STAT340
Data Modelling II

Lecture 0: Introduction and Administrivia

Why study data science?

Q: How effective is the Pfizer vaccine against COVID-19?

FDA NEWS RELEASE

FDA Approves First COVID-19 Vaccine

We frequently see figures in the news about things like vaccine safety, disease mortality rates, infectiousness, etc. What do these figures actually mean, and how do researchers arrive at them?

Why study data science?

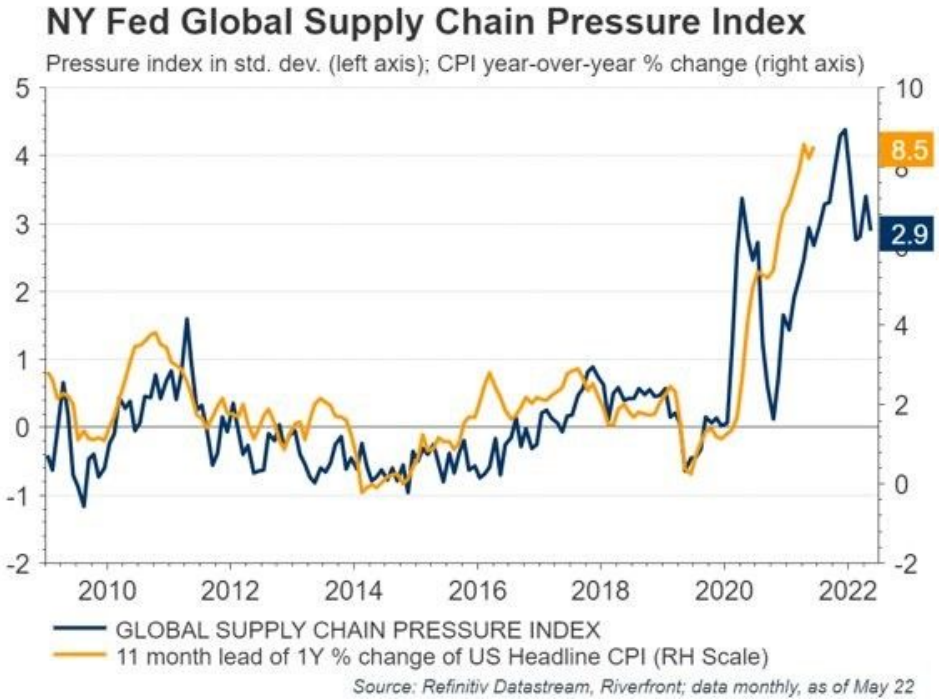
Q: Do early intervention programs like Head Start improve educational outcomes?

Assessing the effectiveness of government programs is of huge importance to policy makers and to offices charged with implementing those programs. How do we measure whether or not an intervention helped?

Why study data science?

Q: What was the rate of inflation in August of 2022?

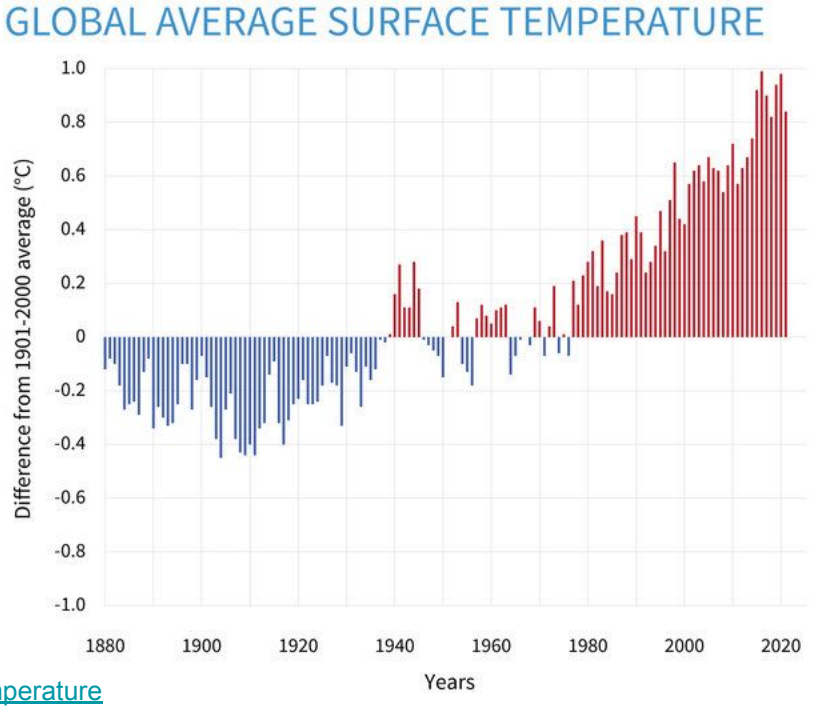
The government and the media report inflation figures like “inflation was X% in August 2022”. What is meant by these numbers, and how are they measured?



Why study data science?

Q: Are record high temperatures in recent years explainable by chance?

When climate scientists make claims about current rates and possible future trajectories of climate change (and likelihood thereof), what do they mean, and how do they arrive at those figures?



https://www.climate.gov/news-features/understanding-climate/climate-change-global-temperature

Why study data science?

These questions are a little too complicated for an introductory course…

…but they are the kinds of questions that data science is equipped to answer!

Our job as data scientists is

- to draw on tools from **statistics**, **computer science** and **mathematics**
- in collaboration with **domain experts**

to answer questions like these.

This hints at why it is so hard to become a truly good data scientist– you need to be pretty good at a lot of different things!

What is data science?

Is “data science” just a rebranding of “applied statistics”? Maybe, sort of.

Computers and the internet have made it easy to collect, share and analyze data

- These technologies are **very new**
- Fast, cheap computers have revolutionized what tools we can use!

Changes in technology have changed how we do science. In this sense, **“data science” is a culture or a way of thinking**, more than it is a field.

Wearing different hats

A “well rounded” data scientist uses multiple approaches to a problem.

Here are a few “hats” you may wear in your day-to-day work:

- **scientist:** understanding data domain, developing questions, “story telling”
- **software dev:** data processing and wrangling / reproducibility
- **mathematician:** linear algebra, probability theory, optimization
- **methodologist:** regression, unsupervised learning, visualizations
- **science communicator:** summarizing results, explaining broader impacts

If nothing else, going into this line of work requires that you be ready and willing to be a life-long student There are always new techniques, methods, frameworks and application domains to be learned!

Communicating your findings clearly!

The most important skill you will learn in this course?

It is **not** a statistical or computational tool (though those are important!)

It is the ability to clearly organize and explain your findings in a way that is appropriate for your intended audience.

That starts with understanding your tools and methods, but it doesn't end there!

The hardest part of this job is **clearly** explaining to your client what your conclusions are and how you arrived at them!

Course goals

- Use the R programming language to analyze data.
- Understand and apply basic concepts in probability; combine basic probability models to build more complicated ones; critique model assumptions.
- Formulate statistical hypotheses for different kinds of research questions and test those hypotheses using both classical and Monte Carlo methods.
- Understand and apply principles of statistical estimation and prediction, including fitting models and assessing model fit.
- Perform exploratory data analysis and create visualizations with `ggplot2`.
- Apply statistical tools to answer research questions using real-world data and present these findings clearly in both spoken and written form to non-experts.

Prerequisites

Required:

STAT240: Introduction to Data Modeling I
One or more of MATH 217, 221, or 275.

In short, you should have a broad familiarity with the R programming language and should be comfortable with basic concepts from calculus.

If you do not meet these prerequisites, I recommend that you wait until you have fulfilled them to take this course. You'll have a much better time, and you will appreciate the material more!

If you do not meet these prerequisites, but feel that you are ready for this course nonetheless, please speak to me **promptly**.

Course information

Instructor: Keith Levin

- Email: `kdlevin@wisc.edu`
- Office: 6170 MSC

Teaching Assistants:

- Alex Hayes
- Joseph Salzer
- Nursultan Azhimuratov
- Shane Huang

Refer to syllabus for office hours

Textbook: No physical textbook

- Lecture notes
- Weekly readings

Grading: ~10 HWs, 3 exams

- HW: 10%
- Exam 1: 25%
- Exam 2: 25%
- Final Exam: 40%
- Late days (see syllabus)

Course website: http://pages.stat.wisc.edu/~kdlevin/teaching/Fall2022/STAT340/

Tentative schedule on course website and Canvas; syllabus at http://pages.stat.wisc.edu/~kdlevin/teaching/Fall2022/STAT340/syllabus.pdf

Course information

Weekly discussion sections

- Attendance is optional but **highly encouraged**

If you run into trouble on homeworks, come to office hours for help

- But also please post to the **discussion board on Canvas**
- If you're having trouble, at least one of your classmates is, too
- You'll learn more by explaining things to each other than by reading stackexchange posts!

Email policy:

I **cannot** provide tech support over email– there are too many of you!

If you are having trouble, post to the discussion board and/or come to OHs!

Policies

Don’t plagiarize!

- You may discuss homeworks with your fellow students...
- ...but you must submit your own work
- Disclose in your homework whom (if anyone) you worked with

Late homeworks are not allowed!

- Instead, we have “late days”, of which you get 5
- One late day extends HW deadline by 24 hours

Refer to the syllabus for details.

Topics covered

We will cover five basic topics this semester:

1. Sampling
2. Estimation
3. Testing
4. Prediction
5. Observational/exploratory data analysis

Let’s briefly discuss each of these.

Sampling

A common polling technique for predicting election outcomes is “random digit dialing”, which is exactly what it sounds like.

A Marquette University poll reached 806 registered voters in Wisconsin:

- 48% of likely voters would choose Biden
- 43% would vote for Trump
- 2% for Jorgensen, and
- 7% remained undecided.

To reach those 806 participants, **many** more numbers needed to be dialed. The response rate was 4.3%. In fact, over 100,000 numbers had to be dialed to get these 806 respondents. The vast majority of those 100,000 calls were **never picked up**. Among those who did pick up, 806 were registered voters and agreed to participate in the survey, but another were 1113 refused to participate (or hung up).

Sampling

Reminder: >100,000 numbers dialed

- 806 picked up and participated; 1113 picked up an declined to participate
- **48%** for Biden; **43%** for Trump; 2% for Jorgensen; 7% undecided.

Actual election results in Wisconsin: Biden **49.45%**; Trump **48.82%**.

Questions:

- How does the outcome compare with the “predicted” vote shares?
- How might we explain the discrepancies?

Sampling

Reminder: >100,000 numbers dialed

- 806 picked up and participated; 1113 picked up an declined to participate
- **48%** for Biden; **43%** for Trump; 2% for Jorgensen; 7% undecided.

Actual election results in Wisconsin: Biden **49.45%**; Trump **48.82%**.

Questions:

- How does the outcome compare with the “predicted” vote shares?
- How might we explain the discrepancies?

Sampling has to do with how we find our “data points” for a study. Are they truly selected at random from the population that we want to measure? What do we mean by a “population”, anyway? What (if anything) can we do if our samples are **biased**, as in this election example?

Estimation

The news is perpetually full of stories about different economic indicators and how they are changing over time.

- The Consumer Price Index (CPI) is meant to measure the change over time in prices of consumer goods and services.
- Most surveys (e.g., public opinion or election surveys) are reported with a +/- 3% “confidence interval” or “sampling error”.

What is that all about?

Estimation: example

In that Wisconsin poll, the margin of error was reported to be +/- 4.3%.

The pollsters predicted **48%** of likely voters would vote Biden and **43%** for Trump; the actual outcome was **49.45%** for Biden and **48.82%** for Trump.

- Is this outcome within the stated margin of error?
- More generally, what does it mean to give a “confidence interval” or a “margin of error” for a quantity of interest like voter share or unemployment rate?

Testing

You have heard that “correlation does not imply causation”. Well, this is true.

- Ice cream consumption correlates with drownings, but we don’t think that eating ice cream causes people to drown.
- Hospitals are full of sick people, but we don’t think hospitals cause sickness.

Still, to paraphrase XKCD, correlation often does give awfully strong hints.

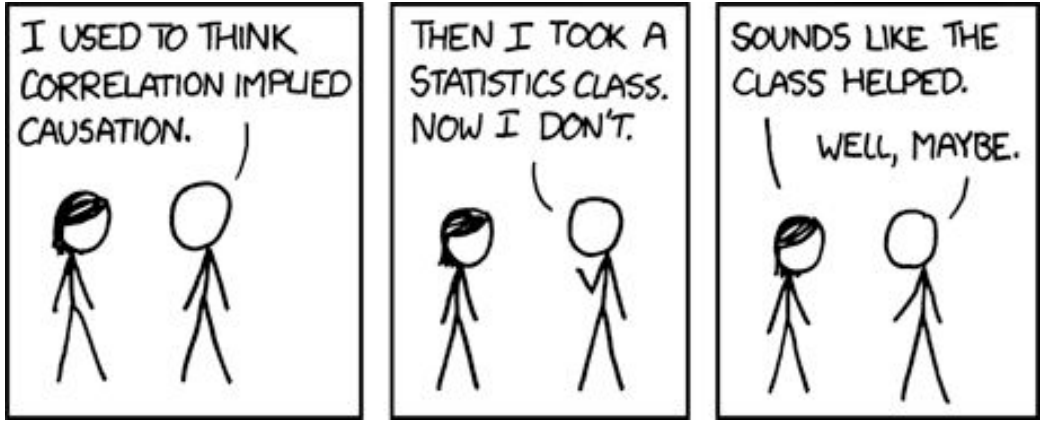


Image credit: Randall Munroe, XKCD (https://xkcd.com/552)

Testing: example

On November 16, 2020 Moderna released results from their phase 3 clinical trial for their COVID-19 vaccine.

There were approximately 30,000 people in the trial, split (approximately) evenly between treatment (got the vaccine) and control (got a placebo).

- In total, there were 95 cases of COVID-19 among the participants; 90 among the placebo group and 5 among the treated group.
- Of the 95 cases, 11 were severe cases, all in the placebo group.

Testing: example

On November 16, 2020 Moderna released results from their phase 3 clinical trial for their COVID-19 vaccine.

There were approximately 30,000 people in the trial, split (approximately) evenly between treatment (got the vaccine) and control (got a placebo).

- In total, there were 95 cases of COVID-19 among the participants; 90 among the placebo group and 5 among the treated group.
- Of the 95 cases, 11 were severe cases, all in the placebo group.

In this study, vaccination is correlated with reduced risk of infection.

- Does this mean vaccination **causally** reduced COVID-19 infection? Why or why not?
- How do we know that we aren't fooling ourselves when we say that the Moderna vaccine is effective? After all, these results **could** just be due to chance!

Prediction

Investing successfully (in real estate, stocks, etc) requires that we be able to predict the future behavior of an asset based on what we know about it currently.

- Based on the size of a house, its proximity to schools or infrastructure, walkability of its neighborhood, etc., we might hope to predict its “true” price.
- Many psychology and sociology studies aim to predict future student outcomes based on performance on a standardized test

In these kinds of problems, our goal is to predict an ***outcome*** or ***response*** (e.g., house price) based on one or more ***predictors*** (e.g., square footage).

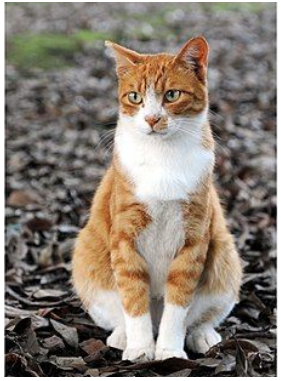
Prediction

Most of machine learning is concerned with prediction problems. For example, detecting whether an image contains a cat can be stated as a prediction problem.





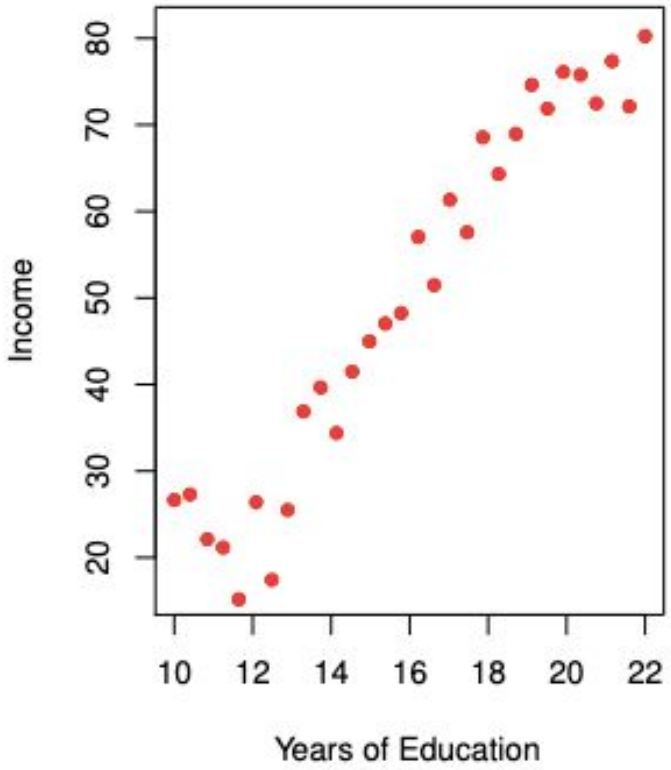





Prediction: example

This plot shows different people's incomes (tens of thousands of dollars per year) as a function of their years of education.

It certainly looks like more years of education correlate with higher income.



Question: Suppose I tell you that someone has 18 years of education. What would you predict their income to be?

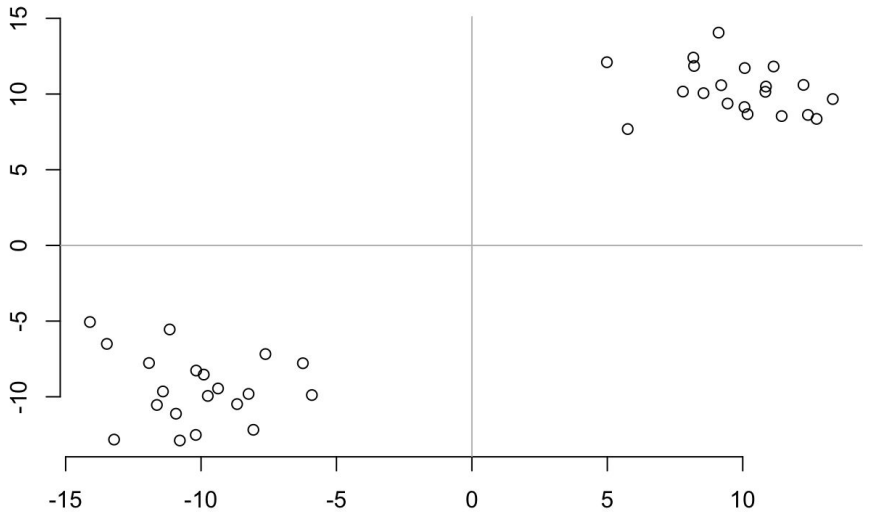
Image credit: ISLR2 https://rdrr.io/cran/ISLR2/

Observatory/exploratory analysis

Suppose that a client gives you a data set that looks like this:

What would you do?

There is clearly some kind of a **cluster structure** present here.



The goal of **exploratory data analysis** is to identify interesting structures in our data that might warrant further study.

Observatory/exploratory analysis: example

In my own research, I collaborate a lot with neuroscientists, who are interested in identifying **functional subnetworks** of the brain

- groups of neurons that work together
- associated with the same activity (e.g., attention, motion, speech).

This is an example of **clustering**, in which our goal is to group data points in a sensible way, without necessarily saying ahead of time what those groups mean

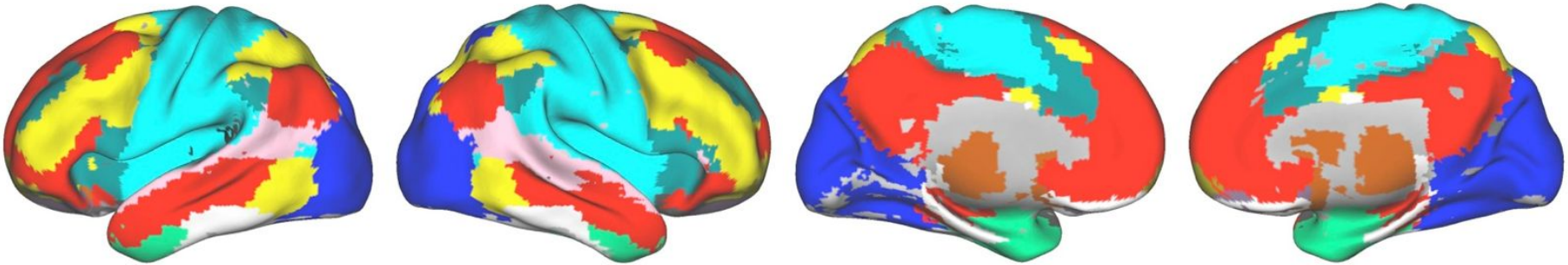


Image credit: Power, et al., 2011, “Functional network organization of the human brain”

Observatory/exploratory analysis: example

Oftentimes, we obtain **observational** data. That is, data that does not come from a carefully-designed experiment.

Lots of “big data” is collected by scraping the web or in other “messy” ways

- Scraping data from a social media site such as Twitter.
- Measuring the socioeconomic status of people in different zip codes

There are lots of interesting scientific questions we might like to ask about such a data set. Unfortunately, because this data isn’t the result of a carefully controlled experiment, we are often much more limited in what we can say.

World, Data, Model (hat tip to Karl Rohe)

To figure things out about the world, we take measurements.

That is, we collect data. These data describe the world…

…but it remains to build a **model** that explains how these data came to be.

The process of **inference** is how we use a specific set of data to guide our beliefs about the world.

This is an iterative process. We collect data, use it to guide our beliefs, and then let those beliefs guide our choices for future data collection.

Example: What is the average human height?

This is a question about the world.

We could answer it exactly: go out and measure the height of every adult human!

- This would be complicated and expensive
- **Alternative:** just measure heights of a few adult humans

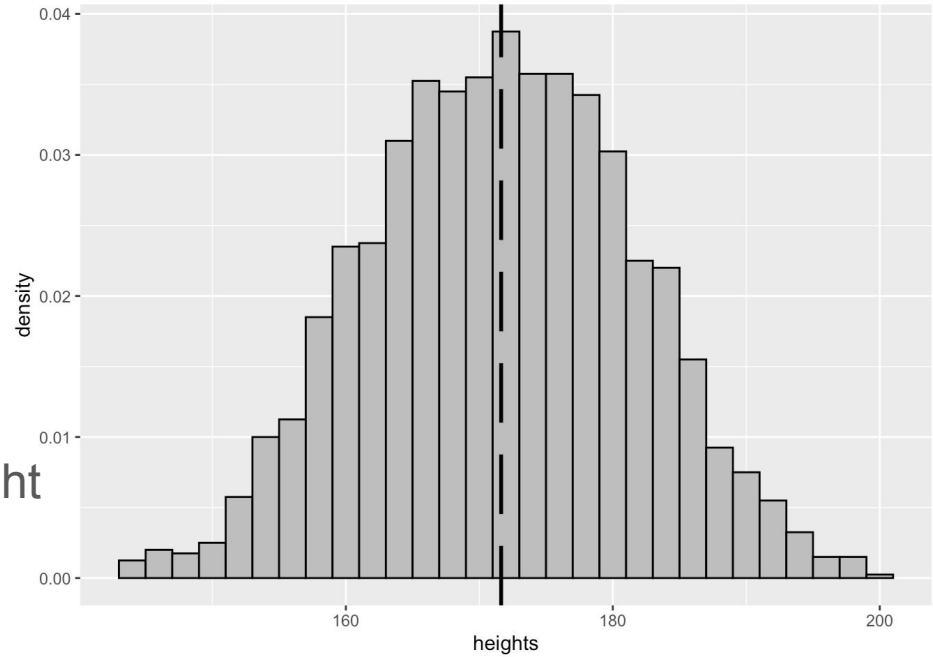
Of course, this small collection of humans would need to be chosen randomly and in such a way that they would be representative of the population as a whole, but let’s ignore that concern until later in the course…

What is the average human height?

Say we measure just a few thousand adult human heights.

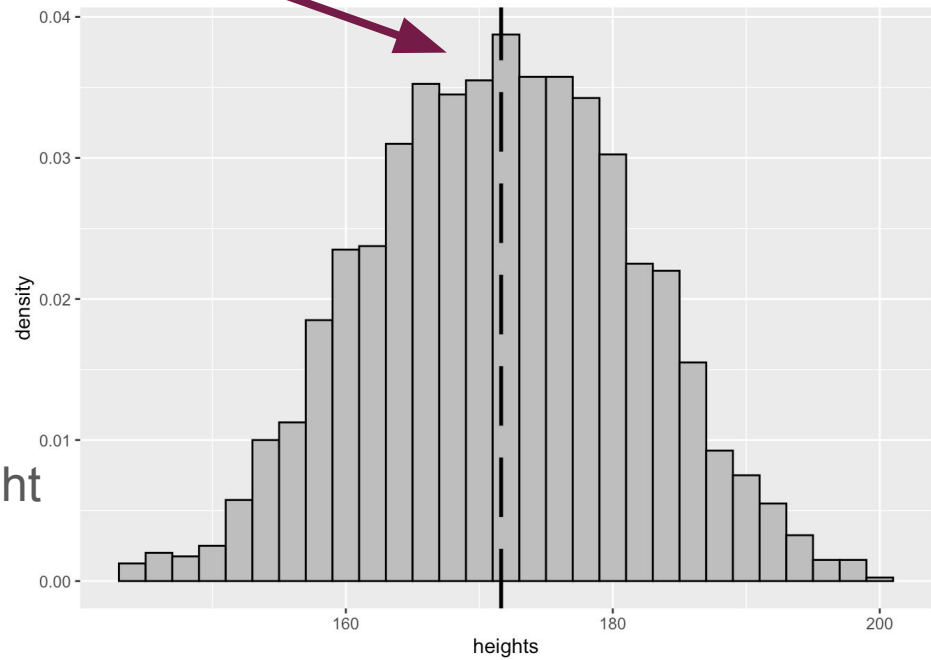
Those heights we measure would constitute our **data**– measurements taken out there in the world.

Here is a simulation of what that data might look like with a sample of size 2000.



What is the average human height?

Vertical dashed line indicates **mean** of our 2000 samples heights

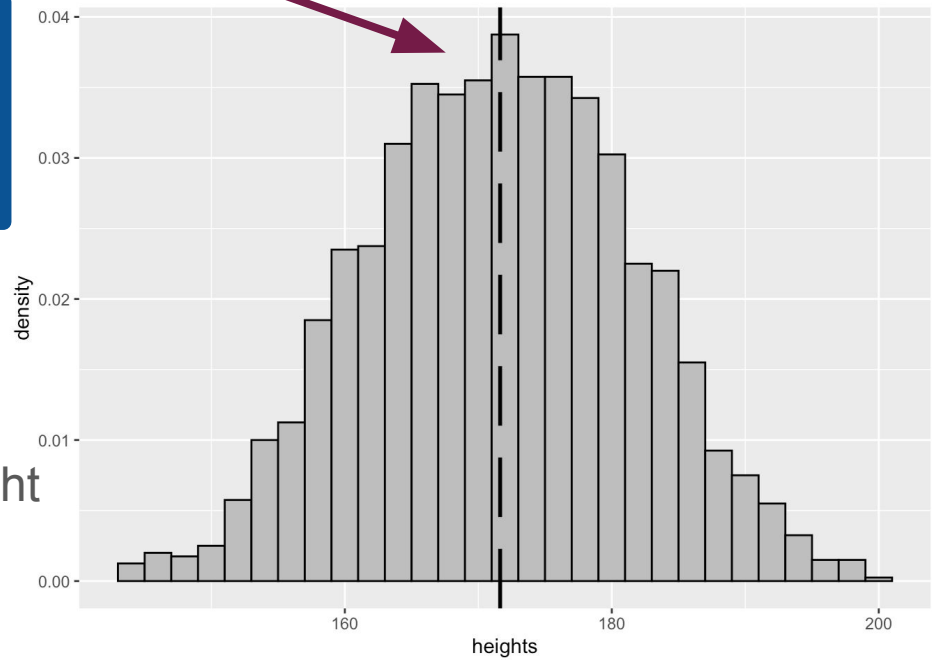


Here is a simulation of what that data might look like with a sample of size 2000.

What is the average human height?

Vertical dashed line indicates **mean** of our 2000 samples heights

Because this is a **random sample** (and heights vary due to, e.g., nutrition and genetics), the **sample mean** need not be equal to the population mean (i.e., the true average adult human height).



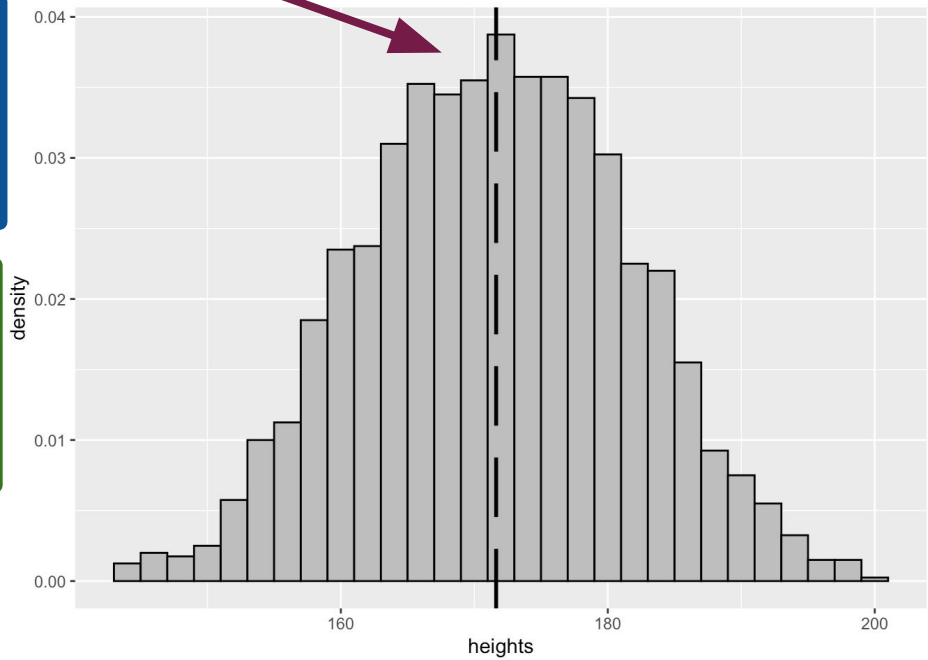
Here is a simulation of what that data might look like with a sample of size 2000.

What is the average human height?

Vertical dashed line indicates **mean** of our 2000 samples heights

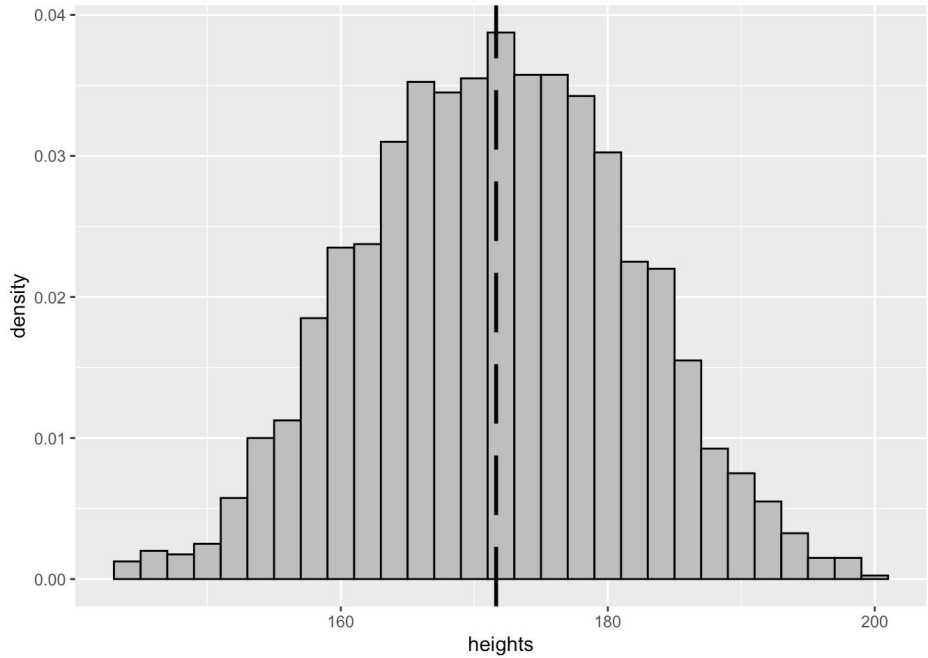
Because this is a **random sample** (and heights vary due to, e.g., nutrition and genetics), the **sample mean** need not be equal to the population mean (i.e., the true average adult human height).

Instead, the heights in our sample (and their mean) will vary randomly about the population average height. We use a **statistical model** (i.e., probability theory) to describe this variation.



What is the average human height?

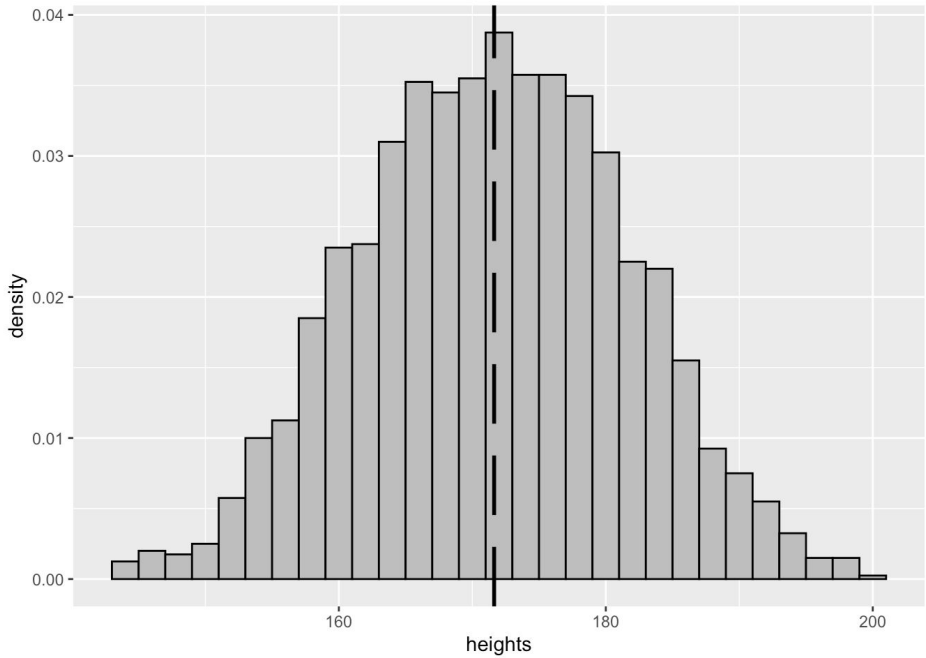
Common choice for modeling random variation like this is the **normal distribution**



What is the average human height?

Common choice for modeling random variation like this is the **normal distribution**

We **assume** our data are generated by a normal distribution with mean μ (i.e., the average height) and standard deviation σ. These are the **parameters** of the model.

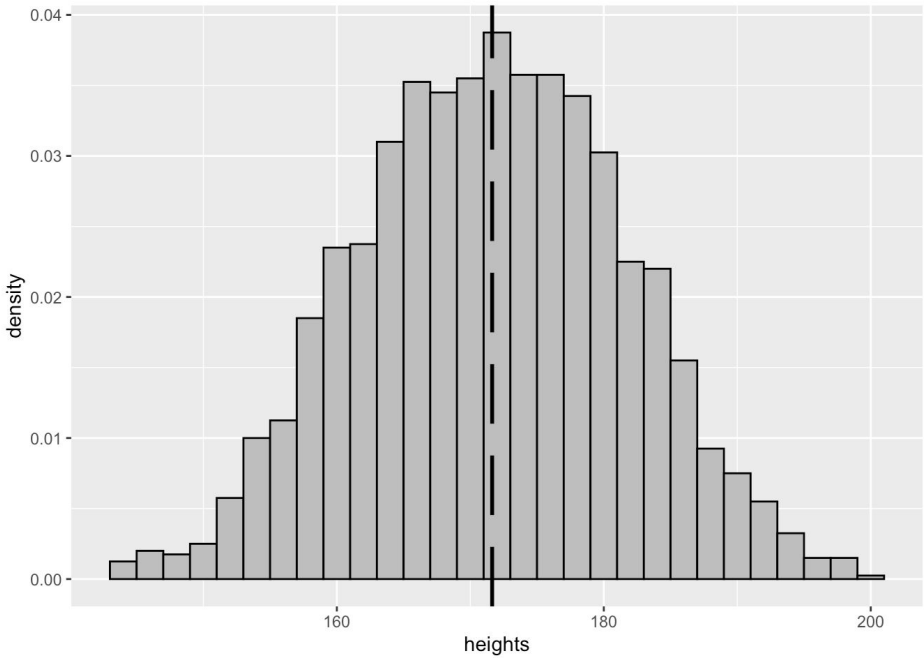


What is the average human height?

Common choice for modeling random variation like this is the **normal distribution**

We **assume** our data are generated by a normal distribution with mean μ (i.e., the average height) and standard deviation σ. These are the **parameters** of the model.

Estimating the population average height reduces to estimating the “true” value of μ from the data.



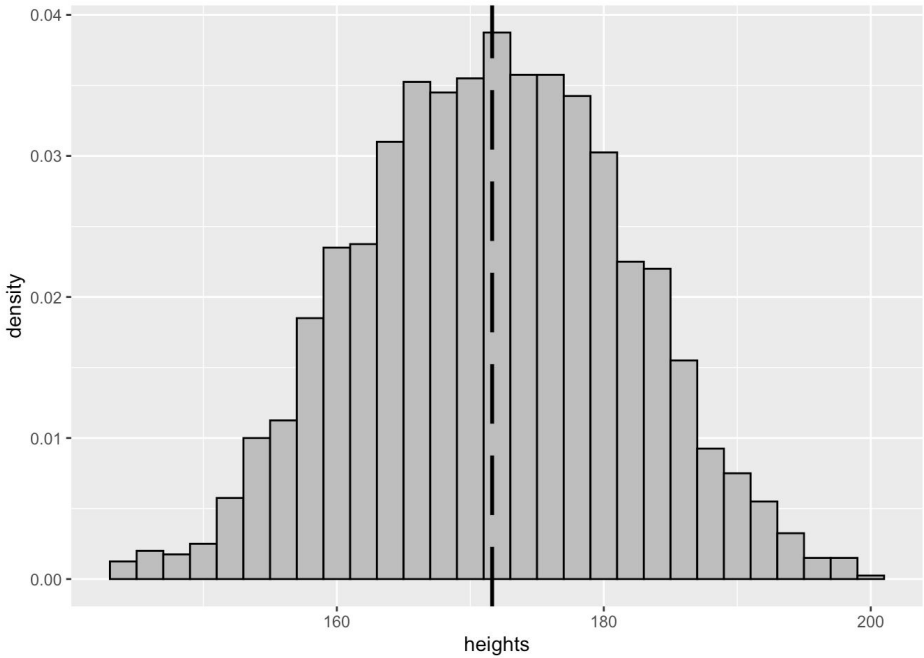
What is the average human height?

Common choice for modeling random variation like this is the **normal distribution**

We **assume** our data are generated by a normal distribution with mean μ (i.e., the average height) and standard deviation σ. These are the **parameters** of the model.

Estimating the population average height reduces to estimating the "true" value of μ from the data.

This step of using data to determine something about our model is called **inference**. Here, our goal is to estimate the value of μ, which will in turn be our estimate of the average adult human height.



All models are wrong, but some are useful*

An important point: we **assumed** that heights are normally distributed.

In practice, modeling assumptions like these are never strictly true.

Our **model** is just that– a model of the world; a set of **simplifying assumptions** that we hope are at least a good approximation to the truth.

*https://en.wikipedia.org/wiki/All_models_are_wrong

All models are wrong, but some are useful*

An important point: we **assumed** that heights are normally distributed.

In practice, modeling assumptions like these are never strictly true.

Our **model** is just that– a model of the world; a set of **simplifying assumptions** that we hope are at least a good approximation to the truth.

Example: Think of your physics courses, where we assume that things happen in a frictionless void and use Newtonian mechanics instead of quantum mechanics.

*https://en.wikipedia.org/wiki/All_models_are_wrong

All models are wrong, but some are useful*

An important point: we **assumed** that heights are normally distributed.

In practice, modeling assumptions like these are never strictly true.

Our **model** is just that– a model of the world; a set of **simplifying assumptions** that we hope are at least a good approximation to the truth.

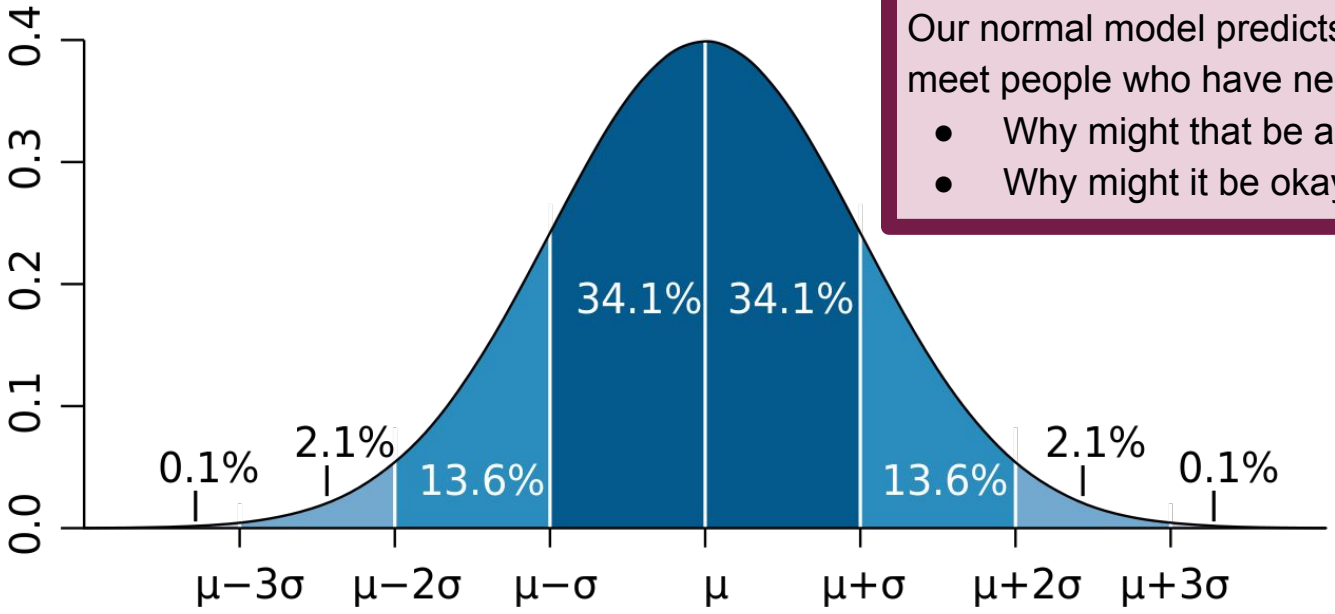
Example: Think of your physics courses, where we assume that things happen in a frictionless void and use Newtonian mechanics instead of quantum mechanics.

We make assumptions like these because they often make the math easier while still being a good approximation to reality.

*https://en.wikipedia.org/wiki/All_models_are_wrong

All models are wrong, but some are useful

In our human heights example, the normal distribution with mean μ and standard deviation σ says that with some (perhaps very small) probability, we **might** observe a negative number.



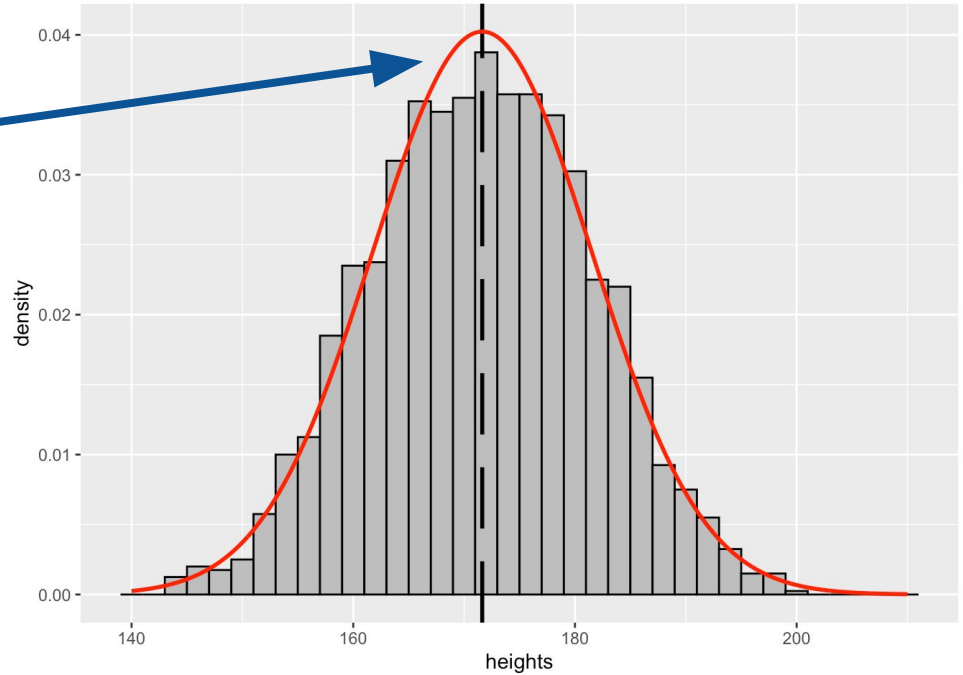
Our normal model predicts that we should, on occasion, meet people who have negative height.

- Why might that be a problem?
- Why might it be okay to use this model anyway?

Let’s push ahead and “fit” a normal to our data

We’ll have lots to say about this later. For now, think of this as choosing, out of all the possible normal distributions, the one that “best agrees”" with our data.

The red curve is the density of the normal distribution that best fits* our observed data.

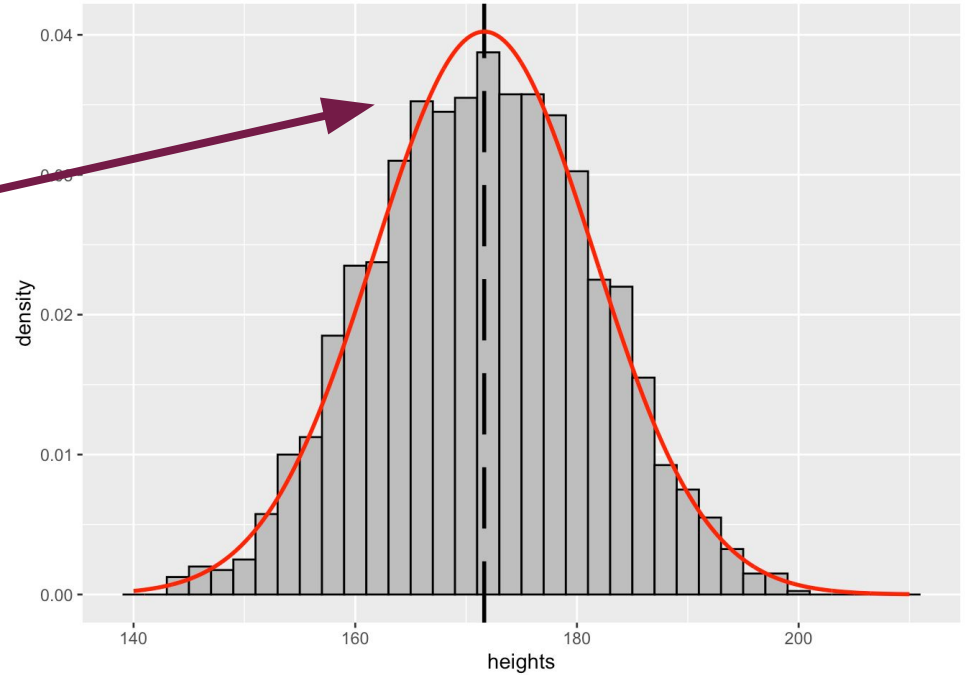


*What do we mean by “best”? We’ll have lots to say about this later in the course.

Let’s push ahead and “fit” a normal to our data

We’ll have lots to say about this later. For now, think of this as choosing, out of all the possible normal distributions, the one that “best agrees”" with our data.

Hmm… our data looks a bit “flatter” than the normal distribution predicts that it should…

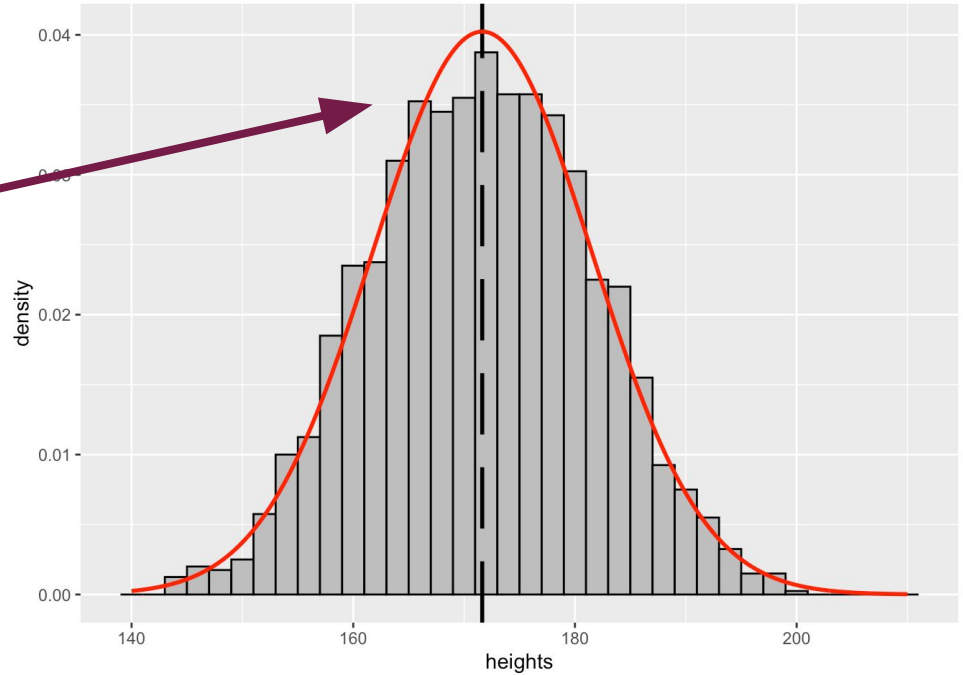


Let’s push ahead and “fit” a normal to our data

We’ll have lots to say about this later. For now, think of this as choosing, out of all the possible normal distributions, the one that “best agrees”" with our data.

Hmm… our data looks a bit “flatter” than the normal distribution predicts that it should…

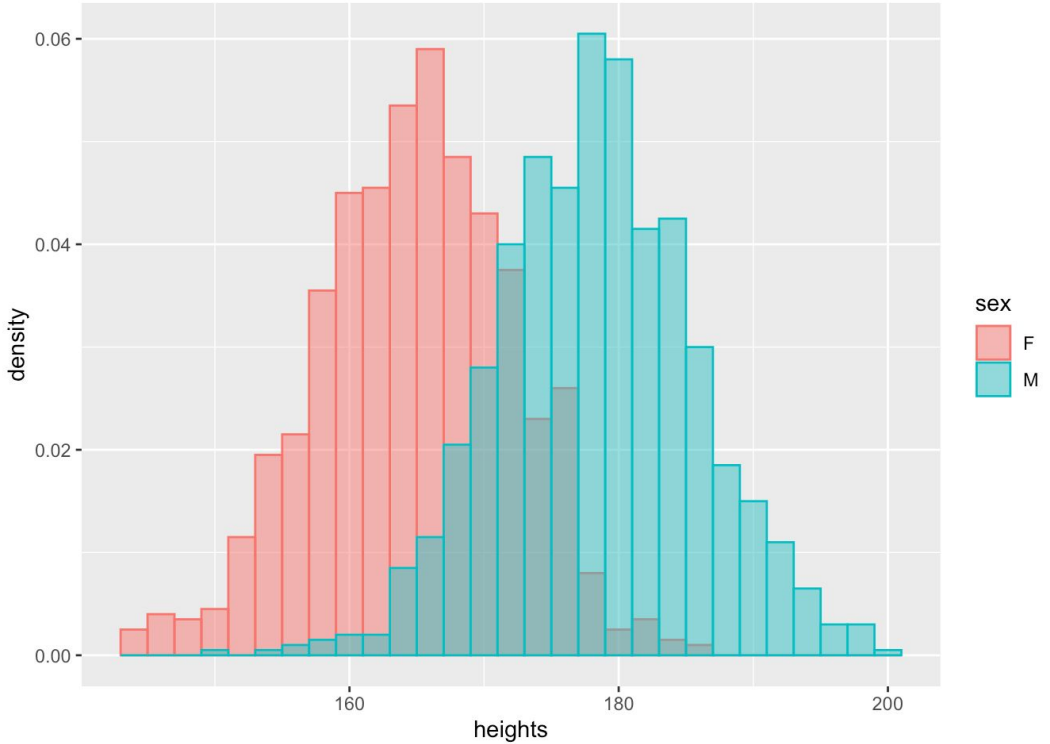
Perhaps this is just due to random fluctuation, but in fact there is a very simple reason for this: I didn’t tell you about it, but this sample includes **both males and females**.



Refining our model

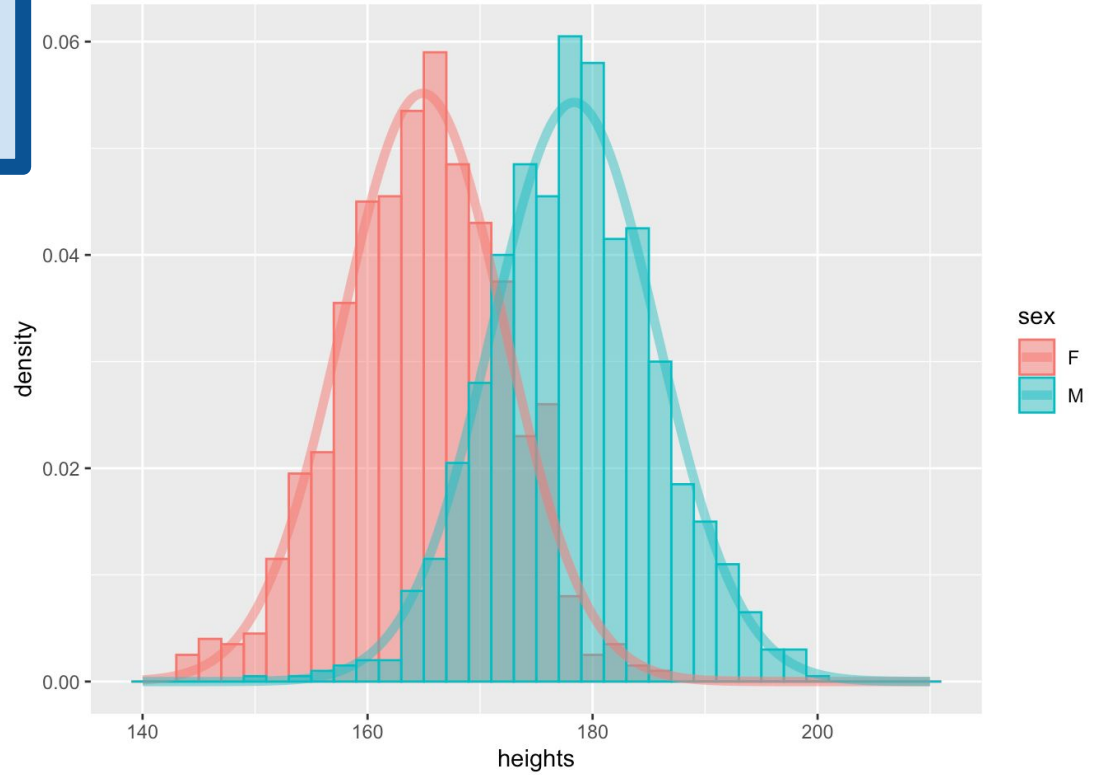
Let's plot the same histogram, but now let's break out heights according to sex.

Human heights are bimodal– female heights are approximately normal about some mean, and male heights are approximately normal about another.



Refining our model

We can fit a normal to the male and female subpopulations separately and we see that our model agrees with the data much better.



Refining our model

This is a good (albeit simple) illustration of the kind of iterative workflow that we typically use in data science

- We obtain our data, fit a model to it, and we examine the shortcomings of that model.
- After some thought, it becomes clear how to improve our model.
- We implement those changes (in this case, we incorporated the variable sex into our model), and examine our findings again.

Typically, we repeat this cycle several times before reaching a conclusion that we are reasonably confident in.

Refining our model

This is a good (albeit simple) illustration of the kind of iterative workflow that we typically use in data science

- We obtain our data, fit a model to it, and we examine the shortcomings of that model.
- After some thought, it becomes clear how to improve our model.
- We implement those changes (in this case, we incorporated the variable sex into our model), and examine our findings again.

Typically, we repeat this cycle several times before reaching a conclusion that we are reasonably confident in.

By the end of this semester, you will be able to perform an analysis just like this (and indeed, do much more complicated and interesting things, too!).