Statistics for Data Science

UC Berkeley, School of Information

2022-01-03

Contents

Live Session Introduction

This is the live session work space for the course. Our goal with this repository, is that we're able to communicate *ahead of time* our aims for each week, and that you can prepare accordingly.

Bloom's Taxonomy

An effective rubric for student understanding is attributed to Bloom (1956). Referred to as *Bloom's Taxonomy*, this proposes that there is a hierarchy of student understanding; that a student may have one *level* of reasoning skill with a concept, but not another. The taxonomy proposes to be ordered: some levels of reasoning build upon other levels of reasoning.

In the learning objective that we present in for each live session, we will also identify the level of reasoning that we hope students will achieve at the conclusion of the live session.

1. **Remember** A student can remember that the concept exists. This might require the student to define, duplicate, or memorize a set of concepts or facts.
2. **Understand** A student can understand the concept, and can produce a working technical and non-technical statement of the concept. The student can explain why the concept *is*, or why the concept works in the way that it does.
3. **Apply** A student can use the concept as it is intended to be used against a novel problem.
4. **Analyze** A student can assess whether the concept has worked as it should have. This requires both an understanding of the intended goal, an application against a novel problem, and then the ability to introspect or reflect on whether the result is as it should be.
5. **Evaluate** A student can analyze multiple approaches, and from this analysis evaluate whether one or another approach has better succeeded at achieving its goals.

6. **Create** A student can create a new or novel method from axioms or experience, and can evaluate the performance of this new method against existing approaches or methods.

Chapter 1

Probability Spaces

Probability is a system of reasoning about the world in the face of incomplete information. In this course, we're going to develop an understanding of the implications of core parts of this theory, how this theory was developed, and how these implications relate to every other part of the practice of data science.

1.1 Learning Objectives

At the end of this week's learning, student will be able to:

1. **Find** and *access* all of the course materials
2. **Develop** a course of study that is builds toward success
3. **Apply** the axioms of probability to make a valid statement
4. **Solve** word problems through the *application* of probability and math rules

1.2 Course Learning Objectives

At the end of this course, students will be able to:

1. **Understand the building blocks of probability theory that prepare learners for the study of statistical models.**
    1. Understand the mathematical objects of probability theory and be able to apply their properties.
    2. Understand how high-level concepts from calculus and linear algebra are related to common procedures in data science.

3. Translate between problems that are defined in business or research terms into problems that can be solved with math.

2. **Understand and apply statistical models in common situations.**
    1. Understand the theory of statistics to prepare students for inferrential statements.
    2. Understand model parameters and high level strategies to estimate them: means, least squares, and maximum likelihood.
    3. Choose an appropriate statistic, and conduct a hypothesis test in the Neyman-Pearson framework.
    4. Interpret the results of a statistical test, including statistical significance and practical significance.
    5. Recognize limitations of the Neyman-Pearson hypothesis testing framework and be a conscientious participant in the scientific process
3. **Analyze a research question using a linear regression framework.**
    1. Explore and wrangle data with the intention of understanding the information and relationships that are (and are not) present
    2. Identify the goals of your analysis
    3. Build a model that achieves the goals of an analysis
4. **Interpret the results of a model and communicate them in manner appropriate to the audience.**
    1. Identify their audience and report process and findings in a manner appropriate to that audience.
    2. Construct regression oriented reports that provide insight for stakeholders.
    3. Construct technical documents of process and code for collaboration and reproducability with peer data scientists.
    4. Read, understand, and assess the claims that are made in technical, regression oriented reports
5. **Contribute proficient, basic work, using industry standard tools and coding practices to a modern data science team.** Demonstrate programming proficiency by translating statistical problems into code.
    1. Understand and incorporate best practices for coding style and data carpentry
    2. Utilize industry standard tooling for collaboration

1.3 Introductions

1.3.1 Instructor Introductions

The instructors for the course come to the program, and to statistics from different backgrounds. Instructors hold PhDs in statistics, astrophysics, biology,

political science, information.

1.3.2 What does a statistician look like?

A statistician looks like YOU!

Identity shapes how people approach and understand their world. We would like to acknowledge that we have limited diversity of identity among the instructors for this course.

However, every one of the instructors shares a core identity as an empathetic educator that wants to understand your strengths, areas for growth, and unique point of view that is shaped by who you are.

- It doesn't matter if you've never taken a stats class before, or if you're reviewing using this class. There will be challenges for everyone to overcome.
- It doesn't matter how old or young you are. We will all be learning frequentist statistics which is timeless.
- The color of your skin doesn't matter; nor does whether you identify as a woman or a man or trans or non-binary; neither does your sexual orientation. There are legacies of exclusion and discrimination against people due to these identities. We will not continue to propagate those legacies and instead will work to controvert those discriminations to build a diverse community of learning in line with the University's Principles of Community.

1.4 Student Introductions

Please take 90 seconds to tell us:

- Your name as you would like to be called;
- Where you dial in from;
- What brings you to an interest in data science; and,
- Any other interests, or identities that you would like your classmates and instructor to know about.

Please, try to keep these intros to just 90 seconds. We've a lot to cover this week!

1.5 Probability Theory

Probability

Probability is a system of reasoning that we use to model the world under incomplete information. This model underlies virtually *every* other model you'll ever use as a data scientist.

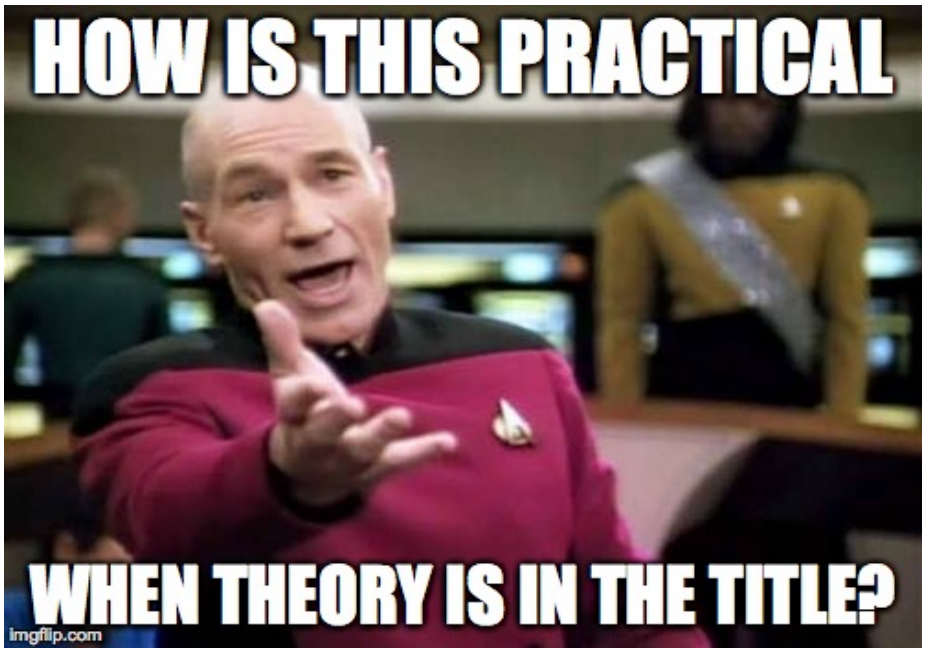


Figure 1.1: picard

In this course, probability theory builds out to random variables; when combined with sampling theory we are able to develop p-values (which are also random variables) and an inferential paradigm to communicate what we know and how certain a statement we can make about it.

In introduction to machine learning, literally the first model that you will train is a naive bayes classifier, which is an application of Bayes' Theorem, trained using an iterative fitting algorithm. Later in machine learning, you'll be fitting non-linear models, but at every point the input data that you are supplying to your models are generated from samples from random variables. That the world can be represented by random variables (which we will cover in the coming weeks) means that you can transform – squeeze and smush, or stretch and pull – variables to heighten different aspects of the variables to produce the most useful *information* from your data.

As you move into NLP, you might think of generative text as a conditional probability problem: given some particular set of words as an input, what is the most likely *next* word or words that someone might type?

Beyond the direct instrumental value that we see working with probability, there

are two additional aims that we have in starting the course in the manner.

First, because we are starting with the axioms of probability as they apply to data science statistics, students in this course develop a *much* fuller understanding of classical statistics than students in most other programs. Unfortunately, it is very common for students and then professionals to see statistics as a series of rules that have to be followed absolutely and without deviation. In this view of statistics, there are distributions to memorize; there are repeated problems to solve that require the rote application of some algebraic rule (i.e. compute the sample average and standard deviation of some vector); and, there are myriad, byzantine statistical tests to memorize and apply. In this view of statistics, if the real-world problem that comes to you as a data scientist doesn't clearly fit into a box, there's no way to move forward.

> Statistics like this is not fun.

In the way that we are approaching this course, we hope that you're able to learn *why* certain distributions (like the normal distribution) arise repeatedly, and why we can use them. We also hope that because you know how sampling theory and random variables combine, that you can be more creative and inventive to solve problems that you haven't seen before.

The second additional aim that we have for this course is that it can serve as either an introduction or a re-introduction to reading and making arguments using the language of math. For some, this will be a new language; for others, it may have been some years since they have worked with the language; for some, this will feel quite familiar. New algorithms and data science model advancements *nearly always* developed in the math first, and then applied into algorithms second. In our view, being a literate reader of graduate- and professional-level math is a necessary skill for any data scientist that is going to keep astride of the field as it continues to develop and these first weeks of the course are designed to bring everyone back into reading and reasoning in the language.

1.6 Working with a Sample Space

1.6.1 Working with a Sample Space, Part I

1. **You roll two six-sided dice**:
    1. How would you define an appropriate sample space, Ω?
    2. How many elements exist in Ω?
    3. What is an appropriate event space, and how many elements does it have?
    4. Give an example of an event.

1.6.2 Working with a Sample Space, Part II

2. **For a random sample of 1,000 Berkeley students**:
    1. How would you define an appropriate sample space, Ω?
    2. How big is Ω? How many elements does it contain?
    3. What is an example of an event for this scenario?
    4. Can a single person be represented in the space twice? Why or why not?

1.6.3 Working with a Sample Sapce, Part III

3. **Suppose that you're sitting in a surf lineup, and you have to pick a wave that is the right height. Too small, and you won't get anywhere, too large and you'll get crushed.**
    1. What sample space is appropriate to represent the height of a single wave, Ω?
    2. How big is Ω? How many elements does it contain?
    3. What is an example of an event that could be part of the event space?
    4. What sample space is appropriate to represent the height of the next 10 waves? How large is this sample space?

To represent 10 waves, you should use $\mathbb{R}^{10}$. It is curious mathematical fact that $\mathbb{R}$ and $\mathbb{R}^{10}$ actually have the same cardinality – there are the same number of elements in each of these sets and there exists a function – i.e. a one-to-one mapping – between these sets.

1.7 Proofs: Style Counts

In each week of a class, you are either caught up or behind.

- The probability that you are caught up in Week 1 is 0.7.
- If you are caught up in a given week, the probability that you will be caught up in the next week is 0.7.
- If you are behind in a given week, the probability that you will be caught up in the next week is 0.4.
- **What is the probability that you are caught up in week 3?**

Identify as many ways to improve this proof as you can:

If you are caught up in a week, there are two possibilities for the previous week: caught up and behind. Let $P(C)$ be the probability of being caught up. In week

1, $P(C) = .7$. The probability of being behind is $P(B) = 1 - .7 = .3$. We first break down the probability for week 2:

$$P(C) = .7 \cdot .7 + .3 \cdot .4 = .65$$

Now we can repeat the process for week 3:

$$P(C) = .65 * .7 + .35 * .4 = .595$$

1.8 A practice problem

Let's go and work on a practice problem over on the course practice problem website.

link here

1.9 Concluding Reminders

1. Welcome!
2. Before next live session:
    1. Complete the homework that builds on this unit
    2. Complete all videos and reading for unit 2

Chapter 2

Defining Random Variables

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.6     v dplyr   1.0.7
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.1.1     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

2.1 Learning Objectives

1.
2.
3.

2.2 Class Announcements

2.3 Roadmap

Rearview Mirror

Today

Looking Ahead

2.4 Random Variables

- Random variables are objects that we use to hold numerical representations of real-world phenomena
- We can use a probability model to model how frequently events in the random variable will occur
- From the axioms of probability, we can build a whole, expressive modeling system (that need not be grounded **at all** in the minutia of the world) that is useful for making predictions
- We assert that there is a probability density (or distribution) function, but we never get to see it.

2.5 Random Variables: Questions

2.5.1 Randomness

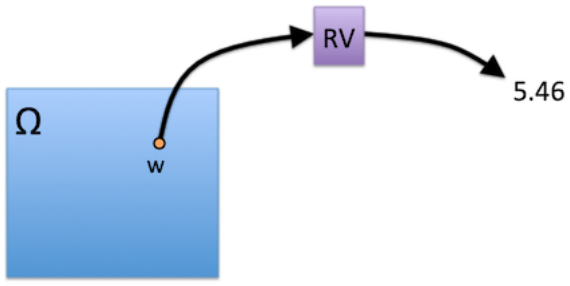


Figure 2.1: sample space

1. Where does a RV's randomness come from?
2. If you select a person at random from the US population and measure their blood pressure, what does Ω represent? What does w represent? What is the random variable?
3. Why do we need a random variable to represent blood pressure? Why can't we make predictions about blood pressure without random variables?

2.5.2 Car crashes and Random Variables

Suppose that you stand at an intersection for six hours one day and report the time of each crash you observe during the observation period.

- What sample space, Ω, is appropriate to represent this scenario?

- Is Ω a finite or an infinite set? Is it countable or uncountable?
- If the police department happens to care only about the number of crashes during the six-hour observation period, what random variable X would represent this summary?
- Can you describe the range of your X? Is it finite or infinite?
- Is your X a discrete or a continuous random variable?

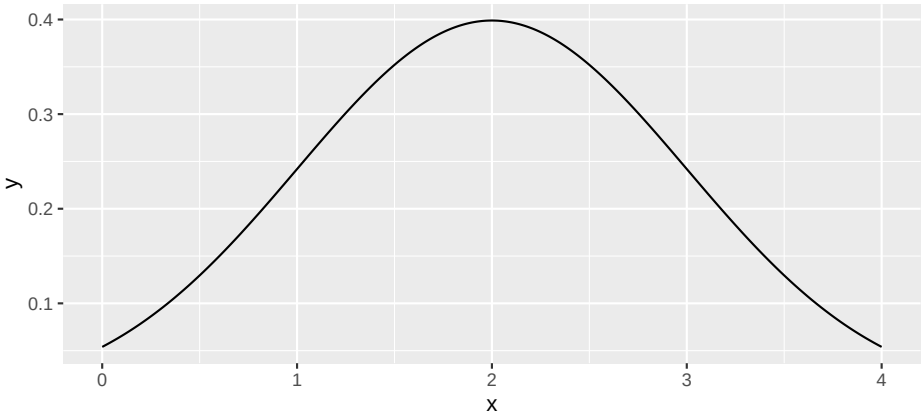
2.5.3 Discrete & Continuous Random Variables

- Suppose X is a random variable that describes the time a student spends on w203 homework 1.
  - If you have only granular measurement – i.e. the number of nights spent working on the homework – is this discrete or continuous?
  - If you have the number of hours, is it discrete or continuous?
  - If you have the number of seconds? Or milliseconds?
- Is it possible that $P(X = a) = 0$ for every point a? For example, that $P(X = 3600) = 0$.

2.6 Visualizing Distributions Via Simulation

2.6.1 The Visualization Trick

Here is the true density function for a normal random variable.



Simulate Draws

There's another way to get an **approximate** idea of what the distribution looks like. Here's how we take a single draw from a normal distribution with a specific set of features:

```
rnorm(n = 1, mean = 2, sd = 1)
```

```
## [1] 1.819932
```

Repeating the Experiment

We want to rerun that experiment 10 times. We take a draw, then rewind time, clear our memory and start over with fresh randomness. To do this in R, an easy way is with the `replicate()` function. Change the code below so that it repeats the experiment above 10 times, then use `hist()` to display a plot of the result.

```
simulation <- replicate(
  n    = 10,      # should you change this line?
  expr = 1 + .2   # or this line?
  )

simulation
```

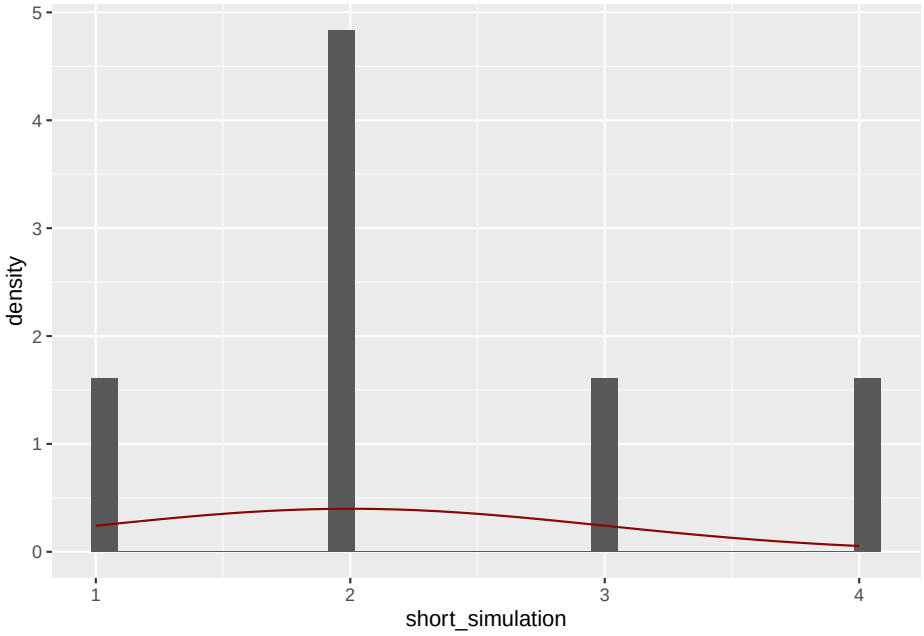
```
##  [1] 1.2 1.2 1.2 1.2 1.2 1.2 1.2 1.2 1.2 1.2
```

Better Visualization

Here's some fancy ggplot code to draw a nice histogram of the result, along with the true density. Remove the first line to make it work with your simulation.

```
short_simulation <-  c(1,2,3,2,4,2)

true_density <- function(x) {
  dnorm(x = x, mean = 2, sd = 1)
  }

dat_hist <- data.frame(short_simulation)

dat_hist %>%
  ggplot() +
  geom_histogram(
    aes(x = short_simulation, y = ..density..)) +
  stat_function(
    aes(x = short_simulation), fun = true_density, color = 'darkred')
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Repeating the Experiment: Questions

- What happens to your plot as you increase the number of draws from 10 to 100 to 1000. . . ?
- In your own words, what is the difference between the distribution and the sample you are taking?

The Visualization Trick

- This is a pretty useful *trick*.

- The repetition we're using has no analogue in the real world – we don't get to shake up the world like a snow globe a number of times in a row to see what it does.
- But, when we say "take a draw from the distribution" another way to say this is that we're *simulating* the random variable.

2.6.2 Apply the Visualization Trick

Part I

- How can the visualization trick help us? Here's a problem:

  - Suppose X and Y are independent normal random variables, both with mean 2 and standard deviation 1. Say $Z = X + Y$.

  - What is the distribution of Z?

- We could do some math to compute the density function of Z, but it's actually quite messy. Instead, let's use the visualization trick to get an approximate idea.

Part II

- First, write an R function to simulate a single draw from Z.

1. Simulate a draw for X.
2. Simulate a draw Y.
3. Return the sum of the previous draws.

```
rz <- function() {
  2 # Replace with your code
}
```

- Use the previous code to repeat this experiment 10,000 times and plot a histogram.
- See if you can guess what the the distribution is, and plot your guess on the histogram.

2.6.3 How is simulation useful?

- Are there situations that you think simulation of this sort might be useful?

2.7 Computing Different Distributions.

Suppose that random variables X and Y are jointly continuous, with joint density function given by,

$$f(x,y) = \begin{cases} c, & 0 \le x \le 1, 0 \le y \le x \\ 0, & otherwise \end{cases}$$

where c is a constant.

1. Draw a graph showing the region of the X-Y plane with positive probability density.

2. What is the constant c?
3. Compute the marginal density function for X. (Be sure to write a complete expression)
4. Compute the conditional density function for Y, conditional on $X = x$. (Be sure to specify for what values of x this is defined)

2.8 Understanding Joint Distributions

- In this picture, we imagine putting a cake down on the X-Y plane.
- Take a sharp knife and make two cuts parallel to the X-axis. one is at $Y = y$, the other at $Y = y + dy$.

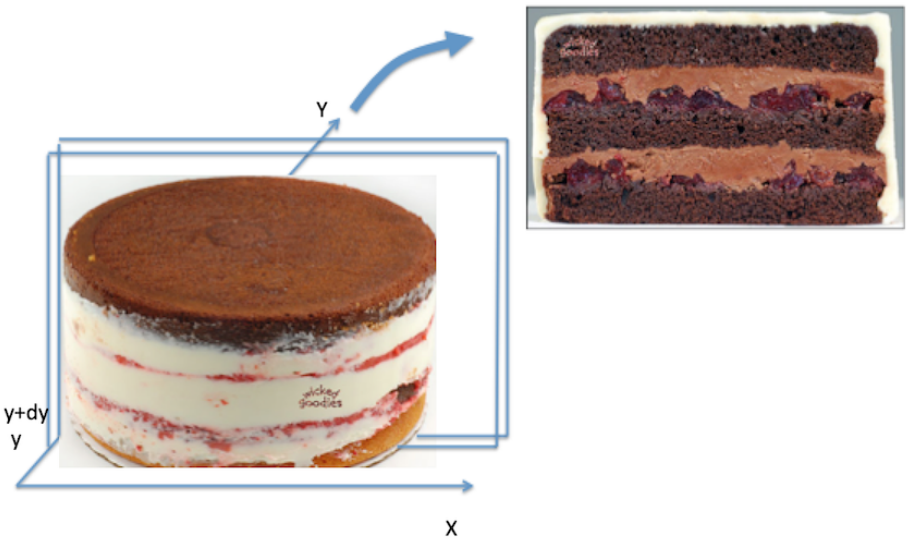


Figure 2.2: cake

Review of Terms

Remember some of the key terms we learned in the async:

- Joint Density Function
- Conditional Distribution
- Marginal Distribution

Explain each of these three in terms of the cake metaphor.

2.9 Working with a Shiny App

```
knitr::include_app(url = 'http://statistics.wtf/betahat/', height = '1200px')
```

Chapter 3

Summarizing Distributions

3.1 Learning Objectives

1.
2.
3.

3.2 Class Announcements

3.3 Roadmap

Roadmap – Rearview mirror

- Statisticians create a population model to represent the world.
- Random variables are the building blocks of a model.
- We can describe the distribution of a random variable using:
  - a cdf (all random varibles)
  - a pmf (discrete random variables)
  - a pdf (continuous random variables)
- When we have multiple random variables,
  - a joint pmf / pdf describes how they behave together
  - a marginal pmf / pdf describes one variable in isolation
  - a conditonal pmf / pdf describes one variable given the value of another

Roadmap – Today's Lesson

- A joint distribution has more information than we can ever use (or estimate with finite data).
- To do useful things, we need to summarize certain features we care about:
  - Expectation
  - Variance
  - Covariance
  - Correlation

Roadmap – Coming Attractions

- A predictor is a function that provides a value for one variable, given values of some others.
- Using our summary tools, we will define a predictor's error and then minimize it.
- This is a basis for linear regression

3.4 Proof Strategy Workshop: Expectation

Let Y be the random variable given by X^2 where X is a uniform distribution on the support $[0, 1]$.

What is the expected value of Y?

- Your instructor will assign you to Proof 1, Proof 2, OR Proof 3.
- First, study your proof.
- Second, in breakout rooms, teach your proof to your classmates.

Proof 1

We first compute the CDF of Y. For $0 \leq y \leq 1$,

$$F_Y(y) = P(Y \leq y) = P(X^2 \leq y) = P(X \leq \sqrt{y}) = \sqrt{y}$$

Next, we take the derivative to get Y's density for $0 \leq y \leq 1$:

$$f_Y(y) = \frac{\partial}{\partial y} F_Y(y) = \frac{1}{2} y^{-1/2}$$

(On your assignments, make sure you write a complete expression) Finally,

$$E[Y] = \int_0^1 y f_Y(y) dy = \int_0^1 \frac{1}{2} y^{1/2} dy = \frac{1}{3} y^{3/2} \Big|_{y=0}^{1} = \frac{1}{3} - 0 = \frac{1}{3}$$

Proof 2

Apply the bonus result from HW 2: Note that $Y = h(x)$ where $h(x) = x^2$. For $0 \leq y \leq 1$,

$$f_Y(y) = f_X(h^{-1}(y)) \cdot \left| \frac{d}{dy} h^{-1}(y) \right| = f_X(\sqrt{y}) \left| \frac{d}{dy} \sqrt{y} \right| = 1 \cdot \frac{1}{2} y^{-1/2}$$

(On your assignments, make sure you write a complete expression) Finally, we find the expectation:

$$E[Y] = \int_0^1 y f_Y(y) dy = \int_0^1 \frac{1}{2} y^{1/2} dy = \frac{1}{3} y^{3/2} \Big|_{y=0}^1 = \frac{1}{3} - 0 = \frac{1}{3}$$

Proof 3

Apply the Law of the Unthinking Statistician (LOTUS):

$$E[Y] = \int_{-\infty}^{\infty} x^2 f_X(x) dx = \int_{-\infty}^{0} x^2 \cdot 0 + \int_0^1 x^2 \cdot 1 dx + \int_1^{\infty} x^2 \cdot 0$$

$$E[Y] = 0 + \frac{1}{3} x^3 \Big|_0^1 + 0 = \frac{1}{3}$$

Proof Debrief

- In words, what is the power of the Law of the Unthinking Statistician?
- Which was your favorite proof?

3.5 Linearity of Expectation

Expectation is Linear

- For a random variable X and constants a and b:

$$E(aX + b) = aE(X) + b$$

- For two random variables X and Y:

$$E(X + Y) = E(X) + E(Y)$$

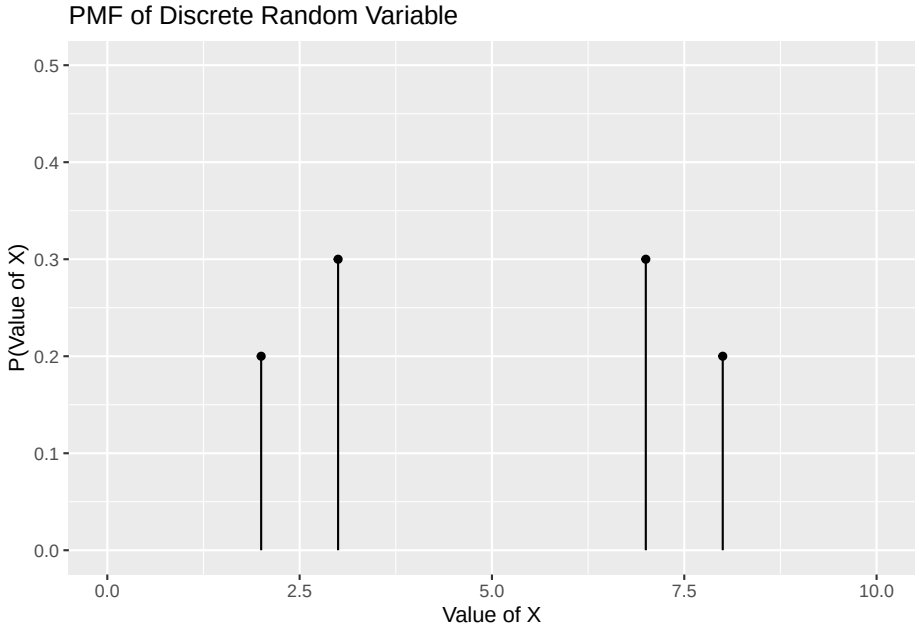
.

Using Linearity, Part I

- Linearity is a very powerful property. Using it can often make your solutions much cleaner.
- Say you want to find the expectation of a symmetric variable, X. For example, here's a random variable that's symmetric around 5.

Using Linearity, Part II

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.6     v dplyr   1.0.7
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.1.1     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```



Chapter 4

Conditional Expectation and The BLP

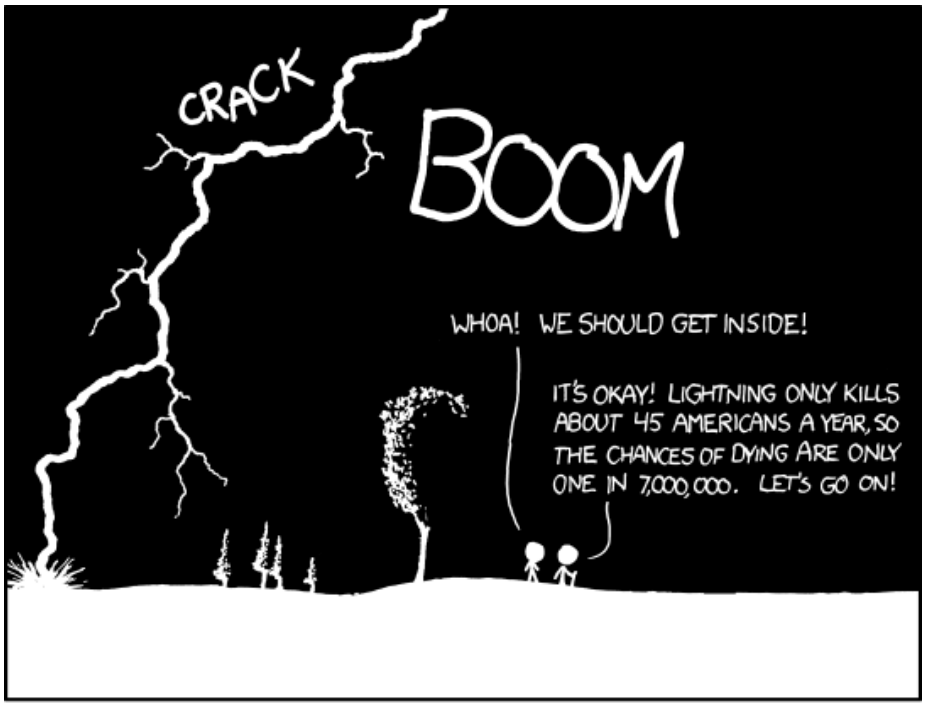


Figure 4.1: thunder struck

4.1 Learning Objectives

1.
2.
3.

4.2 Class Announcements

4.3 Roadmap

Roadmap – Rearview Mirror

- Statisticians create a population model to represent the world.
- $E[X], V[X], Cov[X, Y]$ are "simple" summaries of complex joint distributions, which are hooks for our analyses.
- They also have useful properties – for example, $E[X + Y] = E[X] + E[Y]$.

Roadmap – This week

- We look at situations with one or more "input" random variables, and one "output."
- Conditional expectation summarizes the output, given values for the inputs.
- The conditional expectation function (CEF) is a predictor – a function that yields a value for the output, give values for the inputs.
- The best linear predictor (BLP) summarizes a relationship using a line / linear function.

Roadmap – Coming Attractions

- OLS regression is a workhorse of modern statistics, causal analysis, etc
  - It is also the basis for many other models in classical stats and machine learning
- The target that OLS estimates is exactly the BLP, which we're learning about this week.

4.4 Conditional Expectation Function (CEF)

- Expectation of Y:

$$E[Y] = \int_{-\infty}^{\infty} y \cdot f_Y(y) dy$$

- Conditional expectation of Y given $X = x \in \text{Supp}[X]$:

$$E[Y|X = x] = \int_{-\infty}^{\infty} y \cdot f_{Y|X}(y|x)dy$$

- Compare and contrast $E[Y]$ and $E[Y|X]$. For example, how are their components similar or different?

- What is $E[Y|X]$ a function of? What are "input" variables to this function?

- What is $E[E[Y|X]]$ a function of?

- (Question 1) In both, we have a probability density function multiplied by the values that we realize; this is basically serving as a "weighting" function, we're merely changing what that weighting function is!

- In both cases we're looking at a "mean" of a distribution – in one it is just a "conditional mean" than a "marginal mean".

- (Question 2) We integrate Y out completely by using its every value in the integral, so E will not have Y in the answer – instead, it will remain a function of X.

- $f_{Y|X=x}$ is a function of x! So, the conditional expectation is some function x as well.

- (Optional to discuss; probably too minute) An observation that some parsing students will make:

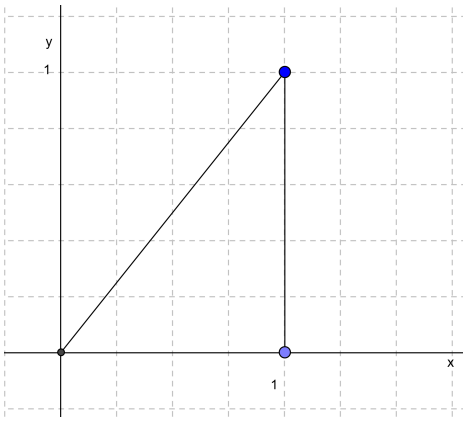
  - Without fixing X to some realization (denoted as a "little x", x, then $E[Y|X]$ is a function of X and so the whole statement is a function of a random variable (i.e. $E[Y|X]$ is *itself* a RV).
  - Once we fix $X = x$, then $E[Y|X = x]$ is fixed to some constant – there is one value that $E[Y|X = x]$ maps to.

- (Question 3) That isn't a function of anything! Once you've computed $E[E[Y|X]]$, you've integrated out all variables!

4.5 Computing the CEF

- Suppose that random variables X and Y are jointly continuous, with joint density function given by,

$$f(x,y) = \begin{cases} 2, & 0 \leq x \leq 1, 0 \leq y \leq x \\ 0, & otherwise \end{cases}$$



- In week 2 we computed the conditional density of this function:

$$f_{Y|X}(y|x) = \begin{cases} \frac{1}{x}, & 0 \leq y \leq x \\ 0, & \text{otherwise} \end{cases}$$

- What is the conditional expectation function?
- What is the conditional variance function?

4.6 Group Exercise

4.6.1 Minimizing MSE

- Theorem 2.2.20 states:

  The CEF $E[Y|X]$ is the "best" predictor of Y given X, where "best" means it has the smallest mean squared error (MSE).

- **Task:** Justify every transition ("=" sign) of the proof below using earlier FOAS concepts, definitions, theorems, calculus, and algebraic operations.
- **Proof:**
  - We need to find such function $g : R \to R$ that gives the smallest $E[(Y - g(X))^2]$.

  - It should turn out that $g(X)$ is actually $E[Y|X]$.

- Deriving a Function to Minimize MSE

$$\begin{aligned}
E[(Y-g(X))^2|X] &= E[Y^2 - 2Yg(X) + g^2(X)|X] \\
&= E[Y^2|X] + E[-2Yg(X)|X] + E[g^2(X)|X] \\
&= E[Y^2|X] - 2g(X)E[Y|X] + g^2(X)E[1|X] \\
&= (E[Y^2|X] - E^2[Y|X]) + (E^2[Y|X] - 2g(X)E[Y|X] + g^2(X)) \\
&= V[Y|X] + (E^2[Y|X] - 2g(X)E[Y|X] + g^2(X)) \\
&= V[Y|X] + (E[Y|X] - g(X))^2
\end{aligned}$$

- Then we have:

$$\begin{aligned}
E[(Y-g(X))^2] &= E\big[E[(Y-g(X))^2|X]\big] \\
&= E\big[V[Y|X] + (E[Y|X] - g(X))^2\big] \\
&= E\big[V[Y|X]\big] + E\big[(E[Y|X] - g(X))^2\big]
\end{aligned}$$

- $E[V[Y|X]]$ doesn't depend on g; and,
- $E[(E[Y|X] - g(X))^2] \geq 0$.

$\therefore g(X) = E[Y|X]$ gives the smallest $E[(Y-g(X))^2]$

- **Implication:** If you are choosing some g, you can't do better than $g(x) = E[Y|X = x]$.

4.6.2 Working with the BLP

Why Linear?

- In some cases, we might try to estimate the CEF. More commonly, however, we work with linear predictors. Why?

- We don't know joint density function of Y. So, it is "difficult" to derive a suitable CEF.
- To estimate *flexible* functions requires considerably more data. Assumptions about distribution (e.g. a linear form) allow you to leverage those assumptions to learn 'more' from the same amount of data.

- Other times, the CEF, even if we *could* produce an estimate, might be so complex that it isn't useful or would be difficult to work with.
- And, many times, linear predictors (which might seem trivially simple) actually do a very good job of producing predictions that are 'close' or useful.

4.6.3 Joint Distribution Practice

4.6.3.1 Professorial Mistakes (Discrete RVs)

- Let the number of questions that students ask be a RV, X.
- Let X take on values: $\{1, 2, 3\}$, each with probability $1/3$.
- Every time a student asks a question, the instructor answers incorrectly with probability $1/4$, independently of other questions.
- Let the RV Y be number of incorrect responses.
- **Questions:**
  - Compute the expectation of Y, conditional on X, $E[Y|X]$
  - Using the law of iterated expectations, compute $E[Y] = E\big[E[Y|X]\big]$.

4.6.3.2 (Bonus Questions)

- Working with the same question from the last slide:
  - Compute the expectation of the product of X and Y, $E(XY)$
  - Using the previous result, compute $\text{cov}(X, Y)$.

Chapter 5

Learning from Random Samples

Figure 5.1: south hall

5.1 Learning Objectives

1.

2.
3.

5.2 Class Announcements

- You're done with probability theory. Congrats!
- You're also done with your first test. Congrats!
- We're going to have a second test in a few weeks; then we're done testing for the semester

5.3 Roadmap

Where We're Going – Coming Attractions

- We're going to start bringing data into our work
- First, we're going to develop a testing framework that is built on sampling theory and reference distributions – `t.tests`, `wilcox.test` and the like
- Second, we're going to show that OLS regression is the sample estimator of the BLP
- Third, we're going to use the testing distribution to test regression coefficients

Where We've Been – Random Variables and Probability Theory

- Statisticians create a model (A.K.A. population) to represent the world.
- That model can be described by parameters like expecation, covariance.
- So far, these parameter values have come from our imaginations

Where we Are

- We want to fit models – use data to set their parameter values.
- A sample is a set of random variables
- Sample statistics are functions of a sample, and they are random variables
- Under iid and other assumptions, we get useful properties:
  - Statistics may be consistent estimators for population parameters
  - The distribution of sample statistics may be asymptotically normal

5.4 Key Terms and Assumptions

5.4.1 Definitions

Define each of the following:

- Sample
- Sample Statistic
- Estimator
- Bias
- Efficiency
- Consistency
- Convergence in Probability
- Convergence in Distribution

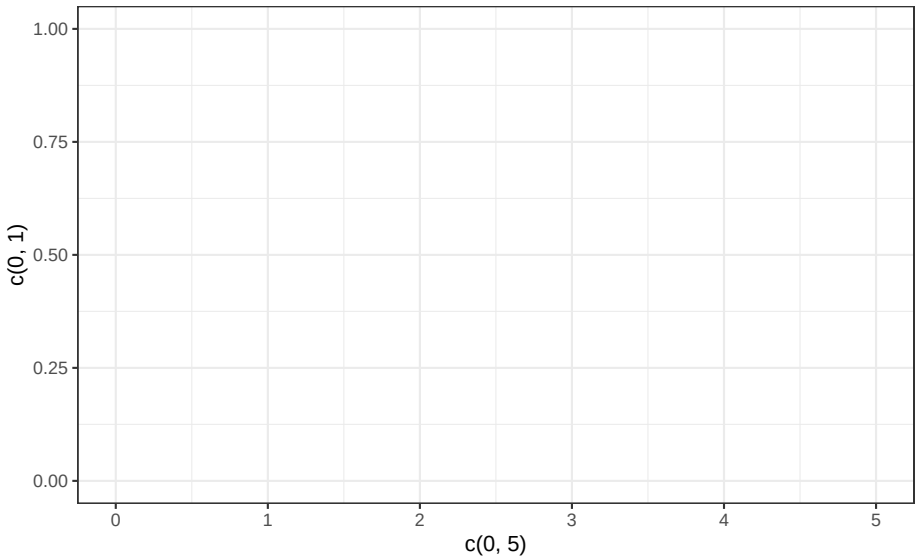
5.4.2 IID

For each scenario, is the IID assumption plausible?

- Call a random phone number. If someone answers, interview all persons in the household. Repeat until you have data on 100 people.
- Call a random phone number, interview the person if they are over 30. Repeat until you have data on 100 people.
- Record year-to-date price change for 20 largest car manufacturers.
- Measure net exports per GDP for all 195 countries recognized by the UN.

5.4.3 Understanding Sampling Distributions

- Let X be a Bernoulli random variable representing an unfair coin with $P(X = 1) = 0.7$.
- You have an iid sample of size 2, (X_1, X_2).
- Compute the sampling distribution of $\overline{X} = \frac{X_1+X_2}{2}$.



Questions:

- Explain the difference between a population distribution and the sampling distribution of a statistic.
- As we toss more and more coins, $\overline{X}_{(100)} \rightarrow \overline{X}_{(10000)}$ what will the value of $\overline{X}$ get closer to? What law generates this, and why does this law generate this result?
- Why do we want to know things about the sampling distribution of a statistic?

5.5 Uncertainty

Which Result is Better?

- Suppose that you measure salary data among individuals who try different strategies
- Report out in the following table:

	Early Rising	Mindfulness Retreat	MIDS Degree
Increase in Salary	\$1020	\$5130	\$9200
SE	(\$350)	(\$4560)	
$N =$	1,000	77	700

(Standard errors in parentheses when available)

5.6 Write Code to Demo the Central Limit Theorem (CLT)

5.6.1 Motivating the Central Limit Theor (CLT)

- Standard Errors tell us a lot about the uncertainty in our statistics
- But we want to say more:
  - How confident are we that this vitamin has a positive effect?
  - How plausible is a mean income $1000 below our estimate?
- For these questions, we need to know the sampling distribution of our statistic.
- How is this possible when we don't know the population distribution?

5.6.2 Sampling from the Bernoulli Distribution in R

- To demonstrate the CLT, we chose a Bernoulli distribution with parameter p.
  - This distribution is very simple
  - This distribution is non-normal, and can be very skewed depending on p.
- First, set `p=0.5` so your population distribution is symmetric. Use a variable `n` to represent your sample size. Initially, set `n=3`.

```
n <- 3
p <- 0.5
```

5.6.3 Useful R Commands

sample() or rbinom()

- R doesn't have a `bernoulli` function.
- To simulate draws from a Bernoulli variable, you can either:
  - Use `sample`
  - Or, use `rbinom` (the Bernoulli distribution is a special case of a binomial distribution. In this function, `size` refers to a distribution parameter, not the number of draws.)

```
sample(x=0:1, size=n, replace=TRUE, prob=c(1-p, p))
rbinom(n=n, size=1, prob=p)
```

```
## [1] 0 0 1
## [1] 0 0 0
```

replicate()

- To repeat an action, you can use `replicate`

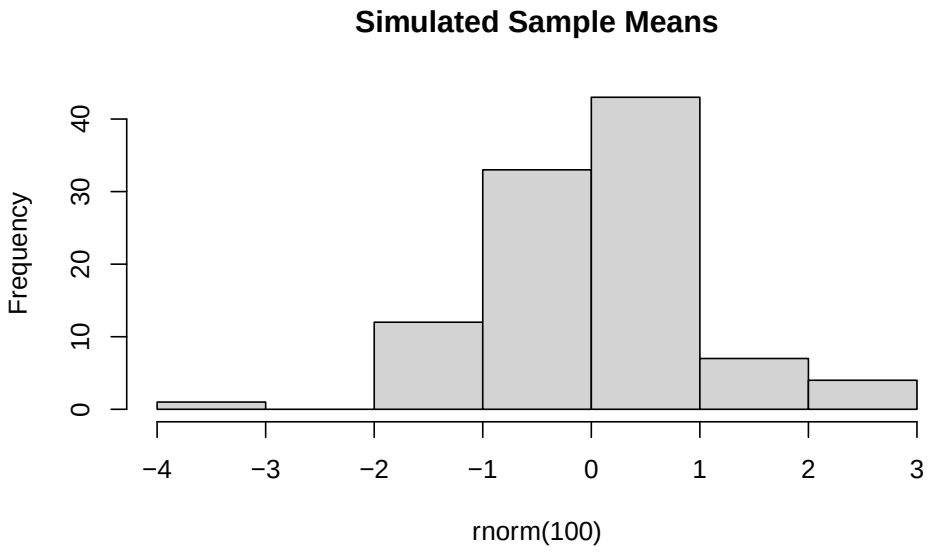
```
replicate(10, log(10))
```

```
##  [1] 2.302585 2.302585 2.302585 2.302585 2.302585 2.302585 2.302585 2.302585
##  [9] 2.302585 2.302585
```

hist()

- To quickly visualize your results, try `hist`

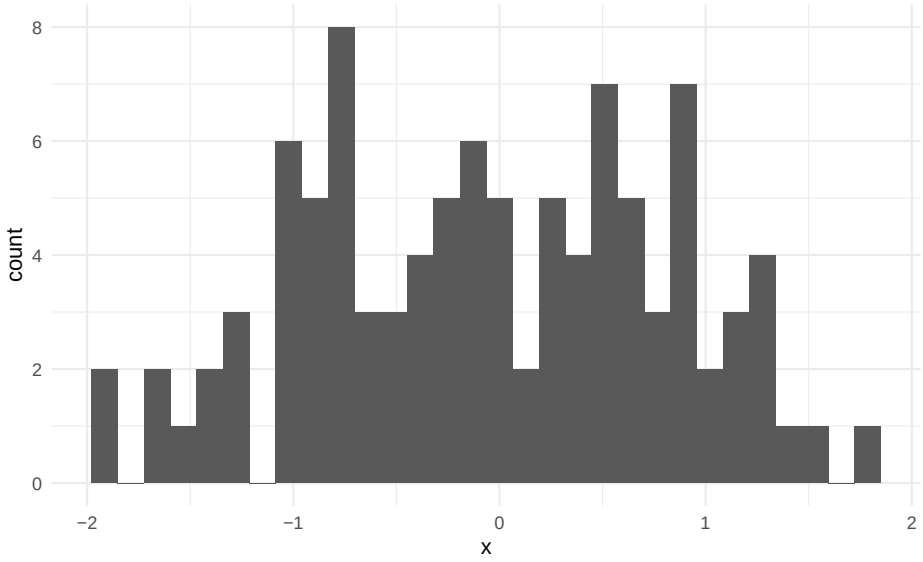
```
hist(x = rnorm(100), main = "Simulated Sample Means")
```



ggplot()

- Or, to work with `ggplot` store these results in a data.frame.

```
d <- data.frame(x = rnorm(100))
d %>%
  ggplot(aes(x=x)) +
  geom_histogram()
```



5.7 Exercise

Part 1

Throughout this part, we will use fair coins (`p = 0.5`).

1. Fill in the function below so that it simulates taking n draws from a Bernoulli distribution with parameter p. This is like tossing n coins at the same time. Use the `mean` function to compute the sample mean – the average of the number of heads that are showing. Make sure that when you run it, you return values in $\{0, 1/3, 2/3, 1\}$.

```
experiment = function(n, p){
}
```

2. The sample mean is a random variable. To understand it, use the visualization trick from a few weeks ago. Use the `replicate` function to run the above experiment 1000 times, and plot a histogram of the results.

3. If you replicate the experiment enough times, will the distribution ever look normal? Why or why not?

4. Use `sd()` to check the standard deviation of the sampling distribution of the mean for `number_of_coins = 3`. What sample size is needed to decrease the standard deviation by a factor of 10? Check that your answer is correct.

Part 2

For this part, we'll continue to study a fair coin.

5. Try different values for the sample size n, and examine the shape of the sampling distribution of the mean. At what point does it look normal to you?

Part 3

For this part, we'll study a very unfair coin. `p = 0.01`.

This is an example of a highly skewed random variable. That roughly means that one tail is a lot longer than the other.

For this activity, you can simply use your eyes to gauge how skewed a distribution is. If you prefer, you can also use the skewness command in the univar package to measure skewness. You may hear a rule of thumb that a skewness above 1 or below -1 is a highly skewed distribution.

6. Start with n=3 as before. What do you notice about the shape of the sampling distribution?
7. Try different values for the sample size n, and examine the shape of the sampling distribution of the mean. At what point does it look normal to you?

5.8 Discussion Questions

1. How does the skewness of the population distribution affect the applicability of the Central Limit Theorem? What lesson can you take for your practice of statistics?
2. Name a variable you would be interested in measuring that has a substantially skewed distribution.
3. One definition of a heavy tailed distribution is one with infinite variance. For example, you can use the `rcauchy` command in R to take draws from a Cauchy distribution, which has heavy tails. Do you think a "heavy tails" distribution will follow the CLT? What leads you to this intuition?

Chapter 6

Hypothesis Testing

6.1 Learning Objectives

1.
2.
3.

6.2 Class Announcements

- Test 2 is this week
  - Includes units 4-5 (not 6)
  - There is another practice test on Gradescope
- Lab 1 starts after live session 7
  - 2 week lab
  - You will work in a group to conduct hypothesis tests

6.3 Roadmap

Looking Backwards

- Statisticians create a model to represent the world
- We saw examples of estimators, which approximate model parameters we're interested in.
- By itself, an estimate isn't much good; we need to capture the uncertainty in the estimate.

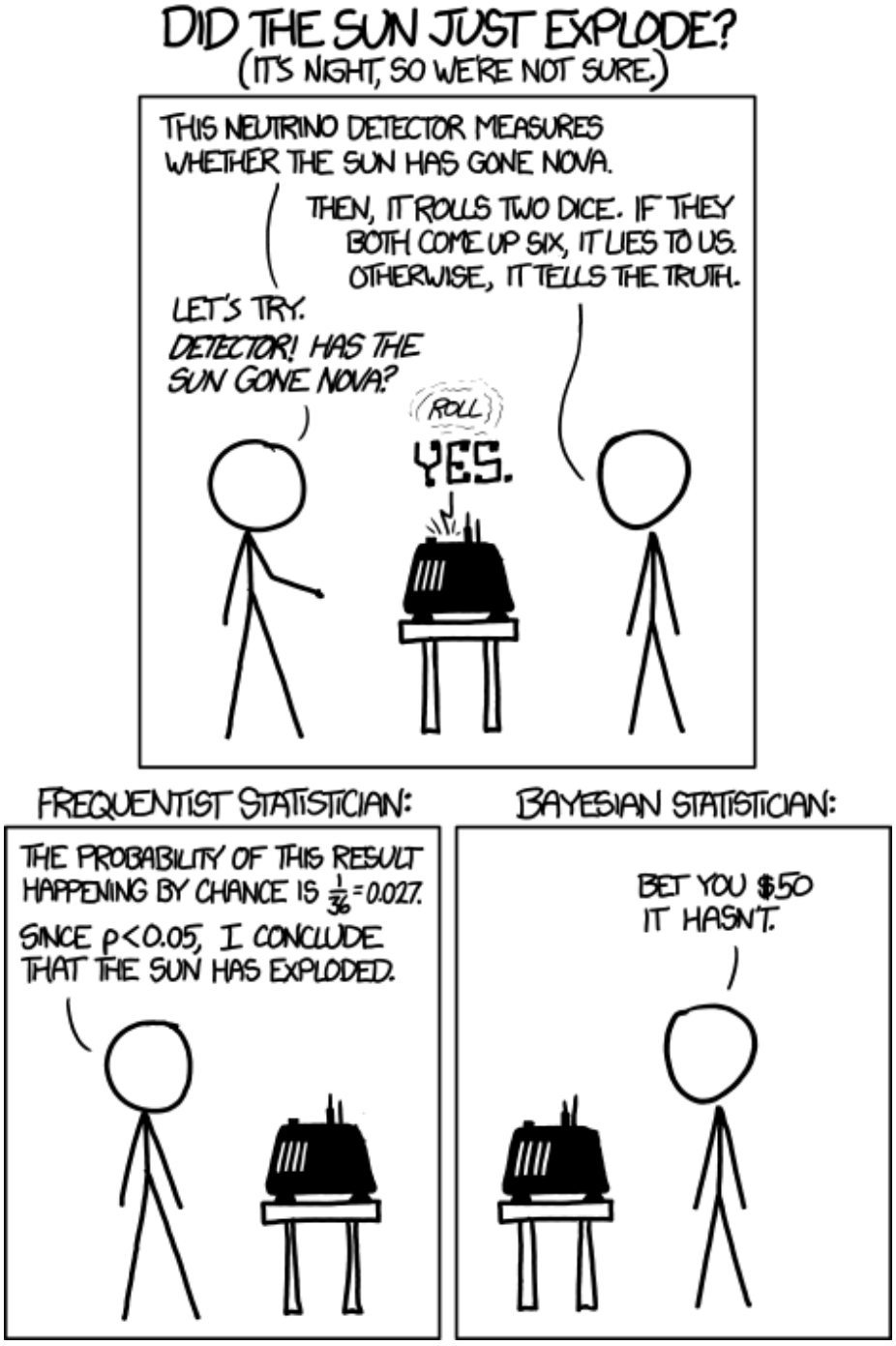


Figure 6.1: update!

- We've seen two ways to express uncertainty in an estimator: standard errors and confidence intervals.

Today

- We introduce hypothesis testing
  - A hypothesis test also captures uncertainty, but in relation to a specific hypothesis.

Looking Ahead

- We'll build on the one-sample t-test, to introduce several other statistical tests.
- We'll see how to choose a test from different alternatives, with an eye on meeting the required assumptions, and maximizing power.

6.4 Discussion

6.4.1 Discussion Questions 1

- What are the two possible outcomes of a hypothesis test?
- What guarantee do you get if you follow the decision rules properly?
- Why do we standardize the mean to create a test statistic?

$$t = \frac{\overline{X}_n - \mu}{\sqrt{\frac{s^2}{n}}}$$

6.4.2 Discussion Questions 2

- Explain this joke:

6.5 Manual Computation of a t-Test

In a warehouse full of power packs labeled as 12 volts we randomly measured the voltage of 7. Here is the data:

```
voltage <- c(11.77, 11.90, 11.64, 11.84, 12.13, 11.99,  11.77)
```

1. Find the mean and the standard deviation.
2. Using `qt()`, compute the t critical value for a hypothesis test for this sample. (Following convention, set $\alpha = .05$.)
3. Define a test statistic, t, for testing whether the population mean is 12.
4. Calculate the p-value using the t statistic.
5. Should you reject the null? Argue this in two different ways.
6. Suppose you were to use a normal distribution instead of a t-distribution to test your hypothesis. What would your p-value be for the z-test?
7. Without actually computing it, say whether a 95% confidence interval for the mean would include 12 volts.
8. Compute a 95% confidence interval for the mean.

6.6 Data Exercise

t-Test Micro Cheat Sheet

- Key t-Test Assumptions
  - Metric variable
  - IID
  - No major deviations from normality, considering sample size

Testing the Home Team Advantage

The file athlet2.Rdata contains data on college football games. The data is provided by Wooldridge and was collected by Paul Anderson, an MSU economics major, for a term project. Football records and scores are from 1993 football season.

```
load("data/athlet2.RData")
data
```

```
##    dscore dinstt doutstt htpriv vtpriv dapps htwrd vtwrd dwinrec dpriv
## 1      10   -409   -4679      0      0 -1038     1     1       0     0
## 2     -14     NA     -66      0      0 -7051     1     1       0     0
## 3      23   -654    -637      0      0  6209     1     0       1     0
## 4       8   -222     456      0      0  -129     1     1       0     0
## 5     -12    -10     208      0      0   794     1     1       0     0
## 6       7    494      17      0      0   411     0     0       0     0
## 7     -21      2       2      0      0 -4363     1     1       0     0
```

```
## 8      -5     96     -333     0     0  1144     1     0      1     0
## 9      -3    223     2526     0     0  3956     0     0      0     0
## 10    -32    -20        0     0     0  -641     0     1     -1     0
## 11      9     66        0     0     0  -278     1     0      1     0
## 12      1     56     -346     0     0 -2223     1     0      1     0
## 13      7    556      717     0     0 -5217     1     0      1     0
## 14    -20    169     -461     0     0  1772     0     1     -1     0
## 15     35   -135      396     0     0    85     1     0      1     0
## 16     35    -40        0     0     0  -988     1     0      1     0
## 17    -25     24        0     0     0 -8140     1     1      0     0
## 18     -9     90        0     0     0  8418     0     1     -1     0
## 19    -33     27      900     0     0 -3273     0     0      0     0
## 20      7    -89      -31     0     0  1906     1     0      1     0
## 21     -3    536     2352     0     0  -151     1     1      0     0
## 22     -6  13261     9111     1     0 -9936     1     1      0     1
## 23    -29  13809    10076     1     0 -6265     0     1     -1     1
## 24     14 -17631   -10589     0     1  1252     1     0      1    -1
## 25    -18  14885     9983     1     0 -4529     1     1      0     1
## 26     48 -15220   -11400     0     1  -318     1     0      1    -1
## 27     -3     99      -29     0     0  -797     0     1     -1     0
## 28     -3    -54      -88     0     0  -372     0     1     -1     0
## 29     -3    -98    -4175     1     0  2460     1     1      0     1
## 30      2   -304     2987     0     1 -3035     1     1      0    -1
```

We are especially interested in the variable, dscore, which represents the score differential, home team score - visiting team score. We would like to test whether a home team really has an advantage over the visiting team.

1. The instructor will assign you to one of two teams. Team 1 will argue that the t-test is appropriate to this scenario. Team 2 will argue that the t-test is invalid. Take a few minutes to examine the data, then formulate your best argument.
2. Should you perform a one-tailed test or a two-tailed test? What is the strongest argument for your answer?
3. Execute the t-test and interpret every component of the output.
4. Based on your output, suggest a different hypothesis that would have led to a different test result. Try executing the test to confirm that you are correct.

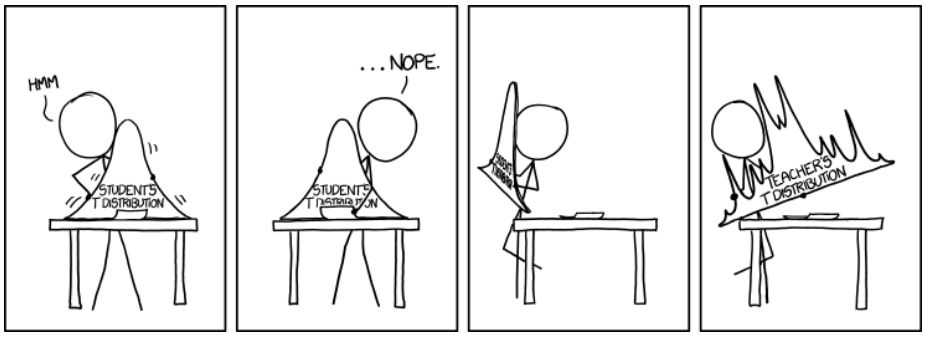
6.7 Assumptions Behind the t-test

For the following scenarios, what is the strongest argument against the validity of a t-test?

- You have a sample of 50 CEO salaries, and you want to know whether the mean salary is greater than $1 million.
- A nonprofit organization measures the percentage of students that pass an 8th grade reading test in 40 neighboring California counties. You are interested in whether the percentage of students that pass in California is over 80%
- You have survey data in which respondents assess their own opinion of corgis, with options ranging from "1 - extreme disgust" to "5 - affection so intense it threatens my career." You want to know whether people on the average like corgis more than 3, representing neutrality.

Chapter 7

Comparing Two Groups



7.1 Learning Objectives

1.
2.
3.

7.2 Class Announcements

- Great work completing your final w203 test!
- There is **no** unit 7 homework!
- The Hypothesis Testing Lab is released today!
  - Lab is due at Unit 09 Live Session (two weeks)
  - Group lab in two parts:

    * *Part 1*: Work as a team to engage the fundamentals of hypothesis tests
    * *Part 2*: Apply these fundamentals to analyze 2020 election data and write a single, three-page analysis

7.3 Roadmap

Rearview Mirror

- Statisticians create a population model to represent the world
- A population model has parameters we are interested in
    - Ex: A parameter might represent the effect that a vitamin has on test performance
- A null hypothesis is a specific statement about a parameter
    - Ex: The vitamin has zero effect on performance
- A hypothesis test is a procedure for rejecting or not rejecting a null, such the probability of a type 1 error is constrained.

Today

- There are often multiple hypothesis tests you can apply to a scenario.
- Our primary concern is choosing a test with assumptions we can defend.
- Secondarily, we want to maximize power.

Looking Ahead

- Next week, we start working with models for linear regression
- We will see how hypothesis testing is also used for regression parameters.

7.4 Teamwork Discussion

Working on Data Science Teams

- Data science is a *beautiful* combination of team-work and individual-work
    - Teams:
        * Define research ambitions and scope
        * Imagine/envision the landscape of what is possible
        * Support, discuss, review and integrate individual contributions

  - Individuals:
    * Conduct the heads-down work that moves question answering forward

The Problematic Psychology of Data Science

- People talk about the *impostor syndrome* – a feeling of inadequacy or interloping that is sometimes also associated with a fear of under-performing relative to the expectation of others on the team.
  - These emotions are common through data science, academics, computer science.
  - But, the emotions are also commonplace in journalism, film-making, and public speaking
- What might be generating these feelings?

What Makes an Effective Team?

- This reading on *effective* teams summarizes academic research to argue:

  What really matters is less about who is on the team, and more about how the team works together.

- In your live session, your section might take 7 minutes to read this brief, reading the problem statement, the proposed solution, and the framework for team effectiveness (stopping at the section titled *"Tool: Help teams determine their own needs."*)

  "Psychological safety refers to an individual's perception of the consequences of taking an interpersonal risk. It is a belief that a team is safe for risk taking in the face of being seen as ignorant, incompetent, negative, or disruptive."

  "In a team with high psychological safety, teammates feel safe to take risks around their team members. They feel confident that no one on the team will embarrass or punish anyone else for admitting a mistake, asking a question, or offering a new idea."

7.4.1 We All Belong

- From your experience, can you give an example of taking a personal risk as part of a team?
  - Can you describe your emotions when contemplating this risk?
  - If you did take the risk, how did the reactions of your teammates affect you?
- Knowing the circumstances that generate feelings of anxiety – what steps can we take as a section, or a team, to recognize and respond to these circumstances?

  How can you add to the psychological safety of your peers in the section and lab teammates?

- Only 26% of data science jobs are held by women.
  - Morgan Ames (a professor in the program, teaching w231), in *The Charisma Machine* (available for free through the library, and with the introduction chapter here) studies the One Laptop Per Child initiative, and argues that the failure of OLPC is (another) example of:
- Sociological barriers that make broad-based, inclusive collaboration in tech challenging.

  "One Laptop Per Child implicitly invokes the social imaginary of the technically precocious boy... This imaginary shows a "natural" mastery of technical toys as well as a particular kind of rebellious sensibility that enables technical tinkering—but is still exclusionary by connecting technical prowess to boys in particular."

  "In contrast, I found that each [successful] student had a constellation of resources that encouraged them down this path: families that steered them toward creative and critical thinking, a focus on the importance of education, and in many cases another computer in the home. This account circumscribes the role of technology, [... and] instead highlights the importance of social worlds."

- In the face of uneven challenges faced by under-represented peoples in data science, what actions can we take to produce an inclusive, integrated data science team?

7.5 Team Kick-Off

Lab 1 Teams

- Here are teams for Lab 1!

Team Kick-Off Conversation

- In a 10 minute breakout with your team, please discuss the following questions:

1. How much time will you invest in the lab each week?
2. How often will you meet and for how long?
3. How will you discuss, review, and integrate individual work into the team deliverable?
4. What do you see as the biggest risks when working on a team? How can you contribute to an effective team dynamic?

7.6 A Quick Review

Review of Key Terms

- Define each of the following:
  - Population Parameter
  - Null Hypothesis
  - Test Statistic
  - Null Distribution

Comparing Groups Review

Take a moment to recall the tests you learned this week. Here is a quick cheat-sheet to their key assumptions.

paired/unpaired	paired metric	non-parametric
unpaired	**unpaired t-test** - metric var - i.i.d. - (not too un-)normal	**Wicoxon rank-sum** ordinal var i.i.d.
paired	**paired t-test** metric var i.i.d. (not too un-)normal	**Wicoxon signed-rank** metric var i.i.d. difference is symmetric **sign test** ordinal var i.i.d.

7.7 Comparing Groups R Exercise

The General Social Survey (GSS) is one of the longest running and extensive survey projects in the US. The full dataset includes over 1000 variables spanning demographics, attitudes, and behaviors. The file `GSS_w203.RData` contains a small selection of a variables from the 2018 GSS.

To learn about each variable, you can enter it into the search bar at the GSS data explorer

```
load('data/GSS_w203.RData')
summary(GSS)
```

```
##            rincome               happy                              sexnow
##  $25000 or more: 851   very happy    : 701   women                     :758
##  $20000 - 24999: 107   pretty happy :1307   man                       :640
##  $10000 - 14999:  94   not too happy: 336   transgender               :  2
##  $15000 - 19999:  61   DK           :   0   a gender not listed here:  1
##  lt $1000       :  33   IAP          :   0   Don't know                :  0
##  (Other)        : 169   NA           :   0   (Other)                   :  0
##  NA's           :1033   NA's         :   4   NA's                      :947
##      wwwhr           emailhr                     socrel
##  Min.   :  0.00   Min.   :  0.000   sev times a week:382
##  1st Qu.:  3.00   1st Qu.:  0.000   sev times a mnth:287
##  Median :  8.00   Median :  2.000   once a month    :259
##  Mean   : 13.91   Mean   :  7.152   sev times a year:240
##  3rd Qu.: 20.00   3rd Qu.: 10.000   almost daily    :217
##  Max.   :140.00   Max.   :100.000   (Other)         :171
##  NA's   :986      NA's   :929       NA's            :792
##              socommun      numpets          tvhours
##  never           :510   Min.   : 0.000   Min.   : 0.000
##  once a month    :243   1st Qu.: 0.000   1st Qu.: 1.000
##  sev times a week:219   Median : 1.000   Median : 2.000
##  sev times a year:196   Mean   : 1.718   Mean   : 2.938
##  sev times a mnth:174   3rd Qu.: 2.000   3rd Qu.: 4.000
##  (Other)         :215   Max.   :20.000   Max.   :24.000
##  NA's            :791   NA's   :1201     NA's   :793
##                      major1         owngun
##  business administration: 138   yes    :537
##  education              :  79   no     :993
##  engineering            :  54   refused: 39
##  nursing                :  51   DK     :  0
##  health                 :  42   IAP    :  0
##  (Other)                : 546   NA     :  0
##  NA's                   :1438   NA's   :779
```

You have a set of questions that you would like to answer with a statistical test. **For each question**:

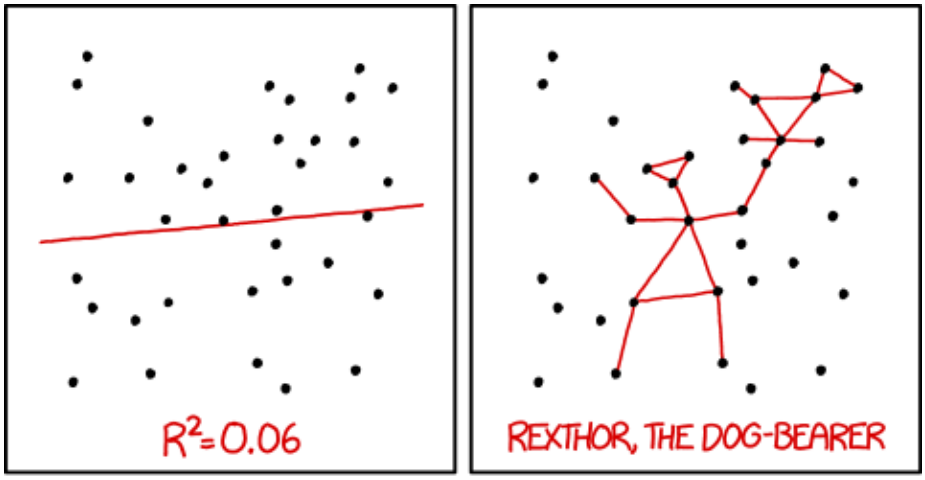
1. Choose the most appropriate test.
2. List and evaluate the assumptions for your test.
3. Conduct your test.
4. Discuss statistical and practical significance.

The Questions

- Do Americans with pets watch more or less TV than Americans without pets?
- Do economics majors watch more or less tv than computer science majors?
- Are Americans that own guns or Americans that don't own guns more likely to have pets?
- Do Americans spend more time emailing or using the web?
- Are Americans with pets happier than Americans without pets?
- Do Americans spend more evenings with neighbors or with relatives?

Chapter 8

OLS Regression Estimates



I DON'T TRUST LINEAR REGRESSIONS WHEN IT'S HARDER TO GUESS THE DIRECTION OF THE CORRELATION FROM THE SCATTER PLOT THAN TO FIND NEW CONSTELLATIONS ON IT.

8.1 Learning Objectives

1.
2.
3.

8.2 Class Announcements

1. Lab 1 is due next week.
2. There is no HW 8. We will have HW 9 as usual.
3. You're doing great - keep it up!

8.3 Roadmap

Rear-View Mirror

- Statisticians create a population model to represent the world.
- Sometimes, the model includes an "outcome" random variable Y and "input" random variables $X_1, X_2, ..., X_k$.
- The joint distribution of Y and $X_1, X_2, ..., X_k$ is complicated.
- The best linear predictor (BLP) is the canonical way to summarize the relationship.

Today

- OLS regression is an estimator for the BLP
- We'll discuss the *mechanics* of OLS

Looking Ahead

- To make regression estimates useful, we need measures of uncertainty (standard errors, tests...).
- The process of building a regression model looks different, depending on whether the goal is prediction, description, or explanation.

8.4 Regression Discussion

8.4.1 Discussion Questions

Suppose we have random variables X and Y.

- Why do we care about the BLP?
- What assumptions are needed for OLS to consistently estimate the BLP?
- What assumptions are needed in terms of causality (X causes Y, Y causes X, etc.) in order to compute the regression of Y on X?

8.4.2 Reasoning by Analogies

Here are some phrases about regression "in the population." Convert each of them to its sample counterpart.

- Population :: Sample
- Error ϵ :: residuals :: e_i
- The BLP is the predictor that minimizes expected squared error.
- $\beta_1 = \frac{Cov[X,Y]}{V[X]}$.
- $Cov[X, \epsilon] = 0$
- $E[\epsilon] = 0$
- The population moment conditions uniquely specify one line, which is the BLP.

8.5 Coding Activity:R Cheat Sheet

Suppose `x` and `y` are variables in dataframe `d`.

To fit an ols regression of Y on X:

```
mod <- lm(y ~ x, data = d)
```

To access **coefficients** from the model object:

```
mod$coefficients
or coef(mod)
```

To access **fitted values** from the model object:

```
mod$fitted
or fitted(mod)
or predict(mod)
```

To access **residuals** from the model object:

```
mod$residuals
or resid(mod)
```

To create a scatterplot that includes the regression line:

```
plot(d['x'], d['y'])
abline(mod)
or
d %>%
  ggplot() +
  aes(x = x, y = y) +
  geom_point() +
  geom_smooth(method = lm)
```

8.6 R Exercise

Real Estate in Boston

The file `hprice1.Rdata` contains 88 observations of homes in the Boston area, taken from the real estate pages of the Boston Globe during 1990. This data was provided by Wooldridge.

```
load('data/hprice1.RData') # provides 3 objects
```

```
head(data)
```

```
##     price assess bdrms lotsize sqrft colonial   lprice  lassess llotsize
## 1 300.000  349.1     4    6126  2438        1 5.703783 5.855359 8.720297
## 2 370.000  351.5     3    9903  2076        1 5.913503 5.862210 9.200593
## 3 191.000  217.7     3    5200  1374        0 5.252274 5.383118 8.556414
## 4 195.000  231.8     3    4600  1448        1 5.273000 5.445875 8.433811
## 5 373.000  319.1     4    6095  2514        1 5.921578 5.765504 8.715224
## 6 466.275  414.5     5    8566  2754        1 6.144775 6.027073 9.055556
##     lsqrft
## 1 7.798934
## 2 7.638198
## 3 7.225482
## 4 7.277938
## 5 7.829630
## 6 7.920810
```

- Are there variables that would *not* be valid outcomes for an OLS regression? If so, why?
- Are there variables that would *not* be valid inputs for an OLS regression? If so, why?

8.6.1 Assess the Relationship between Price and Square Footage

Suppose that you're interested in knowing the relationship between price and square footage.

0. Assess the assumptions of the Large-Sample Linear Model.
1. Create a scatterplot of `price` and `sqrft`. Like every plot you make, ensure that the plot *minimally* has a title and meaningful axes.
2. Find the correlation between the two variables.
3. Recall the equation for the slope of the OLS regression line – here you can either use Variance and Covariance, or if you're bold, the linear algebra. Compute the slope manually (without using `lm()`)
4. Regress `price` on `sqrft` using the `lm` function. This will produce an estimate for the following model:

$$price = \beta_0 + \beta_1 sqrft + e$$

5. Create a scatterplot that includes the fitted regression.
6. Interpret what the coefficient means.

- State what features you are allowing to change and what features you're requiring do not change.
- For each additional square foot, how much more (or less) is the house worth?

7. Estimate a new model (and save it into another object) that includes the size of the lot and whether the house is a colonial. This will estimate the model:

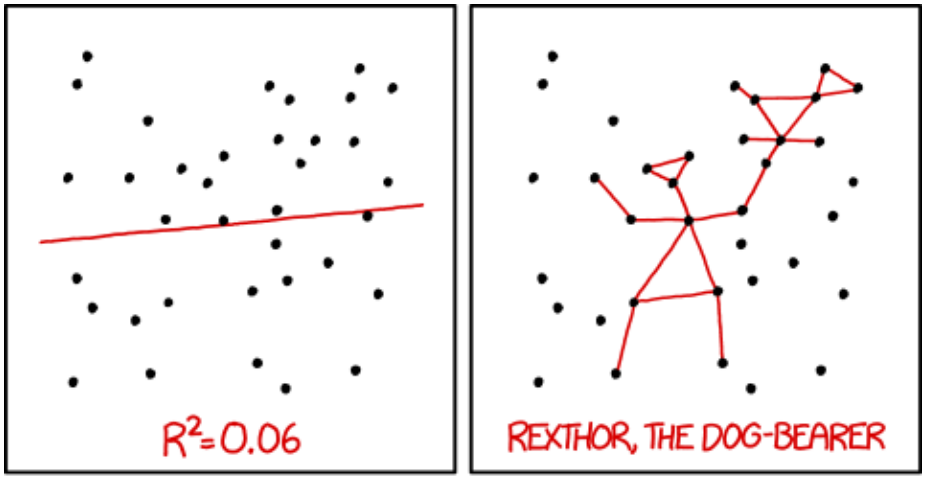
$$price = \beta_0 + \beta_1 sqrft + \beta_2 lotsize + \beta_3 colonial? + e$$

- *BUT BEFORE YOU DO*, make a prediction: What do you think is going to happen to the coefficient that relates square footage and price?
  - Will the coefficient increase, decrease, or stay the same?

7. Compute the sample correlation between X and e_i. What guarantees do we have from the book about this correlation? Does the data seem to bear this out?

Chapter 9

OLS Regression Inference



9.1 Learning Objectives

1.
2.
3.

9.2 Class Announcements

1. Congratulations on finishing your first lab!
2. The next (and the last) lab is coming up in two weeks.
3. Homework 09 has been assigned today, and it's due in a week.

9.3 Roadmap

Rear-View Mirror

- Statisticians create a population model to represent the world.
- Sometimes, the model includes an "outcome" random variable Y and "input" random variables $X_1, X_2, ..., X_k$.
- The joint distribution of Y and $X_1, X_2, ..., X_k$ is complicated.
- The best linear predictor (BLP) is the canonical way to summarize the relationship.
- OLS provides a point estimate of the BLP

Today

- Robust Standard Error: quantify the uncertainty of OLS coefficients
- Hypothesis testing with OLS coefficients
- Bootstrapping

Looking Ahead

- Regression is a foundational tool that can be applied to different contexts
- The process of building a regression model looks different, depending on whether the goal is prediction, description, or explanation.

9.4 Uncertainty in OLS

9.4.1 Discussion Questions

- List as many differences between the BLP and the OLS line as you can.
- In the following regression table, explain in your own words what the standard error in parentheses means.

	outcome: sleep hours
mg. melatonin	0.52
	(0.31)

9.4.2 Understanding Uncertainty

Under the relatively stricter assumptions of constant error variance, the variance of a slope coefficient is given by

$$V(\hat{\beta}_j) = \frac{\sigma^2}{SST_j(1 - R_j^2)}$$

A similar formulation is given in *FOAS* as definition 4.2.3,

$$\hat{V}_C[\hat{\beta}] = \hat{\sigma}^2 \left(X^T X\right)^{-1} \rightsquigarrow \frac{\hat{\sigma}^2}{(X^T X)}$$

Explain why each term makes the variance higher or lower:

- σ^2 is the variance of the error ϵ
- SST_j is (unscaled) variance of X_j
- R_j^2 is R^2 for a regression of X_j on the other X's

9.5 R Exercise

Real Estate in Boston

The file `hprice1.RData` contains 88 observations of homes in the Boston area, taken from the real estate pages of the Boston Globe during 1990. This data was provided by Wooldridge.

```
load('data/hprice1.RData') # provides 3 objects
```

Last week, we fit a regression of price on square feet.

```
model_one <- lm(price ~ sqrft, data = data)
model_one
```

```
##
## Call:
## lm(formula = price ~ sqrft, data = data)
```

```
##
## Coefficients:
## (Intercept)        sqrft
##     11.2041       0.1402
```

Questions

1. Estimate a new model (and save it into another object) that includes the size of the lot and whether the house is a colonial. This will estimate the model:

$$price = \beta_0 + \beta_1 sqrft + \beta_2 lotsize + \beta_3 colonial? + e$$

- *BUT BEFORE YOU DO*, make a prediction: What do you think is going to happen to the coefficient that relates square footage and price?
  - Will the coefficient increase, decrease, or stay the same?
  - Will the *uncertainty* about the coefficient increase, decrease, or stay the same?
  - Conduct an F-test that evaluates whether the model *as a whole* does better when the coefficients on `colonial` and `lotsize` are allowed to estimate freely, or instead are restricted to be zero (i.e. $\beta_2 = \beta_3 = 0$.

2. Use the function `vcovHC` from the `sandwich` package to estimate (a) the the heteroskedastic consistent (i.e. "robust") variance covariance matrix; and (b) the robust standard errors for the intercept and slope of this regression. Recall, what is the relationship between the VCOV and SE in a regression?

3. Perform a hypothesis test to check whether the population relationship between `sqrft` and `price` is zero. Use `coeftest()` with the robust standard errors computed above.

4. Use the robust standard error and `qt` to compute a 95% confidence interval for the coefficient `sqrft` in the second model that you estimated. $price = \beta_0 + \beta_1 sqrft + \beta_2 lotsize + \beta_3 colonial$.

5. **Bootstrap.** The book *very* quickly talks about bootstrapping which is the process of sampling *with replacement* and fitting a model. The idea behind the bootstrap is that since the data is generated via an iid sample from the population, that you can simulate re-running your analysis by drawing repeated samples from the data that you have.

Below is code that will conduct a boostrapping estimator of the uncertainty of the `sqrft` variable when `lotsize` and `colonial` are included in the model.

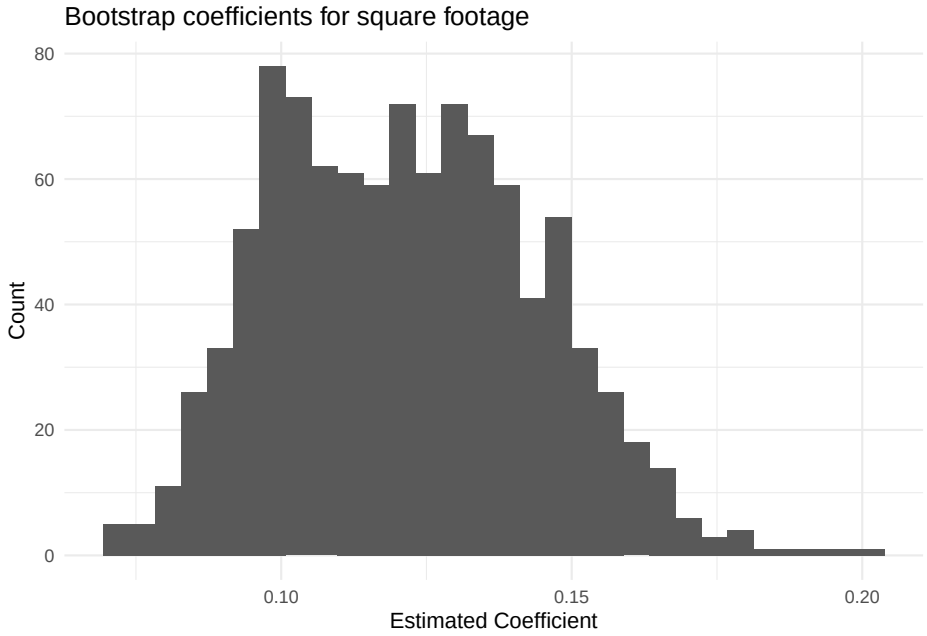
```
bootstrap_sqft <- function(d = data, number_of_bootstraps = 1000) {
  number_of_rows <- nrow(d)

    coef_sqft <- rep(NA, number_of_bootstraps)

    for(i in 1:number_of_bootstraps) {
      bootstrap_data <- d[sample(x=1:number_of_rows, size=number_of_rows, replace=TRUE), ]
      estimated_model <- lm(price ~ sqrft + lotsize + colonial, data = bootstrap_data)
      coef_sqft[i]    <- coef(estimated_model)['sqrft']
    }
  return(coef_sqft)
  }
```

```
bootstrap_result <- bootstrap_sqft(number_of_bootstraps = 1000)
```

With this, it is possible to plot the distribution of these regression coefficients:

```
ggplot() +
  aes(x = bootstrap_result) +
  geom_histogram() +
  labs(
    x = 'Estimated Coefficient',
    y = 'Count',
    title = 'Bootstrap coefficients for square footage'
  )
```

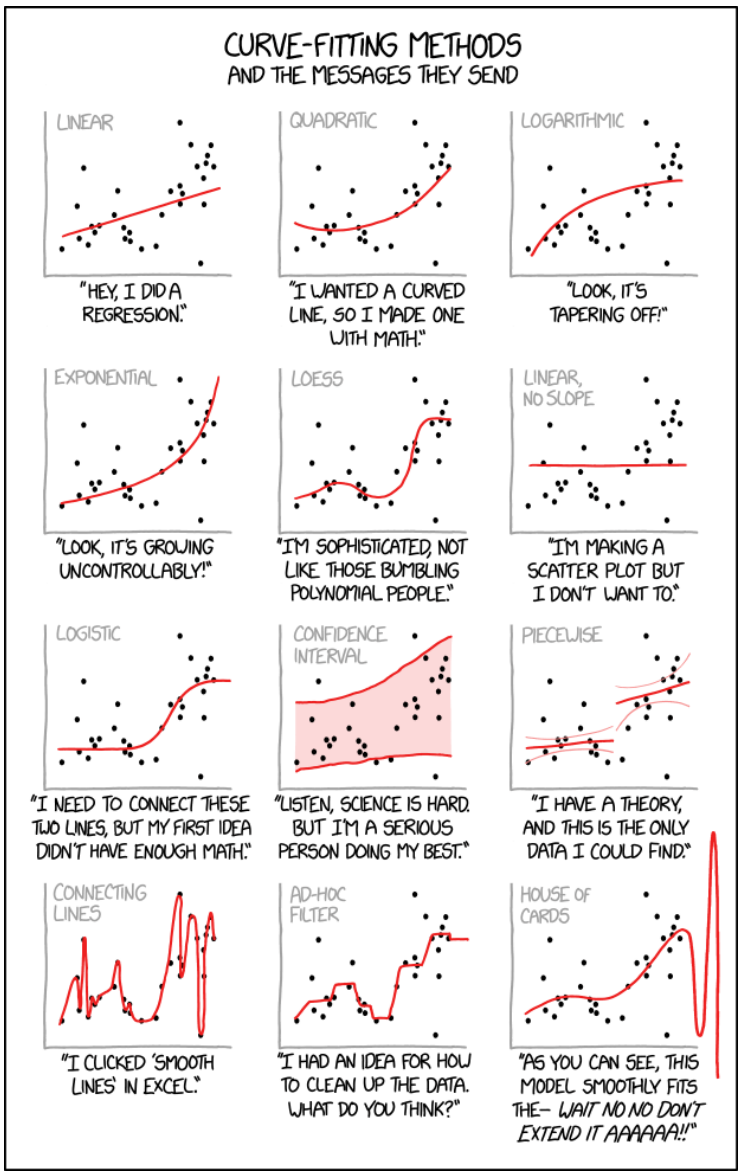
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Compute the standard deviation of the bootstrapped regression coefficients. How does this compare to the robust standard errors you computed above?

Chapter 10

Descriptive Model Building



10.1 Learning Objectives

1.
2.
3.

10.2 Class Announcements

1. The Regression Lab begins next week.
    - Your instructor will divide you into teams.
    - As part of the lab, you will perform a statistical analysis using linear regression models.

10.3 Roadmap

Rearview Mirror

- Statisticians create a population model to represent the world.
- The BLP is a useful way to summarize the relationship between one outcome random variable Y and input random varibles $X_1, ..., X_k$
- OLS regression is an estimator for the Best Linear Predictor (BLP)
- We can capture the sampling uncertainty in an OLS regression with standard errors, and tests for model parameters.

Today

- The research goal determines the strategy for building a linear model.
- Description means summarizing or representing data in a compact, human-understandable way.
- We will capture complex relationships by transforming data, including using indicator variables and interaction terms.

Looking Ahead

- We will see how model building for explanation is different from building for description.
- The famous Classical Linear Model (CLM) allows us to apply regression to smaller samples.

10.4 Discussion

10.4.1 Three modes of model building

- Recall the three major modes of model building: Prediction, Description, Explanation.
- What is the appropriate mode for each of the following questions?

1. What is going on?
2. Why is something going on?
3. What is going to happen?

- Think of a research question you are interested in. Which mode is it aligned with?

10.4.2 The statistical modeling process in different modes

- How does the modeling goal influence each of the following steps in the statistical modeling process?
  - Choice of variables and transformation
  - Choice of model (ols regression, neural nets, random forest, etc.)
  - Model evaluation

10.5 R Activity: Measuring the return to education

- In labor economics, a key concept is *returns to education*.
- Our goal is description: what is the relationship between education and wages? We will proceed in two steps:
  - First, we will discuss what the appropriate specifications are.
  - Then we will estimate the different models to answer this question.
- We will use wage1 dataset in the wooldridge package in the following sections.

```
#?wage1
#names(wage1)
```

10.5.1 Transformations

10.5.1.1 Applying and Interpreting Logarithms

- Which of the following specifications best capture the relationship between education and hourly wage? (Hint: Do a quick a EDA)
  - level-level: $wage = \beta_0 + \beta_1 educ + u$
  - Level-log: $wage = \beta_0 + \beta_1 \ln(educ) + u$
  - log-level: $\ln(wage) = \beta_0 + \beta_1 educ + u$
  - log-log: $\ln(wage) = \beta_0 + \beta_1 \ln(educ) + u$
- What is the interpretation of β_0 and β_1 in your selected specification?
- Can we use R^2 or Adjusted R^2 to choose between level-level or log-level specifications?

Remember

- **Doing a log transformation for any reason essentially implies a fundamentally different relationship between outcome (Y) and predictor (X) that we need to capture**

10.5.1.2 Applying and Interpreting Polynomials

- The following specifications include two control variables: years of experience (exper) and years at current company (tenure).
- Do a quick EDA and select the specification that better suits our description goal.
  - $wage = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 tenure + u$
  - $wage = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 exper^2 + \beta_4 tenure + \beta_5 tenure^2 + u$
- How do you interpret the β coefficients?

10.5.1.3 Applying and Interpreting Indicator variables and interaction terms

- In the following models, first, explain why the indicator variables or interaction terms have been included. Then identify the reference group (if any) and interpret all coefficients.
  - $wage = \beta_0 + \beta_1 educ + \beta_2 I(educ \geq 12) + u$

- $wage = \beta_0 + \beta_1 educ + \beta_2 female + u$
- $wage = \beta_0 + \beta_1 educ + \beta_2 female + \beta_3 educ * female + u$
- $wage = \beta_0 + \beta_1 female + \beta_2 I(educ = 2) + \beta_3 I(educ = 3)$ $... + \beta_{20} I(educ = 20) + u$

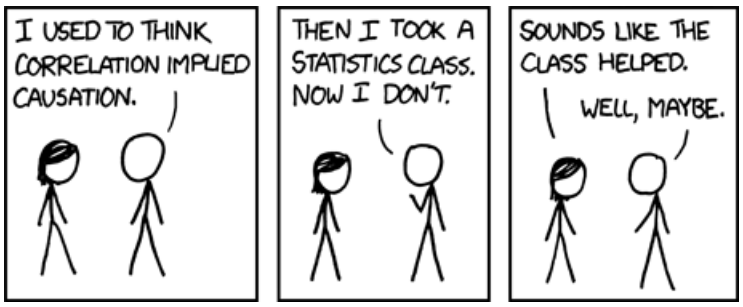
10.5.2 Estimation

Estimating Returns to Education

- Answer the following questions using an appropriate hypothesis test.
  1. Is a year of education associated with changes to hourly wage? (Include experience and tenure without polynomial terms).
  2. Is the association between wage and experience / wage and tenure non-linear?
  3. Is there evidence for gender wage discrimination in the U.S.?
  4. Is there any evidence for a graduation effect on wage?
- Display all estimated models in a regression table, and discuss the robustness of your results.

Chapter 11

Explanatory Model Building



11.1 Learning Objectives

1.
2.
3.

11.2 Class Announcements

Lab 2-Regression

Overview

- **Setting**: You are data scientists for a maker of products.

- **Task**: You select your own research question
  - Your X should be an aspect of product design
  - Your Y should be a metric of product success
- **Deliverable**: A statistical analysis that includes
  - An introduction that motivates your research question
  - A description of your model-building process
  - A discussion of statistical assumptions that may be problematic
  - A well-formatted regression table with a minimum of 3 specifications
  - A conclusion that extracts key lessons from your statistical results

The Report
- Writing for a collaborating data scientist, what research question have you asked, what answers have you found, and how did you find them?

Deliverable Name	Week Due	Grade Weight
Research Proposal	Week 12	10%
Within-Team Review	Week 12	5%
Final Presentation	Week 14	10%
Final Report	Week 14	75%

Team Work Evaluation

- Most data science work happens on teams.
- Our educational goals include helping you improve in your role as a teammate.
- We'll ask you to fill out a confidential evaluation regarding your team dynamics.

Final Presentation

- Team will present their work in live session 14.
  - Teams have between 10-15 min dedicated to discussing their work (depending on section size)
  - Two-thirds of the time can be the team presenting
  - **BUT** at least one-third should be asking and answering questions with your peers
  - For example, if teams have 15 minutes total, then plan to present for no more than 10 minutes and structure 5 minutes of questions.

11.3 Roadmap

Rearview Mirror

- Statisticians create a population model to represent the world.
- The BLP is a useful way to summarize relationships in a model, and OLS regression is a way to estimate the BLP.
- OLS regression is a foundational tool that can be applied to questions of description

Today

- Questions of explanation require a substantially different modeling process.
- To answer causal questions, we must work within a causal theory
- OLS regression is sometimes appropriate for measuring a causal effect,
- But, only when the model estimated matches the causal theory.
- So, we must watch out for omitted variable bias, reverse causality, and outcome variables on the right hand side.

Looking Ahead

- The famous Classical Linear Model (CLM) allows us to apply regression to smaller samples.
- We will address the pervasive issue of false discovery, and ways to be a responsible member of the scientific community.

11.4 Discussion

11.4.1 Path Diagrams

$$\text{Sleep} \rightarrow \text{Feelings of Stress}$$

- How would the following fit into this causal path diagram?
  1. All the other factors in the world that also cause stress but don't have a causal relationship with sleep.
  2. A factor: Coffee Intake
     - What happens if you omit it in your regression?
  3. Reverse causality
  4. An outcome variable on the RHS: Job Performance
     - What happens if you include it in your regression?

11.4.2 Omitted Variable Bias

- Recall the equation for omitted variable bias

$$\text{estimate} = \text{true parameter} + \text{omitted variable bias}$$

$$\alpha_1 = \beta_1 + \beta_2\delta_1$$

How much does omitted variable affect outcome?

How related are measured and omitted variables?

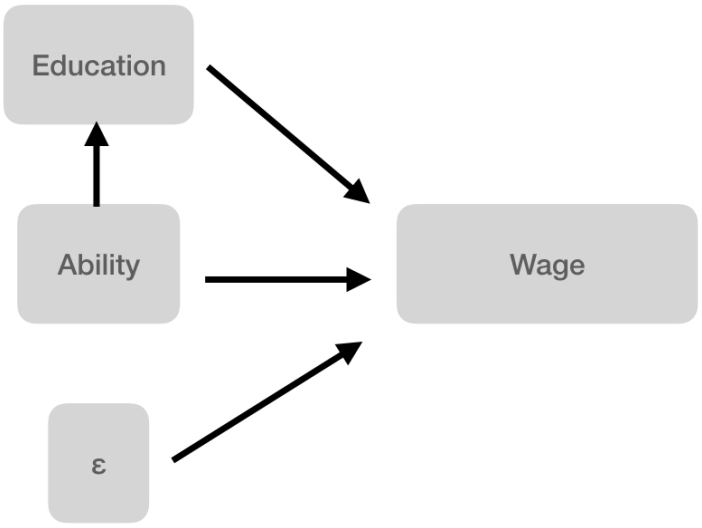
- What specific regressions do β_2 and γ_1 come from?

11.5 R Exercise

Omitted Variable Bias in R

The file `htv.RData` contains data from the 1991 National Longitudinal Survey of Youth, provided by Wooldridge. All people in the sample are males age 26 to 34. The data is interesting here, because it includes education, stored in the variable `educ`, and also a score on an ability test, stored in the variable `abil`.

Assume that the true model is,



Questions:

1- Are we able to *directly* measure ability? If so, how would you propose to measure it?

2- If not, what *do* we measure and how is this measurement related to ability? And there is a lot of evidence to suggest that standardized tests are not a very good proxy. But for now, let's pretend that we really are measuring ability.

3- Using R, estimate (a) the true model, and (b) the regression of ability on education.

- Write down the expression for what omitted variable bias would be if you couldn't measure ability.
- Add this omitted variable bias to the coefficient for education to see what it would be.

4- Now evaluate your previous result by fitting the model,

$$wage = \alpha_0 + \alpha_1 educ + w$$

- Does the coefficient for the relationship between education and wages match what you estimated earlier?
- Why or why not?

5- Reflect on your results:

- What does the direction of omitted variable bias suggest about OLS estimates of returns to education?
- What does this suggest about the reported statistical significance of education?

11.6 Discussion

The Direction of Omitted Variable Bias

- For each regression, estimate whether omitted variable bias is towards zero or away from zero.

Regression Output	Omitted Variable
$\widehat{grade} = 72.1 + 0.4\ attendance$	$time\_studying$

Regression Output	Omitted Variable
$\widehat{lifespan} = 87.4 - 1.2\ cigarettes$	*exercise*
$\widehat{lifespan} = 87.4 - 1.2\ cigarettes$	*time_socializing*
$\widehat{wage} = 14.0 + 2.1\ grad\_education$	*experience*
$\widehat{wage} = 14.0 + 2.1\ grad\_education$	desire to effect *social_good*
$\widehat{literacy} = 54 + 12\ network\_access$	*wealth*

Chapter 12

The Classical Linear Model

12.1 Learning Objectives

1.
2.
3.

12.2 Class Announcements

- Lab 2 Deliverable and Dates
  - Research Proposal (Today)
  - Within-Team Review (Today)
  - Final Report (Week 14)
  - Final Presentation (Week 14)

12.3 Roadmap

Rearview Mirror

- Statisticians create a population model to represent the world.
- The BLP is a useful summary for a relationship among random variables.
- OLS regression is an estimator for the Best Linear Predictor (BLP).
- For a large sample, we only need two mild assumptions to work with OLS
  - To know coefficients are consistent
  - To have valid standard errors, hypothesis tests

Today

- The Classical Linear Model (CLM) allows us to apply regression to smaller samples.
- The CLM requires more to be true of the data generating process, to make coefficients, standard errors, and tests *meaningful* in small samples.
- Understanding if the data meets these requirements (often called assumptions) requires considerable care.

Looking Ahead

- The CLM – and the methods that we use to evaluate the CLM – are the basis of advanced models (*inter alia* time-series)
- (Week 13) In a regression studies (and other studies), false discovery is a widespread problem. Understanding its causes can make you a better member of the scientific community.

12.4 The Classical Linear Model

12.4.1 Comparing the Large Sample Model and the CLM

Part I

- We say that in small samples, more needs be true of our data for OLS regression to "work."
  - What do we mean when we say "work"?
    - If our goals are descriptive, how is a "working" estimator useful?
    - If our goals are explanatory, how is a "working" estimator useful?
    - If our goals are predictive, are the requirements the same?

Part II

- Suppose that you're interested in understanding how subsidized school meals benefit under-resourced students.
  - Using the tools from 201, refine this question to a data science question.
  - Suppose that to answer the question you have identified, there are two data sources:
    - Individual-level data about income, nutrition and test scores, self-reported by individual families who have opted in to the study.

        - Government data about school district characteristics, including district-level college achievement; county-level home prices, and state-level tax receipts.
    - What are the tradeoffs to these different sources?

Part III

- Suppose you use individual-level data (you have a large sample).
    - Which of the large-sample assumptions do you expect are valid, and which are problematic?
- Say you use school-district-level data (you have a small sample).
    - Which of the CLM assumptions do you expect are valid, and which do you expect are most problematic?
- Which dataset do you think will give you more precise estimates?

12.5 Problems with the CLM Requirements

- There are five requirements for the CLM
    1. IID Sampling
    2. Linear Conditional Expectation
    3. No Perfect Collinearity
    4. Homoskedastic Errors
    5. Normally Distributed Errors
- For each of these requirements:
    - Identify one **concrete** way that the data might not satisfy the requirement.
    - Identify what the consequence of failing to satisfy the requirement would be.
    - Identify a path forward to satisfy the requirement.

12.6 R Exercise

```
library(tidyverse)
library(wooldridge)
library(car)
library(lmtest)
library(sandwich)
library(stargazer)
```

If you haven't used the `mtcars` dataset, you haven't been through an intro applied stats class!

In this analysis, we will use the mtcars dataset which is a dataset that was extracted from the 1974 Motor Trend US magazine, and comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973-74 models). The dataset is automatically available when you start R. For more information about the dataset, use the R command: help(mtcars)

```
data(mtcars)

glimpse(mtcars)
```

```
## Rows: 32
## Columns: 11
## $ mpg  <dbl> 21.0, 21.0, 22.8, 21.4, 18.7, 18.1, 14.3, 24.4, 22.8, 19.2, 17.8,~
## $ cyl  <dbl> 6, 6, 4, 6, 8, 6, 8, 4, 4, 6, 6, 8, 8, 8, 8, 8, 8, 4, 4, 4, 4, 8,~
## $ disp <dbl> 160.0, 160.0, 108.0, 258.0, 360.0, 225.0, 360.0, 146.7, 140.8, 16~
## $ hp   <dbl> 110, 110, 93, 110, 175, 105, 245, 62, 95, 123, 123, 180, 180, 180~
## $ drat <dbl> 3.90, 3.90, 3.85, 3.08, 3.15, 2.76, 3.21, 3.69, 3.92, 3.92, 3.92,~
## $ wt   <dbl> 2.620, 2.875, 2.320, 3.215, 3.440, 3.460, 3.570, 3.190, 3.150, 3.~
## $ qsec <dbl> 16.46, 17.02, 18.61, 19.44, 17.02, 20.22, 15.84, 20.00, 22.90, 18~
## $ vs   <dbl> 0, 0, 1, 1, 0, 1, 0, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 0,~
## $ am   <dbl> 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 0, 0,~
## $ gear <dbl> 4, 4, 4, 3, 3, 3, 3, 4, 4, 4, 4, 3, 3, 3, 3, 3, 3, 4, 4, 4, 3, 3,~
## $ carb <dbl> 4, 4, 1, 1, 2, 1, 4, 2, 2, 4, 4, 3, 3, 3, 4, 4, 4, 1, 2, 1, 1, 2,~
```

Questions:

1. Using the mtcars data, run a multiple linear regression to find the relationship between displacement (`disp`), gross horsepower (`hp`), weight (`wt`), and rear axle ratio (`drat`) on the miles per gallon (`mpg`).
2. For **each** of the following CLM assumptions, assess whether the assumption holds. Where possible, demonstrate multiple ways of assessing an assumption. When an assumption appears violated, state what steps you would take in response.
    - I.I.D. data
    - Linear conditional expectation
    - No perfect collinearity
    - Homoskedastic errors
    - Normally distributed errors
3. In addition to the above, assess to what extent (imperfect) collinearity is affecting your inference.

4. Interpret the coefficient on horsepower.

5. Perform a hypothesis test to assess whether rear axle ratio has an effect on mpg. What assumptions need to be true for this hypothesis test to be informative? Are they?

6. Choose variable transformations (if any) for each variable, and try to better meet the assumptions of the CLM (which also maintaining the readability of your model).

7. (As time allows) report the results of both models in a nicely formatted regression table.

Chapter 13

Reproducible Research

13.1 Learning Objectives

1.
2.
3.

13.2 Class Announcements

13.3 Roadmap

Rearview Mirror

Today

Looking Ahead

13.4 What data science hopes to accomplish

- As a data scientist, our goal is to learn about the world:
  - *Theorists* and *theologians* build systems of explanations that are consistent with themselves
  - *Analysts* build systems of explanations that are consistent with the past
  - *Scientists* build systems of explanations that usefully predict events, **or data**, that hasn't yet been seen

13.5 Learning from Data

- As a data scientist, the way we learn about the world is through the streams of data that **real world** events produce
  - Machine processes
  - Political outcomes
  - Customer actions
- The watershed moment in our field has been the profusion of data available, from many places, that is richer than at any other point in our past.
  - In 251, and 266 we place structure on data series like audio, video and text that are *transcendently* rich
  - In 261 we bring together flows of data that are generated at massive scales
  - In 209 we ask, "How can we take data, and produce a *new* form of it that is most effectively understood by the human visual and interactive mind?

13.6 Data Science and Statistics

- So why statistics?
- And why the way we've chosen to approach statistics in 203?

13.7 Why Statistics?: A Closing Argument for Statistics

- Business, policy, education and medical decisions are made *by humans* based on data
- A central task when we observe some pattern in data is to **infer** whether the pattern will occur in some novel context
- Statistics, as we practice it in 203, allows us to characterize:
  - What we have seen
  - What we *could have seen*
  - Whether any guarantees exist about what we have seen
  - What we can infer about the population
- So that we can either describe, explain or predict behavior.

13.8 Course Goals

13.8.1 Course Section III: Purpose-Driven Models

- Statistical models are unknowing transformations of data
  - Because they're built on the foundation of probability, we have certain guarantees what a model "says"
  - Because they're unknowing, the models themselves know-not what they say.
- As the data scientist, bring them alive to achieve our modeling goals
- In Lab 2 we have expanded our ability to parse the world using regression, built a model that accomplishes our goals, and done so in a way that brings the ability to test under a *"null"* scenario
  - **Key insight**: regression is little more than conditional averages

13.8.2 Course Section II: Sampling Theory and Testing

- Under **very** general assumptions, sample averages follow a predictable, known, distribution – the *Gaussian distribution*
- This is true, even when the underlying probability distribution is *very* complex, or unknown!
- Due to this common distribution, we can produce reliable, general tests!
- In Lab 1 we computed simple statistics, and used guarantees from sampling theory to **test** whether these differences were likely to arise under a *"null"* scenario

13.8.3 Course Section I: Probability Theory

- Probability theory
  - Underlies modeling and regression (Part III);
  - Underlies sampling, inference, and testing (Part II)
  - **Every** model built in **every** corner of data science

We can:

- Model the complex world that we live in using probability theory;
- Move from a probability density function that is defined in terms of a single variable, into a function that is defined in terms of many variables
- Compute useful summaries – i.e. the BLP, expected value, and covariance – even with *highly* complex probability density functions.

13.8.4 Statistics as a Foundation for MIDS

- In w203, we hope to have laid a foundation in probability that can be used not only in statistical applications, but also in every other machine learning application that are likely to ever encounter

13.9 Reproducibility Discussion

Green Jelly Beans

What went wrong here?

13.9.1 Discussion

Status Update You have a dataset of the number of Facebook status updates by day of the week. You run 7 different t-tests, one for posts on Monday (versus all other days), or for Tuesday (versus all other days), etc. Only the test for Sunday is significant, with a p-value of .045, so you throw out the other tests.

Should you conclude that Sunday has a significant effect on number of posts? (How can you address this situation responsibly when you publish your results?)

Such Update As before, you have a dataset of the number of Facebook status updates by day of the week. You do a little EDA and notice that Sunday seems to have more "status updates" than all other days, so you recode your "day of the week" variable into a binary one: Sunday = 1, All other days = 0. You run a t-test and get a p-value of .045. Should you conclude that Sunday has a significant effect on number of posts?

Sunday Funday Suppose researcher A tests if Monday has an effect (versus all other days), Researcher B tests Tuesday (versus all other days), and so forth. Only Researcher G, who tests Sunday finds a significant effect with a p-value of .045. Only Researcher G gets to publish her work. If you read the paper, should you conclude that Sunday has a significant effect on number of posts?

Sunday Repentence What if researcher G above is a sociologist that chooses to measure the effect of Sunday based on years of observing the way people behave on weekends? Researcher G is not interested in the other tests, because Sunday is the interesting day from her perspective, and she wouldn't expect any of the other tests to be significant.

Decreasing Effect Sizes Many observers have noted that as studies yielding statistically significant results are repeated, estimated effect sizes go down and often become insignificant. Why is this the case?

Chapter 14

Maximum Likelihood Estimation

Figure 14.1: salvation mountain

14.1 Learning Objectives

1.
2.
3.

14.2 Class Announcements

14.3 Roadmap

Rearview Mirror: What We've Seen

- **WLLN**: $\lim_{n\to\infty} \overline{X}_n \stackrel{p}{=} E[X]$
- **CLT** $\lim_{n\to\infty} \bar{X}_n \stackrel{d}{=} N(E[X], \text{Var}[X])$

Today

- Use maximum likelihood to generate a good guess for model parameters;
- Use a confidence interval to indicate a range of plausible parameter values

14.4 What is a model?

- A data science model is:
  - A representation of the world built from random variables
  - FOIS: "agnostic" models place minimal restrictions on joint distribution
  - Parametric models (i.e. MLE) are models based on a family of distributions.
  - $f_{Y|X}(y|\mathbf{x}) \sim g(y, \mathbf{x}; \theta)$

14.5 Estimation

- We have the tools to use data to infer information about the (joint) distribution
- Because the joint distribution is complicated, we'll usually estimate simpler summaries of the joint distribution – e.g. $E[X]$, $V[X]$, $E[Y|X]$, $Cov[X, Y]$

- There are a number of techniques that you can use to develop an estimator for a parameter. These techniques vary in terms of the principle used to arrive at the estimator and the strength of the assumptions needed to support it.
- However, all of these estimators are statistics meaning they are functions of the data $\{X_i\}_{i=1}^n$

14.6 Discussion of Maximum Likelihood Estimation

1. What is the goal of estimating a parameter? Why is this something that we are interested in as data scientists?
2. In your own words, describe how the method of maximum likelihood is used to estimate the unknown parameters.
3. Why does a likelihood function have a Π (product operator) within it?
4. Is it possible to estimate using maximum likelihood without writing down a model for the data?
5. What happens if your model for the data is wrong? Are your estimates for the parameters "incorrect"? Or, are they "correct" within the context of the model that you've written down?

14.7 Optimization in R

- The method of maximum likelihood requires an optimization routine.
- For a few very simple probability models, a closed-form solution exists and the MLE can be derived by hand. (This is also *potentially* the case for OLS regression.)
- But, instead lets use some machine learning to find the estimates that maximize the likelihood function.
- There are many optimizers (e.g. `optimize`, and `optim`). `optimize` is the simplest to use, but only works in one dimension.

14.7.1 Optimization Example: Optimum Price

- Suppose that a firm's profit from selling a product is related to price, p, and cost, c, as follows:

$$\text{profit} = (p - p^2) - c + 100$$

1. Explain how you would use calculus to find the maximizing price. Assume that cost is fixed.

2. What is the firms revenue as `p=0, cost = 2`? What is it at `p=10, cost = 2`?

3. Create a plot with the following characteristics:

   - On the x-axis is a sequence (`seq()`) of prices from [0, 10].
   - On the y-axis is the revenue as a function of those prices. Hold cost constant at `c=2`.

   - What does the best price seem to be?

4. Solve this numerically in *R*, using the *optimize()* function.

   - Take note: using the default arguments, will `optimize` try to find a maximum or a minimum?
   - Check into the help documentation.

```
profit <- function(p, c) {
  r = (p - p^2) - c + 100
  return(r)
  }
```

```
profit(p=2, c=2)
```

```
## [1] 96
```

```
best_price <- optimize(
  profit,                    # profit is the function
  lower = 0, upper = 1000,   # this is the low and high we consider
  c = 2,                     # here, we're passing cost into profit
  maximum = TRUE)            # we'd like to maximize, not minimize
best_price
```

```
## $maximum
## [1] 0.5
##
## $objective
## [1] 98.25
```

14.8 MLE for Poisson Random Variables

- Suppose we use a camera to record an intersection for a particular length of time, and we write down the number of cars accidents in that interval.
- This process can be modeled by a *Poisson* random variable (now we are non-agnostic), that has a well-known probability mass function given by,

$$f(x; \lambda) = \frac{\lambda^x e^{-\lambda}}{x!}$$

Here is an example of a string of outcomes generated by a Poisson RV, with parameter $\lambda = 2$.

```
rpois(n = 10, lambda = 2)
```

```
##  [1] 3 4 4 2 3 1 0 2 2 1
```

14.8.1 MLE for Poisson Random Variables: Data

- Suppose that we conduct an iid sample, and gather the following number of accidents. (It is a busy street!)

```
data <- c(
  2, 6, 2, 1, 3, 3, 4, 4, 24, 1, 5, 4, 5, 1,  2, 2, 5, 2, 1, 5,
  2, 1, 2, 9, 9, 1, 3, 2, 1,  1, 3, 1, 3, 2,  2, 4, 1, 1, 5, 3,
  3, 2, 2, 1, 1, 1, 5, 1, 3,  1, 1, 1, 1, 2,  2, 4, 2, 1, 2, 2,
  3, 1, 2, 6, 2, 2, 3, 2, 3,  5, 1, 3, 2, 5,  2, 1, 3, 2, 1, 2,
  4, 2, 6, 1, 2, 2, 3, 5, 2,  1, 4, 2, 2, 1,  3, 2, 2, 4, 1, 1,
  1, 1, 2, 3, 5, 1, 2, 2, 3,  1, 4, 1, 3, 2,  2, 2, 2, 2, 2, 3,
  3, 1, 1, 2, 2, 4, 1, 5, 2,  7, 5, 2, 3, 2,  5, 3, 1, 2, 1, 1,
  2, 3, 1, 5, 3, 4, 6, 3, 3,  2, 2, 1, 2, 2,  4, 2, 3, 4, 3, 1,
  6, 3, 1, 2, 3, 2, 2, 3, 1,  1, 1, 1, 1, 10, 3, 2, 1, 1, 3, 2,
  2, 3, 1, 1, 2, 2, 2, 4, 2,  2, 3, 3, 6, 1,  3, 2, 3, 2, 2, 2
  )

table(data)
```

```
## data
##  1  2  3  4  5  6  7  9 10 24
## 54 69 38 14 14  6  1  2  1  1
```

14.8.2 MLE Estimation

- Use the data that is stored in `data`, together with a Poisson model to estimate the λ values that produce the "good" model from the Poisson family.

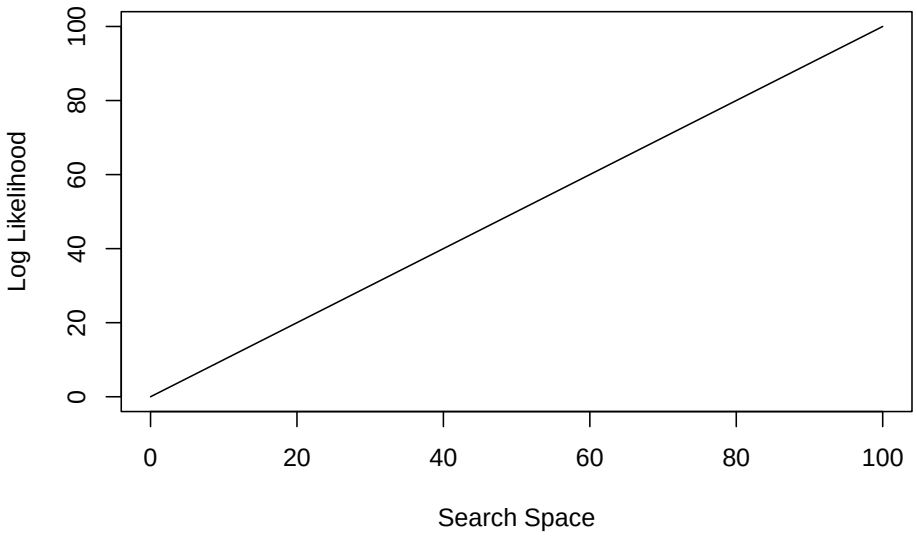
- That is, use MLE to estimate λ.

- Here is your work flow:

1. Define your random variables.
2. Write down the likelihood function for a sample of data that is generated by a *Poisson* process.
3. To make the math easier, take the log of this likelihood function.
4. Optimize this log-likelihood using calculus – what is the value of λ that results? Compute this value, given the data that you have.
5. Maximize this log-likelihood numerically, and report the value for λ that produces the highest likelihood of seeing this data.
6. Comment on your answers from parts 4 and 5. Are you surprised or not by what you see?

```
poisson_ll <- function(data, lambda) {
  ## fill this in:
  lambda # this is a placeholder, change this
}
```

```
search_space <- seq(0,100, by = 0.1)
plot(
  x = search_space, xlab = 'Search Space',
  y = poisson_ll(data=data, lambda=search_space), ylab = 'Log Likelihood',
  type = 'l'
)
```




```
# optimize(poisson_ll, lower = 0, upper = 100, data = data, maximum = TRUE)
```

14.9 Confidence Intervals

This exercise is meant to demonstrate what the confidence level in a confidence interval represents. We will assume a standard normal population distribution and simulate what happens when we draw a sample and compute a confidence interval.

Your task is to complete the following function so that it,

1) simulates and storesn draws from a standard normal distribution
2) based on those draws, computes a valid confidence interval with confidence level α, a parameter that you pass to the function.

Your function should return a vector of length 2, containing the lower bound and upper bound of the confidence interval.

$$CI_\alpha = \overline{X} \pm t_{\alpha/2} \cdot \frac{s}{\sqrt{n}}$$

where:

- CI_α is the confidence interval that you're seeking to produce
- $\overline{X}$ is the sample average,

- $t_{\alpha/2}$ is your critical value (accessible through `qt`),
- and s is your sample standard deviation. Notice that you'll need each of these pieces in the code that you're about to write.

```
sim_conf_int <- function(n, alpha) {
  # Fill in your code to:
  # 1. simulate n draws from a standard normal dist.
  # 2. compute a confidence interval with confidence level alpha

  sample_draws <- 'fill this in'
  sample_mean  <- 'fill this in'
  sample_sd    <- 'fill this in'

  critical_t <- 'fill this in'

  ci_95 <- 'fill this in'

  return(ci_95)
}

sim_conf_int(n = 100, alpha = 0.25)
```

```
## [1] "fill this in"
```

When your function is complete, you can use the following code to run your function 100 times and plot the results.

```
many_confidence_intervals <- function(num_simulations, n, alpha) {
  ## args:
  ##  - num_simulations: the number of simulated confidence intervals
  ##  - n: the number of observations in each simulation that will pass
  ##           into your `sim_conf_int` function
  ##  - alpha: the confidence interval that you will pass into
  ##           your `sim_conf_int` function

  results <- NULL
  for(i in 1:num_simulations) {
    interval = sim_conf_int(n, alpha)
    results = rbind(results, c(interval[1], interval[2], interval[1]<0 & interval[2]>0)
  }
  resultsdf = data.frame(results)
  names(resultsdf) = c("low", "high", "captured")
  return(resultsdf)
}
```

```
n = 20
confidence_intervals = many_confidence_intervals(100, n, .05)
```

```
plot_many_confidence_intervals <- function(c) {
  plot(NULL, type = "n",
       xlim = c(1,100), xlab = 'Trial',
       ylim = c(min(c$low), max(c$high)), ylab=expression(mu),pch=19)

  abline(h = 0, col = 'gray')
  abline(h = qt(0.975, n-1)/sqrt(n), lty = 2, col = 'gray')
  abline(h = qt(0.025, n-1)/sqrt(n), lty = 2, col = 'gray')

  points(c$high, col = 2+c$captured, pch = 20)
  points(c$low,  col = 2+c$captured, pch = 20)
  for(i in 1:nrow(c)) {
    lines(c(i,i), c(c$low[i],c$high[i]), col = 2+c$captured[i], pch = 19)
  }

  title(expression(paste("Simulation of t-Confidence Intervals for ", mu,
                         " with Sample Size 20")))

  legend(0,-.65, legend = c(expression(paste(mu," Captured")),
                            expression(paste(mu," Not Captured"))), fill = c(3,2))
  }
```

```
# plot_many_confidence_intervals(confidence_intervals)
```

1. How many of the simulated confidence intervals contain the true mean, zero?

2. Suppose you run a single study. Based on what you've just written above, why is it incorrect to say that, "There is a 95% probability that the true mean is inside this (single) confidence interval"?

14.10 Maximum Likelihood Example: Printers

Part I

Suppose that you've got a particular sequence of values: $1, 0, 0, 1, 0, 1, 1, 1, 1, 1$ that indicate whether a printer any particular time you try to print.

You have data from the last 10 times you tried.

Question:

- What is the probability (p) that the printer jams on the next print job?

Figure 14.2: bbc, office space

Part II

The data resembles draws from a Bernoulli distribution.

However, even if we want to model this as a Bernoulli distribution, we do not know the value of the parameter, p.

1- Define your random variable.

2- Write down the likelihood function

3- If it will make the math easier, log the likelihood function.

4- *Path 1:* Maximize the likelihood using calculus

5- *Path 2:* Maximize using numeric methods.