Statistics for Data Science

D. Alex Hughes, Paul Laskowski & The 203 Teaching Team[1]

2024-10-08

[1]UC Berkeley, School of Information

Contents

Live Session

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.

```
# library(mids203)
```

Chapter 1

Probability Spaces

```
source('./src/blank_lines.R')
```

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.

Figure 1.1: probability, the final frontier

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 this point in the course, there is so much that is before us! As we settle in to study for the semester, it is useful to have a point of view of where we're trying to go, and what we are going to see along the way.

Allow a justification by analogy:

> Suppose that you decide that you would like to be a chef – all of the time watching cooking shows has revealed to you that this is your life's true calling – and so you enroll in a culinary program.
>
> One does not begin such a program by baking croissants and souffle. They begin the program with knife skills, breaking down ingredients and the basic techniques that build up to produce someone who is not a *cook*, but a *chef* – someone who can combine ingredients and techniques to produce novel ideas.
>
> At the same time, however, one has not gone to school just to become a cucumber slicer. The knife skills are instrumental to the eventual goal – of being a chef – but not the goal itself.

At the beginning of the program, we're teaching these core, fundamental skills. How to read and reason with mathematical objects, how to use conditional probability with the goal of producing a model, and eventually, **eventually** to create novel work as a data scientist.

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

1.2.1 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.

1.2.2 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

1.2.3 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

1.2.4 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

1.2.5 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, computer science, and information.

1.3.2 What does a statistician look 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. We each have been fortunate to be able to study, but we want to acknowledge that the education system in the US has systematically benefited the hegemonic groups and marginalized others voices.

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. We want to see a field of data scientists who embrace each others voices, and respects people for the identies that they hold.

- 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 [Breakout One]

In a breakout room of between three and four students introduce yourself!

Breakout One. A *name story* is the unique, and individual story that describes how you came to have the name that you do. While there may be many people are called the same thing, each of their name stories is unique.

Please share: *What is your name story?*

1.5 Student Introductions [Breakout Two]

In the same breakout room:

Breakout Two. Like our names, the reasons that we joined this program, our goals and our histories are different.

Please share: *What is your data science story? How did you wind up here, in this room today?*

1.6 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.

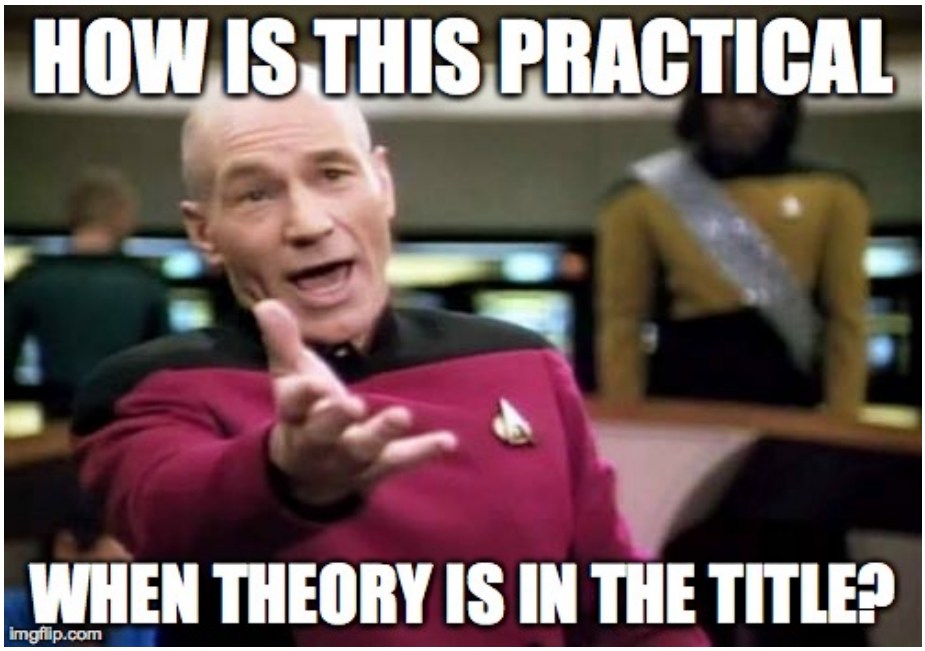


Figure 1.2: told you this would be spacey

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.

1.7 Axiomatic Probability

The book makes a point of defining our axioms of probability, calling them them

Definition 1.1. *Kolmogorov Axioms*

Let Ω be a sample space, S be an event space, and P be a probability measure. Then, (Ω, S, P) is a *probability space* if it satisfies the following:

- Non-negativity: $\forall A \in S, P(A) \geq 0$, where $P(A)$ is finite and real.
- Unitarity: $P(\Omega) = 1$.
- Countable additivity: if $A_1, A_2, A_3, \cdots \in S$ are pairwise disjoint, then

$$P(A_1 \cup A_2 \cup A_3 \cup \dots) = P(A_1) + P(A_2) + P(A_3) = \sum_i P(A_i)$$

There is a lot going on in this definition!

First things first, these are the **axioms of probability** (read aloud in the booming voice of a god).

This means that these are things that we begin from, sort of the foundational principles of the entire system of reasoning that we are going to use. In the style of argument that we're going to make, these are things that are sort of off-limits to question. Instead, these serve as the grounding assumptions, and we see what happens as we flow forward from these statements.

Second, and importantly, from these axioms there are a *very large* set of things that we can build. The first set of things that we will build are probability statements about atomic outcomes (Theorem 1.1.4 in the book), and collections of events. But, these statements, are not the only thing that we're limited to. We can also build *Frequentist Statistics*, and *Bayesian Statistics* and *Language Models*.

In many ways, these axioms are the fundamental particles that hold our system of probabilistic reasoning together. These are to probability what the *fermions* and and *bosons* are to physics.

1.8 Definition vs. Theorem

What is the difference between a definition and a theorem? On pages 10 and 11 of the textbook, there is a rapid fire collection of pink boxes. We reproduce them here (notice that they may have different index numbers than the book – this live session book autoindexes and we're not including every theorem and definition in this live session discussion guide).

Definition 1.2. *Conditional Probability* For $A, B \in S$ with $P(B) > 0$, the *conditional probablity* of A given B is

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

Theorem 1.1. Multiplicative Law of Probability *For $A, B \in S$ with $P(B) > 0$,*

$$P(A|B)P(B) = P(A \cap B)$$

Theorem 1.2. Baye's Rule *For $A, B \in S$ with $P(A) > 0$ and $P(B) > 0$,*

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

- What would happen to the statement of the *Multiplicative Law of Probability* if we did not have the definition of *Conditional Probability*?
- How does one get from the definition, to the law?
- Can one get to *Baye's Rule* wihtout using the *Multiplicative Law of Probability*?

1.9 Working with a Sample Space

As a way to begin lets define terms that we will use for the next activities.

Group Discussion Question

- What is the definition of a sample space?
- What is the definition of an event?
- How are sample spaces, and event spaces related?

1.9.1 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.

1.9.2 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?

1.10 Independence

The book provides a (characteristically) terse statement of what it means for two events to be independent of one another.

Definition 1.3. *Independence of Events* Events $A, B \in S$ are *independent* if

$$P(A \cap B) = P(A)P(B)$$

.

In your own words:

- What does it mean for two events to be independent of one another?
- How do you **know** if two events are independent of one another?
- How do you **test** if two events are independent of one another?

Try using this idea of independent in two places:

1. Suppose that you are creating a model to predict an outcome. Further, suppose that two events A and B are independent of one another. *Can you use B to predict A?*
2. If two events, A and B are independent, then what happens if you work through a statement of conditional probability, $P(A|B)$?

1.11 A practice problem

The last task for us to complete today is working through a practice problem on the course practice problem website. Please, click the link below, and follow us over to the the course's practice problem website.

link here

1.12 Student Tasks to Complete

Before next live session, please complete the homework that builds on this unit. There are two parts, an *applied* and a *proof* part. You can submit these homework as many times as you like before the due date (you will not receive feedback), and you can access this homework through bCourses.

The *applied* homework will be marked either `Correct` or `Incorrect` without partial credit applied. These are meant to be problems that you solve, and that have a single straightforward solution concept. The *proof* homework will be marked for partial credit (out of three points) that evaluates your argument for your solution concept.

Chapter 2

Defining Random Variables

```
## -- Attaching core tidyverse packages ------------------------------------------------------
## v dplyr     1.1.3     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.4     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.0
## v purrr     1.0.2
## -- Conflicts ------------------------------------------------------------------------------
## x purrr::%||%()   masks base::%||%()
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become
```

2.1 Learning Objectives

At the end of this week's course of study (which includes the async, sync, and homework) students should be able to

1. **Remember** that random variable are neither random, or variables, but instead that they are a foundational object that we can use to reason about a world.
2. **Understand** that the intuition developed by the use of set-theory probability maps into the more expressive space of random variables
3. **Apply** the appropriate mathematical transformations to move between joint, marginal, and conditional distributions.

This week's materials are theoretical tooling to build toward one of the first notable results of the course, **conditional probability**. This is the idea that, if we know that one event has occurred, we can make a conditional statement about the probability distribution for another, dependent distribution.

Figure 2.1: yosemite valley

2.2 Introduction to the Materirals

From the axioms of probability, it is possible to build a whole, expressive modeling system (that need not be grounded **at all** in the minutia of the world). With this probability model in place, we can describe how frequently events in the random variable will occur. When variable are dependent upon each other, we can utilize information that is encoded in this dependence in order to make predictions that are *closer to the truth* than predictions made without this information.

There is both a beauty and a tragedy when reasoning about random variables: we describe random variables using their joint density function.

The **beauty** is that by reasoning with such general objects – the definitions that we create, and the theorems that we derive in this section of the course – produce guarantees that hold in every case, no matter the function that stands in for the joint density function. We will compute several examples of *specific* functions to provide a chance to reason about these objects and how they "work".

The **tragedy** is that in the "real world", the world where we are going to eventually going to train and deploy our models, we are never provided with this joint density function. Perhaps this is the creation myth for probability theory: in a perfect world, we can produce a perfect result. But, in the "fallen" world of data, we will only be able to produce approximations.

2.3 Class Announcements

Homework

1. You should have turned in your first homework. The solution set for this homework is scheduled to be released to you in two days. The solution set contains a full explanation of how we solved the questions posed to you. You can expect that feedback for this homework will be released back to you within seven days.
2. You can start working on your second homework when we are out of this class.

Study Groups

It is a **very** good idea for you to create a recurring time to work with a set of your classmates. Working together will help you solve questions more effectively, quickly, and will also help you to learn how to communicate what you do and do not understand about a problem to a group of collaborating data scientists. And, working together with a group will help you to find people who share data science interests with you.

Course Resources

There are several resources to support your learning. A learning object last week was that you would be introduced to each of these systems. Please continue to make sure that you have access to the:

- Library VPN to read all of the scholarly content in the known universe, including the course textbook.
- Course LMS Page

2.4 Using Definitions of Random Variables

2.4.1 Random Varaible

What is a random variable? Does this definition help you?

Definition 2.1 (Random Variable)**.** A random variable is a function $X : \Omega \to \mathbb{R}$, such that $\forall r \in \mathbb{R}, \{\omega \in \Omega : X(\omega) \leq r\} \in S$.

Someone, please, read that without using a single "omega", $\mathbb{R}$, or other jargon terminology. Instead, someone read this aloud and tell us what each of the concepts mean.

The goal of writing with math symbols like this is to be *absolutely* clear what concepts the author does and does not mean to invoke when they write a definition or a theorem. In a very real sense, this is a language that has specific meaning attached to specific symbols; there is a correspondence between the mathematical language and each of our home languages, but exactly what the relationship is needs to be defined into each student's home language.

- What are the key things that random variables allow you to accomplish?
  - Suppose that you were going to try to make a model that predicts the probability of winning "big money" on a slot machine. Big money might be that you get :cherries: :cherries: :cherries:. Can you do *math* with :cherries:?
  - Suppose that you wanted to build a chatbort that uses a language model so that you don't have to do your homework anymore. How would you go about it?
  - Suppose you want to direct class support to students in 203, but their grades are scored `[A, A-, ..., ]` and features include prior statistics classes grades, also scored `A, A-, ...]`

2.5 Pieces of a Random Variable

Definition 2.2 (Random Variable, Suite)**.** A random variable is a function $X : \Omega \to \mathbb{R}$, such that $\forall r \in \mathbb{R}, \{\omega \in \Omega\} : X(\omega) \leq r\} \in S$.

There are two key pieces that must exist for every random variable. What are these pieces? The first of these pieces is provided to us in **Definition 1.2.1** *Random Variable* (on page 16). The second is provided to us in **Definition 1.2.5** *Probability Mass Function* (on page 18).

1.
2.

Suppose that a random variable is simple and discrete. For concreteness, you could think of this random variable as the answer to the question, "Is the grass wet outside?".

1. What is the sample space?
2. What is a sensible function that you might use to map from the sample space to real values?
3. What is a insensible function that you might use to map from the sample space to real values? (A student well-seasoned in Maths might use (and define for the rest of the class) the concept of a *bijective function*).
4. If you simply had the values that the random variable function maps to are you guaranteed to be able to describe the entire sample space? Why or why not?
5. How would you go about determining the probability mass function for this random variable?

2.5.1 Functions of Functions

Why do we say that random variables are functions? Is there some useful property of these being functions rather than any other quantity? What else *could* they be if not a function?

What about a function of a random variable, which is a function of a function.

Definition 2.3 (Function of a Random Variable)**.** Let $g : U \to \mathbb{R}$ be some function, where $X(\Omega) \subset U \subset \mathbb{R}$. Then, if $g \circ X : \Omega \to \mathbb{R}$ is a random variable, we say that g is a *function* of X and write $g(X)$ to denote the random variable $g \circ X$.

If a random variable is a function from the real world, or the sample space, or the outcome space to a real number, then what does it mean to define a function of a random variable?

- At what point does this function work? Does this function change the sample space that is possible to observe? Or, does this function change the real-number that each outcome points to?

Example 2.1 (MNIST)**.** Suppose that you are doing some image processing work. To keep things simple, that you are doing image classification in the style of the MNIST dataset.

- Can someone describe what this task is trying to accomplish?

- Has anyone done work like this?

However, suppose that rather than having good clean indicators for whether a pixel is on or off, instead you have weak indicators – there's a lot of grey. A lot of the cells are marked in the range $0.2 - 0.3$.

1. How might creating a function that re-maps this grey into more extreme values help your model?
2. Is it possible to "blur" events that are in the outcome space? Does this "blurring" meet the requirements of a function of a random variable, as provided above?

2.5.2 Probability Density Functions and Cumulative Distribution Functions

- What is a probability mass function?
- What do the **Kolmogorov Axioms** mean must be true about any probability mass function (*pmf*)?

Example 2.2 (Berkeley Drivers, No Survivors)**.** You should try driving in Berkeley some time. It is a **trip**! Without being deliberately ageist, the city is full of ageing hippies driving Subaru Outbacks and making what seem to be stochastic right-or-left turns to buy incense, pottery, or just sourdough bread.

Suppose that you are walking to campus, and you have to cross 10 crosswalks, each of which are spaced a block apart. Further, suppose that as you get closer to campus, there are fewer aging hippies, and therefore, there is decreasing risk that you're hit by a Subaru as you cross the street. Specifically, and fortunately for our math, the risk of being hit decreases linearly with each block that you cross.

Finally, campus provides you with the safety reports from last year, and reports that there were 120 student-Subaru incidents last year, out of 10,000 student-crosswalk crossings.

1. What is the *pmf* for the probability that you are involved in a student-Subaru incident as you walk across these 10 blocks? What sample space, Ω is appropriate to represent this scenario?
2. Suppose that you don't leave your house – this is a remote program after all! What is your cumulative probability of being involved in a student-subaru incident?

3. What is the cumulative probability *cmf* for the probability that you are involved in a student-Subaru incident?
4. Suppose that you live three blocks from campus, but your classmate lives five blocks from campus. What is the difference in the cumulative probability?
5. How would you describe the cumulative probability of being hit as you walk closer to campus? That is, suppose that you start 10 blocks away

from campus, and are walking to get closer. Is your cumulative probability of being hit on your way to campus increasing or decreasing as you get closer to campus?

6. How would you describe the cumulative probability of being hit as you walk **further** from campus? That is, suppose that you start on campus, and you're walking to a bar after classes. Is your cumulative probability of being hit on your way away from campus increasing or decreasing as you get further from campus?

2.6 Discrete & Continuous Random Variables

What, if anything is fundamentally different between discrete and continuous random variables? As a way of starting the conversation, consider the following cases:

- 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 $P(X = 3600) = 0$.
- Does one of these measures have more *information* in it than another?
  - How are measurement choices that we make as designers of information capture systems – i.e. the machine processes, human processes, or other processes that we are going to work with as data scientists – reflected in both the amount of information that is gathered, the type of information that is gathered, and the types of random variables that are manifest as a result?

2.7 Moving Between PDF and CDF

The book defines *pmf* and *cmf* first as a way of developing intuition and a way of reasoning about these concepts. It then moves to defining continuous density functions, which is many ways are easier to work with although they lack the means of reasoning about them intuitively. Continuous distributions are defined in the book, and more generally, in terms of the *cdf*, which is the cumulative distribution function. There are technical reasons for this choice of definition, some of which are signed in the footnotes on the page where the book presents it.

More importantly for this course, in **Definition 1.2.15** the book defines the relationship between *cdf* and *pdf* in the following way:

Definition 2.4 (Probability Density Function (PDF))**.** For a continuous random

variable X with CDF F, the *probability density function* of X is

$$f(x) = \left.\frac{dF(u)}{du}\right|_{u=x}, \forall x \in \mathbb{R}.$$

- How does this definition, which relates *pdf* and *cdf* by a means of differentiation and integration, fit with the ideas that we just developed in the context of walking to and from campus?

Example 2.3 (Working with a continuous pdf and cdf)**.** Suppose that you learn than a particular random variable, X has the following function that describes its *pdf*, $f_x(x) = \frac{1}{10}x$. Also, suppose that you know that the smallest value that is possible for this random variable to obtain is 0.

1. What is the CDF of X?
2. What is the maximum possible value that x can obtain? How did you develop this answer, using the Kolmogorov axioms of probability?
3. What is the cumulative probability of an outcome up to 0.5?
4. What is the probability of an outcome between 0.25 and 0.75? Produce an answer to this in two ways:
5. Using the *pdf*
6. Using the *cdf*

2.8 Joint Density

Working with a single random variable helps to develop our understanding of how to relate the different features of a *pdf* and a *cdf* through differentiation and integration. However, there's not really *that* much else that we can do; and, there is probably very little in our professional worlds that would look like a single random variable in isolation.

We really start to get to something useful when we consider joint density functions. Joint density functions describe the probability that *both* of two random variables. That is, if we are working with random variables X and Y, then the joint density function provides a probability statement for $P(X \cap Y)$.

In this course, we might typically write this joint density function as $f_{X,Y}(x, y) = f(\cdot)$ where $f(\cdot)$ is the actual function that represents the joint probability. The $f(\cdot)$ means, essentially, "some function" where we just have not designated the specifics of the function; you might think of this as a generic function.

2.8.1 Example: Uniform Joint Density

Suppose that we know that two variables, X and Y are jointly uniformly distributed within the the *support* $x \in [0, 4], y \in [0, 4]$. We have a requirement, imposed by the *Kolmogorov Axioms* that all probabilities must be non-zero, and that the total probability across the whole support must be one.

- Can you use these facts to determine answers to the following:
  - What kind of shape does this joint *pdf* have?
  - What is the specific function that describes this shape?
  - If you draw this shape on three axes, and X, and Y, and a $P(X, Y)$, what does this plot look like?
  - How do you get from the joint density function, to a marginal density function for X?
  - How do you get form the joint density function, to a marginal density function for Y?
  - How do you get from these marginal density functions of X and Y back to the joint density? Is this always possible?

2.8.2 Examples: Thinking Through Many Plots

An alumni of the MIDS program, and a former instructor of this course, Todd Young built this nifty tool that lets us consider several different joint probability functions.

As a class, lets consider a few of these PDFs, beginning with this "triangle" distribution.

```
knitr::include_app('http://www.statistics.wtf/PDF_Explorer/', height="1000px")
```

2.8.3 Triangle Math

After considering the intuition for the triangle distribution, do the following: Write down the function that accords with the figure that you're seeing above.[1]

- What is a full statement of the PDF of this image?
- What is the marginal distribution of X, $f_X(x)$?
- What is the marginal distribution of Y, $f_Y(y)$?
- Using the definition of independence, are X and Y independent of each other?
- What is the CDF of X, $F_X(x)$?

2.8.4 Saddle Sores

Suppose that you know that two random variables, X and Y are jointly distributed with the following *pdf*:

$$f_{X,Y}(x, y) = \begin{cases} a * x^2 * y^2 & 0 < x < 1, 0 < y < 1 \\ 0 & otherwise \end{cases}$$

[1] Notice, that in general, this kind of *curve fitting* isn't really a common data science task. Instead, this is just a learning task that lets the class assess their understanding of the definitions of random variables.

This joint pdf is similar to the pdf that you can visualize above, under the distribution called "saddle". The difference between this function and the image above is that the function bounds the with support of x and y on the range $[0, 1]$. This is to make the math easier for us in the next step.

- Can you use these facts to determine the following?
  - What value of a makes this a valid joint pdf?
  - What is the marginal pdf of x? That is, what is $f_x(x)$?
  - What is the conditional pdf of X given Y? That is, what is $f_{x|y}(x, y)$?
  - Given these facts, would you say that X and Y are dependent or independent?
  - If the support for this joint distribution were instead $[0, 4]$ (rather than $[0, 1]$), how would the shape of the distribution change?

2.9 Computing Different Distributions.

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

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

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)

2.10 Conditional Probability

Conditional probability is **incredible**. In fact, without exaggeration, almost **all** of data science is an exercise in making statements about conditional probability distributions. *Don't believe us?*

- What is the goal of a "customer churn" model or a conversion model?
- What is the goal of a language-completion model?
- What is the goal of flight-departures model?

If we possessed the whole information about a process; **if** we had the CDF that governed probability of occurrences, what kinds of statements would we be able to make? Would we even need data?

Using the distribution above, produce a statement of conditional probability, $f_{Y|X}(y|x)$.

2.11 Visualizing Distributions Via Simulation

To this point in the course, we have focused on concepts in "the population" with no reference to samples. This is on purpose! We want to develop the theory that defines the **best possible** predictor if we knew **everything** (if we know formula of the function that maps from $\omega \to \mathbb{R}$, and we know the probability of each $\omega \in \Omega$ then we know everything). Beginning in week 5 of the course, we will talk about "approximating" (which we will call estimating) this best possible predictor with a limited sample of data.

However, at this point, to help build your working understanding, or intuition, for what is happening, we are going to work on a way to *simulate* draws from a population. In some places, people might refer to these as *Monte Carlo* methods – this is because the method was developed by von Neumann & Ulam during World War II, and they needed a way to talk about it using a code name. They chose *Monte Carlo* after a famous casino in Monaco.

2.11.1 Example: The Uniform Distribution

> You: "Gosh. There sure are a lot of examples that use the uniform distribution. That must be a really important statistical distribution."
>
> Instructor: "Nah. Not really. We're just using the uniform a bunch so that we don't get too lost in doing math while we're working with these concepts."

We'll start with a simple uniform distribution, but then we'll make it a little more complex in a moment.

We can use R to simulate draws from a probability distribution function by providing it with the name of the distribution that we're considering, the support of that distribution, or other features of the distribution. In the case of the uniform, the entire distribution is can be described just from it support.

So, suppose that you had a uniform distribution that had positive probability on the range $[1.1, 4.3]$. Why these? No particular reason. That is, suppose

$$f_X(x) = \begin{cases} a & 1.1 \leq x \leq 4.3 \\ 0 & otherwise \end{cases}$$

What does this distribution "look like"? Because it is a uniform, you might have a sense that it will be a horizontal line. But, what is the height of that line? Aha! We could do the math to figure it out, or we could generate an approximation using a simulation.

In the code below, we are going to create an object called `samples_uniform` that stores the results of the `runif` function call.

```
samples_uniform <- runif(n=1000, min=1.1, max=4.3)
```

What is happening inside `runif`?

When you're writing you own code, you can pull up the documentation for this (and any) function using a question mark, i.e. `?`, followed by the function name – `?runif`.

But, we can speed this up slightly by simply telling you that `n` is the number of samples to take from the population; `min` is the low-end of the support, and `max` is the high-end of the support.

If we look into this object, we can see the results of the function call. Below, we will show the first 20 elements of the `samples_uniform` object.

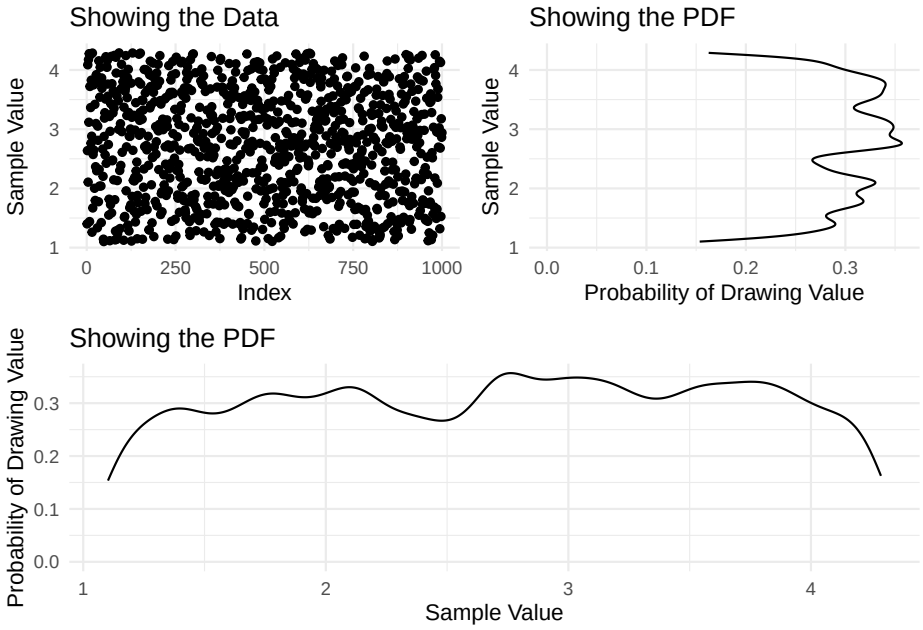
```
samples_uniform[1:20]
```

```
##  [1] 1.401044 2.637784 4.087598 2.105674 4.174847 3.116951 4.275360 1.439993
##  [9] 3.722005 4.253039 3.365360 2.762892 3.828352 2.881831 1.260404 4.285310
## [17] 1.669873 2.573297 3.747595 2.646536
```

(Notice that R is a 1 index language (python is a zero-index language).)

With this object created, we can plot a density of the data and then learn from this histogram what the pdf looks like.

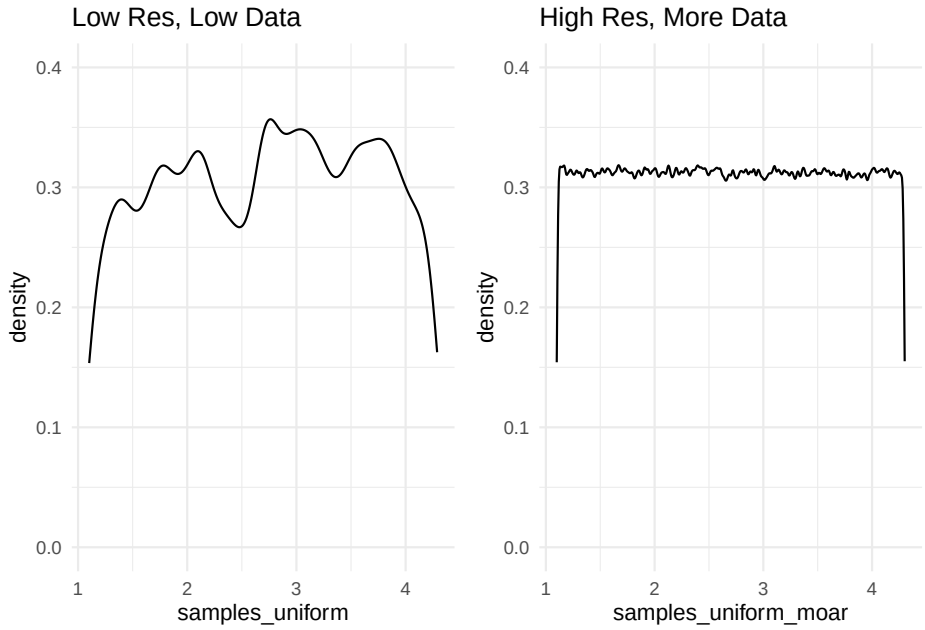
```
plot_full_data <- ggplot() +
  aes(x=1:length(samples_uniform), y=samples_uniform) +
  geom_point()  +
  labs(
    title = 'Showing the Data',
    y     = 'Sample Value',
    x     = 'Index')

plot_density <- ggplot() +
  aes(x=samples_uniform) +
  geom_density(bw=0.1)   +
  labs(
    title = 'Showing the PDF',
    y     = 'Probability of Drawing Value',
    x     = 'Sample Value')

(plot_full_data | (plot_density + coord_flip())) /
  plot_density
```



Interesting. From what we can see here, there does not appear to be any discernible pattern. This leaves us with two options: either, we might reduce the resolution that we're using to view this pattern, or we might take more samples and hold the resolution constant. Below, two different plots show these differing approaches, and are *very* explicit about the code that creates them.

```
samples_uniform_moar <- runif(n=1000000, min=1.1, max=4.3)
```

```
plot_low_res <- ggplot()  +
  aes(x=samples_uniform)  +
  geom_density(bw=0.1)    +
  lims(y=c(0,0.4))        +
  labs(title = 'Low Res, Low Data')

plot_high_res <- ggplot()     +
  aes(x=samples_uniform_moar) +
  geom_density(bw=0.01)       +
  lims(y=c(0,0.4))            +
  labs(title = 'High Res, More Data')

plot_low_res | plot_high_res
```



2.11.2 Example: The Normal Distribution

Folks might have some prior beliefs about the Normal distribution. Don't worry, we'll cover this later in the course. But, this is the distribution that you have in mind when you're thinking of a "bell curve".

We can use the same method to visualize a normal distribution as we did for a uniform distribution. In this case, we would issue the call `rnorm`, together with the population parameters that define the population. At this point in the course, we do not expect that you will know these (and, actually memorizing these facts are not a core focus of the course), but you can look them up if you like. Truthfully, statistics wikipedia is *very* good.

Do do you notice anything about the `runif` and the `rnorm` calls that we have identified? Both seem to name the distribution: $unif \approx uniform$ and $norm \approx normal$, but prepened with a `r`? This is for "random draw".

Base R is loaded with a *pile* of basic statistics distributions, which you can look into using `?distributions`.

```
samples_normal <- rnorm(n=100000, mean=18, sd=4)
```

Like before, we could look at the first 20 of these samples.

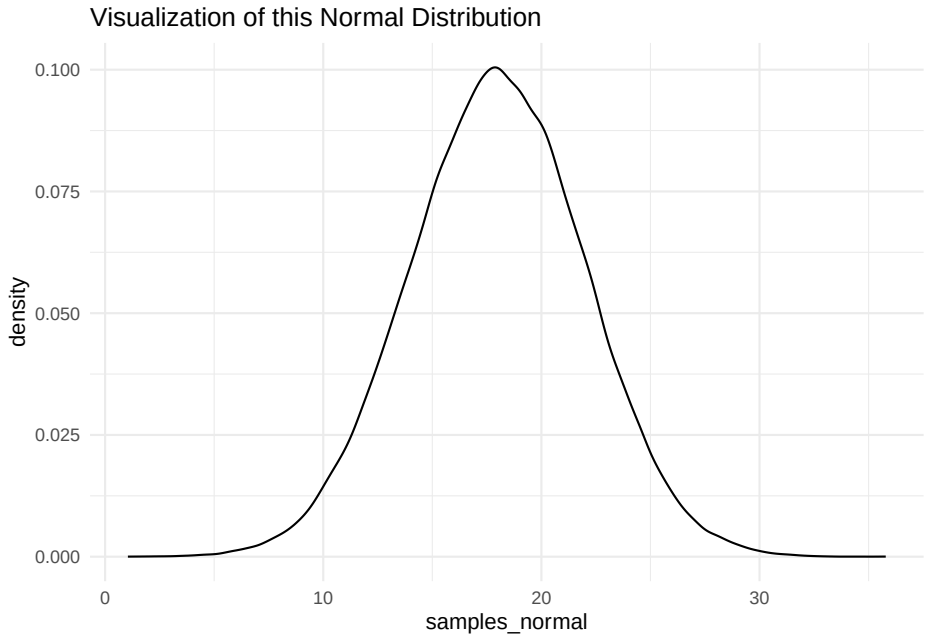
```
samples_normal[1:20]
```

```
##  [1] 20.29549 18.47785 23.29018 11.28234 18.93817 18.60745 23.21143 23.19206
##  [9] 18.83010 16.34688  9.68090 17.99866 22.65670 21.99189 22.02152 19.42756
```

```
## [17] 22.59535 15.92612 20.34516 17.73428
```

And, from here we could visualize this distribution.

```
ggplot() +
  aes(x=samples_normal) +
  geom_density() +
  labs(title='Visualization of this Normal Distribution')
```



2.11.2.1 Combining This Ability

Consider three random variables A, B, C. Suppose,

$$\begin{aligned} A &\sim Uniform(min = 1.1, max = 4.3) \\ B &\sim Normal(mean = 18, sd = 4) \\ C &= A + B \end{aligned}$$

And, suppose that B is a random variable that is described by the normal density that we considered earlier. Suppose that A and B are independent of each other.

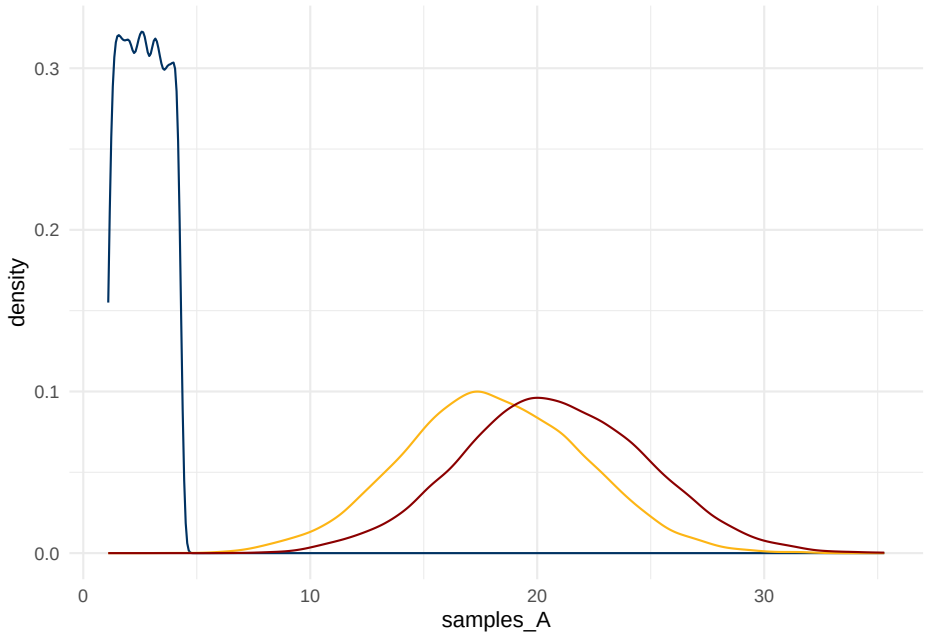
Finally, suppose that $C = A + 2B$.

What does C look like?

Although this is a simple function applied to a random variable – a legal move – the math would be tedious. What if, instead, one used this simulation method to get a sense for the distribution?

```
samples_A <- runif(n=10000, min=1.1, max=4.3)
samples_B <- rnorm(n=10000, mean=18, sd=4)

samples_C <- samples_A + samples_B
```

```
plot_C <- ggplot() +
  aes(x=samples_C) +
  geom_density()

plot_C_and_A_and_B <- ggplot()   +
  geom_density(aes(x=samples_A), color = '#003262') +
  geom_density(aes(x=samples_B), color = '#FDB515') +
  geom_density(aes(x=samples_C), color = 'darkred')

plot_C_and_A_and_B
```



2.12 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.

Chapter 3

Summarizing Distributions

In the last live session, we introduced random variables; probability density and cumulative density; and, made the connection between joint, marginal, and conditional distributions. All of these concepts work with the **entire** distribution.

Take, for example, the idea of conditional probability. We noted that conditional probability is defined to be:

$$f_{Y|X}(y|x) = \frac{f_{Y,X}(y,x)}{f_X(x)}$$

This is a powerful concept that shows a lot of the range of the reasoning system that we've built to this point! The probability distribution of Y might change as a result of changes in X. If you unpack that just a little bit more, we might say that $f_{Y|X}(y|x)$ – the probability density of Y – which is itself a function, is *also* a function of X. To say it again, to be very explicit: the function is a function of another input. That might sound wild, but it is all perfectly consistent with the world that we've built to this point.

This concept is **very** expressive. Knowing $f_Y(y)$ gives a full information representation of a variable; knowing $f_{Y|X}(y|x)$ lets you update that information to make an even more informative statement about Y. In *Foundations* and at this point in the class, we deal only with conditional probability conditioning on a single variable, but the process generalizes.

For example, if there were four random variables, A, B, C, D, we could make a statement about A that conditions on B, C, D:

$$f_{A|\{B,C,D\}}(a|\{b,c,d\}) = \frac{f_{A,B,C,D}(a,b,c,d)}{f_{B,C,D}(b,c,d)}$$

Figure 3.1: a majestic valley

In this week's materials we are going to go in the *opposite* direction: Rather than producing a very expressive system of probabilities, we're going to attempt to summarize all of the information contained in a pdf into lower-dimensional representations. Our first task will be summarizing a single random variable in two ways:

1. Where is the "center" of the random variable; and,
2. How dispersed, "on average" is the random variable from this center.

After developing the concepts of *expectation* and *variance* (which are 1 & 2 above, respectively), we will develop a summary of a joint distribution: the *covariance*. The particular definitions that we choose to call expectation, variance, and covariance require justification. Why should we use these *particular* formulae as measures of the "center" and "dispersion"?

We ground these summaries in the **Mean Squared Error** evaluative metric, as well as justifying this metric.

3.1 Learning Objectives

At the end of the live session and homework this week, students will be able to:

1. **Understand** the importance of thinking in terms of random variables, while;
2. Being able to **appreciate** that it is not typically possible to fully model the world with a single function.
3. **Articulate** why we need a target for a model, and propose several possible such targets.
4. **Justify** why expectation is a good model, why variance is a reasonable model, and how covariance relates two-random variables with a common joint distribution.
5. **Produce** summaries of location and relationship given a particular functional form for a random variable.

3.2 Class Announcements

Where have we come from, and where are we going?

3.2.1 What is in the rearview mirror?

- Statisticians create a population model to represent the world; random variables are the building blocks of such a model.
- We can describe the distribution of a random variable using:
  - A *CDF* for all random variables
  - A *PMF* for discrete random variables
  - A *PDF* for continuous random variables
- When we have multiple random variables,

- The joint PMF/PDF describes how they behave together
- The marginal PMF/PDF describes one variable in isolation
- The conditional PMF/PDF describes one variable given the value of another

3.2.2 Today's Lesson

What might seem frustrating about this probability theory system of reasoning is that we are building a castle in the sky – a fiction. We're supposing that there is some function that describes the probability that values are generated. In reality, there is no such generative function; it is *extremely unlikely* (though we'll acknowledge that it is possible) that the physical reality we believe we exist within is just a complex simulation that has been programmed with functions by some unknown designer.

Especially frustrating is that we're supposing this function, and then we're further saying,

> "If only we had this impossible function; and if only we also had the ability to take an impossible derivative of this impossible function, then we could..."

3.2.2.1 Single number summaries of a single random variable

But, here's the upshot!

What we are doing today is laying the baseline for models that we will introduce next week. Here, we are going to suggest that there are radical simplifications that we can produce that hold specific guarantees, no matter how complex the function that we're reasoning about.

In particular, in one specific usage of the term *best* we will prove that the Expectation operation is the best one-number summary of any distribution. To do so, we will define a term, *variance*, which is the squared deviations from the expectation of a variable that describes how "spread out" is a variable. Then, we will define a concept that is the *mean squared error* that is the square of the distance between a model prediction and a random variable's realization. The key realization is that when the model predicts the expectation, then the MSE is equal to the variance of the random variable, which is the smallest possible value it could realize.

3.2.2.2 Single number summaries of relationships between random variables

Although the single number summaries are **incredibly** powerful, that's not enough for today's lesson! We're also going to suggest that we can create a measure of linear dependence between two variables that we call the "covariance", and a related, re-scaled version of this relationship that is called the correlation.

3.2.3 Future 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

3.3 Discussion of Terms

3.3.1 Expected Value

We define the expected value to be the following for a continuous random variable:

Definition 3.1.

3.4 Expected Value

For a continuous random variable X with PDF f, the *expected value* of X, written $E[X]$ is

$$E[X] = \int_{-\infty}^{\infty} x f_X(x) dx$$

Oh, ok. If you say so. (We do...).

There are two really important things to grasp here:

1. What does this mean about a particular PDF?
2. What is the justification for this *particualar* definition?

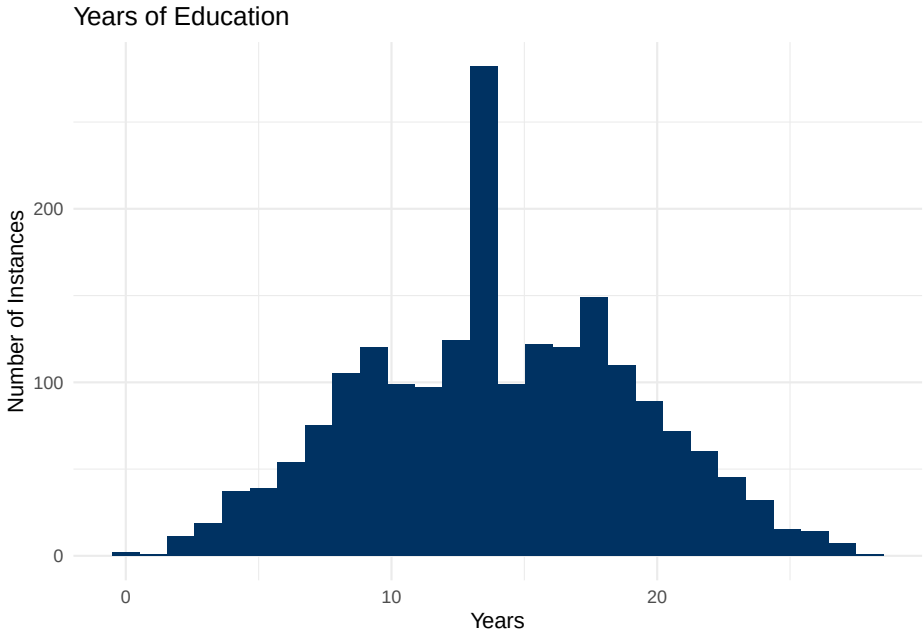
With your instructor, talk about what each of the following definitions mean in your own words. For key concepts, you might also formalize this intuition into a formula that can be computed.

- Expected Value, or Expectation
- Central Moments → Variance → Standard Deviation
- Set aside for later: Chebyshev's Inequality and the Normal Distribution
- Mean Squared Error and its alternative formula
- Covariance and Correlation

3.5 Computing Examples

3.5.1 Expected Value of Education [discrete random variable]

- The expected value of a discrete random variable X is the weighted average of the values in the range of X.
- Suppose that X represents the number of years of education that someone has completed, and so has a support that ranges from 0 years of education, up to 28 years of education. (Incidentally, Mark Labovitz has about 28 years of education.)
- You can then think of



- Without using specific numbers, describe the process you would use to calculate the expected value of this distribution.

3.5.2 Using a formula

- Does the following formula match with your intuitive description of the *expected value*? Why, or why not?

$$\begin{aligned} E[X] &= \sum_{x \in \{EDU\}} x \cdot f(x) \\ &= \sum_{x=0}^{x=28} x \cdot P(X = x) \end{aligned}$$

3.6 Computing by Hand

3.6.1 Compute the Expected Value

Let X represent the result of one roll of a 6 sided die where the events $\omega \in \Omega$ are mapped using a straightforward function: $X(\omega)$: is a function that counts the number of spots that are showing, and maps the number of dots to the corresponding integer, $\mathbb{Z}$.

- Calculating by hand, what is the expected value X, which we write as $E[X]$?
- After you have calculated $E[X]$: Is it possible that the result of a roll is this value?

```
blank_lines(20)
```

3.6.2 Playing a Gnome Game, Part 1

- Suppose that, out on a hike in the hills above campus, you happen across a gnome who asks you if you would like to play the following game:
  - You pay the gnome a dollar, and guess a number between 0 and 6. So, let $g \in \mathbb{R} : 0 \leq g \leq 6$.
  - After you make your guess, the gnome rolls a dice, which comes up with a value $d \in \mathbb{Z} : d \in \{1, 2, 3, 4, 5, 6\}$.
  - The gnome pays you $p = 0.25 \times |d - g|$.
  - **First question**: What is the best guess you can make?
  - **Second Question**: Should you play this game?

Fill this in by hand.

```
blank_lines(20)
```

3.6.3 Compute the Variance

Let X represent the result of one roll of a 6 sided die.

- Calculating by hand, what is the variance of X?

```
blank_lines(20)
```

3.6.4 Playing a Gnome Game, Part 2

- How much do you expect to make on any particular time that you play the game with the best strategy?

```
blank_lines(20)
```

3.7 Expected Value by Code

3.7.1 Expected Value of a Six-Sided Die

Let X represent the result of one roll of a 6 sided die.

- Build an object to represent the whole sample space, Ω of a six sided die.
- Determine what probabilities to assign to each value of that object.
- Write the code to run the expectation algorithm that you just performed by hand.

```
die <- data.frame(
  value = 'fill this in',
  prob  = 'fill this in'
)
```

3.7.2 Variance of a Six-Sided Die

Let X represent the result of one roll of a 6 sided die. Using what you know about the definition of variance, write a function that will compute the variance of your `die` object.

```
variance_function <- function(die) {
  ## fill this in
  mu = 'fill this in'   ## you should index to the correct column
  var = 'fill this in'  ## for each, and use the correct function

  return(var)
}

variance_function(die)
```

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

Suppose that you had to keep the values the same on the die (that is the domain of the outcome still had to be the countable set of integers from one to six), but that you could modify the actual random process. Maybe you could sand off some of the corners on the die, or you could place weights on one side so that the side is less likely to come up. In this case, $\omega \in \{1, 2, 3, 4, 5, 6\}$, but you're able to make a new $f_D(d)$.

- How would you change the probability distribution to decrease the variance of this random variable?
- What is the smallest value that you can generate for this random variable? Use the `variance_function` from above to actually compute this variance.
- What is the largest value of variance that you can generate for this random variable? Use the `variance_function` from above to actually compute this variance.

Now suppose that you again had an equal probability of every outcome, but you were to apply a function to the number of spots that are showing on the die. Rather that each dot contributing one value to the random variable, instead the random variable's outcome is the square of the number of spots.

- How would this change the mean?
- How would this change the variance?

3.8 Practice Computing

3.8.1 Single Variable

Suppose that X has the following density function:

$$f_X(x) = \begin{cases} 6x(1-x), & 0 < x < 1 \\ 0, & otherwise \end{cases}$$

- Find $E[X]$.
- Find $E[X^2]$.
- Find $V[X]$.

3.8.2 Joint Density

3.8.2.1 Discrete Case: Calculate Covariance

In the reading, you saw that we define **covariance** to be:

$$\begin{aligned} Cov[X,Y] &= E[(E[X]-X)^2(E[Y]-Y)^2] \\ &= E[XY] - E[X]E[Y] \end{aligned}$$

And, **correlation** to be a rescaled version of *covariance*:

$$\begin{aligned} Cor[X,Y] &\equiv \rho[X,Y] \\ &= \frac{Cov[X,Y]}{\sigma_X \sigma_Y} \end{aligned}$$

Suppose that X and Y are discrete random variables, where X represents number of office hours attended, and Y represents owning a cat. Furthermore, suppose that X and Y have the joint pmf,

f(x,y)	y=0	y=1
x=0	0.10	0.35
x=1	0.05	0.05
x=2	0.10	0.35

1. Calculate the covariance of X and Y.
2. Are X and Y independent? Why or why not?

3.8.2.2 Continuous Case: Calculate Covariance

Suppose that X and Y have joint density $f_{X,Y}(x,y) = 8xy, 0 \leq y < x \leq 1$.

- Break into groups to find $\text{Cov}[X,Y]$

Suppose that X and Y are random variables with joint density

$$f_{X,Y}(x,y) = \begin{cases} 1, & -y < x < y, 0 < y < 1 \\ 0, & \text{elsewhere} \end{cases}$$

Show that $\text{Cov}[X,Y] = 0$ but that X and Y are dependent.

3.9 Write Code

Suppose that you have a random variable with a **gnarly** probability distribution function:

$$f_X(x) = \frac{3 * \left(x - 2x^2 + x^3\right)}{2}, 0 \leq x \leq 2$$

If you had to pick a single value that minimizes the MSE of this function, what would it be?

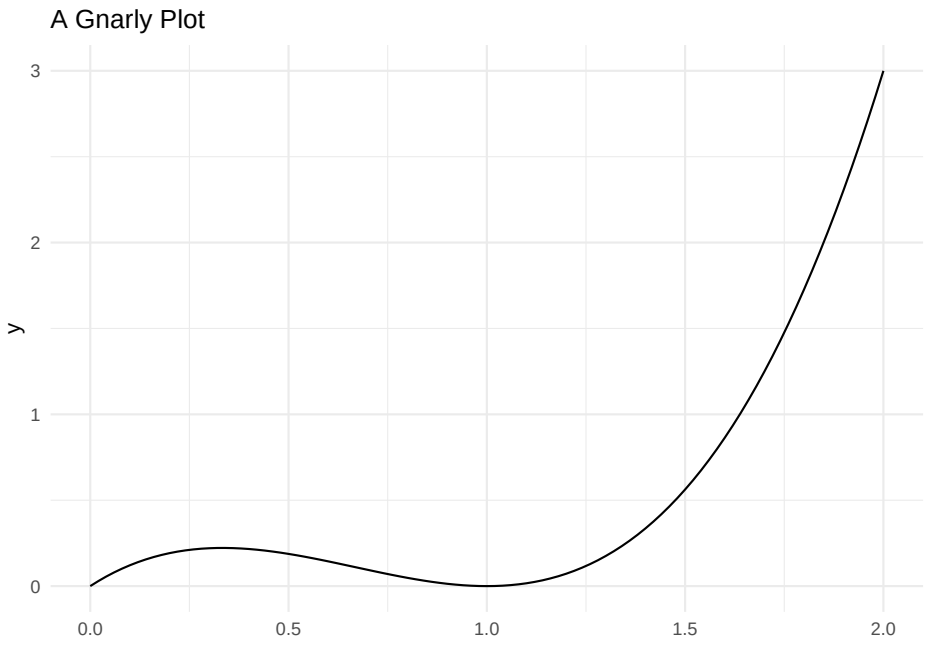
- First, how would you approach this problem *analytically*. By this, we mean, "how would you solve this with the closed form answer?
- Second, how might you approach this problem *computationally*. By this, we mean, "how might you write code that would produce a numeric approximation of the closed form solution?" Don't worry about actually writing the code – we'll have done that for you, but what is the *process* (called in our world, *algorithm*) that you would use to determine the value that produces the smallest MSE?

```
pdf_fun <- function(x) {
  (3/2)*(x - (2*x^2) + x^3)
}
```

```
support <- seq(from=0, to=2, by=0.01)
```

```
ggplot() +
  geom_function(fun = pdf_fun) +
  xlim(min(support), max(support)) +
  labs(
```

```
    title = "A Gnarly Plot"
  )
```




```
expected_value <- function(value, prob){
  sum(value * prob)
}

mse <- function(c) {
  expected_value(
    value = (support - c)^2,
    prob  = pdf_fun(support)
  )
}

mpe <- function(c, power) {
  expected_value(
    value = (support - c)^power,
    prob  = pdf_fun(support)
  )
}
```

```
mean_absolute_error <- function(c) {
  x_values <- pdf_fun(support)
  mae_     <- mean(abs(x_values - c))
}
```

```r
mean_square_error <- function(c) {
  x_values <- pdf_fun(support)
  mse_     <- sum(((x_values - c)^2) * x_values)
  return(mse_)
}

mean_cubic_error <- function(c) {
  x_values <- pdf_fun(support)
  mce_     <- mean((x_values - c)^3)
}

mean_quadratic_error <- function(c) {
  x_values <- pdf_fun(support)
  mqe_     <- mean((x_values - c)^4)
  return(mqe_)
}

mean_power_error <- function(c, power) {
  x_values <- pdf_fun(support)
  m_power_e_     <- mean((x_values - c)^power)
  return(m_power_e_)
}
```

```r
mean_absolute_error  <- Vectorize(mean_absolute_error)
mean_square_error    <- Vectorize(mean_square_error)
mean_cubic_error     <- Vectorize(mean_cubic_error)
mean_quadratic_error <- Vectorize(mean_quadratic_error)
mean_power_error     <- Vectorize(mean_power_error)
```

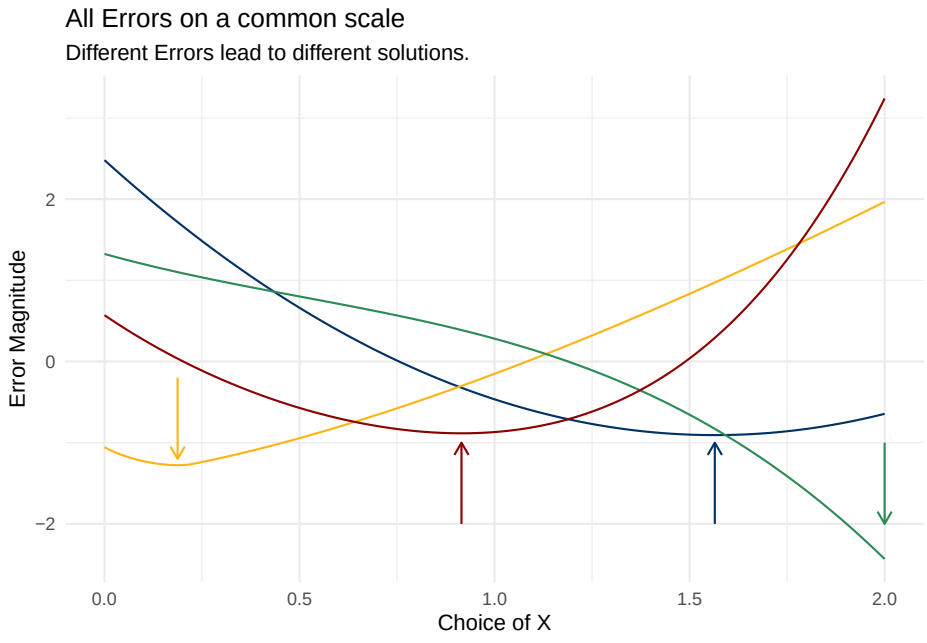
```r
mae_ <- mean_absolute_error(
  c = support
)
mse_ <- mean_square_error(
  c = support
  )
mce_ <- mean_cubic_error(
  c = support
)
mqe_ <- mean_quadratic_error(
  c = support
)
```

```r
absolute_error_ <- optim(
  par = 0,
  fn = mean_absolute_error,
```

```
  method = 'Brent',
  lower = 0, upper = 2
  )$par

squared_error_ <- optim(
  par = 0,
  fn = mean_square_error,
  method = "Brent",
  lower = 0, upper = 2
  )$par
cubic_error_ <- optim(
  par = 0,
  fn = mean_cubic_error,
  method = "Brent",
  lower = 0, upper = 2
  )$par
quadratic_error_ <- optim(
  par = 0,
  fn = mean_quadratic_error,
  method = "Brent",
  lower = 0, upper = 2
  )$par
```

```
all_plots <- ggplot() +
  ## add lines
  geom_line(aes(x=support, y=scale(mse_)), color = "#003262")  +
  geom_line(aes(x=support, y=scale(mae_)), color = "#FDB515")  +
  geom_line(aes(x=support, y=scale(mce_)), color = "seagreen") +
  geom_line(aes(x=support, y=scale(mqe_)), color = "darkred")  +
  ## add optimal solution indicators
  geom_segment(
    aes(x = squared_error_,
    xend = squared_error_,
    y = -2,
    yend = -1),
    arrow = arrow(length = unit(0.25, "cm")),
    color = "#003262")  +
  geom_segment(
    aes(x = absolute_error_,
    xend = absolute_error_,
    y = -.2,
    yend = -1.2),
    arrow = arrow(length = unit(0.25, "cm")),
    color = "#FDB515")  +
  geom_segment(
```

```
    aes(x = cubic_error_,
    xend = cubic_error_,
    y = -1,
    yend = -2),
    arrow = arrow(length = unit(0.25, "cm")),
    color = "seagreen")  +
  geom_segment(
    aes(x = quadratic_error_,
    xend = quadratic_error_,
    y = -2,
    yend = -1),
    arrow = arrow(length = unit(0.25, "cm")),
    color = "darkred")  +
  labs(
    title    = "All Errors on a common scale",
    subtitle = "Different Errors lead to different solutions.",
    y        = "Error Magnitude",
    x        = "Choice of X"
  )

all_plots
```



Chapter 4

Conditional Expectation and The BLP

Figure 4.1: mt. tamalpais

One of our most fundamental goals as data scientists is to produce predictions that are *good*. In this week's async, we make a statement of performance that we

can use to evaluate how good a job a predictor is doing, choosing Mean Squared Error.

With the goal of minimizing MSE, then we then present, justify, and prove that the conditional expectation function (*the CEF*) is the globally best possible predictor. This is an incredibly powerful result, and one that serves as the backstop for **every** other predictor that you will ever fit, whether that predictor is a "simple" regression, or that predictor is a machine learning algorithms (e.g. a random forest) or a deep learning algorithm. Read that again:

> Even the most technologically advanced machine learning algorithms *cannot possibly* perform better than the conditional expectation function at making a prediction.

Why does the CEF do so well? Because it can contain a *vast* amount of complex information and relationships; in fact, the complexity of the CEF is a product of the complexity of the underlying probability space. If that is the case, then why don't we just use the CEF as our predictor every time?

Well, this is one of the core problems of applied data science work: we are never given the function that describes the behavior of the random variable. And so, we're left in a world where we are forced to produce predictions from simplifications of the CEF. A very strong simplification, but one that is useful for our puny human brains, is to restrict ourselves to predictors that make predictions from a linear combination of input variables.

Why should we make such a strong restriction? After all, the conditional expectation function might be a fantastically complex combination of input features, why should we entertain functions that are only linear combinations? Essentially, this is because we're limited in our ability to reason about anything more complex than a linear combination.

4.1 Thunder Struck

4.2 Learning Objectives

At the end of this weeks learning, which includes the asynchronous lectures, reading the textbook, this live session, and the homework associated with the concepts, student should be able to

1. **Recognize** that the conditional expectation function, the *CEF*, is a the pure-form, best-possible predictor of a target variable given information about other variables.
2. **Recall** that all other predictors, be they linear predictors, non-linear predictors, branching predictors, or deep learning predictors, are an attempt to approximate the CEF.
3. **Produce** the conditional expectation function as a predictor, given joint densities of random variables.

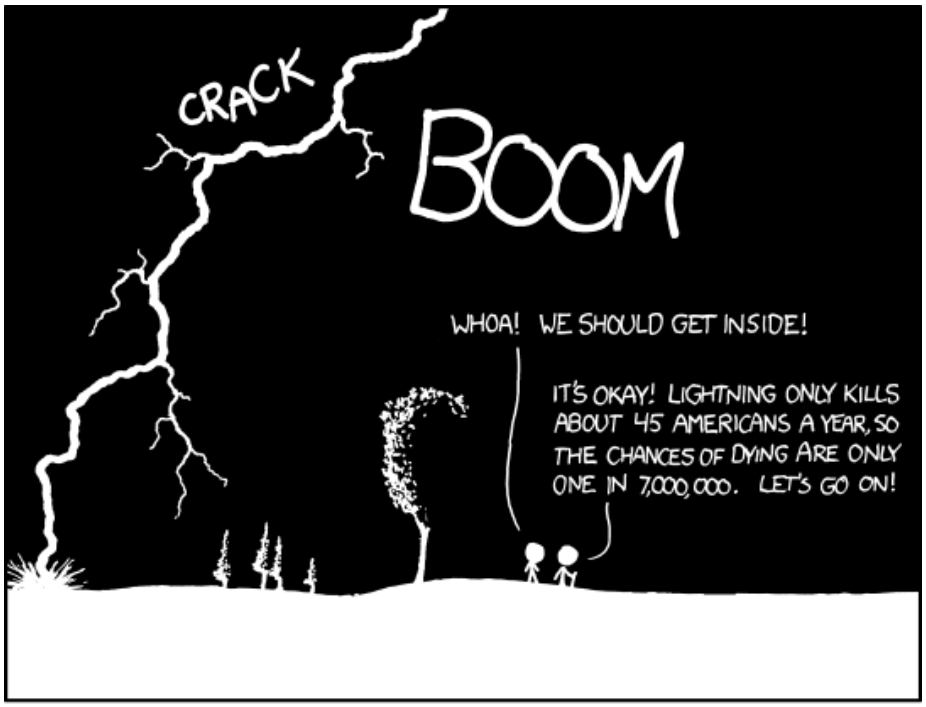


Figure 4.2: thunder struck

4. **Appreciate** that the best linear predictor, which is a restriction of predictors to include only those that are linear combinations of variables, can produce reasonable predictions, and **anticipate** that the BLP forms the target of inquiry for regression.

4.3 Class Announcements

4.3.1 Test 1 is releasing to you today.

The first test is releasing today. There are review sessions scheduled for this week, practice tests available, and practice problems available. The format for the test is posted in the course discussion channel. In addition to your test, your instructor will describe your responsibilities that are due next week.

4.4 Roadmap

4.4.1 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, $E[X + Y] = E[X] + E[Y]$.

4.4.2 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.

4.4.3 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.

4.5 Conditional Expectation Function (CEF),

4.5.1 Part I

Think back to remember the definition of the expectation of Y:

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

This week, in the async reading and lectures we added a new concept, the conditional expectation of Y given $X = x \in \text{Supp}[X]$:

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

4.5.2 Part II

1. What desirable properties of a predictor does the expectation possess (note, this is thinking *back* by a week)? What makes these properties desirable?
2. Turning to the content from this week, how, if at all, does the conditional expectation improve on these desirable properties?

4.5.3 Part III

- Compare and contrast $E[Y]$ and $E[Y|X]$. For example, when you look at how these operators are "shaped", how are their components similar or different?[1]
- What is $E[Y|X]$ a function of? What are "input" variables to this function?
- What, if anything, is $E[E[Y|X]]$ a function of?

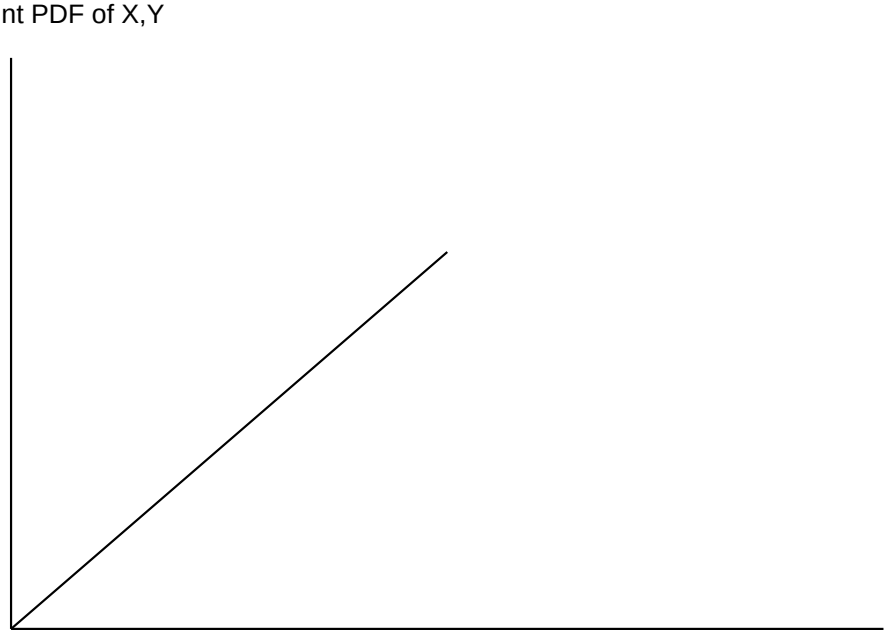
4.6 Computing the CEF

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

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

What does the joint PDF of this function look like?

[1]Note, when we say "shaped" here, we're referring to the deeper concept of a statistical functional. A statistical functional is a function of a function that maps to a real number. So, if T is the functional that we're thinking of, $\mathcal{F}$ is a family of functions that it might operate on, and $\mathbb{R}$ is the set of real numbers, a statistical functional is just $T : \mathcal{F} \to \mathbb{R}$. The Expectation statistical functional, $E[X]$ always has the form $\int x f_X(x)dx$.)



4.6.1 Simple Quantities

To begin with, let's compute the simplest quantities:

- What is the expectation of X?
- What is the expectation of Y?
- How would you compute the variance of X? (We're not going to do it live).

4.6.2 Conditional Quantities

4.6.2.1 Conditional Expectaiton

And then, let's think about how to compute the conditional quantities. To get started, you can use the fact that in week two, we already computed the conditional probability density function:

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

With this knowledge on hand, compute the $CEF[Y|X]$.

Once you have computed the $CEF[Y|X]$, use this function to answer the following questions:

- What is the conditional expectation of Y, given that $X = x = 0$?
- What is the conditional expectation of Y, given that $X = x = 0.5$?
- What is the conditional expectation of X, given that $Y = y = 0.5$?

4.6.2.2 Conditional Variance

- What is the conditional variance function?[2]

[2]Take a moment to strategize just a little bit before you get going on this one. There is a way to compute this value that is easier than another way to compute this value.

- Which of the two of these has a lower conditional variances?
  - $V[Y|X = 0.25]$; or,
  - $V[Y|X = 0.75]$.
- How does $V[Y]$ compare to $V[Y|X = 1]$? Which is larger?

4.6.3 Conditional Expectation

4.7 Minimizing the MSE

4.7.1 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).

Oh yeah? As a breakout group, *ride shotgun* with us as we prove that the conditional expectation is the function that produces the smallest possible Mean Squared Error.

Specifically, **you group's task** is to justify every transition from one line to the next using concepts that we have learned in the course: definitions, theorems, calculus, and algebraic operations.

4.7.2 The pudding (aka: "Where the proof is")

We need to find such function $g(X) : \mathbb{R} \to \mathbb{R}$ that gives the smallest mean squared error.

First, let MSE be defined as it is in **Definition 2.1.22**.

> For a random variable X and constant $c \in \mathbb{R}$, the *mean squared error* of X about c is $E[(x - c)^2]$.

Second, let us note that since $g(X)$ is just a function that maps onto $\mathbb{R}$, that for some particular value of $X = x$, $g(X)$ maps onto a constant value.

- Deriving a Function to Minimize MSE

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

Notice too that we can use the *Law of Iterated Expectations* to do something useful. (This is a good point to talk about how this theorem works in your breakout groups.)

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

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

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

4.7.3 The Implication

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

4.8 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.

4.9 Joint Distribution Practice

4.9.1 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 $E[Y] = E\big[E[Y|X]\big]$.

4.9.2 Continuous BLP

- Recall the PDF that we worked with earlier to produce the $CEF[Y|X].

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

Find the BLP for Y as a function of X. What, if anything, do you notice about this BLP and the CEF?

Chapter 5

Learning from Random Samples

Figure 5.1: south hall

This week, we're coming into the big turn in the class, from probability theory to sampling theory.

In the probability theory section of the course, we developed the *theoretically* **best** possible set of models. Namely, we said that if our goal is to produce a

model that minimizes the Mean Squared Error that *expectation* and *conditional expectation* are as good as it gets. That is, if we only have the outcome series, Y, we cannot possibly improve upon $E[Y]$, the expectation of the random variable Y. If we have additional data on hand, say X and Y, then the best model of Y given X is the conditional expectation, $E[Y|X]$.

We have also said that because this conditional expectation function might be complex, and hard to inform with data, that we might also be interested in a principled simplification of the conditional expectation function – the simplification that requires our model be a line.

With this simplification in mind, we derived the linear system that produces the minimum MSE: the ratio of covariance between variables to variance of the predictor:

$$\beta_{BLP} = \frac{Cov[Y, X]}{V[X]}.$$

We noted, quickly, that the simple case of only two variables – an outcome and a single predictor – generalizes nicely into the (potentially very) many dimensional case. If the many-dimensional BLP is denoted as $g(\mathbf{X}) = b_0 + b_1 X_1 + \cdots + b_k X_k$, then we can arrive at the sloped between one particular predictor, X_k, and the outcome, Y, as:

$$b_k = \frac{\partial g(\mathbf{X})}{\partial X_k}.$$

5.1 Goals, Framework, and Learning Objectives

5.1.1 Class Announcements

- You're done with probability theory. **Yay!**
- You're also done with your first test. **Double Yay!**
- We're going to have a second test in a few weeks. Then we're done testing for the semester **Yay?**

5.1.2 Learning Objectives

At the end of this week, students will be able to

1. **Understand** what iid sampling is, and evaluate whether the assumption of iid sampling is sufficiently plausible to engage in frequentist modeling.
2. **Appreciate** that with iid sampling, summarizing functions of random variables are, themselves, random variables with probability distributions and values that they obtain.
3. **Recall** the definition of an estimator,

4. **Recall** definition of an estimator, **state** and **understand** the desirable properties of estimators, and **evaluate** whether an estimator possesses those desirable properties.
5. **Distinguish** between the concepts of {expectation & sample mean}, {variance & unbiased sample variance estimator, sampling-based variance in the sample mean}.

5.1.3 Roadmap

5.1.3.1 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: these are the **frequentist tests**.
- Second, we're going to show that OLS regression is the sample estimator of the BLP. This means that OLS regression produces estimates of the BLP that have known convergence properties.
- Third, we're going combine the frequentist testing framework with OLS estimation to produce a full regression testing framework.

5.1.3.2 Where We've Been – Random Variables and Probability Theory

Statisticians create a model (also known as the population model) to represent the world. This model exists as joint probability densities that govern the probabilities that any series of events occurs at the same time. This joint probability of outcomes can be summarized and described with lower-dimensional summaries like the expectation, variance, covariance. While the expectation is a summary that contains information on about one marginal distribution (i.e. the outcome we are interested in) we can produce predictive models that update, or *condition* the expectation based on other random variables. This summary, the **conditional expectation** is the best possible (measured in terms of minimizing mean squared error) predictor of an outcome. We might simplify this conditional expectation predictor in many ways; the most common is to simplify to the point that the predictor is constrained to be a line or plane. This is known as the Best Linear Predictor.

5.1.3.3 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

5.2 Key Terms and Assumptions

5.2.1 IID

We use an abbreviation for the sampling process that under girds our frequentist statistics. That abbreviation, **IID**, while short, contains two powerful requirements of our data sampling process.

Definition 5.1. IID sampling is:

1. **Independent**. The first **I** in the abbreviation, this independence requirement is similar to the independence concept that we've used in the probability theory section of the course. When samples are independent, the result of any one sample is not informative about the value of any of the other samples.
2. **Identically Distributed**. The **ID** in the abbreviation, this requirement means that all samples are drawn from the same distribution.

It might be tempting to imagine that IID samples are just *"random samples"*, but it is worth noting that IID sampling has the two specific requirements noted above, and that these requirements are more stringent than a "randomness" criteria.

When we are thinking about IID samples, and evaluating whether the sample do, in point of fact, meet both of the requirements, it is *crucial* to make an explicit statement about the reference population that is under consideration.

For example, suppose that you were interested in learning about life-satisfaction and your reference population are the peoples who live in the United States. Further, suppose that you decide to produce an estimate of this using a sample drawn from UC Berkeley undergraduate students during RRR week? There are several flaws in this design:

1. There is a key **research design** issue: a sample drawn from Berkeley undergraduates is going to be *essentially* uninformative of a US resident reference population!
2. There is a key **statistical** issue: the population of Berkeley undergraduates are not really an independent sample from the entire US resident reference population. Once you learn the age of someone from the Berkeley student population, you can make an conditional guess about the age of the next sample that will be closer than was possible before the first sample. The same goes for life-satisfaction: When you learn about the life-satisfaction from the first undergrad (who is miserable because they have their Stat 140 final coming up) while they are studying for their finals) you can make a conditionally better guess about the satisfaction of the next undergrad.

Notice that these violations of the IID requirements only arise because our reference population is the US resident population. If, instead, the reference population were "Berkeley undergrads" then the sampling process *would* satisfy

the requirements of an IID process.

- How, or why, can a change in the reference population make an identical sampling process move from one that we can consider IID to one that we cannot consider IID?

5.2.1.1 It this IID?

For each of the following scenarios, is the IID assumption plausible?

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

5.3 Estimators

In our presentation of this week's materials, we choose to switch the presentation of statistics and estimators, electing to discuss properties that we would like estimators to posses, before we actually introduce any specific estimators.

5.3.1 Three properties of estimators

What are the three desirable properties of estimators?

1.
2.
3.

Is one these properties *more* important than another? If you had to force-rank these properties in terms of their importance, which is the most, and which the least important? Why?

5.4 Estimator Property: Biased or Unbiased?

1. First, for a general case: Suppose that you have chosen some particular estimator, $\hat{\theta}$ to estimate some characteristic, θ of a random variable. How do you know if this estimator is unbiased?
2. Second, for a specific case: Define the "sample average" to be the following: $\frac{1}{n}\sum_{i=1}^{N} x_i$. Prove that this sample average estimator is an unbiased estimator of $E[X]$.
3. Third (easier), for a different specific case: Define the "smample smaverage" to be the following $\frac{1}{n^2}\sum_{i=1}^{N} x_i$. Prove that the smample smaverage is a biased estimator of $E[X]$.

4. Fourth (harder): Define the geometric mean to be

$$\left(\prod_{i=1}^{N} x_i\right)^{\frac{1}{N}}$$

. Prove that the geometric mean is a biased estimator of $E[X]$.

5.4.1 Is it unbiased, with data?

Suppose that you're getting data from the following process:

```
random_distribution <- function(number_samples) {

  d1 <- c(1.0, 2.0)
  d2 <- c(1.1, 2.1)
  d3 <- c(1.5, 2.5)

  distribution_chooser = sample(x=1:3, size=1)

  if(distribution_chooser == 1) {
    x_ <- runif(n=number_samples, min=d1[1], max=d1[2])
  } else if(distribution_chooser == 2) {
    x_ <- runif(n=number_samples, min=d2[1], max=d2[2])
  } else if(distribution_chooser == 3) {
    x_ <- runif(n=number_samples, min=d3[1], max=d3[2])
  }
  return(x_)
}

random_distribution(number_samples=10)
```

```
##  [1] 2.467197 2.366916 1.937715 1.691938 1.582294 2.083452 1.570361 2.027663
##  [9] 1.972288 1.548191
```

```
mean(random_distribution(number_samples=10000))
```

```
## [1] 1.501582
```

Notice that, there are two forms of inherent uncertainty in this function:

1. There is uncertainty about the distribution that we are getting draws from; and,
2. Within a distribution, we're getting draws at random from a population distribution.

This class of function, the `r*` functions, are the implementation of random generative processes within the R language. Look into `?distributions` as a class to see more about this process.

Suppose that you chose to use the same sample average estimator as a means of producing an estimate of the population expected value, $E[X]$. Suppose that you get the following draws:

```
draws <- random_distribution(number_samples=10)
draws
```

```
##  [1] 1.946123 1.913084 1.711113 1.768129 1.573630 1.916366 1.149307 1.556135
##  [9] 1.358160 1.889717
```

```
mean(draws)
```

```
## [1] 1.678176
```

Is this sample average an unbiased estimator for the population expected value? How do you know?

5.5 Estimator Property: Consistency

Foundations makes another of their jokes when they write, on page 105,

> "Consistency is a simple notion: if we had enough data, the probability that our estimate $\hat{\theta}$ would be far from the truth, θ, would be small."

How do we determine if a particular estimator, $\hat{\theta}$ is a consistent estimator for our parameter of interest?

There are at least two ways:

1. The estimator is unbiased, and has a sampling variance that decreases as we add data; or,
2. We can use Chebyshev's to place a bound on the estimator, showing that as we add data, the estimator converges in probability to θ.

The first notion of convergence requires an understanding of sampling variance:

The sampling variance of an estimator is a statement about how much dispersion due to random sampling, is present in the estimator. We defined the variance of a random variable to be $E\left[(X - E[X])^2\right]$, The sampling variance uses this same definition, but we work with it slightly differently when we are considering sampling variance.

In particular, when we are considering sampling variance, we do not typically got as far as actually computing the variance of the underlying random variable? *Why?* Because, if we're working in a sampling scenario, it is unlikely that we have access to the underlying function that governs the PDF of the random variable.

Instead, we typically start from a statement of the estimator that is under consideration, and apply the variance operator against that estimator. Consider,

for example, forming a statement about the sampling variance of the sample average.

Let $\overline{X} \equiv$ "sample average" $\equiv \frac{1}{n}\sum_{i=1}^{n} X_i$ be the normal form of the sample average.

Earlier, we proved that $\overline{X}$ is an unbiased estimator of $E[X]$.

What is the sampling variance of the sample average?

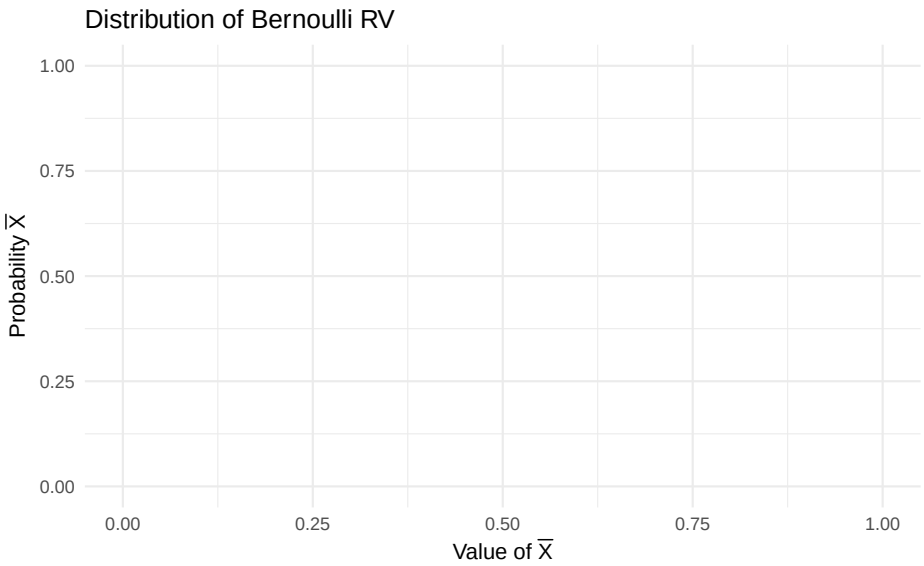
Using the statement that you have just produced, would you say that the sample average is a consistent estimator for the population expectation of a random variable?

5.6 Understanding Sampling Distributions

How do sampling distributions change as we add data to them? This is going to both motivate convergence, and also play forward into the Central Limit Theorem. Let's work through an example that begins with a case that we can think through and draw ourselves. Once we feel pretty good about the *very* small sample, then we will rely on R to do the work when we expand the example beyond what we can draw ourselves.

Suppose that X is a Bernoulli random variable representing an unfair coin. Suppose that the coin has a 70% chance of coming up heads: $P(X = 1) = 0.7$.

- To begin, suppose that you take that coin, and you toss it two times: you have an iid sample of size 2, (X_1, X_2).
- *What is the sampling distribution of the sample average, of this sample of size two?*
- On the axes below, draw the probability distribution of $\overline{X} = \frac{X_1+X_2}{2}$.



- What if you took four samples? What would the sampling distribution of $\overline{X}$ look like? *Draw this onto the axis above.*
- Explain the difference between a population distribution and the sampling distribution of a statistic.
- Why do we want to know things about the sampling distribution of a statistic?

We are going to write a function that, essentially, just wraps a built-in function with a new name and new function arguments. This is, generally, bad coding practice – because it is changing the default lexicon than a collaborator needs to be aware of – but it is useful for teaching purposes here.

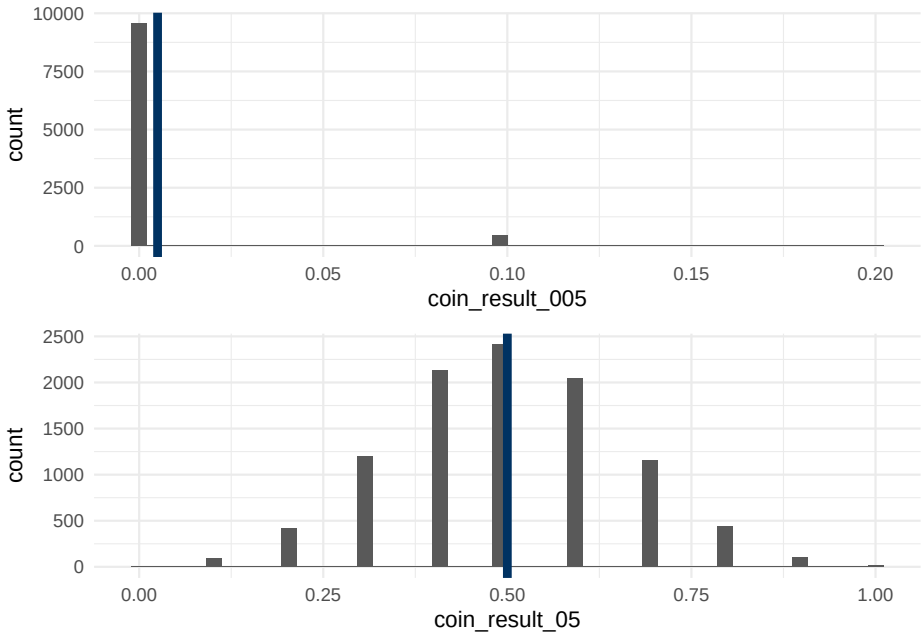
- The `number_of_simulations` argument to the `toss_coin` function basically just adjusts the precision of our simulation study.
- Let's set, and keep this at 1000 simulations. But, if you're curious, you could set this to be 5, or 10 and evaluate what happens.

```
toss_coin <- function(
    number_of_simulations=1000,
    number_of_tosses=2,
    prob_heads=0.7) {

  ## number of simulations is just how many times we want to re-run the experiment
  ## number of tosses is the number of coins we're going to toss.
  number_of_heads <- rbinom(n=number_of_simulations, size=number_of_tosses, prob=prob_
  sample_average  <- number_of_heads / number_of_tosses
  return(sample_average)
}

toss_coin(number_of_simulations=10, number_of_tosses=2, prob_heads=0.7)
```

```
##  [1] 1.0 0.5 1.0 0.5 0.5 0.5 0.5 0.5 1.0 1.0
```

```
ncoins <- 10

coin_result_005 <- toss_coin(
  number_of_simulations = 10000,
  number_of_tosses = ncoins,
  prob_heads = 0.005
  )

coin_result_05 <- toss_coin(
  number_of_simulations = 10000,
  number_of_tosses = ncoins,
  prob_heads = 0.5
  )

plot_005 <- ggplot() +
    aes(x=coin_result_005) +
    geom_histogram(bins=50) +
    geom_vline(xintercept=0.005, color='#003262', linewidth=2)

plot_05 <- ggplot() +
    aes(x=coin_result_05) +
```

```
    geom_histogram(bins=50) +
    geom_vline(xintercept=0.5, color='#003262', linewidth=2)

plot_005 /
  plot_05
```



In the plot that you have drawn above, pick some value, ϵ that is the distance away from the true expected value of this distribution.

- What proportion of the sampling distribution is further away than $E[X] \pm \epsilon$?
- When we toss the coin only two times, we can quickly draw out the distribution of $\overline{X}$, and can form a statement about the $P(E[X] - \epsilon \leq \overline{X} \leq E[X] + \epsilon)$.
- What if we toss the coin ten times? We can still use the IID nature of the coin to figure out the *true* $P(\overline{X} = 0), P(\overline{X} = 1), \ldots, P(\overline{X} = 10)$, but it is going to start to take some time. This is where we rely on the simulation to start speeding up our learning.
- 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?

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

When you were reading for this week, did you sense the palpable *joy* of the authors when they were writing about the central limit theorem?

> We now preset what is often considered the most profound and important theorem in statistics.

Wow. What excitement.

On its own, the result that *across a broad range of generative functions the distribution of sample averages converges in distribution to follow a normal distribution* would be a statistical curiosity. Along the lines of "did you know that dinosaurs might have had feathers," or, "avocado trees reproduce using 'protogynous dichogamy". While these factoids might be useful on your quiz-bowl team, they don't get us very far down the line as practicing data scientists.

However, there is a **very** useful consequence of this convergence in distribution that we will explore in detail over the coming two weeks: because *so many* distributions produce sample averages that converge in distribution to a normal distribution, we can put together a testing framework for sample averages that works for an *agnostic* set of random variables. Wait for that next week, but know that there's a reason that we're as excited about this statement as we are.

5.7.1 Part 1

To begin with, we will use fair coins that have an equal probability landing heads and tails.

1. Modify the function argument below so that it conduct **one** simulations, and in each simulation tosses **ten** coins, each with an **equal** probability of landing heads and tails. Look into the `toss_coin` function: is there a point that this function is producing a sample average? If so, where?
2. Save values from the `toss_coin` function into an object, called `sample_mean`.

```
# toss_coin()
```

The sample mean is a random variable – it is a function that maps from a random generative process' sample space (the number of heads shown on dice) to the real numbers. To try to make this clear, visualize a larger number of simulations on the `toss_coin` function. That is, increase the `number_of_simulations` to be 10, or 100. and plot a histogram of the results. This is quite similar to what we have done earlier.

```
# toss_coin()
```

If you replicate the simulation with **ten** coins enough times, will the distribution

ever look normal? Why or why not?

5.7.2 Part 2

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

What happens if you change the number of coins that you're tossing? Here, set `number_of_simulations=1000`, and examine what happens if you alter the number of coins that you're tossing? Is there a point where this distribution starts to "look normal" to you? (Later in the semester, we'll formalize a test for this "looks normal" concept).

5.7.3 Part 3

What would happen if the coin was very, very unfair? For this part, study a coin that has a `prob_heads=0.01`. This is an example of a highly skewed random variable.

Start your study here tossing three coins, `number_of_coins=3`. What does this distribution look like?

What happens as you increase the number of coins that you're tossing? Is there a point that the distribution starts to look normal?

5.7.4 Discussion Questions About the CLT

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?

5.8 Errors with Standard Errors

Talking about variance and sampling variance is hard, because the terms sound **very** similar, but have important distinctions in what they mean. For example, the "variance" is not the same as the "unbiased sample variance" which is also not the same as the "sampling variance of the sample average". :sob:

Standard errors are a statement about the sampling variance of the sample average. But, related to this concept are the ideas of the *Population Variance*, the *Plug-In Estimator for the Sample Variance*, the *Unbiased Sample Variance*, and, finally, the *Sampling Variance of the Sample Average* (i.e the *Standard Error*).

How are each of these concepts related to one another, and how can we keep them all straight? As a group, fill out the following columns?

For the **Estimator Properties** column, here we're considering, principally biased/unbiased and consistent/inconsistent.

Population Concept	Pop Notation	Sample Estimator	Sample Notation	Estimator Properties	Sampling Variance of Sample Estimator
Expected Value					
Population Variance					
Population Covariance					
CEF					
BLP					

Chapter 6

Hypothesis Testing

Figure 6.1: more lake

Frequentist Hypothesis testing is a very well established framework in the applied practice, and scientific literature. Sometimes (often, currently) referred to as Null Hypothesis Significance Testing (NHST), this framework essentially makes an absurd assertion and asks the data to overturn that assertion.

Like a petulant child, NHST essentially proclaims,

> "If you really loved me, you would let me watch this screen one-

hundred hours every day."

Here the absurdity is that a parent might not love their child, and the criteria to overturn that assertion is noted to be "buy me an iPad" '.

What is Frequentist testing doing?

This testing framework works on **samples** of data, and applies **estimators** to produce **estimates** of **population parameters** that are fundamentally unknown and unknowable. Despite this unknown and unknowable population target, with some carefully written down estimators we can rely on the convergence characteristics of some estimators to produce *useful*, *reliable* results.

We begin with the one-sample t-test. The one-sample t-test relies on the sample average as an estimator of a population expectation. In doing so, it relies on the effectiveness of the **Weak Law of Large Numbers** and the **Central Limit Theorem** to guarantee that the estimator that **converges in probability** to the population expectation, while also **converging in distribution** to a Gaussian distribution.

These two convergence concepts permit a data scientist to make several **inferences** based on data:

1. The probability of generating data that *"looks like what is observed"*, if the null-hypothesis were true. This is often referred to as the **p-value** of a test, and is the petulant statement identified above.
2. An interval of values that, with some stated probability (e.g. 95%), contains the true population parameter.

This framework begins a **exceedingly important** task that we must understand, and undertake when we are working as data scientists: Producing our best-estimate, communicating how we arrived at that estimate, what (if any) guarantees that estimate provides, and *crucially* all **limitations** of our estimate.

6.1 Learning Objectives

1. **Understand** the connection between random variables, sampling, and statistical tests.
2. **Apply** the Frequentist testing framework in a simple test – the one-sample t-test.
3. **Anticipate** that every additional Frequentist test is a closely related variant of this test.

6.2 Class Announcements

1. You will be taking your second test this week. The test follows the same format as the first test, which we will discuss in live session. The test will

cover:
   1. ***Unit 4****: Conditional Expectation and the Best Linear Predictor;* and,
   2. ***Unit 5****: Learning from Random Samples.*

- Like the first test, our goal is to communicate to you what concepts we think are important, and then to test those concepts directly, and fairly. The purpose of the test is to give you an incentive to review what you have learned through probability theory, and then to demonstrate that you can produce work based on that knowledge.
- There is another practice test on Gradescope, and in the GitHub repository.

2. In rosier news, we're moving out of the *only* pencil and paper section of this course, and bringing what we have learned out into the dirty world of data. This means a few things:

- If you haven't yet worked through the **R Bridge Course** that is available to you, working on this bridge course will be useful for you (after you complete your test). The goal of the course is to get you up and running with reasonably successful code and workflows for the data-based portion of the course.

3. We will assign teams, and begin our work on **Lab 1** in Live Session next week. This is a two-week, group lab that you will work on with three total team-mates. The lab will cover some of the fundamentals of 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.
- 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 What does a hypothesis test do?

- What are the two possible outcomes of a hypothesis test?
- What are the four-possible combinations of (a) hypothesis test result; and (b) state of the world?
- Does a hypothesis test always have to report a result that is consistent with the state of the world? What does it mean if it *does*, and what does it mean if it *does not*.
- What if you made up your own testing framework, called the {Your Last Name's} Groundhog test. Which is literally a groundhog looking to see its shadow. Because you made the test, suppose that you know that it is *totally random* whether a groundhog sees its shadow. *How useful would this test be at separating states of the world?*

- What guarantee do you get if you follow the decision rules properly?
- Why do we standardize the mean to create a test statistic?

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

6.5 Madlib prompt

```
tone_of_voice     <- ''
mode_of_speech    <- ''
superlative       <- ''
score_on_test     <- 'percent' # should end with percent
name_of_classmate <- ''
emotion           <- ''
eating_verb       <- '' #slurp
vessel            <- ''
thing_found_in_compost <- ''
```

6.6 Madlib completed

Suppose that a classmate comes to you, and, in a voice , "Hey, I've got something that is for statistics test preperation. All you've got to do to get percent on Test 2 and make is to this of .

You're skeptical, but also curious because that last test was tough. Good.

6.7 "Accepting the Null"

(For the purposes of this class, and while you're talking about testing after the class: the preferred language is to either (a) *Reject the null hypothesis*, or (b) *Fail to reject the null hypothesis.*

Acknowledging that we're only 40% of the way through the course, you decide to hire a hungry, underpaid PhD student the School to conduct the experiment to evaluate this claim. They report back, no details about the test, but they do tell you, "We're sure there's no effect of ."

- Do you believe them?
- What, if any, reasons can you imagine not to believe this conclusion?

6.8 Manually Computing a t-Test

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

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

```
## [1] 11.77 11.90 11.64 11.84 12.13 11.99 11.77
```

1. Find the mean and the standard deviation.

```
sample_mean <- mean(voltage)
sample_sd   <- sd(voltage)
n           <- length(voltage)

test_statistic <- (sample_mean - 12) / (sample_sd / sqrt(n))
test_statistic
```

```
## [1] -2.247806
```

2. Using `qt()`, compute the t critical value for a hypothesis test for this sample.

```
qt(0.025, df=n-1)
```

```
## [1] -2.446912
```

3. Define a test statistic, t, for testing whether the population mean is 12.

```
test_statistic
```

```
## [1] -2.247806
```

4. Calculate the p-value using the t statistic.

```
pt(test_statistic, df=n-1)
```

```
## [1] 0.03281943
```

5. Should you reject the null? Argue this in two different ways. (Following convention, set $\alpha = .05$.)

```
test_stat_function <- function(data, null_hypothesis) {
  sample_mean <- mean(data)
  sample_sd   <- sd(data)
  n           <- length(data)

  test_statistic <- (sample_mean - null_hypothesis) / (sample_sd / sqrt(n))
  return(test_statistic)
}
```

```
test_stat_function(data=voltage, null_hypothesis=12) %>%
  pt(df=length(voltage)-1) * 2
```

```
## [1] 0.06563885
```

```
t.test(
  x           = voltage,
  alternative = 'two.sided',
  mu          = 12)
```

```
##
##  One Sample t-test
##
## data:  voltage
## t = -2.2478, df = 6, p-value = 0.06564
## alternative hypothesis: true mean is not equal to 12
## 95 percent confidence interval:
##  11.71357 12.01215
## sample estimates:
## mean of x
##  11.86286
```

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.

6.9 Falling Ill (The General Form of a Hypothesis Test)

In the async content for the week, we're really, *really* clear that we're only working with the *t-distribution*. But, the general "form" of a frequentist hypothesis test is always the same: produce a test statistic; produce a distribution of that test statistic if the null hypothesis *were* true; then compare the two. Let's stretch this application a little bit.

There is a theory that upcoming tests cause students to fall ill. We have been collecting wellness data from our students for several years (not really...) and we have found the following distribution of illnesses (Notice that this does not tell you anything about how many students we have enrolled over the years):

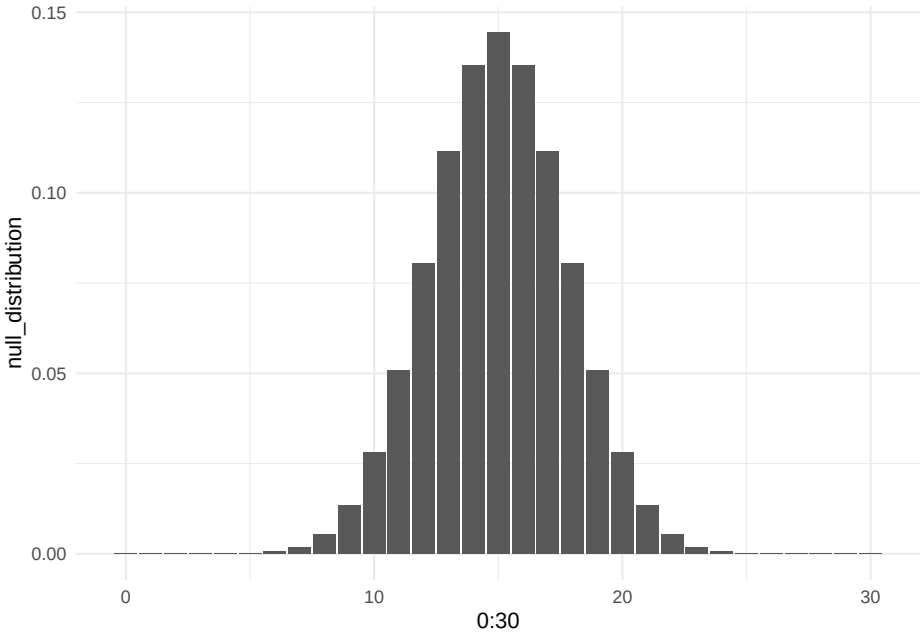
- 20 students have reported being ill in the week *before* Test 2

- 10 students have reported being ill in the week *after* Test 2

Think of wellness/illness as a dichotomous statement.

- **State an appropriate null hypothesis.** After you have stated this null hypothesis, can you think about (or, even better) can you produce a distribution of the probability of {0, 1, 2, 3, ... 30} of the illnesses reported before the test?

```
null_distribution <- dbinom(0:30, 30, prob = 0.5)

ggplot() +
  aes(x=0:30, y=null_distribution) +
  geom_col()
```



- **State a rejection criteria**. What occurrence in the data would cause you do doubt the plausibility of your null hypothesis?
- **What do you conclude?** Given the data that is presented to you and the null hypothesis, what do you conclude?

6.10 Data Exercise

t-Test Micro Cheat Sheet

In order for a t-test to produce valid results, a set of conditions must be satisfied. While the literature refers to these as *assumptions*, you might do better to refer to these for yourselves as *requirements*. Meaning, if these requirements for the data generating process are not satisfied, the test does not produce results that hold any statistical guarantees.

- **Metric variable**: The data needs to be numeric
- **IID**: The data needs to be sampled using an independent, identically distributed sampling process.
- **Well-behaved**: The data need to demonstate no major deviations from normality, considering sample size

Testing the Home Team Advantage

The file `./data/home_team.csv` 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.

```
home_team <- read.csv('./data/home_team.csv') %>%
  select(dscore, dinstt, doutstt) %>%
  rename(
    score_diff               = dscore,
    in_state_tuition_diff    = dinstt,
    out_state_tuition_diff   = doutstt
  )

glimpse(home_team, width = 80)
```

```
## Rows: 30
## Columns: 3
## $ score_diff             <int> 10, -14, 23, 8, -12, 7, -21, -5, -3, -32, 9, 1,~
## $ in_state_tuition_diff  <int> -409, NA, -654, -222, -10, 494, 2, 96, 223, -20~
## $ out_state_tuition_diff <int> -4679, -66, -637, 456, 208, 17, 2, -333, 2526, ~
```

We are especially interested in the variable, `score_diff`, 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?

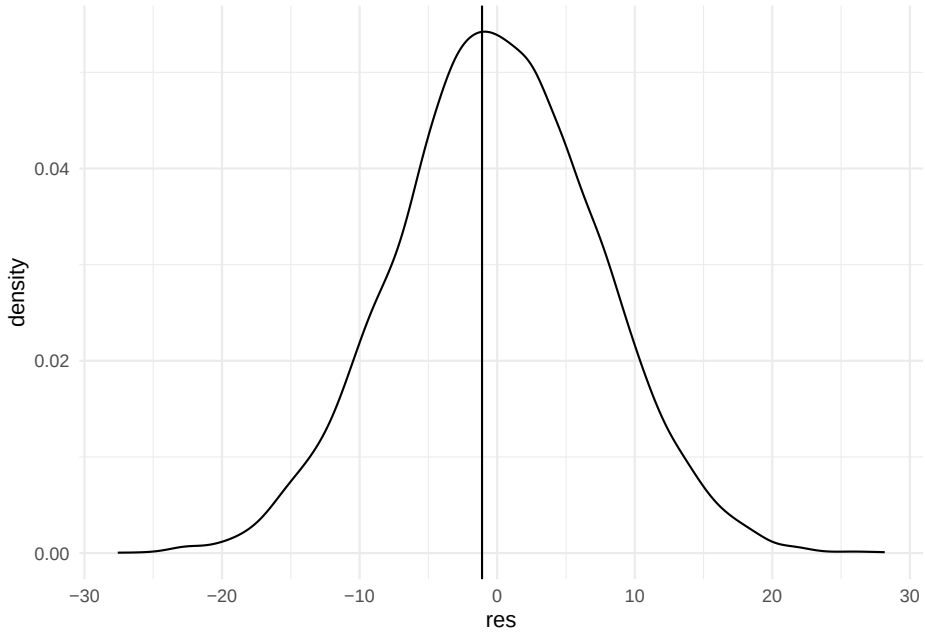
```
## I'm going two-tailed.
## H0 : No effect of being home or away
## HA : There IS some effect.
```

3. Execute the t-test and interpret every component of the output.

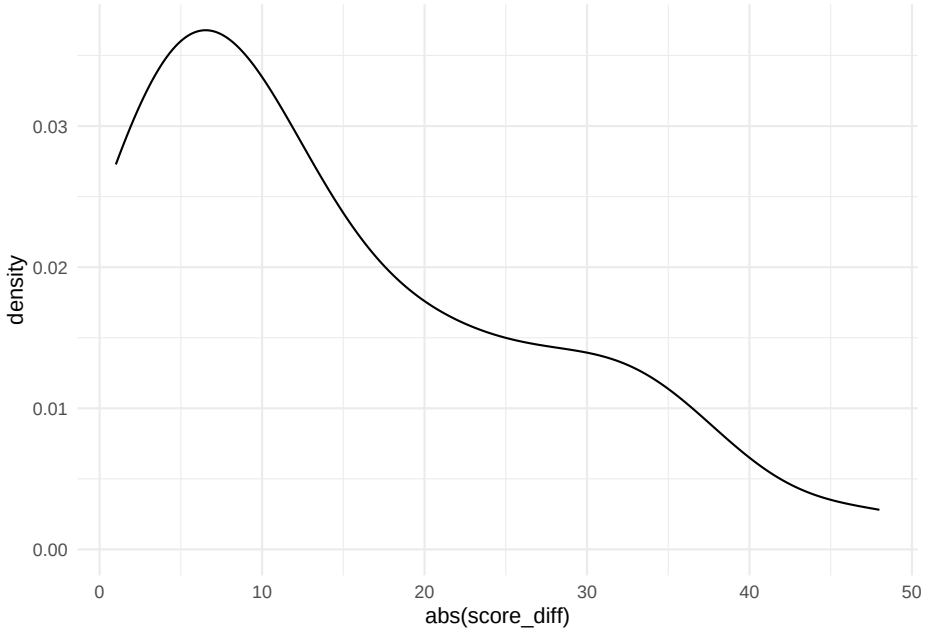
```
t.test(x=home_team$score_diff, mu=0, alternative = 'two.sided')
```

```
##
##  One Sample t-test
##
## data:  home_team$score_diff
## t = -0.30781, df = 29, p-value = 0.7604
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -8.408919  6.208919
## sample estimates:
## mean of x
##      -1.1
```

```
res <- NA
for(i in 1:10000) {
  res[i] <- mean(rnorm(n=7, sd=sd(home_team$score_diff)))
}

ggplot() +
  aes(x=res) +
  geom_density() +
  geom_vline(xintercept=mean(home_team$score_diff))
```




```
home_team %>%
  ggplot() +
  aes(x=abs(score_diff)) +
  geom_density()
```




```
mean(home_team$score_diff)
```

```
## [1] -1.1
```

```
mean((res < mean(home_team$score_diff))) + mean(res > abs(mean(home_team$score_diff)))
```

```
## [1] 0.881
```

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.

6.11 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

This week, we introduce the idea of comparing two groups to evaluate whether the data that we have sampled lead us to believe that the population distribution of the random variables are different. Of course, because we don't get access to the function that describes the random variable, we can't *actually* know if the populations are different. It is for this reason that we call it statistical inference – we are inferring from a sample some belief about the populations.

At the conclusion of this week, students will be able to:

1. *Recognize* the similarities between all frequentist hypothesis tests.
2. *Evaluate* the conditions that surround the data, and choose a test that is best-powered and justifiable.
3. *Perform* and *Interpret* the results of the most common statistical tests.

7.2 Class Announcements

- Great work completing your final w203 test!
- Unit 7 Homework is Group Homework, due next week.
- The Hypothesis Testing Lab is released today!
  - Lab is due at Unit 09 Live Session (two weeks): Apply these fundamentals to analyze 2022 election data and write a single, three-page analysis

7.3 Roadmap

7.3.1 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.

7.3.2 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.

7.3.3 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

7.4.1 Working on Data Science Teams

Data science is a *beautiful* combination of team-work and individual-work. It provides the opportunity to work together on a data pipeline with people from all over the organization, to deal with technical, statistical, and social questions that are always interesting. While we expect that everyone on a team will be a professional, there is so much range within the pursuit of data science that projects are nearly always collaborative exercises.

Together as teams, we

- Define research ambitions and scope
- Imagine/envision the landscape of what is possible
- Support, discuss, review and integrate individual contributions

Together as individuals we conduct the heads-down work that moves question answering forward. This might be reading papers to determine the most appropriate method to bring to bear on the question, or researching the data that is available, or understanding the technical requirements that we have to meet for this answer to be useful to our organization in real time.

What is your instructor *uniquely* capable of? Let them tell you! But, at the same time, what would they acknowledge that others are better than them?

See, the thing is, because there is so much that has to be done, there literally are very, very few people who are one-stop data science shops. Instead, teams rely on collaboration and joint expertise in order to get good work done.

7.4.2 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, these types of emotions are also commonplace in journalism, film-making, and public speaking.

Has anybody ever had the dream that they're late to a test? Or, that that they've got to give a speech that they're unprepared for? Does anybody remember playing an instrument as a kid and having to go to recitals? Or, play for a championship on a youth sports team? Or, go into a test two?

What are the feelings associated with those events? What might be generating these feelings?

7.4.3 What Makes an Effective Team?

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

  What really matters to creating an effective tema 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. If so, please read the following sections:

- The problem statement;
- The proposed solution;
- 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."

7.4.4 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?

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

7.7 Rank Based Tests

Darrin Speegle has a nice talk-through, walk through of rank based testing procedures, linked here. We'll talk through a few examples of this, and then move to estimating against data for the class.

7.8 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.

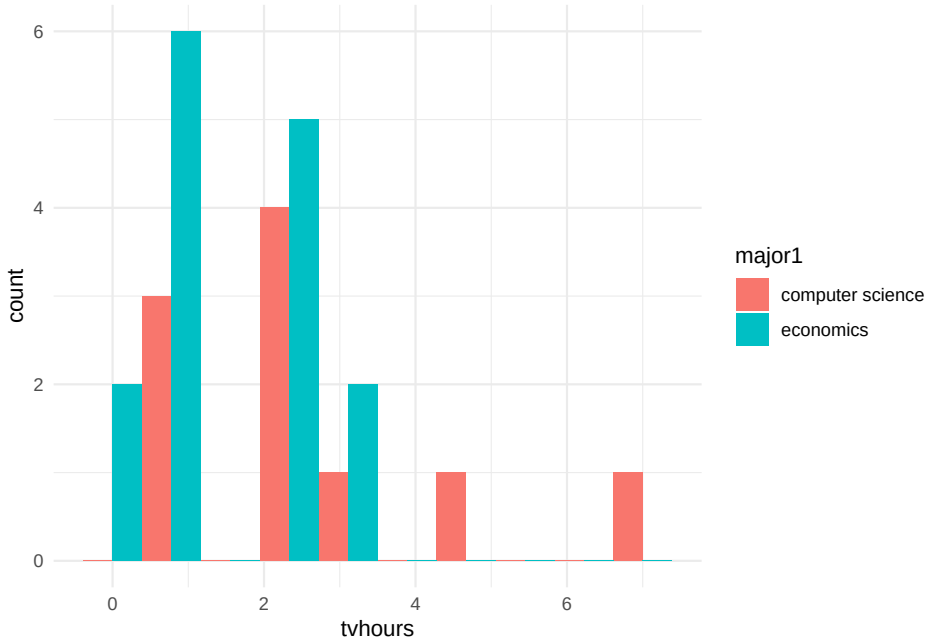
7.9 The Questions

7.9.1 Set 1

- Do economics majors watch more or less TV than computer science majors?

```
GSS %>%
  filter(major1 %in% c('computer science', 'economics')) %>%
  ggplot() +
  aes(x = tvhours, fill = major1) +
  geom_histogram(bins = 10, position = 'dodge')
```

```
## Warning: Removed 11 rows containing non-finite values (`stat_bin()`).
```



What kinds of tests *could* be reasonable to conduct? For what part of the data would we conduct these tests?

```
## The assumptions about the data drive us to the correct test.
## But, let's ask all the tests that could *possibly* make sense, and see how
##     matching or mis-matching assumptions changes what we learn.
```

```
## Answers are in the next chunk... but don't jump to them right away.
```

- Do Americans with pets watch more or less TV than Americans without pets?

7.9.2 Set 2

- Do Americans spend more time emailing or using the web?

```
GSS %>%
  select(wwwhr, emailhr) %>%
  drop_na() %$%
  t.test(x = wwwhr, y = emailhr, paired = TRUE)
```

```
##
##  Paired t-test
##
## data:  wwwhr and emailhr
## t = 13.44, df = 1360, p-value < 2.2e-16
## alternative hypothesis: true mean difference is not equal to 0
## 95 percent confidence interval:
##  5.530219 7.420553
## sample estimates:
## mean difference
##        6.475386
```

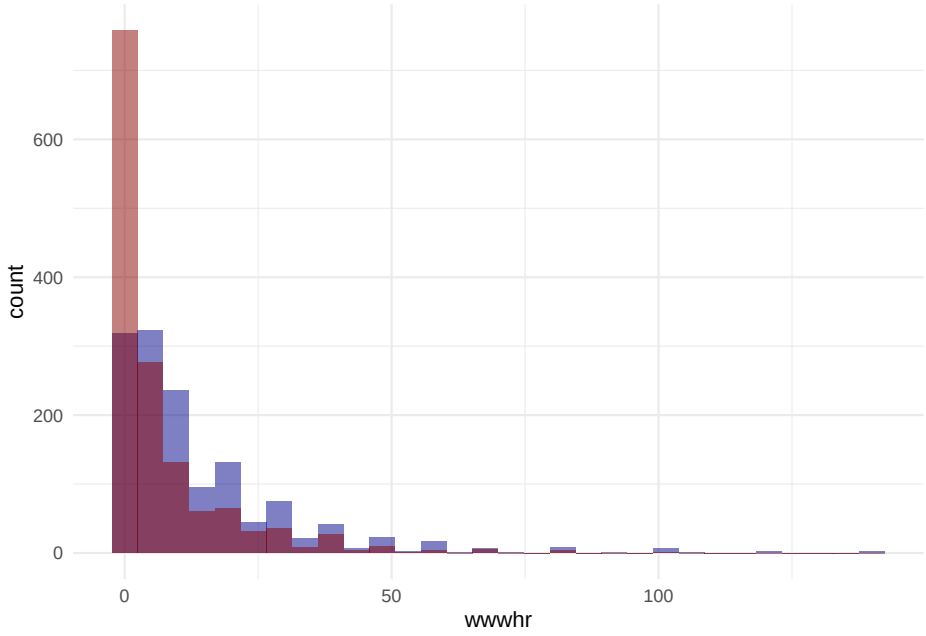
```
GSS %>%
  ggplot() +
  geom_histogram(aes(x = wwwhr),   fill = 'darkblue', alpha = 0.5) +
  geom_histogram(aes(x = emailhr), fill = 'darkred',  alpha = 0.5)
```

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

```
## Warning: Removed 986 rows containing non-finite values (`stat_bin()`).
```

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

```
## Warning: Removed 929 rows containing non-finite values (`stat_bin()`).
```




```
t.test(
  x = GSS$wwwhr,
  y = GSS$emailhr,
  paired = FALSE
)
```

```
##
##  Welch Two Sample t-test
##
## data:  GSS$wwwhr and GSS$emailhr
## t = 12.073, df = 2398.5, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  5.657397 7.851614
## sample estimates:
## mean of x mean of y
## 13.906021  7.151515
```

- Do Americans spend more evenings with neighbors or with relatives?

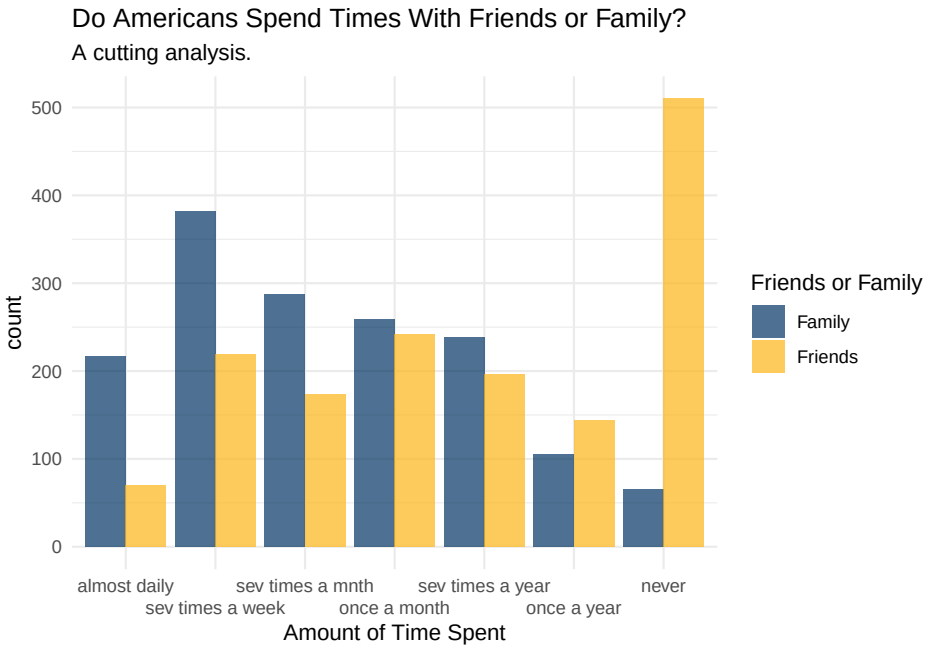
```
wilcox_test_data <- GSS %>%
  select(socrel, socommun) %>%
  mutate(
    family_ordered = factor(
      x      = socrel,
      levels = c('almost daily', 'sev times a week',
```

```
                'sev times a mnth', 'once a month',
                'sev times a year', 'once a year', 'never')),
    friends_ordered = factor(
      x      = socommun,
      levels = c('almost daily', 'sev times a week',
                 'sev times a mnth', 'once a month',
                 'sev times a year', 'once a year', 'never')))
```

To begin this investigation, we've got to look at the data and see what is in it. If you look below, you'll note that it sure seems that people are spending more time with their family... erp, actually no. They're "hanging out" with their friends rather than taking their mother out to dinner.

```
wilcox_test_data %>%
  select(friends_ordered, family_ordered) %>%
  rename(
    Friends = friends_ordered,
    Family  = family_ordered
  ) %>%
  drop_na() %>%
  pivot_longer(cols = c(Friends, Family)) %>%
  ggplot() +
    aes(x=value, fill=name) +
    geom_histogram(stat='count', position='dodge', alpha=0.7) +
  scale_fill_manual(values = c('#003262', '#FDB515')) +
  labs(
    title    = 'Do Americans Spend Times With Friends or Family?',
    subtitle = 'A cutting analysis.',
    fill     = 'Friends or Family',
    x        = 'Amount of Time Spent') +
  scale_x_discrete(guide = guide_axis(n.dodge = 2)) +
  theme_minimal()
```

```
## Warning in geom_histogram(stat = "count", position = "dodge", alpha = 0.7):
## Ignoring unknown parameters: `binwidth`, `bins`, and `pad`
```



With this plot created, we can ask if what we observe in the plot is the produce of what could just be sampling error, or if this is something that was unlikely to arise due if the null hypothesis were true. What is the null hypothesis? Well, lets suppose that if we didn't know anything about the data that we would expect there to be no difference between the amount of time spent with friends or families.

```
## risky choice -- casting the factor to a numeric without checking what happens.
wilcox_test_data %$%
  wilcox.test(
    x = as.numeric(family_ordered),
    y = as.numeric(friends_ordered),
    paired = FALSE
  )
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  as.numeric(family_ordered) and as.numeric(friends_ordered)
## W = 716676, p-value < 2.2e-16
## alternative hypothesis: true location shift is not equal to 0
```

7.9.3 Set 3

- Are Americans that own guns or Americans that don't own guns more likely to have pets?

- Are Americans with pets happier than Americans without pets?

7.9.4 Apply to a New Type of Data

- Is there a relationship between college major and gun ownership?

7.10 Simulating the Effects of Test Choices

```
theme_set(theme_minimal())

berkeley_blue   <- '#003262'
berkeley_gold   <- '#FDB515'
berkeley_sather <- '#B9D3B6'
```

7.10.1 Should we use a t-test or a wilcox sign-rank?

There is some open discussion in the applied statistics literature about whether we should *ever* be using a t-test. In particular, if the underlying distribution that generates the data is **not** normal, than the assumptions of a t-test are not, technically satisfied and the test does not produce results that have nominal p-value coverage. This is both *technically* and *theoretically* true; and yet, researchers, data scientists, your instructors, and the entire world runs t-tests as "test of first recourse."

What is the alternative to conducting a t-test as the test of first recourse? It might be the Wilcox test. The Wilcox test makes a weaker assumption – of symmetry around the mean or median – which is weaker than the assumption of normality.

Additional points of argument, which you will investigate in this worksheet:

- If the underlying data **is** normal, then the Wilcox test is *nearly* as well powered as the t-test.
- If the underlying data **is not** normal, then the Wilcox test still maintains nominal p-value coverage, whereas the t-test might lose this guarantee.

7.11

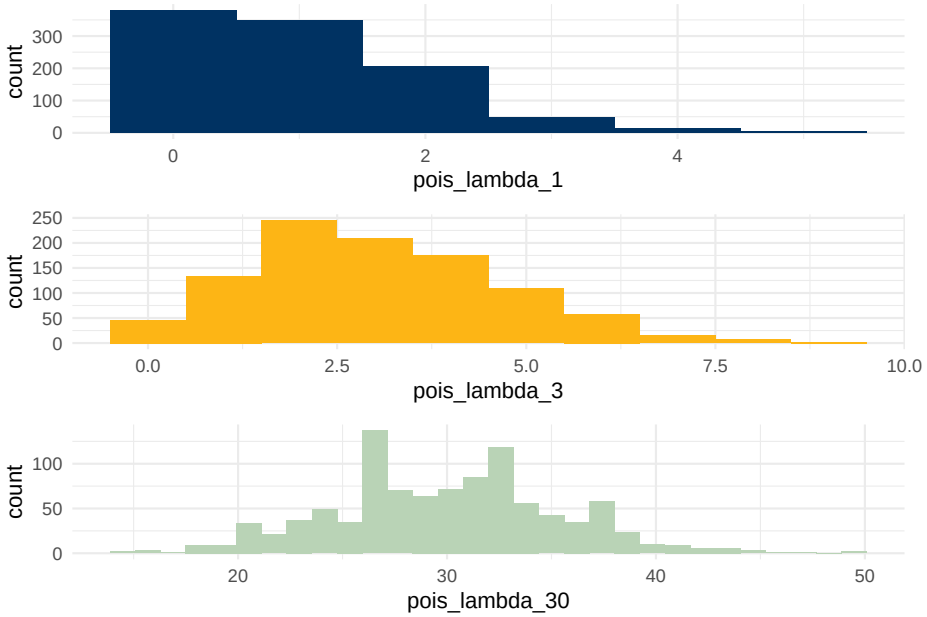
7.11.1 The Poisson Distribution

The poisson distribution has the following PDF:

$$f_X(x) = \frac{\lambda^n e^{-\lambda}}{n!}$$

The key shape parameter for a poisson function is λ; we show three different distributions, setting this shape parameter to be 1, 3, and 30 respectively. Notice that the limits on these plots are not set to be the same; for example, the range in the third plot is considerably larger than the first.

```
pois_lambda_1  <- rpois(n=1000, lambda=1)
pois_lambda_3  <- rpois(n=1000, lambda=3)
pois_lambda_30 <- rpois(n=1000, lambda=30)

plot_1  <- ggplot() + aes(x=pois_lambda_1) + geom_histogram(bins=6, fill = berkeley_blue)
plot_3  <- ggplot() + aes(x=pois_lambda_3) + geom_histogram(bins=10, fill = berkeley_gold)
plot_30 <- ggplot() + aes(x=pois_lambda_30) + geom_histogram(bins=30, fill = berkeley_sather)

plot_1 / plot_3 / plot_30
```



What does this changing distribution do to the p-values?

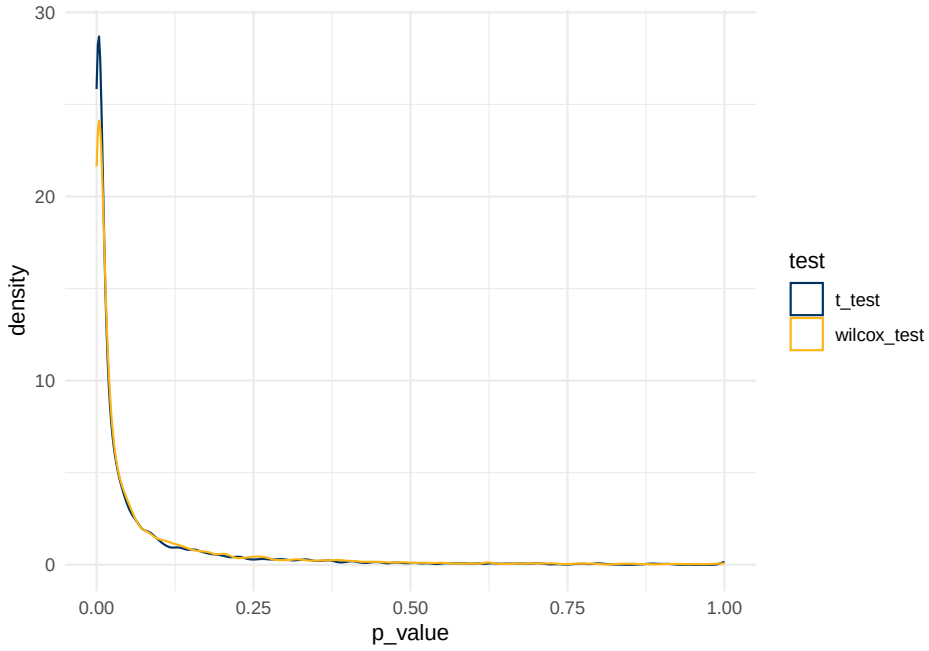
7.11.2 Write a Simulation

```
pois_sim <- function(num_observations, lambda_one, lambda_two) {

  t_test_result <- rep(NA, 10000)
  wilcox_result <- rep(NA, 10000)

  for(i in 1:10000) {
    group_one <- rpois(n=num_observations, lambda=lambda_one)
```

```
    group_two <- rpois(n=num_observations, lambda=lambda_two)

    t_test_result[i] <- t.test(group_one, group_two)$p.value
    wilcox_result[i] <- wilcox.test(x=group_one, y=group_two)$p.value
  }

  df <- data.table(
    p_value = c(t_test_result, wilcox_result),
    test    = rep(c('t_test', 'wilcox_test'), each = 10000)
  )

  return(df)
}
```

```
foo <- pois_sim(20, 1, 2.0)
```

```
foo %>%
  ggplot() +
  geom_density(aes(x=p_value, color = test)) +
  scale_color_manual(values = c(berkeley_blue, berkeley_gold))
```



And so, the simulation rejects the null at the following rates:

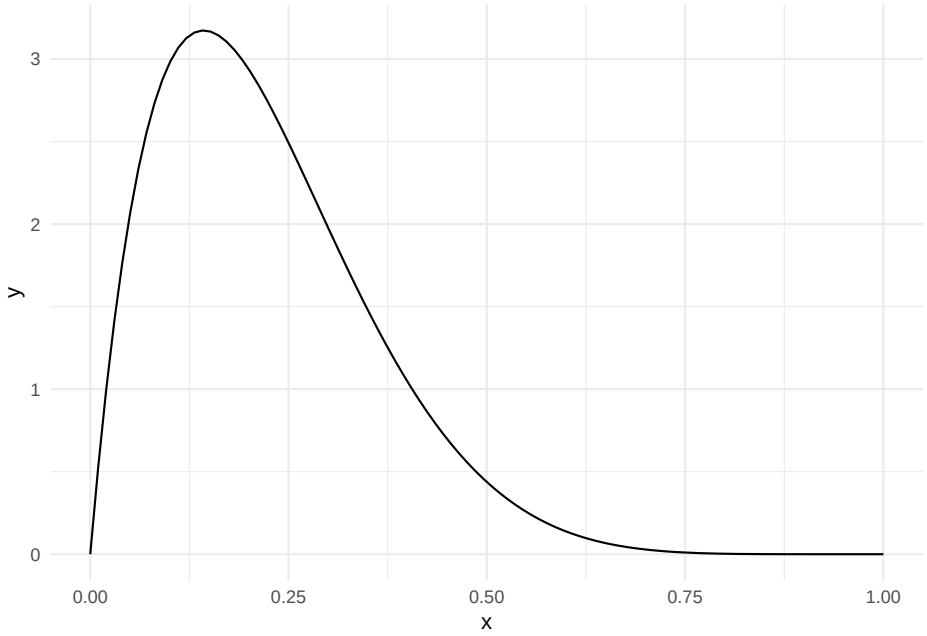
- For the t-test, at a rate of 0.0650033
- For the Wilcox test, at a rate of 0.0748107

```r
skewed_sim <- function(num_sims=1000, num_observations, alpha_1, beta_1, alpha_2, beta_2) {

  t_test_result <- rep(NA, num_sims)
  wilcox_result <- rep(NA, num_sims)

  for(i in 1:num_sims) {
    group_one <- rbeta(n=num_observations, shape1 = alpha_1, shape2 = beta_1)
    group_two <- rbeta(n=num_observations, shape1 = alpha_2, shape2 = beta_2)

    t_test_result[i] <- t.test(group_one, group_two)$p.value
    wilcox_result[i] <- wilcox.test(x=group_one, y=group_two)$p.value
  }

  dt <- data.table(
    p_value = c(t_test_result, wilcox_result),
    test    = rep(c('t_test', 'wilcox_test'), each = num_sims)
  )

  return(dt)
}
```

7.11.4 False Rejection Rates

Start with a distribution that has parameters `alpha=2`, `beta=7`.

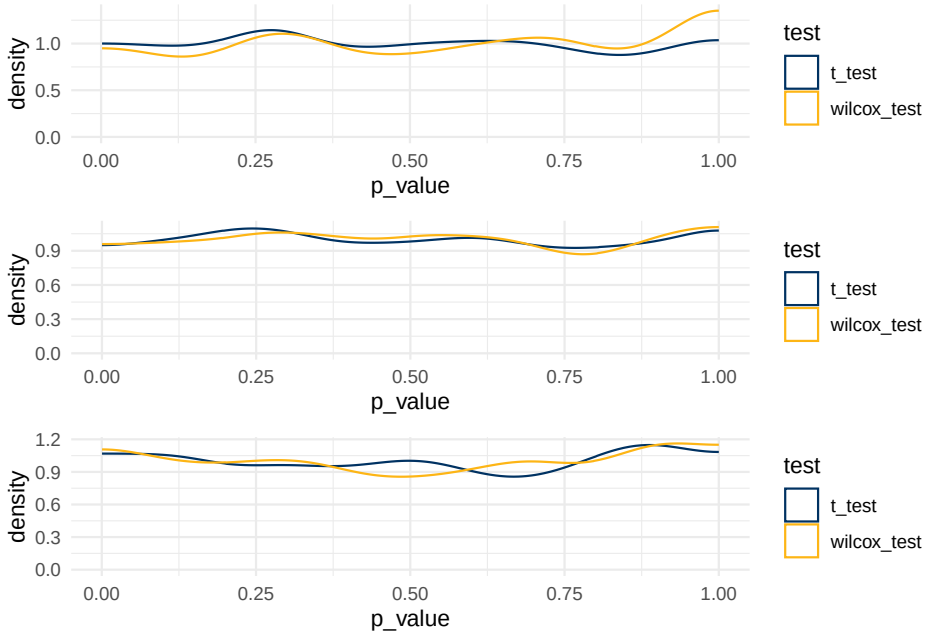
```r
ggplot(data.frame(x=c(0,1)), aes(x)) +
  stat_function(fun = dbeta, n=100, args=list(shape1=2, shape2=7))
```




```
same_dist_small_data <- skewed_sim(
  num_observations=10,
  alpha_1=2, beta_1=7,
  alpha_2=2, beta_2=7
  )
same_dist_med_data <- skewed_sim(
  num_observations=50,
  alpha_1=2, beta_1=7,
  alpha_2=2, beta_2=7
  )
same_dist_big_data <- skewed_sim( # haha, "big data"
  num_observations=100,
  alpha_1=2, beta_1=7,
  alpha_2=2, beta_2=7
  )
```

```
plot_1 <- same_dist_small_data %>%
  ggplot() +
  geom_density(aes(x=p_value, color = test), bounds=c(0,1)) +
  scale_color_manual(values = c(berkeley_blue, berkeley_gold))
plot_2 <- same_dist_med_data %>%
  ggplot() +
  geom_density(aes(x=p_value, color = test), bounds=c(0,1)) +
  scale_color_manual(values = c(berkeley_blue, berkeley_gold))
plot_3 <- same_dist_big_data %>%
```

```
  ggplot() +
  geom_density(aes(x=p_value, color = test), bounds=c(0,1)) +
  scale_color_manual(values = c(berkeley_blue, berkeley_gold))

plot_1 / plot_2 / plot_3
```



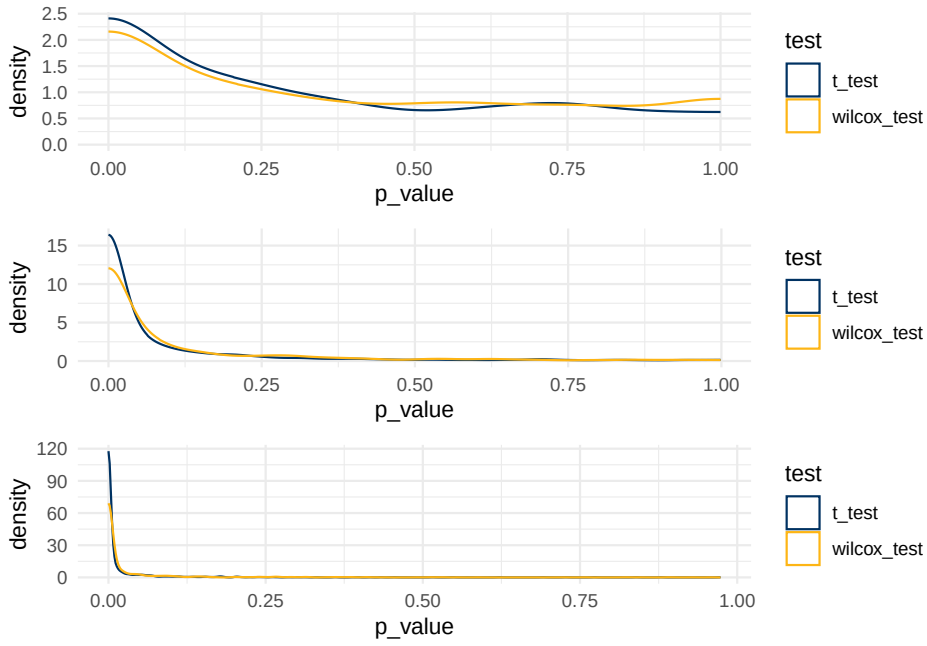
- T-tests
  - 0.055
  - 0.046
  - 0.056
- Wilcox Tests
  - 0.046
  - 0.048
  - 0.061

7.11.5 What about Power to Reject

```
small_diff_small_data <- skewed_sim(
  num_observations=10,
  alpha_1=2, beta_1=7,
  alpha_2=2, beta_2=5
  )
small_diff_med_data <- skewed_sim(
  num_observations=50,
```

```
  alpha_1=2, beta_1=7,
  alpha_2=2, beta_2=5
  )
small_diff_big_data <- skewed_sim( # haha, "big data"
  num_observations=100,
  alpha_1=2, beta_1=7,
  alpha_2=2, beta_2=5
  )
```

```
plot_1 <- small_diff_small_data %>%
  ggplot() +
  geom_density(aes(x=p_value, color = test), bounds=c(0,1)) +
  scale_color_manual(values = c(berkeley_blue, berkeley_gold))
plot_2 <- small_diff_med_data %>%
  ggplot() +
  geom_density(aes(x=p_value, color = test), bounds=c(0,1)) +
  scale_color_manual(values = c(berkeley_blue, berkeley_gold))
plot_3 <- small_diff_big_data %>%
  ggplot() +
  geom_density(aes(x=p_value, color = test), bounds=c(0,1)) +
  scale_color_manual(values = c(berkeley_blue, berkeley_gold))

plot_1 / plot_2 / plot_3
```



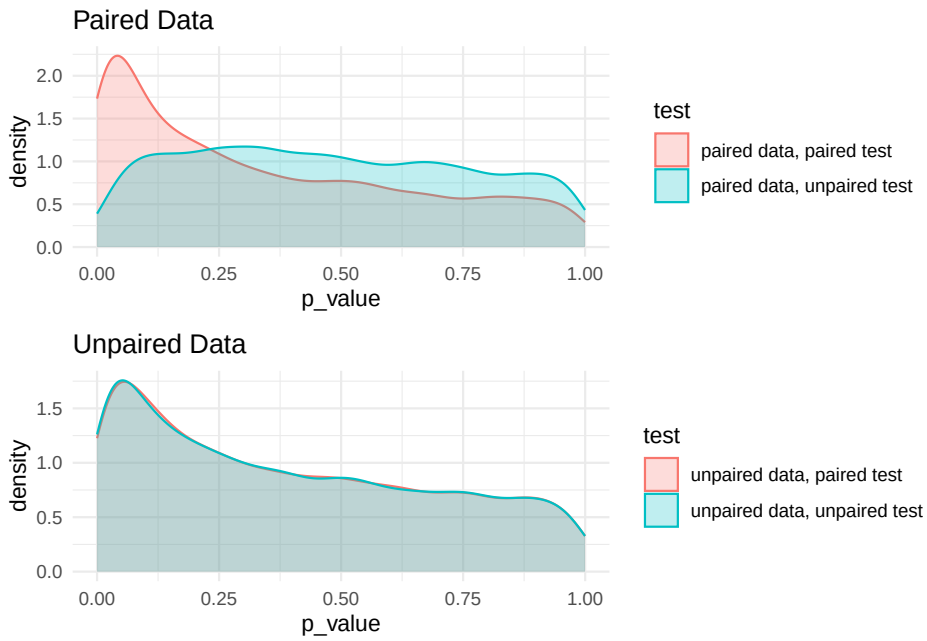
7.11.6 Paired compared to unpaired tests

```
paired_sim <- function(num_sims=10000, num_observations, mean_one, mean_two, paired_diff, sd_one,

  unpaired_test_unpaired_data <- rep(NA, num_sims)
  unpaired_test_paired_data   <- rep(NA, num_sims)
  paired_test_unpaired_data   <- rep(NA, num_sims)
  paired_test_paired_data     <- rep(NA, num_sims)

  for(i in 1:num_sims) {
    observation_a1 <- rnorm(n = num_observations, mean = mean_one, sd = sd_one)
    ## first create unpaired data
    observation_b <- rnorm(n = num_observations, mean = mean_two, sd = sd_two)
    ## then, create paired data
    observation_a2 <- observation_a1 + rnorm(n = num_observations, mean = paired_diff, sd=sd_two)

    ## run tests
    unpaired_test_unpaired_data[i] <- t.test(x=observation_a1, y=observation_b,  paired=FALSE)$p.
    unpaired_test_paired_data[i]   <- t.test(x=observation_a1, y=observation_a2,  paired=FALSE)$p
    paired_test_unpaired_data[i]   <- t.test(x=observation_a1, y=observation_b,  paired=TRUE)$p.v
    paired_test_paired_data[i]     <- t.test(x=observation_a1, y=observation_a2, paired=TRUE)$p.v
  }

  dt <- data.table(
    p_value = c(unpaired_test_unpaired_data, unpaired_test_paired_data,
                paired_test_unpaired_data,   paired_test_paired_data),
    test    = rep(c('unpaired data, unpaired test', 'paired data, unpaired test',
                    'unpaired data, paired test', 'paired data, paired test'), each = num_sims)
  )

  return(dt)
}
```

```
bar <- paired_sim(num_observations = 30,  mean_one = 10, mean_two = 11, paired_diff = 1, sd_one =
```

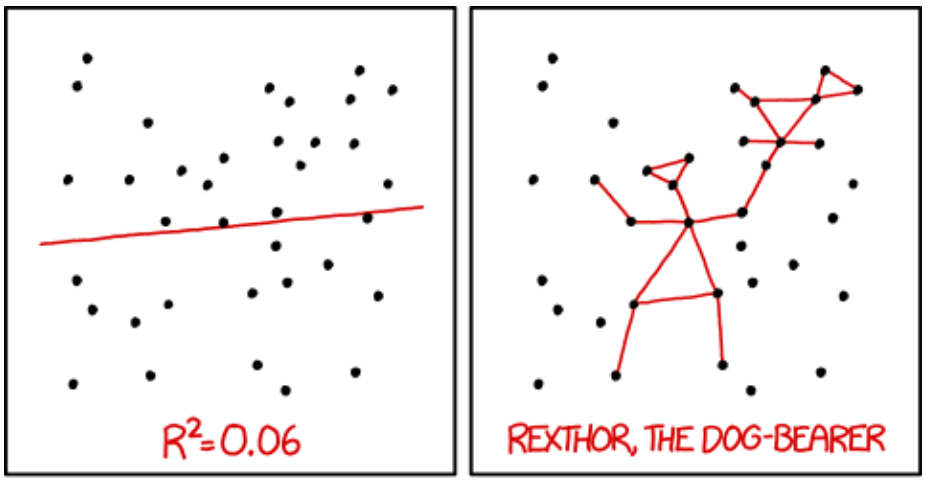
```
paired_data_plot <- bar[grep('unpaired data', test, invert=TRUE)] %>%
  ggplot() +
    aes(x=p_value, color = test, fill = test) +
    geom_density(alpha=0.25) +
  labs(title = 'Paired Data')
unpaired_data_plot <- bar[grep('unpaired data', test, invert=FALSE)] %>%
  ggplot() +
    aes(x=p_value, color = test, fill = test) +
    geom_density(alpha=0.25) +
  labs(title = 'Unpaired Data')
```

```
paired_data_plot / unpaired_data_plot
```



Chapter 8

OLS Regression Estimates




```
library(tidyverse)
library(broom)
library(testthat)
```

```
##
## Attaching package: 'testthat'
```

```
## The following objects are masked from 'package:magrittr':
##
##     equals, is_less_than, not
```

```
## The following object is masked from 'package:dplyr':
##
##     matches

## The following object is masked from 'package:purrr':
##
##     is_null

## The following objects are masked from 'package:readr':
##
##     edition_get, local_edition

## The following object is masked from 'package:tidyr':
##
##     matches
```

```
theme_set(theme_minimal())
```

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 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?

8.5 The Regression Anatomy Formula

We make the claim in live session that we can re-represent a coefficient that we're interested in as a function of all the other variable in a regression. That is, suppose that we were interested, initially, in estimating the model:

$$Y = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \hat{\beta}_3 X_3 + e$$

that we can produce an estimate for $\hat{\beta}_1$ by fitting this auxiliary regression,

$$X_1 = \hat{\delta}_0 + \hat{\delta}_2 X_2 + \hat{\delta}_3 X_3 + r_1$$

And then using the residuals, noted as r_1 above, in a second auxiliary regression,

$$Y = \gamma_0 + \gamma_1 r_1$$

The claim that we make in the live session is that there is a guarantee that $\beta_1 = \gamma_1$. Here, we are first going to show that this is true, and then we're going to reason about what this means, and why this feature is interesting (or at least useful) when we are estimating a regression.

Suppose that the population model is the following:

$$Y = -3 + (1 \cdot X_1) + (2 \cdot X_2) + (3 \cdot X_3)$$

```
d <- data.frame(
  x1 = runif(n=100, min=0, max=10),
  x2 = runif(n=100, min=0, max=10),
  x3 = runif(n=100, min=0, max=10)
)
```

```
## because we know the population model, we can produce a single sample from it
## using the following code:

d <- d %>%
  mutate(y = -3 + 1*x1 + 2*x2 + 3*x3 + rnorm(n=n(), mean=0, sd=1))

head(d)
```

```
##         x1       x2       x3        y
## 1 2.627549 6.111438 9.461474 39.71841
## 2 7.586222 5.339518 3.055064 24.80873
## 3 3.187276 8.444794 6.609802 35.97645
## 4 1.547751 3.650995 4.952615 19.91845
## 5 3.060822 1.922273 2.994374 11.68472
## 6 4.496636 3.194693 2.985301 16.76039
```

Notice that when we made this data, we included a set of random noise at the end. The idea here is that there are other "things" in this universe that also affect Y, but that we don't have access to them. By assumption, what we *have* measured in this world, X_1, X_2, X_3 are uncorrelated with these other features.

```
model_main <- lm(y ~ x1 + x2 + x3, data = d)
coef(model_main)
```

```
## (Intercept)          x1          x2          x3
##   -3.488590    1.062987    2.033696    3.012998
```

The claim is that we can produce an estimate of β_1 using an auxiliary set of regression estimates, and then using the regression from that auxiliary regression.

```
model_aux <- lm(x1 ~ x2 + x3, data = d)
```

If we look into the structure of `model_aux` we can see that there are *a ton* of pieces in here.

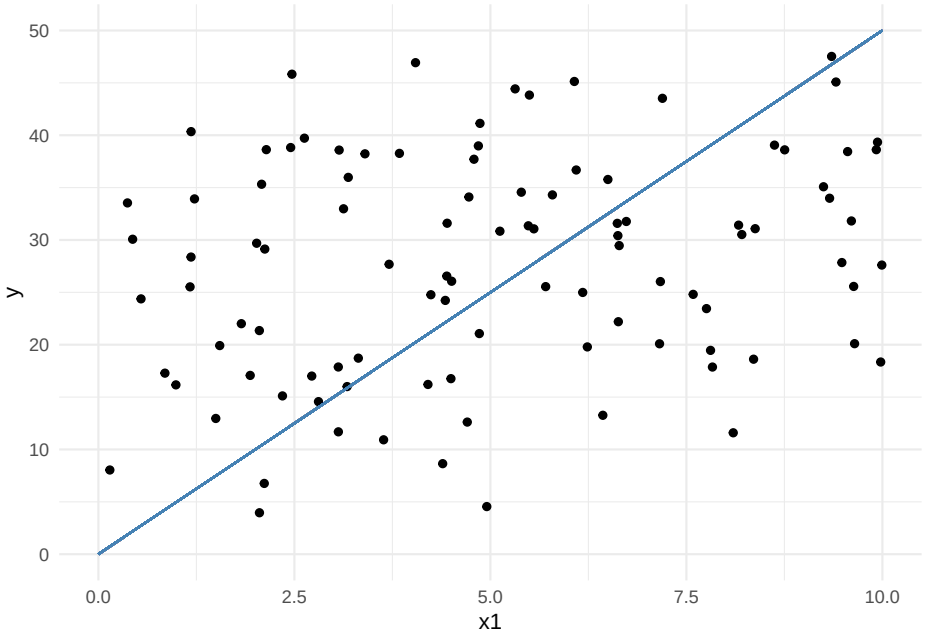
```
coef(model_aux)
```

```
## (Intercept)          x2          x3
##  5.37777970  0.04654908 -0.11105934
```

To evaluate our claim, we need to find the residuals from this regression. As a knowledge check, what is it that we mean when we say, "residual" in this sense?

To make talking about these easier, here is a plot that might be useful.

```
d %>%
  ggplot() +
  aes(x=x1, y=y) +
  geom_point() +
  geom_segment(aes(x=0, xend=10, y=0, yend=50), color = 'steelblue')
```



In order to access these residuals, we can "augment" the dataframe that we used in the model, using the `broom::augment` function call.

```
d_augmented <- augment(model_aux)
d_augmented$y <- d$y
d_augmented
```

```
## # A tibble: 100 x 10
##       x1    x2    x3 .fitted .resid   .hat .sigma  .cooksd .std.resid     y
##    <dbl> <dbl> <dbl>   <dbl>  <dbl>  <dbl>  <dbl>    <dbl>      <dbl> <dbl>
##  1 2.63   6.11  9.46    4.61 -1.98  0.0380   2.82 0.00679      -0.718 39.7
##  2 7.59   5.34  3.06    5.29  2.30  0.0161   2.82 0.00370       0.823 24.8
##  3 3.19   8.44  6.61    5.04 -1.85  0.0256   2.83 0.00388      -0.665 36.0
##  4 1.55   3.65  4.95    5.00 -3.45  0.0127   2.81 0.00652      -1.23  19.9
##  5 3.06   1.92  2.99    5.13 -2.07  0.0276   2.82 0.00527      -0.747 11.7
##  6 4.50   3.19  2.99    5.19 -0.698 0.0202   2.83 0.000430     -0.250 16.8
##  7 1.94   2.75  4.24    5.03 -3.10  0.0176   2.81 0.00736      -1.11  17.1
##  8 0.146  3.68  1.38    5.40 -5.25  0.0314   2.78 0.0388       -1.89   8.04
##  9 0.544  3.87  5.99    4.89 -4.35  0.0134   2.80 0.0109       -1.55  24.4
## 10 0.989  6.61  1.67    5.50 -4.51  0.0307   2.79 0.0279       -1.63  16.2
## # i 90 more rows
```

And finally, with this augmented data that has information from the model, we can estimate the model that includes only the residuals as predictors of Y.

```
model_two <- lm(y ~ .resid, data = d_augmented)
coef(model_two)
```

```
## (Intercept)       .resid
##   27.667081    1.062987
```

Our claim was that the coefficients from `model_main` and `model_two` should be the same.

```
test_that(
  'the model coefficients are equal',
  expect_equal(
    as.numeric(coef(model_main)['x1']),
    as.numeric(coef(model_two)['.resid']))
)
```

```
## Test passed
```

But, why is this an interesting, or at least useful, feature to appreciate?

8.6 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)
```

8.7 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
```

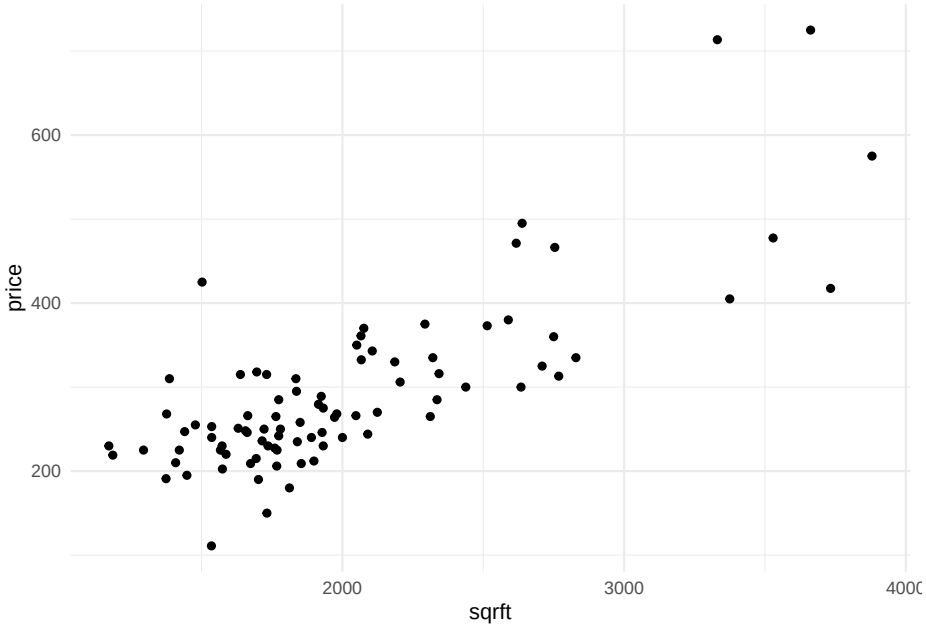
```
head(data)
```

```
##     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?

8.7.1 Assess the Relationship between Price and Square Footage

```
data %>%
  ggplot() +
  aes(x=sqrft, y=price) +
  geom_point()
```



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.

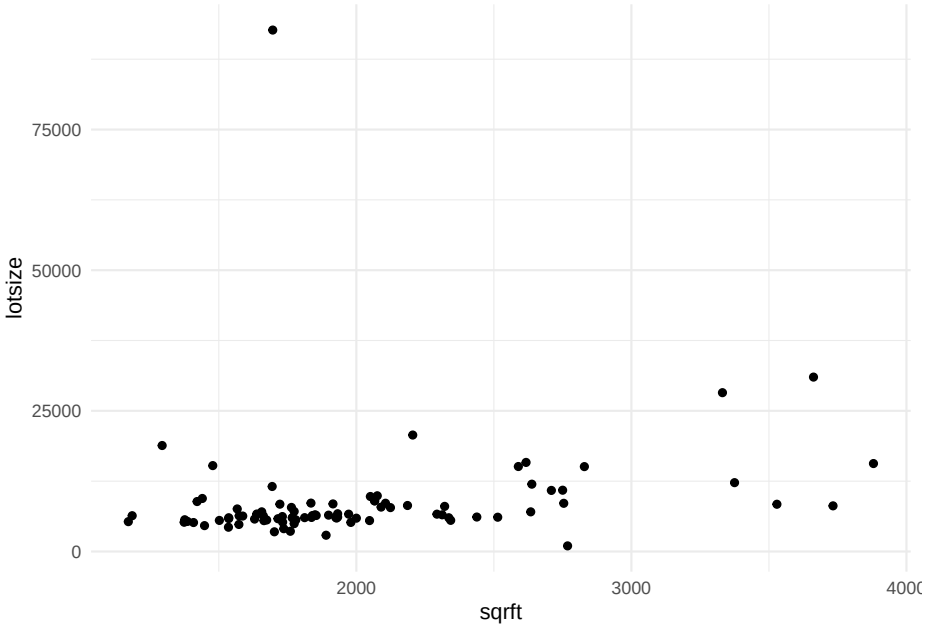
2. Find the correlation between the two variables.

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:

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

```
data %>%
  ggplot() +
  aes(x=sqrft, y=lotsize) +
  geom_point()
```



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?

8.8 Regression Plots and Discussion

In this next set of notes, we're going to give some data, displayed in plots, and we will try to apply what we have learned in the async and reading for this week

to answer questions about each of the scatter plots.

8.8.1 Plot 1

Consider data that is generated according to the following function:

$$Y = 1 + 2x_1 + 3x_2 + e,$$

where $x_1 \sim N(0, 2)$, $x_2 \sim N(0, 2)$ and e is a constant equal to zero.

From this population, you might consider taking a sample of 100 observations, and representing this data in the following 3d scatter plot. In this plot, there are three dimensions, an x_1, x_2, and y dimensions.

```
knitr::include_app(url ="http://www.statistics.wtf/minibeta01/")
```

1. Rotate the cube and explore the data, looking at each face of the cube, including from the top down.
2. One of the lessons that we learned during the random variables section of the course is that every random variable that has been measured can also be marginalized off. You might think of this as "casting down" data from three dimensions, to only two.
3. Sketch the following 2d scatter plots, taking care the label your axes. You need not represent all 100 points, but rather create the *gestalt* of what you see.
   1. $Y = f(x_1)$ (but not x_2) 2. $Y = f(x_2)$ (but not x_1) 3. $x2 = f(x_1)$
4. Once you have sketched the scatter plots, what line would you fit that minimizes the sum of squared residuals in the vertical direction. Define a residual, ϵ, to be the vertical distance between the line you draw, and the corresponding point on the input data.
5. What is the *average* of the residuals for each of the lines that you have fitted? How does this correspond to the *moment conditions* discussed in the async? What would happen if you translated this line vertically?
6. Rotate the cube so that the points "fall into line". When you see this line, how does it help you describe the function that governs this data?

Chapter 9

OLS Regression Inference

Figure 9.1: sunset on golden gate

9.1 Learning Objectives

After this week's learning, student will be able to

1. *Describe* how sampling based uncertainty is reflected in OLS regression parameter estimates.
2. *Report* standard errors, and *conduct* tests for NHST of regression coefficients against zero.

3. *Conduct* a regression based analysis, on real data, in ways that begin to explore regression as a modeling tool.

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 is 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)

9.5 Understanding Uncertainty

Imagine three different regression models, each of the following form:

$$Y = 0 + \beta X + \epsilon$$

The only difference is in the error term. The conditional distribution is given by:

Model	Distribution of ϵ cond. on X
A	Uniform on $[-.5, +.5]$
B	Uniform on $[-\|X\|, \|X\|]$
C	Uniform on $[-1 + \|X\|, 1 - \|X\|]$

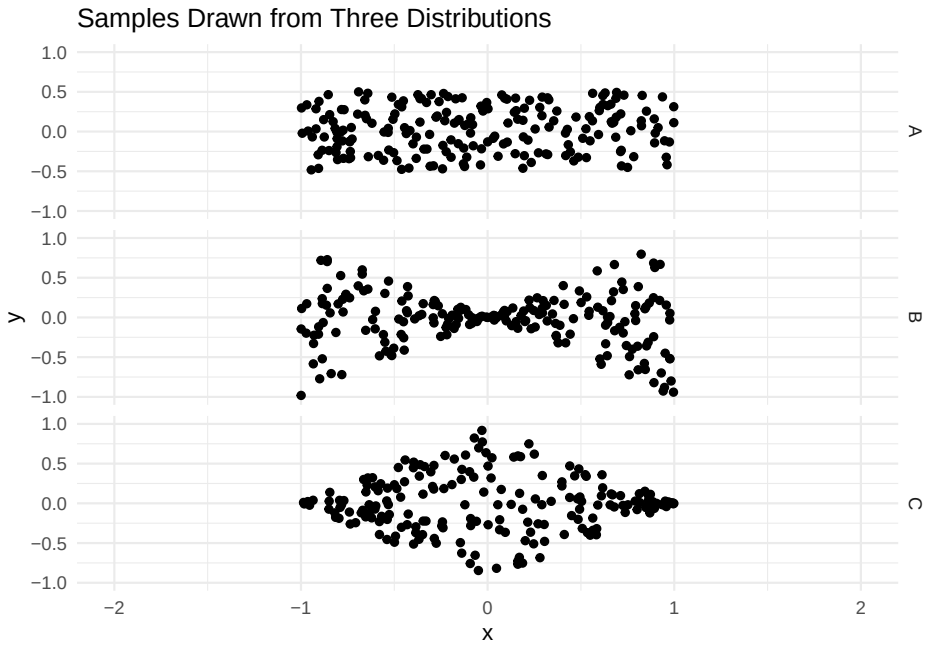
A is what we call a homoskedastic distribution. B and C are what we call heteroskedastic. Below, we define R functions that simulate draws from these three distributions.

```
rA <- function(n, slope=0){
  x       = runif(n, min=-1, max = 1)
  epsilon = runif(n, min=-.5, max=.5)
  y       = 0 + slope*x + epsilon
  return( data.frame(x=x,y=y) )
}

rB <- function(n, slope=0){
  x       = runif(n, min=-1, max = 1)
  epsilon = runif(n, min=- abs(x), max=abs(x))
  y       = 0 + slope*x + epsilon
  return( data.frame(x=x,y=y) )
}

rC <- function(n, slope=0){
  x       = runif(n, min=-1, max = 1)
  epsilon = runif(n, min= -1 + abs(x), max=1 - abs(x))
  y       = 0 + slope*x + epsilon
  return( data.frame(x=x,y=y) )
}
```

```
data <- rbind(
  data.frame( rA(200), label = 'A'),
  data.frame( rB(200), label = 'B'),
  data.frame( rC(200), label = 'C'))
```

```
data %>%
  ggplot(aes(x=x, y=y)) +
```

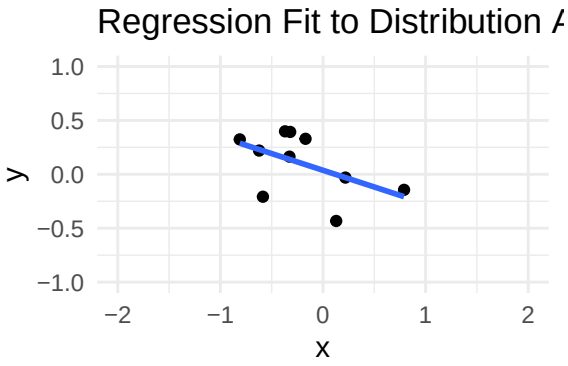
```
  geom_point() +
  lims(
    x = c(-2,2),
    y = c(-1,1)) +
  labs(title = 'Samples Drawn from Three Distributions') +
  facet_grid(rows=vars(label))
```



9.5.1 Question 1

The following code draws a sample from distribution A, fits a regression line, and plots it. Run it a few times to see what happens. Now explain how you would visually estimate the standard error of the slope coefficient. Why is this standard error important?

```
data <-  rA(10, slope=0)

data %>%
  ggplot() +
  aes(x=x, y=y) +
  geom_point() +
  geom_smooth(method='lm', formula = 'y ~ x', se=FALSE) +
  lims(
    x = c(-2,2),
    y = c(-1,1)) +
  labs(title = 'Regression Fit to Distribution A')
```

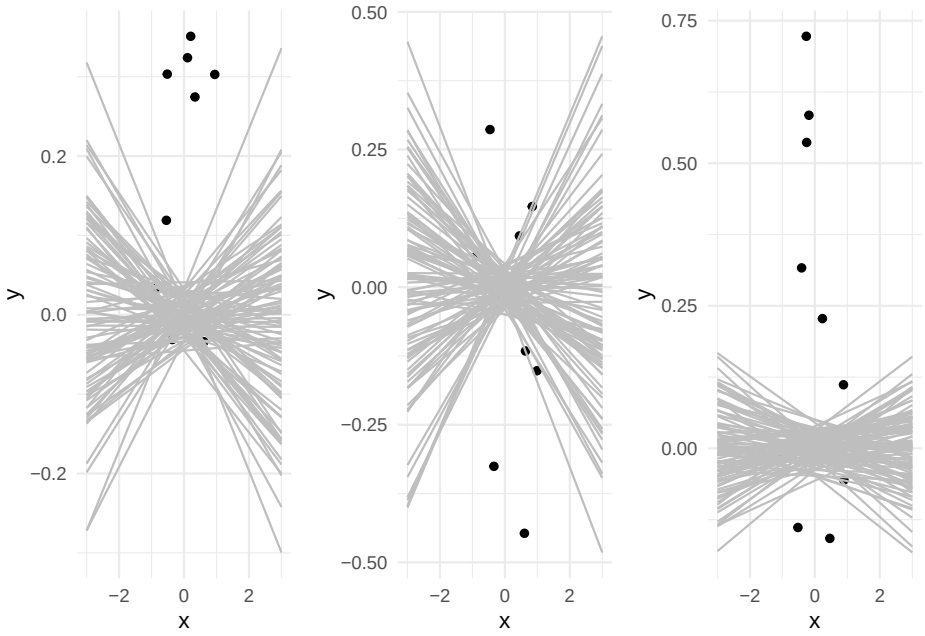



```
data_points <- 200

base_plot_a <- rA(10) %>%
  ggplot() +
  aes(x=x, y=y) +
  geom_point() +
  scale_x_continuous(limits = c(-3, 3))

for(i in 1:100) {
    base_plot_a <- base_plot_a + rA(data_points) %>%
      stat_smooth(
        mapping = aes(x=x, y=y),
        method  = 'lm',         se = FALSE,
        formula = 'y~x', fullrange = TRUE,
        color   = 'grey',    alpha = 0.5,
        size    = 0.5
      )
}
```

```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use `linewidth` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was generated.
```

```
base_plot_b <- rB(10) %>%
  ggplot() +
  aes(x=x, y=y) +
  geom_point() +
  scale_x_continuous(limits = c(-3, 3))

for(i in 1:100) {
    base_plot_b <- base_plot_b + rB(data_points) %>%
      stat_smooth(
        mapping = aes(x=x, y=y),
        method  = 'lm',         se = FALSE,
```

```
        formula = 'y~x', fullrange = TRUE,
        color   = 'grey',    alpha = 0.5,
        size    = 0.5
      )
  }

base_plot_c <- rC(10) %>%
  ggplot() +
  aes(x=x, y=y) +
  geom_point() +
  scale_x_continuous(limits = c(-3, 3))

for(i in 1:100) {
    base_plot_c <- base_plot_c + rC(data_points) %>%
      stat_smooth(
        mapping = aes(x=x, y=y),
        method  = 'lm',         se = FALSE,
        formula = 'y~x', fullrange = TRUE,
        color   = 'grey',    alpha = 0.5,
        size    = 0.5
      )
}

base_plot_a | base_plot_b  | base_plot_c
```



9.5.2 Question 2

You have a sample from each distribution, A, B, and C and you fit a regression of Y on X. Which will have the highest standard error for the slope coefficient? Which will have the lowest standard error? Why? (You may want to try experimenting with the function defined above)

9.5.3 Question 3

For distribution A, perform a simulated experiment. Draw a large number of samples, and for each sample fit a linear regression. Store the slope coefficient from each regression in a vector. Finally, compute the standard deviation for the slope coefficients.

Repeat this process for distributions B and C. Do the results match your intuition?

9.6 Understanding Uncertainty

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

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

Definition 9.1. A similar formulation is given in *FOAS* as definition 4.2.3,

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

where $\hat{\sigma}^2 = V[\hat{\epsilon}]$

Explain why each term makes the variance higher or lower:

- $\hat{\sigma}^2$ is the variance of the error $\hat{\epsilon}$
- 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

9.7 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$df.residual
```

```
## [1] 86
```

Can you use the pieces that you're familiar with to produce a p-value using robust standard errors?

```
regression_p_value <- function(model, variable) {
  ## this function takes a model
  ## and computes a test-statistic,
  ## then compares that test-statistic against the
  ## appropriate t-distribution

  ## you can use the following helper functions:
  ##  - coef()
  ##  - vcovHC()
  df <- model$df.residual

  # numerator   <- 'fill this in'
  # denominator <- 'fill this in'

  numerator   <- coef(model)[variable]
  denominator <- sqrt(diag(vcovHC(model)))[variable]

  test_stat_  <- numerator / denominator
  p_val_      <- 'fill this in'
  p_val_      <- pt(test_stat_, df = df, lower.tail = FALSE) * 2

  return(p_val_)
  }
```

If you want to confirm that what you have written is correct, you can compare against the value that you receive from the line below.

```
coeftest(model_one, vcov. = vcovHC(model_one))
```

```
##
## t test of coefficients:
##
##              Estimate Std. Error t value  Pr(>|t|)
## (Intercept) 11.204145  39.450563  0.2840    0.7771
## sqrft        0.140211   0.021111  6.6417 2.673e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
p_value_ <- broom::tidy(coeftest(model_one, vcov. = vcovHC(model_one))) %>%
  filter(term == 'sqrft') %>%
```

```
  select('p.value') %>%
  as.numeric()

test_that(
  'test that hand coded p-value is the same as the pre-rolled',
  expect_equal(
    object   = as.numeric(regression_p_value(model_one, 'sqrft')),
    expected = p_value_
  )
)
```

```
## Test passed
```

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.

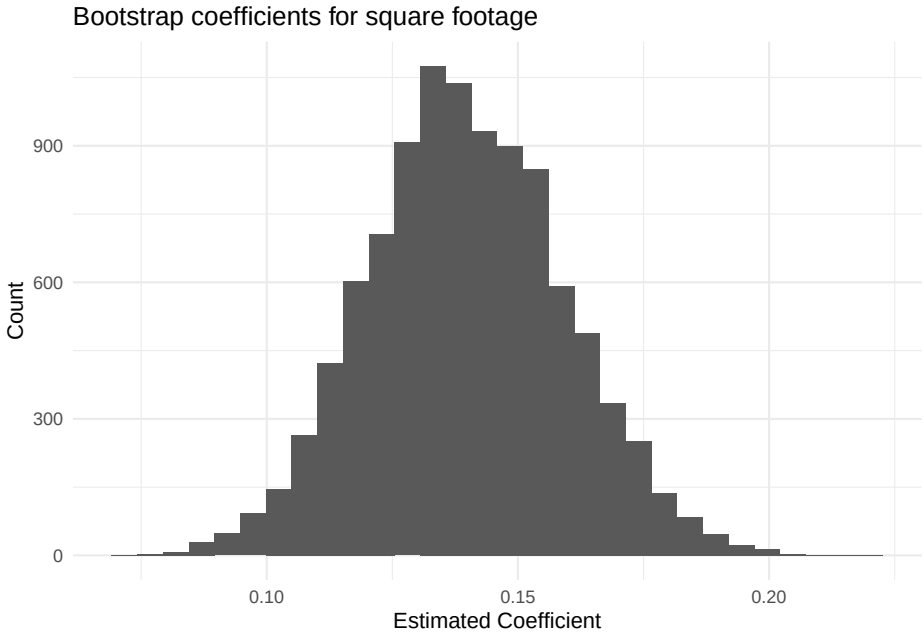
```
bootstrap_sqft <- function(d = data, number_of_bootstraps = 1000) {
  number_of_rows <- nrow(d)

    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, data = bootstrap_data)
      coef_sqft[i]    <- coef(estimated_model)['sqrft']
    }
  return(coef_sqft)
}
```

```
bootstrap_result <- bootstrap_sqft(d = data, number_of_bootstraps = 10000)
```

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?

```
coeftest(model_one, vcov. = vcovHC(model_one))
```

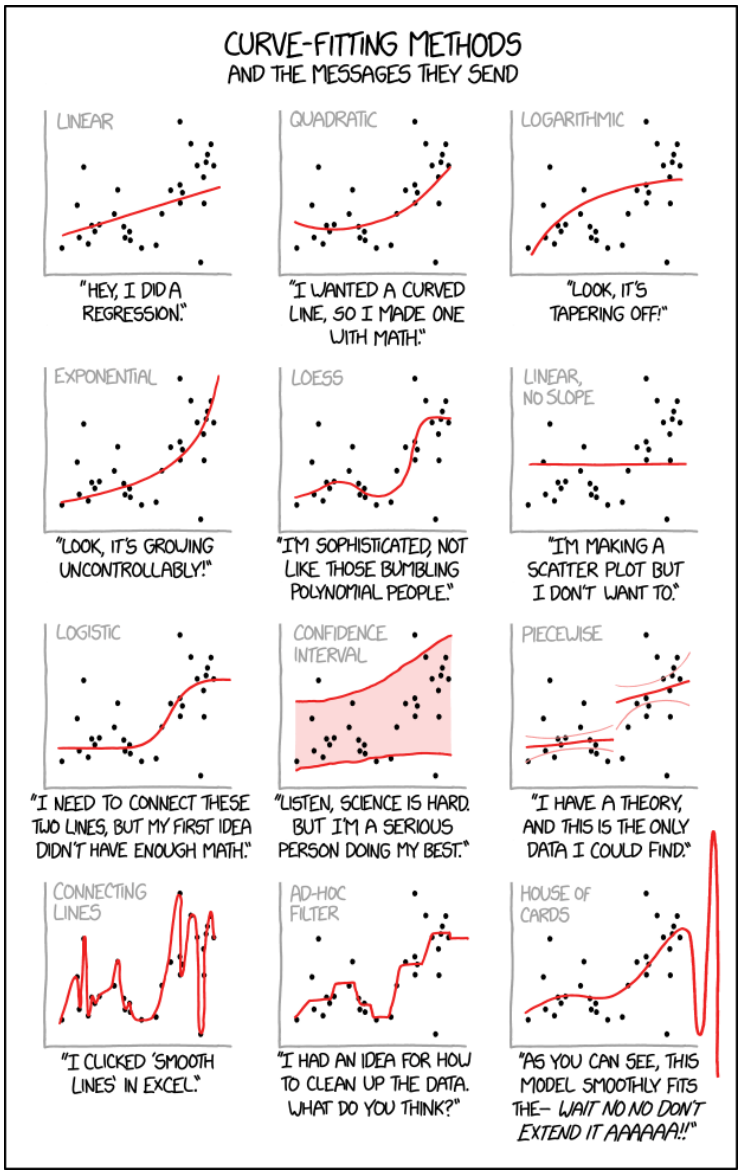
```
##
## t test of coefficients:
##
##              Estimate Std. Error t value  Pr(>|t|)
## (Intercept) 11.204145  39.450563  0.2840    0.7771
## sqrft        0.140211   0.021111  6.6417 2.673e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
sd(bootstrap_result)
```

```
## [1] 0.01947353
```

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

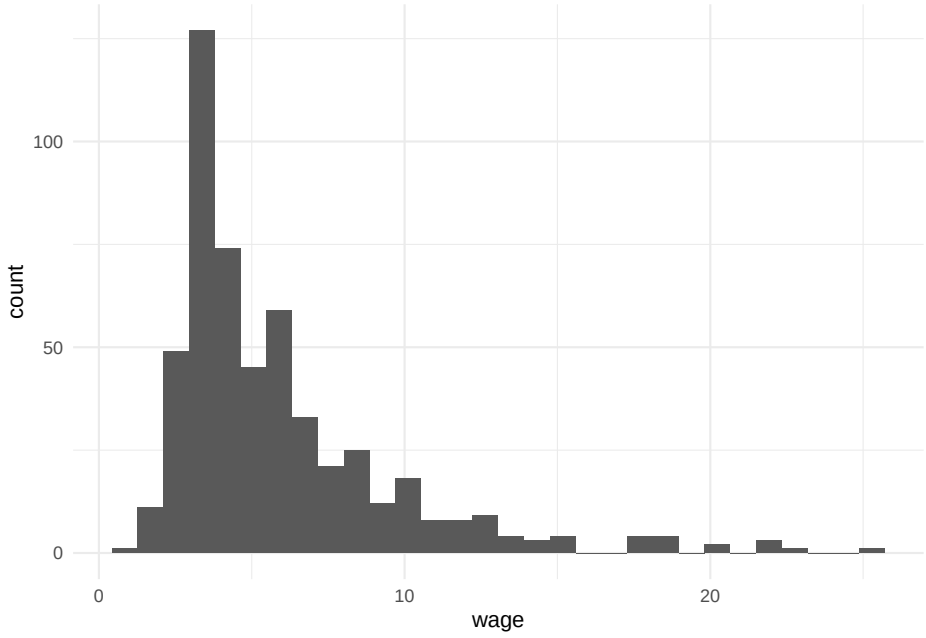
10.5 R Activity: Measuring the return to education

- 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 <- wooldridge::wage1
#names(wage1)

wage1 %>%
  ggplot() +
  aes(x=wage) +
  geom_histogram()
```

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



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

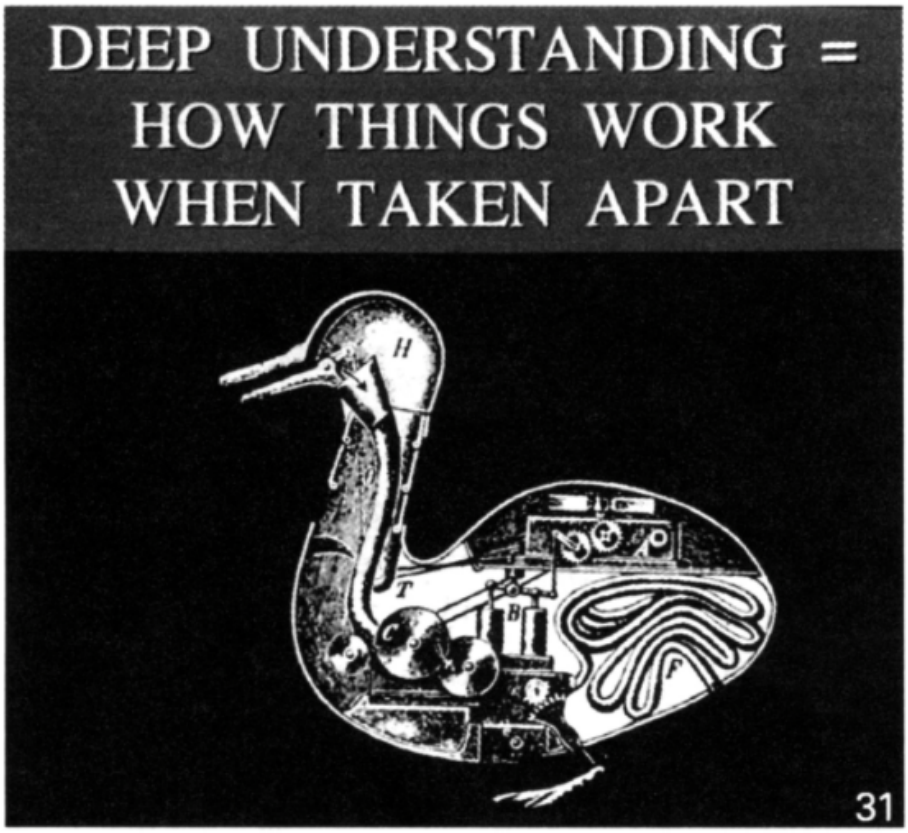


Figure 11.1: duck yeah

What does it mean for **this** to cause **that**? This question has flummoxed the discipline of statistics for a *very* long time; but, more than statistics, it has also flummoxed philosophers for even longer!

Why is something that seems so natural to us in our limited, daily lives, so difficult to formalize? If it is difficult for us to formalize in conversation, how can we hope to formalize this so that a model can *discover* and *evluate* causal effects from data?

Of all the weeks in this class, this is perhaps the most conceptually challening.

11.1 Learning Objectives

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

1. **Remember** that most interesting questions in their data analysis are actually causal questions.
2. **Articulate** a particular causal model that describes the world, and **evaluate** whether a research design and a statistical analysis does an adequate job answering a question about a causal model.
3. **Appreciate** the deep difficulty of causal questions, and how research design guides data collection.

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 14	5%

Deliverable Name	Week Due	Grade Weight
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.

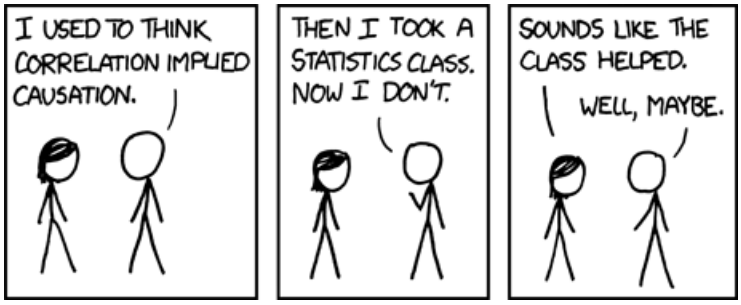
11.4 Discussion

11.4.1 Path Diagrams

Sleep → 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?

11.5 An Interlude



11.5.1 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?

11.6 R Exercise

11.6.1 Omitted Variable Bias in R

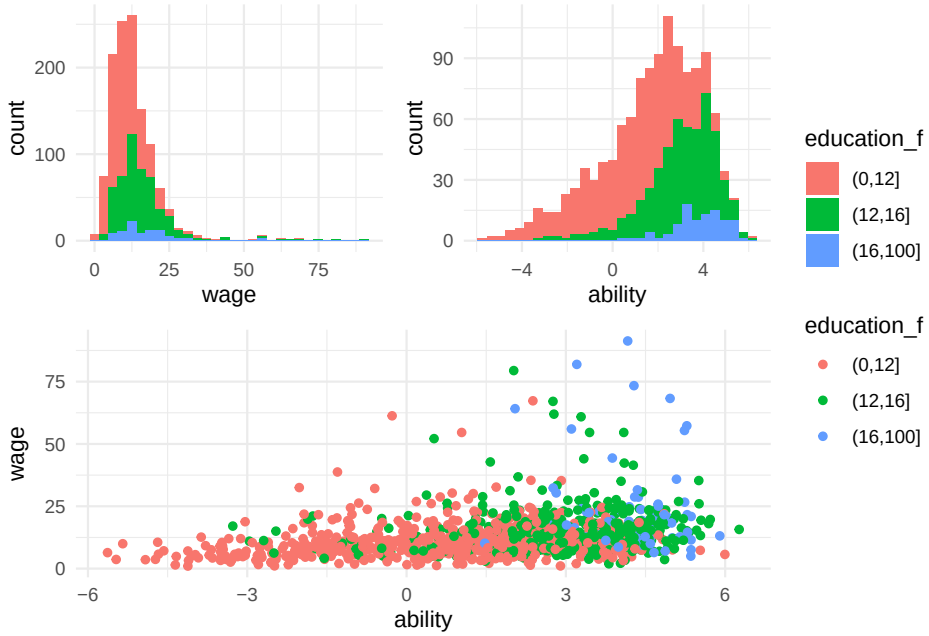
The file `htv.RData` contains data from the 1991 National Longitudinal Survey of Youth, provided by Wooldridge. All people in the sample are males age 26 to 34. The data is interesting here, because it includes education, stored in the variable `educ`, and also a score on an ability test, stored in the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

```
load('./data/htv.RData')

data <- data %>%
  rename(
    ability    = abil,
    education  = educ,
    north_east = ne,
    north_cent = nc,
    potential_experience = exper,
    edu_mother = motheduc,
    edu_father = fatheduc,
    divorce_14 = brkhme14,
    siblings   = sibs,
    tuition_17 = tuit17,
    tuition_18 = tuit18) %>%
  mutate(
    education_f = cut(education, breaks = c(0,12,16,100))) %>%
  select(-c(ctuit, expersq, lwage))

glimpse(data)
```

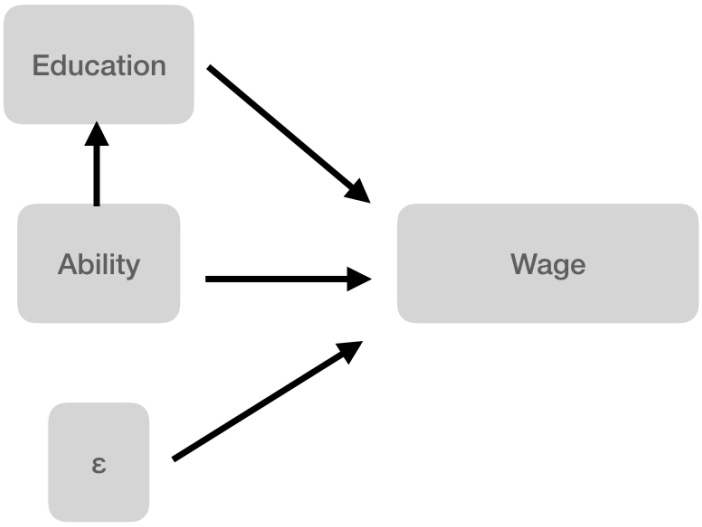
```
## Rows: 1,230
## Columns: 21
## $ wage                <dbl> 12.019231, 8.912656, 15.514334, 13.333333, 11.070~
## $ ability             <dbl> 5.0277381, 2.0371704, 2.4758952, 3.6092398, 2.636~
## $ education           <int> 15, 13, 15, 15, 13, 18, 13, 12, 13, 12, 12, 12, 1~
## $ north_east          <int> 0, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 0, 1, 1, 1, 1~
## $ north_cent          <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ west                <int> 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0~
## $ south               <int> 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ potential_experience <int> 9, 8, 11, 6, 15, 8, 13, 14, 9, 9, 13, 14, 4, 8, 7~
## $ edu_mother          <int> 12, 12, 12, 12, 12, 12, 13, 12, 10, 14, 9, 12, 17~
## $ edu_father          <int> 12, 10, 16, 12, 15, 12, 12, 12, 12, 12, 10, 16, 1~
## $ divorce_14          <int> 0, 1, 0, 0, 1, 0, 0, 1, 1, 0, 1, 0, 0, 0, 0, 0, 0~
## $ siblings            <int> 1, 4, 2, 1, 2, 2, 5, 4, 3, 1, 2, 1, 1, 3, 2, 2, 1~
## $ urban               <int> 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 0, 1, 1, 1, 1, 1~
## $ ne18                <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1~
## $ nc18                <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
```

```
## $ south18              <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ west18               <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ urban18              <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1~
## $ tuition_17           <dbl> 7.582914, 8.595144, 7.311346, 9.499537, 7.311346,~
## $ tuition_18           <dbl> 7.260242, 9.499537, 7.311346, 10.162070, 7.311346~
## $ education_f          <fct> "(12,16]", "(12,16]", "(12,16]", "(12,16]", "(12,~
```

```
wage_plot <- data %>%
  ggplot() +
  aes(x=wage, fill=education_f) +
  geom_histogram(bins=30)
ability_plot <- data %>%
  ggplot() +
  aes(x=ability, fill=education_f) +
  geom_histogram(bins=30)
wage_by_ability_plot <- data %>%
  ggplot() +
  aes(x=ability, y=wage, color=education_f) +
  geom_point()


(wage_plot | ability_plot) /
  wage_by_ability_plot +
  plot_layout(guides = 'collect')
```



Assume that the true model is,



11.6.2 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.
4. Write down the expression for what omitted variable bias would be if you couldn't measure ability.

5. Add this omitted variable bias to the coefficient for education to see what it would be.
6. Now evaluate your previous result by fitting the model,

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

7. Does the coefficient for the relationship between education and wages match what you estimated earlier?
8. Why or why not?
9. Reflect on your results:
10. What does the direction of omitted variable bias suggest about OLS estimates of returns to education?

11. What does this suggest about the reported statistical significance of education?

11.7 Research Design Strategies

Hopefully you feel like, "Golly. It would be really, *really* hard to assert some causal model and *know that it is actually true*." How does this lead you to think about the role of research design in setting up your data collection?

1. If you could **do the experiment** to determine the effect of education on wages, how would you do it?
2. If you cannot **do the experiment** to determine the effect of education on wages, what are some options for where to look for data? What would you hope these areas provide to you?

11.8 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*
$\widehat{lifespan} = 87.4 - 1.2\ cigarettes$	*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

```
# install.packages("corrgram")
```

12.1 Learning Objectives

At the end of this week's learning students will be able to

1. **Describe** the assumptions of the classical linear model (sometimes referred to as the Gauss-Markov Assumptions) and what each assumption contributes to the estimator.
2. **Evaluate** using empirical methods, whether each of the assumptions are likely to be true of the population data generating function.
3. **Assess** whether the guarantees that are provided by the classical linear model's requirements are likely to *ever* be true, including within data the student is likely to encounter.

12.2 Class Announcements

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

Comparing the Large Sample Model and the CLM

12.4.1 Part 1

- 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?

12.4.2 Part 2

- Suppose that you're interested in understanding how subsidized school meals benefit under-resourced students in San Francisco East Bay region.
  - Using the tools from DATASCI 201, refine this question to a data science question.
  - Suppose that there exists two possible data sources to answer the question you have formed:
    - A large amount (e.g. 10,000 data points) of individual-level data about income, nutrition and test scores, self-reported by

individual families who have opted in to the study.

    - A relatively smaller amount (e.g. 500 data points) of 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 using one or the other data source?**

12.4.3 Part 3

- Suppose you elect to use the relatively larger sample of individual-level data.
  - Which of the large-sample assumptions do you expect are valid, and which are problematic?
- Or, suppose that you elect to use the relatively smaller sample of school-district-level data.
  - Which of the CLM assumptions do you expect are valid, and which do you expect are most problematic?
- **What was the research question that you identified?**
- **What would a successful answer accomplish?**

12.4.4 Part 4

- **Which data source, the individual or the district-level, do you think is more likely to produce a successful answer?**

12.4.5 Part 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.

12.5 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)`

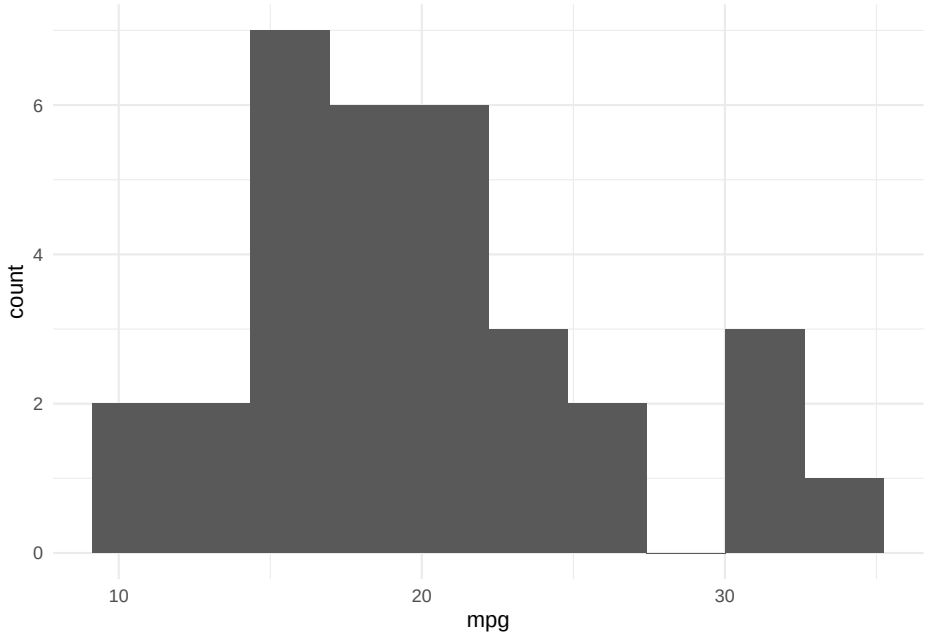
```
data(mtcars)
glimpse(mtcars)
```

```
## 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,~
```

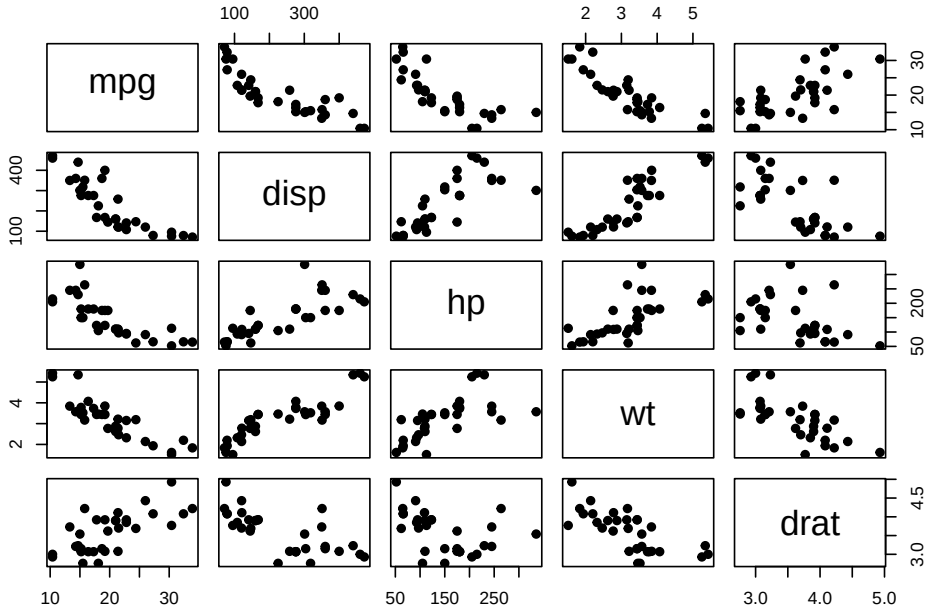
12.5.1 Questions:

0. Using the `mtcars` data, quickly reason about the variables that we're interested in studying:

```
mtcars %>%
  ggplot() +
  aes(x=mpg) +
  geom_histogram(bins=10)
```




```
mtcars %>%
  select(mpg, disp, hp, wt, drat) %>%
  pairs(pch=19)
```



1. Using the `mtcars` data, run a linear regression to find the relationship between miles per gallon (`mpg`) on the left-hand-side as a function of

displacement (`disp`), gross horsepower (`hp`), weight (`wt`), and rear axle ratio (`drat`) on the right-hand-side. That is, fit a regression of the following form:

$$\widehat{mpg} = \hat{\beta}_0 + \hat{\beta}_1 disp + \hat{\beta}_2 horse\_power + \hat{\beta}_3 weight + \hat{\beta}_4 drive\_ratio$$

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

```
# goal:
# consequence if violated:
```

```
# goal:
# consequence if violated:
```

```
# goal:
# consequence if violated:
```

```
# goal:
# consequence if violated:
```

```
# goal:
# consequence if violated:
```

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] 2 2 2 1 2 1 2 2 2 2
```

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
```

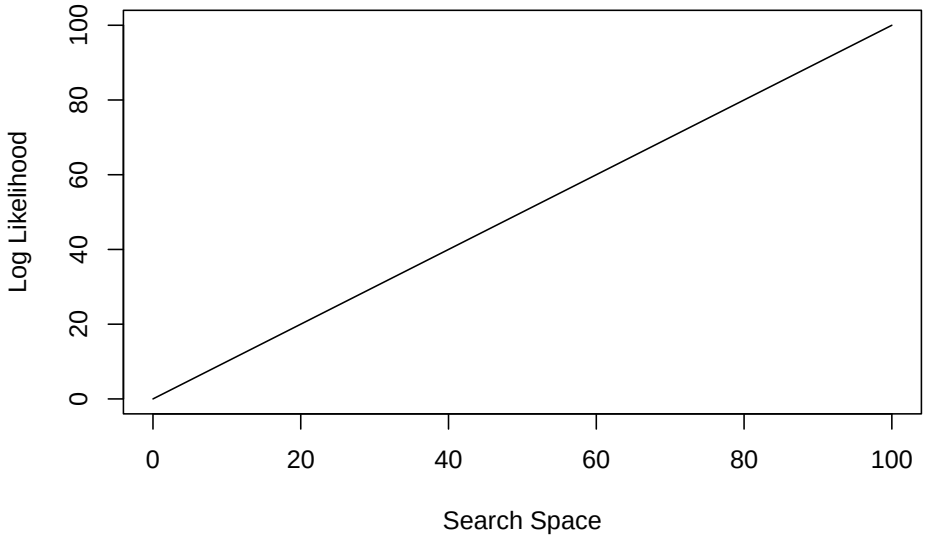
14.8.2 MLE Estimation

- 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?

Part II

The data resembles draws from a Bernoulli distribution.

Figure 14.2: bbc, office space

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.

Appendix

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.