Statistics for Data Science

D. Alex Hughes, Paul Laskowski, The 203 Teaching Team

2025-01-16

Table of 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.

Part I

Probabilty Theory

Probability theory is the basis for all modeling in data science. In the first part of the course, we will cover the basics.

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.

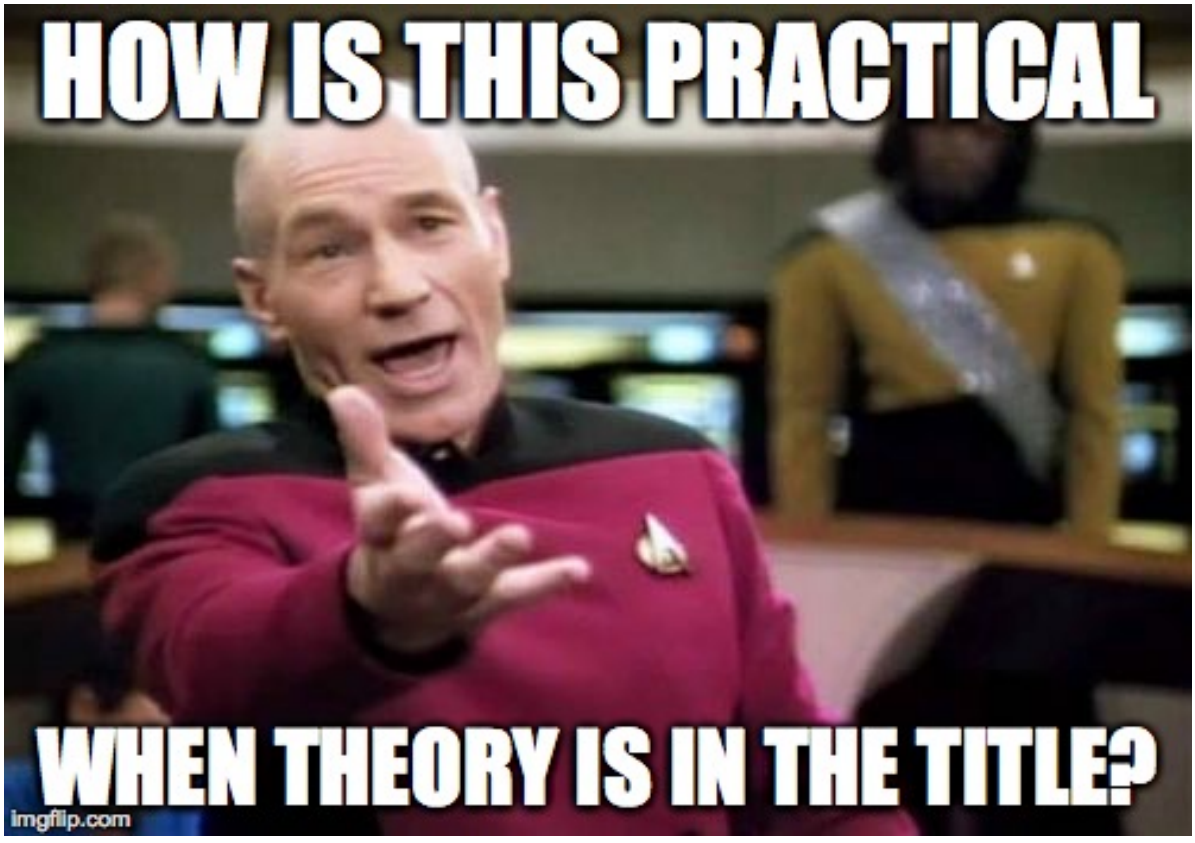


Figure 1.2: told you this would be spacey

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

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

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

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).

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

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

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

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

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.

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.

2 Defining Random Variables

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
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 ------------------------------------------ tidyverse_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 becor

Attaching package: 'plotly'


The following object is masked from 'package:ggplot2':

    last_plot


The following object is masked from 'package:stats':

    filter


The following object is masked from 'package:graphics':

    layout
```

Figure 2.1: yosemite valley

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 objects thare are a foundation 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.

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. Because we don't have access to the joint density function, in later weeks we will try to produce estimates using data. The simpler the estimate, the less data we need; the fuller the representation of the joint density function we desire, the more data we need.

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 Thursday at 2:00p. 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?

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.

> **Note**
>
> You might notice that the $X(\omega) \leq r$ feels kind of weird; why isn't it just $X(\omega) = r$? After all, this is a mapping from an outcome, $\omega \in \Omega$ to a real number, right? So, why not just be direct about it? The answer is a real deep-dive, and one that is better suited to a formal measure theory course.
> The short answer, is that the we use this $\leq r$ range because we're using a Boreal σ algebra to constrain the sets that have probability measures. Why this constraint? If we don't weird stuff can happen. Like, real weird things: you can split a sphere into pieces and create two new spheres of equal volume from the pieces (Banch-Tarski Paradox).

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 . Can you do *math* with ?
  - 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? Can you do math on words or concepts?
  - Suppose you want to direct class support to students in 203, but their grades are scored `[A, A-, ..., C+]` and features include prior statistics classes grades, also scored `A, A-, ..., C+]`.

2.5 Pieces of a Random Variable

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

There are two key pieces that must exist for every random variable. What are these pieces? The new piece is provided to us in **Definition 1.2.1** *Random Variable* (on page 16). The older piece (from last week) that is now useful is a part of the Kolmogorov Axiom, the probabilty triple: (Ω, S, P) where Ω is a sample space, S is an event space, and P is a probability-measure.

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 sensible 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.

Let $g : U \to \mathbb{R}$ be some function, where $X(\Omega) \subseteq U \subseteq \mathbb{R}$. 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?

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*)?

You should try driving in Berkeley some time. It is a **trip**! Without being deliberately ageist, the city is full of ageing hippies driving beater 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 55 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:

For a continuous random variable X with CDF F, the *probability density function* of X is

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

The implies, further, that for a random variable X with PDF f, the *cumulative density function* of X is:

$$F(x) = \int f(x)dx, \forall x \in \mathbb{R}.$$

- 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?

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:
    1. Using the PDF
    2. 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

WebGL is not supported by your browser - visit
https://get.webgl.org for more info

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*:

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

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.

[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.

- 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.

```r
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.

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

```
 [1] 1.614028 1.455171 2.095655 2.082238 2.506946 2.957886 3.452067 1.765094
 [9] 4.164895 1.448327 1.769607 1.659020 2.894359 3.193325 2.511176 2.090183
[17] 1.112637 2.649366 3.657755 3.560751
```

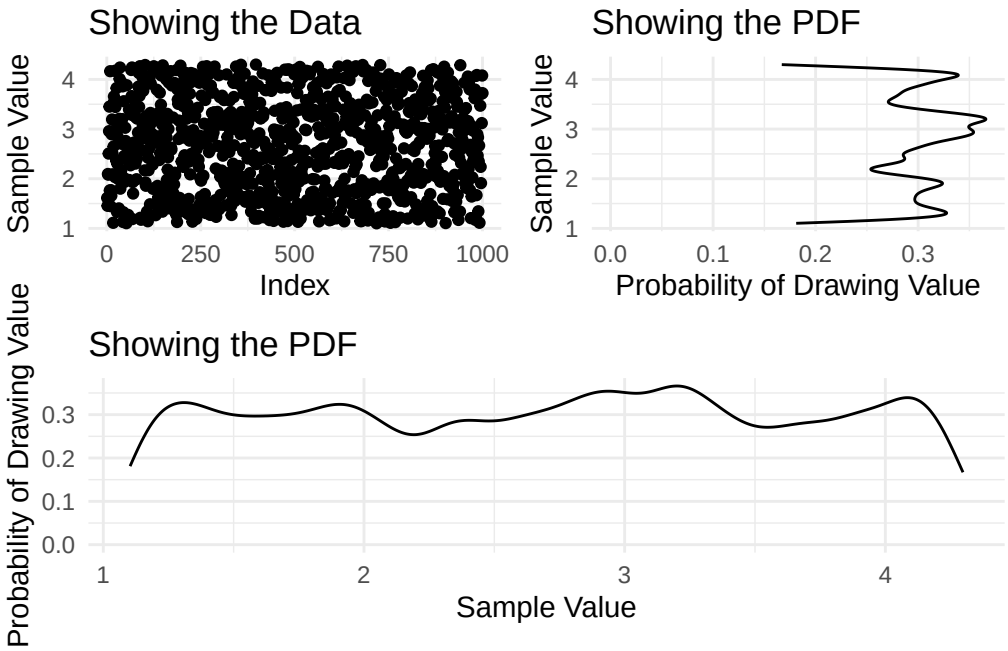
(Notice that R is a 1 index language (python is a zero-index language).)

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

```r
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
```



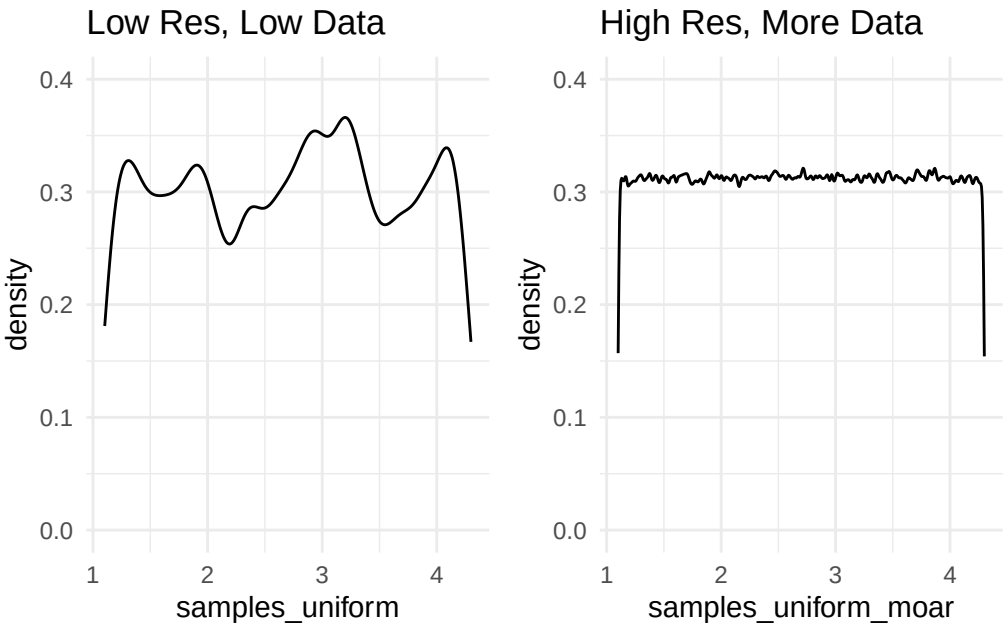
Interesting. From what we can see here, there does not appear to be any discernible pattern. This leaves us with two options: either, we might reduce the resolution that we're using to view this pattern, or we might take more samples and hold the resolution constant. Below, two different plots show these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

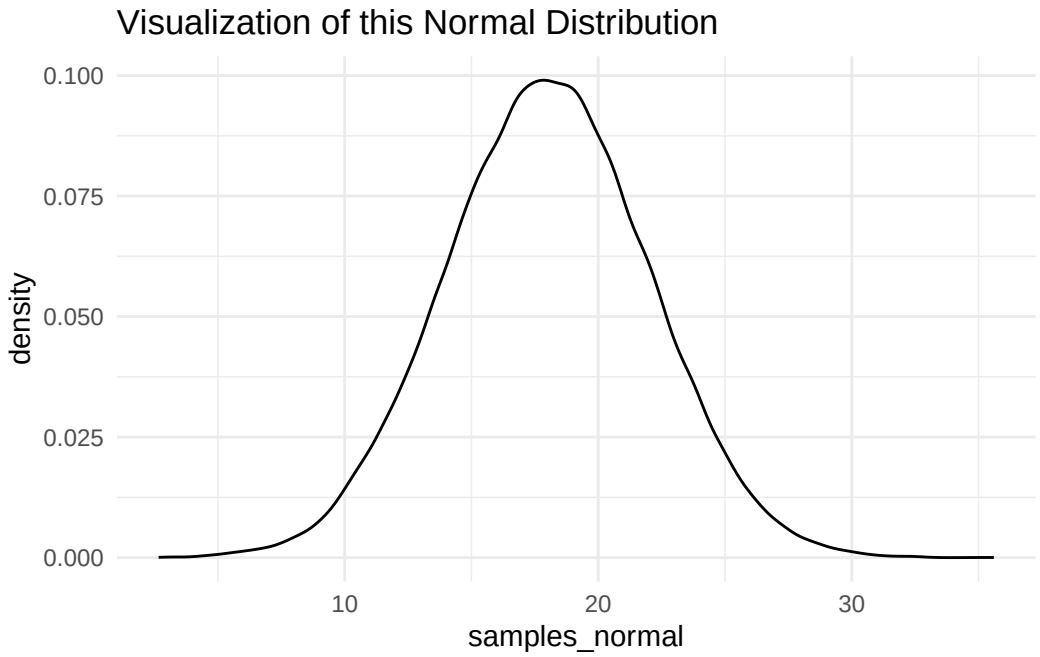
Like before, we could look at the first 20 of these samples.

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

```
 [1] 20.01714 12.06071 18.98264 22.95449 14.66461 16.66013 21.63034 17.68357
 [9] 20.81829 16.47933 20.43985 17.37152 13.74835 18.17729 19.21103 23.63705
[17] 26.01543 15.64904 18.20650 19.18236
```

And, from here we could visualize this distribution.

```r
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,

$$A \sim Uniform(min = 1.1, max = 4.3)$$
$$B \sim Normal(mean = 18, sd = 4)$$
$$C = A + B$$

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?

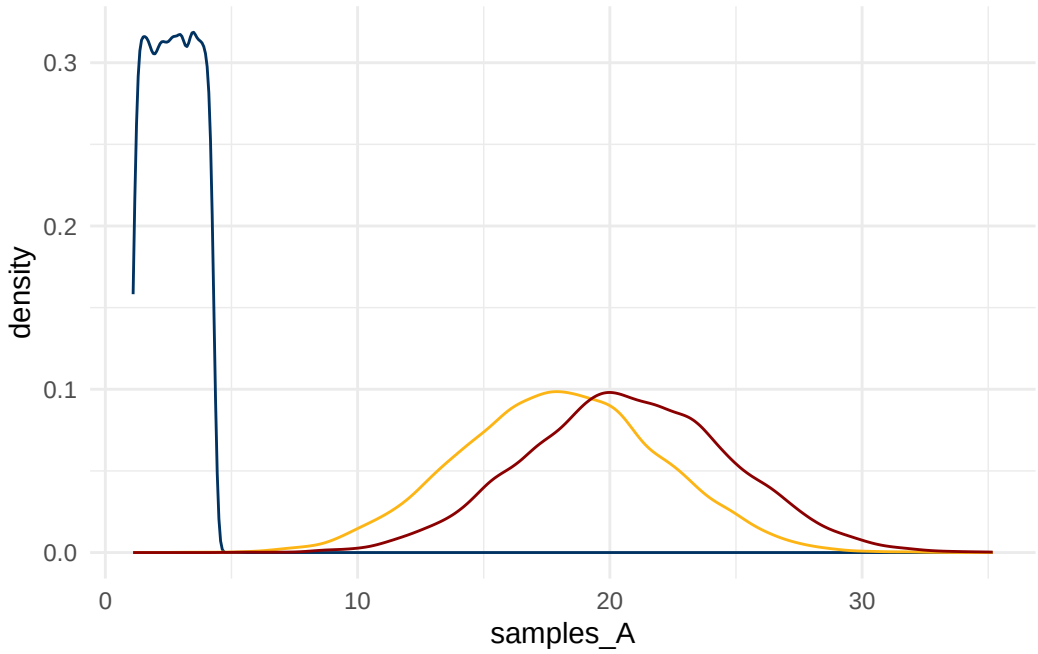
```
samples_A <- runif(n = 10000, min = 1.1, max = 4.3)
samples_B <- rnorm(n = 10000, mean = 18, sd = 4)

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

3 Summarizing Distributions

Figure 3.1: a majestic valley

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:

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

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:

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