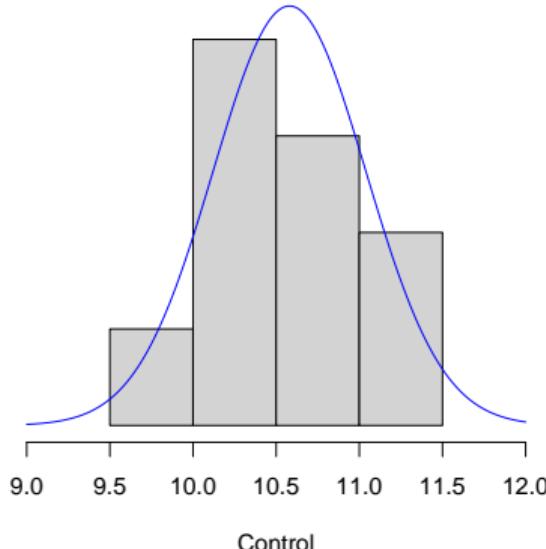
- The confidence interval is

```
out$conf.int  
## [1] 0.297 1.303  
## attr(,"conf.level")  
## [1] 0.95
```

- We are 95% confident that the mean EEG frequency for the control group is between (0.297, 1.303) higher than those in solitary confinement
- The confidence interval has the same properties as before
 - In the long run, we would expect 95% of the confidence intervals we calculate to include the true difference $\mu_1 - \mu_2$
 - If we were to repeatedly sample from the population and repeat this analysis

Checking assumptions

- We are assuming a normal model for each group
- Check fitted model



Checking assumptions

- No major departures from normality
- Enough to make us cautious
 - ▶ Small sample size: normality assumption very important
 - It is hardest to assess normality assumptions, when it matters the most
- Want to be cautious in our conclusions

Hypothesis test

- This study was set up to look into a specific hypothesis
 - ▶ Confirmatory
- Theory was that sensory deprivation changes EEG frequency
- Null hypothesis: status quo / assumption of no difference
 - ▶ The two groups have the same mean: $\mu_1 = \mu_2$
 - ▶ $H_0 : \mu_1 - \mu_2 = 0$
- The alternative hypothesis
 - ▶ The two groups differ: $\mu_1 \neq \mu_2$
 - ▶ $H_A : \mu_1 - \mu_2 \neq 0$

Hypothesis test

- The same function (`t.test`) is used to calculate a hypothesis test

```
out = t.test(control$Freq, solitary$Freq)

out

## 

## Welch Two Sample t-test

## 

## data: control$Freq and solitary$Freq
## t = 3, df = 17, p-value = 0.004
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.297 1.303
## sample estimates:
## mean of x mean of y
##      10.58      9.78
```

Interpretation

- The p -value is 0.004
 - ▶ Evidence of incompatibility between data and null hypothesis
 - ▶ Data provide support for the alternative hypothesis
 - Difference in EEG frequency between the control and solitary groups
- Given the small sample and cautiousness in checking assumptions
 - ▶ We have provided evidence in support of EEG differing
 - ▶ Larger studies desirable to provide further confirmation

Confidence intervals vs hypothesis testing

- In this example we look at both confidence intervals and hypothesis test
- The p -value does not tell us how strong an effect is
 - ▶ We could have p -value of 0.05 with $\bar{y}_1 - \bar{y}_2 = 10$
 - Small sample size
 - ▶ We could have p -value of 0.001 with $\bar{y}_1 - \bar{y}_2 = 0.002$
 - Large sample size
- Confidence interval gives an interval estimate of effect

Independent groups

- We have assumed the two groups are independent
 - ▶ Important assumption
- What does that mean?
 - ▶ The outcome from one group does not affect the outcome from the other group
- This will not always be the case:
 - ▶ Students take a test before undertaking a course
 - ▶ Same students undertake the same test after the course
 - Same participants in each 'group'
 - It is likely that someone who scored well in first test will also score well in the second test
- Look into this more tomorrow

Summary

- First look at relationship between variables
 - ▶ How EEG frequency varies by sensory deprivation
- Relationship between a continuous variable and a categorical variable
 - ▶ EEG frequency (continuous); sensory deprivation yes/no (categorical)

Outline

- Previous:
 - ▶ Started to look at relationships between variables
 - Frequency of brain waves (EEG) and sensory deprivation
 - ▶ Examples of relationship between one continuous and one categorical variable
 - Two groups are independent
- Today:
 - ▶ Look at paired data (two groups are not independent)
 - ▶ Start looking at relationships between two continuous variables

Motivating example

- Reaction time (ms) for 23 participants (press a button after stimulus)
 - ▶ University students
- There are two stimuli:
 - ▶ Auditory (a burst of white noise)
 - ▶ Visual (a circle flashing on a computer screen)
- Each participant exposed to both stimuli
 - ▶ Shouldn't use the approach from previous lecture
 - ▶ The two groups are not independent
 - We might expect someone with fast reaction time (auditory) to have a fast reaction (visual)
- Example of paired data
 - ▶ Each observation in group one has correspondence to an observation in group two
- This is an exploratory study

Data

```
AV = read.csv('AV.csv')
head(AV)
```

```
##      auditory visual
## 1        226     256
## 2        188     309
## 3        280     364
## 4        234     379
## 5        181     268
## 6        178     288
```

Paired: find the difference back to the future

- Look at the difference in the outcomes for each pair

```
AV$differ = AV$visual - AV$auditory  
# this adds another variable (called differ) to the data frame AV  
head(AV)  
  
##   auditory visual differ  
## 1      226    256   29.3  
## 2      188    309  121.9  
## 3      280    364   83.7  
## 4      234    379  144.8  
## 5      181    268   87.1  
## 6      178    288  109.9
```

Paired: back to the future

- Model the differences as if they were a single sample
 - ▶ The data are the differences and are given by y_d
 - ▶ The differences y_d are assumed to be normal with mean μ_d and variance σ_d^2
 - ▶ μ_d is a parameter representing the mean difference in the population
- For our example:
 - ▶ y_d is the difference in reaction time (visual - auditory)
 - ▶ μ_d is the population mean difference in reaction time (visual - auditory)

In R

- For paired data: two ways to find confidence intervals and hypothesis tests in R
- Option 1: use `t.test` on the differenced values

```
t.test(AV$differ)

##
##  One Sample t-test
##
## data: AV$differ
## t = 4, df = 22, p-value = 2e-04
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  32.3 87.9
## sample estimates:
## mean of x
##       60.1
```

In R

- For paired data: two ways to find confidence intervals and hypothesis tests in R
- Option 2: specify the two groups and include option paired = TRUE

```
t.test(AV$visual, AV$auditory, paired = TRUE)

##
##  Paired t-test
##
## data:  AV$visual and AV$auditory
## t = 4, df = 22, p-value = 2e-04
## alternative hypothesis: true mean difference is not equal to 0
## 95 percent confidence interval:
##  32.3 87.9
## sample estimates:
## mean difference
##                 60.1
```

Output and interpretation

- Both approaches give identical confidence intervals
- Minor differences
 - ▶ Input differs: (1) input the differences; (2) input each group
 - ▶ Wording differences in output
 - ‘One sample t-test’ vs ‘Paired t-test’
 - ‘true mean’ vs ‘true mean difference’
 - ‘mean of x’ vs ‘mean difference’
- Interpretation:
 - ▶ We are 95% confident that mean difference in the reaction times between visual and auditory stimuli is between (32.3, 87.9) ms

Hypothesis test

- Often with an exploratory study: use confidence interval
 - ▶ Calculate hypothesis test here as an example
- The hypothesis test is in terms of μ_d
- Null hypothesis: assumption of no difference ($\mu_d = 0$)
 - ▶ $H_0 : \mu_d = 0$
 - ▶ $H_A : \mu_d \neq 0$
- The p -value is 1.85×10^{-4}
 - ▶ Evidence that data are incompatible with the null hypothesis
 - ▶ There is evidence (at the $\alpha = 0.05$ level) that the data are incompatible with assumption of no difference

Extension

- Many applications may have more than two groups
 - ▶ Data from multiple independent groups
 - ▶ Multiple observations of each subject (repeated measures)
- There are statistical models for both cases
 - ▶ Independence: ANOVA (analysis of variance)
 - We will see this later in the course
 - ▶ Repeated measures: complex model
 - Outside the scope of this course

Relationship between continuous variables

- Previous examples: relationship between a continuous variable and a categorical variable
 - ▶ Continuous: reaction time; categorical: stimuli
 - ▶ Continuous: EEG frequency; categorical: sensory status (solitary/control)
- We are now going to consider relationships between two continuous variables

Motivating examples

- We are going to introduce three motivating examples
 1. The size of brushtail possums
 - Compare total length (mm) to head length (cm)
 - $n = 104$ observations
 2. Height of STAT 110 students
 - Compare father's height (cm) to son's height (cm)
 - $n = 279$ observations
 3. Squat weight of international power lifters
 - Comparing body weight (kg) to max squat weight (kg)
 - Photo from powerliftingtechnique.com
 - The athlete pictured (Kelly Branton) is in the dataset
 - $n = 9045$ observations (athletes)
- All of these involve two continuous variables



Brushtail possums

- Import the data

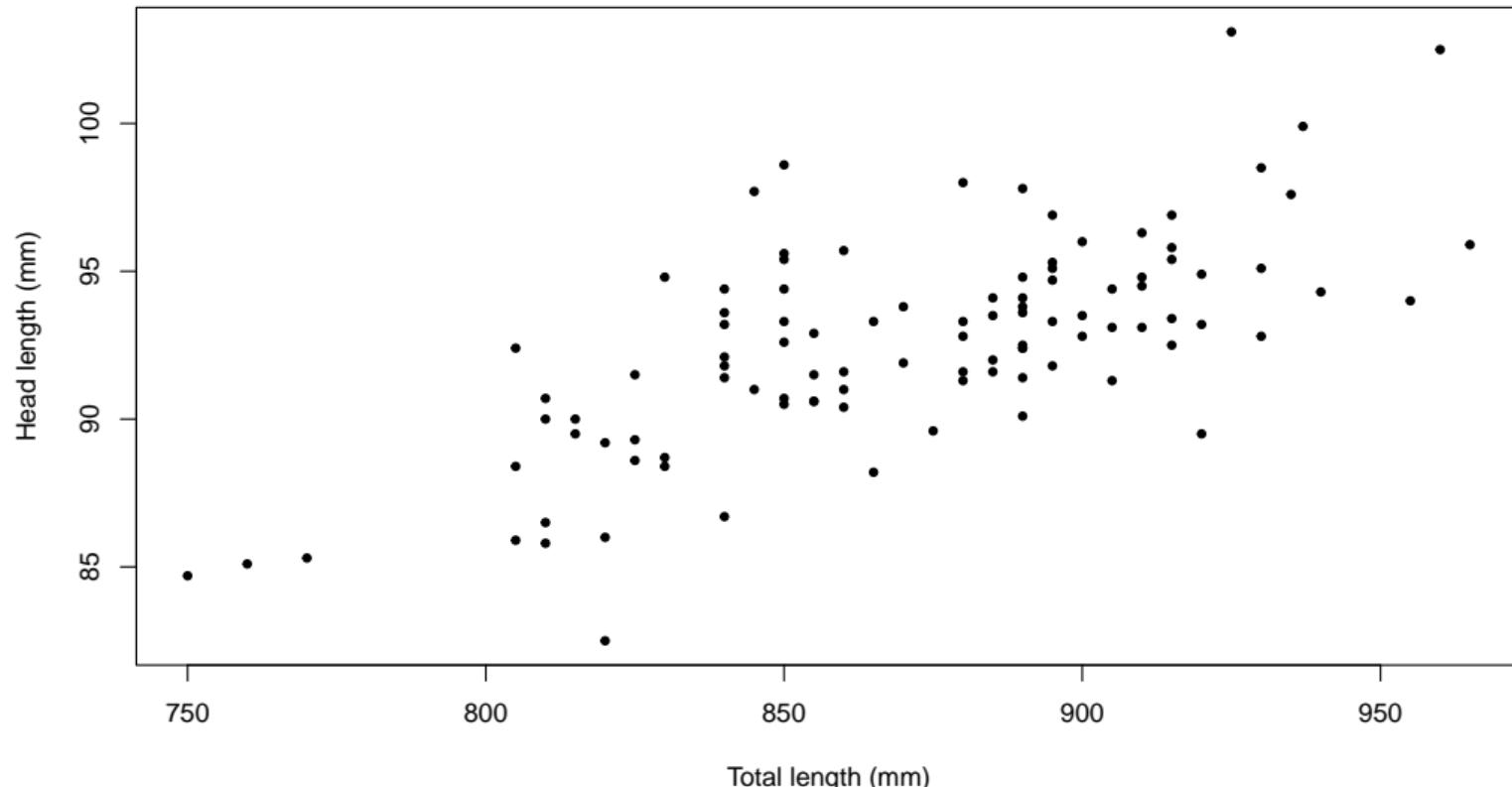
```
possum = read.csv('possum.csv')
```

- Have a look at the data:

```
head(possum)

##      total_l head_l
## 1      890   94.1
## 2      915   92.5
## 3      955   94.0
## 4      920   93.2
## 5      855   91.5
## 6      905   93.1
```

Brushtail possums: scatterplot



Father & son height

- Import the data

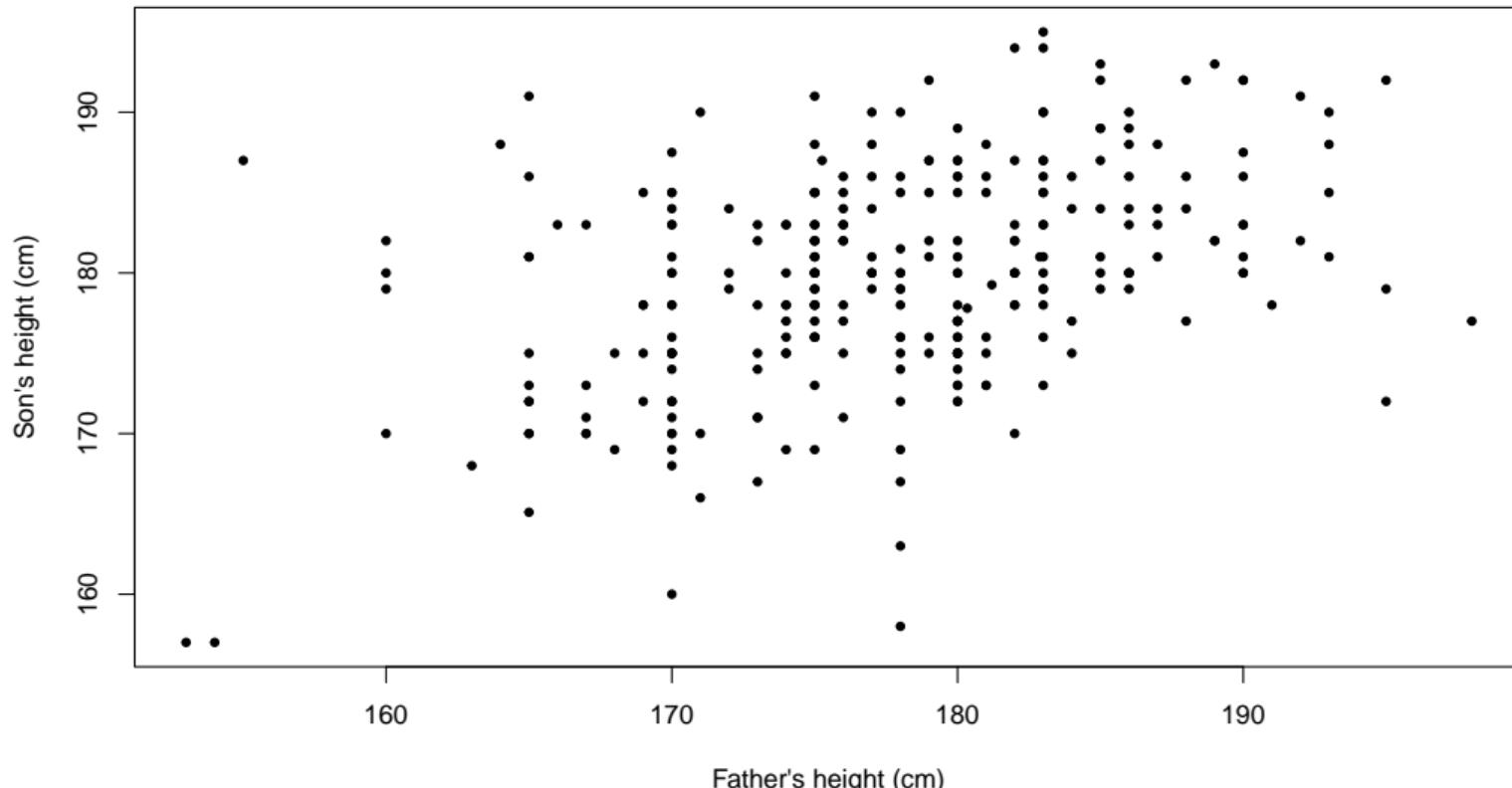
```
height = read.csv('height.csv')
```

- Have a look at the data:

```
head(height)

##      son father
## 1    176    178
## 2    180    190
## 3    180    174
## 4    181    179
## 5    184    187
## 6    180    182
```

Father & son height: scatterplot



Powerlifting

- Import the data

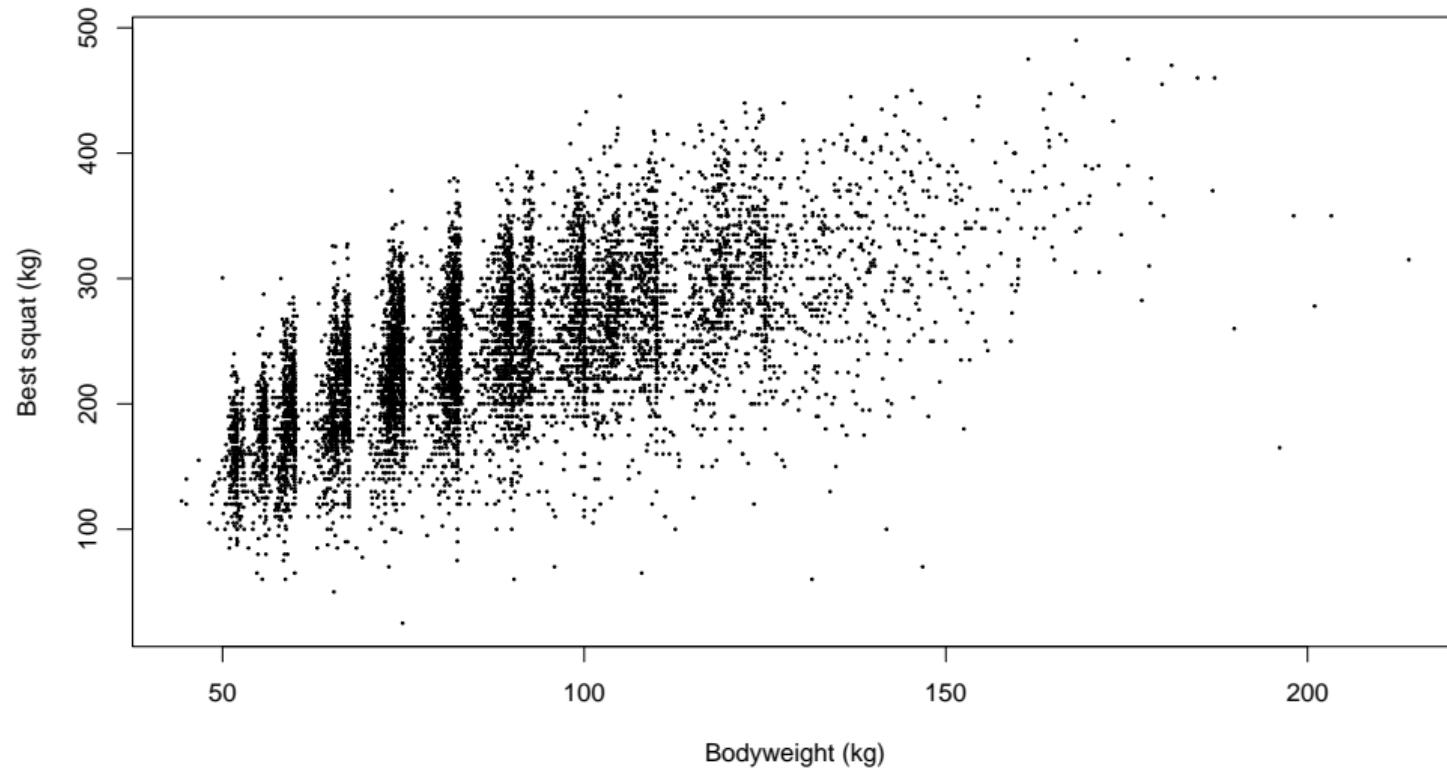
```
powerlift = read.csv('powerlift.csv')
```

- Have a look at the data:

```
head(powerlift)

##      bodyweight bestsquat
## 1        59.6     228
## 2        67.2     255
## 3        67.4     270
## 4        59.9     260
## 5        59.9     250
## 6        56.0     210
```

Powerlift: scatterplot



Back to the beginning

- What was the first thing we did when we first encountered data in STAT 110?
 - ▶ Found data summaries: sample mean and sample variance
- What summary describes the relationship between two continuous variables?

Correlation

- Correlation describes the strength of a linear relationship between two variables (let's call them x and y)
 - ▶ Always takes a value between -1 and 1
 - ▶ Population correlation represented by ρ (greek letter rho)
 - ▶ Sample correlation represented by r
- With data $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$, the correlation is given by

$$r = \frac{1}{n-1} \sum_{i=1}^n \frac{(x_i - \bar{x})}{s_x} \frac{(y_i - \bar{y})}{s_y}$$

- We will calculate the correlation using the R function `cor`

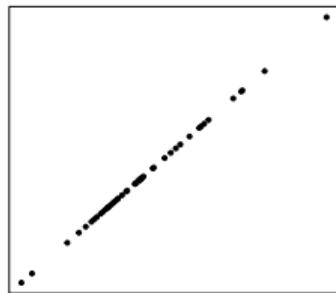
```
cor(possum$total_l, possum$head_l)  
## [1] 0.691
```

Understanding correlation

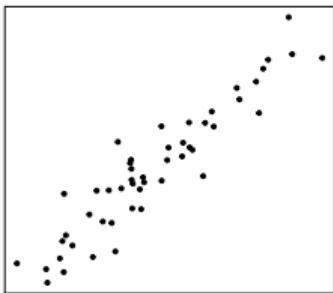
- Positive correlation:
 - ▶ If y is above its mean, then x is likely to be above its mean (and vice versa)
- Negative correlation
 - ▶ If y is above its mean, then x is likely to be below its mean (and vice versa)
- If the relationship is strong and positive
 - ▶ r will be close to 1
- If the relationship is strong and negative
 - ▶ r will be close to -1
- If there is no apparent (linear) relationship between x and y
 - ▶ r will be close to 0

Understanding correlation: graphically I

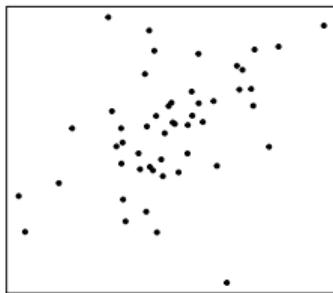
$r = 1$



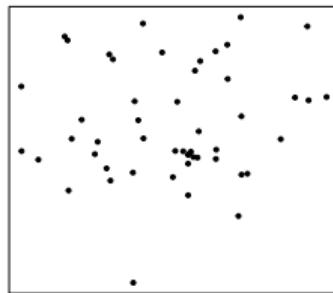
$r = 0.93$



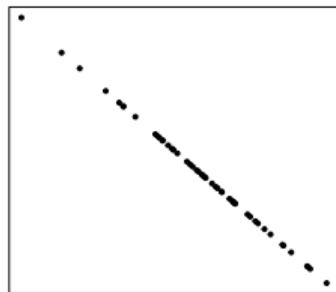
$r = 0.44$



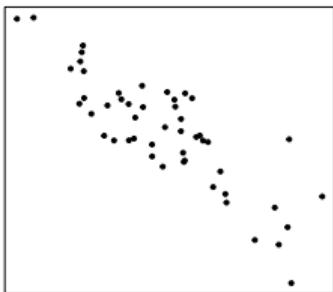
$r = 0.1$



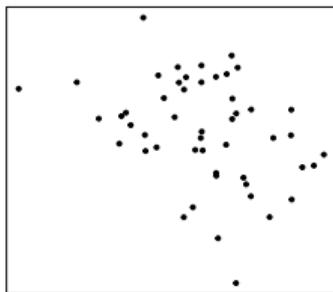
$r = -1$



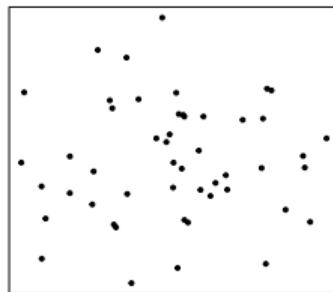
$r = -0.84$



$r = -0.34$



$r = 0.02$



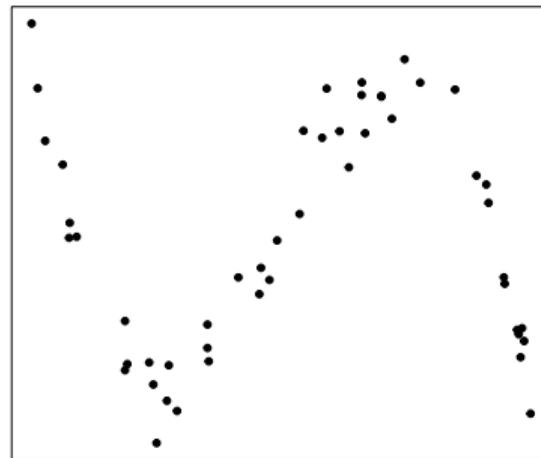
Understanding correlation: graphically II

- r measures the strength of the linear relationship
 - ▶ Strong non-linear relationships can produce r values that do not reflect the strength of the relationship

$r = -0.1$

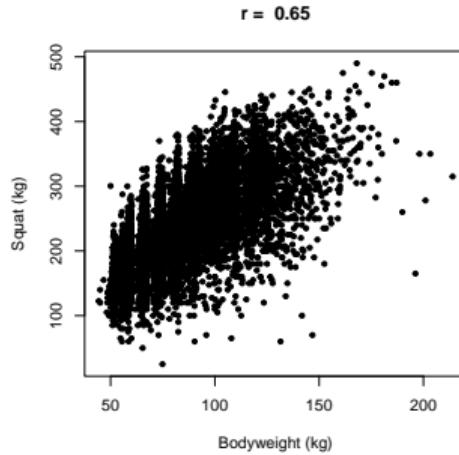
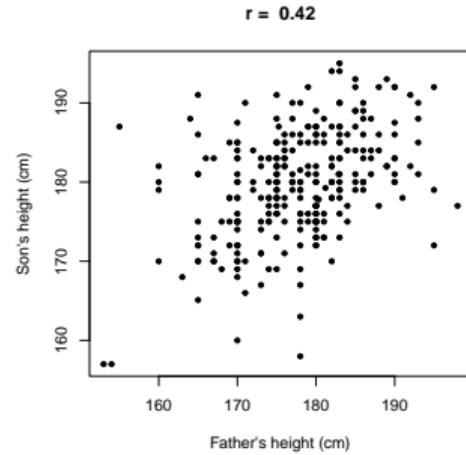
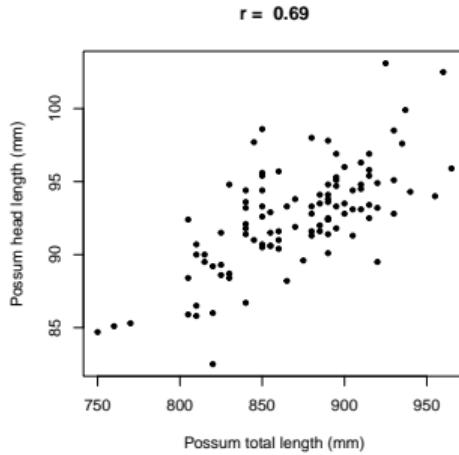


$r = 0.08$



Data

```
rpossum = cor(possum$total_l, possum$head_l)  
rheight = cor(height$son, height$father)  
rpower = cor(powerlift$bodyweight, powerlift$bestsquat)
```



Practice

- Guess the correlation

Limitations

- The correlation r is a useful summary
 - ▶ We may want to learn how precise it is: confidence interval
 - ▶ Such intervals can be found: `cor.test` in R
 - We will not consider them in STAT 110
- The correlation as a summary is limited
- What might we want to know?
 1. Possum data: predict head length from a measurement of total length
 2. Height data: understanding and quantifying heritability of height as a trait
 3. Powerlifting: compare the squat weight of an athlete to their peers of a similar weight
- Correlation does not help us for 1 and 3
 - ▶ Limited for 2: quantifies the linear relationship, but does not describe it
 - What is the expected difference in height between a son with father who is 170 cm tall, and a son with father who is 180 cm tall?

Summary

- Looked at paired data
 - ▶ Model the difference between the two groups
 - ▶ Confidence intervals
 - ▶ Hypothesis test
- Looked at relationships between two continuous variables
- Explored a data summary: correlation
 - ▶ Gives the strength of a linear relationship between two variables
 - ▶ Always between -1 and +1
 - ▶ Easy to calculate in R

Outline

- Continue to explore relationships between two variables
- Go beyond summary statistics
 - ▶ Look into a statistical model for the relationship
 - What the model looks like
 - Fitted model
 - Residuals

Recall: motivating examples

- The size of brushtail possums
 - ▶ Compare total length (mm) to head length (cm)
- Height of STAT 110 students
 - ▶ Compare father's height (cm) to son's height (cm)
- Squat weight of international power lifters
 - ▶ Comparing body weight (kg) to max squat weight (kg)

Recall: correlation

- The correlation r measures the strength of linear relationship between two variables x and y
- The correlation is limited
- What might we want to know?
 1. Possum data: predict head length from a measurement of total length
 2. Height data: understanding and quantifying heritability of height as a trait
 3. Powerlifting: compare the squat weight of an athlete to their peers of a similar weight
- Correlation does not help us for 1 and 3
 - ▶ Limited for 2: quantifies the linear relationship, but does not describe it
 - What is the expected difference in height between a son with father who is 170 cm tall, and a son with father who is 180 cm tall?

Statistical model

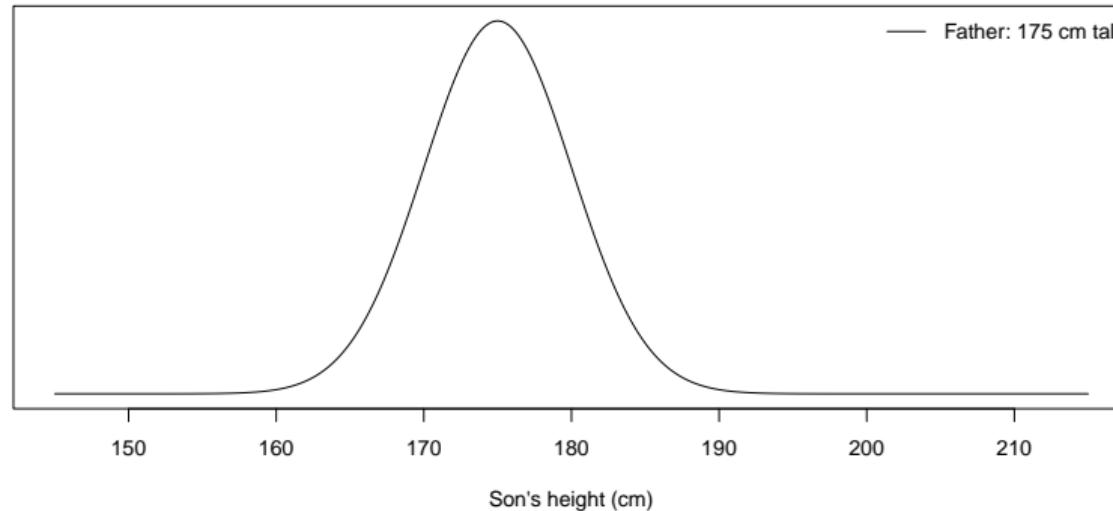
- To overcome these problems we will look to a statistical model
 - ▶ Extension of our previous models
- Explore relationship between continuous variables x and y
 - ▶ e.g. x is father's height, y is son's height
- The variable y is referred to as the outcome variable
 - ▶ Can also be called the response variable, or dependent variable
- The variable x is referred to as the predictor variable
 - ▶ Can also be called the explanatory variable, or independent variable
- The idea: the predictor variable helps us 'predict' the outcome variable

Statistical model

- Our description will make use of the father/son height example
 - ▶ Interest is in understanding the relationship the height of NZ male university students and their fathers
 - ▶ Sample is from (former) students in STAT 110
- Using probability to describe data
- Recall concept of conditional probability: $Pr(A|B)$
 - ▶ Here we are looking at a probability density for $y|x$
 - We have the height of a father (x) and son (y)
 - Given a father's height (x), we specify a model for son's height (y)
 - We will specify a normal model
- Look at it graphically

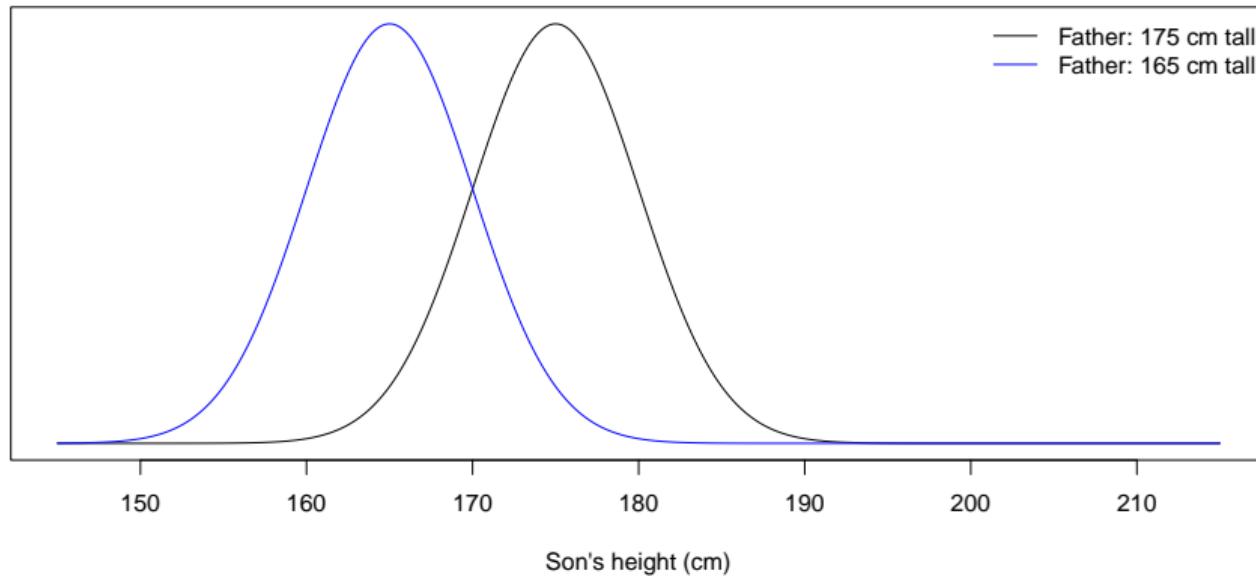
Statistical model

- Consider the subpopulation at particular value of x
 - ▶ e.g. sons with fathers who are 175 cm tall ($x = 175$)
 - ▶ Assume that son's height is normally distribution
 - For the sake of explanation: sons are expected to be the same height as their fathers



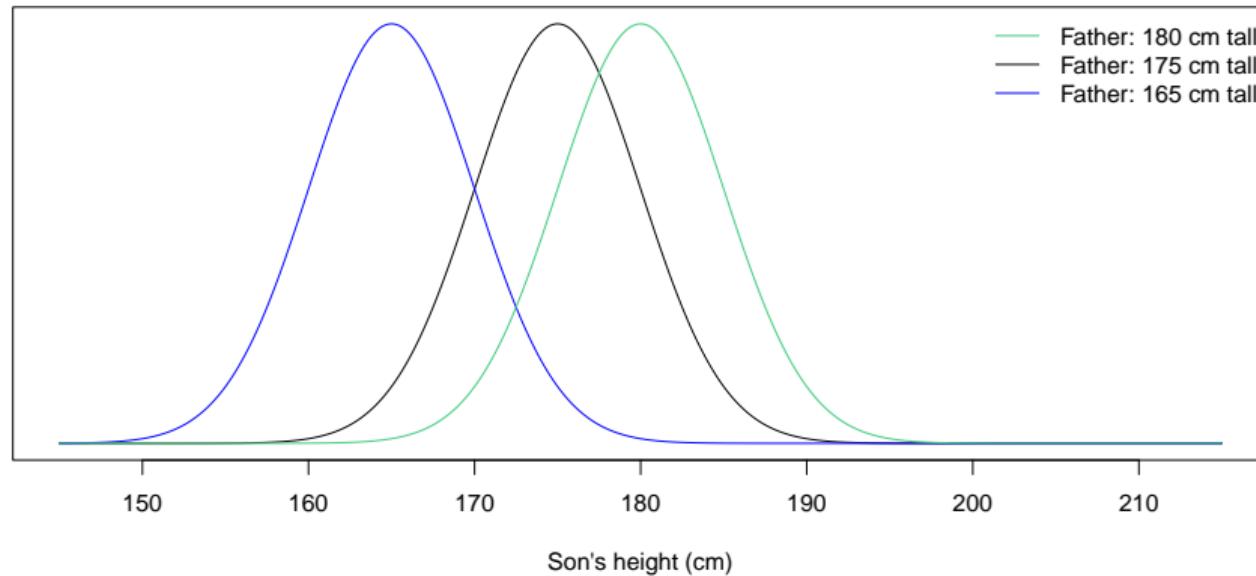
Statistical model

- Subpopulation at a given value of x : outcome variable is normally distributed
- For fathers who are 165 cm tall (blue)



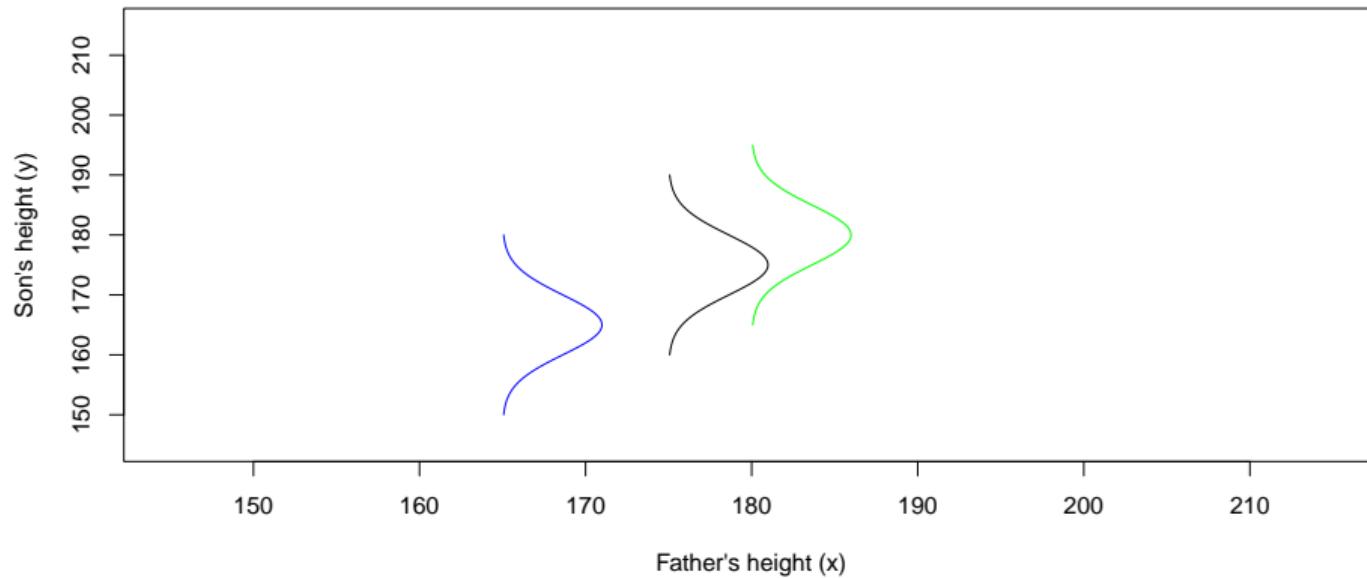
Statistical model

- Subpopulation at a given value of x : outcome variable is normally distributed
- For fathers who are 180 cm tall (green)



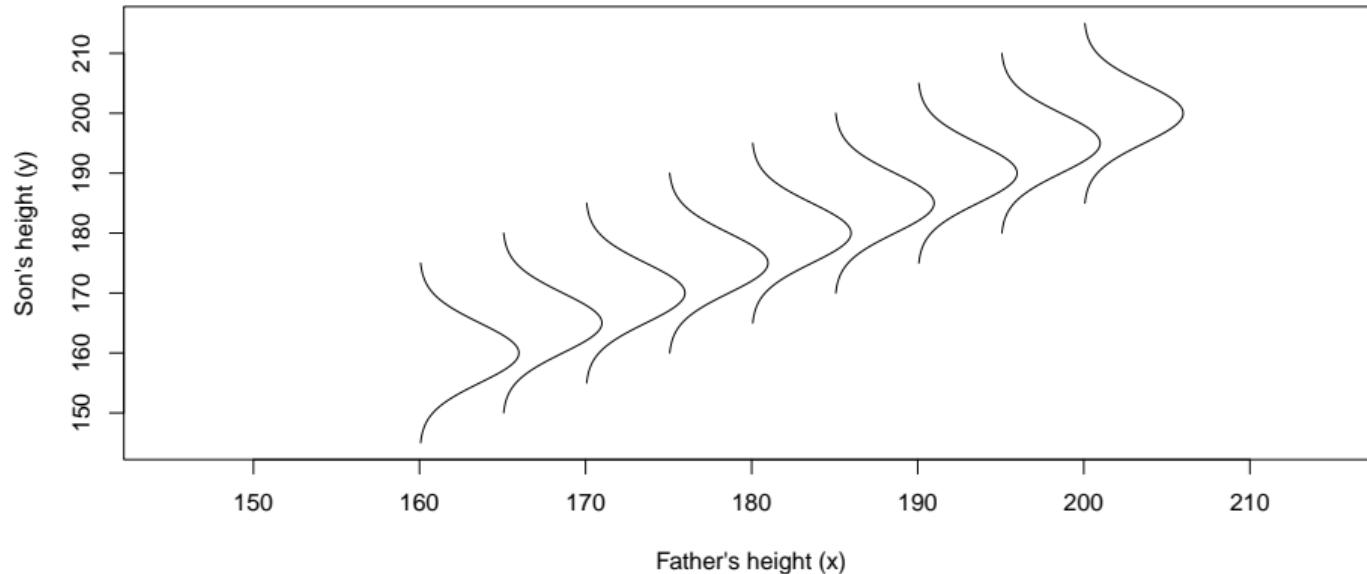
Turning it sideways

- Visualise it with outcome variable on y-axis, and predictor variable on x-axis
 - ▶ The same distributions are given below



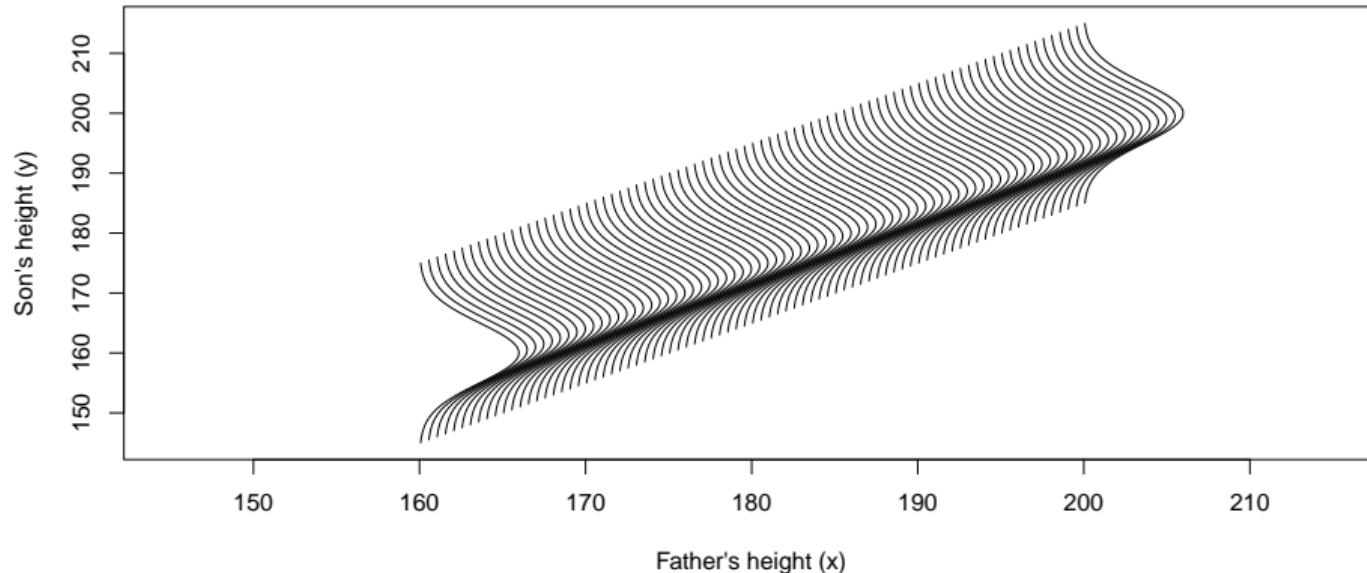
Turning it sideways

- Including some other values of x (father's height)



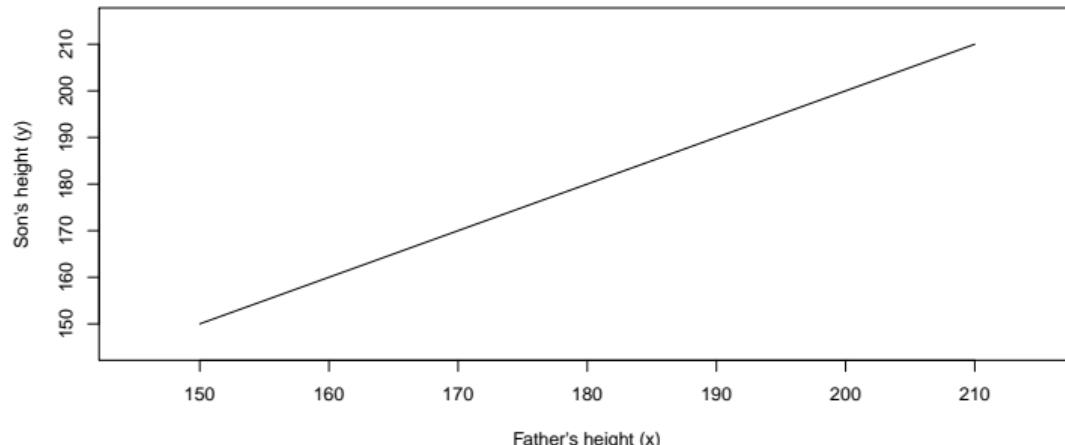
Turning it sideways

- Including even more values of x (father's height)



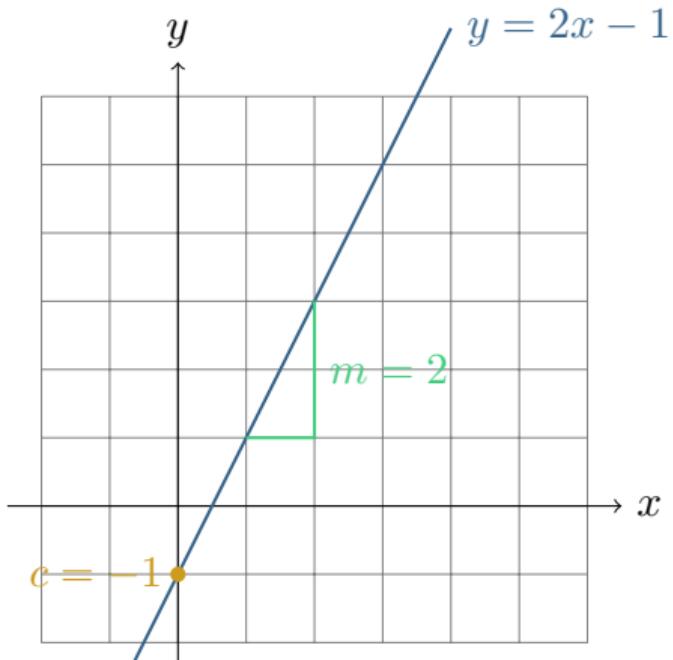
Linear regression

- The outcome variable, y , can be written in terms of two pieces:
 - ▶ outcome = mean response + error
- The mean response (what we expect) is assumed to vary with the predictor x
 - ▶ Expected height of a son is different if father is 165 cm vs father who is 180 cm
- We assume the mean response is a straight line
 - ▶ e.g. continuing the father and son height example, the mean response is



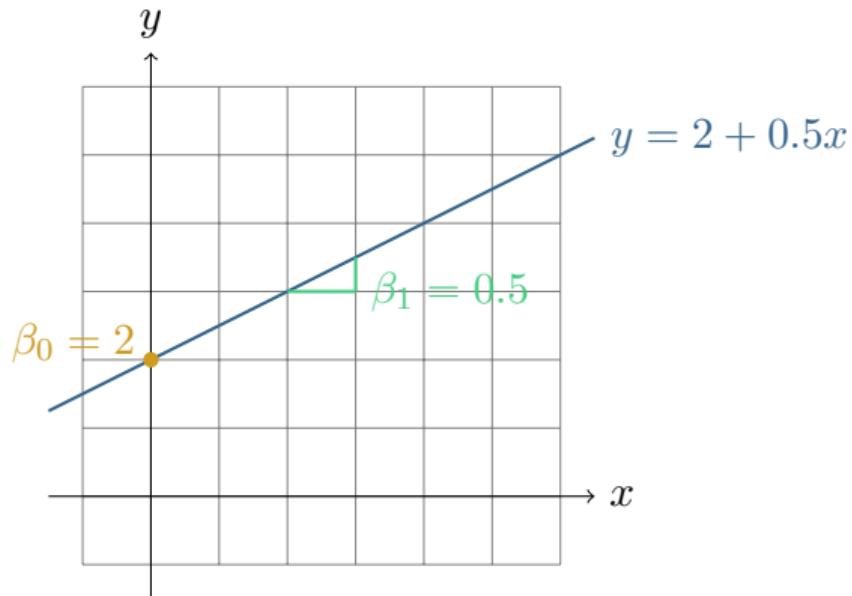
Revision: equation for a straight line

- Mathematical equation: $y = mx + c$
 - ▶ Intercept c : where it crosses the y -axis ($x = 0$)
 - ▶ Slope m



Revision: equation for a straight line

- We will use the equation: $\beta_0 + \beta_1 x$
 - ▶ Convention: use β_0 and β_1 in place of c and m
 - Intercept β_0 : where it crosses the y-axis ($x = 0$)
 - Slope β_1



Understanding the model: population level

- Putting this together we have:

$$\underbrace{y}_{\text{outcome}} = \underbrace{\beta_0 + \beta_1 x}_{\text{mean response}} + \underbrace{\varepsilon}_{\text{error}}$$

- The mean response is given by the straight line: $\mu_y = \beta_0 + \beta_1 x$
 - Gives us the expected value of y in the population for a given value of x
- The mean will be different for two different values of x
- For $x = 165$ cm:
 - Mean is: $\mu_y = \beta_0 + \beta_1 \times 165$
- For $x = 180$ cm:
 - Mean is: $\mu_y = \beta_0 + \beta_1 \times 180$

Interpretation

- What do β_0 and β_1 represent?
- The mean will be different for two different values of x
 - ▶ Mean is: $\mu_y = \beta_0 + \beta_1 x$
- For someone with a father one cm taller ($x + 1$), the mean response is
 - ▶ Mean is: $\mu_y = \beta_0 + \beta_1(x + 1) = \beta_0 + \beta_1x + \beta_1$
- β_1 is the difference between these
 - ▶ β_1 is the change in mean response when x increases by one unit
 - Change in the expected height of two male NZ university students whose fathers differ in height by 1 cm
- β_0 is the mean response when $x = 0$
 - ▶ May make no sense in many examples
 - Mean response for a son with a father of height 0 cm: physically impossible

From mean response to individual response

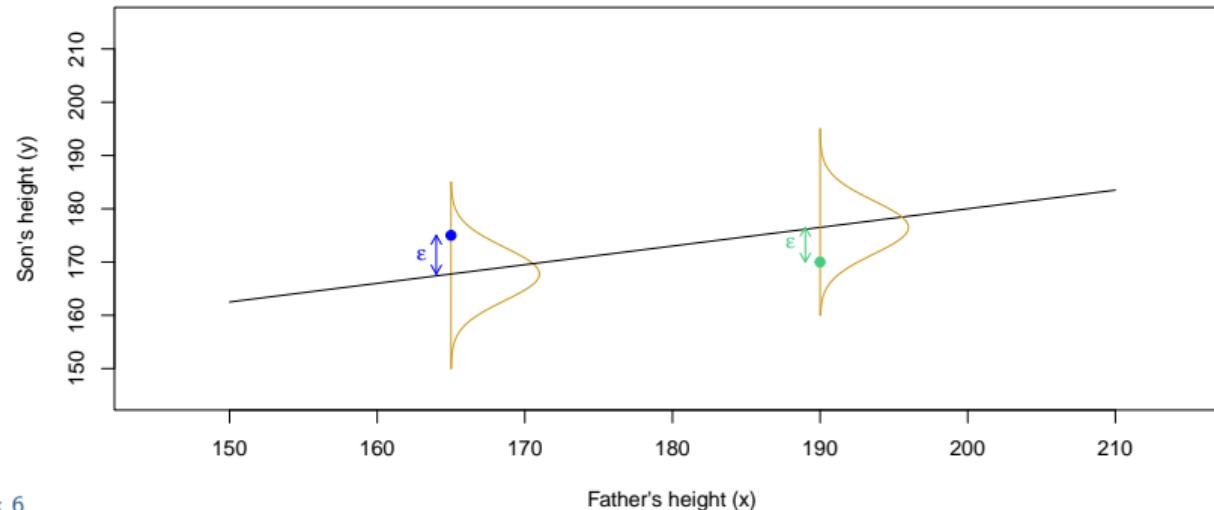
- The linear regression model is

$$y = \beta_0 + \beta_1 x + \varepsilon$$

- Error term ε (greek letter epsilon) describes how an individual response differs from the mean of their subpopulation
 - ▶ Subpopulation: all individuals in the population with the same value of x
- We assume that variation within a given subpopulation is normally distributed
 - ▶ ε is normally distributed with mean 0 and variance σ_ε^2
 - σ_ε tells us how variable individual observations are within their subpopulation

Visualising subpopulation

- Suppose that the true regression model for height is $y = 110 + 0.35x + \varepsilon$
 - Mean response (black line)
 - Normal model for the errors (gold)
 - Individual with $y = 175$ and $x = 165$ (blue point)
 - Individual with $y = 170$ and $x = 190$ (green point)

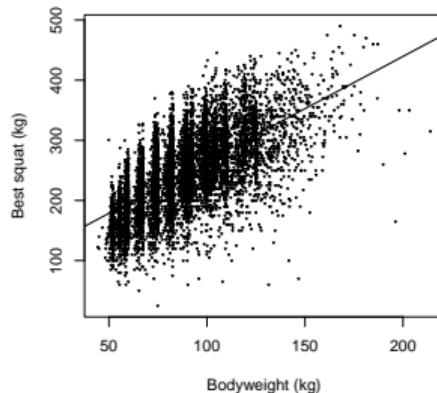
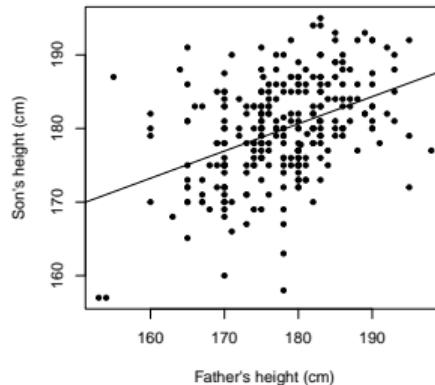
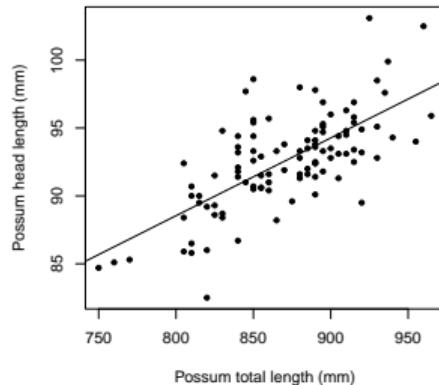


Statistical model: data

- The linear regression model is

$$y = \beta_0 + \beta_1 x + \varepsilon$$

- The errors mean that data will not fall exactly on the line
 - Like the data we have!



It's quiz time!

- Suppose that the true regression model for height is

$$y = 110 + 0.35x + \varepsilon$$

- Decide whether the following statements are true or false:
 - Consider the subpopulation of all students with fathers of height $x = 200$ cm. The mean height of those students is 180 cm.
 - On average, students with fathers of height $x = 201$ cm are 0.35 cm taller than students with fathers of height $x = 200$ cm.
 - All students with fathers of height $x = 190$ cm are taller than all students with fathers of height $x = 170$ cm.
 - Students with fathers of height $x = 0$ cm are 110 cm tall on average

Summary

- Introduced a statistical model for the relationship between x and y
 - ▶ Outcome variable, y
 - ▶ Predictor variable, x
 - ▶ For a given value of x , y is assumed to be normally distributed
- Understand the linear regression model
 - ▶ Mean response
 - ▶ Error
 - ▶ Interpretation
- Looking forward: how do we fit a linear regression to data?

STAT 110: Week 7

University of Otago

Outline

- Previous
 - ▶ Model for linear regression
 - ▶ $y = \beta_0 + \beta_1 x + \varepsilon$
- Today:
 - ▶ Fitting the model
 - Estimating β_0 and β_1
 - Fitted model
 - Residuals

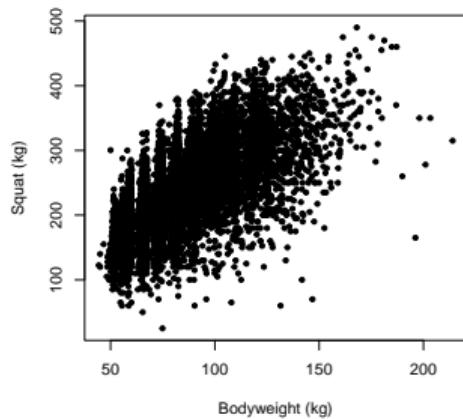
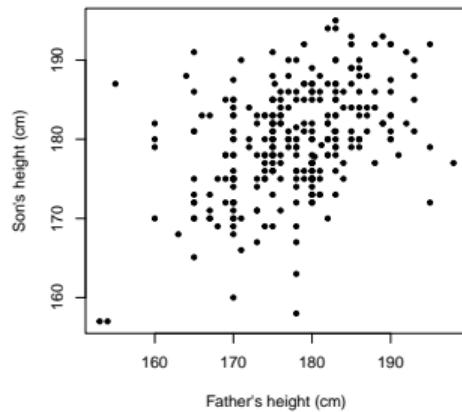
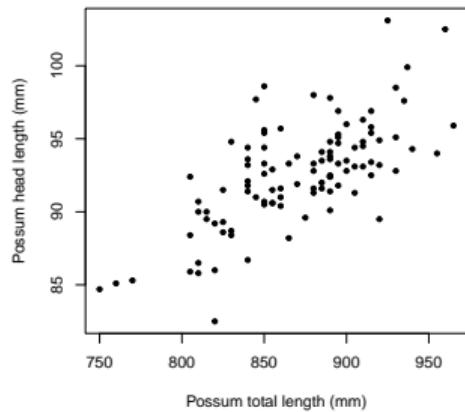
Recall: motivating data

- The size of brushtail possums
 - ▶ Exploring relationship between total length (mm) and head length (mm)
- Height of STAT 110 students
 - ▶ Compare father's height (cm) and son's height (cm)
- Squat weight of international power lifters
 - ▶ Look at the relationship between body weight (kg) and max squat weight (kg)

Recall: importing data into R

- Import the data into R

```
possum = read.csv('possum.csv')
height = read.csv('height.csv')
powerlift = read.csv('powerlift.csv')
```



Fitting a regression model

- The (simple) linear regression model is

$$y = \underbrace{\beta_0 + \beta_1 x}_{\text{mean response}} + \varepsilon$$

- β_0 and β_1 are parameters
 - Estimate parameters (population) with statistics (sample)
 - What statistics could we use to estimate β_0 and β_1 ?

Fitting a regression model

- The (simple) linear regression model is

$$y = \underbrace{\beta_0 + \beta_1 x}_{\text{mean response}} + \varepsilon$$

- β_0 and β_1 are parameters
 - Estimate parameters (population) with statistics (sample)
 - What statistics could we use to estimate β_0 and β_1 ?
 - We could guess by eye: use paper, pencil and ruler (or electronic equivalents)
 - Later in the lecture: find general approach for estimating β_0 and β_1
- For now: assume we have some way to find estimates $\hat{\beta}_0$ and $\hat{\beta}_1$
- Work through using the possum data to illustrate concepts

Fitted model

- The (simple) linear regression model is

$$y = \underbrace{\beta_0 + \beta_1 x}_{\text{mean response}} + \varepsilon$$

- Once we have estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ we can write the fitted model

$$\hat{\mu}_y = \hat{\beta}_0 + \hat{\beta}_1 x$$

- The fitted model is commonly written as

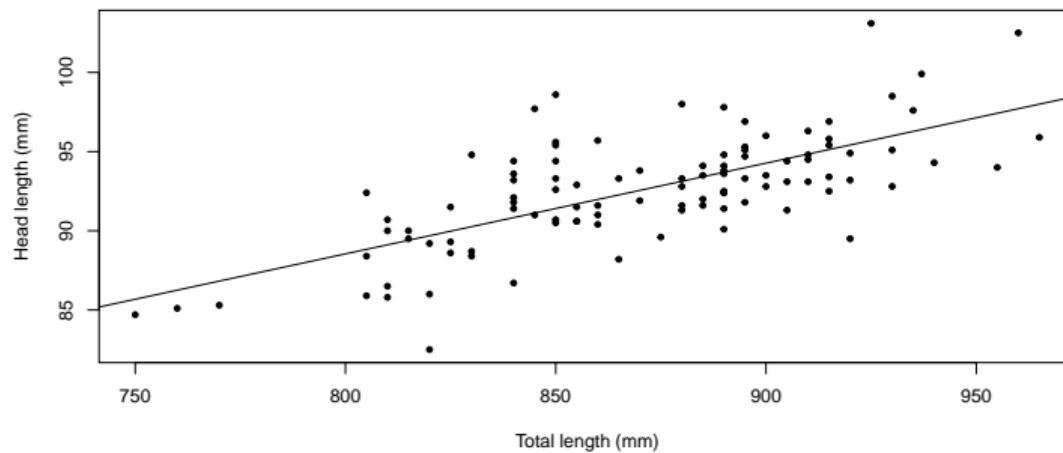
$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

- The fitted model gives the estimate of the mean at a given x value

Fitted model: possum data

- Use estimates $\hat{\beta}_0 = 42.7$ and $\hat{\beta}_1 = 0.057$
- Fitted model is

$$\hat{y} = 42.7 + 0.057x$$



Residuals

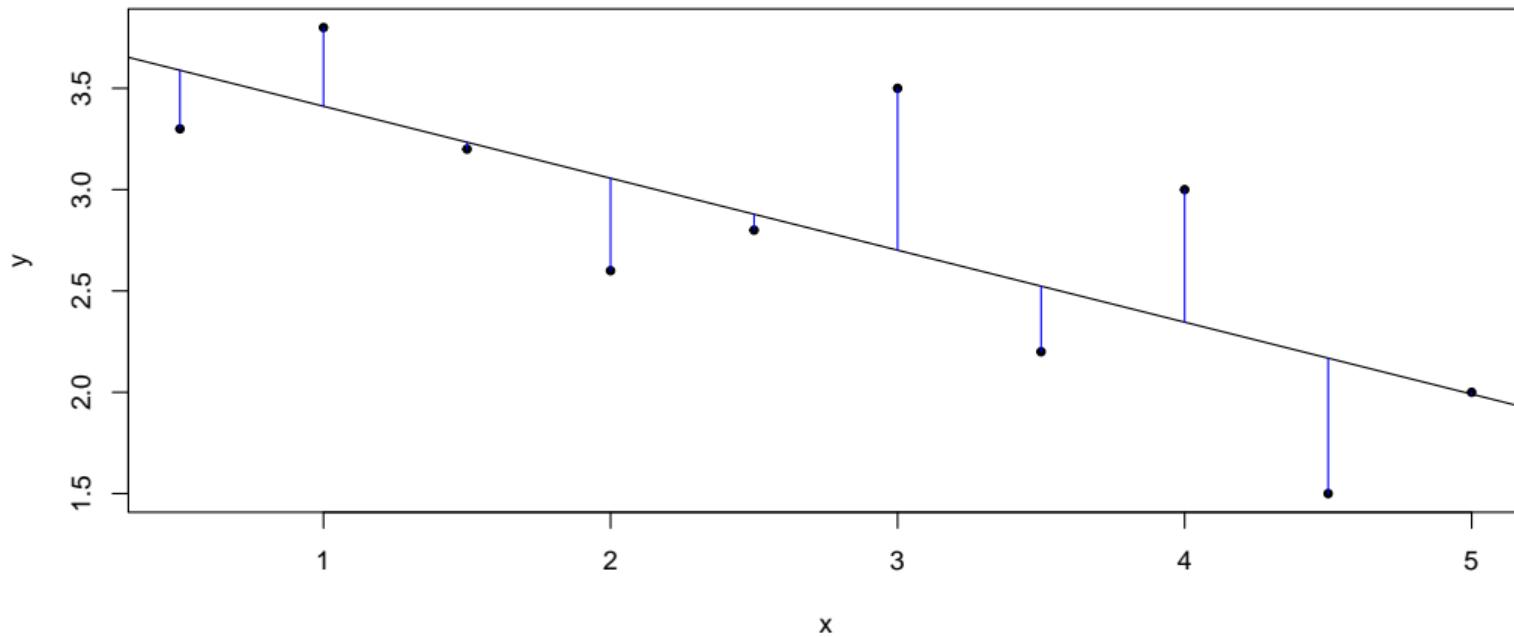
- The statistical model can be expressed as
 - ▶ observation = mean response + error
- After fitting the model, we have
 - ▶ observation = fitted model + residual
- The residual $\hat{\varepsilon}$ is our best guess (estimate) of the error ε
 - ▶ It is the difference between the observation (y) and the mean response (\hat{y})

$$\hat{\varepsilon} = y - \hat{y}$$

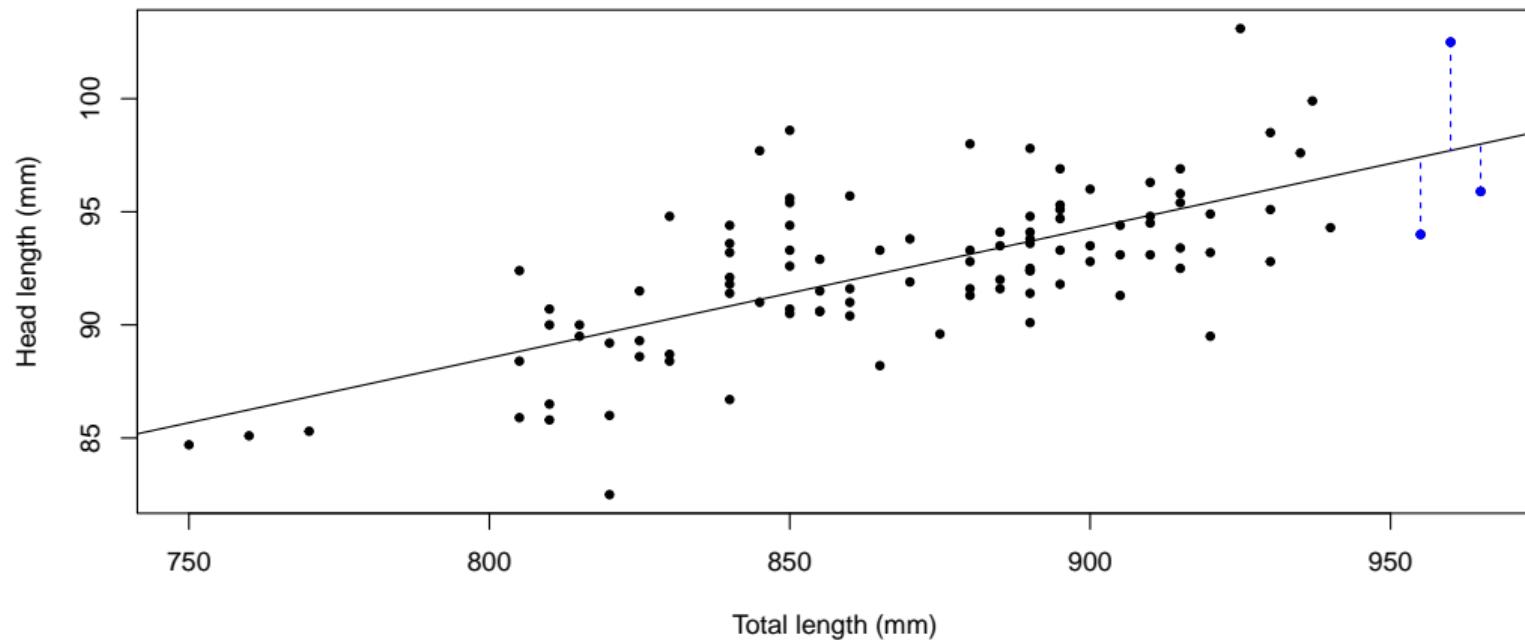
- We often index by i : for the i th observation (x_i, y_i) the residual is

$$\hat{\varepsilon}_i = y_i - \hat{y}_i$$

Residuals: blue lines



Residuals: possum data (three points in blue)



How do we fit the model?

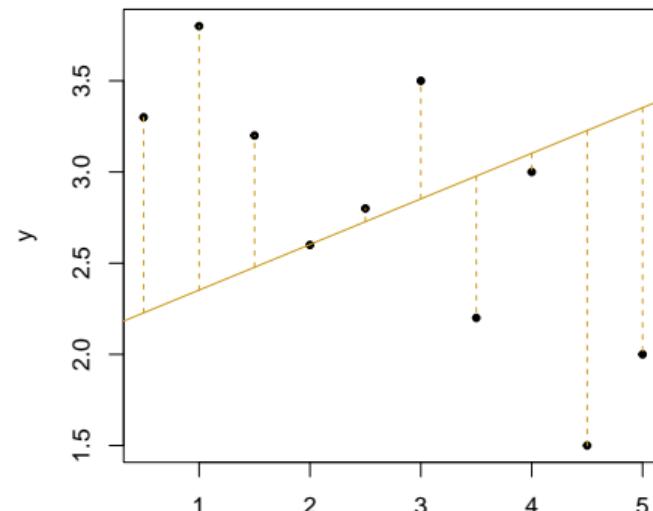
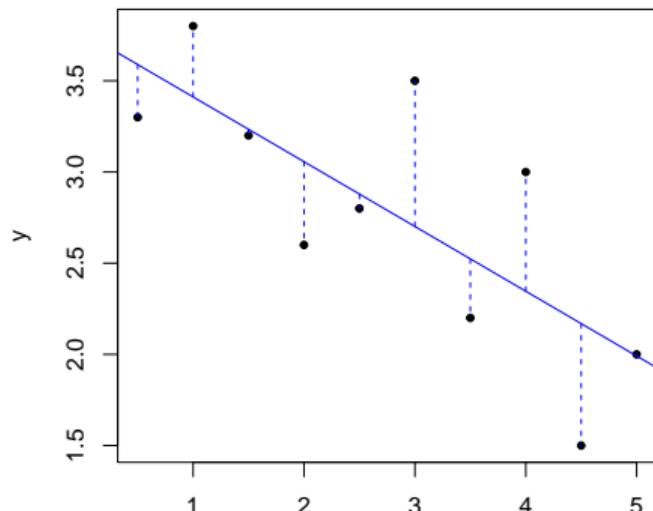
- The (simple) linear regression model is

$$y = \beta_0 + \beta_1 x + \varepsilon$$

- Estimate parameters β_0 and β_1
 - ▶ Find β_0 and β_1 that give the ‘best’ description of relationship between x and y
- Suppose we had a choice between two possible fitted models
 1. One of them has many large residuals (large positive and large negative residuals)
 2. The other one has mostly small residuals (small positive and small negative residuals)
- Which is better?
 - ▶ Look graphically

Graphical representation

- Same data, two possible fitted models
 - ▶ One with larger residuals (magnitude): gold
 - ▶ One with smaller residuals (magnitude): blue
- Which describes the relationship between x and y better?

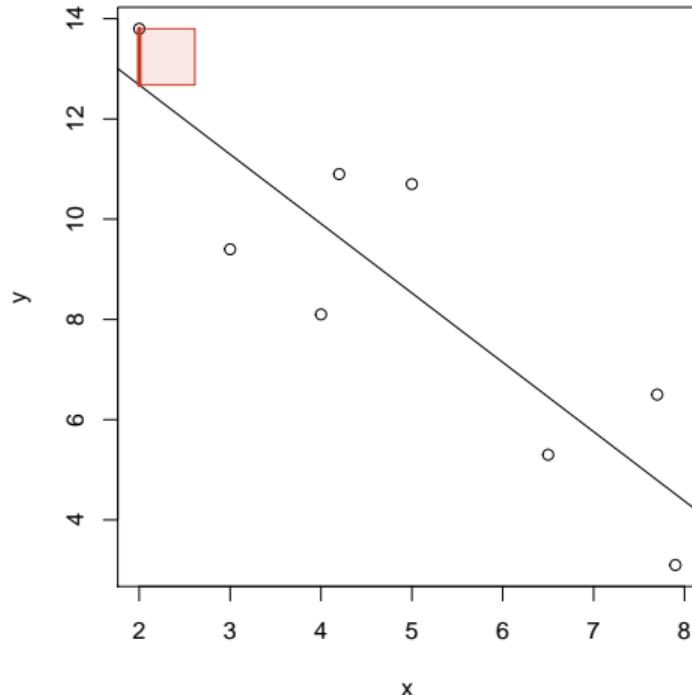


Least squares

- We want the (magnitude of the) residuals to be as small as possible
- We will find estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ using the method of least squares
 - ▶ Find the values $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimize the sum of squared residuals
- Explain the process graphically

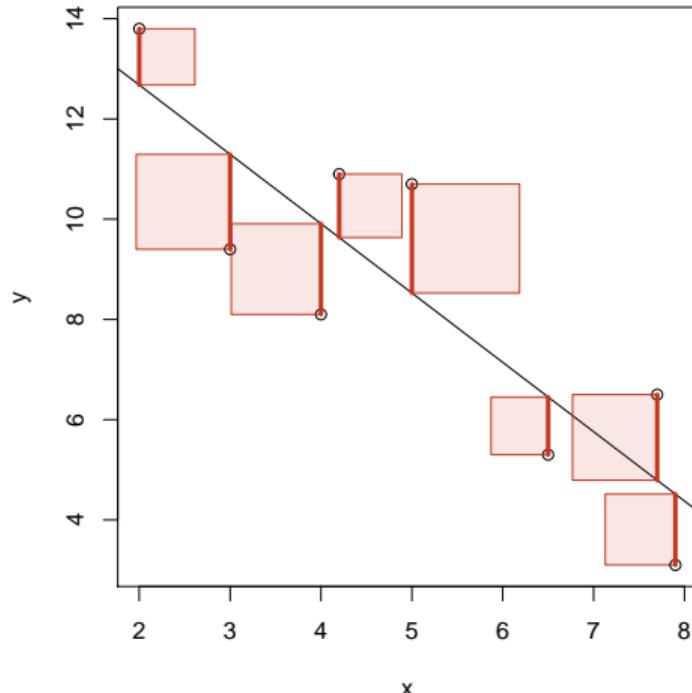
Least squares

- We can visualise the squared residual by drawing a square!
 - ▶ Squared residual is the area of red square



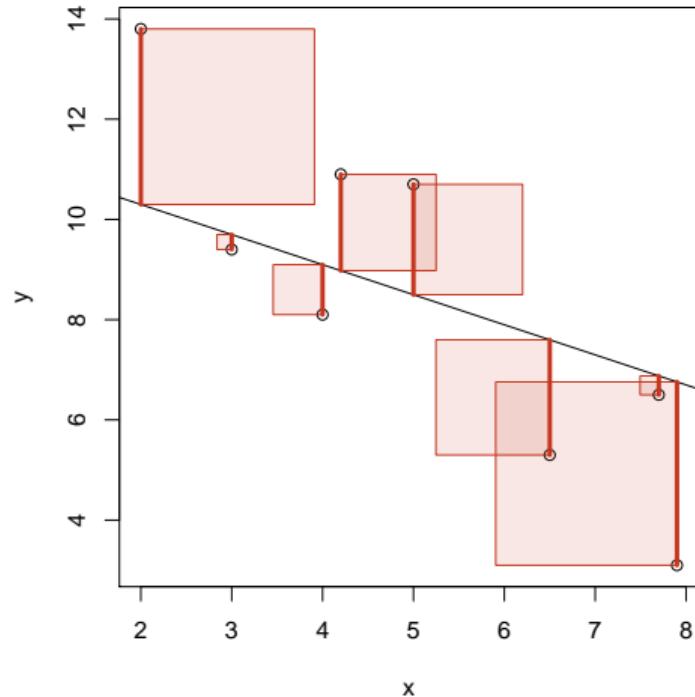
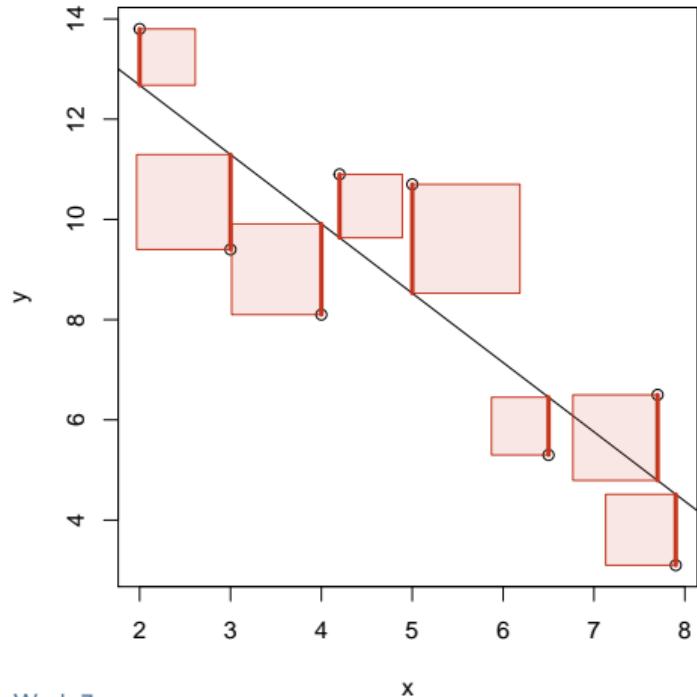
Least squares

- The sum of squared residuals
 - ▶ Combined area of the red squares



Least squares

- Minimise the sum of squared residuals (minimise combined area)
 - Left plot: better fit (to the same data)



Least squares

- The sum of squared residuals:

$$\begin{aligned}\sum_{i=1}^n \hat{\varepsilon}_i^2 &= \sum_{i=1}^n (y_i - \hat{y}_i)^2 \\ &= \sum_{i=1}^n (y_i - [\hat{\beta}_0 + \hat{\beta}_1 x_i])^2\end{aligned}$$

- ▶ Note: $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$
- Find $\hat{\beta}_0$ and $\hat{\beta}_1$ that make $\sum \hat{\varepsilon}_i^2$ as small as possible

Parameter estimates

- We can use calculus to find estimates
 - ▶ $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimise sum of square residuals

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{s_y}{s_x} r$$

- ▶ s_y : sample standard deviation of outcome y
- ▶ s_x : sample standard deviation of predictor x
- ▶ r : sample correlation between x and y

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

- Details of how to find these: outside the scope of the course

In R

- We can find the least squares estimates using R
- The R code is

```
lm(y ~ x)
```

- Look at each piece in turn:
 - ▶ `lm`: function for fitting a linear model
 - ▶ `y`: outcome variable
 - ▶ `x`: predictor variable
 - ▶ `~`: thought of as 'is modelled by'
 - ▶ `lm(y ~ x)`: is saying that we are fitting a linear model where the outcome variable y is modelled in terms of the predictor variable x

Fitting the possum data

```
m_possum = lm(possum$head_1 ~ possum$total_1) # assigned the output to object m_possum
summary(m_possum) # shows a summary of the results

##
## Call:
## lm(formula = possum$head_1 ~ possum$total_1)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -7.188 -1.534 -0.334  1.279  7.397
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 42.70979   5.17281   8.26  5.7e-13 ***
## possum$total_1  0.05729   0.00593   9.66  4.7e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.6 on 102 degrees of freedom
## Multiple R-squared:  0.478, Adjusted R-squared:  0.472
## F-statistic: 93.3 on 1 and 102 DF,  p-value: 4.68e-16
```

Estimates in R

- The estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ are given in column headed Estimate
 - ▶ $\hat{\beta}_0 = 42.71$
 - ▶ $\hat{\beta}_1 = 0.057$
- R labels the estimates in terms of the variable names
 - ▶ (Intercept)
 - ▶ possum\$total.l

Detour: data option in lm

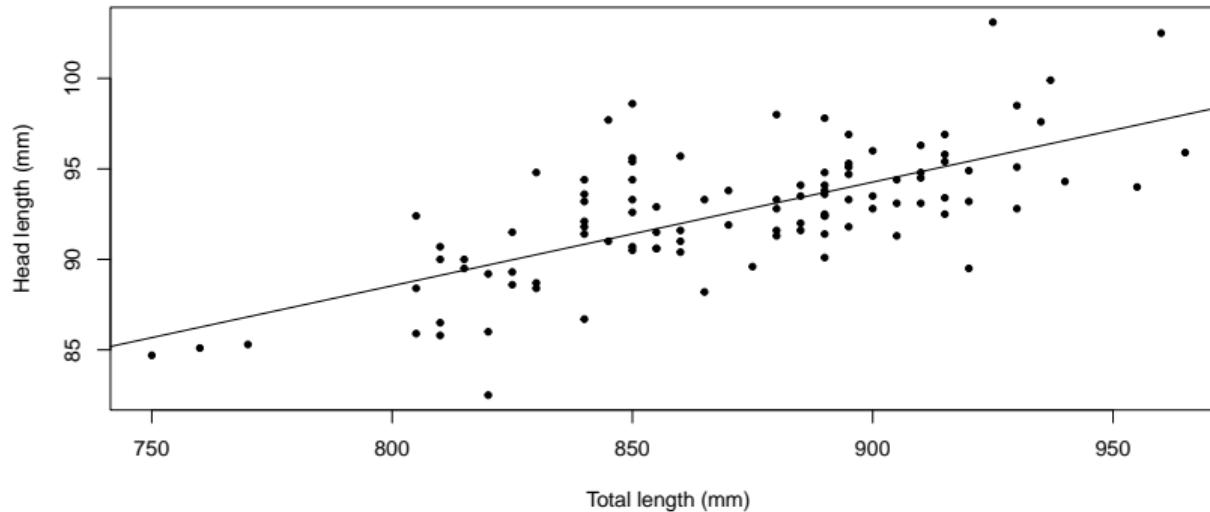
- The lm function includes a data option that can make specification easier
- Separate the variable (e.g. head_l) from the data frame object (possum)
- The code is

```
m_possum2 = lm(head_l ~ total_l, data = possum)
```

- This is fitting the same model as in the slide above

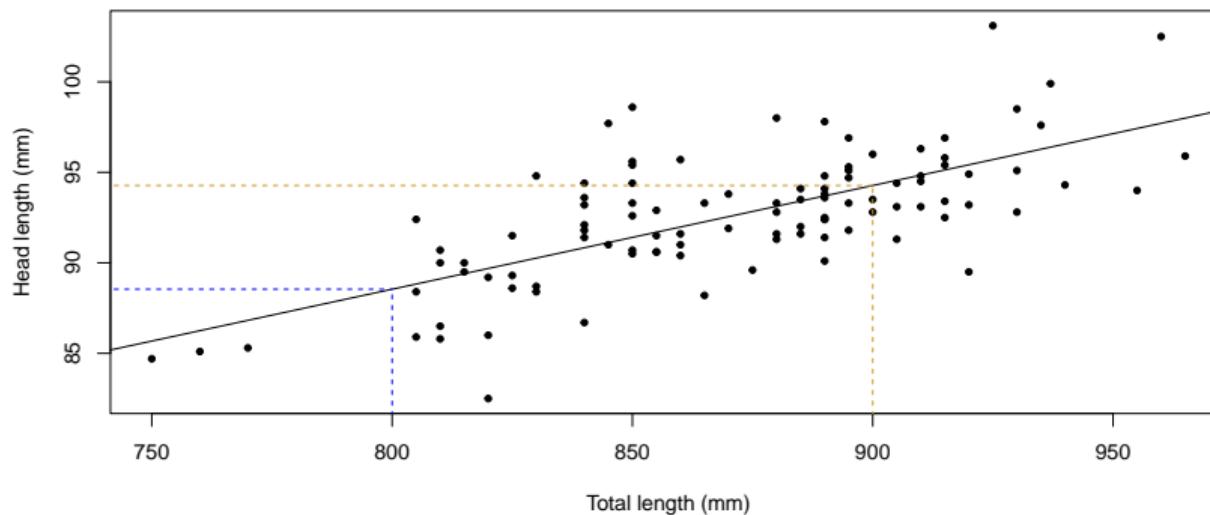
Fitted model: possum data

- The fitted model is $\hat{y} = 42.71 + 0.057x$
 - Recall: y is head length, x is total length
 - We could also write: $\widehat{\text{head}} = 42.71 + 0.057 \text{total}$



Fitted model: possum data

- Fitted model is $\hat{y} = 42.7 + 0.057x$
 - ▶ For $x = 800$ we have $\hat{y} = 42.7 + 0.057 \times 800 = 88.5$
 - ▶ For $x = 900$ we have $\hat{y} = 42.7 + 0.057 \times 900 = 94.3$



Interpretation

- Fitted model $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$
 - ▶ For the possum data: $\hat{y} = 42.7 + 0.057x$
- Our interest is $\hat{\beta}_1$:
 - ▶ We estimate that the average head length of a possum will increase by 0.057 mm for a 1 mm increase in total length.
- This is a comparison of two subpopulations
 - ▶ If we compare possums whose total length is x mm to possums whose total length is $x + 1$ mm, the estimated increase in their expected (or mean) head length is 0.057 mm.
- $\hat{\beta}_0$ is the estimated mean head length of possums with total length 0 mm
 - ▶ Makes no biological sense
 - ▶ Do not interpret in this case

Summary

- Fitting a linear regression model
 - ▶ Fitted values: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$
 - ▶ Residuals: $\hat{\epsilon} = y - \hat{y}$
- Method of least squares
 - ▶ Minimise the sum of squared residuals
 - ▶ Fit the model using `lm` in R: `lm(y ~ x)`

Outline

- Previous:
 - ▶ Fitting a statistical model
 - ▶ Method of least squares
- Today:
 - ▶ Assumptions underlying linear regression
 - What are the assumptions?
 - How do we check the assumptions?

Motivation

- Exploring relationship between total length (mm) and head length (mm) of brushtail possums
- Recall: fitting linear model

```
m_possum = lm(head_l ~ total_l, data = possum) # possum data
```

- Linear regression model allows us to:
 - ▶ Estimate the effect of x (total length) on y (head length)
 - ▶ Estimate the mean response of y (head length) given x (total length)
 - E.g. estimate mean head length of possums that have total length $x = 820$ mm
- Problem: the model relies on assumptions
 - ▶ Interpretations and conclusions may be invalid if assumptions are badly wrong
- We need to test the model assumptions (so far as possible)

Assumptions for Simple Linear Regression

- Recall that the linear regression model is

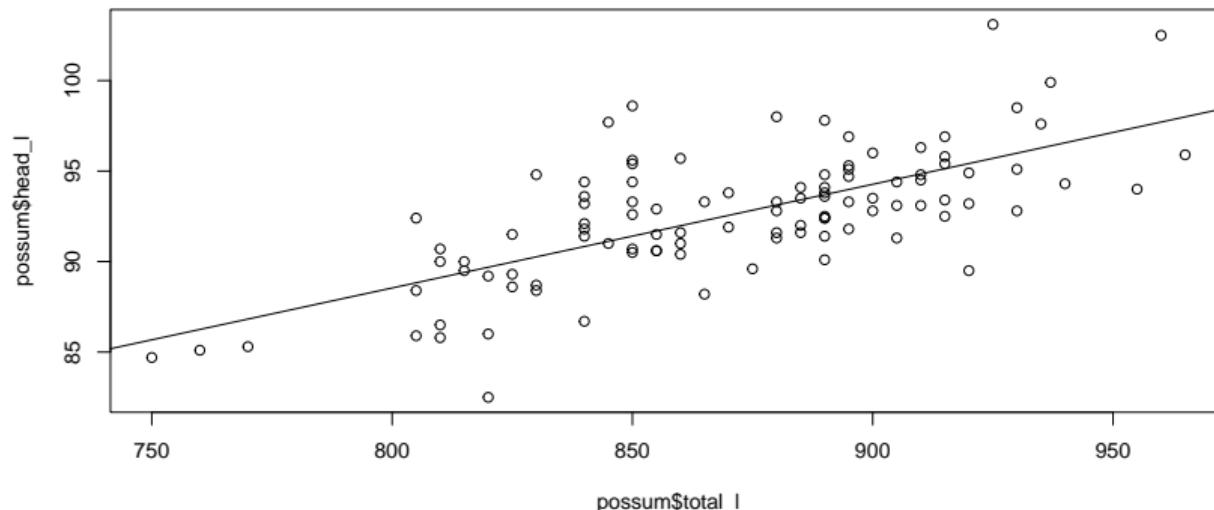
$$y = \underbrace{\beta_0 + \beta_1 x}_{\mu_y} + \varepsilon$$

- The underlying assumptions are:
 - Linearity:** The mean response μ_y is described by a straight line
 - Independence:** The errors $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n$ are independent
 - Normality:** The error terms ε are normally distributed
 - Equal variance:** The error terms all have the same variance, σ_ε^2 ('homoscedastic')
- These are often remembered using the mnemonic **LINE**.

Tools for checking assumptions

- Fitted line plot: compare the observed data to the fitted model
 - ▶ Useful, but not extensively used for checking assumptions

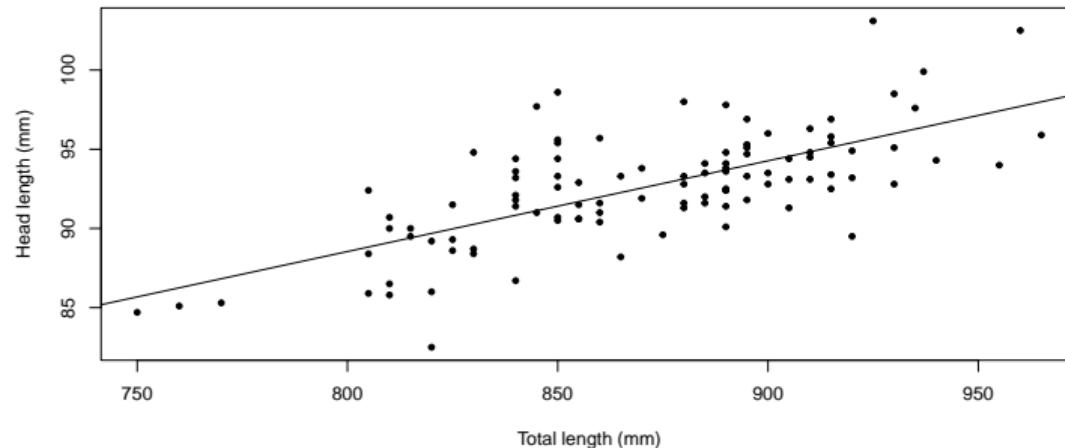
```
plot(possum$total_l, possum$head_l) # plot(x,y): x gives x values, y gives y values  
abline(m_possum) # draws the fitted regression line
```



Detour: plotting in R

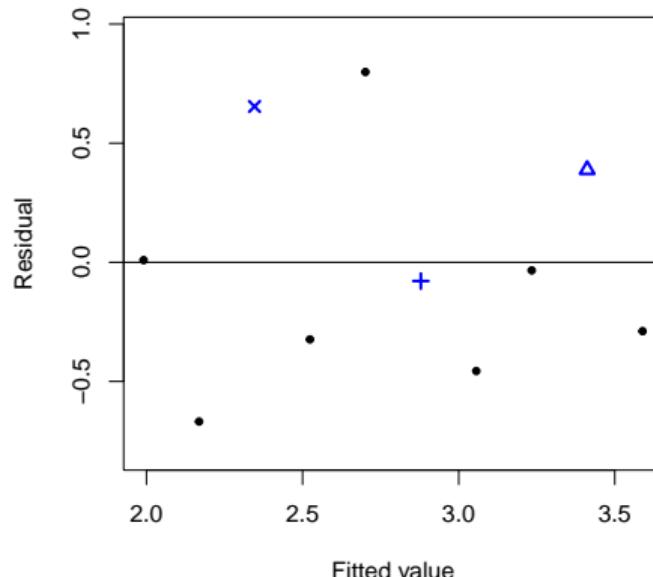
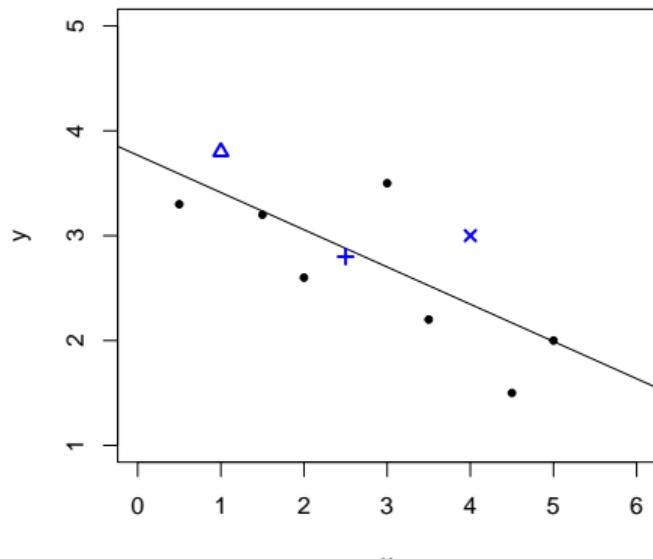
- Show code for 'default' plots: default points, colours, axis labels, etc
 - ▶ All that is needed for this course (STAT 260 explores plotting and visualisation of data)
- For interest: present same plot as above with some modifications

```
plot(possum$total_l, possum$head_l, pch = 20, xlab = "Total length (mm)",  
      ylab = "Head length (mm)")  
abline(m_possum)
```



Residual plots

- It is more common to use a residual plot
 - ▶ Residuals $\hat{\varepsilon}$ are on the y-axis
 - Recall: $\hat{\varepsilon} = y - \hat{y}$
- Look at a small example

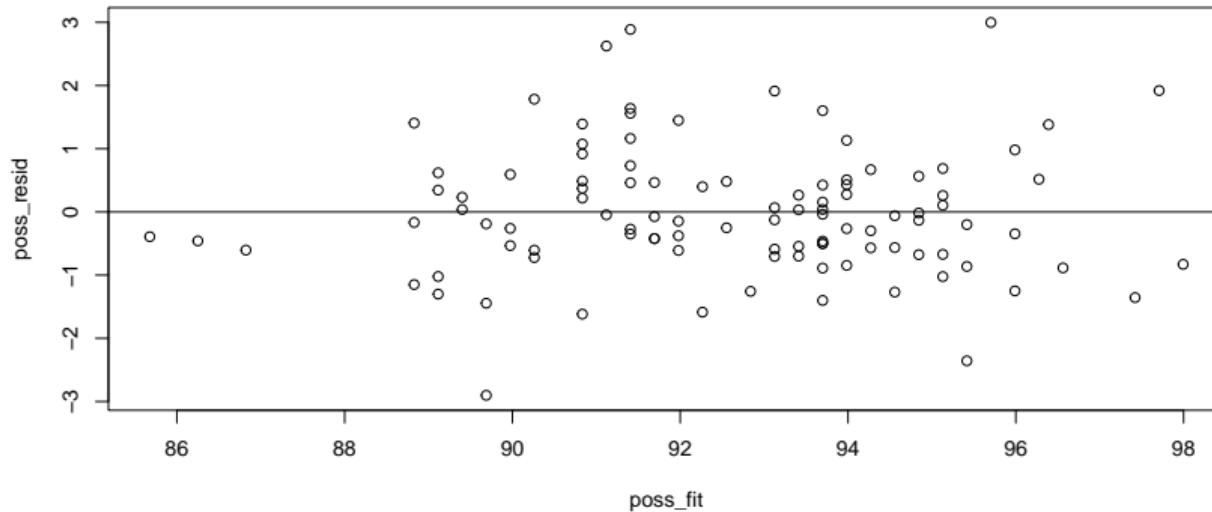


More on residuals: $\hat{\varepsilon} = y - \hat{y}$

- The residual is $\hat{\varepsilon} = y - \hat{\beta}_0 - \hat{\beta}_1 x$
- Residuals are estimates of error terms (ε)
 - ▶ Can be used to check assumptions about error terms (ε)
- The residual $\hat{\varepsilon}$ is often called a raw residual
 - ▶ Standardised or studentised residuals are often preferred
 - We will use studentised residuals in this course
 - ▶ What are studentised (or standardised) residuals?
 - Transformed to have standard deviation ≈ 1
 - (Mathematical) details are beyond the scope of the course
 - ▶ Find them in R using function `rstudent`
 - e.g. for model object `m_possum` we find studentised residuals using `rstudent(m_possum)`

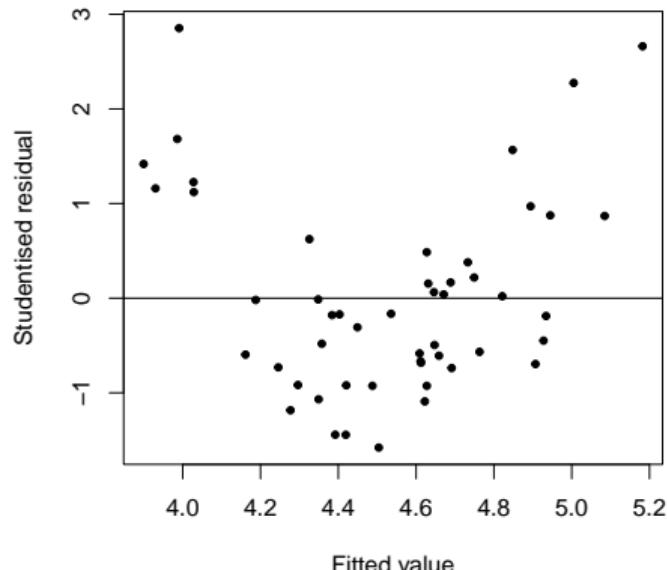
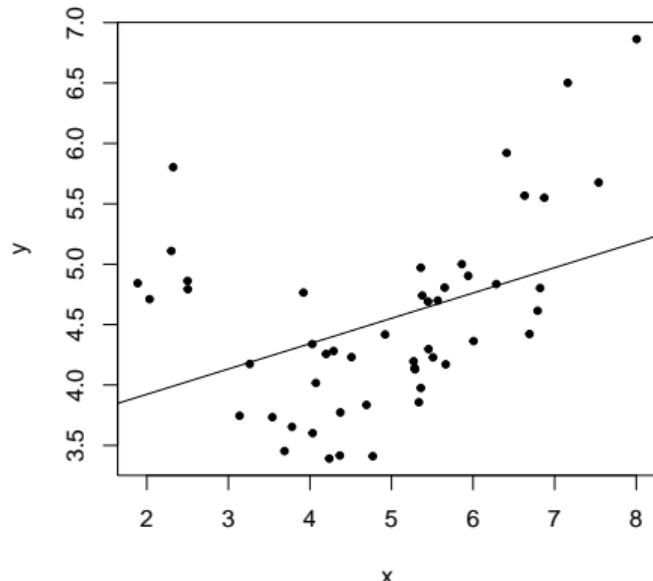
Plotting residuals in R

```
poss_fit = fitted(m_possum) # finds the fitted values of the model m_possum  
poss_resid = rstudent(m_possum) # finds the studentized residuals of the model m_possum  
plot(poss_fit, poss_resid) # plots residuals against fitted values  
abline(h=0) # draws a horizontal line at 0
```



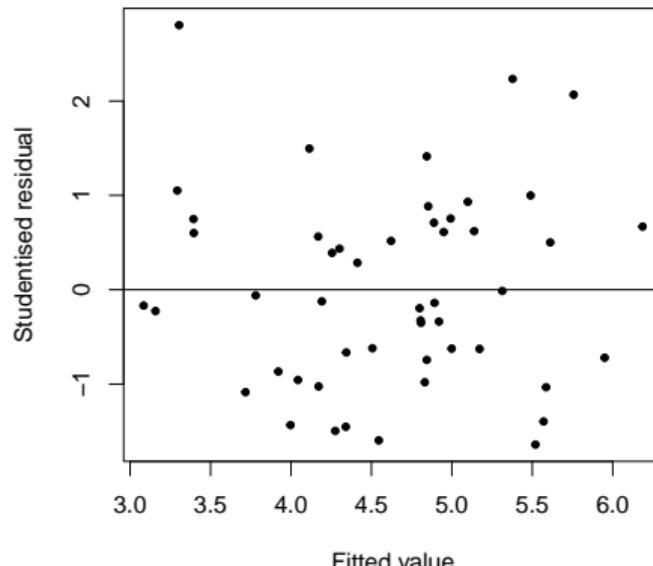
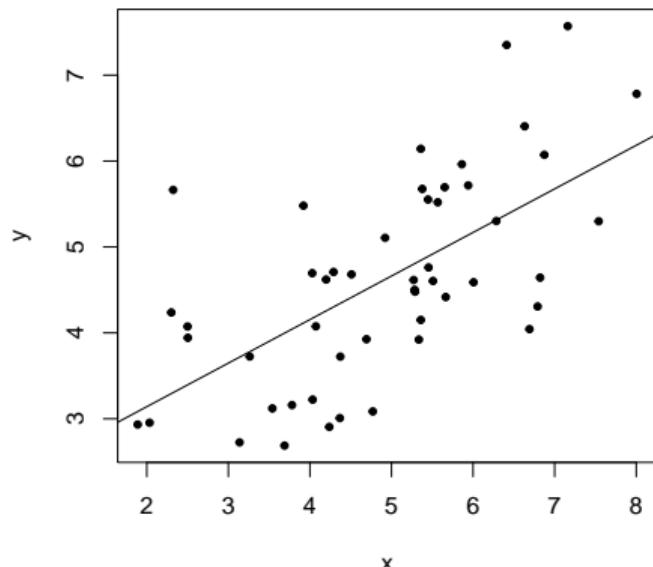
Checking the linearity assumption

- Looking for clear departure for linearity in trend of data.
 - ▶ Look for patterns in plot of residuals against fitted values
- Plots below illustrate failure of linearity assumption (bad)



Checking the linearity assumption

- Looking for clear departure for linearity in trend of data.
 - ▶ Look for patterns in plot of residuals against fitted values
- Plots below: no evidence of failure of linearity assumption (good)



The independence assumption

- Independence assumption: errors $\varepsilon_1, \dots, \varepsilon_n$ are independent
- What does it mean that errors ε_1 and ε_2 are independent?
 - ▶ Knowing ε_1 tells us nothing about ε_2
 - $\varepsilon_i = y_i - (\beta_0 + \beta_1 x_i)$
- For the possum example, independence means
 - ▶ Knowing how much above average one possum's head length is, gives no information about how far above average another possum's head length is.

Checking the independence assumption

- In general: difficult to assess
 - ▶ We are unable to check it by looking at fitted line or residual plots.
- In certain situations, we may be able to check it
 - ▶ If the data are collected in time (time series)
 - Expect observations close together in time to be correlated
 - ▶ If the data are collected in space (spatial data)
 - Expect observations close together in space to be correlated
 - ▶ If there are multiple measurements from each participant (repeated measures)
 - Expect observations from a given participant to be correlated
- We can look at more complex statistical models for each of the cases above
 - ▶ Outside the scope of this course

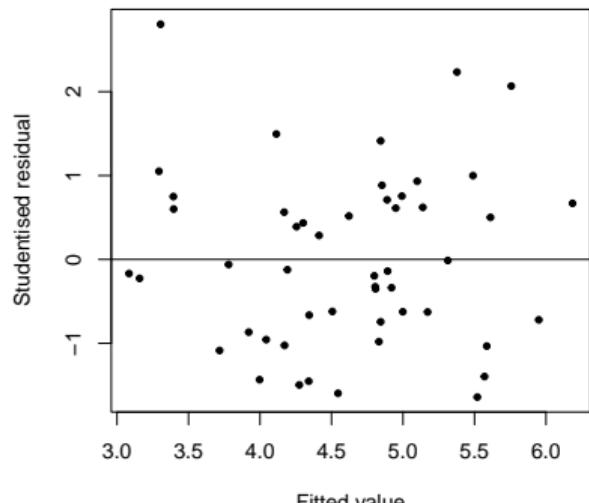
Checking the normality assumption

- Assumption: errors ε are normally distributed
- The importance of the normality assumption depends on sample size
 - ▶ Sample size small: important, but hard to check
 - ▶ As sample size increases (say $n > 50$) it becomes increasingly less important
 - Looking for large violations of normality
 - Are there one (or more) extreme values: outliers
- We assess outliers using the residual plot

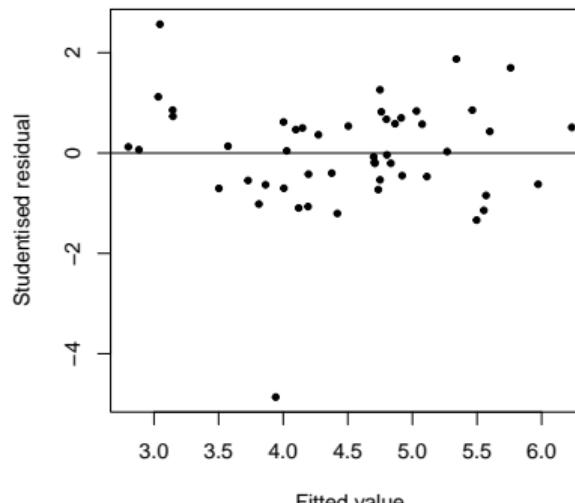
Checking the normality assumption

- Studentized residuals should be approximately normal with standard deviation 1:
 - ▶ Most (approx 95%) within ± 2
 - ▶ Nearly all ($> 99\%$) within ± 3
 - ▶ Values exceeding ± 4 are unusual

No apparent outliers



Apparent outlier

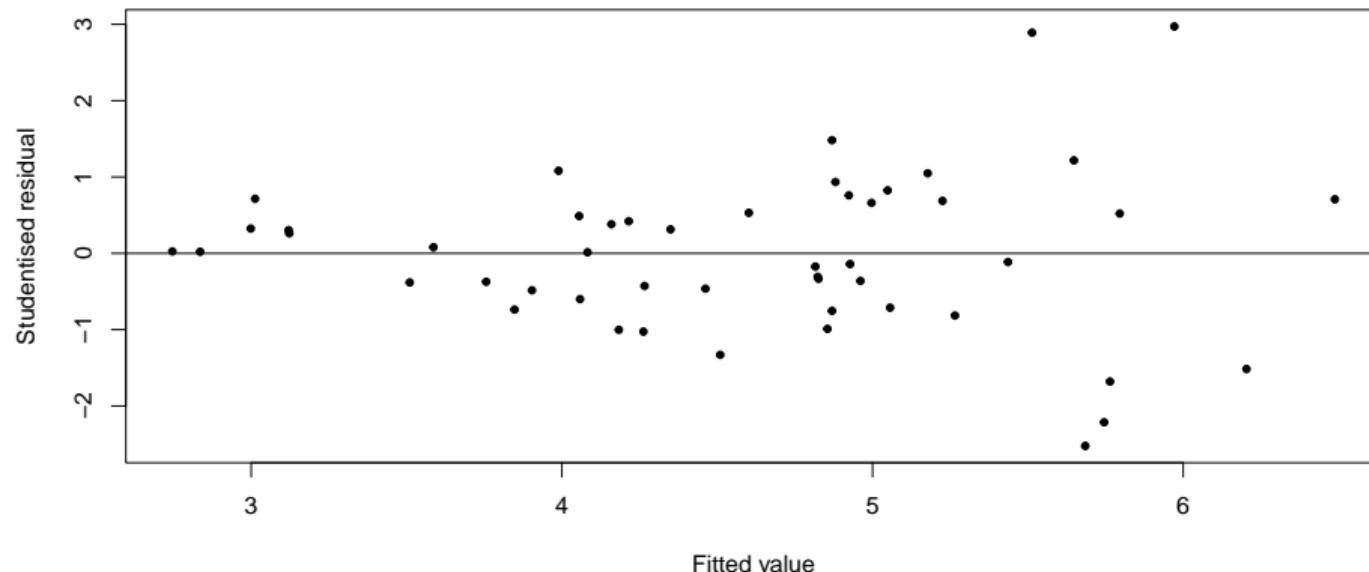


Checking equal variance assumption (homoscedasticity)

- Assumption: error terms $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n$ have the same variance
 - ▶ The magnitude of spread of data about regression line should not change too much with x .
- In contrast, if (say) variance of error terms increases with x
 - ▶ We would expect to see data more dispersion as x increases.
- Best seen with residual plot against fitted values.

Checking equal variance

- Example where there is evidence of non-constant variance
 - ▶ Variance of residuals increases with fitted value



What to do when assumptions fail: linearity

- Failure of the linearity assumption is critical
 - ▶ Conclusions drawn from the model will be invalid
- Paths forward include
 - ▶ Consider transforming outcome or predictor variables (where appropriate)
 - ▶ Explore more sophisticated models
 - Move beyond a simple linear regression model
- Both of these are outside the scope of the course
 - ▶ Considered further in STAT 210, 310

What to do when assumptions fail: independence or equal variance

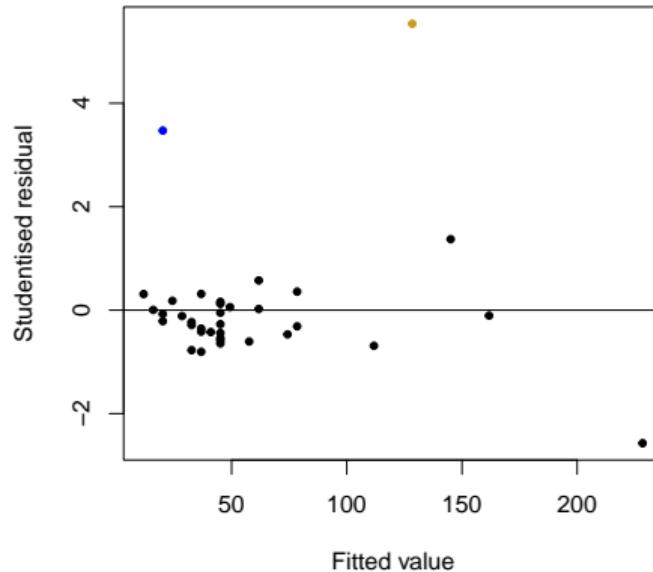
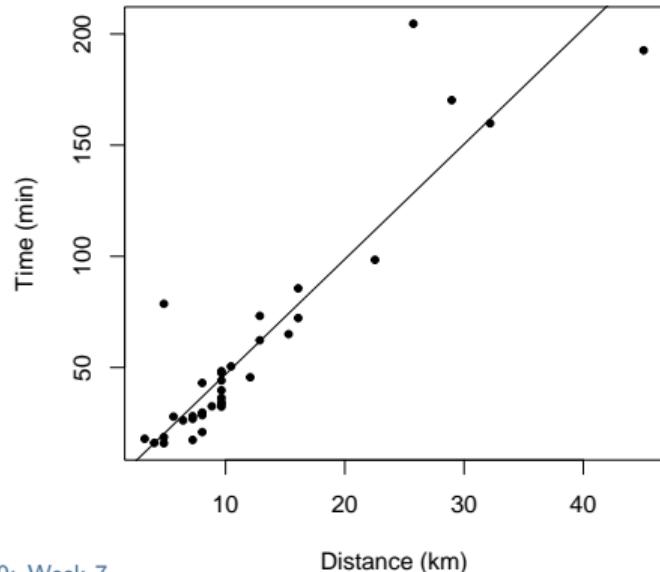
- When independence or equal variance assumptions fail
 - ▶ Estimates of parameters remain valid
 - ▶ Estimates can be inefficient
 - They can be improved
- Assumes that fitted regression line is useable
- Confidence intervals and hypothesis tests will be invalid.
- Failure of assumptions can be rectified by sophisticated modelling techniques.
 - ▶ Details beyond this course.

What to do when assumptions fail: normality / outliers

- Outliers can have a dramatic effect on the estimated regression
 - ▶ Such values are called influential points
- If outliers are present: check that the data are correctly recorded.
- If outliers remain we may consider removing them, however:
 - ▶ Think carefully first
 - Often outliers (or unexpected values in general) are the most interesting
 - They could be revealing something important about what we are studying
 - ▶ We should first assess if they are influential
 - If removing them has little effect: leave them in
 - ▶ If we do remove observations, we must be transparent
 - It should clear and obvious that values were removed and why
- Look at an example

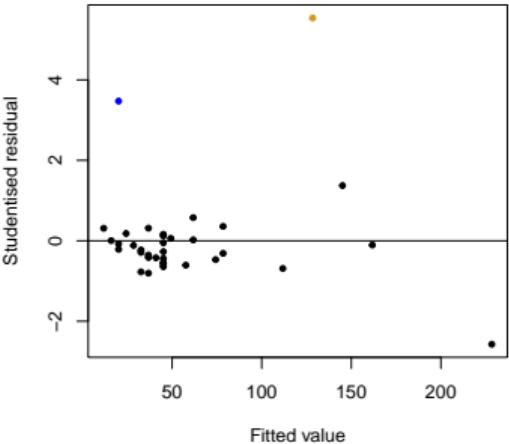
Scottish hill racing

- Data are the record times in 1984 for 35 Scottish hill races (running)
- Interested in the relationship between distance and record time
 - ▶ Outcome variable (y): record time (in minutes)
 - ▶ Predictor variable (x): distance (in km)



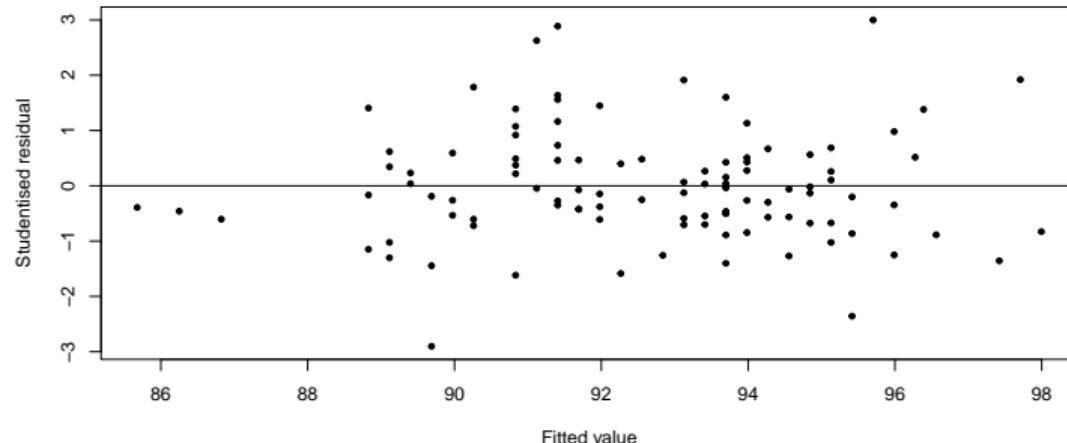
Scottish hill races: Investigate the outliers

- Knock Hill: record incorrectly recorded
 - ▶ Recorded as 78 minutes 39 seconds
 - ▶ It should have been 18 minutes 39 seconds.
- Bens of Jura: other important information?
 - ▶ This race has the largest climb by over 700 m
 - ▶ Consider (extended) model that includes climb?
- General: we may want to think about whether it is reasonable to describe the relationship between time and distance as linear for all races between 3 km and 40+ km



Residuals: possum data

```
plot(fitted(m_pos), rstudent(m_pos), pch = 20, xlab = "Fitted value",
      ylab = "Studentised residual")
abline(h = 0)
```



- Linearity: no evidence of a trend
- Outliers: no apparent outliers
- Constant variance: no obvious change in magnitude of spread of residuals

Recall: weightlifting data

- Maximum squat weight of international power lifters
 - ▶ Found the maximum squat for each athlete across competitions
- Data from 9045 athletes
- Look at the relationship between body weight (kg) and max squat weight (kg)
 - ▶ Outcome variable (y): (best recorded) squat weight
 - ▶ Predictor variable (x): body weight
- Import the data

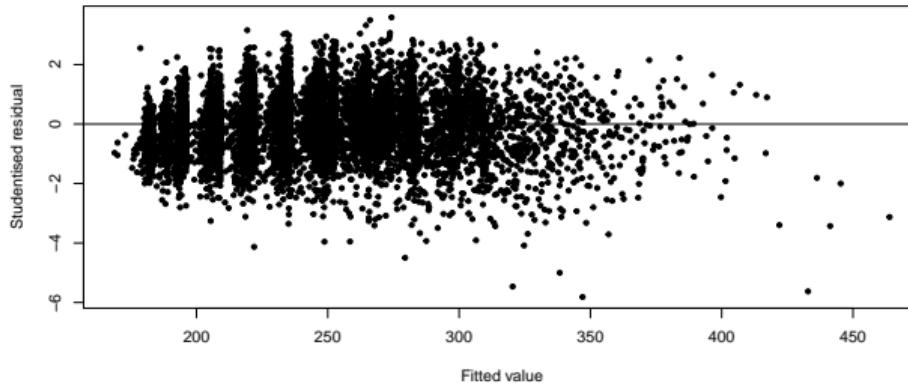
```
powerlift = read.csv('powerlift.csv')
```

- Fit linear regression model

```
m_power = lm(bestsquat ~ bodyweight, data = powerlift)
```

Residuals: powerlift data

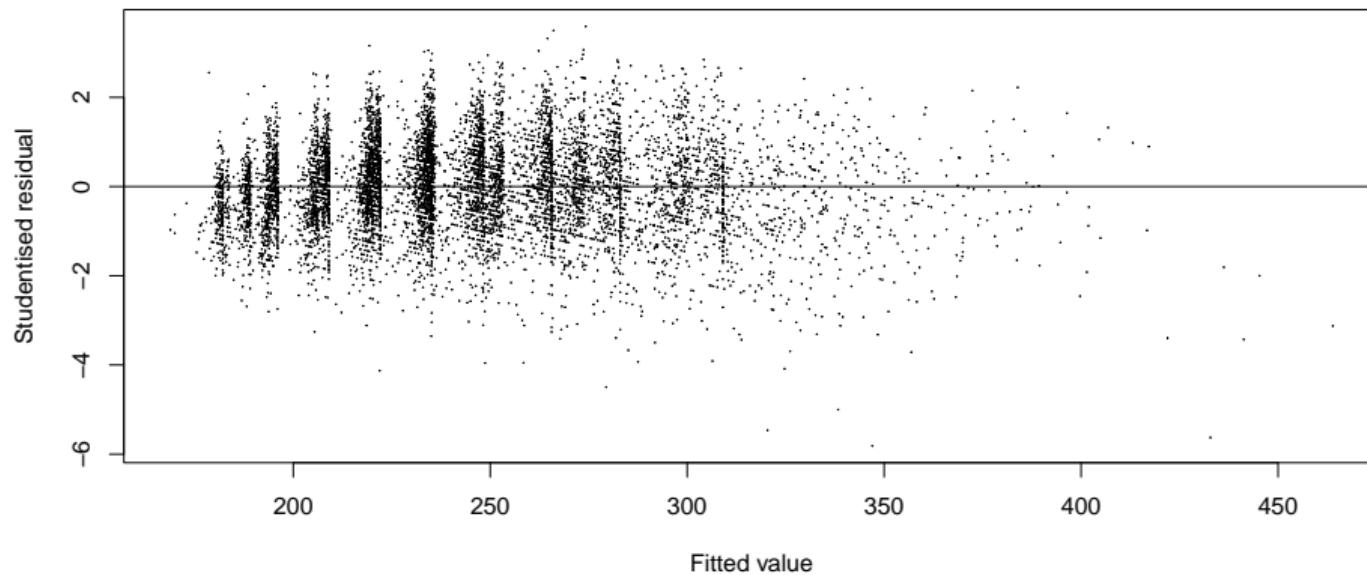
```
plot(fitted(m_power), rstudent(m_power), pch = 20, xlab = "Fitted value",
      ylab = "Studentised residual")
abline(h = 0)
```



- Linearity: very hard to tell
 - ▶ Too many points: draw points smaller to distinguish observations
- Outliers: some large negative residuals
- Constant variance: no obvious change in magnitude of spread of residuals

Residuals: powerlift data

- To better assess linearity
 - ▶ Draw points smaller (better see the number of points)



Residuals: powerlift data

- There is an apparent trend in the residuals
 - ▶ Residuals tend to be negative for low and high fitted values
- A more complex model may be required
 - ▶ e.g. there may be an upper ‘physiological’ limit that a human can squat
 - Consider a model where mean response increases to a maximum value
 - Outside the scope of the course
- Investigate the outliers
 - ▶ Data: maximum squat for each athlete across all recorded competitions
 - ▶ Outliers may have been from competitors with a single competition
 - Possible option: restrict to competitors with data from at least 5 competitions

Summary

- Assumptions of linear regression
 - ▶ LINE
 - Linearity
 - Independence
 - Normality
 - Equal variance
- Introduced residual plots
 - ▶ Can be used to check assumptions of linear regression model

Outline

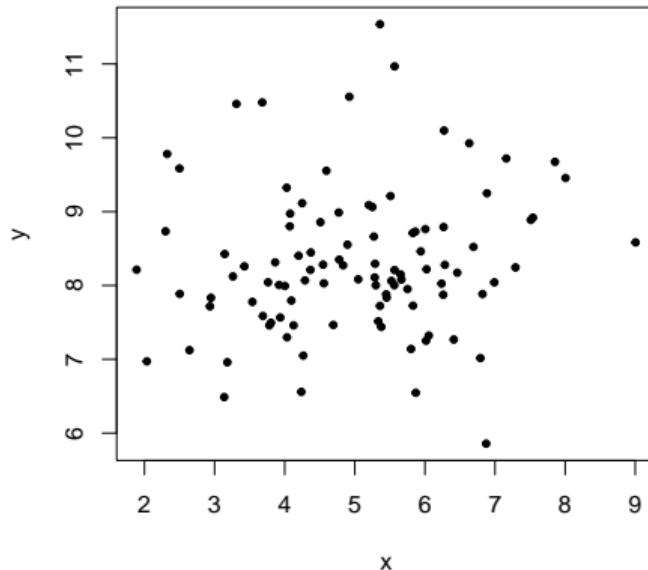
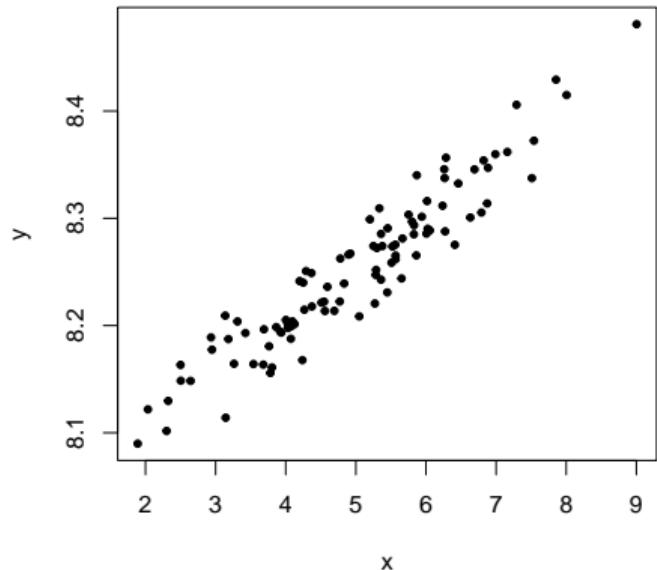
- Previous:
 - ▶ Fitting statistical models
 - ▶ Checking model assumptions
- Today:
 - ▶ Standard error
 - ▶ Confidence interval
 - ▶ Hypothesis test

What does a regression model tell us?

- Consider the height of fathers and sons data
- The fitted model is an estimate of the true regression line in population
 - ▶ Population may be all male NZ university students (and their fathers)
- We need to assess the precision of the estimated parameters
 - ▶ Standard errors of the regression parameters
- Use standard errors to find confidence intervals and conduct hypothesis tests

The importance of the error variance

- Both sets of data come from populations with identical trend: $\mu_y = 8 + 0.05x$.



The importance of error variance

- The linear regression model is $y = \beta_0 + \beta_1 x + \varepsilon$
 - ▶ The error ε is assumed to be normal with mean 0 and variance σ_ε^2
- The larger the error variance (all else equal)
 - ▶ The larger the spread of points around the true regression line
 - ▶ The more uncertain we are about the fitted regression line
 - That is, the estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ are less precise
 - ▶ To quantify our uncertainty about a fitted model
 - We need to estimate the error variance σ_ε^2

Estimation of the error variance

- The residuals ($\hat{\varepsilon}$) are estimates of the true errors (ε)
- Good estimate of error variance σ_{ε}^2 : sample variance of the residuals $\hat{\varepsilon}_1, \hat{\varepsilon}_2, \dots, \hat{\varepsilon}_n$
- We need a few minor technical modifications
- The sample variance of the residuals is $\frac{1}{n-1} \sum_{i=1}^n (\hat{\varepsilon}_i - \bar{\hat{\varepsilon}})^2$.
 - ▶ The sample mean of the residuals is 0: $\bar{\hat{\varepsilon}} = 0$
 - ▶ The correct divisor for simple linear regression is $n - 2$ (rather than $n - 1$)
- So estimate of error variance is

$$s_{\varepsilon}^2 = \frac{1}{n-2} \sum_{i=1}^n \hat{\varepsilon}_i^2 = \frac{RSS}{n-2}$$

- ▶ $RSS = \sum_{i=1}^n \hat{\varepsilon}_i^2$ is called the residual sum of squares

In R: father/son height data

- We can get s_ε from the R output (called Residual standard error)

```
m_height = lm(son ~ father, data = height)
summary(m_height)

##
## Call:
## lm(formula = son ~ father, data = height)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -21.89  -3.89  -0.41   4.59  15.92
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 114.0533    8.4979  13.42 < 2e-16 ***
## father       0.3699    0.0478   7.74  1.9e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.13 on 277 degrees of freedom
## Multiple R-squared:  0.178, Adjusted R-squared:  0.175
## F-statistic: 59.9 on 1 and 277 DF,  p-value: 1.9e-13
```

Standard error of $\hat{\beta}_1$

- In many studies β_1 is the parameter we are most interested in
 - ▶ Change in the expected value of y for changing x in the population
- We estimate $\hat{\beta}_1$ from the observed data (sample)
- Measure precision of estimate by standard error $\sigma_{\hat{\beta}_1}$
 - ▶ Standard deviation of the sampling distribution of $\hat{\beta}_1$
 - Variation in $\hat{\beta}_1$ if there were many data sets (of the same size) from the population

Standard error of $\hat{\beta}_1$

- The standard error for $\hat{\beta}_1$ is

$$\sigma_{\hat{\beta}_1} = \frac{\sigma_\varepsilon}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}}$$

- The standard error is proportional to the error standard deviation σ_ε
 - As σ_ε^2 increases, the standard error of $\hat{\beta}_1$ also increases

- In principle this tells us about the precision of our estimated slope, $\hat{\beta}_1$
- In practice the formula is useless, since we don't know σ_ε
- We can handle that by estimating σ_ε by s_ε
- In practice, we will then use (estimated) standard error

$$s_{\hat{\beta}_1} = \frac{s_\varepsilon}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}}$$

In R

- We can get $s_{\hat{\beta}_1}$ from the R output (column called Std. Error)

```
summary(m_height)

##
## Call:
## lm(formula = son ~ father, data = height)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -21.89  -3.89  -0.41   4.59  15.92
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 114.0533    8.4979 13.42 < 2e-16 ***
## father       0.3699    0.0478   7.74 1.9e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.13 on 277 degrees of freedom
## Multiple R-squared:  0.178, Adjusted R-squared:  0.175
## F-statistic: 59.9 on 1 and 277 DF,  p-value: 1.9e-13
```

Confidence intervals and hypothesis tests

- The standard error is needed to find confidence intervals and test statistics
- Earlier in semester we have seen that confidence intervals take the form

$$\text{estimate} \pm \text{multiplier} \times \text{std. error}$$

- For testing $H_0: \beta_1 = \text{null}$ we use the test statistic

$$t = \frac{\text{estimate} - \text{null}}{\text{std. error}}$$

- These continue to apply for a simple linear regression model

Confidence interval for slope

estimate \pm multiplier \times std. error

- Estimate is $\hat{\beta}_1$
- Multiplier comes from a t -distribution with $\nu = n - 2$ degrees of freedom.
 - ▶ Degrees of freedom match denominator in equation $s_{\varepsilon}^2 = RSS/(n - 2)$.
 - ▶ So for $100(1 - \alpha)\%$ confidence interval, multiplier is $t_{(1 - \frac{\alpha}{2}, \nu)}$.
- Standard error is

$$s_{\hat{\beta}_1} = \frac{s_{\varepsilon}}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}}$$

Confidence interval 'by hand'

$$\hat{\beta}_1 \pm t_{(1-\frac{\alpha}{2}, n-2)} s_{\hat{\beta}_1}$$

$$0.37 \pm t_{(0.975, 277)} \times 0.048$$

- There are $n = 279$ observations
- From R: $qt(0.975, 277) = 1.969$

$$0.37 \pm 1.969 \times 0.048$$

$$0.37 \pm 0.094$$

$$(0.276, 0.464)$$

- We are 95% confident that the true slope is between 0.276 and 0.464
 - ▶ We estimate that the expected height of a son will increase by between 0.276 and 0.464 cm for a 1 cm increase in height of father

In R

- We typically find confidence intervals in R using `confint` function
 - ▶ It is important to understand how the confidence interval is found

```
confint(m_height)

##              2.5 % 97.5 %
## (Intercept) 97.325 130.782
## father      0.276  0.464
```

- Confidence interval for `father` is identical to that calculated on previous slide
- For a 99% confidence interval

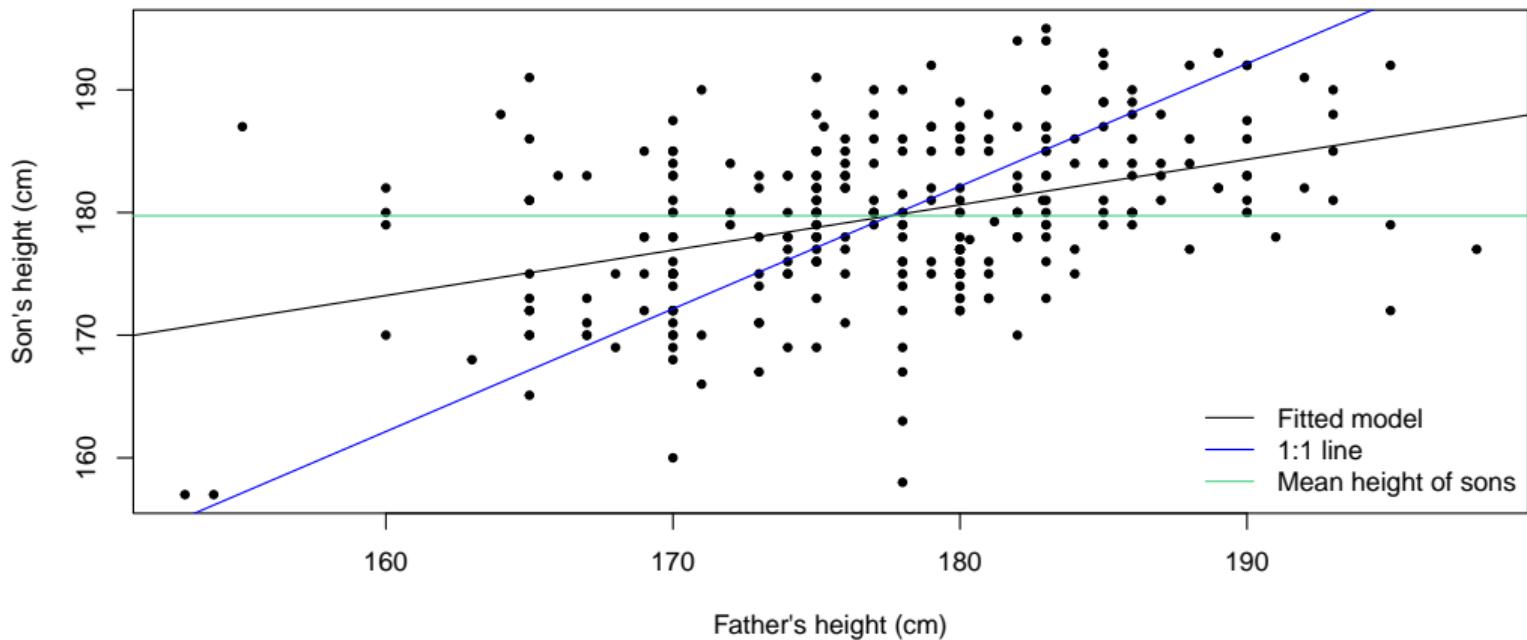
```
confint(m_height, level = 0.99)

##              0.5 % 99.5 %
## (Intercept) 92.012 136.094
## father      0.246  0.494
```

Regression

- We might have expected the average height of a son to increase by 1 cm for a 1 cm increase in father's height.
- That it does not, is the origin of the label: regression (to the mean)
 - ▶ The son of a short father tends to be short, but on average he is taller than his father
 - ▶ The son of a tall father tends to be tall, but on average he is shorter than his father
 - ▶ Extreme traits tend to regress to the mean
- 'Regression' introduced by Francis Galton when comparing the heights of parents and children
 - ▶ Galton, F. (1886). Regression Towards Mediocrity in Hereditary Stature. *The Journal of the Anthropological Institute of Great Britain and Ireland*, **15**, 246–263

Regression to the mean



Hypothesis test for the slope

- Recall: $y = \beta_0 + \beta_1 x + \varepsilon$
 - ▶ β_1 describes how the mean response μ_y changes with x at population level
- If $\beta_1 = 0$ then $y = \beta_0 + \varepsilon$
 - ▶ $\mu_y = \beta_0$: μ_y does not depend on x
 - ▶ Outcome variable is not (linearly) related to the predictor variable
- A hypothesis test about β_1 assesses the hypothesis that two variables are related
 - ▶ Null hypothesis: statement of no relationship between x and y
 - $H_0 : \beta_1 = 0$
 - ▶ Alternative hypothesis: relationship exists
 - $H_A : \beta_1 \neq 0$

The test statistic

- To compute the p -value, we need a test statistic
- The test statistic is

$$t = \frac{\text{estimate} - \text{null}}{\text{std. error}}$$

- The estimate is $\hat{\beta}_1$
- The null value is 0 (previous slide)
- The standard error is $s_{\hat{\beta}_1} = s_\varepsilon / \sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}$
 - ▶ See previous lecture
- So for testing hypothesis about β_1 , we use the test statistic

$$t = \frac{\hat{\beta}_1}{s_{\hat{\beta}_1}}$$

Example: PCB in trout

- Concern that polychlorinated biphenyls (PCBs) polluting waterways and accumulating in food chain
 - ▶ PCBs used to be commonly found in transformers, capacitors, paints, etc
 - ▶ 28 trout collected¹ from Cayuga Lake, NY in 1970
 - Fish were marked and annually stocked (age was known)
 - Each trout was (mechanically) chopped, ground, and mixed before a 5 gm sample taken
 - Chromatography used to find PCB residue in ppm (parts per million)
- Scientific question: is there evidence that (log) PCB residue increases with age?
 - ▶ Null hypothesis: $H_0 : \beta_1 = 0$
 - ▶ Alternative hypothesis: $H_A : \beta_1 \neq 0$
- Treat it as a confirmatory study (specific hypothesis to assess)

¹ Science (1972), 177, 1191–1192.

Example: PCB in trout

- Import the data into R

```
pcb = read.csv('pcb.csv')
```

- Look at the data

```
head(pcb)

##      age logpcb
## 1     1 -0.511
## 2     1  0.470
## 3     1 -0.693
## 4     1  0.182
## 5     2  0.693
## 6     2  0.262
```

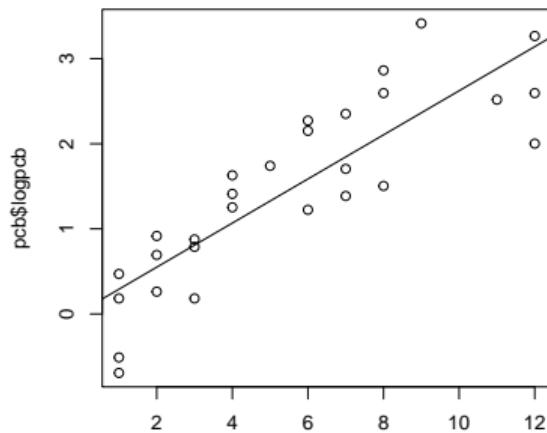
Example: PCB in trout

- Fit simple linear regression

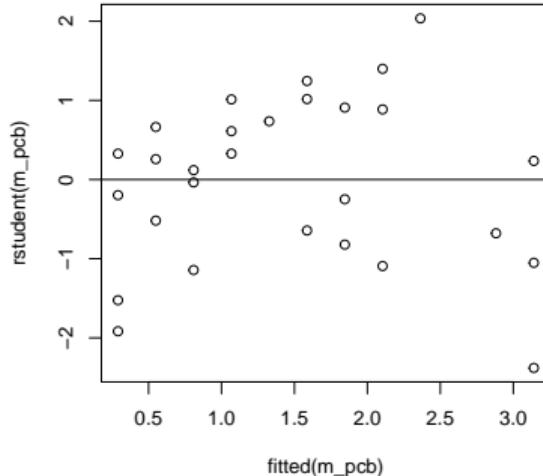
```
m_pcb = lm(logpcb ~ age, data = pcb)
```

- Plot fitted model and residuals: any concerns?

```
plot(pcb$age, pcb$logpcb)  
abline(m_pcb)
```



```
plot(fitted(m_pcb), rstudent(m_pcb))  
abline(h = 0)
```



R model output

```
summary(m_pcb)

##
## Call:
## lm(formula = logpcb ~ age, data = pcb)
##
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -1.1395 -0.3879  0.0957  0.4327  1.0508 
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  0.0315    0.2014    0.16    0.88    
## age         0.2591    0.0308    8.41  6.8e-09 *** 
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 
##
## Residual standard error: 0.567 on 26 degrees of freedom
## Multiple R-squared:  0.731, Adjusted R-squared:  0.721 
## F-statistic: 70.8 on 1 and 26 DF,  p-value: 6.78e-09
```

Interpretation

- For a confirmatory study
 - ▶ Formal test
- Compare the p -value to α
 - ▶ If p -value < α : reject H_0
 - Evidence in favour of H_A
 - ▶ If p -value > α : fail to reject H_0
- For an exploratory study
 - ▶ Interpret the p -value as a degree of incompatibility between data and null hypothesis
 - Use α as a guide
 - Try to avoid making a decision between hypotheses
 - ▶ Often prefer to use confidence intervals

Interpretation PCB: $\alpha = 0.05$

- The test statistic t is given in column t value: 8.41
- The p -value is given in the column $\text{Pr}(>|t|)$: 6.8e-09
 - ▶ These are found assuming the hypothesis: $H_0 : \beta_i = 0$
- p -value $< \alpha$: evidence of incompatibility between the data and null hypothesis
 - ▶ Data are incompatible with assumption of no relationship between PCB and age
 - ▶ Data are unusual compared to what we would expect if the null hypothesis were correct
- As this is a confirmatory study, we conclude that
 - ▶ There is evidence against H_0
 - ▶ There is evidence of a relationship between (log) PCB and age of fish (H_A)

Summary

- We want to quantify how precise our estimate is
 - ▶ Estimate of error variance
 - ▶ Estimate of standard error for $\hat{\beta}_1$
 - ▶ Found confidence interval for β_1
 - ▶ Hypothesis test for β_1
- Discussed origin of 'regression'

STAT 110: Week 8

University of Otago

Outline

- R^2 : the proportion of variance explained
- Another look at estimating the mean response
- Predicting a new observation
- Extrapolation

Recall: possum data

- The size of brushtail possums
 - ▶ Exploring relationship between total length (mm) and head length (mm)
- If we have a total length measurement
 - ▶ Can we predict the head length?
- Import the data into R

```
possum = read.csv('possum.csv')
```

- Fit a simple linear regression

```
m_possum = lm(head_l ~ total_l, data = possum)
```

Output

```
summary(m_possum)

##
## Call:
## lm(formula = head_l ~ total_l, data = possum)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -7.188 -1.534 -0.334  1.279  7.397
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 42.70979   5.17281   8.26  5.7e-13 ***
## total_l      0.05729   0.00593   9.66  4.7e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.6 on 102 degrees of freedom
## Multiple R-squared:  0.478, Adjusted R-squared:  0.472
## F-statistic: 93.3 on 1 and 102 DF,  p-value: 4.68e-16
```

R^2 : Coefficient of determination

- R^2 is a commonly used measure of how well a regression model describes the data
 - ▶ In R summary: Multiple R-squared = 0.478
- Look at two descriptions of R^2
 - ▶ Give us different perspectives on what it represents

R^2 : squared correlation

- R^2 is the squared correlation between y and \hat{y}

```
y = possum$head_1 # y values  
yhat = fitted(m_possum) # y-hat values  
R = cor(y, yhat)  
R^2 # correlation^2  
## [1] 0.478
```

- Since $-1 \leq r \leq 1$ we have $0 < R^2 < 1$
 - ▶ The larger the value of R^2 , the better the regression model describes the data
 - The fitted values are 'close' to the observations

R^2 : percentage of variance explained

- The total sum of squares is $TSS = \sum_{i=1}^n (y_i - \bar{y})^2$
 - ▶ Measures the variability of the outcome variable
- (Recall) the residual sum of squares $RSS = \sum_{i=1}^n (y_i - \hat{y}_i)^2$
 - ▶ Measures the variability of the outcome variable after fitting regression model
- The explained sum of squares $ESS = TSS - RSS$
 - ▶ Amount of variation in the outcome variable that is explained by the regression model
- R^2 can be expressed as

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS}$$

- The proportion of variance explained by the model
 - ▶ R^2 is often reported as a percentage: $R^2 = 47.8\%$

Interpreting R^2

- R^2 is often reported when fitting a linear regression
- No absolute rule for what a good (or bad) R^2 value is
 - ▶ In one particular area of application: an R^2 of 0.3 might be good
 - ▶ In another area of application: an R^2 of 0.8 might be poor

Mean response

- Recall: linear regression model

$$y = \beta_0 + \beta_1 x + \varepsilon$$

- Mean response at a given x value: $\mu_y = \beta_0 + \beta_1 x$
- The fitted model is an estimate of the mean response

$$\hat{\mu}_y = \hat{\beta}_0 + \hat{\beta}_1 x$$

- How precise is this estimate?
- Can we find a confidence interval for μ_y ?
 - ▶ e.g. what is the confidence interval for mean head length of the subpopulation of possums with total length 850 mm

Confidence interval for mean response

- Goal: find a confidence interval for μ_{y_0} , the mean response when $x = x_0$
- Confidence interval will have the form

estimate \pm multiplier \times std. error

- Estimate: $\hat{\mu}_{y_0} = \hat{\beta}_0 + \hat{\beta}_1 x_0$.
- The (estimated) standard error for $\hat{\mu}_{y_0}$ is

$$s_{\hat{\mu}_{y_0}} = s_\varepsilon \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}$$

- Multiplier: t-distribution with $\nu = n - 2$ degrees of freedom

Confidence interval for mean response

- A $100(1 - \alpha)\%$ confidence interval for μ_{y_0} is given by

$$\hat{\mu}_{y_0} \pm t_{(1 - \frac{\alpha}{2}, n-2)} \times s_{\hat{\mu}_{y_0}}$$

- This is an interval estimate for the mean response μ_{y_0}
- Finding this confidence interval by hand is tedious
 - ▶ Use R to help us
 - ▶ predict function
- The predict function requires a data frame
 - ▶ Contains x_0 : the predictor variable values where we want to find the mean response

Excursion: data frames in R

- You have been using data frames all semester
- When we import data into R: it is in a data frame
 - ▶ Rows: Each row is an observation or data record
 - ▶ Columns: Each column is a variable (typically with a name)
- We can construct a data frame using function `data.frame`

```
first_df = data.frame(name = c("Bob", "Mary", "Lucy"), age = c(19, 17, 23),  
                      height = c(173, 168, 176))  
  
first_df  
  
##   name age height  
## 1 Bob   19    173  
## 2 Mary   17    168  
## 3 Lucy   23    176
```

Data from for predict: possum data

- We need to construct a data frame in R
 - ▶ Contain the x (predictor variable) values where we want to find the mean response
 - ▶ Same variable name as was used to fit the model in `lm`
- Recall:

```
m_possum = lm(head_l ~ total_l, data = possum)
```

- Predictor variable name: `total_l`
- Let's say we want to estimate the mean response at 850 mm

```
predictor1 = data.frame(total_l = 850)
```

- If we wanted to find the mean response at 850 mm and 900 mm

```
predictor2 = data.frame(total_l = c(850, 900))
```

Mean response in R

- Use the `predict` function, with option `interval = "confidence"`

```
mean_resp = predict(m_possum, newdata = predictor1, interval = "confidence")
mean_resp
##      fit    lwr   upr
## 1 91.4 90.8 92
```

- First argument: model we are using (`m_possum`)
- Second argument (`newdata`): data frame of predictor values
- Third argument (`interval`): the kind of interval
 - ▶ Confidence interval for mean response: `interval = "confidence"`

Mean response: possum

- The estimated mean response is

$$\hat{\mu}_{y_0} = \hat{\beta}_0 + \hat{\beta}_1 x_0 = 42.71 + 0.057 \times 850 = 91.4$$

- Estimated mean head length for possums with total length 850 mm is 91.4 mm
 - Given by fit from predict output

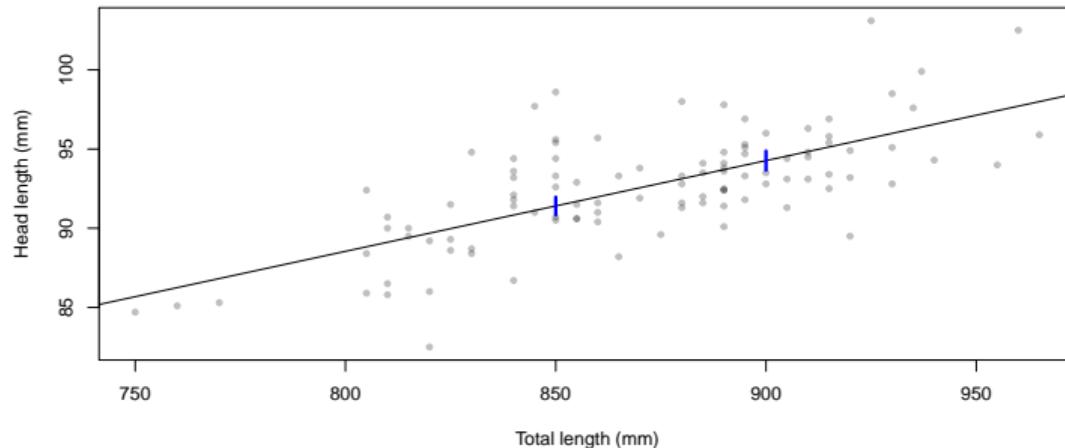
```
mean_resp  
##     fit   lwr upr  
## 1 91.4 90.8 92
```

- We are 95% confident that the mean head length for possums with total length 850 mm is between 90.8 mm and 92 mm
 - Given by lwr and upr in predict output

Mean response: visual

```
mean_resp2 = predict(m_possum, newdata = predictor2, interval = "confidence")
mean_resp2

##      fit    lwr   upr
## 1 91.4 90.8 92.0
## 2 94.3 93.7 94.9
```



Prediction

- We can also use the model to predict a new observation y_0
- At a given value of $x = x_0$ (say $x_0 = 850$ mm)
 - ▶ The prediction (\hat{y}_0) is the same as the estimated mean response ($\hat{\mu}_{y_0}$)
 - Recall: fitted line was $\hat{y} = \hat{\mu}_y = \hat{\beta}_0 + \hat{\beta}_1 x$
- That means that at $x_0 = 850$ mm we have

$$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0 = 42.71 + 0.057 \times 850 = 91.4$$

- We predict that a (new) possum of 850 mm would have a head length of 91.4 mm
 - ▶ What about the possible error in the prediction?
 - ▶ We want to find a prediction interval?

Prediction error

- The prediction uncertainty is larger than the uncertainty about mean response
 - ▶ It needs to combine uncertainty about the mean response and individual variability
- Eg. if we are predicting the head length of a possum with total length 850 mm
 - ▶ The mean head length among the subpopulation of possums with total length 850 mm is uncertain
 - Standard error for mean response
 - ▶ There is possum to possum variability in head length among the subpopulation of possums with total length 850 mm
 - Not all possums with total length 850 mm will have the same head length
 - Given by the error ε in the linear regression model

Prediction error

- The prediction error takes account of both sources of uncertainty
- For prediction at $x = x_0$, the prediction error is

$$PE(\hat{y}_0) = s_\varepsilon \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}.$$

- ▶ Looks like standard error for mean response
 - Has an extra term in the square root: $1 + \frac{1}{n}$
 - Accounts for individual variation about the mean
- A $100(1 - \alpha)\%$ prediction interval for y_0 is $\hat{y}_0 \pm t_{(1-\frac{\alpha}{2}, n-2)} \times PE(\hat{y}_0)$
- The prediction interval is a probability interval
 - ▶ There is a probability of $(1 - \alpha)$ that y_0 will lie in this interval

Prediction in R

- Use the `predict` function, with option `interval = "prediction"`

```
pred = predict(m_possum, newdata = predictor1, interval = "prediction")
pred
##     fit    lwr   upr
## 1 91.4 86.2 96.6
```

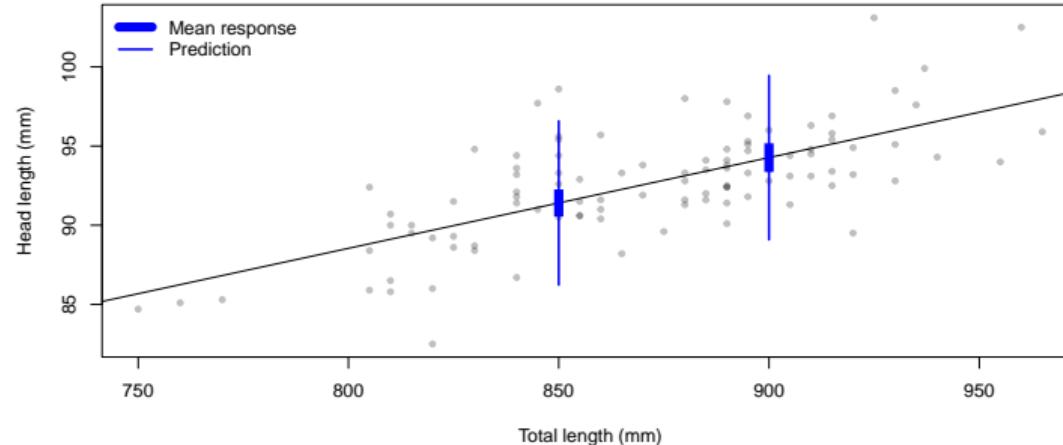
- There is a probability of 0.95 that a possum with total length 850 mm will have head length between 86.2 mm and 96.6 mm
- Note: we can find a 90% or 99% interval by including the argument `level`
 - ▶ Also applies when finding confidence interval for mean response

```
predict(m_possum, newdata = predictor1, interval = "prediction", level = 0.99)
##     fit    lwr   upr
## 1 91.4 84.6 98.3
```

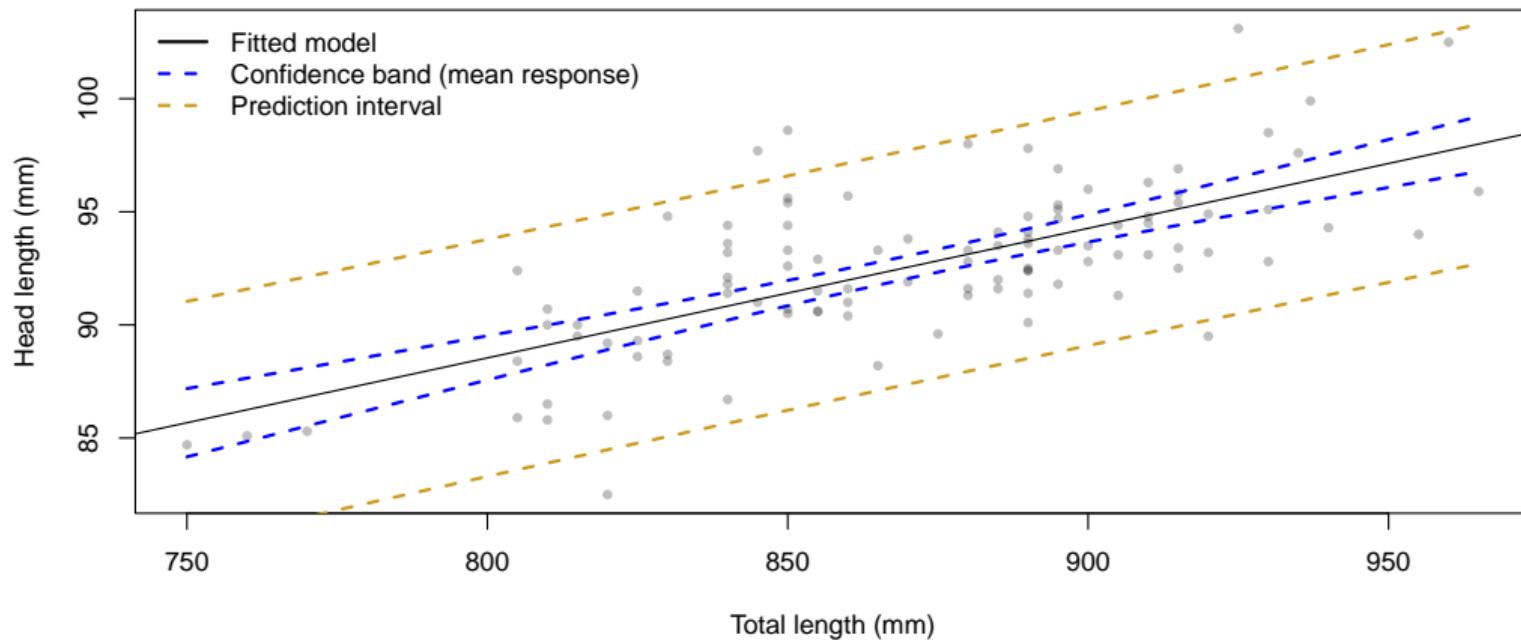
Prediction: visual

```
pred2 = predict(m_possum, newdata = predictor2, interval = "prediction")
pred2

##      fit    lwr   upr
## 1 91.4 86.2 96.6
## 2 94.3 89.1 99.5
```



Mean response and prediction: visual



Mean response and prediction

- The mean response is most precise in middle of plot
 - ▶ Confidence interval is narrower
- Same is true of prediction interval (harder to see on plot)
- The standard error and prediction error both include the term

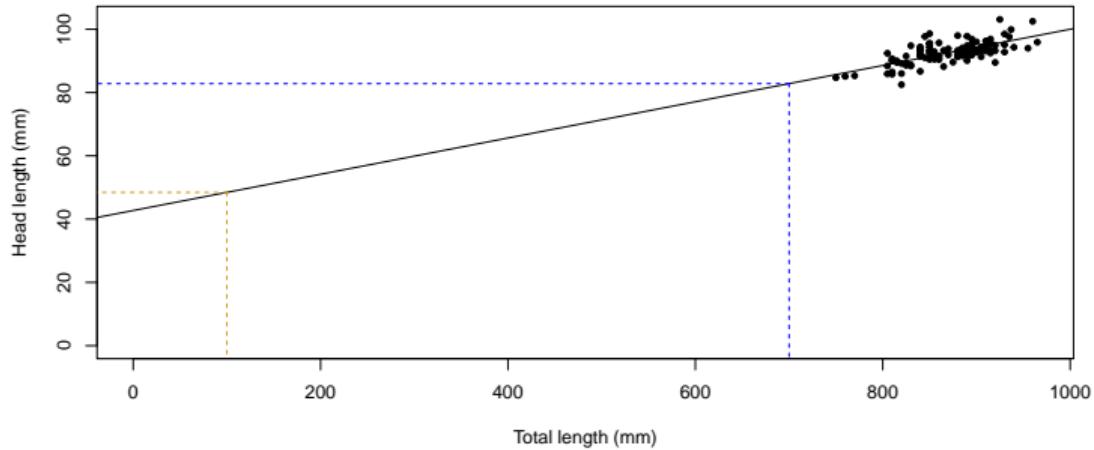
$$\frac{(x_0 - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

- This is smallest when $x_0 = \bar{x}$
 - ▶ Estimation of mean response and prediction is most precise at $x_0 = \bar{x}$
 - ▶ Errors increase the further x_0 is from sample mean \bar{x}

Extrapolation

- When using linear regression models
 - ▶ Care is needed if extrapolating!
- Extrapolation: predicting values outside the range of the observed data
- Why is this a problem?
 - ▶ The linear regression model has limitations
 - It approximates the relationship between x and y across the range of data we observe
 - We don't necessarily believe it describes the true relationship between x and y
 - We don't know how data will behave outside the range we have observed
- If we decide to extrapolate
 - ▶ Important to know the risks and limitations

Extrapolation: possum



- The linear regression model provides a description of the relationship between total length and head length across the range of observed data
 - ▶ Total length between 750 mm and 950 mm
- We don't believe it describes the true relationship
 - ▶ We wouldn't use it to predict head length when total length is 100 mm
 - ▶ What about predicting head length when total length is 700 mm?

Summary

- Model summary: R^2
 - ▶ Squared correlation between fitted values and observations
 - ▶ Gives the percentage of variance explained by regression
- Looked again at mean response
 - ▶ Found confidence interval for mean response at $x = x_0$
- Looked at predicting a new observation
 - ▶ $\hat{y} = \hat{\mu}_y$
 - ▶ Prediction interval wider than confidence interval for mean response
- Looked at dangers of extrapolating

Outline

- Explore multiple linear regression
 - ▶ Where there is more than one predictor variable
- How to fit in R
- How to interpret the estimates
- How to find confidence intervals and conduct hypothesis tests
- Estimating mean response and predicting new observation
- Assessing model fit

Neurocognitive scores

- Neurocognitive function evaluated with MATRICS Consensus Cognitive Battery¹
 - ▶ Measures cognitive performance in seven domains
- To start, we will focus on one domain: speed of processing
 - ▶ Explore how does it relate to age?
- We will use data from 145 'healthy' participants
 - ▶ Screen for medical and psychiatric illness
 - ▶ No history of substance abuse
- Subset of a larger study that had different aims²
 - ▶ Assess how cognitive scores varied between individuals with schizophrenia, individuals with schizoaffective disorder, and healthy controls

¹*American Journal of Psychiatry*, 165, 203–213, 2008.

²*Schizophrenia Research: Cognition*, 2, 227–232, 2015.

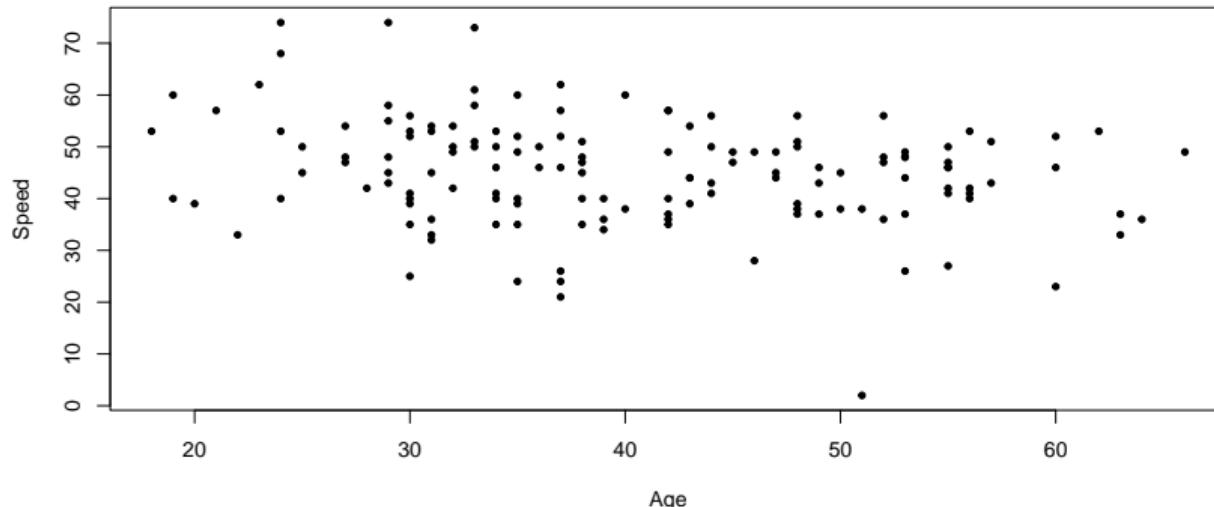
Neurocognitive scores: data

- Import the data

```
neuro = read.csv('neuro.csv')
```

- Look at scatterplot of speed score and age

```
plot(neuro$age, neuro$speed, xlab = "Age", ylab = "Speed", pch = 20)
```



Neurocognitive scores: regression model

- Consider the model: $\text{speed} = \beta_0 + \beta_1 \text{age} + \varepsilon$
 - ▶ Score in the speed of processing test: outcome variable y
 - ▶ Age of participant: predictor variable x
- If we take $y = \text{speed}$ and $x = \text{age}$ we have the usual model: $y = \beta_0 + \beta_1 x + \varepsilon$
- The parameters:
 - ▶ β_0 is the expected outcome when the predictor variable is 0
 - How useful (or meaningful) the parameter is, depends on application
 - Neurocognitive example: expected speed score when age is 0 (not meaningful to interpret)
 - ▶ β_1 is the change in the expected outcome for a one unit increase in the predictor
 - Change in the expected speed score for a one year increase in age
 - Comparing two subpopulations that are one year apart in age

Neurocognitive scores: fitted regression model

```
m_neuro = lm(speed ~ age, data = neuro)
summary(m_neuro)

##
## Call:
## lm(formula = speed ~ age, data = neuro)
##
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -40.72   -6.17    0.40    5.80   26.35 
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 54.1468    3.1646   17.11 <2e-16 ***
## age         -0.2240    0.0757   -2.96  0.0036 **  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.2 on 143 degrees of freedom
## Multiple R-squared:  0.0578, Adjusted R-squared:  0.0512 
## F-statistic: 8.77 on 1 and 143 DF,  p-value: 0.00359
```

Interpret the effect

- Find confidence intervals

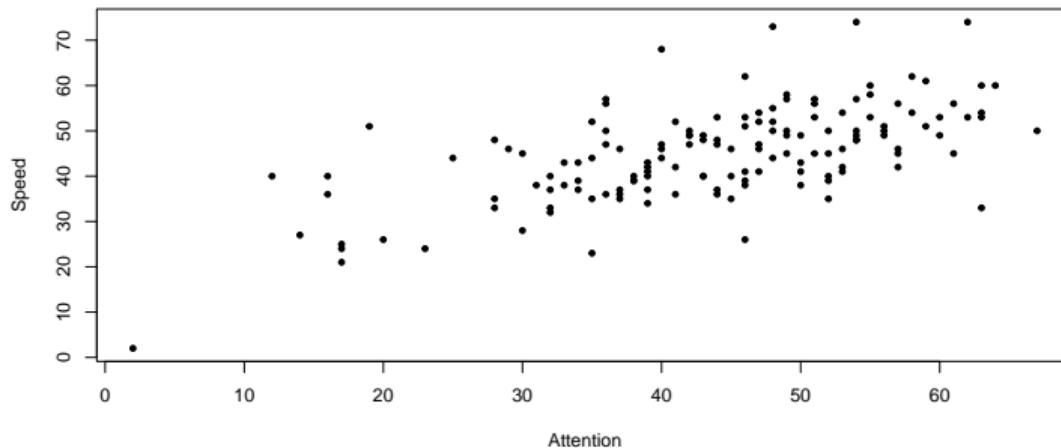
```
confint(m_neuro)

##             2.5 %   97.5 %
## (Intercept) 47.891 60.4022
## age         -0.374 -0.0745
```

- We are 95% confident that the increase in expected speed score is between -0.374 and -0.074 for a one year increase in age
- As $\hat{\beta}_1$ is negative: represents a decrease in expected score
 - ▶ We are 95% confident that the decrease in expected speed score is between 0.074 and 0.374 for a one year increase in age

We have more information...

- The regression is explaining $R^2 = 5.8\%$ of the variation in speed score
- There are other variables that could potentially help explain the speed score
 - ▶ e.g. the score on the other domains: we will look at scores from the attention domain



- Can we use attention and age together to describe the speed scores?

Multiple linear regression

- In multiple linear regression we have multiple predictors
 - ▶ We call them x_1, x_2, \dots, x_k
 - ▶ k denotes the number of predictor variables
- The multiple regression model is $y = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k + \varepsilon$
 - ▶ $\beta_0, \beta_1, \dots, \beta_k$ are parameters (regression coefficients)
 - ▶ ε is an error term following a $N(0, \sigma_\varepsilon^2)$ distribution.
- The mean response is $\mu_y = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k$
 - ▶ This is a conditional mean, given the values of the predictor variables x_1, \dots, x_k
- For the neurocognitive scores we have

$$\text{speed} = \beta_0 + \beta_1 \text{age} + \beta_2 \text{attention} + \varepsilon$$

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

Model fitting

- Once we have parameter estimates $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k$, the fitted model is

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k$$

- \hat{y} is also an estimate $\hat{\mu}_y$ of the mean response
- We can find the residuals: $\hat{\varepsilon}_i = y_i - \hat{y}_i$
 - Estimate of the error term ε_i
 - Identical to simple linear regression
- We can use least squares to find estimates $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k$
 - Minimise the squared residuals $\sum_{i=1}^n \hat{\varepsilon}_i^2$
 - Same as with simple linear regression

Multiple regression: in R

- Use the same function to fit multiple linear regression: `lm`
- Add another predictor variable: `+ attention`

```
m_neuro2 = lm(speed ~ age + attention, data = neuro)
```

- We will see that much remains the same with multiple linear regression
 - ▶ Highlight differences with simple linear regression
- One difference is that it is much harder to visualise multiple linear regression
 - ▶ We now have two predictor variables (and we could potentially have more!)

Neurocognitive scores: in R

```
summary(m_neuro2)

##
## Call:
## lm(formula = speed ~ age + attention, data = neuro)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -21.176 -5.495 -0.466  4.458 23.770
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 31.6661    3.2885   9.63 <2e-16 ***
## age         -0.2459    0.0579  -4.24  4e-05 ***
## attention    0.5349    0.0529  10.11 <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.79 on 142 degrees of freedom
## Multiple R-squared:  0.452, Adjusted R-squared:  0.444
## F-statistic: 58.6 on 2 and 142 DF,  p-value: <2e-16
```

Interpretation

- There are some (minor) changes in how we interpret the parameters
- β_0 : expected outcome when *all* predictor variables are 0
- Other coefficients are specific to the associated explanatory variable
 - ▶ e.g. β_2 is the change in the expected outcome when variable x_2 is increased by one unit, *and all other predictor variables remain unchanged*
 - Often say: all else held fixed
- In the neurocognitive scores example: β_2 is the change in the expected speed score when the attention score is increased by one, all else held fixed
 - ▶ All else held fixed: age unchanged
- Sometimes expressed as: β_2 is the effect of x_2 *having adjusted for* all other predictor variables

Interpretation: neurocognitive scores

- The fitted model is

$$\widehat{\text{speed}} = 31.67 - 0.25 \text{ age} + 0.53 \text{ attention}$$

- Interpretation of $\hat{\beta}_1$: the decrease in expected speed score is estimated to be 0.25 for a one year increase in age, holding the attention score fixed
- Interpretation of $\hat{\beta}_2$: the increase in average speed score is estimated to be 0.53 for a one year increase in age, having adjusted for age
- It doesn't make sense to interpret $\hat{\beta}_0$, but if we did
 - ▶ The average speed score for a participant of age 0, with attention score of 0 is 31.67
 - ▶ Why does it not make sense to interpret this?

Confidence interval

- We can find confidence intervals for the parameter β_j

- ▶ Minor changes from simple linear regression

- We still use

$$\text{estimate} \pm \text{multiplier} \times \text{standard error}$$

- The estimate is $\hat{\beta}_j$
- The multiplier comes from a t -distribution with $\nu = n - k - 1$ degrees of freedom
- The (estimated) standard error $s_{\hat{\beta}_j}$ is complicated
 - ▶ It can be obtained from R output: column Std. error
- We can still find confidence interval directly with `confint`

Confidence interval: neurocognitive scores

- The confidence intervals are

```
confint(m_neuro2, level = 0.9)

##           5 %    95 %
## (Intercept) 26.222 37.111
## age         -0.342 -0.150
## attention   0.447  0.623
```

- Interpreting the confidence interval for β_2
 - We are 90% confident that the average speed score will increase by between 0.447 and 0.623 for a one unit increase in the attention score, holding age fixed.

Hypothesis testing

- The multiple linear regression model is

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \varepsilon$$

- The mean response is $\mu_y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$
 - ▶ This depends on variable x_j only if β_j is not 0
- Testing $\beta_j = 0$ is equivalent to testing if mean response depends on x_j
 - ▶ Having adjusted for all the other variables in the model

Setting up the hypothesis test

- We set up a null hypothesis indicating ‘no effect’
 - ▶ $H_0 : \beta_j = 0$
 - ▶ $H_A : \beta_j \neq 0$
- The test statistic is of the usual form:

$$t = \frac{\text{estimate} - \text{null}}{\text{standard error}} = \frac{\hat{\beta}_j}{s_{\hat{\beta}_j}}$$

- The t statistic, estimate $\hat{\beta}_j$, estimate standard error $s_{\hat{\beta}_j}$ and p -value are all available in the R output
- The p -value quantifies the incompatibility between the data and null hypothesis
 - ▶ A small p -value suggests the data are unusual assuming the null hypothesis is true

Prediction and mean estimation in multiple regression

- As with simple linear regression, the fitted model can be interpreted as both
 - ▶ An estimate of the mean response $\hat{\mu}_y$, and
 - ▶ A prediction of the response for a new data point \hat{y}
- If $x_{01}, x_{02}, \dots, x_{0k}$ give the value of the predictor variables at which we wish to predict/estimate, then

$$\hat{y}_0 = \hat{\mu}_{y_0} = \hat{\beta}_0 + \hat{\beta}_1 x_{01} + \cdots + \hat{\beta}_k x_{0k}$$

- The estimated mean response and predicted value are the same

Prediction and mean estimation: neurocognitive scores

- The fitted model is

$$\widehat{\text{speed}} = 31.67 - 0.25 \text{ age} + 0.53 \text{ attention}$$

- The estimated mean response (and prediction) for participant aged 40, with attention score of 50 is

$$\begin{aligned}\widehat{\text{speed}} &= 31.67 - 0.25 \times 40 + 0.53 \times 50 \\ &= 48.58\end{aligned}$$

Prediction and mean estimation in multiple regression

- The general structure of the intervals is the same as with simple linear regression
 - ▶ A $100(1 - \alpha)\%$ confidence interval for mean response μ_{y_0} is

$$\hat{\mu}_{y_0} \pm t_{(1-\frac{\alpha}{2}, n-k-1)} \times s_{\hat{\mu}_{y_0}}$$

- ▶ A $100(1 - \alpha)\%$ prediction interval for y_0 is

$$\hat{y}_0 \pm t_{(1-\frac{\alpha}{2}, n-k-1)} \times PE(\hat{y}_0)$$

- These are minor changes from simple linear regression:
 - ▶ Multiplier degrees of freedom are now $n - k - 1$
 - ▶ The formulae for standard error $s_{\hat{\mu}_{y_0}}$ and prediction error $PE(\hat{y}_0)$ are more complicated
- The way in which we find these in R remains the same

Mean response and prediction in R

- Mean response and prediction for participant aged 40 with attention score 50
- Set up data frame

```
to_pred = data.frame(age = 40, attention = 50)
```

- Estimated mean response with confidence interval (`interval = "confidence"`)

```
predict(m_neuro2, newdata = to_pred, interval = "confidence")  
##     fit    lwr   upr  
## 1 48.6 47.1 50
```

- Prediction with prediction interval (`interval = "predict"`)

```
predict(m_neuro2, newdata = to_pred, interval = "predict")  
##     fit    lwr   upr  
## 1 48.6 33.1 64
```

Model assumptions

- The multiple linear regression model is

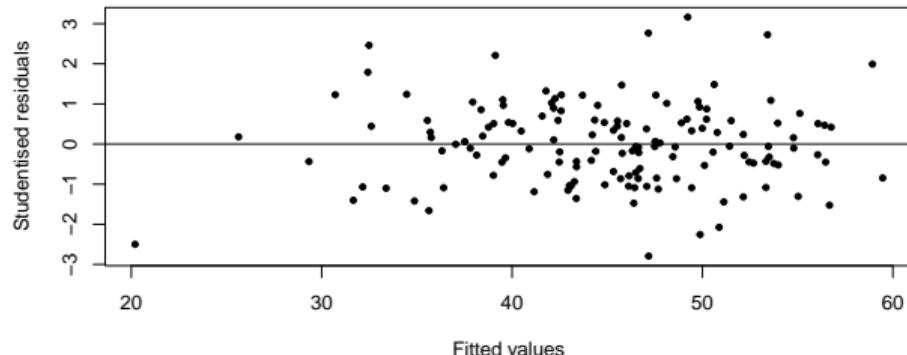
$$y = \underbrace{\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k}_{\mu_y} + \varepsilon$$

- We are making the following assumptions:
 - Linearity:** There is a linear line relationship between μ_y and x_j when all other predictor variables are held constant
 - Independence:** The error terms $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n$ are independent
 - Normality:** The error terms $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n$ are normally distributed
 - Equal variance:** The errors terms all have the same variance, σ_ε^2 ('homoscedastic').

Checking assumptions: same as simple linear regression

- Check assumptions by plotting studentised residuals against fitted values
- Violation of assumptions given by
 - ▶ A trend (linearity), changing variance (equal variance), outliers (normality)
- Are there any obvious violations of assumptions?

```
plot(fitted(m_neuro2), rstudent(m_neuro2), xlab = "Fitted values",
      ylab = "Studentised residuals", pch = 20)
abline(h = 0)
```



Coefficient of determination R^2

- Definition of R^2 the same as for simple linear regression
 - ▶ The squared correlation between outcome y and fitted values \hat{y}
 - ▶ The percentage of variance explained by the regression model
- For neurocognitive example:
 - ▶ Age (simple linear regression) explains $R^2 = 5.8\%$ of the variation in speed scores
 - ▶ Age and the attention score (multiple linear regression) explain $R^2 = 45.2\%$ of the variation in speed scores
- Both of these can be read off the summaries in slides above

Big picture

- Multiple linear regression is an incredibly powerful tool
 - ▶ We've only just scratched the surface
- There are a lot of important topics we haven't covered, including
 - ▶ Model building
 - ▶ Variable selection
 - ▶ Collinearity (this is when two predictors explain similar variation)
 - ▶ Interactions (when effect of one variable depends on value of another)
 - ▶ ...
- There are lots of possible extensions
- There are also lots of ways to get ourselves into trouble
- STAT 210 explores the use of multiple linear regression for scientific problems

Summary

- Looked at multiple linear regression
 - ▶ Where we have more than one predictor variable
- Only scratched the surface
- We have looked at
 - ▶ Fitting the model
 - ▶ Interpreting the parameters
 - ▶ Finding confidence interval or performing a hypothesis test
 - ▶ Estimating the mean response and predicting a new observation
 - ▶ Assessing model fit

Outline

- Think again about categorical predictor variables
- Categorical predictors with two levels
 - ▶ Include them in a linear regression model
 - ▶ Compare to the difference in means of two independent groups
- Categorical predictors with more than two levels
 - ▶ Introduce ANOVA (analysis of variance) model

Predictor variables

- We have looked at lots of linear regression examples
- The predictor variables in these examples were
 - ▶ Height: father's height
 - ▶ Possums: total length of possum
 - ▶ Powerlifting: weight of athlete
 - ▶ Neurocognitive scores: age and attention score
- All of these are continuous variables
- Linear regression can also be used when the predictor variable is categorical
 - ▶ Represent groups or categories, e.g. sex, country of birth, blood type, etc.
 - ▶ Start with categorical variables with two levels (or groups)
 - e.g. sex: male and female

Mario Kart

- Ebay auctions for video game: Mario Kart for Nintendo Wii
 - ▶ Ebay is similar to trademe
 - ▶ Online auction website
- Two variables:
 - ▶ Total auction price: continuous outcome variable y
 - ▶ Game condition: categorical predictor variable x taking values used and new
- Another example is comparing EEG frequencies (brain waves) according to sensory deprivation (control or solitary confinement)
 - ▶ Example we considered in an earlier lecture

Hang on a minute...

- We already know how to model these data!
 - ▶ Two independent groups
 - Group 1: normally distributed with mean μ_1 and variance σ_1^2
 - Group 2: normally distributed with mean μ_2 and variance σ_2^2
 - ▶ Find confidence interval for $\mu_2 - \mu_1$ using `t.test` in R
- Why are we looking at this in the context of linear regression?
 1. Understanding: see how two independent groups is ‘special case’ of linear regression
 2. Useful: use categorical variables in multiple regression
 - e.g. for Mario Kart auction data: we could explore how auction length, and the number of bids, as well as game condition relate to auction price
- We will look at only one outcome variable and one categorical predictor
 - ▶ See STAT 210 for more elaborate models

Data: Mario Kart

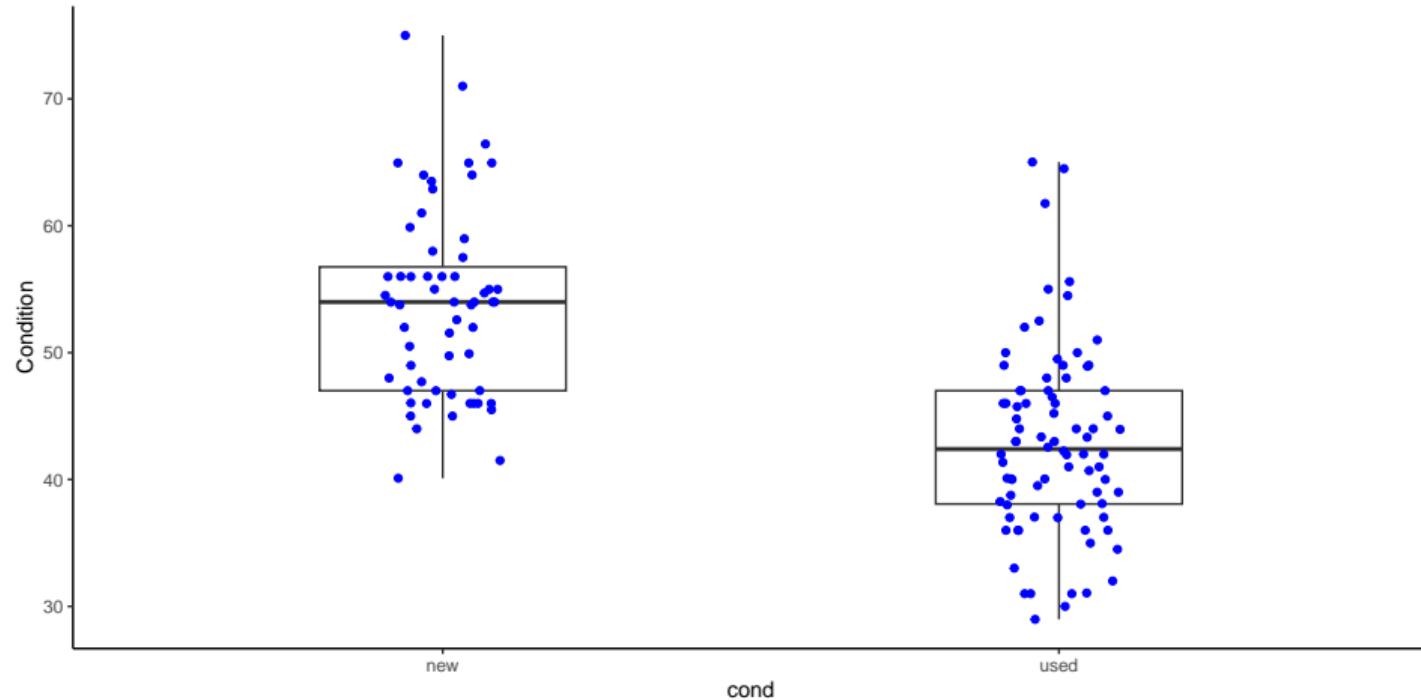
- Import the data into R

```
mario = read.csv('mario.csv')
```

- The data have had two observations / outliers removed
 - The data are from a full week of auctions in October 2009
 - Removed observations: auctions where multiple games (incl. Mario Kart) were sold
- Look at the data

```
head(mario)  
##   cond price  
## 1  new  51.5  
## 2 used  37.0  
## 3  new  45.5  
## 4  new  44.0  
## 5  new  71.0  
## 6  new  45.0
```

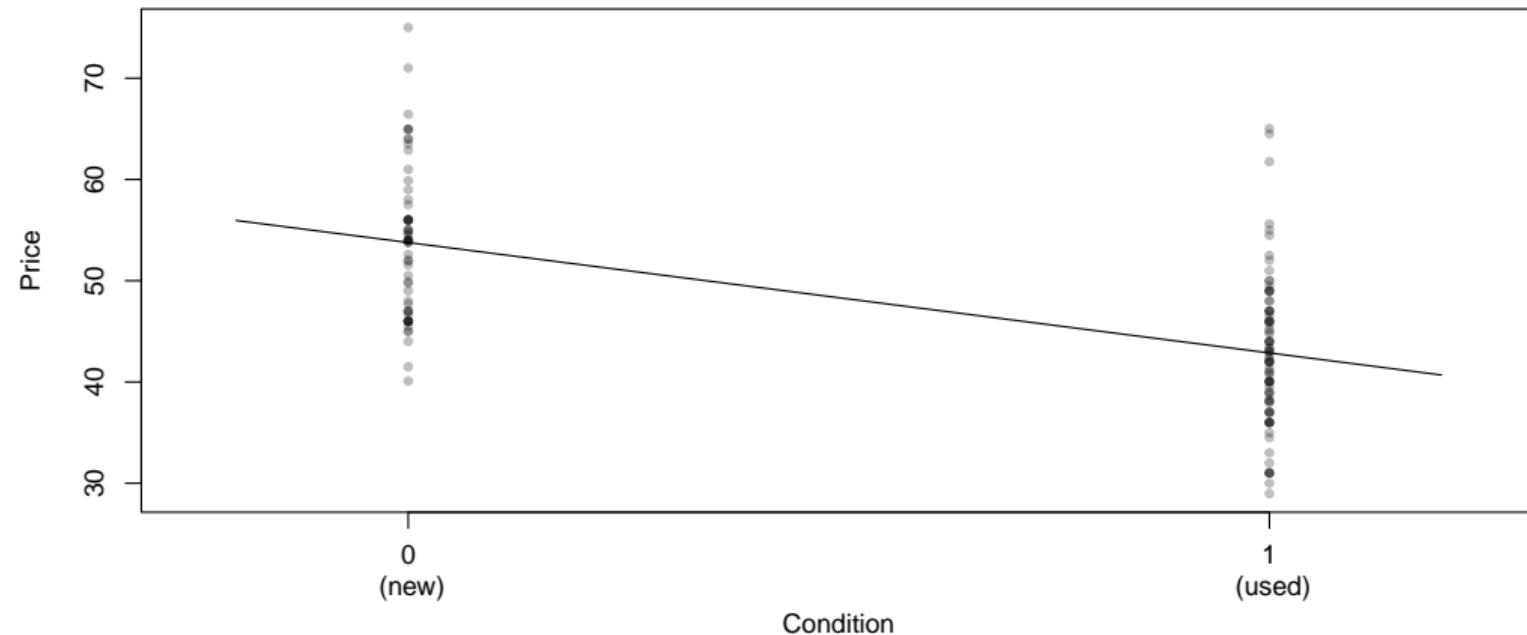
Visualisation: Mario Kart



Dummy (or indicator) variables

- The boxplot suggests a way forward
- Relabel (or encode) the condition variable to take numeric values
 - ▶ One level takes the value 0 (new)
 - ▶ Other level takes the value 1 (used)
- That is, our predictor variable x is
 - ▶ 0 if cond = new
 - ▶ 1 if cond = used
- Referred to as a dummy (or indicator) variable
- We now have a quantitative variable and can fit a regression model

Another visualisation: fitted regression



Regression model

- The mean response from a linear regression model: $\mu_y = \beta_0 + \beta_1 x$
 - ▶ The mean response when $x = 0$ (condition = new)

$$\mu_y = \beta_0 + \beta_1 x = \beta_0 + \beta_1 \times 0 = \beta_0$$

- ▶ The mean response when $x = 1$ (condition = used)

$$\mu_y = \beta_0 + \beta_1 x = \beta_0 + \beta_1 \times 1 = \beta_0 + \beta_1$$

- β_0 is the mean response when $x = 0$
 - ▶ β_0 is the expected price when the game is new
- β_1 is the difference in mean response for $x = 1$ compared to $x = 0$
 - ▶ β_1 is the difference in the expected price between used and new games

Fitting the model in R

- To fit the model in R we could obtain the dummy variable ourselves
 - ▶ We don't have to
 - ▶ We will let R do it for us
- We make use of the data type **factor** in R
 - ▶ Used to represent categorical data
- When using a factor in R it automatically includes a dummy variable for us
 - ▶ Value 0: level that comes first in alphabet (for us this is new)
 - ▶ Value 1: other level (for us this is used)
 - This order can be changed: no reason to change it in this course
- We make `cond` a factor variable using `as.factor`

```
mario$cond = as.factor(mario$cond) # cond is now a factor variable
```

Fitting the model in R

```
m_mario = lm(price ~ cond, data = mario)
summary(m_mario)

##
## Call:
## lm(formula = price ~ cond, data = mario)
##

## Residuals:
##    Min     1Q Median     3Q    Max 
## -13.891 -5.831  0.129  4.129 22.149 
## 

## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  53.77      0.96   56.03 < 2e-16 ***
## condused    -10.90      1.26   -8.66  1.1e-14 ***
## ---    
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 7.37 on 139 degrees of freedom
## Multiple R-squared:  0.351, Adjusted R-squared:  0.346 
## F-statistic: 75 on 1 and 139 DF,  p-value: 1.06e-14
```

Mario Kart: interpretation

- The fitted model is

$$\hat{y} = 53.77 - 10.9x, \quad \text{or}$$
$$\widehat{\text{price}} = 53.77 - 10.9 \text{ used}$$

- The estimated expected price for new games is $\hat{\beta}_0 = 53.77$
- The estimated change in expected price for used games (compared to new games) is $\hat{\beta}_1 = -10.9$
 - We could refer to this as an estimated decrease in expected price of 10.9
- Using what we learned for linear regression:
 - We can find confidence intervals for β_1 (or β_0): see below
 - We can conduct hypothesis tests for β_1

Comparison with t.test

- Comparing linear regression (with dummy variable) to the model with two independent groups we find:
 - ▶ The parameter $\beta_0 = \mu_1$, the mean of the first group
 - ▶ The parameter $\beta_1 = \mu_2 - \mu_1$, the difference in means between the groups
- Regression model assumes equal variance: both groups have the same variance
- The independent group model allowed the two groups to have different variances
 - ▶ We can assume both groups have same variance when using t.test
 - Next slide
 - ▶ We can extend regression model to have different variance
 - Actually quite difficult

Comparison with t.test

- To use t.test we find the two groups

```
new = subset(mario, cond == "new")
used = subset(mario, cond == "used")
```

- We then use t.test with option var.equal = TRUE

```
t_mario = t.test(used$price, new$price, var.equal = TRUE)
t_mario
##
##  Two Sample t-test
##
## data: used$price and new$price
## t = -9, df = 139, p-value = 1e-14
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
## -13.39 -8.41
## sample estimates:
## mean of x mean of y
##      42.9      53.8
```

Comparison with t.test

- The confidence interval for $\mu_{\text{used}} - \mu_{\text{new}}$ from t.test

```
t_mario$conf.int  
## [1] -13.387540 -8.411621  
## attr(,"conf.level")  
## [1] 0.95
```

- The confidence interval for β_1 when using linear regression

```
confint(m_mario, parm = 2) # parm = 2 gives CI for 2nd parameter only  
## 2.5 % 97.5 %  
## condused -13.38754 -8.411621
```

- They are identical!

Categorical variable: more than 2 groups

- We may be interested in categorical predictor variables with more than two groups, e.g.
 - ▶ Prioritised ethnicity (assigned to one ethnic group, even if they identify with multiple ethnicities, based on a predefined order of priority)
 - ▶ Highest education level attained (primary, high school, undergraduate, postgraduate)
 - ▶ Fertilizer (in agricultural trial)
 - ▶ Drug (control, drug A, drug B)
 - ▶ etc
- How can we extend the approach above for categorical predictors with more than two groups?

Example

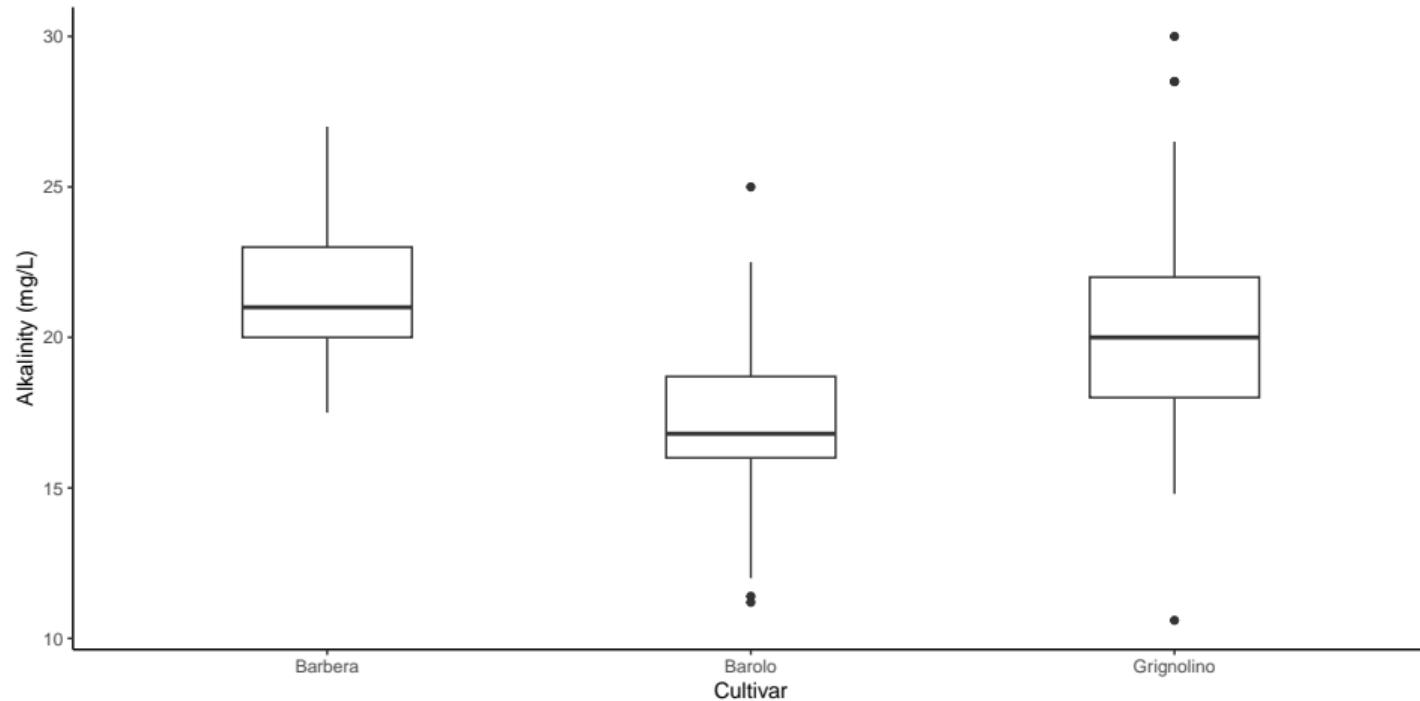
- Data on chemical composition of Italian wines
 - ▶ Three cultivars: barbera, barolo, grignolino
- We will focus on the alkalinity of the wine (measured in mg/L)
- Import the data

```
wine = read.csv('wine.csv')
```

- Look at the data

```
head(wine)  
##   cultivar alkalinity  
## 1 Barolo     15.6  
## 2 Barolo     18.6  
## 3 Barolo     16.0  
## 4 Barolo     18.0  
## 5 Barolo     16.8  
## 6 Barolo     16.0
```

Visualise the data



Statistical model: categorical predictor with K levels

- We can extend the independent group model we have seen earlier
 - ▶ Outcome variable in group 1 is normally distributed with mean μ_1 and variance σ^2
 - ▶ Outcome variable in group 2 is normally distributed with mean μ_2 and variance σ^2
 - ▶ ...
 - ▶ Outcome variable in group K is normally distributed with mean μ_K and variance σ^2
- Assume the variance is the same for all groups
- This is called an ANOVA (analysis of variance) model
 - ▶ More precisely, it is a one-way ANOVA model
- Again, this model is a special case of a linear regression
 - ▶ STAT 210 explores (and exploits) the connection in more detail

Big picture: what do we want to know

- What do we want to know: how do the mean outcome differ between groups?
 - ▶ We could look at pairwise differences in the means
 - Is there a difference in the mean alkalinity between Barbera and Grignolino
 - ▶ This approach is unreliable, particularly when there are a lot of groups (large K)
 - End up making many comparisons: with 10 groups there are 45 pairwise comparisons
 - Increased chance of finding a difference, even if there is no difference in the population
 - Look at this more in the next lecture, and later in course

Hypothesis test

- Start with a slightly different question: does the mean outcome from any group differ from the mean outcome in the other groups?
 - ▶ Is there a difference in the mean alkalinity among any of the cultivars?
- We can express this as a set of hypotheses
 - ▶ $H_0 : \mu_1 = \mu_2 = \dots = \mu_K$
 - ▶ $H_A : \text{at least one mean is different}$
- Develop a hypothesis test to simultaneously compare the mean of all groups
 - ▶ Next lecture

Summary

- Categorical predictor variables
- Include them in a linear regression
 - ▶ Dummy (indicator) variables
 - ▶ Relabel the two groups as 0/1
- Equivalence of linear regression (with categorical predictor) and difference in two means (independent groups)
- Introduced categorical variables with more than two groups
 - ▶ ANOVA model

STAT 110: Week 9

University of Otago

Outline

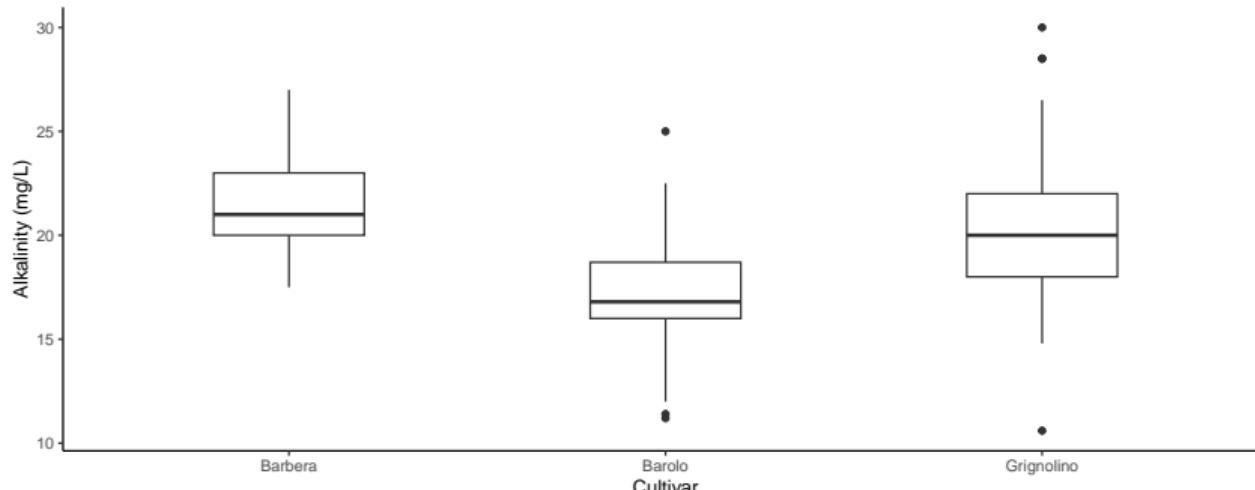
- Fitting ANOVA model
- Understanding ANOVA table
 - ▶ Comparing the variance within a group, to the variance between groups
- Look at multiple comparisons
 - ▶ Pairwise differences

Recall: chemical composition of Italian wines

- We are looking at alkalinity of the wine (measured in mg/L)
 - ▶ Three cultivars: barbera, barolo, grignolino
- Import the data

```
wine = read.csv('wine.csv')
```

- Look at the data



Recall: ANOVA

- One-way ANOVA model with K groups
 - ▶ Outcome variable in group j is normally distributed with mean μ_j and variance σ^2
- We want to know how the mean outcome differs among groups
 - ▶ Potential problems with multiple comparisons
- Are there any differences in mean outcome among the groups?
- This takes the form of a hypothesis test
 - ▶ $H_0 : \mu_1 = \mu_2 = \dots = \mu_K$
 - ▶ $H_A : \text{at least one mean is different}$

In R

- As with categorical variables with 2 levels
 - ▶ Special case of linear regression
 - ▶ Categorical variables can be included in R as factors

```
wine$cultivar = as.factor(wine$cultivar)
```

- We can then fit a linear regression model

```
m_wine = lm(alkalinity ~ cultivar, data = wine)
```

- This fits the ANOVA model
- Problem: output from `m_wine` is not in a convenient form
 - ▶ Output is in terms of particular pairwise comparisons

In R

- We use the `aov` function to get the results in more convenient form

```
a_wine_lm = aov(m_wine)
```

- We can also use `aov` directly

```
a_wine = aov(alkalinity ~ cultivar, data = wine)
```

- The output we will consider is an ANOVA table

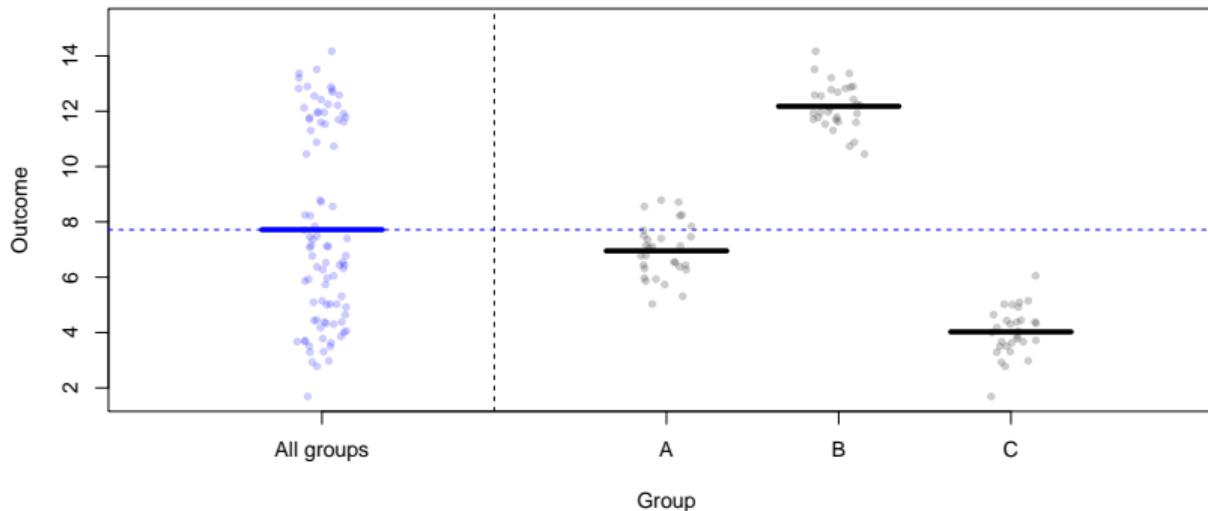
```
summary(a_wine)

##           Df Sum Sq Mean Sq F value    Pr(>F)
## cultivar     2    573     286    35.8 9.4e-14 ***
## Residuals 175   1401      8
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

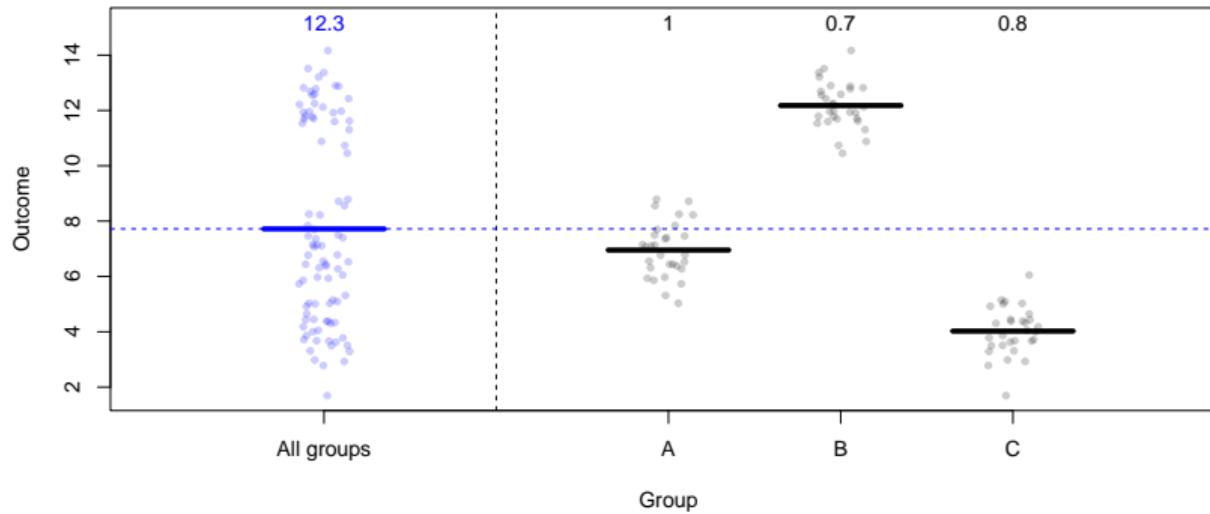
- Take a graphical look at the ANOVA model to help explain what this tells us

Understanding ANOVA (analysis of variance)

- Left plot (blue): plot of all outcome variables (irrespective of group)
- Right three plots (black): plot of outcome variables by group
- Solid horizontal lines: means
 - ▶ Dashed blue line is the overall mean

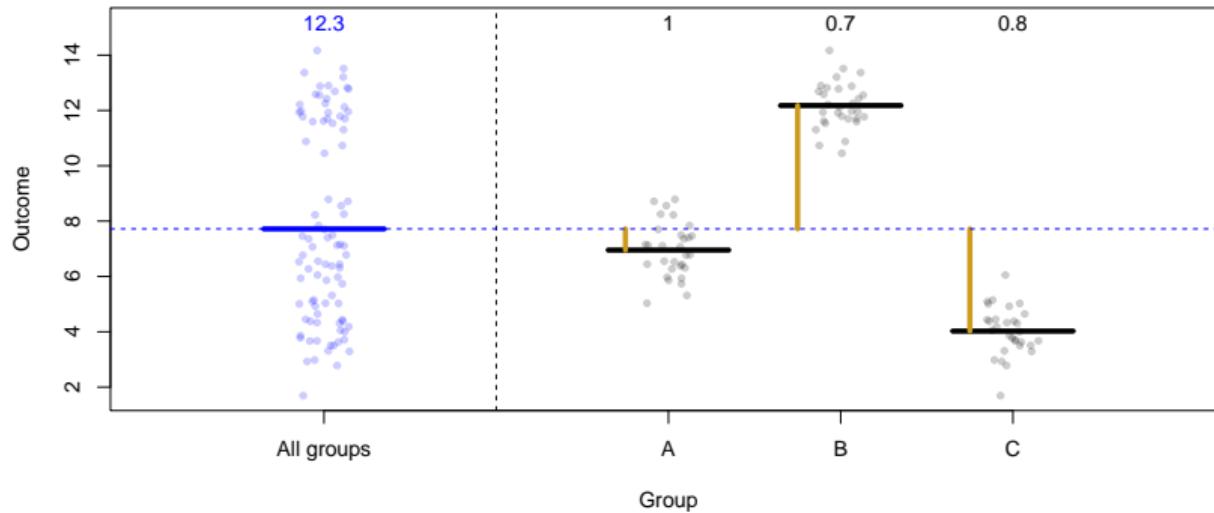


Comparing variance



- The sample variance for each group is given on the plot above
 - ▶ Combined data (blue): outcomes are highly variable
 - ▶ Data from each group (black; A, B, C): outcomes have much lower variability
- The group variable has explained a lot of the variability in the data

Comparing variance



- Overall variability partitioned into:
 - ▶ Variability in group means (indicated by gold lines)
 - ▶ Variability within the groups (points around their mean)
- This is the information summarized in the ANOVA table

ANOVA table

- The ANOVA table for the wine data is

```
##                Df Sum Sq Mean Sq F value Pr(>F)
## cultivar        2    573     286    35.8 9.4e-14 ***
## Residuals     175   1401      8
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- To explain what this represents we will use the table:

Source	Df	Sum Sq	Mean Sq	F value
Group	$K - 1$	GSS	$GMS = \frac{GSS}{DF}$	$F = \frac{GMS}{RMS}$
Residuals	$n - K$	RSS	$RMS = \frac{RSS}{DF}$	
Total	$n - 1$	TSS		

ANOVA table: rows

Source	Df	Sum Sq	Mean Sq	F value
Group	$K - 1$	GSS	$GMS = \frac{GSS}{DF}$	$F = \frac{GMS}{RMS}$
Residuals	$n - K$	RSS	$RMS = \frac{RSS}{DF}$	
Total	$n - 1$	TSS		

- Group row: describes the variation between group means
 - ▶ Variation represented by gold bar in plot above
- Residuals row: describes the variation within each group
- Total row: describes the variation when we combine across groups
 - ▶ Data represented in blue in plot above
 - ▶ This row is not in R output

ANOVA table: columns

Source	Df	Sum Sq	Mean Sq	F value
Group	$K - 1$	GSS	$GMS = \frac{GSS}{DF}$	$F = \frac{GMS}{RMS}$
Residuals	$n - K$	RSS	$RMS = \frac{RSS}{DF}$	
Total	$n - 1$	TSS		

- Mean Sq[uare]
 - ▶ Group (GMS): related to the between-group variance
 - ▶ Residual (RMS): estimate of within-group variance
- F value: ratio of group mean square and residual mean square
- Df: degrees of freedom
- Sum Sq: sum of squares
 - ▶ Convenient when calculating by hand

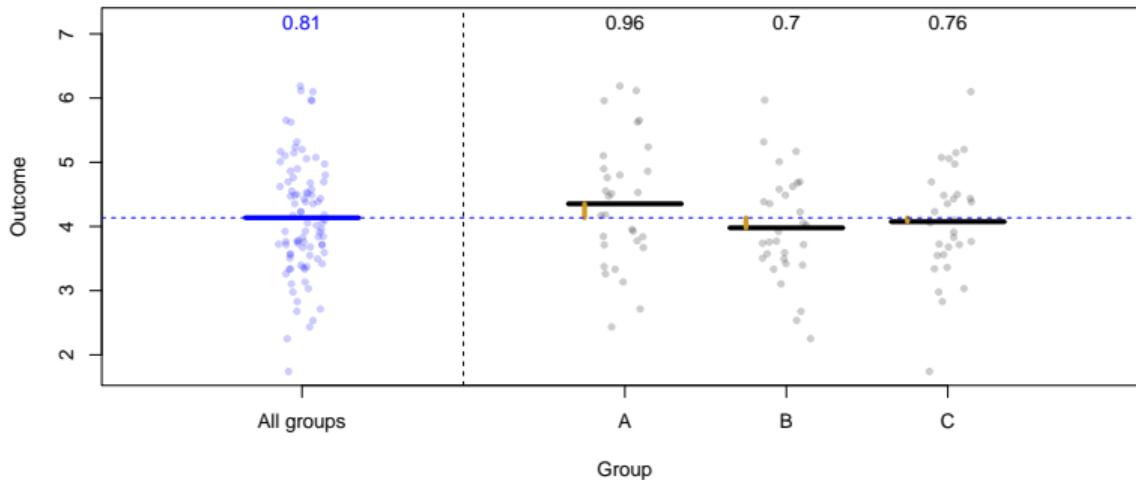
ANOVA table

Source	Df	Sum Sq	Mean Sq	F value
Group	$K - 1$	GSS	$GMS = \frac{GSS}{DF}$	$F = \frac{GMS}{RMS}$
Residuals	$n - K$	RSS	$RMS = \frac{RSS}{DF}$	
Total	$n - 1$	TSS		

- If the groups explain a lot of variability (like our plots above)
 - ▶ The group mean square will be large relative to residual mean square
 - ▶ F-value will be relatively large
 - ANOVA table below is for data from plots above

```
##          Df Sum Sq Mean Sq F value Pr(>F)
## group      2   1024    512     635 <2e-16 ***
## Residuals  87     70      1
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Example II: group does not explain much variation



- The group mean square will not be large relative to residual mean square
- The F-value is not large

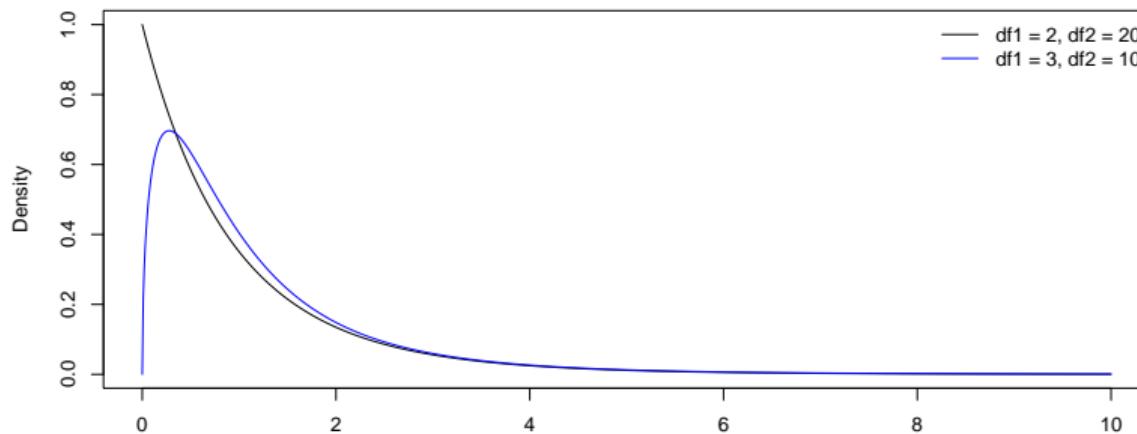
```
##           Df Sum Sq Mean Sq F value Pr(>F)
## group       2    2.3   1.135   1.41   0.25
## Residuals  87  70.1   0.805
```

ANOVA table: F column

- The F-value is comparing the variance among groups (the variability in the group means) to the variance within the groups
 - ▶ It is a measure of how much variation in the data is explained by the groups compared to unexplained variation
- If the null hypothesis is true
 - ▶ Data come from the ANOVA model with all means equal ($\mu_1 = \mu_2 = \dots = \mu_k$)
 - The data are normally distributed with the same mean and variance
 - ▶ F-statistic will have an F-distribution with Df (group), Df (residual) degrees of freedom
- We can use this to find a *p*-value
 - ▶ Quantify the incompatibility between the data and null hypothesis
 - ▶ Are the data unusual given that the null hypothesis is true (group means are the same)
- If null hypothesis is true, we expect an F-value of around 1

Detour: F-distribution

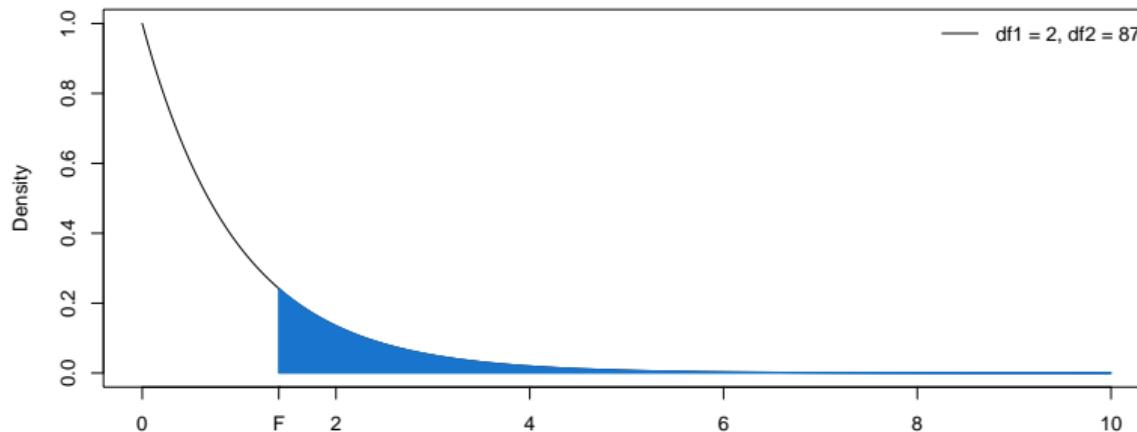
- The F-distribution is a distribution for positive random variables



- ▶ It is asymmetric (positively skewed)
- ▶ It has two parameters:
 - Degrees of freedom for the numerator (df1)
 - Degrees of freedom for the denominator (df2)

Finding a p -value

- An extreme F-value is as large, or larger, than that observed
 - ▶ Indicative of groups explaining as much, or more, variation in the data



- The blue area is given by $1 - pf(F, df_1, df_2)$
 - ▶ $pf(F, df_1, df_2)$ gives probability of a value less than F

Example II

- The ANOVA table for example II is

```
##          Df Sum Sq Mean Sq F value Pr(>F)
## group      2     2.3   1.135   1.41   0.25
## Residuals  87    70.1   0.805
```

- The observed F-statistic is 1.41
 - df1 is degrees of freedom for group: 2
 - df2 is degrees of freedom for residuals: 87
- The p-value is

```
1-pf(1.41, 2, 87)
## [1] 0.25
```

- In practice: refer to the Pr(>F) column in the output

In R: wine data

- The ANOVA table for the wine data is

```
##          Df Sum Sq Mean Sq F value Pr(>F)
## cultivar     2    573     286    35.8 9.4e-14 ***
## Residuals   175   1401      8
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- The F-value is large, p -value is small
 - p -value $< \alpha$: evidence of incompatibility between data and null hypothesis
 - Data are (highly) unusual if all the means were truly the same
 - Providing evidence that at least one of the means differ
- Which groups have means that appear to differ?

Pairwise comparisons of group means

- To compare each group, there are (potentially) many comparisons
 - ▶ If we have $K = 3$ groups: 3 comparisons
 - ▶ If we have $K = 5$ groups: 10 comparisons
 - ▶ If we have $K = 10$ groups: 45 comparisons
- E.g. for $K = 3$: conduct hypothesis tests or find confidence intervals:
 - ▶ CI for $\mu_1 - \mu_2$; hypothesis test with $H_0 : \mu_1 - \mu_2 = 0$
 - ▶ CI for $\mu_1 - \mu_3$; hypothesis test with $H_0 : \mu_1 - \mu_3 = 0$
 - ▶ CI for $\mu_2 - \mu_3$; hypothesis test with $H_0 : \mu_2 - \mu_3 = 0$

Multiple comparisons

- The problem with multiple tests (or multiple confidence intervals) is that properties no longer hold. For hypothesis testing:
 - ▶ α gives the type I error rate for a single test
 - Probability of α of a 'false positive' given that the null hypothesis is true
 - ▶ In each test, there is a chance of a false positive (type I error)
 - ▶ With multiple tests, the overall chance of a type I error increases
 - ▶ Overall type I error rate: referred to as the family-wise error rate
 - Probability of making at least one type I error when performing multiple tests
 - ▶ Multiple comparisons increase the family wise error rate
 - e.g. if we perform 10 independent tests with $\alpha = 0.05$, then the probability of at least one type I error is $1 - 0.95^{10} = 0.4$, if the null hypothesis is true in each instance
 - Probability found using complements

Tukey HSD

- Tukey's honest significant difference (HSD) is a multiple comparison approach designed for ANOVA models
- If the sample sizes are the same in each group
 - ▶ Family-wise error rate is exactly α
- If the sample sizes are different among groups
 - ▶ It is conservative (family-wise error rate is less than α)
- The Tukey approach finds corrected confidence intervals and p -values
- It is easily implemented in R: `TukeyHSD`

In R: wine data

```
TukeyHSD(a_wine)

## Tukey multiple comparisons of means
## 95% family-wise confidence level
##
## Fit: aov(formula = alkalinity ~ cultivar, data = wine)
##
## $cultivar
##          diff    lwr     upr p adj
## Barolo-Barbera -4.38 -5.68 -3.0792 0.000
## Grignolino-Barbera -1.18 -2.43  0.0712 0.069
## Grignolino-Barolo   3.20  2.02  4.3791 0.000
```

Interpretation: wine data

- Interpret the adjusted confidence intervals, e.g.
 - ▶ We are 95% confident that the difference in mean alkalinity between the Grignolino and Barolo cultivars is between 2.02 and 4.38
- Interpret the adjusted p -values, e.g.
 - ▶ The p -value for the difference between Grignolino and Barbera cultivars is 0.069.
 - ▶ As p -value $> \alpha$ there is no evidence that the observed difference is unusual given the null hypothesis that the two means are the same
 - ▶ Note: the uncorrected p -value is 0.027

ANOVA: big picture

- We have looked at fitting one-way a ANOVA model
 - ▶ One-way refers to one categorical predictors: cultivar (for wine example)
 - ▶ Two-way ANOVA: have two categorical predictors
- There might be many other potential predictors (categorical or continuous)
 - ▶ e.g. vineyard, climate (temperature, rainfall), fertilizer used, etc
- Recall: ANOVA is a special case of linear regression
 - ▶ We can use multiple linear regression to include these other variables
- There are lots of possible extensions
- There are also lots of ways to get ourselves into trouble
- These more complex models are explored in STAT 210

Summary

- Looked at the ANOVA summary table
 - ▶ Group: the variation between group means
 - ▶ Residuals: the variation within a group
 - ▶ F-value: comparing the variance within a group, to the variance between groups
- F-distribution to find p -value
- Look at multiple comparisons for pairwise differences
 - ▶ Tukey's honest significant difference
 - ▶ See multiple comparisons in general context later in the course

Outline

- Previous
 - ▶ Exploring (normal) models for continuous data
 - Single mean
 - Two independent groups
 - Paired data
 - Multiple independent groups
 - Linear regression
- Today
 - ▶ Consider data that are not continuous
 - ▶ Explore models for binary data

How well can you putt?

- What is the probability a pro golfer will sink a 6 ft putt?
- Data on professional golfers from 6 feet:
 - ▶ 272 attempts, 149 successes

Problem

- We have been working with models for continuous outcome variables
- This is not continuous data
- It is binary data
 - ▶ Each observation is yes/no, success/failure, 1/0
 - ▶ Each putt will either go in (success), or not (failure)
- Such data arises all the time
 - ▶ Will you support candidate X in the next election?
 - ▶ Did the chick successfully fledge?
 - ▶ Did the participant select option A (or B)?
 - ▶ Did the home team win the football match?
- We need a model for binary data
 - ▶ Probability distribution for binary data

Bernoulli distribution

- Recall: discrete probability distributions
- Random variable Y with two possible outcomes: success/failure
 - ▶ Represent success with 1
 - ▶ Represent failure with 0
- These two outcomes have associated probabilities
 - ▶ Earlier in semester: we assigned them actual numbers, e.g. 0.6 and 0.4
 - ▶ Now: represent the probability of success with an (unknown) parameter: p
- That gives the probability distribution

i	1	2	Total
y_i	0	1	
$\Pr(Y = y_i)$	$1 - p$	p	1

Bernoulli distribution: properties

- Recall: we found means and variances of discrete probability distributions

$$E[Y] = \sum_{i=1}^k y_i \Pr(Y = y_i)$$

$$\text{Var}(Y) = \sum_{i=1}^k (y_i - E[Y])^2 \Pr(Y = y_i)$$

- Using these we can find the mean and variance of a Bernoulli distribution

$$E[Y] = p$$

$$\text{Var}(Y) = p(1 - p)$$

- Extension: Confirm these using the expectation and variance formulae above

Binary to binomial

- We may be interested in cases where there are many binary trials
 - ▶ Flip a coin 15 times
 - ▶ Record the success/failure of 272 putts
- The number of successes from multiple trials has a binomial distribution, if:
 1. The trials are binary
 - The outcome can be represented as success / failure
 2. The number of trials n , is fixed
 - e.g. the number of trials does not depend on the number of successes (or failures) you see
 3. The trials are independent
 - The outcome of one trial does not affect the outcome of another
 4. The probability of success, p , is the same for each trial
 - The probability of success does not change from one trial to another

Binary to binomial

- Let's think about the simplest case
 - ▶ Y_1 and Y_2 are two (independent) random variables
 - ▶ Each of them has a Bernoulli distribution with probability of success p
- Our interest is in the random variable $X = Y_1 + Y_2$
 - ▶ Number of successes from two trials
- If we had two professionals putting from 6 foot
 - ▶ X is a random variable that represents how many putts go in

Binomial distribution: $n = 2$

- The probability distribution of $X = Y_1 + Y_2$ is

i	1	2	3	Total
x_i	0	1	2	
$\Pr(X = x_i)$	$(1 - p)^2$	$2p(1 - p)$	p^2	1

$$\begin{aligned}\Pr(X = 0) &= \Pr(Y_1 = 0 \text{ and } Y_2 = 0) \\&= \Pr(Y_1 = 0) \Pr(Y_2 = 0) \quad \text{multiplication rule: independence} \\&= (1 - p) \times (1 - p)\end{aligned}$$

Binomial distribution: $n = 2$

- The probability distribution of $X = Y_1 + Y_2$ is

i	1	2	3	Total
x_i	0	1	2	
$\Pr(X = x_i)$	$(1 - p)^2$	$2p(1 - p)$	p^2	1

$$\begin{aligned}\Pr(X = 1) &= \Pr(Y_1 = 1 \text{ and } Y_2 = 0) + \Pr(Y_1 = 0 \text{ and } Y_2 = 1) \\&= \Pr(Y_1 = 1) \Pr(Y_2 = 0) + \Pr(Y_1 = 0) \Pr(Y_2 = 1) \quad \text{independence} \\&= p(1 - p) + (1 - p)p\end{aligned}$$

Binomial distribution: general

- In general, the number of successes from n independent Bernoulli trials is:
 - ▶ $X = Y_1 + Y_2 + \dots + Y_n$
- For moderate or large values of n
 - ▶ Possible, but extremely tedious, to write out full probability distribution
- We have a shortcut: we can find the probability of x successes from n independent Bernoulli trials

$$\Pr(X = x) = \binom{n}{x} p^x (1 - p)^{n-x}$$

Binomial distribution: general

- The probability of x successes from n independent Bernoulli trials is

$$\Pr(X = x) = \binom{n}{x} p^x (1 - p)^{n-x}$$

- $\binom{n}{x} = \frac{n!}{x!(n-x)!}$ is the number of ways to obtain x successes from n trials¹
- For each of these, the probability of observing those x successes is $p^x(1 - p)^{n-x}$
 - E.g. there are two ways to see $x = 1$ success from $n = 2$ trials (see above)
 - Each of those has probability $p(1 - p)$
 - E.g. there are 3003 ways to see $x = 5$ successes from $n = 15$ trials
 - Each of these has probability $p^5(1 - p)^{10}$

¹ $x! = x \times (x - 1) \times \dots \times 3 \times 2 \times 1$, e.g. $3! = 3 \times 2 \times 1 = 6$. $x!$ is read as x factorial.

Binomial distribution: general

- The probability of x successes from n independent Bernoulli trials is

$$\Pr(X = x) = \binom{n}{x} p^x (1 - p)^{n-x}$$

- We can use this to find the expectation and variance
 - ▶ The mean of a binomial distribution is $E[X] = np$
 - ▶ The variance of a binomial distribution $\text{Var}(X) = np(1 - p)$
- If there are $n = 100$ putts with probability of success $p = 0.2$, then
 - ▶ $E[X] = np = 100 \times 0.2 = 20$
 - ▶ $\text{Var}(X) = np(1 - p) = 100 \times 0.2 \times 0.8 = 16$
 - ▶ $\text{sd}(X) = \sqrt{\text{Var}(X)} = 4$

Binomial probabilities in R

- We don't have to calculate the long form of that equation
 - ▶ We can use the R function `dbinom`
- Example: what is $\Pr(X = 1)$ when $p = 0.2$ and $n = 2$

```
dbinom(x = 1, size = 2, prob = 0.2)  
## [1] 0.32
```

- The arguments are:
 - ▶ `x = 1`: the number of successes x
 - ▶ `size = 2`: the number of trials n
 - ▶ `prob = 0.2`: the probability of success p
- Check that it gives the correct answer: we know it should be $2p(1 - p)$

```
2*0.2*(1-0.2)  
## [1] 0.32
```

More examples

- If we take 15 putts where there is a probability of 0.7 of making the putt
- What is the probability that we make 10 putts?
- We have $x = 10$, $n = 15$, $p = 0.7$

```
dbinom(x = 10, size = 15, prob = 0.7)  
## [1] 0.206
```

- What is the probability of making 70 putts out of 100 putts with probability 0.6

```
dbinom(x = 70, size = 100, prob = 0.6)  
## [1] 0.01001
```

Back to the data

- We want to estimate the probability of a professional golfer making a 6 foot putt
- What is our statistical model?
 - ▶ Each putt is the outcome of an independent Bernoulli trial with probability p
 - ▶ Equivalently, the total number of successful putts is binomially distributed
- We want to estimate a parameter (population) with a statistic (sample)
 - ▶ (Reasonably) obvious statistic: sample proportion x/n
- For golf data:

$$\hat{p} = \frac{x}{n} = \frac{149}{272} = 0.548$$

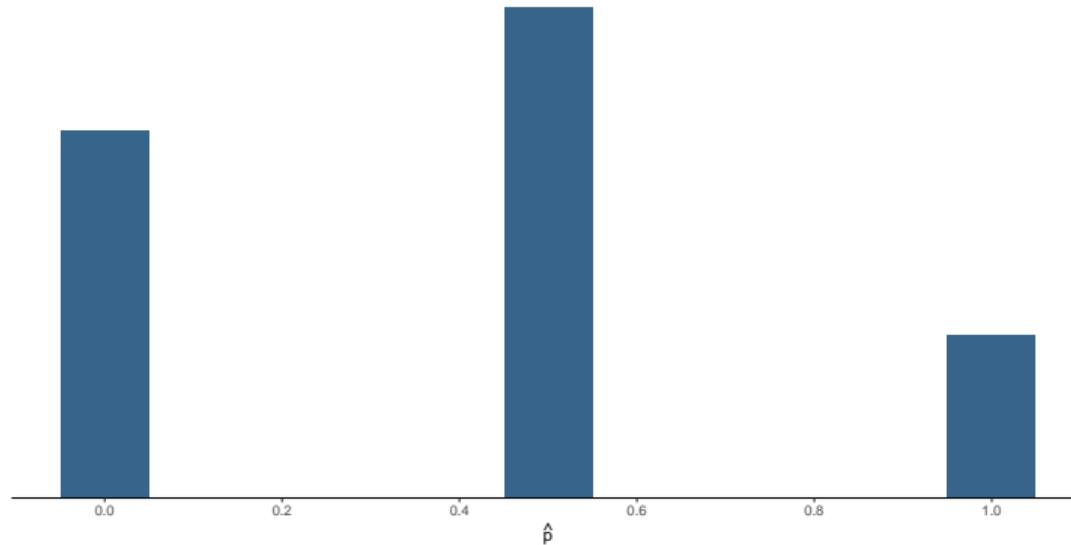
- Recall: \hat{p} is the estimate of parameter p

Confidence interval

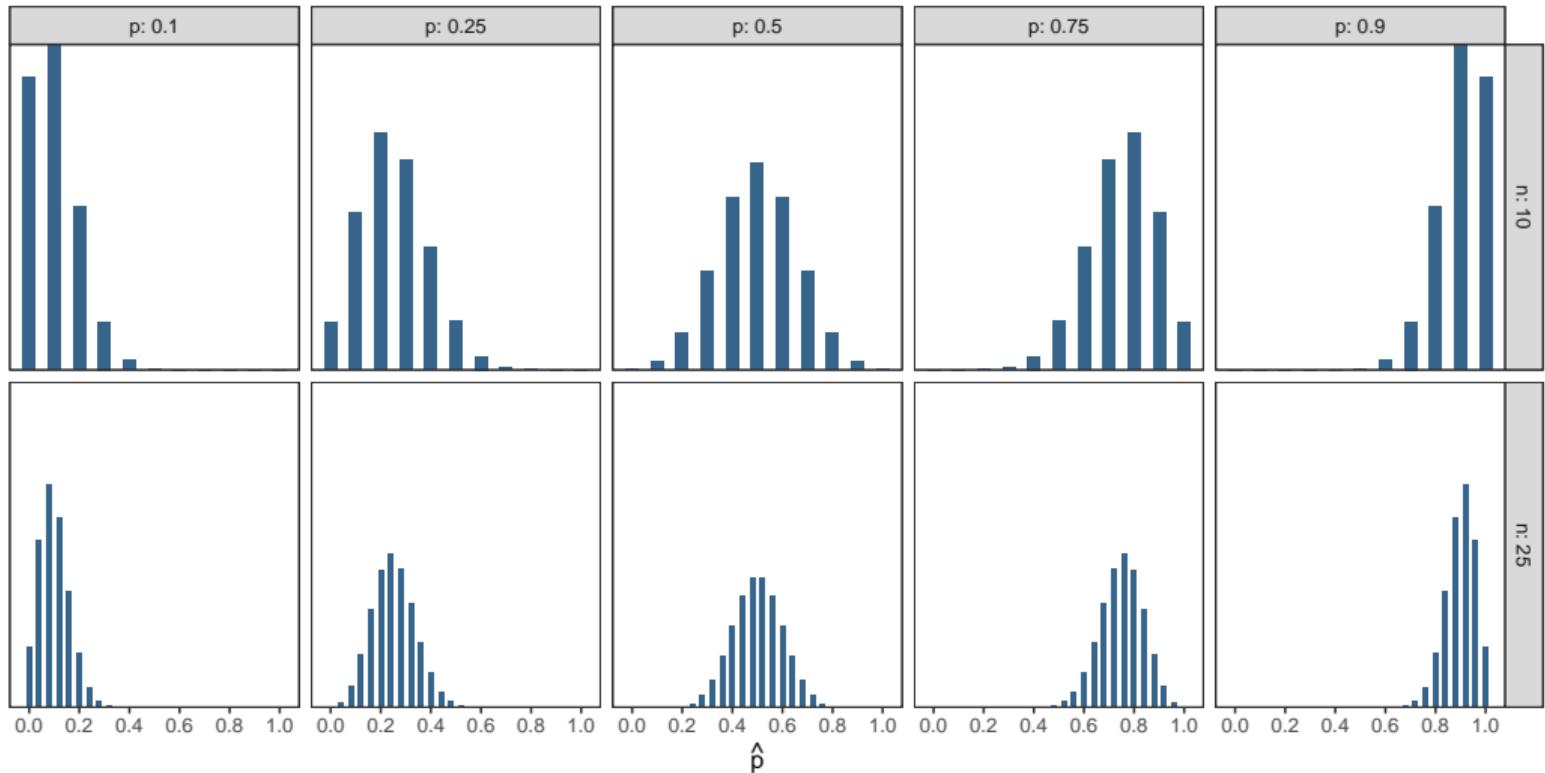
- How do we find a confidence interval?
- Recall: normal model
 - ▶ Found the sampling distribution
 - ▶ Obtained a confidence interval from the sampling distribution
- Can we do the same thing here?
 - ▶ The sampling distribution is the distribution of \hat{p} if we take repeated samples
- Look at it graphically

Sampling distribution for \hat{p} : Start small with $n = 2$ and $p = 0.4$

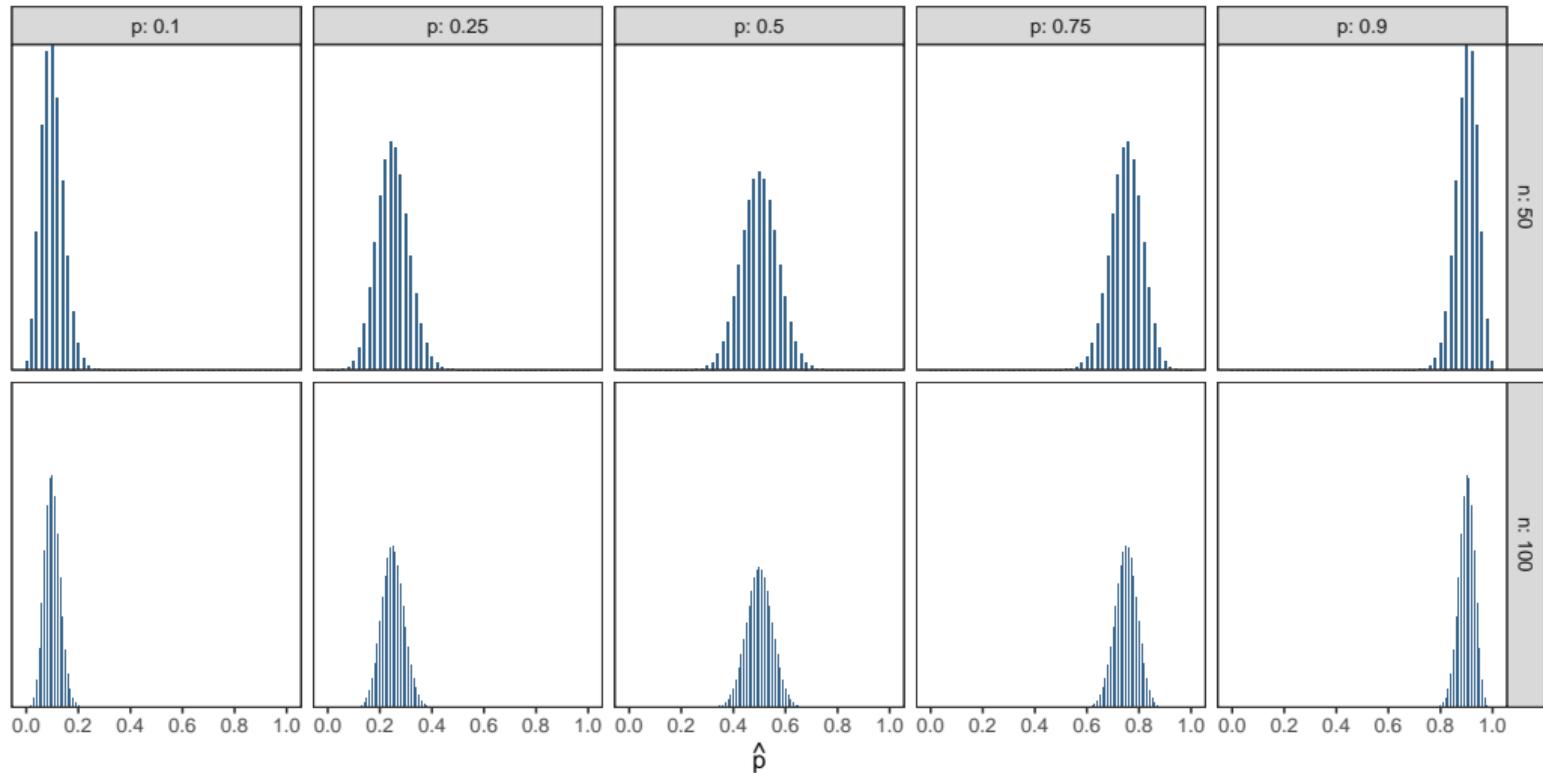
- There are three possibilities:
 - ▶ Observe $x = 0$ with probability 0.36: estimate $\hat{p} = 0$
 - ▶ Observe $x = 1$ with probability 0.48: estimate $\hat{p} = 0.5$
 - ▶ Observe $x = 2$ with probability 0.16: estimate $\hat{p} = 1$



Same principle, but increase the number of trials

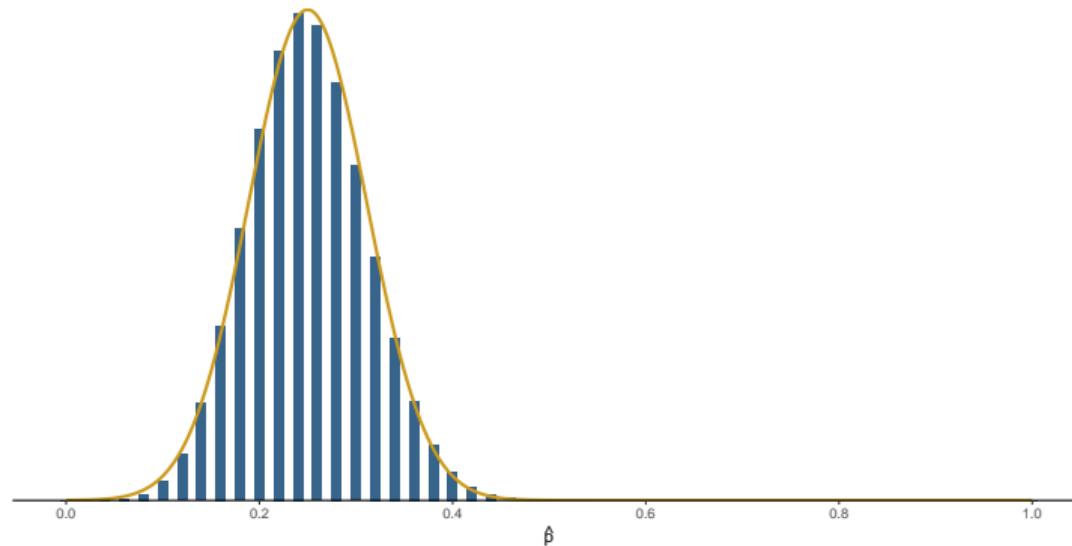


Increase the number of trials some more



Sampling distribution

- As the sample size gets larger, the sampling distribution looks increasingly normal
 - ▶ Normal pdf given in gold
- Example: $n = 50, p = 0.25$



Sampling distribution

- We can approximate the sampling distribution by a normal distribution
 - ▶ Provided n is large enough
- There are various rules of thumb used to determine if the normal approximation is appropriate
- One of these is
 - ▶ $np > 10$ and $n(1 - p) > 10$
- As we saw on the plots above, this reflects that
 - ▶ The sampling distribution is increasingly normal as n increases
 - ▶ When p is close to 0 or 1 it takes a larger n for it to approach normality
- In practice we use $n\hat{p}$ and $n(1 - \hat{p})$ to check if a normal approximation is reasonable

Sampling distribution

- We can approximate the sampling distribution by a normal distribution
 - ▶ Provided n is large enough
- The mean and variance are

$$E[\hat{p}] = p$$

$$\text{Var}(\hat{p}) = \frac{p(1-p)}{n}$$

- So the standard error: $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$
- Extension: Derive $E[\hat{p}]$ and $\text{Var}(\hat{p})$
 - ▶ We have $\hat{P} = \frac{X}{n}$ where $E[X] = np$ and $\text{Var}(X) = np(1-p)$

Confidence interval in R

- We use the normal approximation to find a confidence interval: `prop.test`

```
n = 272; x = 149
prop.test(x, n)

##
## 1-sample proportions test with continuity correction
##
## data: x out of n, null probability 0.5
## X-squared = 2.3, df = 1, p-value = 0.13
## alternative hypothesis: true p is not equal to 0.5
## 95 percent confidence interval:
## 0.48656 0.60766
## sample estimates:
##      p
## 0.54779
```

- We are 95% confident that the probability of a professional golfer making a putt from 6 feet is between 0.487 and 0.608

Hypothesis test

- We can also test the hypothesis
 - ▶ $H_0 : p = p_0$
 - ▶ $H_A : p \neq p_0$
- `prop.test` defaults to $p_0 = 0.5$
 - ▶ It can be changed with option `p`, e.g. `p = 0.4`

```
prop.test(x, n, p = 0.4)
```

- For the putting data with $p_0 = 0.5$ we have a p-value of 0.13
 - ▶ This quantifies the incompatibility between the data and null hypothesis
 - ▶ Since $p\text{-value} > \alpha = 0.05$ there is no evidence that the data are unusual given the null hypothesis is true
 - The data we have observed would not be unusual if professionals truly sank 50% of their putts from 6 feet

Summary

- Introduced binary data
- Bernoulli distribution for binary observations
- The number of successes from multiple binary trials have binomial distribution
 - ▶ Several conditions need to be satisfied
- Use a binomial model to find:
 - ▶ Confidence interval for p
 - ▶ Hypothesis test
 - We will look more into these in the next lecture

Outline

- A closer look at confidence intervals and hypothesis tests for p
- Extending the model
 - ▶ Compare probabilities between two (independent) groups
- Difference in proportions: $p_1 - p_2$
 - ▶ Confidence interval
 - ▶ Hypothesis test

Recall: Golf putting

- What is the probability a professional golfer makes a putt from 6 feet?
 - ▶ $n = 272$ putts with $x = 149$ made

```
n = 272; x = 149
prop.test(x, n)

##
## 1-sample proportions test with continuity correction
##
## data: x out of n, null probability 0.5
## X-squared = 2.3, df = 1, p-value = 0.13
## alternative hypothesis: true p is not equal to 0.5
## 95 percent confidence interval:
## 0.48656 0.60766
## sample estimates:
##      p
## 0.54779
```

Finding confidence interval for p

- We found the confidence interval in R
 - ▶ We haven't yet described where it comes from (like we normally do)
- It turns out there are many possible confidence intervals for p
 - ▶ The `binomCI` package in R gives the choice of 15 (!) different intervals
- Why are there so many many intervals?
 - ▶ There are many reasons
 - ▶ Most obvious: because the 'standard' confidence interval doesn't work well

Confidence intervals for p

- The ‘standard’ confidence interval can be written as

$$\text{estimate} \pm \text{multiplier} \times \text{std. error}$$

- ▶ Estimate: \hat{p}
- ▶ Multiplier: sampling distribution is approximate normal
 - Multiplier is $z_{1-\alpha/2}$
- ▶ Standard error: $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$
 - Estimate this: $s_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$
- Commonly called a Wald interval
- Similar to what we had for μ

Problems with the Wald interval

- The Wald interval is not very reliable, particularly when n not large, and p close to 0 or 1
 - ▶ Despite this it is still commonly used and seen in textbooks
- Recall: what is a confidence interval?
 - ▶ If we collect multiple datasets with n binary observations from the population of interest and calculate a confidence interval for each:
 - Then 95% of the intervals, on average, should contain the true p ($\alpha = 0.05$)
- The Wald interval does a poor job of this
 - ▶ The interval tends to contain the true value (p) less often than it is supposed to
 - e.g. when $n = 50$ and $p = 0.06$ fewer than 81% of intervals will contain the true p
 - Particularly poor when np or $n(1 - p)$ is small

What about the interval that R gives?

- `prop.test` finds the Wilson (score) interval
- Comparing the Wilson interval to the Wald interval:
 - ▶ Both are based on a normal approximation to the binomial
 - ▶ The Wilson interval is asymmetric
 - It is not found using: $\text{estimate} \pm \text{multiplier} \times \text{standard error}$
 - ▶ It has improved performance when p is close to 0 or 1
 - It is reasonable to use even if $np < 10$ or $n(1 - p) < 10$
 - ▶ We will not delve into the detail
 - It is more complicated
 - Extension: more information is provided at [this link](#) for those who may be interested
- In practice: use Wilson interval found using `prop.test`

Continuity correction

- By default `prop.test` adopts a continuity correction
 - ▶ For confidence intervals and hypothesis tests
- A continuity correction is adjustment that reflects that we are approximating a discrete distribution (binomial) with a continuous distribution (normal)
 - ▶ We make an adjustment of ± 0.5
- If X is a random variable with a binomial distribution, and Z is a random variable with a normal distribution that approximates X , a continuity correction is
 - ▶ $\Pr(X \leq 10) \approx \Pr(Z < 10.5)$
 - ▶ $\Pr(X \geq 5) \approx \Pr(Z > 4.5)$
- It is conservative: makes confidence intervals wider (increases p-value)
- It can be turned off using option `correct = FALSE`
 - ▶ We will use the default settings in `prop.test`

What about the hypothesis test?

- We may wish to test the hypotheses:
 - ▶ $H_0 : p = p_0$
 - ▶ $H_A : p \neq p_0$
- A test statistic can be found using:

$$z = \frac{\text{estimate} - \text{null}}{\text{standard error}} = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

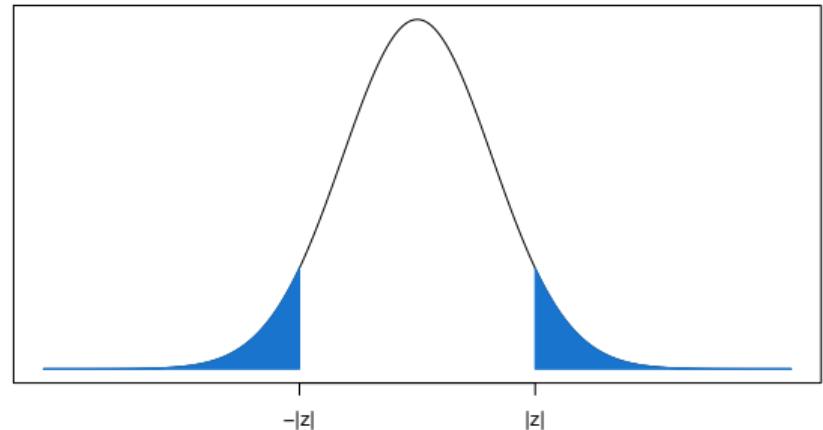
- Two things to note:
 - ▶ Find standard error assuming null hypothesis is true: $\sigma_{\hat{p}} = \sqrt{\frac{p_0(1-p_0)}{n}}$
 - ▶ Find p -value from a (standard) normal distribution
 - That's why the test statistic is z , not t

Hypothesis test: golf

- To test if putting probability is different from 50/50: $p_0 = 0.5$

```
# estimate of p
phat = x/n
p0 = 0.5
# Find standard error under H0
se = sqrt(p0*(1-p0)/n)
# Find test statistic
z = (phat - p0)/se
# Find pvalue
pval = 2*pnorm(-abs(z))
pval
## [1] 0.115
```

```
z
## [1] 1.58
```



Hypothesis testing in R

- `prop.test` conducts the hypothesis test in a slightly different way
 - ▶ By default it uses a continuity correction
 - ▶ Uses χ^2 test statistic² rather than z
 - Performing the same test, but in a different way
 - Details are outside the scope of the course (see STAT 270)
 - ▶ If the correction was turned off (`correct = FALSE`)
 - Obtain an identical p -value to our procedure above
 - ▶ Alternatively, we could include a continuity correction in our p -value calculation
 - We would find an identical p -value to that from `prop.test`
 - Details outside the scope of the course

² χ is the greek letter chi, pronounced kai (rhymes with sky).

Data: Smallpox in Boston

- Data are 6224 observations from individuals in Boston in 1721 who were exposed to smallpox³
 - ▶ Inoculated: yes or no
 - ▶ Result: lived or died
- We are interested in comparing the probability of death for those who were inoculated to those who were not

		inoculated		
		yes	no	Total
result	lived	238	5136	5374
	died	6	844	850
	Total	244	5980	6224

³This is the same data that we saw in week 2.

Models for binomial data

- We don't have the tools to answer the question
 - ▶ We only know how to estimate p , not compare p across two groups
- We can look at model extensions for binomial data that parallel those we explored for normal models, e.g.
 - ▶ Comparing two or more independent groups
 - ▶ Regression-type models: probability of success depends on predictor variables
 - Called logistic regression
 - ▶ Defer many of these extensions to later courses (i.e. STAT 210)
- For smallpox data: two independent binomials
 - ▶ Inoculated: modelled as binomial with probability p_1
 - $x_1 = 238, n_1 = 244$
 - ▶ Not inoculated: modelled as binomial with probability p_2
 - $x_2 = 5136, n_2 = 5980$

Big picture

- We want to compare the survival between inoculated and uninoculated
- There are multiple ways we could do this, e.g.
 - ▶ Difference in probabilities: $p_1 - p_2$
 - ▶ Ratio of probabilities (also called relative risk): p_1/p_2
- We will focus on $p_1 - p_2$
- It is straightforward to estimate this difference
 - ▶ $\hat{p}_1 - \hat{p}_2$
- We also know those estimates are uncertain
 - ▶ Found from data (a sample from the population)
 - ▶ Find a confidence interval

Confidence interval for $p_1 - p_2$

- Find a confidence interval using

estimate \pm multiplier \times standard error

- Estimate: $\hat{p}_1 - \hat{p}_2$
- Multiplier: we again approximate the sampling distribution with normal
 - ▶ Multiplier is $z_{1-\alpha/2}$
- Standard error: $\sigma_{\hat{p}_1 - \hat{p}_2} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$
 - ▶ Estimate this with: $s_{\hat{p}_1 - \hat{p}_2} = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$

Wald confidence interval for $p_1 - p_2$

- Putting this together we have the $100(1 - \alpha)\%$ Wald confidence interval:

$$\hat{p}_1 - \hat{p}_2 \pm z_{1-\alpha/2} \sqrt{\frac{\hat{p}_1(1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2}}$$

- This is the interval returned by `prop.test` when we have two groups
- As with the Wald interval for p
 - The interval is not that reliable if either n_1 or n_2 is small and either p_1 or p_2 is close to 0 or 1
 - Improved confidence intervals do exist
 - e.g. the Newcombe interval is based on Wilson interval
 - Such intervals can be found in other R packages
- We will use the Wald interval in `prop.test`

In R

```
x = c(238, 5136); n = c(244, 5980) # smallpox data
prop.test(x, n)

##
## 2-sample test for equality of proportions with continuity correction
##
## data: x out of n
## X-squared = 26, df = 1, p-value = 3e-07
## alternative hypothesis: two.sided
## 95 percent confidence interval:
## 0.0931 0.1400
## sample estimates:
## prop 1 prop 2
## 0.975 0.859
```

- We are 95% confident that the probability of survival was between 0.093 and 0.14 higher for those who were inoculated compared to those who were not

Hypothesis test

- Both p_1 and p_2 are conditional probabilities
 - ▶ p_1 is the survival probability given inoculated
 - ▶ p_2 is the survival probability given not inoculated
- If $p_1 = p_2$ then survival does not depend on inoculation
 - ▶ Survival and inoculation are independent
- We can test the hypotheses:
 - ▶ $H_0 : p_1 - p_2 = 0$ (this is equivalent to $p_1 = p_2$)
 - ▶ $H_A : p_1 - p_2 \neq 0$ (this is equivalent to $p_1 \neq p_2$)

Hypothesis test

- A test statistic can be found using:

$$z = \frac{\text{estimate} - \text{null}}{\text{standard error}}$$

- Estimate is $\hat{p}_1 - \hat{p}_2$
- Null value is 0
- We need the standard error assuming null hypothesis is true
 - ▶ The two groups have the same probability: $p_1 = p_2$
 - ▶ The null hypothesis doesn't specify what this value is
 - Let's call it p^*

Hypothesis test

- The standard error is: $\sigma_{\hat{p}_1 - \hat{p}_2} = \sqrt{\frac{p^*(1-p^*)}{n_1} + \frac{p^*(1-p^*)}{n_2}}$
 - ▶ This is the standard error above evaluated at $p_1 = p_2 = p^*$
- We don't know p^*
 - ▶ Estimate it: $\hat{p}^* = \frac{\text{total success}}{\text{total trials}} = \frac{x_1 + x_2}{n_1 + n_2} = \frac{n_1 \hat{p}_1 + n_2 \hat{p}_2}{n_1 + n_2}$
 - ▶ \hat{p}^* is sometimes call the pooled proportion
- Use this to estimate the standard error: $s_{\hat{p}_1 - \hat{p}_2} = \sqrt{\frac{\hat{p}^*(1-\hat{p}^*)}{n_1} + \frac{\hat{p}^*(1-\hat{p}^*)}{n_2}}$
- This hypothesis test is found using `prop.test`. As with the test for p :
 - ▶ It uses a different test statistic (χ^2 vs z)
 - ▶ Includes a continuity correct by default

Hypothesis test: in R

- Using `prop.test` to find the *p*-value

```
prop.test(x,n)

##
## 2-sample test for equality of proportions with continuity correction
##
## data: x out of n
## X-squared = 26, df = 1, p-value = 3e-07
## alternative hypothesis: two.sided
## 95 percent confidence interval:
## 0.0931 0.1400
## sample estimates:
## prop 1 prop 2
## 0.975 0.859
```

Interpretation

- The p -value quantifies the incompatibility between the null hypothesis and the data
 - ▶ The p -value $< \alpha = 0.05$, which suggests the data are unusual if the two groups (inoculated and uninoculated) truly had the same probability of survival

Summary

- Look at estimating p
 - ▶ Confidence intervals:
 - Wald interval can be unreliable
 - `prop.test` using more reliable alternative
 - ▶ Hypothesis tests
- Explored comparison between two groups: $p_1 - p_2$
 - ▶ Confidence intervals
 - ▶ Hypothesis test

STAT 110: Week 10

University of Otago

Outline

- Contingency table
 - ▶ Looking at the relationship between two categorical variables
 - ▶ Investigate approaches to test independence of two categorical variables
 - ▶ Compare observed and expected counts
 - ▶ Introduce χ^2 distribution
- Central limit theorem
 - ▶ Investigate the sampling distribution for non-normal data
 - ▶ Generalise what was done for binomial data

Data: Passengers on the Titanic

- Data from the adult passengers on the titanic. Two variables:
 - ▶ Class: 1st, 2nd, 3rd or crew
 - ▶ Survived: yes or no

Class	survived			Total
	no	yes		
1st	122	197	319	
2nd	167	94	261	
3rd	476	151	627	
Crew	673	212	885	
Total	1438	654	2092	

- Do survival probabilities depend on the class?

Big picture

- We have investigated when both variables have two levels (groups)
- Here one of the variables has four levels
 - ▶ 1st – 3rd class, crew
- If the survival probabilities vary by class
 - ▶ The two variables (class and survival) are related
- If the survival probabilities do not vary by class
 - ▶ The two variables (class and survival) are independent
 - ▶ Knowing the class of a passenger tells us nothing about their survival probability
 - ▶ Recall: Definition of independence when we looked at probability
- Idea: Compare the observed data to what we would expect if two variables were independent

Expected counts

- We can use the margin totals to find the expected counts under independence

$$\text{expected count} = \frac{\text{row total} \times \text{column total}}{\text{table total}}$$

- Work through the Titanic table to understand this

Expected counts: Titanic

$$\text{expected count} = \frac{\text{row total} \times \text{column total}}{\text{table total}} = \frac{319 \times 654}{2092} = 99.73$$

Class	survived		Total
	no	yes	
1st		99.73	319
2nd			261
3rd			627
Crew			885
Total	1438	654	2092

- Proportion of passengers who are 1st class
 - ▶ $\frac{\text{row total}}{\text{table total}} = \frac{319}{2092} = 0.1525$
 - ▶ 15.25% of passengers are 1st class
- If survival and class are independent
 - ▶ Expected number is the total number of passengers who survive \times the proportion of passengers who are 1st class
 - ▶ Or column total \times $\frac{\text{row total}}{\text{table total}}$

Expected counts: Titanic

$$\text{expected count} = \frac{\text{row total} \times \text{column total}}{\text{table total}} = \frac{627 \times 1438}{2092} = 430.99$$

Class	survived		Total
	no	yes	
1st		99.73	319
2nd			261
3rd	430.99		627
Crew			885
Total	1438	654	2092

- Proportion of passengers who are 3rd class
 - ▶ $\frac{\text{row total}}{\text{table total}} = \frac{627}{2092} = 0.2997$
 - ▶ 29.97% of passengers are 3rd class
- If survival and class are independent
 - ▶ Expected number is the total number of passengers who died \times the proportion of passengers who are 3rd class
 - ▶ Or column total \times $\frac{\text{row total}}{\text{table total}}$

Expected counts: Titanic

- Put it all together to give observed (black) and expected (blue)

Class		survived		Total
		no	yes	
1st		122 (219.27)	197 (99.73)	319
2nd		167 (179.41)	94 (81.59)	261
3rd		476 (430.99)	151 (196.01)	627
Crew		673 (608.33)	212 (276.67)	885
	Total	1438	654	2092

- The observed and expected counts will vary: there is natural variation in the data
 - Do they vary more than we would expect if variables are truly independent?

Test for independence

- We can look at this with a hypothesis test
 - ▶ H_0 : the two variables are independent
 - ▶ H_A : the two variables are related
- The test statistic we will use is

$$X^2 = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$$

- ▶ For each cell we calculate $\frac{(\text{observed} - \text{expected})^2}{\text{expected}}$ and add them up

Test statistic

Class		survived		Total
		no	yes	
	1st	122 (219.27)	197 (99.73)	319
	2nd	167 (179.41)	94 (81.59)	261
	3rd	476 (430.99)	151 (196.01)	627
	Crew	673 (608.33)	212 (276.67)	885
	Total	1438	654	2092

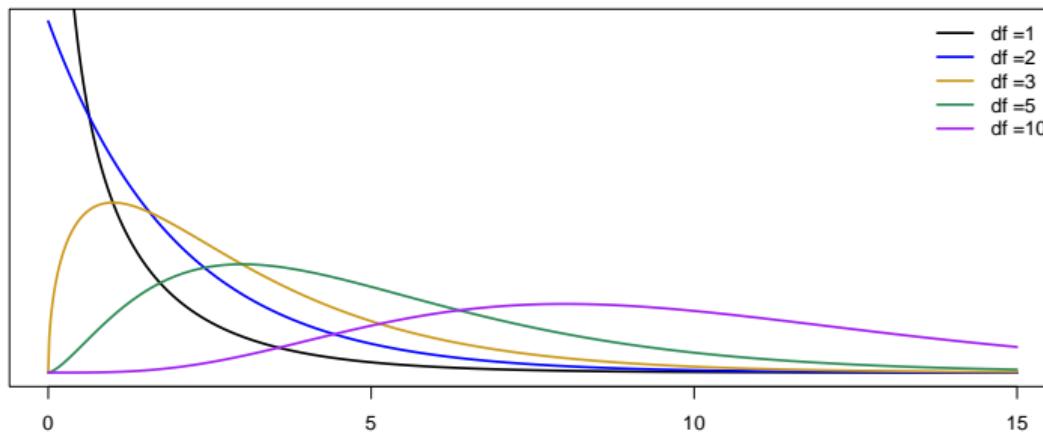
$$\begin{aligned} X^2 &= \frac{(122 - 219.27)^2}{219.27} + \frac{(197 - 99.73)^2}{99.73} + \dots + \frac{(212 - 276.67)^2}{276.67} \\ &= 177.8 \end{aligned}$$

Test statistic

- If the null hypothesis is true
 - ▶ The test statistic, X^2 , will be a realisation from a χ^2 -distribution with $(R - 1) \times (C - 1)$ degrees of freedom
 - R is the number of rows; C is the number of columns
- Titanic data: $R = 4$, $C = 2$
 - ▶ $df = (4 - 1) \times (2 - 1) = 3$

Detour: χ^2 -distribution

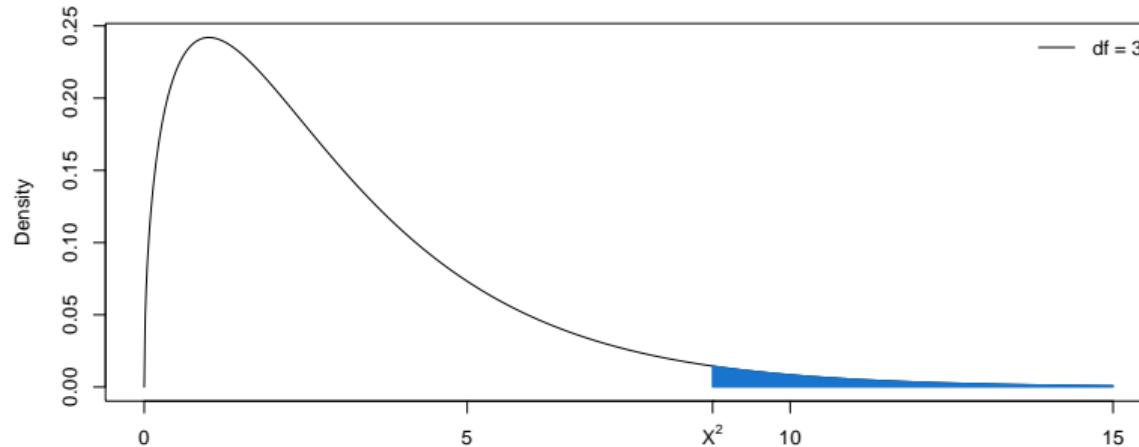
- The χ^2 -distribution is a distribution for positive random variables



- It is asymmetric (positively skewed)
- It has one parameters: degrees of freedom

Finding a p -value

- An extreme X^2 -value is one that is as large, or larger, than that observed
 - ▶ Indicative of increased divergence between observed and expected counts



- The p -value (blue area) is given by $1 - \text{pchisq}(X2, \text{df})$
 - ▶ $\text{pchisq}(X2, \text{df})$ gives probability of a value less than X^2

In R

- Data: each row is an observation
 - ▶ Titanic data: each row is a passenger
- Import into R

```
titanic = read.csv('titanic.csv')  
head(titanic)
```

```
##   Class Survived  
## 1   Crew      Yes  
## 2   Crew      Yes  
## 3   2nd       No  
## 4   1st       Yes  
## 5   Crew      Yes  
## 6   3rd       No
```

In R

- We use the `table` function to obtain contingency table

```
titan = table(titanic$Class, titanic$Survived)
```

- ▶ First argument: variable 1 (class of passenger)
- ▶ Second argument: variable 2 (survived: yes / no)

```
titan  
##  
##          No Yes  
## 1st    122 197  
## 2nd    167  94  
## 3rd    476 151  
## Crew   673 212
```

```
addmargins(titan)  
##  
##          No Yes Sum  
## 1st    122 197 319  
## 2nd    167  94 261  
## 3rd    476 151 627  
## Crew   673 212 885  
## Sum    1438 654 2092
```

- The function `addmargins` includes the margins on the table

In R

- The R function `chisq.test` evaluates the test

```
out1 = chisq.test(titan)

out1

##
##  Pearson's Chi-squared test
##
## data: titan
## X-squared = 177.8, df = 3, p-value <2e-16
```

- The p -value $< \alpha = 0.05$. Observing a test statistic as large as we did is unusual if the two variables were independent
 - ▶ Evidence in support of H_A : that the variables are not independent

χ^2 -test

- If $R = 2$ and $C = 2$: we have a 2×2 contingency table, e.g. smallpox in Boston
 - ▶ The χ^2 test is identical to test for difference in proportions
 - ▶ $H_0 : p_1 - p_2 = 0$ and $H_A : p_1 - p_2 \neq 0$
- The χ^2 test can also be used if both $R > 2$ and $C > 2$
- The χ^2 test is unreliable if any of the expected counts < 5
 - ▶ Options for resolving this problem are beyond the scope of course

In R

- The `chisq.test` function can return the expected counts

```
out1$expected  
##  
##           No      Yes  
## 1st  219.27  99.726  
## 2nd  179.41  81.594  
## 3rd  430.99 196.012  
## Crew 608.33 276.668
```

- Still important to know:
 - ▶ How to calculate them
 - ▶ What they represent (expected counts if variables are independent)

Normal approximation

- Binomial: The sampling distribution for \hat{p} was approximated by a normal
 - ▶ Provided n is large, and p is not too close to 0 or 1
- This formed the basis for finding confidence intervals (and conducting hypothesis tests)
- Does this result generalise?
 - ▶ Will this also happen for other ‘non-normal’ distributions?

Central limit theorem

- If we collect a large sample of independent observations from a population with mean μ and standard deviation σ , the sampling distribution of \bar{y} will be approximately normal
 - ▶ Mean μ
 - ▶ Standard error $\frac{\sigma}{\sqrt{n}}$
- This is known as the central limit theorem

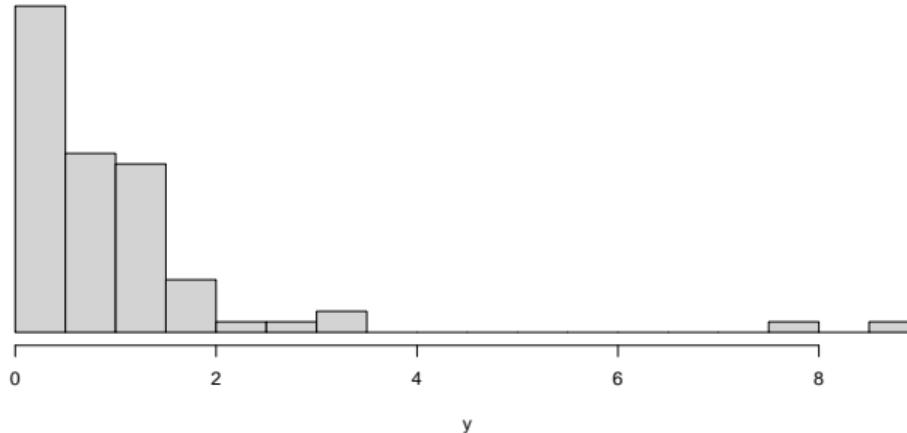
Central limit theorem: notes

- The distribution of y need not be normal
- What is a large sample?
 - ▶ A standard rule of thumb is $n > 30$
 - ▶ Lots of exceptions to this rule, e.g.
 - If the data are highly skewed, we likely need more than 30
 - If there are (extreme) outliers, we likely need more than 30

Central limit theorem: Example

- If data come from a non-normal distribution (an exponential distribution)
- Simulate one data set to see what the data looks like

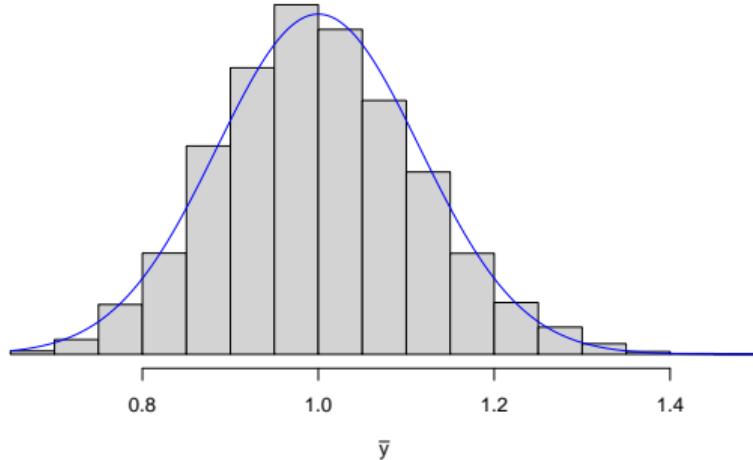
```
### Single sample  
n = 75 # sample size of 75  
y = rexp(n,1) # generate the data  
hist(y) # look at histogram
```



Central limit theorem: Example

- Generate lots of datasets and visualize the sampling distribution
 - ▶ See that it is approximately normal

```
### Taking 10000 samples  
m = 10000; ybar = rep(NA, m)  
for(i in 1:m){ # repeat m times  
  y = rexp(n,1) # simulate data  
  ybar[i] = mean(y) # find the sample mean  
}  
hist(ybar)
```



Central limit theorem: implications

- The approaches we worked through for normal models
 - ▶ Can also be used for non-normal models
 - ▶ Need to ensure a large sample (usually $n > 30$)
- This list includes confidence intervals and hypothesis tests for:
 - ▶ Population mean μ with one sample: use `t.test`
 - ▶ Difference in two means $\mu_1 - \mu_2$: use `t.test`
 - ▶ ANOVA: use `aov`
 - ▶ Linear regression: use `lm`

Central limit theorem: implications

- Model checking: this is why we were only concerned about major departures from normality when the sample size was large
 - ▶ Linear regression
 - ▶ Normal models
- The central limit theorem underpins a lot of statistical practice
 - ▶ Often in the background

Summary

- χ^2 test for independence of contingency table
 - ▶ Idea: compare observed counts with those expected under independence
- Central limit theorem
 - ▶ Sampling distribution is normal
 - ▶ The approaches we have already developed can be used for non-normal data
- CLT holds if sample size is large
 - ▶ Usually $n > 30$

Outline

- Explore some non-parametric methods
- Focus on two examples:
 - ▶ Data from two independent groups
 - ▶ Relationship between two ordinal variables
- Outline other approaches

Data: Hawks

- 100 measurements from two species of hawk
 - ▶ Red-tailed (RT), and Sharp-shinned (SS)
- Hallux measurement (mm): length of the killing talon
- Import the data into R

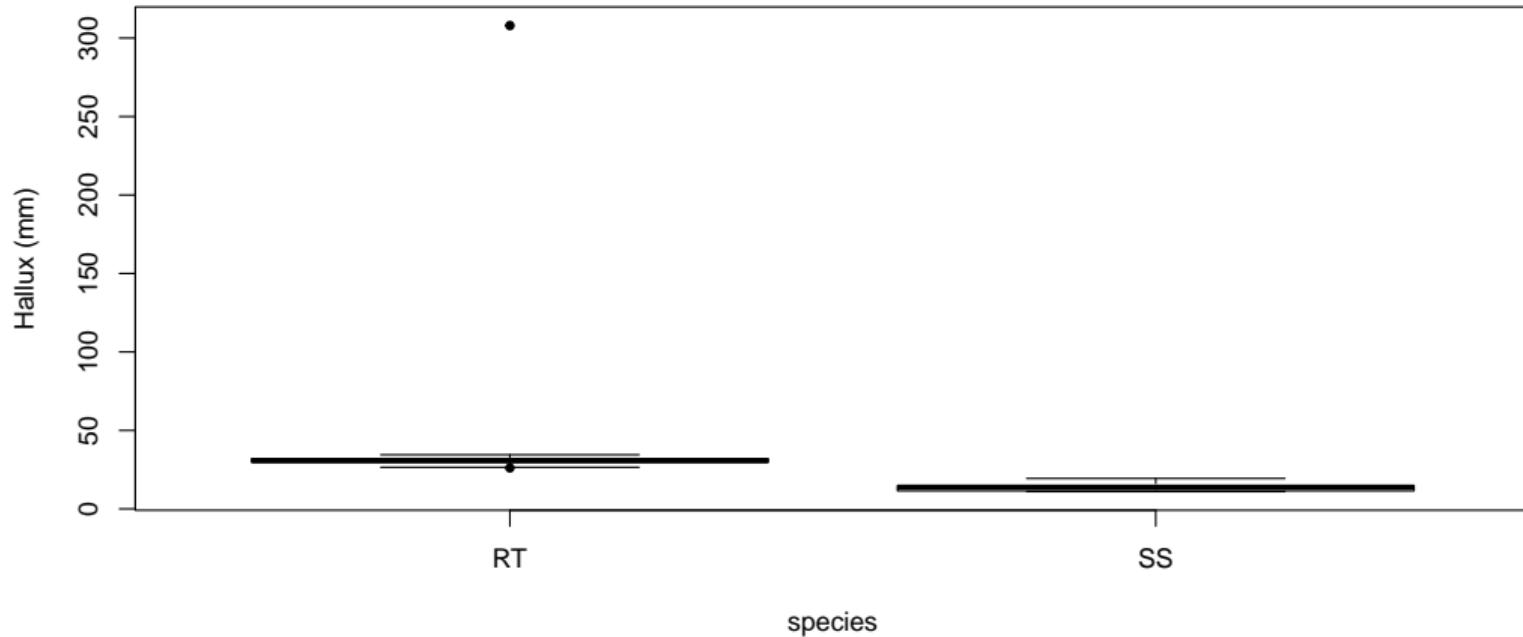
```
hawk = read.csv('hawk.csv')
```

- Look at the first few lines

```
head(hawk)

##   hallux species
## 1   32.9      RT
## 2   29.9      RT
## 3   11.0      SS
## 4   31.2      RT
## 5   33.0      RT
## 6   30.5      RT
```

Data: Hawks



What is the state of play?

- We have developed a variety of statistical models for data
 - ▶ Normal models
 - ▶ Binomial models
 - ▶ Central limit theorem
 - We can use difference in two means, ANOVA, linear regression, etc, even if data are non-normal
 - Require a large sample
- There may be situations where these methods may be inappropriate
 - ▶ We may be unwilling to assume the data is normal
 - ▶ We may be unwilling to rely on the CLT
 - e.g. outliers or skew
- Introduce non-parametric methods

Idea: look at ranks

- We rank the observations
 - ▶ From 1 to n , smallest to largest (or vice versa)
- Work with the ranks rather than the actual observations
- It can be useful with (extreme) outliers
 - ▶ Same rank irrespective of whether the largest observation is 0.1 units larger than 2nd biggest observation, or 10000 units larger
- It can be useful if there is a lot of skew
 - ▶ All ranks are equally far apart from each other

Example: ranking data

- Suppose that we had the following data

Group A	Group B
1.2	5.5
4.3	1.7
3.1	2.9

- The ranks are given alongside (in blue)

Group A	Group B
1.2 (1)	5.5 (6)
4.3 (5)	1.7 (2)
3.1 (4)	2.9 (3)

In R

- The R function `rank` will rank data

```
hawk$rank = rank(hawk$hallux)
```

- This code: ranks the hallux measurements
 - Inserts a new variable (`rank`) into the `hawk` data frame

```
head(hawk)

##      hallux species rank
## 1    32.9      RT   92
## 2    29.9      RT   55
## 3    11.0      SS    1
## 4    31.2      RT   75
## 5    33.0      RT   93
## 6    30.5      RT   66
```

What now?

- We can compare the ranks of the two groups
- Hypothesis test
 - ▶ H_0 : the distribution for the two groups are the same
 - ▶ H_A : the distribution for the two groups differ
- Sum up the ranks in the two groups
 - ▶ The specific form of the test statistic isn't important (for this course)
 - ▶ We can find a p -value
 - Tells us the probability of observing sum of ranks as extreme or more extreme than that observed if the distribution for the two groups are identical
- This is called the Mann-Whitney U test
 - ▶ It has many other names, such as the Mann-Whitney-Wilcoxon test

In R

- The test can be performed using the `wilcox.test` function in R
- Like when using `t.test` we separate data into two groups

```
rt = subset(hawk, species == "RT") # same function as we used for t.test
ss = subset(hawk, species == "SS")
wilcox.test(rt$hallux, ss$hallux)

##
## Wilcoxon rank sum test with continuity correction
##
## data: rt$hallux and ss$hallux
## W = 2275, p-value <2e-16
## alternative hypothesis: true location shift is not equal to 0
```

Interpretation

- As usual, the p -value is quantifying the incompatibility between the null hypothesis and the data
- Since the p -value $< \alpha = 0.05$, the data are unusual if the null hypothesis were true
 - ▶ The data we have observed would be unusual if there were the distribution of hallux length were the same for the two species

Parametric vs non-parametric

- Most of the models and methods we have seen so far are referred to as parametric
 - ▶ Specify the distribution of the observations: normal, binomial, etc
 - ▶ These models are defined in terms of parameters: μ , p , etc
 - ▶ We find confidence intervals for the parameters
 - ▶ We specify hypothesis tests about the parameters
- With non-parametric models, we make fewer assumptions
 - ▶ We assume the observations come from an unknown distribution
 - There are not specific parameters as above (hence non-parametric)
 - ▶ We can specify hypothesis tests
 - ▶ Confidence intervals are more challenging
- A common misconception is that non-parametric approaches make no assumptions

Non-parametric approaches

- The principle of converting data to ranks can also be used for other cases we have considered
- Single sample (or paired data) → Wilcoxon signed-rank test
- Two samples (independent groups) → Mann-Whitney test
- ANOVA (multiple independent groups) → Kruskal-Wallis test
- Remembering the names isn't important
- The concepts are more important: converting data to ranks
 - ▶ Note: not all non-parametric approaches use ranks
- We won't look at any details regarding the methods in blue above
 - ▶ It is worth knowing that the approaches exist

In R

- Seen `wilcox.test`
 - ▶ Used for single sample or paired data
 - ▶ Can be used for two independent groups
- The function `kruskal.test` can be used for:
 - ▶ Multiple independent groups
 - ▶ Can be used for two independent groups
- When using `kruskal.test` we need to use formula: as in `lm` or `aov`

```
kruskal.test(hallux ~ species, data = hawk)

##
##  Kruskal-Wallis rank sum test
##
## data: hallux by species
## Kruskal-Wallis chi-squared = 68, df = 1, p-value <2e-16
```

Comparing the two tests

- If we have two independent groups, we have a choice
 - ▶ Use `wilcox.test`
 - ▶ Use `kruskal.test`
- These have different test statistics
 - ▶ Give the same p -values
- Note: `wilcox.test` includes a continuity correction when calculating the p -value
 - ▶ The two approaches give the same p -value when this is turned off with `correct = FALSE`

Data: hawk tail measurements

- Look at data from 43 red-tailed hawks
- Data comparing two tail measurements
 - ▶ tail_std: Standard approach for measuring the tail length (mm)
 - ▶ tail: Approach invented by those involved in collecting data (mm)
- Import and view the data

```
hawk_tail = read.csv('hawk_tail.csv')
```

```
head(hawk_tail)  
##   tail tail_std  
## 1  222     229  
## 2  215     217  
## 3  235     236  
## 4  215     215  
## 5  212     221  
## 6  206     217
```

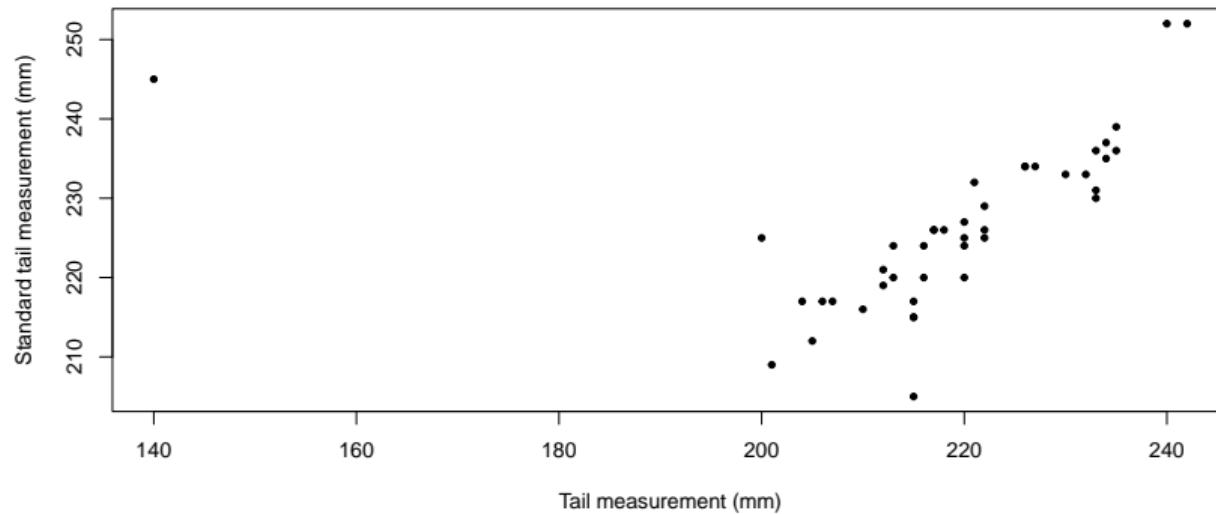
Hawk tail measurements: correlation

- How ‘consistent’ are the two measurement approaches?
 - ▶ We could assess with correlation

```
cor(hawk_tail$tail, hawk_tail$tail_std)  
## [1] 0.326
```

- That does not seem very high
- Look at the data to see what may be going on

Hawk tail measurements



- Seems like a reasonably strong linear relationship
 - ▶ With a large outlier

Back to ranks

- We can again work with ranks
 - ▶ Rank x (new tail measurements)
 - ▶ Rank y (standard tail measurements)
- Find the correlation of the ranks

```
hawk_tail$rank_tail = rank(hawk_tail$tail)
hawk_tail$rank_std = rank(hawk_tail$tail_std)
cor(hawk_tail$rank_tail, hawk_tail$rank_std)
## [1] 0.777
```

Correlation (sorry more names!)

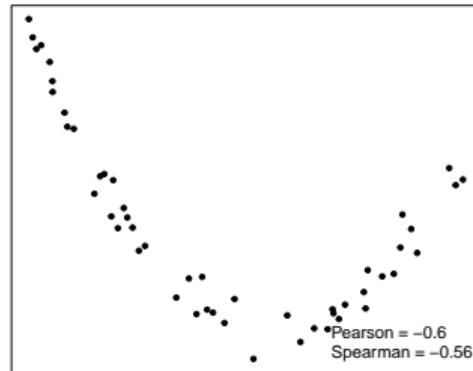
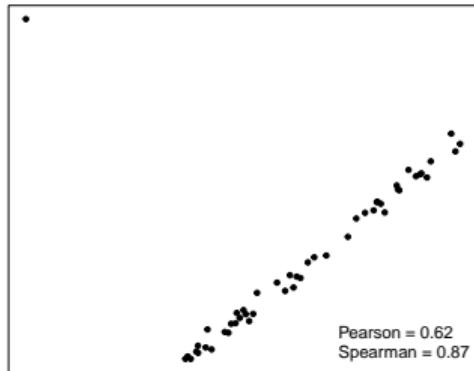
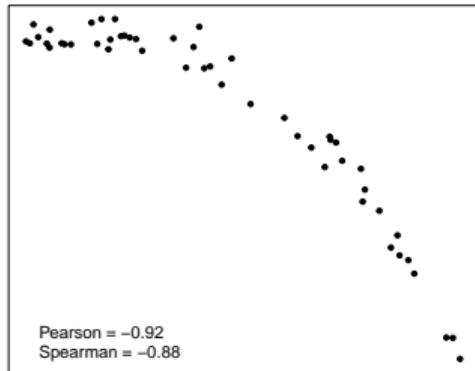
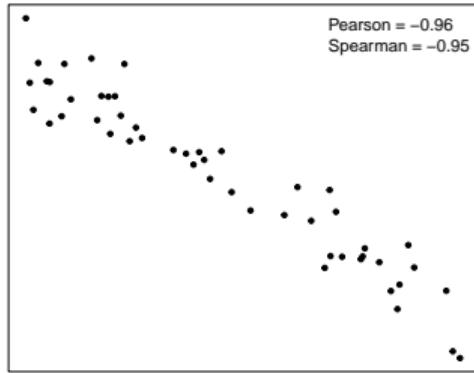
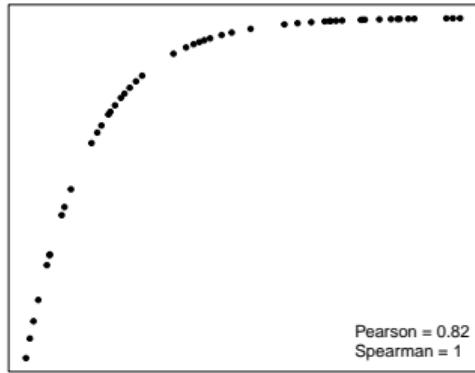
- The correlation based on ranks: Spearman correlation
- The correlation based on data: Pearson correlation
 - ▶ What we looked at when we covered linear regression
- Need not calculate the ranks in R to calculate Spearman correlation:

```
cor(hawk_tail$tail, hawk_tail$tail_std, method = "pearson") # this is the default  
## [1] 0.326  
  
cor(hawk_tail$tail, hawk_tail$tail_std, method = "spearman")  
## [1] 0.777
```

Spearman correlation

- Spearman correlation measures the strength of an increasing or decreasing relationship
 - ▶ It need not be a linear relationship
- Spearman correlation is robust to outliers: using ranks
 - ▶ Spearman correlation an alternative to throwing away outliers without justification
- Spearman correlation can be used with ordinal data
 - ▶ Categorical data where the values have an order
 - ▶ e.g. survey response: 'Excellent', 'Good', 'OK', 'Poor', 'Terrible'
- Spearman and Pearson correlation are often similar
 - ▶ Relationship is approximately linear
 - ▶ Minimal effect of outliers
- Look at some examples

Examples



Big picture

- Seen an introduction to non-parametric methods
- Focused on conceptual understanding
 - ▶ Assuming the data come from an unknown distribution
 - ▶ For the methods we have seen: working with ranks
 - ▶ Skipped over the details
- Advantages of parametric models
 - ▶ More powerful when assumptions hold
 - ▶ Interpret parameter (estimates)
 - ▶ Straightforward confidence intervals
- Advantages of non-parametric methods
 - ▶ Fewer assumptions
 - ▶ More robust to outliers and skewed data

Summary

- Looked at non-parameteric approaches
- Work with ranks
- Two independent groups
 - ▶ Mann-Whitney
- Correlation
 - ▶ Spearman correlation
- Outlined other approaches
 - ▶ Wilcoxon rank-sum (one sample / paired data)
 - ▶ Kruskal-Wallis (multiple independent groups)

STAT 110: Week 11

University of Otago

Outline

- Where does the data come from?
- Up until now, we have assumed the data are representative of the population
 - ▶ What problems can arise if it isn't?
 - ▶ How can we sample to ensure that it is?
- Look at generalizing from one population to another
 - ▶ Example of what can go wrong

Case study: presidential election

- The *Literary Digest* was a general interest weekly American magazine
- In 1936 they ran a presidential election poll
 - ▶ Polled 10 million people
 - ▶ Received 2.38 million responses
- 57.08% of the respondents preferred Candidate A
 - ▶ If we were to find a confidence interval for p
 - 95% CI: (0.5702, 0.5714)
- This was not the first election the *Literary Digest* had run a poll
 - ▶ They had correctly predicted the winner every election since first polling in 1916
- Seems fairly conclusive evidence that Candidate A would win
 - ▶ Do you agree?

Outcome

- Candidate A was Kansas Governor Alfred Langdon
- The other candidate was incumbent Franklin D. Roosevelt
- The literary digest poll was wrong by nearly 20 percentage points
 - ▶ FDR received 60.8% of the popular vote
- How can a sample so large (2.4 million people) be so wrong?

Big picture

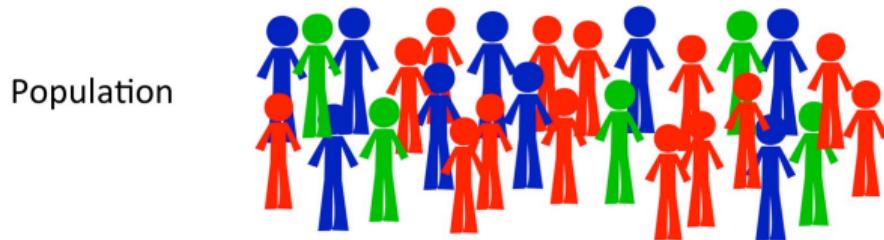
- Up until now, we haven't 'questioned the data'
 - ▶ Assumed it is reliable
 - ▶ Assumed it is representative of the population
- *Literary Digest*: the sample was not representative of the population
 - ▶ Even though the sample size is large
- Other (potentially) non-representative samples:
 - ▶ Measuring Otago Nuggets players (basketball) to learn about height in NZ population
 - ▶ Monitoring captive skinks to learn about movement in the wild
 - ▶ Using prisoners for a neuro study about brain response to stimuli for general population

Sampling

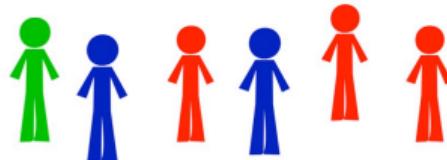
- Goal: to obtain a representative sample from the population
- Population: entire group of interest
- Sample: subset of the population
- A sampling frame is a list from which the sample is drawn
 - ▶ It ideally consists of the population

Simple random sampling

- Draw a sample of size n from the sampling frame such that each possible sample has the same probability of being selected.



Simple random
sample of 6
people



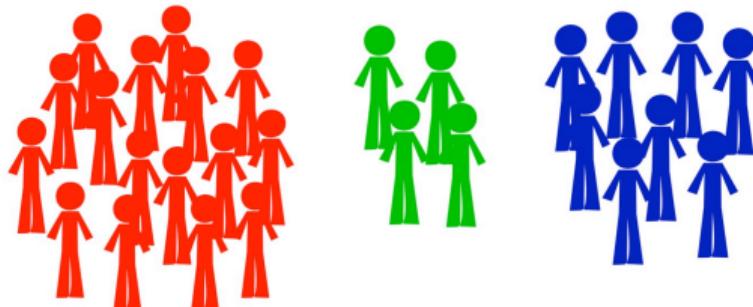
Every possible subgroup of 6 people has the same chance of selection
(you can think of it as every person having the same chance of selection)

Stratified sampling

- In many cases we may have certain groups within a population
 - ▶ e.g. ethnicity
- We can use stratified sampling in this situation
 - ▶ Define strata (or groups)
 - ▶ Take a simple random sample from within each stratum
- All strata are included in the sample

Stratified sampling: probability proportional to size

Population
stratified
(divided into
groups of
similar
people)



Stratified
random
sample

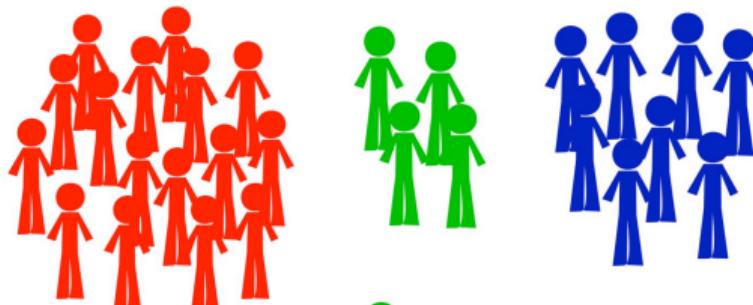


Sampled with probability proportional to size – everyone has the same chance of being selected

- This approach is useful if our interest is in the overall population
 - ▶ Ensuring each strata is represented
 - ▶ More precise than a simple random sample (of same size)

Stratified sampling: equal number from each strata

Population
stratified
(divided into
groups of
similar
people)



Stratified
random
sample



Sampled with equal numbers from each strata – those in smaller strata
are more likely to be selected

- This approach is useful if our interest is understanding the each stratum as well as the overall population
 - ▶ Ensuring accurate estimate within each strata (particularly small strata)

Cluster sampling

- Cluster sampling works using groups within the population
 - ▶ Single stage: take a simple random sample of groups and select all units in group
 - ▶ Two stage:
 1. Take a simple random sample of groups
 2. Take a simple random sample of all units in the group
- Cluster sampling is useful when clusters are easy to sample
 - ▶ e.g. sampling high school students by sampling schools
 - ▶ Usually cheaper but less precise than a simple random sample

One stage cluster sample

Population
divided into
clusters



Simple
random
sample of
clusters



Everyone in each sampled cluster is included in the study

Two stage cluster sample

Population
divided into
clusters



Simple
random
sample of
clusters



Simple random sample of one person from each cluster. Probability of someone being in the study depends on the number in their cluster

Sampling

- There are many, more complex, sampling designs
 - ▶ e.g. the sampling frame may not exist or may be otherwise difficult to find
 - Those in Dunedin addicted to drugs
 - Hedgehogs in the Dunedin Botanic Gardens
- How we analyze and interpret data may depend on how it has been sampled
- What can go wrong:
 1. Sampling error: the natural variation between statistic (sample) and parameter (population)
 - What we have been looking at: confidence intervals, etc
 2. Sampling bias: where systematic bias arises due to how the sample is collected

Sampling bias

- Selection bias: when the sample is not representative of the population
 - ▶ Sampling frame and population differ
 - ▶ *Literary Digest*: sampling frame telephone directories and car registrations
 - Predominantly wealthier people in 1936
 - Poorer voters were less likely to be sampled
- Non-response bias: those who don't participate in the study are systematically different from those who do
 - ▶ *Literary Digest*: empirical evidence was that working class were less likely to respond
 - Tended to favour FDR (Roosevelt)
 - 1936 was during the great depression
- Information bias: information obtained is not reliable
 - ▶ e.g. asking a participant about their diet in 1987
- There are many other possible sources of bias

Study design and interpretation

- We interpret the studies in light of the population of interest
- Care is required if we wish to generalize to other populations
 - ▶ Look at example

Case study: the 'warrior' gene

The screenshot shows the homepage of THE AGE website. At the top, there's a navigation bar with links for NEWS, MY CAREER, DOMAIN, DRIVE, and a member centre. Below the navigation is the THE AGE logo with its URL theage.com.au. A horizontal menu bar below the logo includes links for NEWS, ENTERTAINMENT, BUSINESS, SPORT, TRAVEL, TECH, SECTIONS, and CLASSIFIEDS. Below this menu, a breadcrumb trail shows the path: Home > National > Breaking News > Article. The main headline is 'Warrior gene' blamed for Maori violence' dated August 8, 2006, at 4:59PM.

'Warrior gene' blamed for Maori violence

August 8, 2006 - 4:59PM

- Research led by Dr. Rod Lea (ESR)
- General claim: a variant of the monoamine oxidase-A (MAO-A) gene is strongly associated with aggressive behaviour in Māori

Case study: the 'warrior' gene

- There were some controversial quotes in the media, e.g.
 - ▶ The MAO-A gene “*goes a long way to explaining some of the problems Māori have. Obviously, this means they are going to be more aggressive and violent and more likely to get involved in risk-taking behaviour like gambling.*” The Dominion Post, 9 August 2006¹
- Dr Lea subsequently claimed much of the controversy was unjustified because it stemmed from a combination of misquotes and misunderstandings ²
- Let's take a look at the science, study design and interpretation
 - ▶ Based on the work of Merriman and Cameron (2007; NZMJ; Vol 120–1250; 59–62)

¹ See Crampton and Parkin (2007; NZMJ; Vol 120–1250; 63–65)

² See Lea and Chambers (2007; NZMJ; Vol 120–1250; 5–10)

Case study: the 'warrior' gene

- There is evidence that low levels of MAO-A are associated with antisocial behaviour for males with a prior history of maltreatment
 - ▶ No evidence of association for those with no history of maltreatment
- This association may vary by ethnicity
 - ▶ The association was replicated in NZ (Dunedin study), UK, and USA (for white Americans)
 - ▶ No evidence of the association was found for US 'non-whites'
 - ▶ Consistent with evidence that many other genetic associations vary by ethnicity
- Summary: any association between antisocial behaviour and MAO-A does not appear to be strong, depends on the environment, and may vary by ethnicity

Case study: the 'warrior' gene

- To determine if the gene is associated with aggressive behaviour in Māori men, a study would need to be conducted that assess that
 - ▶ This is not what was done
- What was done: estimated the proportion of the specific MAO-A variant in the Māori male population
 - ▶ Based on a sample of $n = 46$ participants
 - ▶ Found it to be higher than some other ethnic groups (e.g. European ancestry)
- Very risky (at best) to use this and generalize / extrapolate the results of these previous studies

Case study: the 'warrior' gene

- So where does the term: 'warrior' gene, come from?
- It was termed by a scientific journalist
 - ▶ Based on a study of Rhesus macaque monkeys
 - ▶ Found that MAO-A gene was associated aggression
 - In a different way to that found in human studies
- Term 'warrior gene' had never been used for a human population before Dr Lea used it

Bias in AI

- Collecting more data does not solve the issues we have discussed today
 - ▶ *Literary Digest* is an example of that
- Application: Modern AI is backed by (often) huge datasets
 - ▶ If the data are biased (do not represent the population), the AI will be biased
 - Examples in [policing](#), [AI-assisted hiring](#), ...
- Bias in AI can also arise from generalizing from one population to another
 - ▶ Data used to train AI may come from a different population than where it is applied
 - Training data often preferential toward rich, white, male
- This is seen in healthcare, facial recognition, ...

The New York Times

[Artificial Intelligence](#) > [A.I. Forecast](#) [A.I.'s Super Bowl](#) [Google's Anthropic Investment](#) [What Is Vibecoding?](#) [Quiz](#)

*Facial Recognition Is Accurate, if
You're a White Guy*

Summary

- Where the data come from is critical
 - ▶ The fanciest statistical approach is useless if the data do not represent the population
- Look at some of the bias that can arise when sampling
- Looked at simple sampling designs
 - ▶ Simple random sample
 - ▶ Stratified sample
 - ▶ Cluster sample
- Looked at issues that can arise if we generalize to another population

STAT110 2025

Data and Interpretation II: Where do the data come from? (and why this matters...)

Phillip Wilcox

(Ngāti Rakaipaaka, Ngāti Kahungunu ki Wairoa, Rongomaiwahine, Pakeha)

Ahorangi Tuarua (= Associate Professor)

Department of Mathematics and Statistics

(Also: Affiliate Faculty, Bioethics Centre)



Current Roles...

- Associate Professor, Quantitative Genetics, Dept of Mathematics and Statistics
- Established Masters degree programme in Quantitative Genetics that started in 2017 -> now graduating students! See
<https://www.otago.ac.nz/sciences/study/applied-science/majors/otago619406.html>

Quantitative Genetics

[Home](#) / [Studying science](#) / [Applied Science programme](#) / [Majors](#) /

Quantitative Genetics uses statistical methods to understand the complexities of genetic inheritance, and can be applied in human medicine, population genetics and for selective breeding of plants and animals. So Quantitative Genetics is key in improving New Zealand's primary sector productivity and profitability as well as for understanding the genetic basis of a very wide range of diseases.

» Who is it for?
» What does it deliver?
» What is required?
» Contributors and Contacts



Current Roles...

- Kaikōkiri Māori in Genetics Teaching Programme – introducing te ao Māori content
- Kaiawhina Māori Math and Stat Dept + Genetics Teaching Programme
 - Oversee introduction of
- Current Research:
 - Genetic disease risk prediction in health-related characteristics in Māori and Pasifika
 - Co-lead *Aotearoa Variome* Project -> establish database of DNA sequence variation in contemporary Māori
 - Co-lead *Rakeiora* Pathfinder project -> pilot for personalised medicine involving Ngāti Porou Hauora
 - Māori-informed breeding objectives
 - Māori perceptions of gene editing



genomics
aotearoa

About Training Resources Our work Our stories News & events

Home / Our stories / Rakeiora: ground-breaking research guides the path to precision healthcare

Rakeiora: ground-breaking
research guides the path
to precision healthcare

Some of my Māori-related professional activities

- Involved in Māori-specific conversations/research regarding genetic technologies since 2001:
 - Established Te Arotūruki and co-developed TA Process (JRSNZ 2008)
 - Otago University ‘Full Circle Theme’
 - University of Waikato ‘Te Mata Ira’ & ‘Te Hau Mihi Ata’ projects
 - Various consultation with Māori communities
- Māori advisory roles
 - Scion (GE pine, kauri transcriptomics)
 - Royal Society of NZ Gene Editing expert panel
 - Reviewer for Te Tipu o te Wānanga
- Māori-specific education initiatives
 - Genetics modules in Science Wānanga for 11-14 year old Māori high school students
 - Summer Internship of Indigenous Peoples in Genomics (SING-Aotearoa)
 - Trained/ing 8 Māori graduate students in genetics
- Scientific Research
 - 24 years forestry-related genetics research
 - Genetics of gout in Maori and Pacific populations
 - GA-funded Aotearoa Variome project co-lead*
 - Co-led Māori components in successful NZ\$35M Genomics Aotearoa Proposal
 - Māori-specific roles in BioHeritage National Science Challenge and BioProtection CoRE
- Iwi (= tribal) roles
 - Previously mandated representative for Ngāti Rakaipaaka
 - Advisor for various iwi initiatives e.g. Ngāti Kohatu
 - Science advisor for various proposals in Wairoa/Mahia area
- University-specific activities
 - Maths and Stats Dept Kaiawhina
 - Sci Div Māori Leadership roopu
 - Overviewing & teaching UoO Science Wānanga Māori outreach
 - Teaching Māori ethical frameworks and consultation requirements in 300- and 400-level science courses



What we're covering...

Tuesday: Context and Rationale

- Why are we learning this?
- Indigenous peoples and knowledge... and some examples of how that impacts YOUR lives TODAY...
- Indigenous peoples and concepts of mathematics and data

Wednesday: Methods and tools

- Indigenous data sovereignty: key principles and tools
- Study design and conduct with Māori communities: tikanga-informed study design
- Co-design – an emerging area in conduct of research studies

Learning Outcomes...

By the end of these two lectures you should be able to describe:

- Examples of how indigenous peoples used data including use of mathematics
- What Māori and indigenous data sovereignty (IDS) are, why these exist, what are the underpinning principles, and what tools that can be used to implement IDS
- What is ethically appropriate (= tikanga informed) study design in an Māori context
- What ‘Co-design’ is (and why...)
- Why you need to know about all these things...

Why are we learning this?

- A broader understanding of indigenous knowledge and how this interfaces with application of modern statistical methods
- How data are collected impacts validity of study design and conduct – and utility of results...
- In Aotearoa/New Zealand, Māori knowledge, values and world views are increasingly important in all sector of our society...

Growing importance of the ‘Māori economy’

TE AO MĀORI / MONEY

New report highlights dramatic growth in Māori economy

5:39 pm on 11 March 2025

Share this

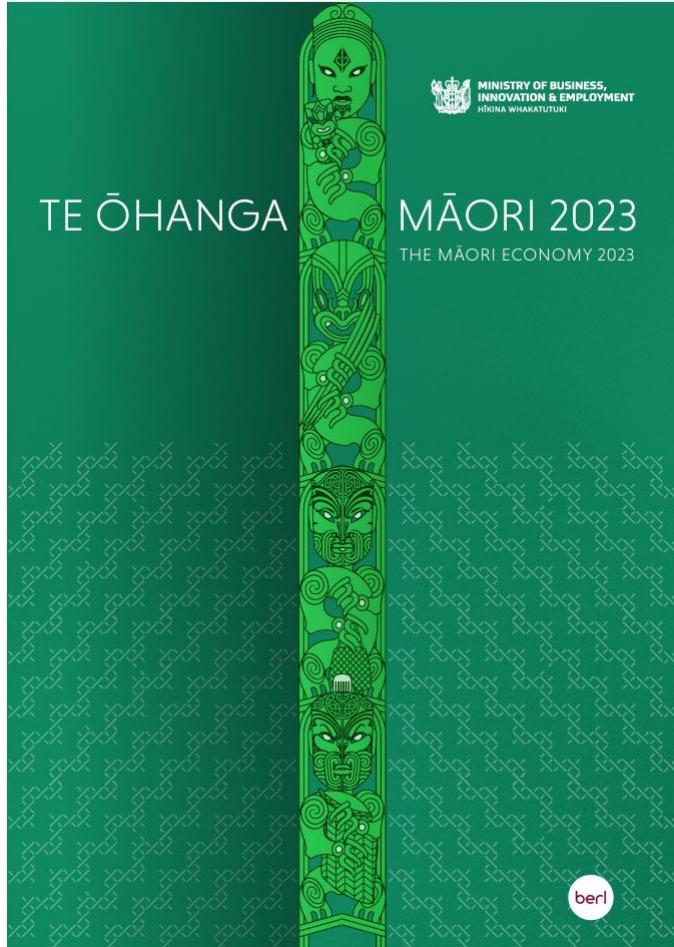


Emma Andrews, Henare te Ua Māori Journalism Intern
emma.andrews@rnz.co.nz



Māori entities have grown from contributing \$17 billion to New Zealand's GDP in 2018 to \$32 billion in 2023. Photo: 123RF

- Te ao Māori contribution to New Zealand society is increasing...
 - growing population (17% and increasing)
 - Cultural renaissance
 - Te Tiriti o Waitangi settlements
 - 23 000 Māori businesses*
 - Māori asset base



3 SNAPSHOT



Te Āhanga Māori is a strong, distinct, growing, and diversified component of the Aotearoa New Zealand economy.

MĀORI POPULATION

- The Māori population experienced substantial growth between 2018 and 2023, increasing by 14 percent from 775,800 to 887,500. This growth rate significantly outpaced the five percent increase observed in the non-Māori population during the same period.

MĀORI WORKFORCE

- The total number of Māori employed (including employers, employees, self-employed, and unpaid workers) totalled 390,700 in 2023, up 19 percent from 329,200 in 2018.
- More Māori were now employed in high-skilled jobs compared to those in low-skilled jobs.

MĀORI ASSET BASE

- In 2023, the asset base within Te Āhanga Māori was valued at \$1.26 billion, following an 83 percent increase from \$69 billion in 2018.
- While agriculture, forestry, and fishing remain significant the Māori asset base is diversifying, with real estate and property services experiencing substantial growth, increasing by 58 percent from \$16.7 billion in 2018 to \$26.3 billion in 2023.

MĀORI VALUE ADD (GDP)

- Value added (production GDP) from Te Āhanga Māori totalled \$32 billion in 2023, up from \$17 billion in 2018.
- The three largest sectors were professional, scientific, and technical services at \$5.1 billion; administrative, support, and other services at \$4.2 billion; and real estate and property services at \$4.1 billion.

MĀORI-OWNED BUSINESSES

- In 2023, there were nearly 24,000 Māori-owned businesses in Aotearoa New Zealand, an increase from 19,200 in 2018.
- The largest number of businesses, 5,934, was located in Tāmaki Makaurau followed closely by Waitaha with 4,215 Māori-owned businesses.



Māori contributions in all sectors of NZ society is growing...



WHAT IS CO-DESIGN IN A MĀORI SPACE?

KOTAHITANGA IN ACTION

Rita Estelle Wakefield
Kriki, British

Co-designing health research in Aotearoa New Zealand

// Lessons from the Healthier Lives National Science Challenge

Maori Data Experts Challenge Government Moves To Offshore New Zealand Data

Wednesday, 27 July 2022, 9:56 am
Press Release: Data Iwi Leaders Group

Data Iwi Leaders Group have challenged Government as it increasingly offshores New Zealand data, saying there are long-term benefits to investing in local data infrastructure instead.

Data Iwi Leaders Group Launches Revolutionary Iwi Data Platform - Te Whata

Thursday, 5 November 2020, 11:52 am
Press Release: Data Iwi Leaders Group

He whata kai, he whata idioro, īnhā he māramatanga

Tīhei Mauri Ora

myNZTE Export guides Courses and events Tools and resources Services

How to use and protect Māori intellectual property

Is whakapapa the answer to better health treatment?



Special Issue: Precision medicine

Rakeiora Genomics Platform: a pathfinder for genomic medicine research in Aotearoa New Zealand

Claire E Rye, Huti Puketapu-Watson, Helen Whongi, Ben Te Aika, Donia Macartney-Coxon, Joep de Ligt, ... show all

Received 14 Jun 2024; Accepted 13 Feb 2025; Published online: 24 Mar 2025

Cite this article | <https://doi.org/10.1080/03036758.2025.2469026>



Indigenous Branding- Creating an emotional connection.

Kellogg Rural Leadership Programme

Course 37 2018
Ashleigh Phillips



Māori in Governance of Agricultural Co-operatives in Aotearoa New Zealand.

Troy Hobson

Indigenous Branding – Creating a point of difference to the New Zealand Primary Sector



Kellogg Rural Leadership 2015
Stephen Thomson



Cultural Safety in Vocational Medical Training

Tipene-Leach D., Haggie-H., Potiki-M., Carter-M.



Health Equity

Experiences perspectives and values of indigenous peoples regarding kidney transplantation: Systematic review and thematic synthesis of qualitative studies

Walker R., Tipene-Leach D., Abel S., Reynolds A., Palmer S., Walker C.



Health Equity

Understanding the experiences perspectives and values of indigenous women around smoking cessation in pregnancy: Systematic review and thematic synthesis of qualitative studies

Walker R., Graham A., Palmer S., Tipene-Leach D., Jagroop A.



The Kaitiaki Intelligence Platform: conceptual foundations for an indigenous environmental sensing network

John Keoh, Katherine Rout, Dennis Wairere-Schulman, Corey Ruha & Jan Hanra

Received 30 Sep 2023; Accepted 17 Feb 2025; Published online: 25 Mar 2025

Cite this article | <https://doi.org/10.1080/03036758.2025.2470423>

Full Article | Figures & data | References | Citations | Metrics | Reprints & Permissions | View PDF | View EPUB



Sustainable Seas

How to Indigenise the Blue Economy in Aotearoa New Zealand

Rout M., Mikia J., Reid J., Whitehead G., Gillies A., Wiremu F., McLellan G., MacDonald T., Ruha C.



Sustainable Seas

Indigenising the Blue Economy: A Case Study of the Moriori of Rēkohu

Gomes D., Gillies A.



Health Equity

Mahi a Atua: A Pathway Forward for Māori Mental Health?

Tipene-Leach D., Rangihuna D., Kopua M.



Cultural Safety

Cultural Safety Training Plan for Vocational Medicine in Aotearoa

Simmonds S., Carter M., Haggie-H., Mills V., Lyndon M., Tipene-Leach D.



Nourishing Hawke's Bay

Assessing the Potential for School Based Programmes Ka Ora Ka Ako to Enhance Education Sustainability, and Health Goals

McKee-Sellizar P., Swinburn B., Rees D., Glassay R., Tipene-Leach D., Garton K.

Also... UoO graduate Attributes

- <https://www.otago.ac.nz/courses/the-university-of-otago-graduate-profile>



Cultural understanding

Knowledge and appreciation of biculturalism within the framework of the Treaty of Waitangi; knowledge and appreciation of multiculturalism; and an ability to apply such knowledge in a culturally appropriate manner.

Ethics

Knowledge of ethics and ethical standards and an ability to apply these with a sense of responsibility within the workplace and community

Todays lecture

- Some context:
 - Indigenous peoples
 - Indigenous knowledge and (some) contributions to our lives
 - Indigenous concepts of mathematics and data

Who are indigenous peoples? What differentiates indigenous from non-indigenous?

Source:

https://www.un.org/esa/socdev/unpfii/documents/5session_factsheet1.pdf



FACTSHEET

Who are indigenous peoples?

It is estimated that there are more than 370 million indigenous people spread across 70 countries worldwide. Practicing unique traditions, they retain social, cultural, economic and political characteristics that are distinct from those of the dominant societies in which they live. Spread across the world from the Arctic to the South Pacific, they are the descendants - according to a common definition - of those who inhabited a country or a geographical region at the time when people of different cultures or ethnic origins arrived. The new arrivals later became dominant through conquest, occupation, settlement or other means.

Among the indigenous peoples are those of the Americas (for example, the Lakota in the USA, the Mayas in Guatemala or the Aymaras in Bolivia), the Inuit and Aleutians of the circumpolar region, the Saami of northern Europe, the Aborigines and Torres Strait Islanders of Australia and the Maori of New Zealand. These and most other indigenous peoples have retained distinct characteristics which are clearly different from those of other segments of the national populations.

Understanding the term “indigenous”

Considering the diversity of indigenous peoples, an official definition of “indigenous” has not been adopted by any UN-system body. Instead the system has developed a modern understanding of this term based on the following:

- Self- identification as indigenous peoples at the individual level and accepted by the community as their member.
- Historical continuity with pre-colonial and/or pre-settler societies
- Strong link to territories and surrounding natural resources
- Distinct social, economic or political systems
- Distinct language, culture and beliefs
- Form non-dominant groups of society
- Resolve to maintain and reproduce their ancestral environments and systems as distinctive peoples and communities.

Indigenous People have always been 'scientists'



Polynesian
navigation



Plant and medicinal knowledge



Diné understanding of genetics

Indigenous people used traditional knowledge from generations of experiential learning, repeated observations and **experimentation** to understand the world around them.

Indigenous knowledge exist pre-dates universities.

Examples of indigenous knowledge applications that affect our everyday lives...

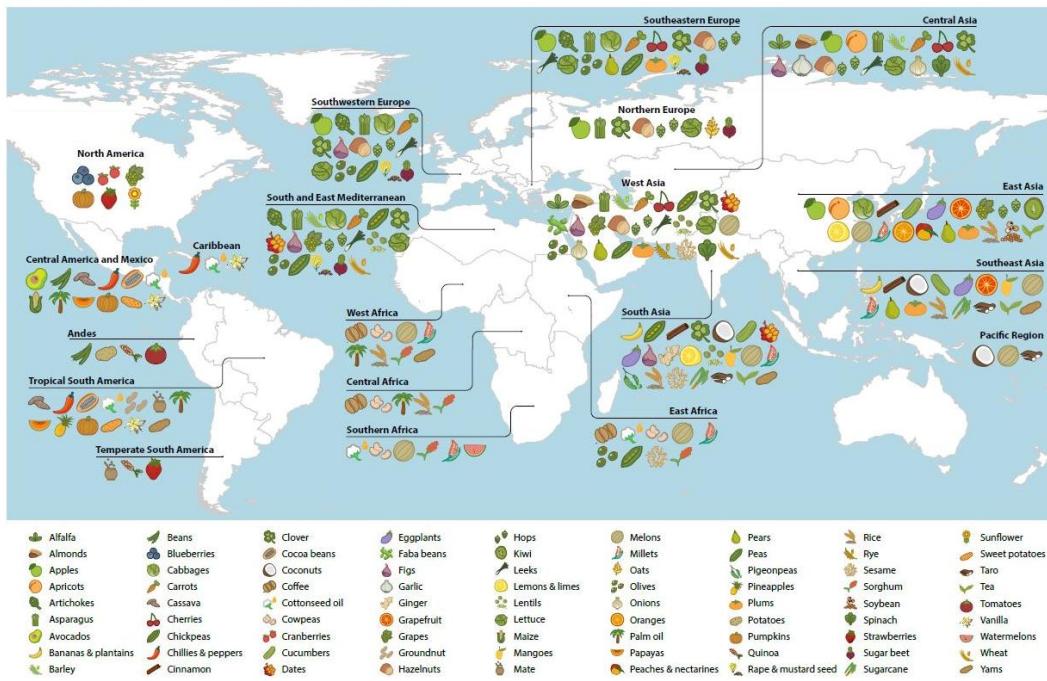
ORIGINS AND PRIMARY REGIONS OF DIVERSITY OF AGRICULTURAL CROPS

Khoury CK, Achianoy HA, Bjorkman AD, Navarro-Racines C, Guarino L, Flores-Palacios X, Engels JMM, Wiersma JH, Dempewolf H, Sotelo S, Ramirez-Villegas J, Castañeda-Alvarez NP, Fowler C, Jarvis A, Rieseberg LH, and Struik PC (2016).

Origins of food crops connect countries worldwide. Proc. R. Soc. B 283: 20160792. DOI: 10.1098/rspb.2016.0792.



International Center for Tropical Agriculture
Since 1967: Science to cultivate change



<https://blog.ciat.cgiar.org/origin-of-crops/>

- Domestication of plant species: **almost every food crop was domesticated by indigenous or ancient peoples**
- **Modern-day breeding programmes use these domesticated varieties** (and modern statistical methods like ANOVA and regression to quantify and rank genotypes...)

Kumara Domestication led to Māori Cultural Evolution...

- Adaptation of kumara – a tropical species – to temperate climates...



Phase 1: Introductory

Introduced species grown using tropical-based production systems



Phase 2: Experimental

Development of agricultural methods like:

- Cold-protection for growing plants
- Long term storage methods
- Cultivation systems for indigenous species



Phase 3: Systemic/Stable

Production systems mature

Well-developed and systemized ceremonial rituals involving these species (*tapu*, *māramataka*)

Major contributions to pre-European Māori economy, especially kumara & harakeke



Evolution of ‘Classic’ Māori Culture

‘Archaic Eastern Polynesian’ (1150 → 1350-1500 AD)

- Aka ‘Moa-hunter’
- Nomadic, hunter-gatherer – based society
- Few permanent year-round settlements
- Widely travelled within Aotearoa
- Less war-like (?) – some evidence of peace traditions
- Some elements still exist today (e.g., southern parts of Te Waipounamu)



Classic Māori (1350 1500 → 1800 AD)

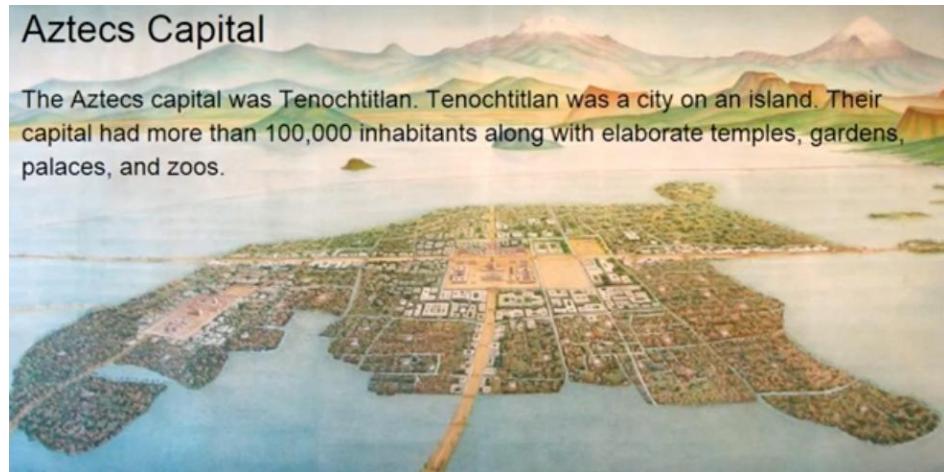
- Place-based permanent settlements
- **Agriculturalists** = role specialisation
- New forms of cultural expression: whakairo, religious developments (e.g., tuahu, local deities such as taniwha, etc.)
- New adaptations: pā (fortified villages)
- Sophisticated tribal structures and role specialisation: rangatira, tohunga, kaititaki, etc.
- Kaitiakitanga (environmental stewardship for the benefit of the tribe)



Indigenous mathematics... some examples

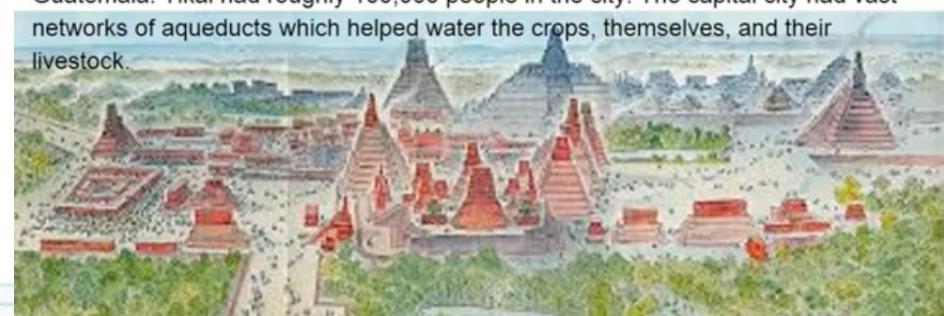
Mayan and Aztec

- Base 20 counting system = enabled mathematical operations to performed with large numbers
- Mathematical & Astronomical Knowledge:
 - The Maya and Aztecs developed complex mathematical and astronomical knowledge, crucial for their calendar systems, architectural planning, and resource management.
- Calendar Systems:
 - The Maya and Aztecs developed sophisticated calendar systems based on astronomical observations, which were used for religious rituals, agricultural planning, and timekeeping.
- Impact on Urbanization:
 - Urban Planning: Their mathematical and astronomical knowledge facilitated the precise planning and construction of their cities, including the alignment of buildings and the creation of infrastructure like aqueducts and roads.
 - Resource Management: They used their knowledge of mathematics and astronomy to manage resources, including land allocation, taxation, and trade.
 - Social Organization: Their mathematical and astronomical systems were also intertwined with their social and religious beliefs, reinforcing the power of the ruling



Mayans Capital

The Mayans capital city was called Tikal, which is located in modern day Guatemala. Tikal had roughly 100,000 people in the city. The capital city had vast networks of aqueducts which helped water the crops, themselves, and their livestock.



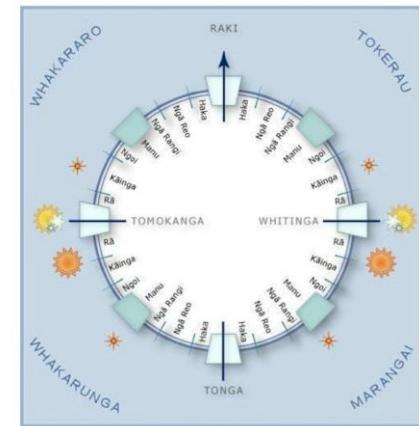
Indigenous mathematics... some examples

Māori

Much knowledge of Māori mathematics is lost, however examples include...

- The commonly used Māori term ‘Pangarau’ means ‘many relations’, reflecting an algorithmic understanding
 - Aspects of mathematics were integrated into various activities, survival practices, spiritual practices, games, and art.
 - Forms of quantification, measurement, utilisation of patterns, means of orientation, and the ability to form and recognise relationships and interactions between people, things, and place.
 - Numbering system
 - Concepts of nothingness, or zero: te kore; concepts of infinity

- Perhaps most well-known feat is navigating using stars



Indigenous Peoples and Data....

- ‘Indigenous peoples have always been data collectors and knowledge holders’

About Us

WHO WE ARE

PURPOSE

HISTORY OF
INDIGENOUS DATA
SOVEREIGNTY

HISTORY OF INDIGENOUS DATA SOVEREIGNTY

While the term Indigenous Data Sovereignty is relatively new, **Indigenous Peoples have always been data collectors and knowledge holders.** The rise of national Indigenous Data Sovereignty networks reflects a growing global concern about the need to protect against the misuse of Indigenous data and to ensure Indigenous Peoples are the primary beneficiaries of their data. GIDA connects these national communities to advocate for shared rights and interests in data.

<http://gida-global.org/history-of-indigenous-data-sovereignty>

WHAT ARE INDIGENOUS DATA?

Data, information and knowledges, in any format, that impacts Indigenous Peoples, nations, and communities at the collective and individual levels:

DATA ABOUT OUR NON-HUMAN RELATIONS

Land, water,
geology, titles, air,
soil, sacred
ecosystems,
territories, plants,
animals, etc.

DATA ABOUT US AS INDIVIDUALS

Administrative, legal,
health, social,
commercial,
corporate, services,
etc.

DATA ABOUT US AS COLLECTIVES

Traditional and
cultural information,
languages
knowledge systems,
ancestral and clan
knowledges, etc.

USINDIGENOUSDATA.ORG
@USIDSN

Informed by British Columbia First Nations Data Governance Institute - BCFNDGI.COM

GIDA-GLOBAL.ORG
@GidaGlobal

This work is licensed under a [Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License](#)

What are traditional Māori concepts of data ?

- Data = knowledge/information...
- Cosmological mythologies:
 1. Tawhaki the demi-god (= a person of human and divine lineage) ascended to the heavens to retrieve four baskets of knowledge:
 - Ritual, prayer, and memory.
 - Knowledge of the natural world and patterns of energy.
 - Knowledge of both terrestrial and celestial forms of knowledge.
 - Incantations, literature, philosophy, and ritual practices.
 2. Tane Mahuta – a god – passed down to humans three baskets of knowledge:
 - Te Kete Aronui:
 - This basket holds the knowledge that benefits humankind, including knowledge of the natural world, and the physical, spiritual, and mental well-being of people.
 - Te Kete Tuauri:
 - This basket contains knowledge of ritual, prayer, memory, and spiritual communication, representing the realm of the spirit and the sacred.
 - Te Kete Tuatea:
 - This basket holds knowledge of evil or mākutu, which is harmful to humanity, representing the darker aspects of life.
- Knowledge is from the gods... therefore tapu (= restricted)

What are contemporary Māori concepts of data ?

- Language (te reo Māori words): data = raraunga, tātauranga
 - raraunga motukore (continuous data), raraunga motumotu (discrete data), and raraunga houanga (time series data), roopu matatini – multivariate categorical data
- Perspectives:
 - ‘Māori data. Māori data refers to information produced by or about Māori, and about the environments we have relationships with.
 - Māori view data as a living taonga (treasure) with immense strategic value. It’s an important tool in understanding our whenua and our tangata whenua. It helps us answer questions like how many people whakapapa to our iwi and where they live, how many Māori live within Aotearoa, and how many live or were born overseas’
- Ethics:
 - ‘Māori data is subject to the rights articulated in the Treaty of Waitangi and the UN’s Declaration on the Rights of Indigenous Peoples, to which Aotearoa New Zealand is a signatory.’

COMMENT & ANALYSIS
Māori data is a taonga
by Ngäpera Riley | May 28, 2023 | 8 min read

Data collected about Māori people and resources is a valuable asset – it can be a powerful mechanism for informing and driving significant change in communities.

But that will only happen if Māori are able to exercise authority over data and treat it as a taonga, as Ngäpera Riley tells us here.



Source: <https://e-tangata.co.nz/comment-and-analysis/maori-data-is-a-taonga/#:~:text=M%C4%81ori%20view%20data%20as%20a,live%20or%20were%20born%20overseas.>

Data are taonga...

- ‘He taonga he tapu... na te tapu i puta mai te tikanga’
- Māori data are treasured, i.e., of emotional value, thus are restricted and should be treated appropriately
- But what does ‘appropriate’ consist of?

HE TAONGA HE TAPU

Tissue is a taonga [precious]
Tissue, DNA and Data are taonga, separately and together
Data refers to both genomic and clinical information

He Taonga, He Tapu
- Protection of taonga
- Na te tapu i puta mō te tikanga
(Physical and Spiritual components)



STAT110 2025 – Lecture B

Data and Interpretation II: Where do the data come from? (and why this matters...)

Phillip Wilcox

(Ngāti Rakaipaaka, Ngāti Kahungunu ki Wairoa, Rongomaiwahine, Pakeha)

Ahorangi Tuarua (= Associate Professor)

Department of Mathematics and Statistics

(Also: Affiliate Faculty, Bioethics Centre)



What we covered last lecture

- Why we teach this material in STAT110
- Indigenous concepts of data, mathematics and their use including in te ao Māori
- Examples of indigenous knowledge applications that impact our lives
- Māori concepts of data – both traditional and contemporary – and why data are a taonga

What we're covering today...

Tuesday: Context and Rationale

- Why are we learning this?
- Indigenous peoples and knowledge... and some examples of how that impacts YOUR lives TODAY...
- Indigenous peoples and concepts of mathematics and data

Wednesday: Methods and tools

- Indigenous data sovereignty (IDS): key principles and tools
- Study design and conduct with Māori communities: tikanga-informed study design
- Co-design – an emerging area in conduct of research studies

WHAT ARE INDIGENOUS DATA?

Data, information and knowledges, in any format, that impacts Indigenous Peoples, nations, and communities at the collective and individual levels:

DATA ABOUT OUR NON-HUMAN RELATIONS

Land, water, geology, titles, air, soil, sacred ecosystems, territories, plants, animals, etc.

DATA ABOUT US AS INDIVIDUALS

Administrative, legal, health, social, commercial, corporate, services, etc.

DATA ABOUT US AS COLLECTIVES

Traditional and cultural information, languages knowledge systems, ancestral and clan knowledges, etc.

USINDIGENOUSDATA.ORG
@USIDSN

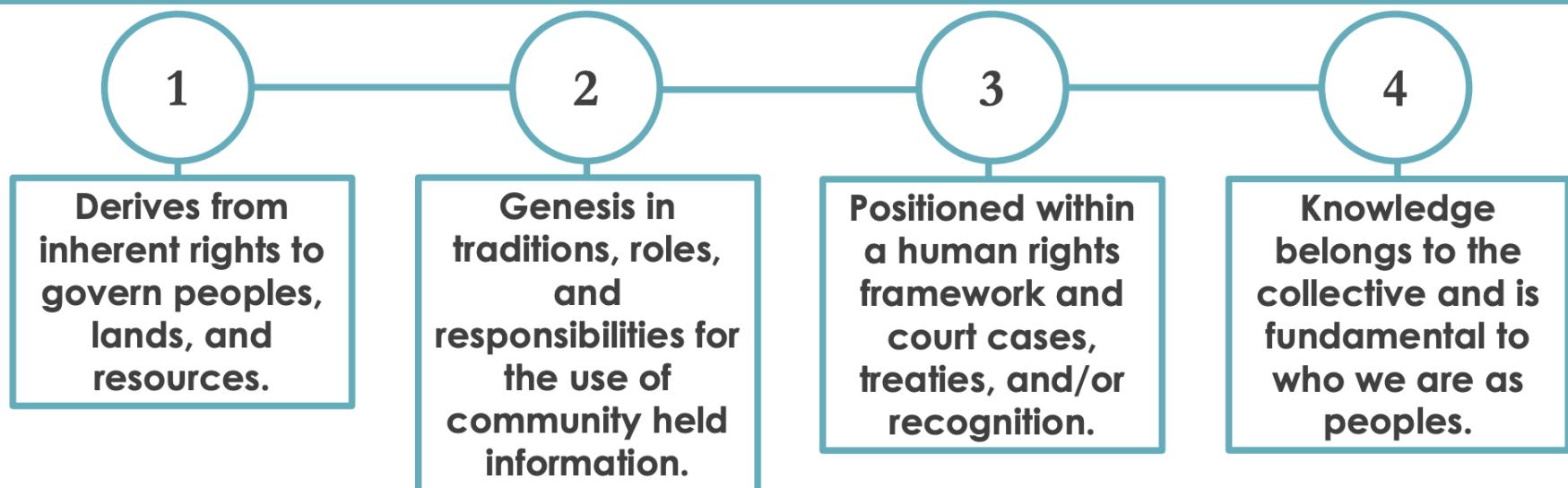
Informed by British Columbia First Nations Data Governance Institute - BCFNDGI.COM

This work is licensed under a [Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License](https://creativecommons.org/licenses/by-nc-nd/4.0/)

GIDA-GLOBAL.ORG
@GidaGlobal

INDIGENOUS DATA SOVEREIGNTY

The *right* of Indigenous Peoples and nations to govern the collection, ownership, and application of their own data.



See Kukutai T & Taylor J. (Eds). (2016). Indigenous Data Sovereignty. Canberra: Australian National University Press.

USINDIGENOUSDATA.ORG | @USIDSN

This work is licensed under a [Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License](https://creativecommons.org/licenses/by-nc-nd/4.0/).

GIDA-GLOBAL.ORG | @GIDAGlobal

The CARE Principles*

C = Collective Benefit: ‘Data ecosystems shall be designed and function in ways that enable Indigenous Peoples to derive benefit from the data’

A = Authority to control: ‘Indigenous Peoples rights and interests in indigenous data must be recognised and their authority to control such data be empowered. Indigenous data governance enables indigenous peoples and governing bodies to determine how Indigenous Peoples, as well as indigenous lands, territories, resources knowledges and geographical indicators, are indicated and identified within data.’

R = Responsibility: ‘Those working with Indigenous data have a responsibility to share how those data are used to support Indigenous Peoples’ self determination and collective benefit. Accountability requires meaningful and openly available evidence of these efforts and the benefits accruing to Indigenous Peoples.’

E = Ethics: ‘Indigenous Peoples’ rights and wellbeing should be the primary concern at all stages of the data life cycle and across the data ecosystem.’



*Source: Global Indigenous Data Alliance
<https://www.gida-global.org/care>

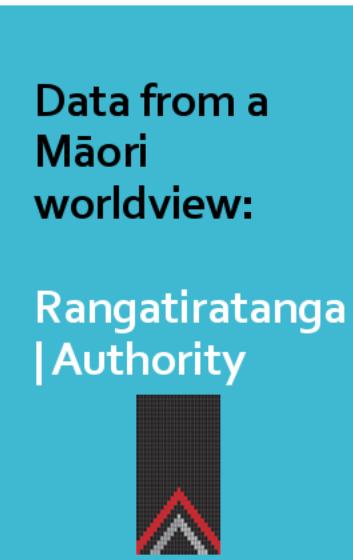
Māori Data and Data Sovereignty

- Māori data (from yesterday...) = ‘digital or digitisable information or knowledge that is about or from Māori people, our language, culture, resources or environments’
- Māori Data Sovereignty (MDS) MDS refers to the inherent rights and interests that Māori have in relation to the collection, ownership, and application of Māori data*

Principles of
Māori Data
Sovereignty

Source: https://www.otago.ac.nz/__data/assets/pdf_file/0014/321044/tmr-maori-data-sovereignty-principles-october-2018-832194.pdf

Māori Data Sovereignty in Action...



- **Control.** Māori have an inherent right to exercise control over Māori data and Māori data ecosystems. This includes but is not limited to data creation, development, stewardship, analysis, dissemination and infrastructure.
- **Jurisdiction.** Decisions about the physical and virtual storage of Māori data should enhance control for current and future generations. Whenever possible, Māori data should be stored in Aotearoa NZ
- **Self-determination.** Māori have the right to data that is relevant and empowers sustainable self-determination and effective self-governance.

Principles of
Māori Data
Sovereignty

Source: https://www.otago.ac.nz/__data/assets/pdf_file/0014/321044/tmr-maori-data-sovereignty-principles-october-2018-832194.pdf

Māori Data Sovereignty in Action...

Data from a
Māori
worldview:

Kaitiakitanga |
Guardianship

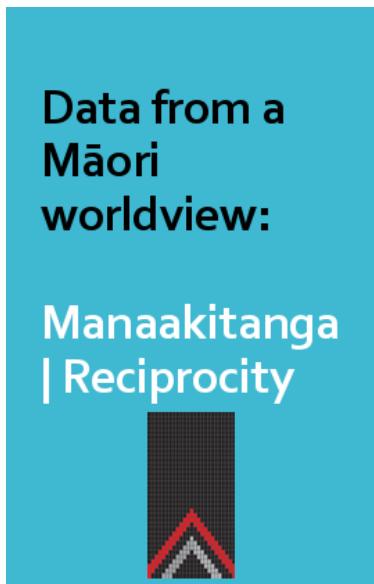


- *Stewardship.* Maori data needs to be stored and transferred in such a way that it enables and reinforces the capacity of Māori to exercise kaitiakitanga over Māori data .
- *Restrictions.* Māori should decide which Māori data sets should be controlled (tapu) or open (noa) access.
- *Ethics.* Tikanga, kawa (protocols) and mātauranga Māori (knowledge) should underpin the protection, access and use of Māori data.

Principles of
Māori Data
Sovereignty

Source: https://www.otago.ac.nz/__data/assets/pdf_file/0014/321044/tmr-maori-data-sovereignty-principles-october-2018-832194.pdf

Māori Data Sovereignty in Action...



- *Respect*. The collection, use and interpretation of data should uphold the intrinsic dignity of Māori individual, groups and communities.
- *Consent*. Free, prior and informed consent should underpin the collection and use of all data from or about Māori. Less defined types of consent must be balanced by stronger governance arrangements.

Principles of
Māori Data
Sovereignty

Source: https://www.otago.ac.nz/__data/assets/pdf_file/0014/321044/tmr-maori-data-sovereignty-principles-october-2018-832194.pdf

DIGITAL TOOLS TO PROTECT INDIGENOUS DATA SOVEREIGNTY

- ✓ **Traditional Knowledge (TK) Labels.** Digital markers that define attribution, access, and use rights for Indigenous cultural heritage
- ✓ **Biocultural (BC) Labels.** Digital markers for provenance, transparency and integrity in research engagements related to community expectations and consent for use of collections and data.
- ✓ **Dynamic Consent Portal.** An Indigenous-led data repository to house Tribally-consented genomic sequence data and manage access and attribution.
- ✓ **Blockchain.** A distributed ledger system that tracks sharing via transactions, can fine-tune user access, attribute provenance, and facilitate data governance.
- ✓ **Federated learning.** To facilitate secure and community-consented data sharing.



Native BioData
consortium

Tools for Protecting Indigenous Data

- **Biocultural (BC) labels and traditional knowledge indicators**
 - metadata tags on data sets that provide provenance and attribution
 - ‘BC Labels define community expectations about appropriate use of biocultural collections and data. The BC Labels focus on accurate provenance, transparency and integrity in research engagements with Indigenous communities.’ source: <https://localcontexts.org/>

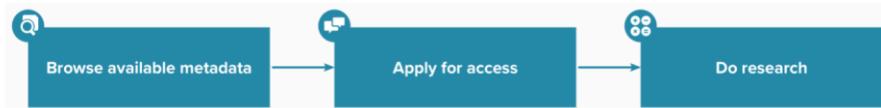


Tools for Protecting Indigenous Data

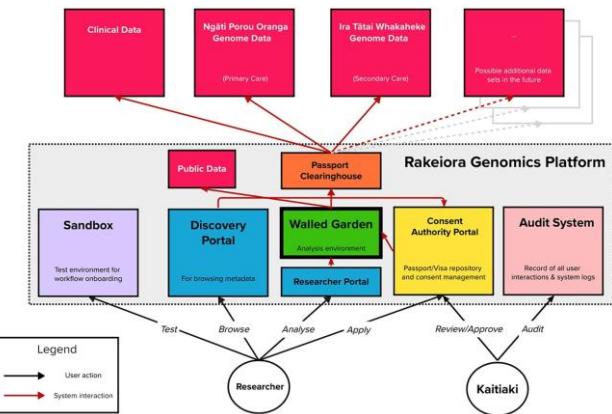
Dynamic Consent portals/Platforms

- Embed indigenous governance and oversight in access to, and analyses of, specific data sets within custom-built computational platforms/online environments

Researcher:



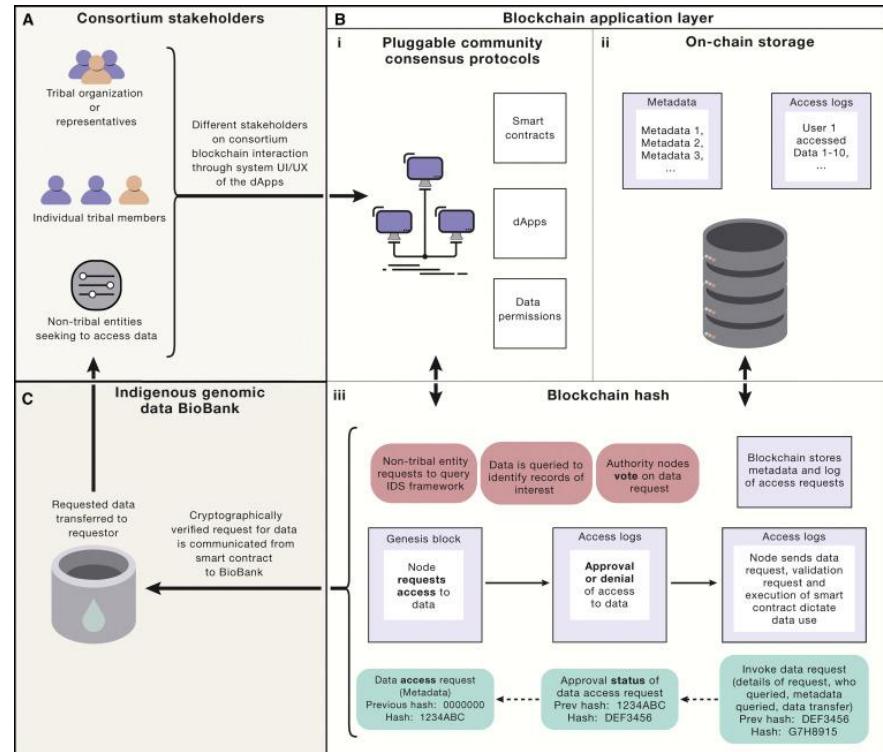
Community representatives:



Tools for Protecting Indigenous Data

Blockchain (also used for cryptocurrencies):

- ‘Distributed ledger that records transactions and is shared and agreed upon by all parties as the sole record of transactions... creating a “blockchain” of timestamped data establishing the agreement, provenance, and finality of the history of a transaction or management of data between a network of users’



Cell

Volume 185, Issue 11, 21 July 2022, Pages 2626-2632

Commentary

Establishing a blockchain-enabled Indigenous data sovereignty framework for genomic data

Tim K. Mocky^{1,2,3*}, Alec J. Colar^{3,4,5}, B. S. Chennu Keshava³, Joseph Ynacheta², Krystal S. Hostie², Keolu Fox^{1,4,6,7,8}



Implications for Study Designs when working with Māori...

- Where the data come from and how they are handled is VERY important...
- Māori have their own knowledge base (typically not well known outside of te ao Māori ‘institutions’)
- Māori have their own concepts of ‘data’ both traditionally and in a contemporary sense
- Data are treasured items (= taonga)
- There are restrictions regarding data – how data are generated stored and utilized
- There are tools and frameworks that now enable this

Study Design – in a Maori Context

Things to consider when working in indigenous subject areas*...

- Ensuring Māori participation in research – including ethics (= tikanga)
- Te Tiriti o Waitangi: data are taonga, and under Article 2 of Te Tiriti, all taonga are to be under Māori control
- 1993 Human Rights Act:
 - Individual's right to freedom from discrimination
 - Right to complain if rights are breached

Things to consider when working in indigenous subject areas*...

- Ensuring Māori participation in research – including ethics (= tikanga)
- Te Tiriti o Waitangi: data are taonga, and under Article 2 of Te Tiriti, all taonga are to be under Māori control
- 1993 Human Rights Act:
 - Individual's right to freedom from discrimination
 - Right to complain if rights are breached

What is ‘tikanga Māori’?

Guiding behaviour and relationships

Tikanga are more than just ‘rules’. They are best described as a form of social control and can guide the way relationships are formed, provide ways for groups to interact, and even guide the way people identify themselves.

Tikanga inform frameworks that address ethical issues. They guide good behaviour and practice when engaging with Māori and the things that matter to them.

Practical applications of tikanga

In Aotearoa New Zealand, tikanga are already present in many domains and have become widely known and accepted for some time. They exist in many corners of our society, are heard on television and radio, and are seen in almost all daily interactions, from social media and classrooms to the sports fields.

References to tikanga and their definitions appear in some of our legislation, education policies, government services, court processes, and political systems. Tikanga principles reflected in areas beyond the marae context is not new, neither is it unheard of in Aotearoa New Zealand.

“Tikanga Māori accompanies Māori wherever they go and whatever they do. Tikanga Māori is adaptable, flexible, transferable, and capable of being applied to entirely new situations.”

<https://data.govt.nz/toolkit/data-ethics/nga-tikanga-paihere/data-and-tikanga>

- Tikanga is underpinned by **values** that provide the cultural logic for **study design and conduct** with Māori communities

Tikanga Māori

- ‘Cultural concepts are conceptual markers, derived from mātauranga Māori (indigenous knowledge) and tikanga Māori (Māori values), which are intrinsic to an indigenous way of viewing and living in the world. These cultural cues provide the basis for describing the cultural logic that underpins engagement in a culturally acceptable manner.’

- Hudson et al 2019

Commonly Used Māori Values in Ethical Frameworks:

- *whakapapa* (genealogy)
- *mauri* (life essence)
- *mana* (power/authority)
- *kaitiakitanga* (guardianship)
- *mātauranga* (indigenous knowledge)
- *tapu* (sacred/restricted) and *noa* (not sacred or restricted)
- *pono* (honest, transparent)
- *mātau* (expertise)
- *wairua* (emotional or spiritual sides)

Tikanga Māori can be applied at all stages of Study Design and Conduct*

- **Pre-proposal Phase**
 - best practice is engage with Māori communities/partners/entities at research conceptualisation phase (ie., before research starts)
 - Evaluate researcher's readiness to engage with Māori
 - Evaluate research project according to Māori values
- **Dialogue (with Māori) Phase**
 - Consult not inform – open-minded dialogue with willingness to change the study design (which can also lead to better designs...)
 - Discuss questions like: who defined the research problem? What are the benefits and risks and to whom do(n't) they apply? Will study participants be treated with respect? How are Māori values being applied in this study? What expertise in te ao Māori does the research team have? Etc etc
- **Data collection/Study Implementation Phase**
 - Māori governance and oversight... how is this implemented
 - Reporting back to communities on research progress
- **Translation/Implementation Phase**
 - Best practice is Indigenous/Māori control over narratives...
 - Translation into benefits for Māori primarily
 - How are contributions from Māori acknowledged?
 - Who owns IP and how are Māori data being protected?

*Examples only...

Example: Tikanga-based evaluation of a Research Proposal Aimed at Identifying Genes in a Native Tree Species (Kauri) for resistance to a new pathogen...

Value	Trigger Question(s) for Researchers to Answer
Kaitiakitanga (the duty of care, for people and the environment)	<p>How does the project take into account respect for people, the environment and organisms involved?</p> <p>How might project assist in the utilisation of wood from infected trees?</p> <p>How does the project incorporate and acknowledge the kaitiakitanga rights and responsibilities of whānau, hapū and iwi over their environment?</p> <p>How might these tools developed assist Maori communities in identifying/managing ngahere (forests) potentially infected with the disease.</p> <p>What other means of evaluation could be included in this project that would be easy to implement/use?</p> <p>How does the research contribute to resource sustainability?</p>
Mauri (lifeforce)	<p>How might the mauri of the organism(s) or the environment be affected? [affects could be beneficial or detrimental]</p> <p>What are the long term effects on the mauri of the whole forest (ie., not just the individual trees or the organisms?)</p> <p>During the site visit at Huia/Watakere we noticed healthy regeneration under thinning canopy of PTA infected trees, but not under the canopy of healthy kauri. What is this telling us?</p>
Whakapapa (interconnectivity including genealogy)	<p>What are the likely affects, if any, on whakapapa or relationships of the research and/or its implementation?</p> <p>This project will reveal some DNA sequence of a taonga species. How are the researchers accommodating sensitivities regarding such information? How will this be safeguarded?</p> <p>There are dynamic interrelationships in the past (what was), now (what is standing), and in the future (what is coming up/what is to come) among humans, ngahere, taiao. This parallels with the Maori belief of interconnectedness of whakapapa between land and people. How might this research impact those interrelationships/interconnectedness?</p>

There are many tikanga-informed study design guidelines and frameworks ...



Journal of the Royal Society of New Zealand
Volume 36, Number 3, September, 2006, pp 213–227

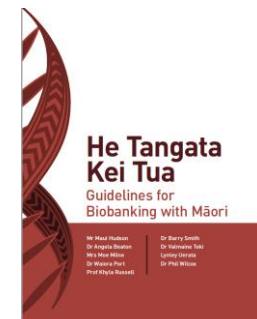
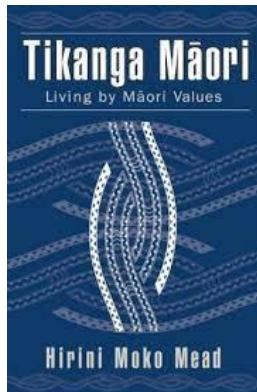
A values-based process for cross-cultural dialogue between scientists and Māori¹

P. L. Wilcox¹, J. A. Chaytor¹, M. R. Roberts^{1,2}, S. Tawharere³, B. Tipene-Matua^{1,3}, E. Kerevuna-Koroi⁴, K. Hunter⁵, H. M. Keast⁶, P. Moles-Denton^{7,8}

Abstract Cross-cultural dialogue is an essential part of the evaluation of controversial technologies and research proposals of significance to indigenous peoples. In this paper we present our experience of developing a process for ensuring that effective processes are developed and implemented to ensure enduring outcomes for their communities. We describe the development of a values-based process for cross-cultural dialogue that starts well before research applications are submitted to funding and/or regulatory agencies. The process begins with processes to 'hui' both the researchers and the Māori partners involved in the research, and continues with the process to have a constructive dialogue with each other concerning the proposal and its intended outcomes.



Ngā Tikanga Paihere draws on 10 tikanga (Te Ao Māori - Māori world concepts) to help you establish goals, boundaries, and principles that guide and inform your data practice.



Mr Mai Hudson

Dr Angela Deaton

Mrs Moi Milne

Dr Waora Part

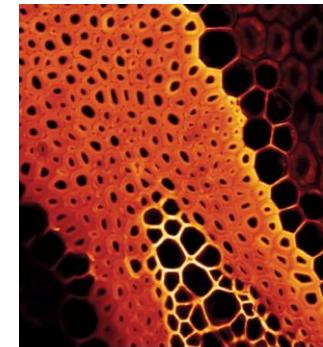
Prof Klyla Russell

Dr Barry Smith

Dr Valentine Toki

Lynley Uerata

Dr Phil Wilcox



Te Nohonga Kaitiaki
Guidelines for Genomic
Research on Taonga Species



RAUIKA MĀNGAI



A WAI 262 BEST PRACTICE
GUIDE FOR SCIENCE
PARTNERSHIPS WITH
KAITIAKI FOR RESEARCH
INVOLVING TAONGA

LESSONS FROM MĀORI VOICES IN THE
NEW ZEALAND SCIENCE SECTOR

JUNE 2022

Co-design – a type of Study Design

What?

- A research process by which researchers, AS WELL AS users, participants and/or communities are involved in defining the research aims, process, analysis and dissemination of research findings
- In other words, *those being studied are part of the team designing and conducting the study...*
- (*Bate & Robert, 2006*): 'True co-design... involves all stakeholders as partners through every stage of the design process—identifying a challenge, engaging people, capturing experiences, understanding experiences, planning improvements and measuring the impact of changes'

Co-design... Why?

- Improved Research Outcomes:
 - Co-design ensures that research aligns with user needs and priorities, leading to more relevant and impactful findings.
- Enhanced Idea Generation:
 - By involving users, co-design facilitates a wider range of perspectives and ideas, leading to more innovative solutions.
- Increased User Satisfaction:
 - Co-design promotes user involvement and empowers them, leading to increased satisfaction with the research process
- Greater Project Success:
 - Co-design can lead to faster implementation, reduced costs, and increased sustainability by aligning research with user needs and priorities from the outset.
- Stronger Relationships:
 - Co-design fosters collaboration and builds trust between researchers, users, and other stakeholders, strengthening partnerships and promoting long-term success.
- Improved Access to Knowledge:
 - Co-design provides researchers with access to user knowledge, expertise, and perspectives, which can be crucial for understanding complex problems and developing effective solutions.
- Increased Relevance and Uptake:
 - By involving users, co-design ensures that research findings and solutions are more relevant and practical, increasing the likelihood of successful implementation and adoption.
- More Equitable Outcomes:
 - Co-design can help address health disparities and improve outcomes for marginalized groups by involving them in the research process and ensuring that their needs are considered.
- Capacity Building:
 - Co-design can empower users and communities, building their capacity to participate in research and contribute to service development.

Adapted from: Goodwin and Boulton (2024) see <https://wairangahau.waipareira.com/wp-content/uploads/2019/11/Rita-Wakefield-What-is-Co-Design-in-a-M%C4%81ori-Space.pdf>

Co-design... how?

- How does this work?
 1. **Learn:** identify right problem, assemble team with right people, and project design = who to work with, what's issue/problem, how to work together to solve
 2. **Design:** how is the research to be conducted? What could the project look like... and what does it look like? Ethical issues are addressed before next phase...
 3. **Do:** undertake the research project... ensuring aspects such as communication are undertaken. Analyse data, communicate results and define implications. Implement.
 4. **Review...** what worked, what didn't...

Co-design... some key things

- Requires trusting relationships
- Must be cognisance of accountabilities of community members to their communities
- Requires time and flexibility
- Transparency in use of frameworks and methods
- Opportunity for reciprocal capacity and capability building
- Must involve the right people
- Appropriate resourcing for communities to undertake study
- Funding for design, implementation and evaluation

Source: https://healthierlives.co.nz/wp-content/uploads/Co-designing-health-research-in-Aotearoa-2024_lessons_digital.pdf

In summary...

- ‘where the data come from’ is not trivial
- There are societal contexts that are VERY important:
 - How data are generated
 - How data are used
 - How data are shared and stored
 - These things can impact lives of people, careers, environments etc etc
- There ways of thinking that can inform what we do to generate data and what we do with them, and tools that can assist...

Learning Outcomes...

By the end of these two lectures you should be able to describe:

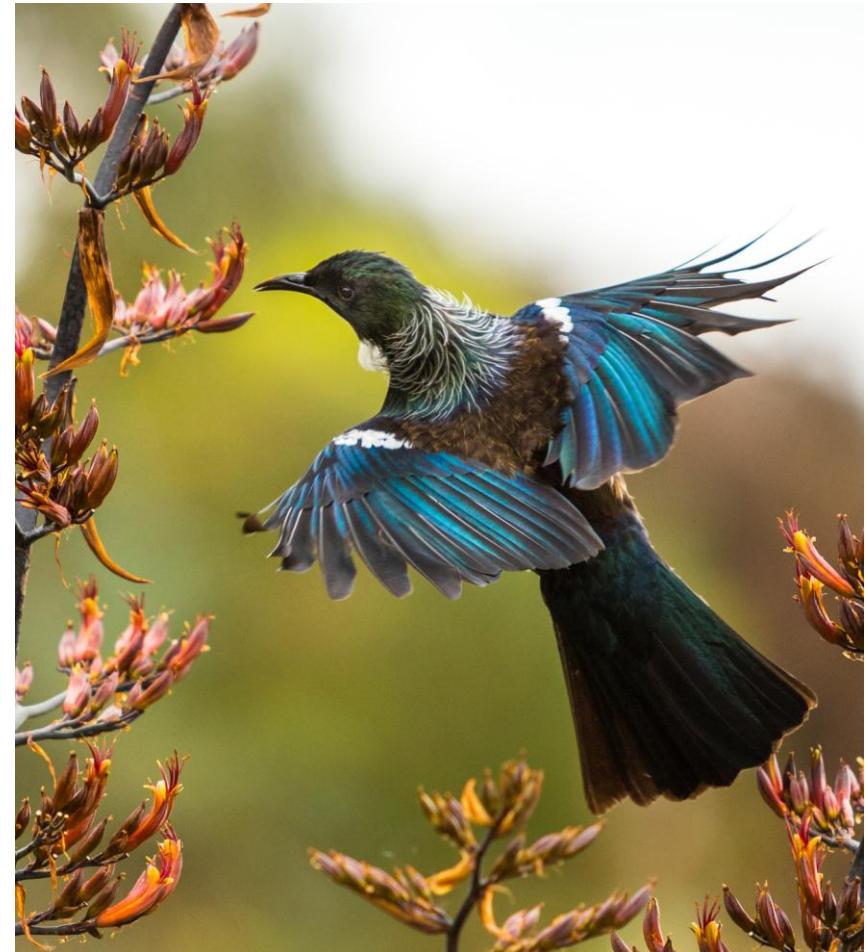
- Examples of how indigenous peoples used data including use of mathematics
- What Māori and indigenous data sovereignty (IDS) are, why these exist, what are the underpinning principles, and what tools that can be used to implement IDS
- What is ethically appropriate (= tikanga informed) study design in an Māori context
- What ‘Co-design’ is (and why...)
- Why you need to know about all these things...

- Final Comment -

Ko te manu e kai ana i te miro, nōna te ngahere.

Engari, ko te manu e kai ana i te mātauranga, nōna te ao.

The bird that eats the fruit of the miro will have the forest as it's domain, but the bird that eat the tree of knowledge will inherit the earth



STAT 110: Week 12

University of Otago

Outline

- Compare different study designs
 - ▶ Focus on understanding relationship between variables
- Experiments
 - ▶ Introduce randomized control trial
- Observational data
 - ▶ Confounding
- Correlation and causation
- Causal inference

Data: Whickham smoking study

- A twenty year study was conducted on 1314 women in Whickham, England¹
 - ▶ Variable 1: Information on mortality (alive/dead)
 - ▶ Variable 2: Smoking status at baseline (smoker: yes/no)
- Represent data in a contingency table

```
##           Smoking_status
##   Outcome    No   Yes  Sum
##   Alive     502  443 945
##   Dead      230  139 369
##   Sum       732  582 1314
```

- We could compare survival probability in the two smoking categories
 - ▶ We know how to do this: prop.test

¹There were also many other variables collected

Whickham smoking study

```
prop.test(x = c(502, 443), n = c(732, 582))

##
## 2-sample test for equality of proportions with continuity correction
##
## data: c(502, 443) out of c(732, 582)
## X-squared = 9, df = 1, p-value = 0.003
## alternative hypothesis: two.sided
## 95 percent confidence interval:
## -0.1252 -0.0256
## sample estimates:
## prop 1 prop 2
## 0.686 0.761
```

Whickham smoking study

- \hat{p}_1 : estimate of 20-year survival probability for non-smokers is 0.686
- \hat{p}_2 : estimate of 20-year survival probability for smokers is 0.761
- The 95% confidence interval for $p_1 - p_2$ is (-0.125, -0.026)
 - ▶ We are 95% confident that the survival probability of non-smokers is between 0.026 and 0.125 lower than that of smokers
- What is going on?!?
 - ▶ Let's ponder that while we look into different ways of collecting data to understand relationships between variables

Experiments and observational data

- Focus on studies that explore relationships between variables
 - ▶ Whickham study: between mortality and smoking status
- Look at two classes of study design:
 - ▶ Experiment
 - Researchers observe outcome after assigning treatments
 - Treatment: variable that could potential change an outcome
 - ▶ Observational study
 - Researchers observe outcome without manipulating any variables

Experiment: Example

- Does studying while listening to classical music improve test performance?
 - ▶ Let's suppose we have 50 participants
- Assign 'treatments'
 - ▶ Assign 25 to study to classical music during a 30 minute study period
 - ▶ Assign other 25 to study in silence during 30 minute study period
- Both groups took a standardized test immediately after the study period
- Compare the scores of the two groups
 - ▶ We have the tools to analyze this data!
 - ▶ Two independent groups

Experiments: randomization

- Randomization is an important principle when designing an experiment
 - ▶ Researchers randomly allocate treatments to experimental units, e.g.
 - allocate fertilizer A or B (treatments) to plots of land (units)
 - allocate new drug or existing drug (treatments) to participants (units)
 - allocate stressful task / neutral task (treatments) to participants (units)
- Idea: avoid systematic differences between the treatment groups
- Example: randomly allocate music / silent study to participants
 - ▶ The distribution of other variables should be approximately the same in the two groups
 - The only difference between the two groups is the treatment
 - Example: Distribution of intelligence should be approximately same in both treatments
- The use of randomization allows us to make causal interpretations
 - ▶ Example: evidence that an increase in test score is caused by studying to classical music

Experiment: Example 2

- Does a new drug reduce deaths in heart attack patients?
- Use randomization to assign ‘treatment’
 - ▶ One group received the new drug
 - ▶ Other group received no drug treatment
- Compare the number of deaths over some time period in the two groups
- This experiment raises a number of other considerations

Experiment: control group

- The control group is an important part of the experimental
 - ▶ It helps determine a baseline (to compare against)
- Example 2: Put yourself into the shoes of someone in the study
 - ▶ In drug group: you receive a brand new drug that you hope will help
 - ▶ In no drug group: downcast, knowing you missed out on an improved chance of survival
- Often studies will introduce a placebo (fake treatment)
 - ▶ Example 2: Participants in the 'no drug' group receive a sugar pill
 - ▶ There can be real improvements in those receiving placebos: placebo effect
- Blind study: participants do not know if they are receiving treatment or placebo
- Double blind study: the doctor does not know if the participant is receiving treatment or placebo

RCTs: issues

- The experiments outlined above are called randomized control trials (RCTs)
 - ▶ They are the gold standard for understanding relationship between variables
- They also have challenges, including
 - ▶ Cost and time: RCTs are expensive and time-consuming to design, implement and monitor
 - ▶ Generalizability: the sampling frame may not match the population of interest
 - e.g. if evaluating a possible treatment for a particular disease, an RCT will generally not include the oldest and sickest individuals
 - ▶ Ethical considerations, including:
 - It may not be ethical to assign 'treatments', e.g. smoking
 - Having participants continue in placebo group once treatment determined to be effective
 - Informed consent

Experiments

- There is much more to experimental design than RCTs
- There are approaches for reducing variability: blocking
 - ▶ Blocks are groups of similar experimental units
 - Example 2: we might have 'low-risk' block and 'high-risk' block
 - ▶ Blocking helps us isolate the effect of treatment by controlling for block variability
- There are experimental designs for more complex situations
 - ▶ e.g. factorial designs (multiple treatments), cross-over designs (each unit receives multiple treatments), etc
- Details of these extensions are not important (for this course)
 - ▶ Good to know that many extensions exist
- Study design explored in STAT 311

Observational studies

- Observational studies: researchers observe participants without intervention
 - ▶ Common in many fields: ecology, earth science, epidemiology, social science, genetics, economics, psychology, ...
- Observational study designs include
 - ▶ Cross-sectional study: collect data at a single point in time
 - ▶ Cohort study
 - Follow groups (cohorts) of participants and observe the occurrence of outcomes
 - ▶ Case-control study
 - Participants grouped on the basis of their outcome status
 - Look back in time for potential factors that might have contributed to the outcome

Observational data

- Observational data: researchers observe participants without intervention
 - ▶ There is no randomization in the variables that could influence the outcome
 - ▶ There may be important variables that can distort the story if omitted
- The Whickham smoking study is an example of observational data
 - ▶ Whickham study: Could there be an important omitted variable?

Back to the Whickham smoking study

- Another variable collected in the study is age (at baseline)
 - ▶ Let's look at two age groups: 18 – 64 and 65+

- Age 18 – 64

```
##          Smoking_status
## Outcome   No  Yes Sum
## Alive    474 437 911
## Dead     65  95 160
## Sum      539 532 1071
```

- Age 65+

```
##          Smoking_status
## Outcome   No  Yes Sum
## Alive    28   6   34
## Dead    165  44 209
## Sum     193  50 243
```

- $\hat{p}_1: 0.879$
- $\hat{p}_2: 0.821$

- $\hat{p}_1: 0.145$
- $\hat{p}_2: 0.12$

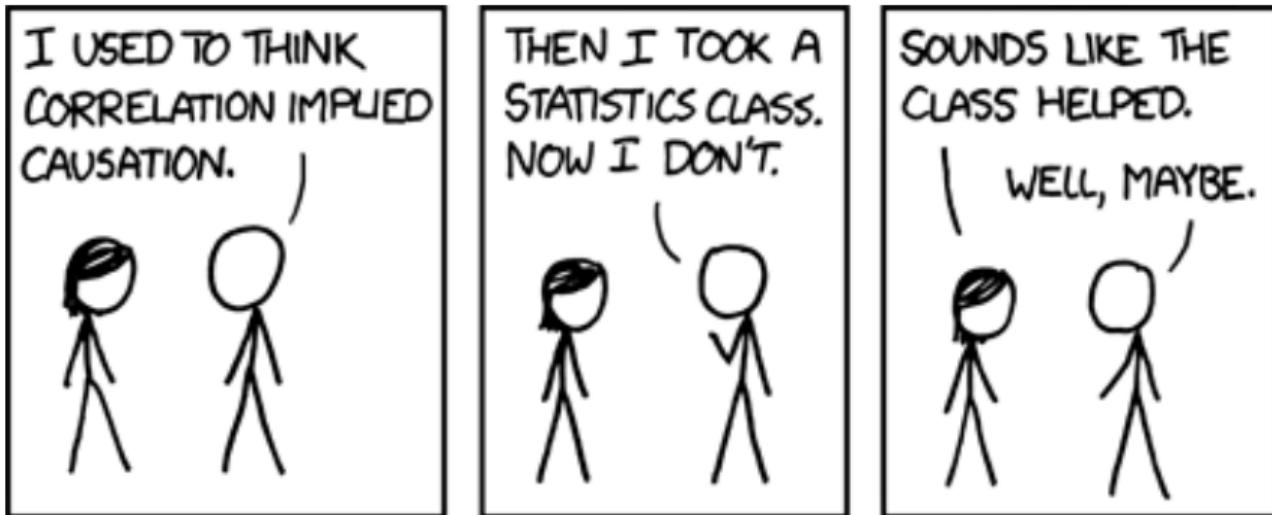
Whickham smoking study

- In each age group
 - ▶ Estimated 20-year survival probability is higher for non-smokers
- The estimated 20-year survival probability is (much) lower for those over 65
 - ▶ Not that surprising
- The proportion of smokers differs considerably between the two age classes
 - ▶ Proportion of smokers (young): $\frac{532}{1071} = 0.497$
 - ▶ Proportion of smokers (old): $\frac{50}{243} = 0.206$
- There are fewer smokers among ‘old’ than ‘young’ (as a proportion)
 - ▶ One possible explanation: many of those who may have been recruited as ‘old’ smokers died before the study began

Confounding

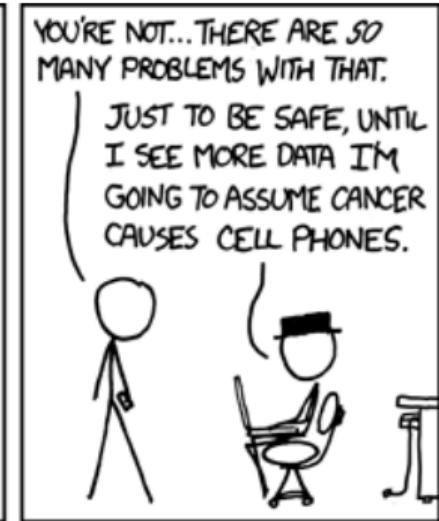
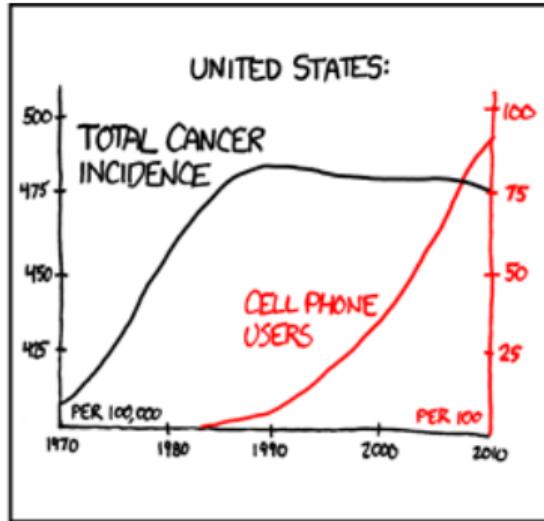
- Whickham study: Age is an example of a confounding variable
- Confounding variable: influences the predictor variable and the outcome variable
 - ▶ Spurious relationship: two variables are associated but not causally related
- Whickham study: Positive association between smoking status and survival
 - ▶ Highly unlikely to be a causal relationship
- We have been careful not to make causal interpretations
 - ▶ Difference in two means, difference in two proportions, linear regression
- Association/correlation: Comparing two (sub)populations
- Causation: a change in x causes a change in y

Correlation and causation²



²<https://xkcd.com/552>

Correlation and causation³



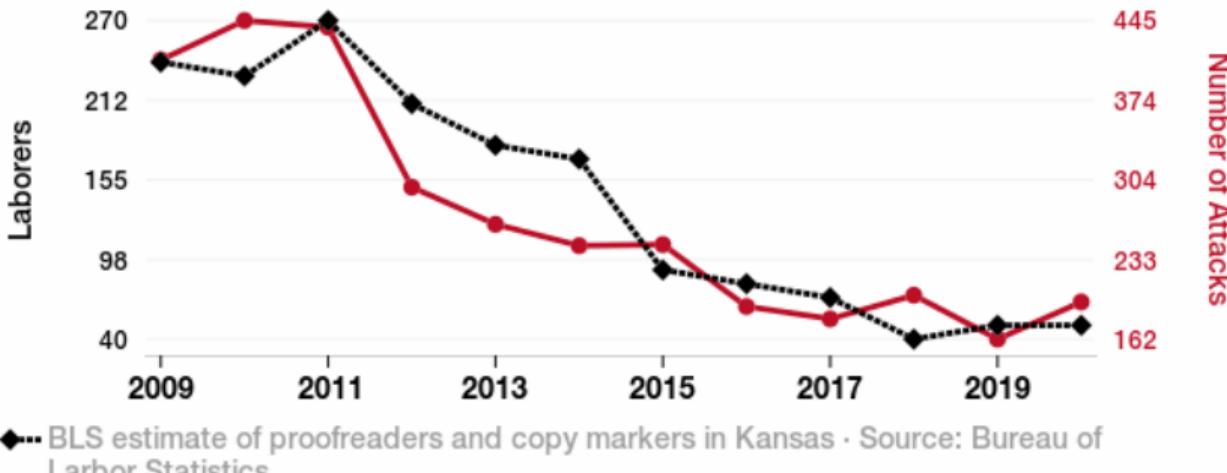
³<https://xkcd.com/925>

Correlation and causation

The number of proofreaders in Kansas

correlates with

Pirate attacks globally

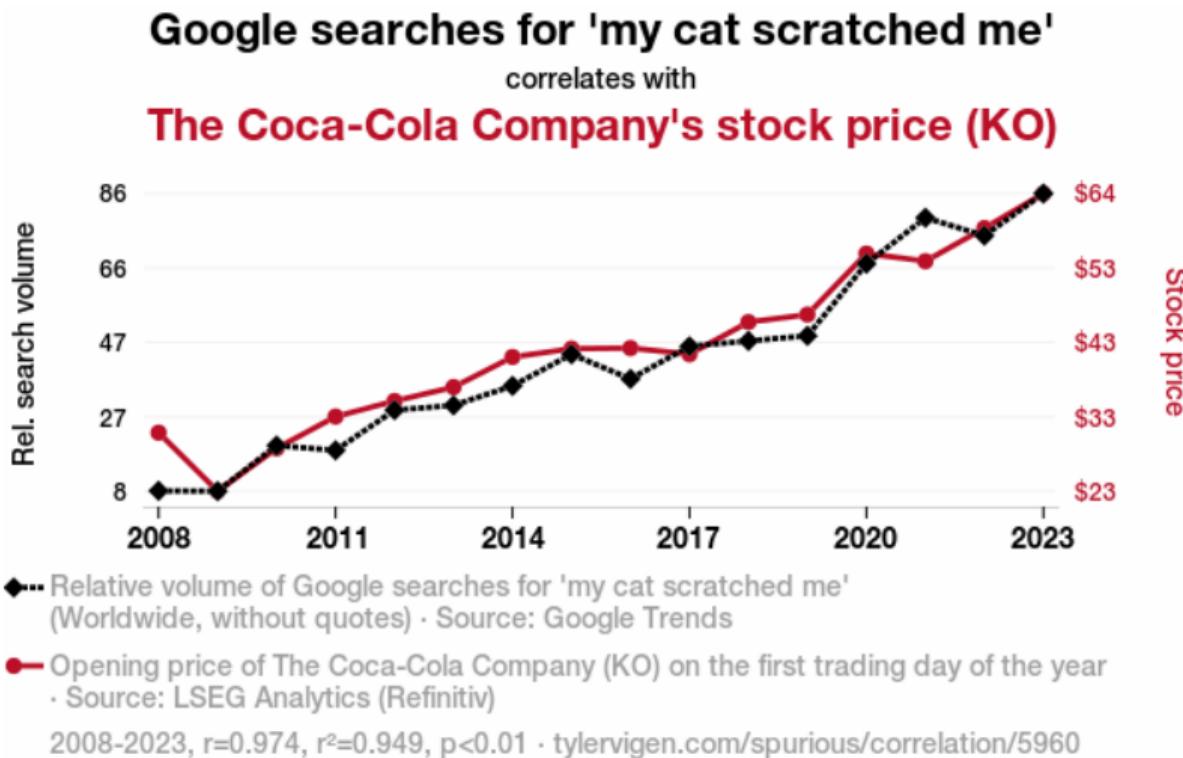


◆ BLS estimate of proofreaders and copy markers in Kansas · Source: Bureau of Labor Statistics

● Global Pirate Attack Count · Source: Statista

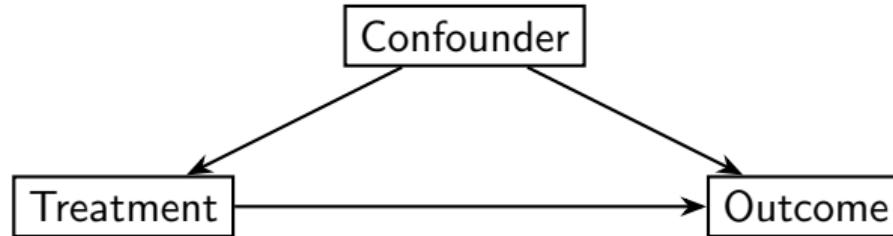
2009-2020, $r=0.914$, $r^2=0.836$, $p<0.01$ · tylervigen.com/spurious/correlation/2334

Correlation and causation

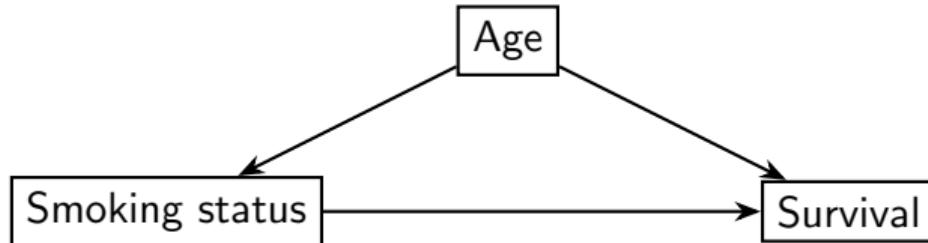


Causal diagram

- Confounding variables can be represented on a causal diagram



- For the Whickham smoking study



Causal inference from observational data

- Observational data is everywhere
 - ▶ Ecology, medical records, genetics, education data, economics
- Are there ways to try and infer causation from observational data?
- Causal inference (from observational data) is a large and active research area
- We must make assumptions about the causal relationship (what causes what)
 - ▶ Can be specified using a causal diagram
- Translate assumptions about causal relationship to an estimate of the causal relationship between a particular variable and the outcome
- More details about causal inference: see STAT 311

Summary

- Compared experiments and observational study designs
- Experiments (RCTs)
 - ▶ Randomization
 - ▶ Blinding
 - ▶ Causal interpretation of effects
- Observational data
 - ▶ Confounding
 - ▶ Correlation and causation
 - ▶ Causal inference
- Whickham example: correlation is not causation with observational data

Outline

- We are looking big picture
- How do we use statistics in 'the real world'
- Discuss the replication crisis
 - ▶ Crisis largely caused by poor statistical and scientific practice
 - ▶ Focused in psychology, but relevant in all disciplines
- Explore how it relates to what we have been taught
- Controversial topic
 - ▶ Everyone has an opinion
 - ▶ Try and provide a balanced view

Replication crisis

- BBC: Most scientists ‘can’t replicate studies by their peers’
- Northwestern: ‘An Existential Crisis’ for Science
 - ▶ “the replication crisis refers to a pattern of scientists being unable to obtain the same results previous investigators found”
- Nature: 1,500 scientists lift the lid on reproducibility
 - ▶ “More than 70% of researchers have tried and failed to reproduce another scientist’s experiments” (based on a survey of 1576 scientists)
- Science: Estimating the reproducibility of psychological science
 - ▶ Authors replicated 100 studies published in three psychology journals
 - ▶ Found 39% of effects were replicated

p-values

- The use of *p*-values and statistical testing took a lot of heat
 - ▶ “The *p*-value plays into the human need for certainty and has led to the reproducibility crisis in many fields”
 - ▶ Scientific American
 - “The concept of statistical significance . . . has emerged as an obvious part of the problem”
 - “The current culture of statistical significance testing, interpretation, and reporting has to go”

p-values

- Others defended *p*-values
- Claim the problem is not with *p*-values themselves, but in how they are used in modern science
- Scientific American (same article as above)
 - ▶ “*p*-values themselves are not necessarily the problem. They are a useful tool when considered in context.”
 - ▶ “Statistical significance is supposed to be like a right swipe on Tinder. It indicates just a certain level of interest. But unfortunately, that’s not what statistical significance has become. People say, ‘I’ve got 0.05, I’m good.’ The science stops.”
 - ▶ “a *p*-value shouldn’t be a gatekeeper . . . Let’s take a more holistic and nuanced and evaluative view.”

A closer look at the problem?

- The problem is often called p-hacking
 - ▶ When we collect or select data or statistical analyses until non-significant results become significant
- Motivated by a desire for $p\text{-value} < \alpha$
 - ▶ Publication bias: it can be difficult to publish null (non-significant) results
 - No evidence of an effect
- There are many things we (as researchers) can do to make a significant results more likely
 - ▶ Destroy the validity of p -values (and confidence intervals) in the process

A fishing expedition?

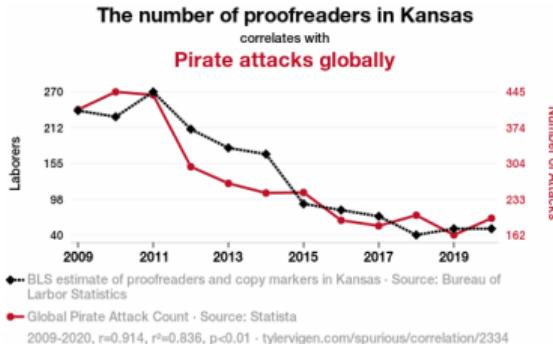
- We might look at many variables until we find combinations that are significantly related
 - ▶ There could be multiple predictor and/or outcome variables
 - e.g. GWAS (genome-wide association study)
 - Compares the DNA of individuals with a specific trait to those without
 - It is common to have more than one million possible predictors
- Often we explore variables ‘formally’
 - ▶ We fit multiple models, calculating CI or p -value until we find significance
- It can also be ‘informal’
 - ▶ Determine which variables are ‘of interest’ while plotting and exploring the data

Multiple comparisons

- This is another example of multiple comparisons
 - ▶ Recall: ANOVA
- Example: suppose we have outcome variable y and we make up (or otherwise procure) k completely unrelated variables
 - ▶ Fit a regression model between y and each variable in turn
 - ▶ If we have k variables with $\alpha = 0.05$, the probability of having at least one significant predictor is
 - $k = 5$: probability = 0.23
 - $k = 10$: probability = 0.4
 - $k = 50$: probability = 0.92
 - $k = 100$: probability = 0.994
 - Probability found using complement

Recall: Spurious relationships

- We saw examples of spurious relationships previously, e.g.



- The website that is from has a section 'Why this works'
 - ▶ "I have 25,153 variables in my database. I compare all these variables against each other to find ones that randomly match up. That's 632,673,409 correlation calculations! This is called 'data dredging'. Instead of starting with a hypothesis and testing it, I instead abused the data to see what correlations shake out. It's a dangerous way to go about analysis"
- If we compare enough variables, we will find a significant correlation

Another xkcd cartoon

- We can't fit this one on the slide!

Multiple Comparisons

- To some degree multiple comparisons can be accounted for
 - ▶ The Bonferroni correction is a popular, general approach (e.g. in GWAS)
- If we conduct m tests, use significance level $\alpha^* = \alpha/m$
- e.g. if we perform $m = 1\,000\,000$ tests with $\alpha = 0.05$
 - ▶ Significance level is $\alpha^* = \frac{0.05}{1000000} = 0.00000005$
- The Bonferroni adjustment is simple to specify and use
 - ▶ It is a general approach
- It ensures the family-wise error rate is less than α
 - ▶ It is conservative
- It is difficult to account for ‘informal’ tests
 - ▶ How do you quantify the decisions made by eye?

Other types of p-hacking

- There are many other choices that we can make when exploring data that can effect the results that can be just as problematic
- Many of these choices are a part of good model building practice
 - ▶ Discuss more in STAT 210
- Easily abused if hunting for a significant result
- Best seen with an example (next slide)

Example

- Example below is taken from [here](#)
 - ▶ We conduct a study testing whether symmetrical faces are more attractive than asymmetrical ones
 - Find no overall difference in attractiveness
 - ▶ So, we test whether the effect differed as a function of the gender of the participant and the gender of the face
 - Find that men found symmetry attractive for faces of both genders, whereas women found symmetry attractive in women's faces but asymmetry attractive in men's faces
 - But, not significant
 - ▶ We examine the data more closely. We notice that some faces were rated as maximally attractive by almost everyone
 - Delete those faces from the analysis because they might obscure a real effect
 - ▶ Other participants were older than the rest and their ratings don't seem to fit the pattern
 - Delete those faces from the analysis because they might obscure a real effect
 - ▶ Now the difference between men and women becomes statistically significant
 - 'Eureka!' we cry

HARKing

- Suppose we then took that significant result and presented it without an explanation of how we got there
 - ▶ Example of HARKing (Hypothesizing after the results are known)
- We explore the data and multiple models (formally or informally) and then present the result as if it was the hypothesis we had in mind all along
- The problem:
 - ▶ p -values and confidence intervals lose their validity

Preregistration

- One approach for improving transparency is to use preregistration
- Simple concept: create a permanent record of our study plans before we look at (or even collect) the data
- It is possible to preregister any (and every) detail of the study
 - ▶ Data-collection plans, analysis code, competing hypotheses, etc
- This does not eliminate exploration
 - ▶ It can make sense to divert from the plan
 - ▶ It is then clear which hypotheses were confirmatory (specified in advance) and what aspects were exploratory and driven by the data
- Open Science: preregistration

ASA statement on *p*-values

- We have seen many mentions about the ASA statement on *p*-values
- There are six principles
 - ▶ Principle 1: P-values can indicate how incompatible the data are with a specified statistical model
 - This is consistent with how we have discussed p-values: they measure incompatibility between the data and the model given by the null hypothesis
 - ▶ Principle 2: P-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.
 - These are common misconceptions

ASA statement on *p*-values

- ▶ Principle 3: Scientific conclusions and business or policy decisions should not be based only on whether a *p*-value passes a specific threshold.
 - “The widespread use of ‘statistical significance’ (generally interpreted as “ < 0.05 ”) as a license for making a claim of a scientific finding (or implied truth) leads to considerable distortion of the scientific process.”
 - “Researchers should bring many contextual factors into play to derive scientific inferences, including the design of a study, the quality of the measurements, the external evidence for the phenomenon under study, and the validity of assumptions that underlie the data analysis.”
- ▶ Principle 4: Proper inference requires full reporting and transparency
 - Exactly what we have been describing above
 - “Conducting multiple analyses of the data and reporting only those with certain *p*-values (typically those passing a significance threshold) renders the reported *p*-values essentially uninterpretable”

ASA statement on *p*-values

- ▶ Principle 5: A p-value, or statistical significance, does not measure the size of an effect or the importance of a result.
 - “Smaller p-values do not necessarily imply the presence of larger or more important effects, and larger p-values do not imply a lack of importance or even lack of effect.”
 - “Any effect, no matter how tiny, can produce a small p-value if the sample size or measurement precision is high enough”
- ▶ Principle 6: By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis
 - “Researchers should recognize that a p-value without context or other evidence provides limited information.”
 - “A p-value near 0.05 taken by itself offers only weak evidence against the null hypothesis”

Concluding remarks

- The ASA statement on *p*-values concludes with:
 - ▶ “Good statistical practice, as an essential component of good scientific practice, emphasizes principles of good study design and conduct, a variety of numerical and graphical summaries of data, understanding of the phenomenon under study, interpretation of results in context, complete reporting and proper logical and quantitative understanding of what data summaries mean. No single index should substitute for scientific reasoning.”
- We like certainty. We like black and white
- The problem is that, with data, we can't be certain. Things are grey

Outline

- Look at approaches for estimation
 - ▶ Maximum likelihood estimation
 - ▶ Bayesian inference

What is statistics?

- How do we collect data to ensure
 - ▶ Representative of the population
 - ▶ We can explore the scientific questions (hypothesis) of interest
- Describe a statistical model for the data
 - ▶ Describes the variability of the data
 - ▶ Estimate the parameters
 - Quantify the uncertainty about the parameters
 - ▶ Interpret these estimates
 - In the context of (scientific) application
 - ▶ Predict new observations
 - ▶ Visualise data and model output

Estimation in context

- Let's focus on how we estimate parameters
 - ▶ Recall: estimate parameters with statistics
- In many of the cases we've seen, we've relied on these as being 'obvious'
 - ▶ Estimate population mean with sample mean
 - ▶ Estimate population variance with sample variance
 - ▶ Estimate p with sample proportion
- Others were more complicated
 - ▶ Used least squares to estimate β_0 and β_1
- There are many complex situations with no obvious estimators
 - ▶ e.g. a model with a parameter related to the skew (shape) of the data
- Can we find general strategies to estimate parameters?

Estimation

- We will look at two such estimation approaches (there are several!)
 - ▶ Maximum likelihood estimation
 - Extensively used in applied statistical work
 - The estimators we have used this semester are maximum likelihood estimators
 - ▶ Bayesian statistics
 - A (very) different approach
 - Use a different definition of probability
 - Seen increasing use in last 30 years
- Looking at a basic understanding of both approaches
 - ▶ We will not sweat the details
 - ▶ Useful to see: if you continue doing research you will likely come across these approaches

Example: Palmer penguins

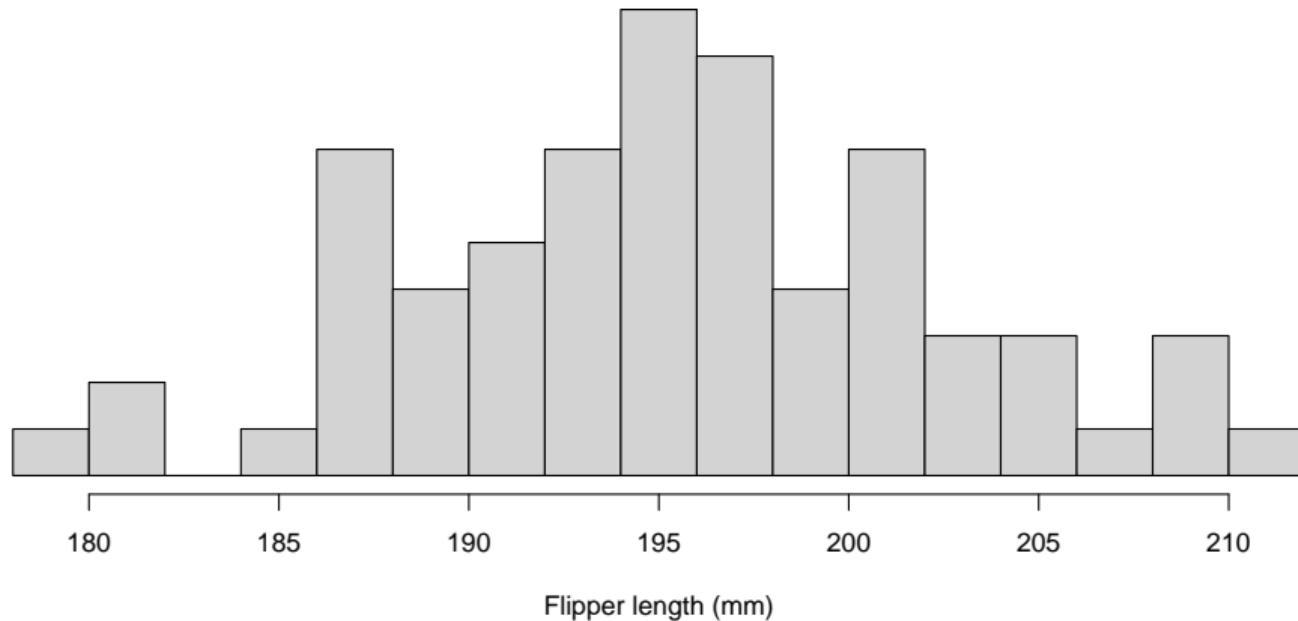
- Flipper and bill length of chinstrap penguins on Palmer archipelago
 - ▶ We will focus on flipper lengths

```
penguin = read.csv('penguin.csv')
```

```
head(penguin)

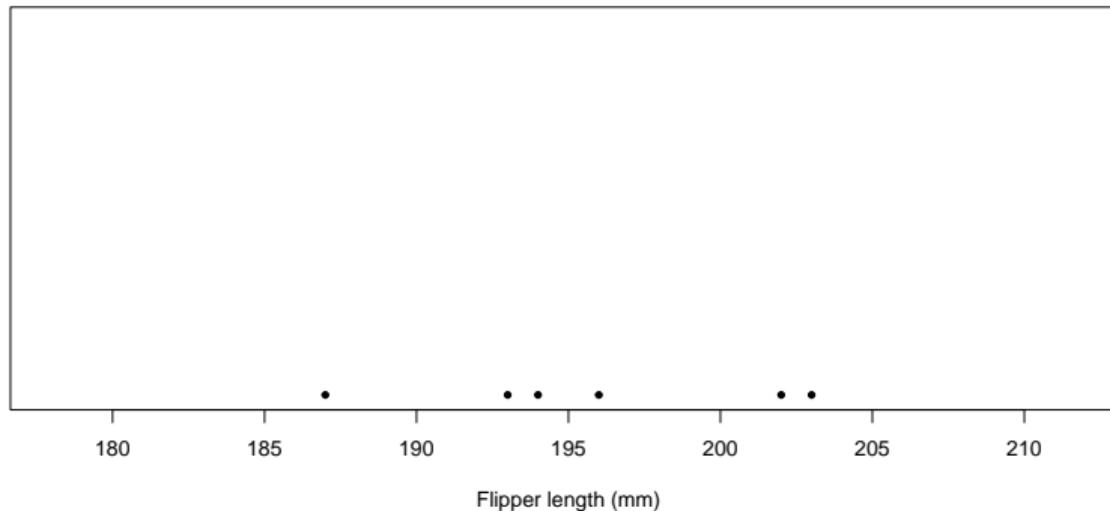
##   bill_length_mm flipper_length_mm
## 1      46.5          192
## 2      50.0          196
## 3      51.3          193
## 4      45.4          188
## 5      52.7          197
## 6      45.2          198
```

Palmer penguins: flipper length



Palmer penguins: flipper length

- To help us understand we reduce the sample down to 6 (randomly chosen) observations

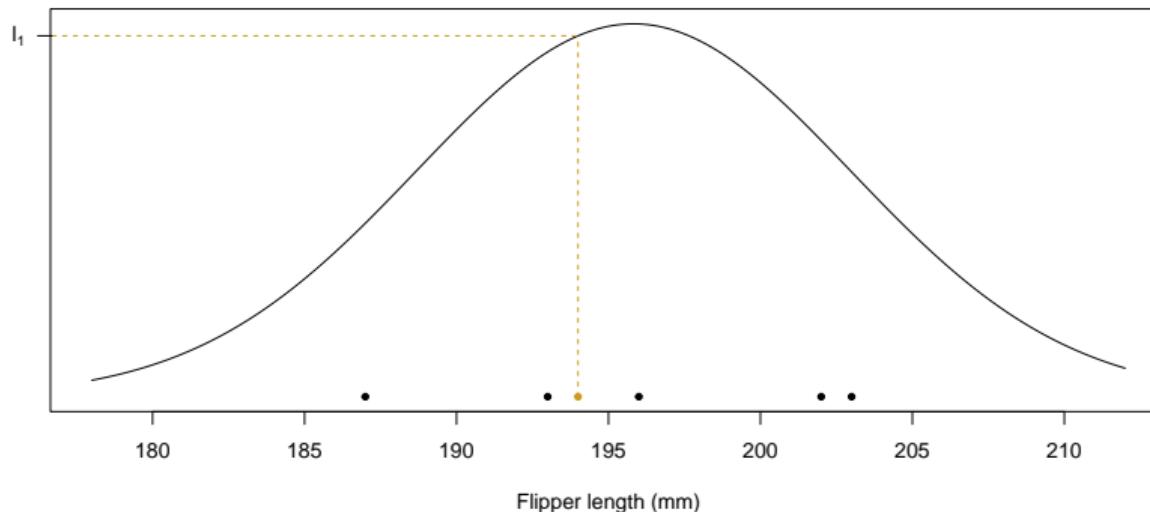


Maximum likelihood estimation

- Idea: we specify a likelihood function
 - ▶ For a given value of the parameters μ and σ
 - ▶ The function gives us a numerical value for how ‘likely’ the parameters are given the data we have observed
- Maximum likelihood
 - ▶ Find the value of the parameters that are most likely
- The likelihood function is given by the probability density function (pdf) of the model
 - ▶ Penguins: normal pdf
- Look at a graphical representation

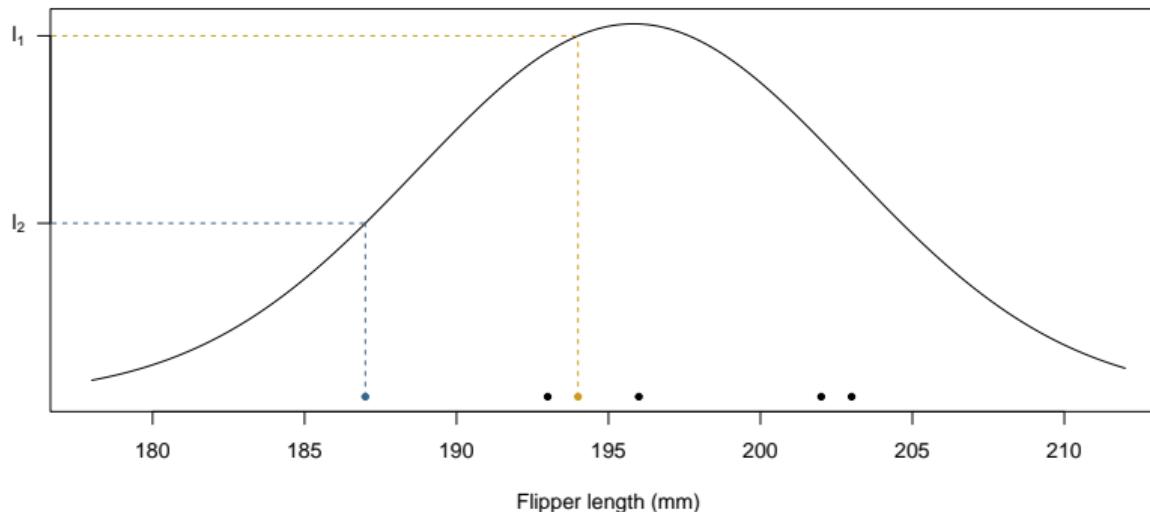
Maximum likelihood: normal model

- Likelihood when $\mu = 195.824$ and $\sigma = 7.132$
 - ▶ Find the likelihood of first observation (in gold)
 - ▶ Given by the value on the y-axis (denoted l_1)



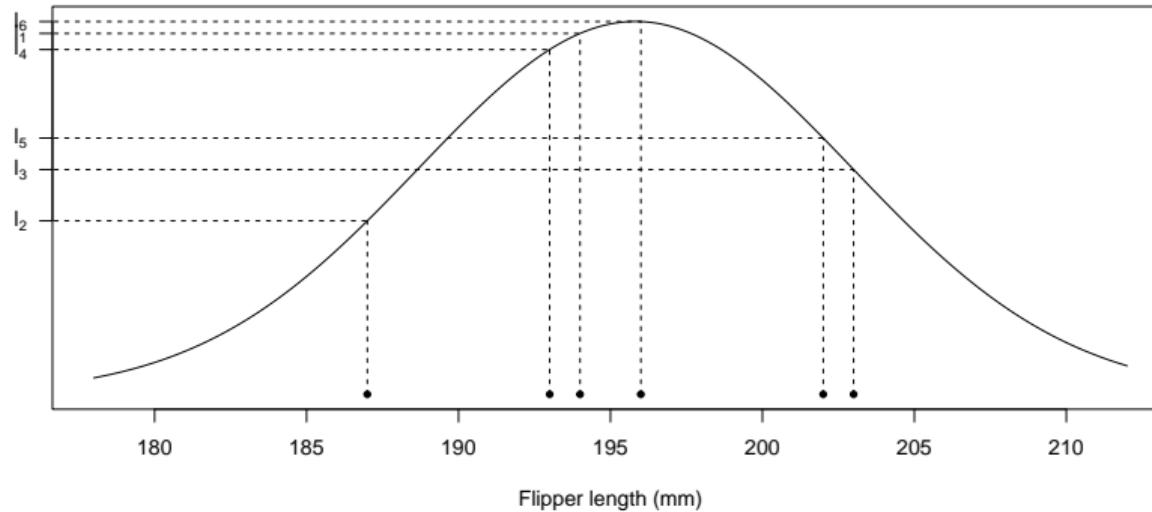
Maximum likelihood: normal model

- Likelihood when $\mu = 195.824$ and $\sigma = 7.132$
 - ▶ Find the likelihood of second observation (in blue)
 - ▶ Given by the value on the y-axis (denoted l_2)



Maximum likelihood: normal model

- Likelihood when $\mu = 195.824$ and $\sigma = 7.132$
 - ▶ Find the likelihood for all observations



Maximum likelihood: normal model

- We want the joint (or combined) likelihood
 - ▶ Multiply together the likelihood for each observation: $l_1 \times l_2 \times \dots \times l_6$
 - ▶ Usually just called the likelihood
- We find the value of μ and σ so that the joint likelihood is as large as possible
 - ▶ Hence the name, maximum likelihood
- For many models we can find the maximum likelihood estimator mathematically
 - ▶ Normal model: $\hat{\mu} = \bar{y}$
 - ▶ Linear regression with normal errors: least squares and maximum likelihood estimators are the same
 - ▶ Binomial model: $\hat{p} = \frac{x}{n}$
- We've been using maximum likelihood without realising it!

Maximum likelihood in practice

- We have demonstrated this for a normal model
- Same process can be used for any statistical model
 - ▶ General approach for estimating a model
- Maximum likelihood estimation is explored more in STAT 270 and 370
 - ▶ How do we find maximum likelihood estimators mathematically?
 - ▶ What is the sampling distribution?
 - ▶ What is the standard error?
 - ▶ Are maximum likelihood estimators 'good' estimators?

Bayesian inference

- In the past 40 years, Bayesian inference has surged in popularity
 - ▶ Increasingly used in application areas



Empirical Article | Open Access |

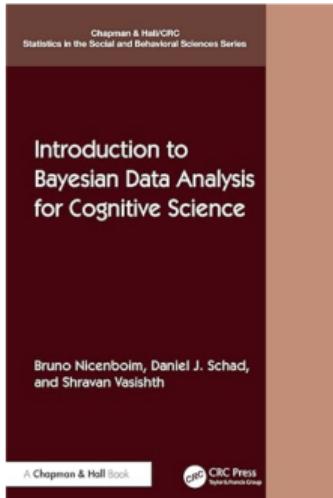
Navigating the Bayes maze: The psychologist's guide to Bayesian statistics, a hands-on tutorial with R code

Udi Alter , Miranda A. Too, Robert A. Cribbie

First published: 19 December 2024 | <https://doi.org/10.1002/ijop.13271>

Bayesian inference

- In the past 40 years, Bayesian inference has surged in popularity
 - ▶ Increasingly used in application areas



Introduction to Bayesian Data Analysis for Cognitive Science
(Chapman & Hall/CRC Statistics in the Social and Behavioral Sciences) 1st Edition

by [Bruno Nicenboim](#) (Author), [Daniel J. Schad](#) (Author), [Shravan Vasishth](#) (Author)



[See all formats and editions](#)

Savings Pre-order Price Guarantee. [Terms](#)

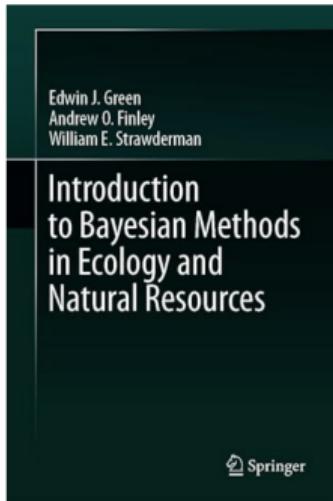
This book introduces Bayesian data analysis and Bayesian cognitive modeling to students and researchers in cognitive science (e.g., linguistics, psycholinguistics, psychology, computer science), with a particular focus on modeling data from planned experiments. The book relies on the probabilistic programming language Stan and the R package brms, which is a front-end to Stan. The book only assumes that the reader is familiar with the statistical programming language R, and has basic high school exposure to pre-calculus mathematics; some of the important mathematical constructs needed for the book are introduced in the first chapter.

Through this book, the reader will be able to develop a practical ability to apply Bayesian modeling within their own field. The book begins with an informal introduction to foundational topics such as probability theory, and univariate and bi-/multivariate discrete and continuous random variables. Then, the application of Bayes' rule for statistical inference is introduced with several simple analytical examples that require no computing software; the main insight here is that the posterior distribution of a parameter is a compromise between the prior and the likelihood functions. The book then gradually builds up the regression framework using the brms package in R,

<https://bruno.nicenboim.me/bayescogsci/>

Bayesian inference

- In the past 40 years, Bayesian inference has surged in popularity
 - ▶ Increasingly used in application areas



Introduction to Bayesian Methods in Ecology and Natural Resources 1st ed. 2020 Edition, Kindle Edition

by [Edwin J. Green](#) (Author), [Andrew O. Finley](#) (Author), [William E. Strawderman](#) (Author) | Format: Kindle Edition
5.0 [See all formats and editions](#)

This book presents modern Bayesian analysis in a format that is accessible to researchers in the fields of ecology, wildlife biology, and natural resource management. Bayesian analysis has undergone a remarkable transformation since the early 1990s. Widespread adoption of Markov chain Monte Carlo techniques has made the Bayesian paradigm the viable alternative to classical statistical procedures for scientific inference. The Bayesian approach has a number of desirable qualities, three chief ones being: i) the mathematical procedure is always the same, allowing the analyst to concentrate on the scientific aspects of the problem; ii) historical information is readily used, when appropriate; and iii) hierarchical models are readily accommodated.

This monograph contains numerous worked examples and the requisite computer programs. The latter are easily modified to meet new situations. A primer on probability distributions is also included because these form the basis of Bayesian inference.

Researchers and graduate students in Ecology and Natural Resource Management will find this book a valuable reference.

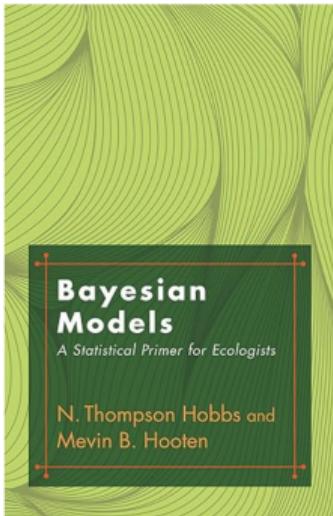
[▼ Read more](#)

[Read sample](#)

ISBN-13	Edition	Publisher	Publication date	Language
978-3030607500	# 1st ed. 2020	Springer	November 26, 2020	English

Bayesian inference

- In the past 40 years, Bayesian inference has surged in popularity
 - ▶ Increasingly used in application areas



Bayesian Models: A Statistical Primer for Ecologists

by [N. Thompson Hobbs](#) (Author), [Mevin B. Hooten](#) (Author)

4.7 ★★★★☆ (40) 4.3 on Goodreads 24 ratings



[See all formats and editions](#)

Bayesian modeling has become an indispensable tool for ecological research because it is uniquely suited to deal with complexity in a statistically coherent way. This textbook provides a comprehensive and accessible introduction to the latest Bayesian methods—in language ecologists can understand. Unlike other books on the subject, this one emphasizes the principles behind the computations, giving ecologists a big-picture understanding of how to implement this powerful statistical approach.

Bayesian Models is an essential primer for non-statisticians. It begins with a definition of probability and develops a step-by-step sequence of connected ideas, including basic distribution theory, network diagrams, hierarchical models, Markov chain Monte Carlo, and inference from single and multiple models. This unique book places less emphasis on computer coding, favoring instead a concise presentation of the mathematical statistics needed to understand how and why Bayesian analysis works. It also explains how to write out properly formulated hierarchical Bayesian models and use them in computing, research papers, and proposals.

[▼ Read more](#)

[Report an issue with this product or seller](#)

ISBN-10



0691159289

ISBN-13



978-0691159287

Publisher



Princeton
University Press

Publication date



August 4, 2015

Language



English



Bayesian inference: Applied probability

- Bayesian statistics return to earlier discussions about probability
- Recall: Mentioned there are several interpretations of probability
 - ▶ We previously relied on a frequentist definition
- Bayesian statistics: interprets probability as a measure of belief in the occurrence of an event
 - ▶ Often called subjective or personal probability
- Example: The All Blacks played France at Lancaster Park on 26 June 1994
 - ▶ What is the probability the All Blacks won?

Bayesian inference: Applied probability

- Bayesian statistics return to earlier discussions about probability
- Recall: Mentioned there are several interpretations of probability
 - ▶ We previously relied on a frequentist definition
- Bayesian statistics: interprets probability as a measure of belief in the occurrence of an event
 - ▶ Often called subjective or personal probability
- Example: The All Blacks played France at Lancaster Park on 26 June 1994
 - ▶ What is the probability the All Blacks won?
 - ▶ Does it make sense to use probability here?

Bayesian inference: Applied probability

- Bayesian statistics return to earlier discussions about probability
- Recall: Mentioned there are several interpretations of probability
 - ▶ We previously relied on a frequentist definition
- Bayesian statistics: interprets probability as a measure of belief in the occurrence of an event
 - ▶ Often called subjective or personal probability
- Example: The All Blacks played France at Lancaster Park on 26 June 1994
 - ▶ What is the probability the All Blacks won?
 - ▶ Does it make sense to use probability here?
 - ▶ The event has happened (it is fixed and not random)
 - Probability makes no sense under frequentist interpretation
 - ▶ It seems reasonable to use probability
 - Using probability as measure of belief (in the All Blacks winning)

Bayesian inference: Applied probability

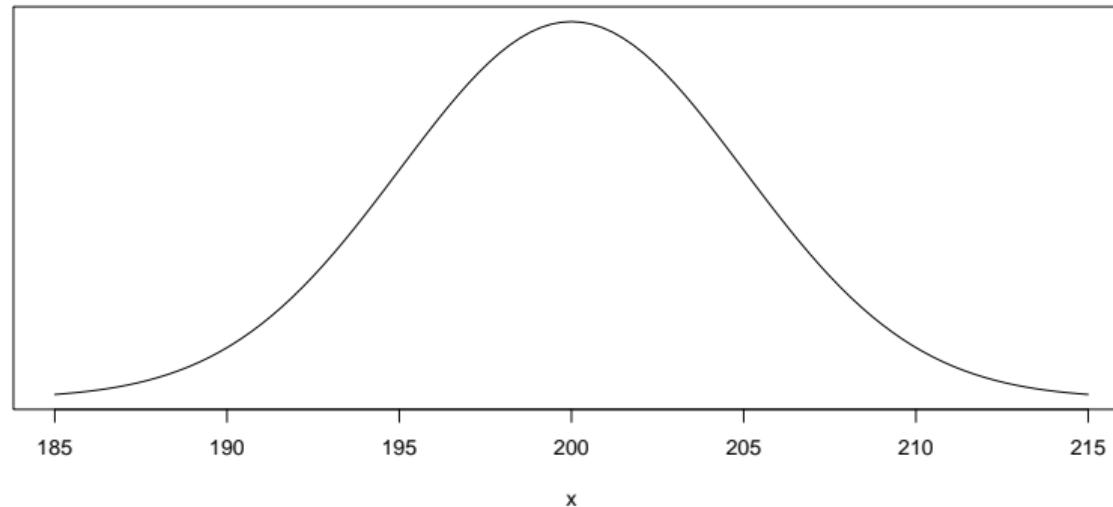
- Use probability to describe our ‘belief’ about the parameters
 - ▶ Use probability to describe uncertainty about the parameters
- There are two such probability distributions
 - ▶ Prior distribution: belief about the parameters before study conducted
 - ▶ Posterior distribution: belief about the parameters given data observed
- These are found using Bayes theorem (we saw this earlier!)
- The posterior distribution is what we use to get
 - ▶ Estimate (a point estimate)
 - ▶ Uncertainty (an interval estimate)

Bayesian inference: posterior distribution

- Posterior distribution found by combining (multiplying) likelihood and prior
 - ▶ Likelihood: same likelihood as above
 - ▶ Note: there is some additional mathematical complexity that we can ignore here
- We will look at the process graphically
- We will ignore many aspects
 - ▶ Mathematical details
 - ▶ In-depth discussion about the prior distribution
 - It has been a (historically) controversial aspect of Bayesian inference

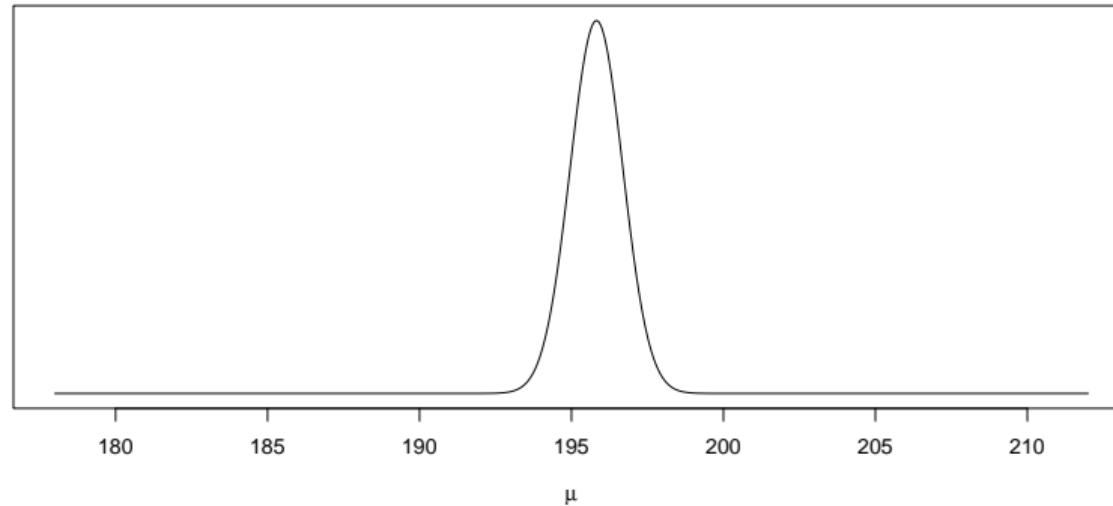
Prior

- The prior distribution describes belief about the parameters before study conducted
 - ▶ We may have a prior centered on mean flipper length of 200 mm



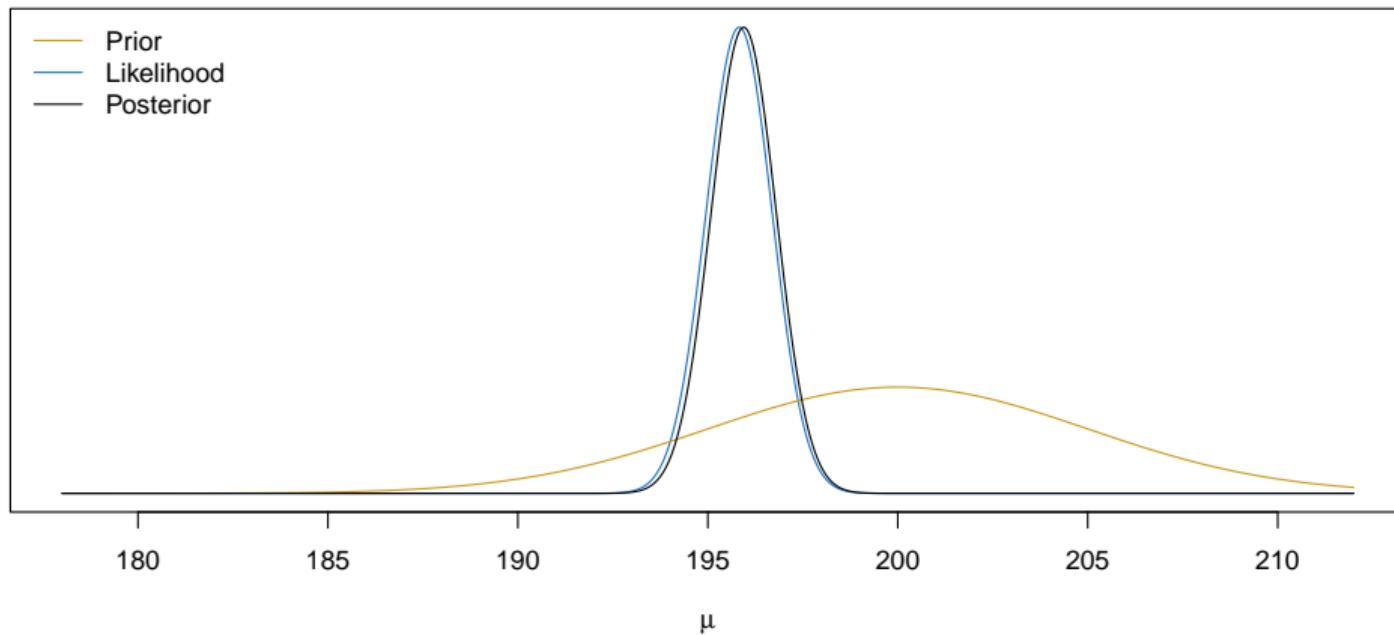
Likelihood

- The likelihood we describe above can be seen graphically
 - ▶ Likelihood for parameter μ (for a given value of σ)
 - The larger the value, the more likely the parameter value given the observed data



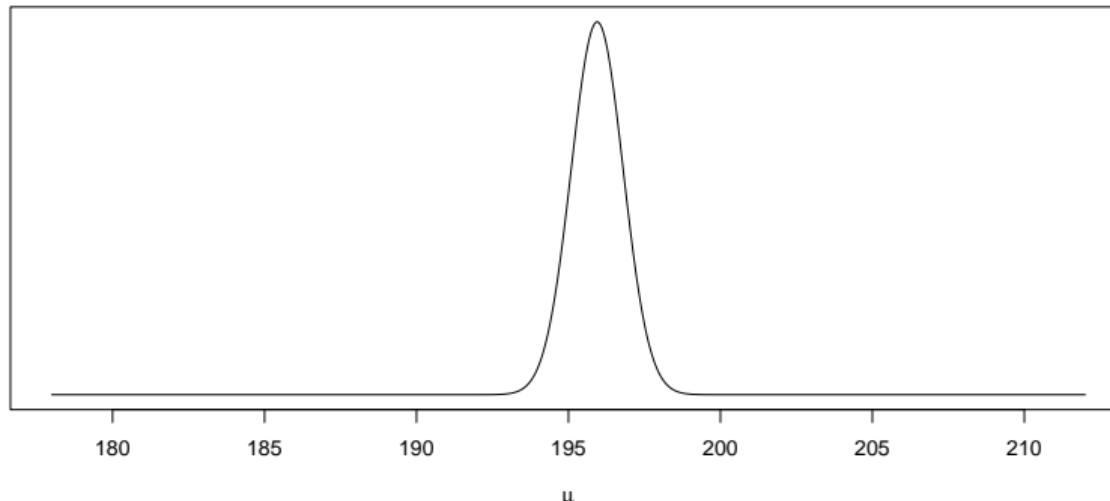
Posterior distribution

- The posterior distribution combines the likelihood and prior
 - ▶ $\text{Posterior} \propto \text{Likelihood} \times \text{Prior}$



Posterior distribution

- The posterior distribution describes our belief about the parameters given data observed
 - ▶ Probabilistic description of what value we think μ is
 - ▶ Can obtain a point and interval estimates
 - Summaries of the posterior distribution



Why has it become popular?

- For many complex (realistic) problems
 - ▶ We can fit Bayesian models, where 'standard' approaches are prohibitively difficult
 - ▶ There are software packages for fitting Bayesian models
- Revolutionized applied statistical modeling in the last 30 years
- Explore Bayesian modeling in STAT 371

Summary

- Introduced maximum likelihood and Bayesian modeling
 - ▶ Heard the terminology
 - ▶ Likely to come across one or both terms if continue into research involving data
- To delve deeper into these approaches
 - ▶ We need a better understanding of probability
 - ▶ We need some understanding of calculus
 - ▶ We explore the approaches in higher level courses (STAT 270, 370, 371)

STAT 110: Review lecture

University of Otago

Outline

- Big picture review of the course
- Connect key elements to (practice) exam questions
- I have not included the context in many cases

Data

- We looked at data, summaries, and R
 - ▶ The R object `penguin` contains information on a random sample of chinstrap penguins from the Palmer archipelago. There are two variables: `bill`, the bill length (mm), and `flipper`, the flipper length (mm). We consider the R code below:

```
mean(penguin$bill)  
sd(penguin$flipper)
```

- What is being evaluated in the first line of R code: `mean(penguin$bill)`?
- ▶ We observe data $y = (39.7, 41.3, 44.4, 39.0, 45.5)$
 - The sample mean \bar{y} is closest to

Probability and random variables

- We want to fit statistical models
- We need knowledge of probability¹
 - ▶ What is the best interpretation of $\Pr(B|V^C)$?
 - ▶ The probability $\Pr(V|B)$ is closest to
 - ▶ Find the quantity $E[Y]$
 - ▶ What is the best description of a random variable?
 - ▶ What is the expected nutrient score per serving, $E[2X - 3Y]$?

¹I haven't included the context for these questions.

The normal distribution

- Looked in detail at the normal distribution
 - ▶ Working memory span refers to the amount of information a person can temporarily hold and manipulate in their mind while performing a cognitive task. A score of working memory span has been developed that is normally distributed with mean $\mu = 40$ and standard deviation $\sigma = 8$ for healthy adults in the population.
 - A randomly selected healthy adult has a working memory score that is 1.5 standard deviations below the mean ($z = -1.5$). Their working memory score is closest to
 - Which of the following options calculates the probability that a randomly selected healthy adult has a score above 48?
- We found the sampling distribution for \bar{y}
 - ▶ If we were to collect a sample of $n = 64$ healthy adults and calculate their working memory score, select the option below that best describes the sampling distribution of the sample mean \bar{y}

Normal models

- One sample & paired data
 - ▶ The R code below carries out the hypothesis test:

$$H_0 : \mu_d = 0; \quad H_A : \mu_d \neq 0,$$

where μ_d is the mean difference in the nitrogen levels (after-before). If $\alpha = 0.05$, select the best interpretation:

- Two independent samples
 - ▶ The sample mean reduction for drug A is $\bar{y}_1 = 19.00$ with sample standard deviation $s_1 = 13.579$. The sample mean reduction for drug B is $\bar{y}_2 = 15.95$ with sample standard deviation $s_2 = 9.054$. The estimated standard error for $\bar{y}_1 - \bar{y}_2$ is closest to
 - ▶ Which of the following options should we use to find a 95% confidence interval for $\mu_1 - \mu_2$
 - ▶ The R output of a suitable model is below. Select the best interpretation

Normal models

- ANOVA
 - ▶ Select the hypotheses that are being tested with ANOVA
 - ▶ The F-value for the appropriate test is closest to
 - ▶ Select the option that is not correct with respect to ANOVA
 - ▶ Select the best interpretation of the p-value from the ANOVA test

Linear regression

- Understanding the linear regression model
 - ▶ What is the best interpretation of β_1 ?
 - ▶ Which of the following is correct for the subpopulation of mammals that have body mass of 20kg?
 - ▶ Does it make sense to interpret $\hat{\beta}_0$ in this application?
- Estimating / fitting a linear regression model
 - ▶ What is the best description of the method used to estimate the parameters in the linear regression model below?
 - ▶ Select the correct expression for the fitted regression model based on the R output
- Assumptions
 - ▶ Suppose that we fit a linear regression model with outcome y and predictor variable x .
Based on the plot below, select the option that best describes which regression assumptions, if any, appear to be violated

Linear regression

- Prediction
 - ▶ The researchers want to use the model to predict the aptitude of a child who first speaks at 60 months. This quantity is closest to
 - ▶ The code below finds two intervals. The type of interval is hidden (we have replaced the type of interval by A and B). Select the best description of these intervals
- Multiple linear regression
 - ▶ Which of the following statements about multiple linear regression is correct?
 - ▶ Researchers fit a model that includes both temperature and activity. Select the option that gives $\hat{\beta}_2$ and the standard error for $\hat{\beta}_2$.
- Categorical predictors: see Assignment 8
- Model fit
 - ▶ The R^2 is 84.8%. Which of the statements below is not correct.

Binary/binomial models

- Assumptions
 - ▶ Researchers are studying how frogs respond to a predator cue. They expose individual frogs to the cue and record whether each frog jumps away (yes/no). They continue collecting data until they observe 20 frogs that run away. Which of the binomial assumptions, if any, are violated?
- Model fitting and interpretation
 - ▶ The sample proportion of field goals made from less than 50 yards is closest to
 - ▶ A confidence interval can be found using `prop.test` as below. Select the best description of the parameter being estimated by the confidence interval shown in the output
 - ▶ What hypothesis test is being carried out when using `prop.test`

Contingency table

- χ^2 -test
 - ▶ If we assume independence between diet and cancer, the expected count of those with a moderate diet of fish and no cancer is closest to
 - ▶ What is the appropriate hypotheses for the χ^2 -test for contingency tables.
 - ▶ What are the degrees of freedom for the χ^2 -test?
 - ▶ Select the best interpretation from the χ^2 -test below if $\alpha = 0.05$

Other methods

- Nonparametric methods
 - ▶ Select the option that best describes how the Mann-Whitney test statistic is found
 - ▶ Which of the following is a benefit of using a non-parametric test such as the Mann-Whitney test?
 - ▶ Interpret the test carried out below if $\alpha = 0.05$
- Central limit theorem
 - ▶ As a summer research project we develop a new working memory score that is not normally distributed but still has mean $\mu = 40$ and standard deviation $\sigma = 8$ (we can assume it is not excessively skewed). If we were to collect a sample of $n = 64$ healthy adults and calculate their working memory score, select the option below that best describes the sampling distribution of \bar{y}

Where is the data from?

- Sampling
 - ▶ Which of the following best describes stratified sampling?
 - ▶ ...which of the following is likely to be the largest source of bias, and why?
- Culturally informed design and governance
 - ▶ Which of the following is not a characteristic of co-designed research studies?
 - ▶ In Indigenous data sovereignty, the CARE acronym refers to which of the following
- Experiments and observational data
 - ▶ What is the best description of a placebo group?
 - ▶ What is a confounding variable?

Other topics

- Replication crisis
 - ▶ What is the replication crisis in science primarily about?
 - ▶ What is the best description of HARKing?
 - ▶ What is the main danger of performing many statistical tests without adjustment?
- Estimation of statistical models
 - ▶ Which of the following best describes maximum likelihood estimation (MLE)?
 - ▶ Which of the following is a feature of Bayesian inference?

Summary

- The final exam is comprehensive
 - ▶ Questions cover the entire course
 - ▶ Questions cover all of the learning outcomes for the course
 - Be able to describe the information contained in a data set
 - Be able to carry out common statistical data analyses in R
 - Be able to interpret the results of common statistical analyses in the context of the scientific study
 - Be aware of the appropriate use of study designs
 - Be able to understand advantages and disadvantages of various statistical procedures
- Keep an eye out for exam help sessions closer to the exam