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

[1] I 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

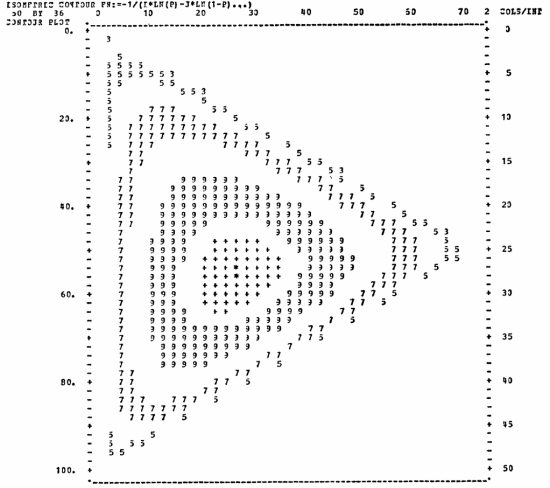
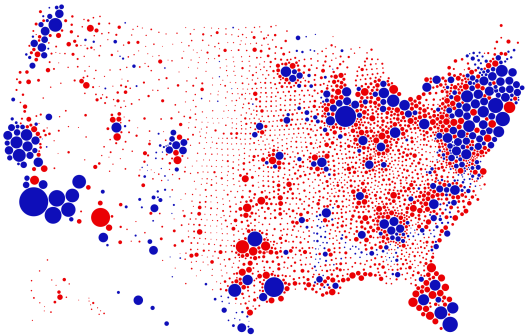


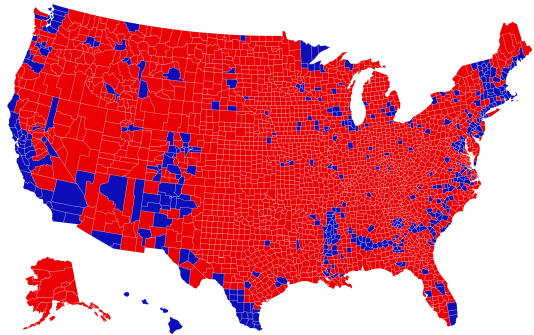
FIG. 1. Isometric contour plot for election data.

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%

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





[1] From @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

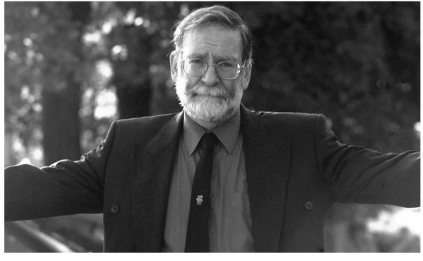
Example 1: Harold Shipman

- 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 Glaave, PEN 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

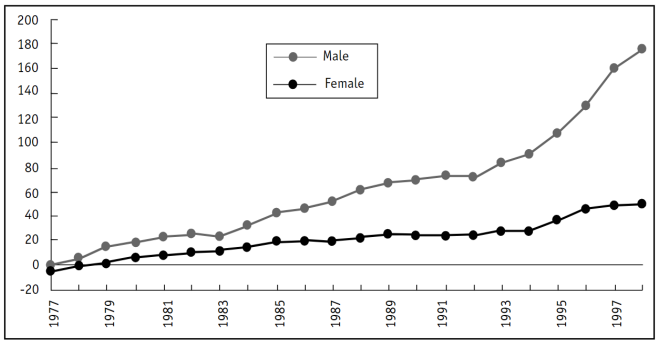
Variability

- 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 Letby[2]:
  - 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.*"

[2]Both 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

[a] We 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 panes[3]:
  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)

[3] The 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 Archipelago[4]
  - Group of islands off the northwestern coast off Antartica
- Measurements of flipper length and bill length for a sample of gentoo penguins[5]
- We will interact with this (and the larger dataset) a few times this semester

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

[5]Photo: 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
```

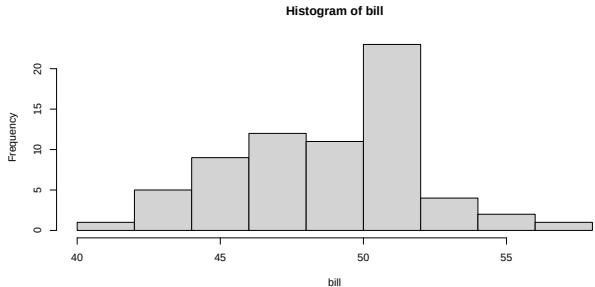
- Assign the new variable to `bill`
  - Use = or <-
  - Use these values later (next slide!)

```
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]

[6] A 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

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

Into R

- Step 1: call data into R
  - Import using menu (File Import Dataset)
  - Use commands

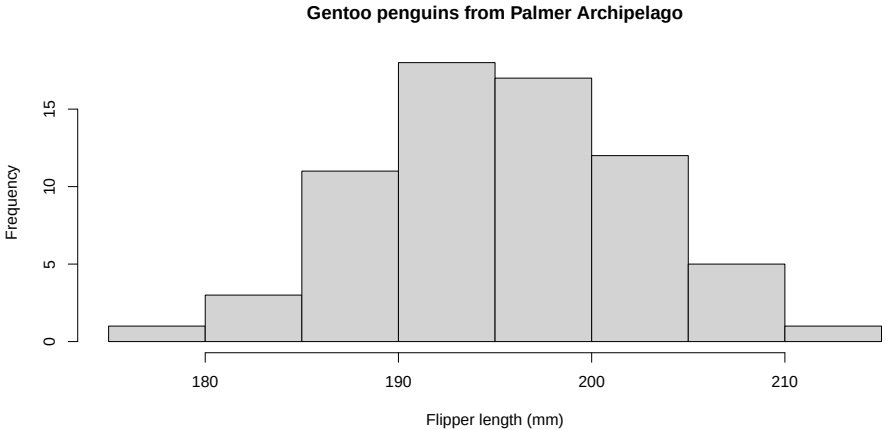
```
peng_lect1 = read.csv('peng_lect1.csv')
```

  - `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


Gentoo penguins from Palmer Archipelago
Frequency
0
5
10
15
180
190
200
210
Flipper length (mm)

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

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

- 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

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

- 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

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

- $\sum$ 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
```

[9] The 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 $y_1, \ldots, y_n$ 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 $y_1, \ldots, y_n$ 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

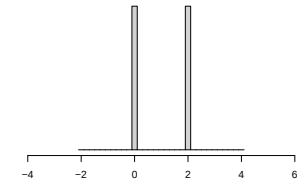
Data summaries: big picture

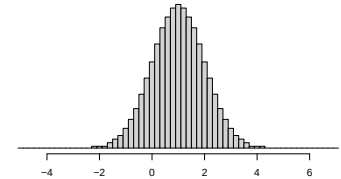
- On one hand: lost a lot of information
  - $n = 68$ into two numbers
- On the other hand: created order out of chaos
  - It is hard for us to get an understanding of $n = 68$ values [10]
  - Summarized the data to gain an understanding about important features of the data
  - Suppose we got $n = 205$ 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

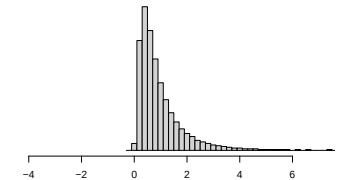
[10] It is even worse if we have $n = 6800$ values!

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