# Inference

Until now, we were talking about the mechanics of describing and summarizing distributions in real life datasets.

The reason we do this is that we want the real life data to teach us something about the larger world. (But not always.)

But, real life datasets are often just snapshots of that larger world.

The process of learning about the larger world using datasets is known as **inference**. There are two main flavors:

1. **Statistical inference**: how can we learn from relationships in datasets when we have some uncertainty about the link between the data and the real world?

# Inference

Until now, we were talking about the mechanics of describing and summarizing distributions in real life datasets.

The reason we do this is that we want the real life data to teach us something about the larger world. (But not always.)

But, real life datasets are often just snapshots of that larger world.

The process of learning about the larger world using datasets is known as **inference**. There are two main flavors:

2. **Causal inference**: how can we learn whether relationships in datasets specifically teach us about *causal* relationships in the real world?

# Where Does Data Come From?

We've already talked a little about the origins of data.

There's a more abstract question we now need to tackle:

↪ What processes in the world create the conditions under which the data we see comes into being?

For example, when a researcher to runs an experiment, what processes make it possible to observe interpretable patterns in the resulting dataset?

# Data Generating Processes

Using Gailmard's definition, a **data generating process** (or **DGP**) "is a rule or set of rules governing the social or political events that an analyst wishes to study and the rules by which observations of its results come to be represented in a dataset."

Think back to Prof. Hubert's dataset: all civil rights cases filed in seven courts from 1995 to 2016.

He was not specifically and solely interested in these seven courts, these 20 years, or the 200 judges in my dataset.

He collected this data because he wanted to study the relationship between partisan appointment of judges and civil rights outcomes.

*He wanted to learn about the larger data generating process.*

# Data Generating Processes

When one uses a dataset to evaluate theories, there are two "metaphysical commitments" one makes.

1. The dataset isn't all there is: there's a DGP that generated the specific dataset.
2. Data arises from a DGP in one of two ways:
   2.1 A DGP is **deterministic** if the data it generates involves no random chance.

   2.2 A DGP is **stochastic** if the data it generates involves some degree of random chance.

These are starting presumptions; there's no way to empirically evaluate whether they are "true."

# Deterministic DGPs

For a deterministic DGP, a researcher typically tries to figure out the **necessary and sufficient conditions** for some event.

| Republican | Pro-Plaintiff |
|---|---|
| 1 | 0 |
| 1 | 0 |
| 1 | 1 |
| 0 | 1 |
| 0 | 1 |

↪ Having a Republican judge is necessary for a pro-defendant outcome. (But it is not sufficient!)

↪ Having a Democratic judge is sufficient for a pro-plaintiff outcome. (But it is not necessary!)

# Deterministic DGPs

For a deterministic DGP, a researcher typically tries to figure out the **necessary and sufficient conditions** for some event.

| Republican | Pro-Plaintiff |
|---|---|
| 1 | 0 |
| 1 | 0 |
| 1 | 0 |
| 0 | 1 |
| 0 | 1 |

↪ Having a Republican judge is necessary and sufficient for a pro-defendant outcome.

↪ And equivalently: having a Democratic judge is necessary and sufficient for a pro-plaintiff outcome. (Why is this equivalent?)

# Stochastic DGPs

In last example, why would we assume a deterministic DGP?

↪ Do we really think that there's no random chance in the way judges make decisions on cases??

There are three possible sources of uncertainty that would lead someone to assume a stochastic DGP:

1. **Sampling uncertainty**: the data is a **probability sample** of the larger population of interest (controlled by researcher)
2. **Theoretical uncertainty**: the positive theory being examined has missing elements, which could be important for the DGP
3. **Fundamental uncertainty**: exactly what it sounds like...

# Stochastic DGPs

A few important notes on this:

↪ Sampling uncertainty is controlled by the researcher, so it's pretty easy to deal with in our statistical analysis.

↪ Most(?) of the time, we're dealing with theoretical or fundamental uncertainty.

↪ Theoretical uncertainty is "good": it's a byproduct of the fact that <u>all</u> theories are simplifications of the world.

↪ It's hard to know exactly how (and whether) theoretical and fundamental uncertainty manifest in real life data.

# Stochastic DGPs

Core dilemma: you typically only observe one realization of a DGP (i.e., one dataset), but to do statistical analysis, we need to have some idea how the uncertainty works.

This is easy when a researcher controls how the uncertainty arises in her data, such as with survey research.

↪ In this case, we say the dataset arises by way of **design-based sampling**, in which sampling uncertainty is the issue.

But, when a researcher does not control how the uncertainty arises in her data, then she has to make assumptions about (i.e., "model") the uncertainty. This is much harder.

↪ The dataset arises by way of **model-based sampling**.

# Stochastic DGPs

In the remainder of this course, we'll assume stochastic DGPs.

Most researchers are comfortable assuming that real life social processes involve some random chance.

But so far, we've been pretty abstract about what exactly a DGP is.

We'll conceptualize them as **probability distributions**.

$\hookrightarrow$ Review the basics of probability from math camp.

# Mathematical Representations of DGPs

Recall: DGPs are the way we talk about the larger social and political processes that generate real life data.

We always like to try to simplify complex realities into models. And we will model DGPs using probability distributions.

A **probability distribution** describes a *hypothetical* (or "theoretical") distribution that arises from a random process.

This is distinct from an **empirical distribution** (or **frequency distribution**) which you've already seen.

We think of empirical distributions as being real-life manifestations of probability distributions (which we cannot directly observe).

# Mathematical Representations of DGPs

Quick aside: the following material in topics 4A and 4B will be relatively more mathematical than in topics 1–3

↪ This material is more foundational in nature – a lot of it is not directly used in applied research.

↪ However, it is important to understand the foundations of probability and statistics to be able to understand the applied research.

↪ Still, if some of the mathematical details are confusing, don't worry too much.

↪ I will try to highlight the key takeaways from the math.

# Probability Spaces/Models

We're all intuitively familiar with the idea of randomness.

But we need to attach some vocabulary, math and notation to our intuitions so we can be more precise.

We'll do this using a simple example: rolling a 6-sided die.

When you roll a die, one of six numbers will be facing up.

↪ We refer to the specific roll as an **experiment***.

↪ We refer to the possible outcomes as the **sample space**.

↪ We usually use $S$ to label a sample space. In this example, the sample space can be written: $S = \{1, 2, 3, 4, 5, 6\}$.

# Probability Spaces/Models

We use the term **probability** to describe the way that the randomness operates in a particular situation.

This can get a little mathematically complicated, but for our purposes, we simply need to know the probability with which each element of $S$ can be chosen in an experiment.

For example: a roll of a (fair) die will yield each of the possible outcomes with equal probability, probability $\frac{1}{6}$.

I will loosely refer to a probabilistic situation (like a roll of a dice) as a **probability space (or model)**.

$\hookrightarrow$ The key thing for a probability space/model: we need to know the sample space and the probabilities.

# Probability Spaces/Models

Here: we are going to jump right into *probability distributions*, which are the link between basic probability and statistics.

# Random Variables

When we do statistics, we will typically want to "translate" a sample space $S$ into numbers.

A **random variable** (or RV) is a function that translates outcomes of an experiment (elements of $S$) into real numbers.

$\hookrightarrow$ By convention, we use uppercase letters (e.g., $X$ or $Y$) to label RVs. So, $X : S \to \mathbb{R}$ is the "mathy" way to write an RV called $X$ for a probability model with sample space $S$.

$\hookrightarrow$ RVs will only return as many numbers as elements in $S$. We label that set of numbers with curly capital letters, like $\mathcal{X}$.

$\hookrightarrow$ The specific value of an RV in an experiment is its **realization**, and we use lower case letters to label realizations, like $x \in \mathcal{X}$.

# Random Variables

A random variable representing the probability space for one roll of a fair six-sided die will only return one of six numbers.

So, if we label that random variable $X$ and the sample space $S$, then: $S = \{1, 2, 3, 4, 5, 6\}$ and $\mathcal{X} = \{1, 2, 3, 4, 5, 6\}$.

But wait! $S$ is already comprised of numbers—why translate?!

$\hookrightarrow$ A bit philosophical: $S$ isn't *really* comprised of numbers.

This is an example of a **discrete random variable** $X$ in which the set $\mathcal{X}$ is countable and finite (i.e., "discrete").

# Random Variables

There are some cases where the sample space $S$ is may not be comprised of numbers.

Example: 5-scale of left-right political ideology.

Here, the sample space is defined as

$$S = \{\text{Strong R}, \text{Weak R}, \text{Moderate}, \text{Weak L}, \text{Strong L}\}$$

We can translate this into a random variable that consists of numbers: $X : S \rightarrow \mathcal{X}$ where $\mathcal{X} = \{1, 2, 3, 4, 5\}$.

# Random Variables

Consider a different random process: spraying water into an empty container many feet away.

$\hookrightarrow$ Assume you spray 1 liter of water toward a 1 liter container.

There is some randomness since spraying water is imprecise. Some of it will not land in the container.

Measuring the amount you get in the container in milliliters, then we have an experiment from a probability space with $S = [0, 1000]$.

If we call the random variable $Y$, then $\mathcal{Y} = [0, 1000]$.

This is an example of a **continuous random variable** $X$ in which the set $\mathcal{Y}$ is non-countable and infinite ("a continuum").

# Probability Distributions

So far, we haven't talked about the randomness of an RV.

For example, we may want to know the probability that a particular value of the RV will be realized.

↪ Using math notation, we usually write the probability that an RV labeled $X$ will take a specific value $x$ as $\Pr[X = x]$ or $\Pr[x]$.

↪ Some people use round brackets. Such is life.

For any RV, there are two important functions that tell us about how the randomness works.

These functions describe the *probability distribution* for that RV.

# Discrete Probability Distributions

A **probability mass function** (or **PMF**) of a discrete random variable $X$ specifies the probability that $X = x$ for all $x$.

We usually call this function $f$, so in math: $f : \mathcal{X} \to [0, 1]$.

$\hookrightarrow$ So, you plug in any element $x$ from $\mathcal{X}$, and the PMF will return a probability for that element: $f(x)$.

But, the function has to return "proper" probabilities, meaning:

1. $0 \leq f(x) \leq 1$ for all $x \in \mathcal{X}$.
2. $\sum_{x \in \mathcal{X}} f(x) = 1$.
3. For any subset of possible outcomes, the probability that any one of those outcomes arises is the sum of the probabilities of each outcome.

# Discrete Probability Distributions

For example, consider our example: one roll of a fair die.

Then, we can write the PMF of this random variable as:

$$f(x) = 1/6 \quad \text{for all } x \in \mathcal{X}$$

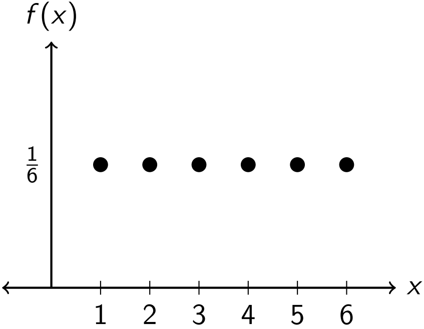
We can verify that this function gives us proper probabilities:

1. For each outcome $x$, $f(x) = 1/6$ which is between 0 and 1.
2. $f(1) + f(2) + f(3) + f(4) + f(5) + f(6) = \frac{1+1+1+1+1+1}{6} = 1$.

Practical note: property 3 will always be satisfied if properties 1 and 2 are satisfied and $X$ is a random variable.

# Discrete Probability Distributions

We can plot this PMF:



# Discrete Probability Distributions

A **cumulative distribution function** (or **CDF**) of a discrete random variable specifies the probability that $X \leq x$.

We usually call this function $F$. So, in math $F : \mathcal{X} \to [0, 1]$ where:

$$F(x) = \Pr[X \leq x] = \sum_{a \leq x} f(a)$$

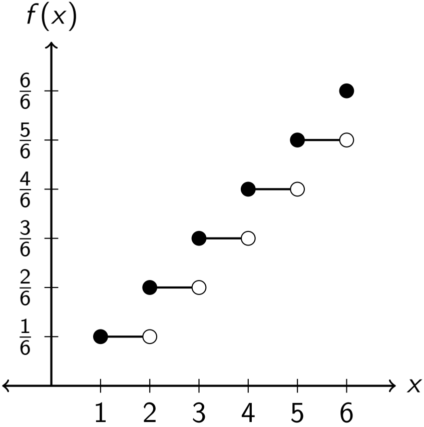
For example, the CDF of our die roll random variable is $F(x) = \frac{x}{6}$.

Suppose we want to know the probability of rolling a 4 or below.

We can calculate it as:
$F(4) = \Pr[X \leq 4] = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{4}{6} = \frac{2}{3}$.

# Discrete Probability Distributions

We can plot this CDF:



# Summary so far

**Random variables:** functions that translate outcomes of an experiment (elements of $S$) into real numbers.

Basically a more "theoretical" way to think about variables that we observe in real life.

$\hookrightarrow$ How do we know how likely each outcome $x$ of RV $X$ is?

$\hookrightarrow$ For discrete RVs, we have probability mass functions (PMFs) and cumulative distribution functions (CDFs).

# Summary so far

**Random variables:** functions that translate outcomes of an experiment (elements of $S$) into real numbers.

Basically a more "theoretical" way to think about variables that we observe in real life.

$\hookrightarrow$ How do we know how likely each outcome $x$ of RV $X$ is?

$\hookrightarrow$ For discrete RVs, we have probability mass functions (PMFs) and cumulative distribution functions (CDFs).

$\hookrightarrow$ Probability mass functions: a function $f(x)$ that tells us the probability $Pr(X = x)$.

$\hookrightarrow$ Cumulative distribution functions: a function $F(x)$ that tells us the probability $Pr(X \leq x)$.

# Continuous Probability Distributions

Things get mathier when we turn to continuous random variables.

The basic “problem” (although not really) is that the set $\mathcal{X}$ has an infinite number of possible outcomes.

**Example:** Amount of ice cream in one scoop – can be 1 oz, 1.1 oz, 1.11 oz etc. – infinitely many values.

Two thorny mathematical questions arise from this:

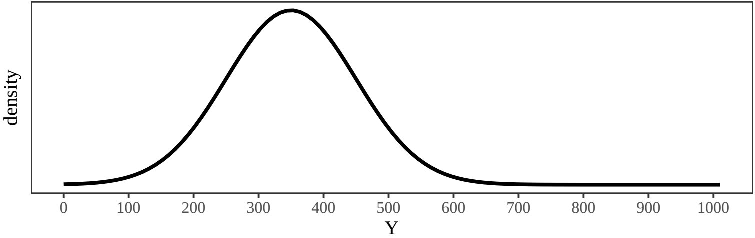
1. For the PMF: how do we come up with “proper” probabilities for each possible $x \in \mathcal{X}$ if there are infinitely many of them?
2. For the CDF: how can we possibly sum together an uncountable set of objects?

These are the kinds of issues that calculus is meant to deal with.

# Continuous Probability Distributions

The **probability density function** (or **PDF**) of a continuous random variable $X$ specifies the density for all $x \in \mathcal{X}$.

Roughly: a PDF gives us the "shape" of a probability distribution.



We don't need to worry too much about the technical details of density, except note: $\Pr[X = x] = 0$ for all $x$ in a continuous RV.

# Continuous Probability Distributions

The **support** of a random variable is all the $x \in \mathcal{X}$ where $f(x) > 0$.

$\hookrightarrow$ For a discrete RV, the support of that RV is all the values of $x$ that arise with strictly positive probability.

$\hookrightarrow$ For a continuous RV, the support of that RV is all the values of $x$ that have strictly positive density.

# Continuous Probability Distributions

The **cumulative distribution function** of a continuous RV also specifies the probability that $X \leq x$ for all $x$.

Because we have a continuum, we can't sum. We integrate:

$$F(x) = \Pr[X \leq x] = \int_{a \leq x} f(a)\, da$$

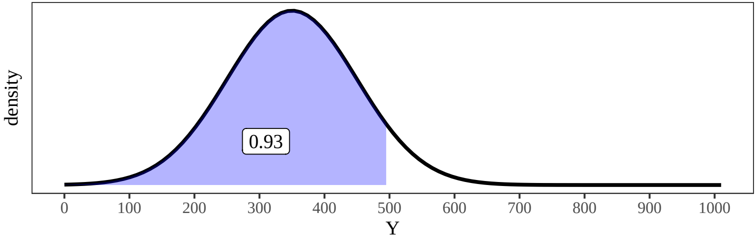
There's a nice geometric way to think about this:

$\hookrightarrow$ Recall from calculus that integrating a function gives the area under that function's curve.

$\hookrightarrow$ Here, we're integrating the PDF, so the CDF gives us the area under the PDF.

# Continuous Probability Distributions

For example, what's the probability the container will be less than half full after we spray water toward it?



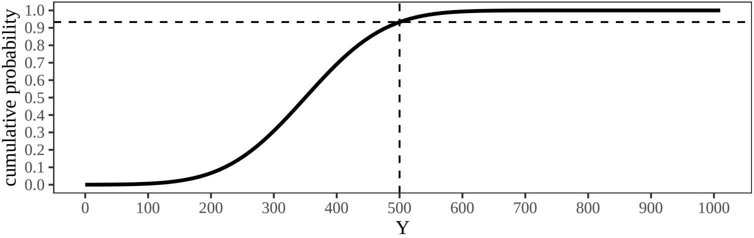
How did I calculate $F(500) = \Pr[Y \leq 500] = 0.93$?

You *could* calculate by hand the definite integral of this PDF from 0 to 500, but I just asked R to do it for me. More on this later.

# Continuous Probability Distributions

The last plot was of the PDF, and I was just showing you the area under the curve corresponding to a particular cutoff (500).

We can also directly plot a continuous CDF showing what the area under the curve would be for each possible cutoff:



# Continuous Probability Distributions

Some important facts about all CDFs:

- ↪ A CDF ends at one. A CDF for a continuous RV starts at zero. Why?
- ↪ If you want the probability of some number *higher than* $x$, you can just calculate $\Pr[X > x] = 1 - F(x)$.
- ↪ If you want the probability of some number *between two numbers* $x_1$ and $x_2$, you can just calculate

$$\Pr[x_1 \leq X \leq x_2] = F(X \leq x_2) - F(X \leq x_1)$$

When you are calculating things with $<$ or $\leq$, be careful to think about whether your RV is discrete or continuous.

Generally, with continuous RVs, we always work with some form of interval – we never calculate $Pr[X = x_1]$ (which is just 0)

# Continuous Probability Distributions

A CDF will be a continuous function if the RV is continuous. (Almost by definition, a discrete RV has a CDF that's not continuous.)

If we have a continuous RV and we assume its CDF is "differentiable everywhere," then:

$\hookrightarrow$ You can get the PDF by differentiating the CDF.

$\hookrightarrow$ You can get the CDF by calculating the definite integral of the PDF from $-\infty$ to $x$.

# Multivariate Probability Distributions

So far, we've talked only about probability distributions for one random variable.

As we discussed in Topic 3, most social scientists are interested in relationships between variables.

Just like we could empirically examine multivariate empirical distributions, we can model their DGPs with multivariate probability distributions.

As we did in Topic 3, we'll build the main ideas for bivariate probability distributions. But, they can extend to distributions with more variables.

# Multivariate Probability Distributions

To model the idea that there's a relationship between two variables, we use the idea of joint distributions.

A **joint distribution of $X$ and $Y$** specifies the probabilistic behavior of each variable and the relationship between them.

- ↪ Again, we typically use $f$ to indicate a PMF/PDF and $F$ to indicate a CDF of a joint distribution.
- ↪ These functions will now have multiple arguments (specifically, two for a bivariate distribution): $f(x, y)$ and $F(x, y)$.
- ↪ They still do the same thing: $f$ takes a value $x$ and a value $y$ and returns a probability/density, and $F$ takes a value $x$ and a value $y$ and returns a cumulative probability.

# Multivariate Probability Distributions

A joint PMF/PDF and a joint CDF each return values for combinations of $X$ and $Y$ realizations.

For example, $f(1, 7)$ gives the probability/density of a joint distribution when the $X$ RV is 1 and the $Y$ RV is 7.

We will build ideas with an example of a simple discrete RV (next slide), but all of these concepts apply to continuous RVs too.

# Multivariate Probability Distributions

Consider an example in which someone flips a fair coin twice, and calculates these quantities:

$$X = \begin{cases} 0 & \text{if } TT \\ 1 & \text{if } HT \text{ or } TH \\ 2 & \text{if } HH \end{cases} \qquad Y = \begin{cases} 0 & \text{if } HH \text{ or } TT \\ 1 & \text{if } HT \text{ or } TH \end{cases}$$

Do you agree these are discrete random variables?

| | $Y = 0$ | $Y = 1$ |
|---|---|---|
| $X = 0$ | | |
| $X = 1$ | | |
| $X = 2$ | | |

# Multivariate Probability Distributions

Consider an example in which someone flips a fair coin twice, and calculates these quantities:

$$X = \begin{cases} 0 & \text{if } TT \\ 1 & \text{if } HT \text{ or } TH \\ 2 & \text{if } HH \end{cases} \qquad Y = \begin{cases} 0 & \text{if } HH \text{ or } TT \\ 1 & \text{if } HT \text{ or } TH \end{cases}$$

Do you agree these are discrete random variables?

| | $Y = 0$ | $Y = 1$ |
|---|---|---|
| $X = 0$ | | 0 |
| $X = 1$ | 0 | |
| $X = 2$ | | 0 |

# Multivariate Probability Distributions

Consider an example in which someone flips a fair coin twice, and calculates these quantities:

$$X = \begin{cases} 0 & \text{if } TT \\ 1 & \text{if } HT \text{ or } TH \\ 2 & \text{if } HH \end{cases} \qquad Y = \begin{cases} 0 & \text{if } HH \text{ or } TT \\ 1 & \text{if } HT \text{ or } TH \end{cases}$$

Do you agree these are discrete random variables?

| | $Y = 0$ | $Y = 1$ |
|---|---|---|
| $X = 0$ | $\frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$ | 0 |
| $X = 1$ | 0 | |
| $X = 2$ | $\frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$ | 0 |

# Multivariate Probability Distributions

Consider an example in which someone flips a fair coin twice, and calculates these quantities:

$$X = \begin{cases} 0 & \text{if } TT \\ 1 & \text{if } HT \text{ or } TH \\ 2 & \text{if } HH \end{cases} \qquad Y = \begin{cases} 0 & \text{if } HH \text{ or } TT \\ 1 & \text{if } HT \text{ or } TH \end{cases}$$

Do you agree these are discrete random variables?

| | $Y = 0$ | $Y = 1$ |
|---|---|---|
| $X = 0$ | $\frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$ | $0$ |
| $X = 1$ | $0$ | $\frac{1}{2} \times \frac{1}{2} + \frac{1}{2} \times \frac{1}{2} = \frac{1}{2}$ |
| $X = 2$ | $\frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$ | $0$ |

# Multivariate Probability Distributions

Consider an example in which someone flips a fair coin twice, and calculates these quantities:

$$X = \begin{cases} 0 & \text{if } TT \\ 1 & \text{if } HT \text{ or } TH \\ 2 & \text{if } HH \end{cases} \qquad Y = \begin{cases} 0 & \text{if } HH \text{ or } TT \\ 1 & \text{if } HT \text{ or } TH \end{cases}$$

Do you agree these are discrete random variables? In math:

$$f(x, y) = \begin{cases} 1/2 & \text{if } x = 1, y = 1 \\ 1/4 & \text{if } x = y = 0 \text{ or } x = 2, y = 0 \\ 0 & \text{otherwise} \end{cases}$$

The CDF is straight-forward to calculate, but ugly to write.

# Multivariate Probability Distributions

The PMF/PDF and the CDF for a joint distribution needs to follow the same rules as for a univariate distribution.

To calculate the CDF for a bivariate joint distribution, you sum/integrate over both variables:

$$\underbrace{F(x,y) = \sum_{a \leq x} \sum_{b \leq y} f(a,b)}_{\text{for discrete RVs}} \quad \underbrace{F(x,y) = \int_{a \leq x} \int_{b \leq y} f(a,b)\ da\ db}_{\text{for continuous RVs}}$$

An example from the previous distribution:

$$\begin{aligned} F(1,1) = \Pr[X \leq 1, Y \leq 1] &= f(0,0) + f(0,1) + f(1,0) + f(1,1) \\ &= \frac{1}{4} + 0 + 0 + \frac{1}{2} = \frac{3}{4} \end{aligned}$$

# Multivariate Probability Distributions

Important note: the joint PMF/PDF and the joint CDF tell us about the *relationship* between the two variables.

$\hookrightarrow$ If the two variables are related, it will be apparent from looking at their joint PMF/PDF/CDF.

Compare two extreme examples of joint PMFs over two RVs:

**no correlation**

| | $Y = 0$ | $Y = 1$ |
|---|---|---|
| $X = 0$ | $\frac{1}{4}$ | $\frac{1}{4}$ |
| $X = 1$ | $\frac{1}{4}$ | $\frac{1}{4}$ |

**perfect negative correlation**

| | $Y = 0$ | $Y = 1$ |
|---|---|---|
| $X = 0$ | 0 | $\frac{1}{2}$ |
| $X = 1$ | $\frac{1}{2}$ | 0 |

# Marginal Distributions

A joint distribution tells us about the relationship between random variables, but it also tells us about each variable separately.

The **marginal distribution of $X$** gives us the distribution of $X$, ignoring how $X$ varies with $Y$. Formally:

$$\underbrace{f_X(x) = \sum_{y \in \mathcal{Y}} f(x, y)}_{\text{for discrete RVs}} \qquad \underbrace{f_X(x) = \int_{y \in \mathcal{Y}} f(x, y)\; dy}_{\text{for continuous RVs}}$$

Important: if we start with joint distribution, we can derive each variable's marginal distribution. But we cannot go the other way.

↪ Going from marginals to a joint distribution is called **ecological inference**, and it creates a whole host of problems.

# Marginal Distributions

Think back to our first example of a joint distribution:

| | $Y = 0$ | $Y = 1$ |
|---|---|---|
| $X = 0$ | $\frac{1}{4}$ | 0 |
| $X = 1$ | 0 | $\frac{1}{2}$ |
| $X = 2$ | $\frac{1}{4}$ | 0 |

$$f_X(x) = \begin{cases} \frac{1}{4} & \text{if } x = 0 \text{ or } x = 2 \\ \frac{1}{2} & \text{if } x = 1 \end{cases} \qquad f_Y(y) = \frac{1}{2} \text{ for all } y \in \mathcal{Y}$$

# Conditional Distribution

Marginal distributions tell us the distribution of a particular variable *ignoring* the other variable(s).

In contrast, the **conditional distribution of $X$ given $Y$** gives you a distribution over $X$ when holding $Y$ fixed at specific values. Formally:

$$f_{X|Y}(x|y) = \frac{\overbrace{f(x,y)}^{\text{joint}}}{\underbrace{f_Y(y)}_{\text{marginal for } Y}}$$

# Conditional Distribution

Recall from basic probability that **Bayes's rule** tells us how prior information and new information can be combined to create an updated assessment.

Applying it to probability distributions, we have:

$$\underbrace{f_{Y|X}(y|x)}_{\text{posterior}} = \frac{\overbrace{f_{X|Y}(x|y)}^{\text{likelihood}}\,\overbrace{f_Y(y)}^{\text{prior}}}{f_X(x)}$$

Practically speaking, Bayes's rule is very useful for reversing the "direction" of the conditional probability.

# Conditional Distribution

Why would someone want to know the conditional distribution?

Because this is where you can most easily see the relationship between the variables!

The conditional distribution tells you how the distribution over one variable changes as the other variable changes.

$\hookrightarrow$ The distribution over the first variable will only change as the second variable changes <u>if</u> they are related!

If they are not related, we say that they are **(stochastically) independent**. Formally, $X \perp\!\!\!\perp Y$ if and only if:

$$f(x,y) = f_X(x)f_Y(y) \Longrightarrow f_{X|Y}(x|y) = f_X(x) \text{ and } f_{Y|X}(y|x) = f_Y(y)$$

# Conditional Distribution

Think back to our first example of a joint distribution:

| | $Y = 0$ | $Y = 1$ |
|---|---|---|
| $X = 0$ | $\frac{1}{4}$ | $0$ |
| $X = 1$ | $0$ | $\frac{1}{2}$ |
| $X = 2$ | $\frac{1}{4}$ | $0$ |

↪ Let's consider the case of $X$ conditional on $Y = 0$

↪ What is $f_Y(Y = 0)$?

↪ $f_Y(Y = 0) = \frac{1}{2}$ (i.e. half of the time, $Y$ is equal to 0

We know that the conditional distribution for $Y = 0$ is defined as:

$$f_{X|Y}(x|Y = 0) = \frac{f(x, Y = 0)}{f_Y(Y = 0)}$$

# Conditional Distribution

| | $Y = 0$ | $Y = 1$ |
|---|---|---|
| $X = 0$ | $\frac{1}{4}$ | 0 |
| $X = 1$ | 0 | $\frac{1}{2}$ |
| $X = 2$ | $\frac{1}{4}$ | 0 |

We can just do this for all possible values of $X$:

$$f_{X|Y}(X = 0|Y = 0) = \frac{f(X = 0, Y = 0)}{f_Y(Y = 0)} = \frac{1/4}{1/2} = \frac{1}{2}$$

And so on - we get:

$$f_{X|Y}(x|Y = 0) = \begin{cases} 1/2 & \text{if } x = 0 \\ 0 & \text{if } x = 1 \\ 1/2 & \text{if } x = 2 \end{cases}$$

# Summary so far

We are working with random variables.

A random variable $X$ can take on a countable (discrete case) or uncountable (continuous case) number of values.

We call the set of possible values that $X$ can take $\mathcal{X}$.

$\hookrightarrow$ Die roll example: $\mathcal{X} = \{1, 2, 3, 4, 5, 6\}$

The possible values of RV $X$ are called $x$.

When we say $\sum_{x \in \mathcal{X}}$ or $\int_{x \in \mathcal{X}}$, we mean that we sum or integrate over all the possible values $x$ of the random variable $\mathcal{X}$.

$\hookrightarrow$ Die roll example:
$\sum_{x \in \mathcal{X}} Pr(X = x) = Pr(X = 1) + \cdots + Pr(X = 6) = 1$

# Summary so far

**Probability distributions:** functions that tell us how likely certain values $x$ of RV $X$ is. We call these functions $f(x)$.

- Discrete case: $f(x) = Pr(X = x)$ (*Probability mass function*)
- Continuous case: $Pr(a \leq X \leq b) = \int_a^b f(x)dx$ (*Probability density function*)
- Continuous case: $Pr(X = x) = 0$ for all $x$

**Marginal distribution:** distribution of a particular variable ignoring the other variable(s) ($f_X(x)$)

**Conditional distribution:** distribution of $X$ when holding $Y$ fixed at specific values ($f_{X|Y}(x|y)$).

# Summary so far

**Some important relationships:**

$\hookrightarrow$ Marginal distribution: $f_X(x) = \sum_{y \in \mathcal{Y}} f(x,y)$ (discrete) or $f_X(x) = \int_{y \in \mathcal{Y}} f(x,y)\, dy$ (continuous)

$\hookrightarrow$ Conditional distribution: $f_{X|Y}(x|y) = \frac{f(x,y)}{f_Y(y)}$

$\hookrightarrow$ Bayes rule: $f_{Y|X}(y|x) = \frac{f_{X|Y}(x|y) f_Y(y)}{f_X(x)}$

$\hookrightarrow$ Independence: $X \perp\!\!\!\perp Y \iff f(x,y) = f_X(x) f_Y(y)$

$\hookrightarrow$ Independence (2nd defintion): $X \perp\!\!\!\perp Y \iff f_{X|Y}(x|y) = f_X(x)$ and $f_{Y|X}(y|x) = f_Y(y)$

Always remember: $X$ is the random variable, $x$ is the value of the random variable.

# Summarizing Probability Distributions

Recall from Topics 2 and 3, we wanted to *summarize* empirical distributions because we wanted to reduce complexity.

We summarized them with *central tendency* and *dispersion*.

We'll do the same thing for "theoretical" probability distributions!

The ideas will be very similar, but will involve some different notation and more abstract mathematical formulas.

To reiterate: we are moving from real world data to theoretical distributions of how data could hypothetically be generated.

$\hookrightarrow$ So, summaries of probability distributions involve *expectations* about what kind of data they would hypothetically generate.

# Central Tendency of Probability Distributions

The **expected value** of a discrete random variable $X$ is:

$$E(X) = \sum_{x \in \mathcal{X}} x f(x)$$

The **expected value** of a continuous random variable $X$ is:

$$E(X) = \int_{x \in \mathcal{X}} x f(x)\ dx$$

# Central Tendency of Probability Distributions

Roughly speaking, you can think about the **expectation operator** $E(\cdot)$ as telling you to take a (weighted) average.

$\hookrightarrow$ Here is a basic example from a discrete PMF:

$$f(x) = \begin{cases} 1/3 & \text{if } x = 0 \\ 1/3 & \text{if } x = 1 \\ 1/3 & \text{if } x = 2 \end{cases}$$

What is the expectation?

# Central Tendency of Probability Distributions

↪ Here is a basic example from a discrete PMF:

$$f(x) = \begin{cases} 1/3 & \text{if } x = 0 \\ 1/3 & \text{if } x = 1 \\ 1/3 & \text{if } x = 2 \end{cases}$$

What is the expectation?

$$E(X) = \frac{1}{3} \cdot 0 + \frac{1}{3} \cdot 1 + \frac{1}{3} \cdot 2 = \frac{1}{3} + \frac{2}{3} = 1$$

We simply multiply all possible values of $X$ by $Pr(X = x)$, which are given above.

↪ $E(X)$ is the value of $g$ that minimizes MSE: $E((X - g)^2)$.

# Central Tendency of Probability Distributions

Expectations have some nice properties:

1. The expected value of a constant $a$ is that constant: $E(a) = a$.
2. The expected value of a constant $b$ times an RV $X$ is that constant times the expected value of $X$: $E(bX) = bE(X)$.
3. Given two random variables, $X$ and $Y$, $E(X + Y) = E(X) + E(Y)$.
4. $E(X)$ always falls between the minimum and maximum values that $X$ can take: $\min \mathcal{X} \leq E(X) \leq \max \mathcal{X}$.

One way to summarize the properties 1 through 3 is that the expectation is a **linear operator**.

# Central Tendency of Probability Distributions

Consider two random variables, depicted in this table:

|  | 0 | 1 | 2 | 3 |
|---|---|---|---|---|
| $X_1$ | $\frac{5}{12}$ | 0 | $\frac{1}{4}$ | $\frac{1}{3}$ |
| $X_2$ | $\frac{1}{4}$ | $\frac{1}{4}$ | $\frac{1}{4}$ | $\frac{1}{4}$ |

(Note that this is **not** a joint distribution)

What are $E(X_1)$, $E(X_2)$ and $E(X_1 + X_2)$?

$\hookrightarrow$ No need to do the actual calculation, but how would we do this if we wanted to?

# Central Tendency of Probability Distributions

| | 0 | 1 | 2 | 3 |
|---|---|---|---|---|
| $X_1$ | $\frac{5}{12}$ | 0 | $\frac{1}{4}$ | $\frac{1}{3}$ |
| $X_2$ | $\frac{1}{4}$ | $\frac{1}{4}$ | $\frac{1}{4}$ | $\frac{1}{4}$ |

$$E(X_1) = \frac{5}{12} \cdot 0 + 0 \cdot 1 + \frac{1}{4} \cdot 2 + \frac{1}{3} \cdot 3 = \frac{1}{2} + 1 = \frac{3}{2}$$

$$E(X_2) = \frac{1}{4} \cdot 0 + \frac{1}{4} \cdot 1 + \frac{1}{4} \cdot 2 + \frac{1}{4} \cdot 3 = \frac{1+2+3}{4} = \frac{6}{4} = \frac{3}{2}$$

$$E(X_1 + X_2) = E(X_1) + E(X_2) = \frac{3}{2} + \frac{3}{2} = 3$$

You can check also do the $E(X_1 + X_2)$ calculation by hand, but it's a bit tedious.

# Central Tendency of Probability Distributions

What if we wanted to transform a random variable by applying a function to it?

$\hookrightarrow$ For example, taking the RV $X$ and doing: $\sqrt{X}$ or $X^2$.

For any function $g$ of a discrete RV $X$:

$$E(g(X)) = \sum_{x \in \mathcal{X}} g(x) f(x)$$

And for any function $g$ of a continuous RV $X$:

$$E(g(X)) = \int_{x \in \mathcal{X}} g(x) f(x) \, dx$$

This is called the **law of the unconscious statistician** (LOTUS).

# Dispersion of Probability Distributions

The **variance** of a random variable $X$ is:

$$Var(X) = E((X - E(X))^2)$$

Some people label variance as: $V(\cdot)$, $\mathbb{V}(\cdot)$, without italics $\mathrm{Var}(\cdot)$, with square brackets or as $\sigma^2$ (Greek "sigma").

Recall that the *sample variance* is defined as:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

$Var(X)$ and $s^2$ are very similar !

# Dispersion of Probability Distributions

Using some algebra and the properties of expectations, we can write the variance in a different way:

$$\begin{aligned} Var(X) &= E((X - E(X))^2) \\ &= E(X^2 - 2XE(X) + E(X)^2) \\ &= E(X^2) - 2E(X)E(X) + E(X)^2 \\ &= E(X^2) - E(X)^2 \end{aligned}$$

# Dispersion of Probability Distributions

↪ Here is a basic example from a discrete PMF:

$$f(x) = \begin{cases} 1/3 & \text{if } x = 0 \\ 1/3 & \text{if } x = 1 \\ 1/3 & \text{if } x = 2 \end{cases}$$

What is the variance? Note that we know that $E(X) = 1$ from before.

$$E((X-E(X))^2) = \frac{1}{3}\cdot(0-1)^2+\frac{1}{3}\cdot(1-1)^2+\frac{1}{3}\cdot(2-1)^2 = \frac{1}{3}+\frac{1}{3} = \frac{2}{3}$$

Alternatively, we can also use the formula from the previous slide:

$$E(X^2) - E(X)^2 = \frac{1}{3}\cdot 0^2 + \frac{1}{3}\cdot 1^2 + \frac{1}{3}\cdot 2^2 - 1^2 = \frac{1}{3} + \frac{4}{3} - 1 = \frac{2}{3}$$

# Dispersion of Probability Distributions

The variance of a random variable also has nice properties:

1. The variance of a constant $a$ is zero: $Var(a) = 0$.
2. $Var(X)$ is always weakly positive: $Var(x) \geq 0$.
3. The variance of a constant $b$ times an RV $X$ is that constant squared times the variance of $X$: $Var(bX) = b^2 Var(X)$.
4. If $X$ and $Y$ are stochastically independent, then $Var(X + Y) = Var(X) + Var(Y)$.

Variance is <u>not</u> a linear operator.

# Other Measures of Central Tendency

You can also calculate the mean and mode of a probability distribution.

The median of a random variable $X$ is just a value $x_M$ such that:

$$F(x_M) = \frac{1}{2}$$

Recall that $F(x)$ is defined as the CDF of $X$, i.e. $F(x) = \Pr(X \leq x)$.

# Other Measures of Central Tendency

The mode of a random variable $X$ is just a value $x_D$ where:

$$f(x_D) = \max_{x \in \mathcal{X}} f(x)$$

Does the last expression (for the mode) make sense for continuous random variables?

Practically speaking, we tend to talk less about medians and modes in the context of theoretical probability distributions.

# Moments

This is a little bit technical, but both the expected value and the variance are **moments** of a distribution.

There is a potentially infinite number of moments of a distribution, but in practice we tend to focus on four:

$\hookrightarrow$ The expected value is the **first moment** of a distribution.

$\hookrightarrow$ The variance is the **second central moment** of a distribution.

The **skew** of a distribution is the **third central moment** of that distribution.

$\hookrightarrow$ Skew measures how "symmetric" the distribution is.

# Moments

This is a little bit technical, but both the expected value and the variance are **moments** of a distribution.

There is a potentially infinite number of moments of a distribution, but in practice we tend to focus on four:

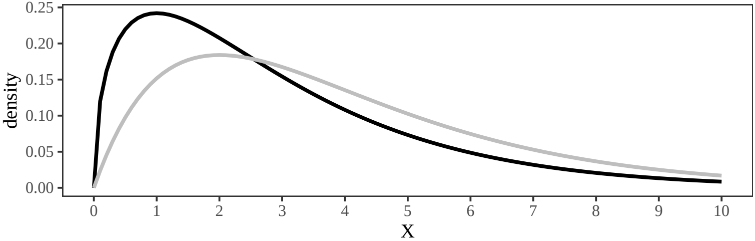
$\hookrightarrow$ The expected value is the **first moment** of a distribution.

$\hookrightarrow$ The variance is the **second central moment** of a distribution.

The **kurtosis** of a distribution is the **fourth central moment** of that distribution.

$\hookrightarrow$ Kurtosis measures how "fat" the tails of a distribution are.

# Moments



These distributions each have a right skew.

# Summarizing Relationships

When we have a multivariate probability distribution, we may also want to summarize the relationship between variables.

Just as we did in Topic 3, we'll focus on three measures of relationships.

But first note: if you have a joint distribution over $X$ and $Y$, you can always "back out" the marginal distributions and calculate:

$$E(X) = \sum_{x \in \mathcal{X}} x f_X(x) \qquad E(Y) = \sum_{y \in \mathcal{Y}} y f_Y(y)$$

(These are for discrete RVs, but the same idea applies to continuous RVs. You just integrate instead.)

# Measuring Relationships (1): Conditional Expectation

The **conditional expectation** of a discrete RV $Y$ given $X = x$ is:

$$E(Y|X = x) = \sum_{y \in \mathcal{Y}} y f_{Y|X}(y|x)$$

The **conditional expectation** of a continuous RV $Y$ given $X = x$ is:

$$E(Y|X = x) = \int_{y \in \mathcal{Y}} y f_{Y|X}(y|x)\ dy$$

What does this do? Holding the $X$ variable fixed at a specific value $x$, calculate the expected value of $Y$.

# Measuring Relationships (1): Conditional Expectation

Let's do an example:

| | $Y = 0$ | $Y = 1$ |
|---|---|---|
| $X = 0$ | $\frac{1}{4}$ | 0 |
| $X = 1$ | 0 | $\frac{1}{2}$ |
| $X = 2$ | $\frac{1}{4}$ | 0 |

What is $E(X|Y = 1)$? Note that we first need the conditional probability distribution $f_{X|Y=1}(x|Y = 1)$.

Basically, we are asking: What is $Pr(X = x|Y = 1)$.

# Measuring Relationships (1): Conditional Expectation

|  | $Y = 0$ | $Y = 1$ |
|---|---|---|
| $X = 0$ | $\frac{1}{4}$ | 0 |
| $X = 1$ | 0 | $\frac{1}{2}$ |
| $X = 2$ | $\frac{1}{4}$ | 0 |

Since $Y = 1$, we have $Pr(X = 0|Y = 1) = 0$,
$Pr(X = 1|Y = 1) = 1$, and $Pr(X = 2|Y = 1) = 0$.

Now, we can just plug this into the formula for the expectation.

$$E(X|Y = 1) = 0 \cdot 0 + 1 \cdot 1 + 2 \cdot 0 = 1$$

# Measuring Relationships (1): Conditional Expectation

Similarly, we can define the **conditional variance** of a RV $Y$ given $X$ as:

$$Var(Y|X = x) = E((Y - E(Y|x))^2|x)$$

What does this do? Holding the $X$ variable fixed at a specific value $x$, calculate the variance of $Y$.

# Measuring Relationships (1): Conditional Expectation

**Brief summary:**

From the joint distribution $f(x, y)$, we can calculate the marginal distributions $f_X(x)$ and $f_Y(y)$.

$\hookrightarrow$ Then, we can calculate the conditional distributions $f_{X|Y}(x|y)$ and $f_{Y|X}(y|x)$.

$\hookrightarrow$ In the discrete case, the conditional distribution of $X$ given $Y$ is defined as:

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

Then, we can calculate the conditional expectations $E(Y|x)$ and $E(X|y)$ (discrete case).

# Measuring Relationships (1): Conditional Expectation

Both $E(Y|X=x)$ and $Var(Y|X=x)$ return only one number.

However, most of the time, we want to know how the expected value of $Y$ changes as we change the value of $X$.

The **conditional mean function** is $E(Y|X)$.

What does this do? For any value of $X$ you plug in, calculate the expected value of $Y$.

We can also define the **conditional variance function**: $Var(Y|X)$.

# Measuring Relationships (1): Conditional Expectation

Think back to our first example of a joint distribution:

| | $Y = 0$ | $Y = 1$ |
|---|---|---|
| $X = 0$ | $\frac{1}{4}$ | 0 |
| $X = 1$ | 0 | $\frac{1}{2}$ |
| $X = 2$ | $\frac{1}{4}$ | 0 |

$$E(Y|X = x) = \begin{cases} 0 & \text{if } x = 0 \\ 1 & \text{if } x = 1 \\ 0 & \text{if } x = 2 \end{cases} \quad E(X|Y = x) = \begin{cases} 1 & \text{if } y = 0 \\ 1 & \text{if } y = 1 \end{cases}$$

# Measuring Relationships (1):
# Conditional Expectation

Think back to our first example of a joint distribution:

|  | $Y = 0$ | $Y = 1$ |
|---|---|---|
| $X = 0$ | $\frac{1}{4}$ | 0 |
| $X = 1$ | 0 | $\frac{1}{2}$ |
| $X = 2$ | $\frac{1}{4}$ | 0 |

$$E(Y|X) = \begin{cases} 0 & \text{if } x = 0 \\ 1/2 & \text{if } x = 1 \\ 0 & \text{if } x = 2 \end{cases}$$

# Measuring Relationships (2): Covariance

The **covariance** of two random variables $X$ and $Y$ is defined as:

$$Cov(X, Y) = E((Y - E(Y))(X - E(X)))$$

Using some algebra and the properties of expectations, we can rearrange this:

$$\begin{aligned} Cov(X, Y) &= E((Y - E(Y))(X - E(X))) \\ &= E(YX - E(Y)X - E(X)Y + E(X)E(Y)) \\ &= E(YX) - E(Y)E(X) - E(X)E(Y) + E(X)E(Y) \\ &= E(YX) - E(Y)E(X) \end{aligned}$$

Note that: $Cov(X, Y) = Cov(Y, X)$ and $Cov(X, X) = Var(X)$.

# Measuring Relationships (3): Correlation

For the same reason as we did in Topic 3, we typically want to transform covariance into a measure that has a common scale.

So, the **correlation** of two random variables $X$ and $Y$ is defined as:

$$Corr(X, Y) = \frac{Cov(X, Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}}$$

Again, note that $Corr(X, Y) = Corr(Y, X)$.

# Stochastic Independence

Since expectations are just ways to summarize probability distributions, then they can also tell us whether variables in a distribution are stochastically independent.

If two RVs $X$ and $Y$ are stochastically independent (i.e., $X \perp\!\!\!\perp Y$) then:

1. $E(Y|X) = E(Y)$ and $E(X|Y) = E(X)$.
2. $Cov(X, Y) = 0$.
3. $Corr(X, Y) = 0$.

# Brief Recap of Where We've Been

We started by calculating statistics for real life datasets.

The goal of doing that is to learn about the larger world.

But we don't see the larger world! Only the dataset(s) we have.

The DGP is the term we give to the forces in the larger world that generate datasets that we actually see.

If we want to learn about the larger world, we need *some* idea of what it looks like. We can't ask too much of our data.

This amounts to making some presumptions about what the DGP looks like. In other words, we want to create a **model** of the DGP.

# Brief Recap of Where We've Been

If we assume it's stochastic (reasonable!), then we are admitting there's some random chance in the things we're studying.

We model stochastic DGPs using probability distributions.

↪ This is known as **statistically modeling** the DGP.

Probability distributions are theoretical models of how data can arise: we specify all the components.

So, a statistical model of a DGP is a theoretical model of how we think the DGP operates (remember: we can't see it!).

Part 4 (this one) is about the process of picking an appropriate statistical model to represent a DGP we're interested in.

# Where We're Going

Why do we even need to make a statistical model of a DGP?

I've hinted at the answer already: we can't expect too much of our data, which is a limited snapshot of the world.

Whether you realize it or not, you're already (implicitly or subconsciously) modeling the DGP when you calculate some statistics with a dataset. For example: when you run a regression.

At a minimum, we should strive to be transparent about doing that.

More importantly, it helps us figure out what exactly we need to do with our dataset when we are trying to learn about the world.

# Where We're Going

Starting with the next part of Topic 4, we'll explore this:

↪ We'll look at **sampling distributions** which link a DGP to a specific dataset.

↪ We'll look at **hypothesis testing** which allow us to evaluate (theoretical) claims about the DGP.

↪ We'll look at **estimation** which allows us to calculate ("estimate") unknown features of the DGP.

All of this is statistical inference: a way for us to learn about the world given that we only have one dataset and have some uncertainty about its connection to the real world.

The tools of statistical inference don't solve all our problems. Topic 5 through the end of the course is about these problems.

# Statistical Models

There are many (infinite!) different probability distributions:

$\hookrightarrow$ Is it discrete or continuous?

$\hookrightarrow$ How many possible values can it take?

$\hookrightarrow$ What are the probabilities/densities that describe how likely different values are?

We're interested in real-life DGPs, so we want to pick probability distributions that "resemble" the thing we're interested in.

The process of picking a probability distribution that resembles the real-life DGP we're studying is **statistical modeling**.

# Statistical Models

Practically speaking, there are a bunch of well-known and widely understood probability distributions that people use.

In these lecture slides, we'll look at a few of them.

When you're doing statistical modeling, you should try to find probability distributions that resemble the things you're studying.

We're going to look at probability distributions that are defined by certain **parameters**.

There are many "**parametric families**" that have similar shapes but vary in the *specific* shape they take.

# Binary Random Variables

Scenario: you have a binary dependent variable, $Y$.

$\hookrightarrow$ You've already assumed it's a RV if you're using statistics!

$\hookrightarrow$ For example, did a court case favor the plaintiff?

A **Bernoulli distribution** is a discrete probability distribution where $\mathcal{Y} = \{0, 1\}$ and the PMF is defined by:

$$f(y) = \begin{cases} 1 - p & \text{if } y = 0 \\ p & \text{if } y = 1 \end{cases}$$

This distribution has <u>one</u> parameter, $p$.

One classic example: a fair coin flip, whose random variable follows a Bernoulli distribution with $p = 1/2$.

# Binary Random Variables

A **binomial distribution** is a discrete probability distribution where an experiment with a binary outcome is repeated $N$ independent times.

$\hookrightarrow$ Bernoulli distribution is a special case of a binomial distribution!

The PMF is more complicated, but it's a function $f$ that specifies the probability of getting any possible combination of 1s and 0s.

This distribution has <u>two</u> parameters, $N$ and $p$.

Extending the example: 3 flips of a fair coin, whose random variable follows a binomial distribution with $N = 3$ and $p = 1/2$.

# Binary Random Variables

A random variable $Y$ that follows a Bernoulli distribution with parameter $p$ can be summarized with the following moments:

$$E(Y) = 1 \times p + 0 \times (1 - p) = p$$
$$Var(Y) = E(Y^2) - E(Y)^2 = p - p^2 = p(1 - p)$$

A random variable $Y$ that follows a binomial distribution with parameters $p$ and $N$ can be summarized with the following moments:

$$E(Y) = Np$$
$$Var(Y) = Np(1 - p)$$

↪ Remember: binomial distribution can be thought of as flipping a coin $N$ times, where the probability of one side is $p$, and the other side is $1 - p$.

# Summary so far – special distributions

Some distributions are common, describe many related scenarios.

$\hookrightarrow$ These can be described by *parameters.*

Bernoulli distribution – flipping a coin with prob. $p$: $Y \sim \mathrm{Bern}(p)$

$$f(y) = \begin{cases} 1-p & \text{if } y = 0 \\ p & \text{if } y = 1 \end{cases}$$

Binomial distribution: flipping a coin $N$ times with prob. $p$: $Y \sim \mathrm{Bin}(N, p)$

$$f(y) = \begin{cases} \binom{N}{y} p^y (1-p)^{N-y} & \text{if } y \in \{0, 1, 2, ..., N\} \\ 0 & \text{otherwise} \end{cases}$$

# Summary so far – special distributions

If we see Bernoulli or binomial distributions, we instantly know a lot about the data.

$\hookrightarrow$ We know their expected value and variance.

Bernoulli RV:

$$E(Y) = p \qquad Var(Y) = p(1-p)$$

Binomial RV:

$$E(Y) = Np \qquad Var(Y) = Np(1-p)$$

These depend only on the *parameters* .

# Some Notes

When we have a random variable that is assumed to be distributed according to some specific probability distribution, we have a special way to write that.

For a binomial RV $Y$ that's distributed according to a binomial distribution with parameters $N$ and $p$, we can write it in one of several ways:

$$Y \sim B(N, p) \qquad Y \sim \mathrm{Bin}(N, p) \qquad Y \sim \mathcal{B}(N, p)$$

These are all saying the same thing.

# Some Notes

Once we model a variable $Y$ as a random variable following a particular distribution, the parameters determine the "fundamentals" and everything else is just random chance.

$\hookrightarrow$ For example, a fair coin flip ($p = 1/2$) will result in approximately 500 heads out of $N = 1,000$ flips

$\hookrightarrow$ For a coin flip with $p = 4/5$ will result in approximately 800 heads out of $N = 1,000$ flips.

# Other Scenarios

It's very common in political science to be interested in a dependent variable that is binary.

- Did a country experience a civil war in a given year?
- Did a person turn out to vote in an election?
- Is a person exposed to misinformation or not?

But, it is not the only scenario that comes up in research!

The wikipedia pages for each distribution are really helpful!

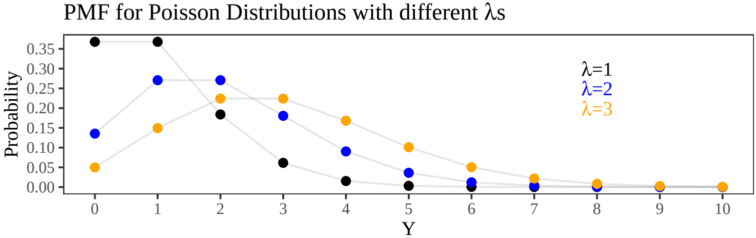
For example: https://en.wikipedia.org/wiki/Bernoulli_distribution

# Other Scenarios

Scenario: you have a dependent variable $Y$ that is an event count.

$\hookrightarrow$ For example: number of countries at war in each year.

You could model this as a **Poisson distribution**: a discrete probability distribution where $\mathcal{Y} = \{0, 1, 2, ...\}$, and which has one parameter usually labeled $\lambda$.

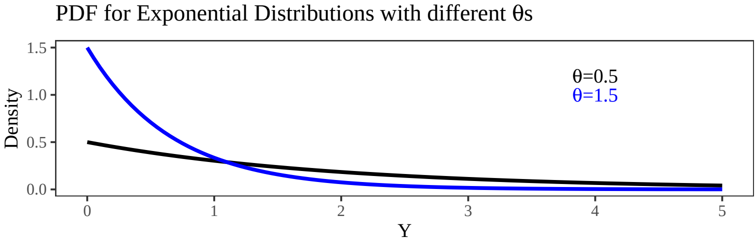


# Other Scenarios

Scenario: you have a dependent variable $Y$ that is a duration.

$\hookrightarrow$ For example: how long a court case lasts.

You could model this as an **exponential distribution**: a continuous probability distribution where $\mathcal{Y} = [0, \infty)$, and which has one parameter labeled $\theta$ here.



# Other Scenarios

Scenario: you have a variable $X$ that can take one of many values, but each is equally likely.

↪ For example, whether one out of three available judges will be (randomly) selected to hear a case.

You could model this as an **uniform distribution**.

↪ You've seen many examples of discrete uniform distributions: fair coin flips, fair die rolls, etc.

# Other Scenarios

Scenario: you have a variable $X$ that can take one of many values, but each is equally likely.

$\hookrightarrow$ For example, whether one out of three available judges will be (randomly) selected to hear a case.

You could model this as an **uniform distribution**.

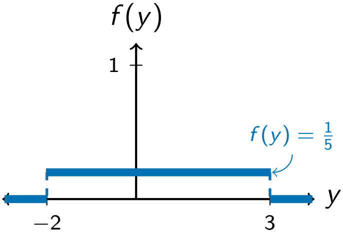
A **continuous uniform distribution** is a continuous probability distribution with two parameters $a$ and $b$, where $\mathcal{Y} = [a, b]$ and:

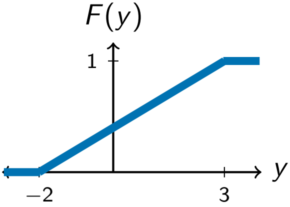
$$f(y) = \begin{cases} \frac{1}{b-a} & \text{if } y \in [a, b] \\ 0 & \text{otherwise} \end{cases} \quad F(y) = \begin{cases} \frac{y-a}{b-a} & \text{if } y \in [a, b] \\ 0 & \text{otherwise} \end{cases}$$

# Other Scenarios

Suppose we have $Y \sim \mathcal{U}[-2, 3]$. Then:

$$f(y) = \frac{1}{5} \qquad F(y) = \frac{y+2}{5}$$





# Other Scenarios

Mean and variance of the uniform distribution are defined as follows:

$$E(Y) = \frac{a+b}{2}$$

$$Var(Y) = \frac{(b-a)^2}{12}$$

# What If You Don't Know?

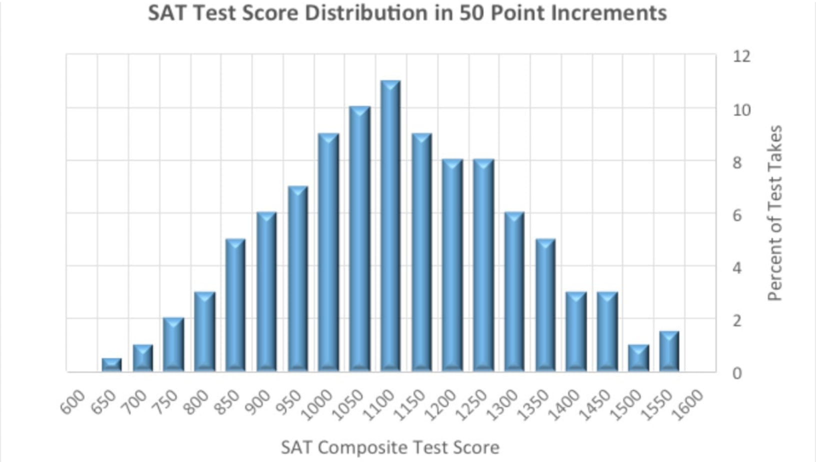
Each of the previous probability distributions made sense in different scenarios.

But what if you don't know how to statistically model the variable you're looking at?

↪ For example: height of students in a class. Not a duration, not a count, not binary, etc.

In cases like this, we might consider using the **normal distribution** as a good approximation for the DGP.

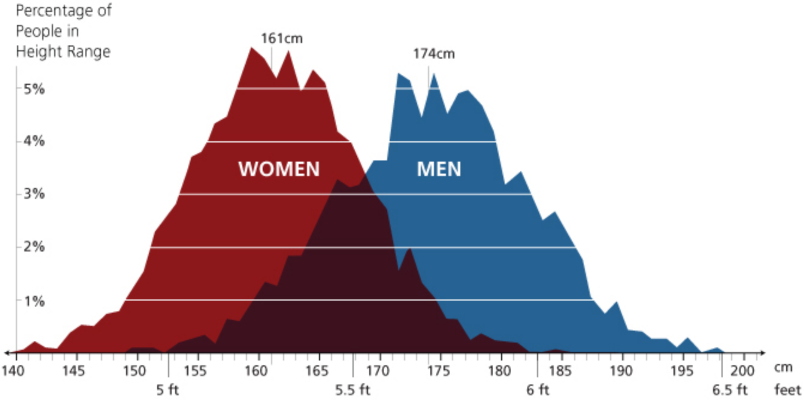
# What If You Don’t Know?



# What If You Don't Know?

**Height of Adult Women and Men**

Within-group variation and between-group overlap are significant



Data from U.S. CDC, adults ages 18-86 in 2007

# What If You Don't Know?

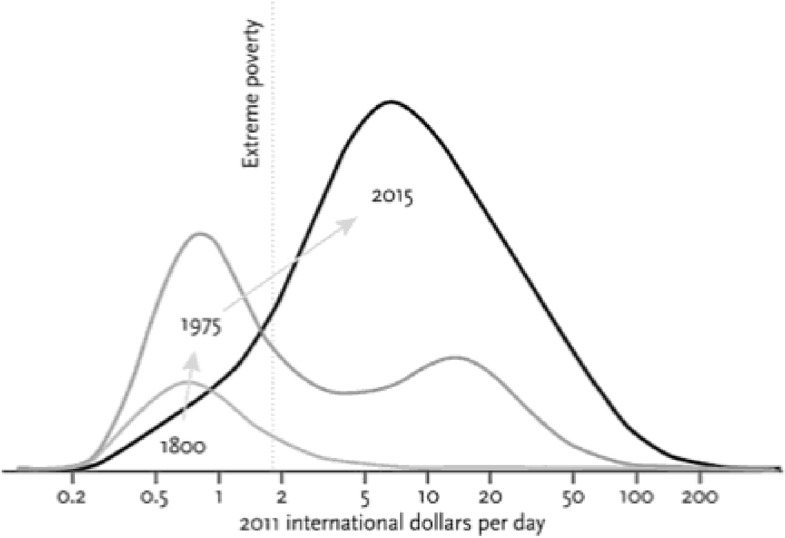


Figure 8-3: World income distribution, 1800, 1975, and 2015

**Source:** *Gapminder,* via Ola Rosling, http://www.gapminder.org/tools/mountain. The scale is in 2011 international dollars.

# What If You Don't Know?

Three reasons why the normal distribution is often used:

$\hookrightarrow$ Useful mathematical properties

$\hookrightarrow$ Shows up often in real life (see previous slides)

$\hookrightarrow$ Functions of observations in large samples are often normal

- We will talk about this in the next topic.

# What If You Don't Know?

If $Y$ is distributed according to a **normal distribution**, then it is a continuous random variable with $\mathcal{Y} = \mathbb{R}$ and with a PDF that has two parameters, usually labeled $\mu$ (mean) and $\sigma^2$ (variance).

$\hookrightarrow$ This is sometimes called a **Gaussian distribution**.

$\hookrightarrow$ The short way to write that a random variable $Y$ is normally distributed with mean $\mu$ and variance $\sigma^2$ is: $Y \sim \mathcal{N}(\mu, \sigma^2)$.

The PDF is ugly, but creates the famous "bell curve" shape:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

# What If You Don't Know?

There's lots to love about the normal distribution, but one thing is that its first two central moments are easy to calculate:

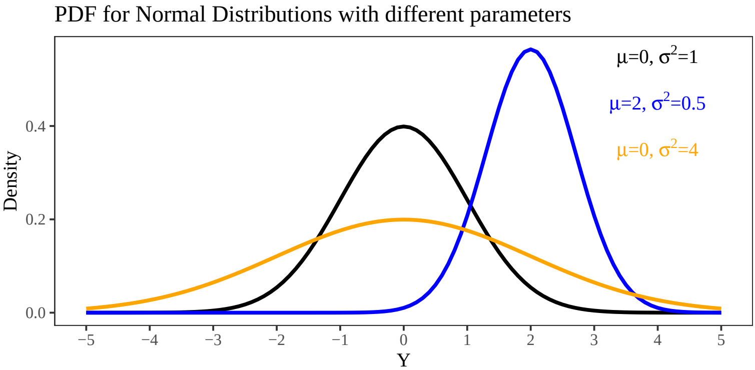
$$E(Y) = \mu$$
$$Var(Y) = \sigma^2$$

In other words, the two parameters that define the shape of the normal PDF *also* happen to be the expected value and variance.

The **standard normal distribution**: $\mu = 0$ and $\sigma^2 = 1$:

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

# What If You Don't Know?



# What If You Don't Know?

The normal distribution has some other useful properties:

For a standard normal distribution $X \sim N(0,1)$:

$P(X < -1.96) = 0.025$ and $P(X > 1.96) = 0.025$

$\hookrightarrow$ 95% of all values are between -1.96 and 1.96.

For multiple normal distributions $X_1, X_2, ..., X_n$, which are distributed with $X_i \sim (\mu_i, \sigma_i^2)$ and are independent:

$$X_1 + X_2 + ... + X_n \sim N(\mu_1 + \mu_2 + \ldots \mu_n, \sigma_1^2 + \sigma_2^2 + \cdots + \sigma_n^2)$$

$\hookrightarrow$ For independent normal RVs, their sum is also normal.

# Introducing an Independent Variable

So far, we've only looked at statistical models of a single variable.

This allows us to come up with distributions that resemble the kind of dependent variable that we might have in a real life dataset.

But usually, social scientists want to know how a dependent variable is related to one or more independent variables.

I'll now show you the most common way that social scientists statistically model relationships. (You've already seen it!)

↪ Here: we'll just focus on situations with one independent variable, but the ideas could extend to more variables.

# Introducing an Independent Variable

Consider a situation with a binary dependent variable $Y$.

You would model this as a Bernoulli distribution with parameter $p$.

You now want to model the relationship between this variable $Y$ and an independent variable $X$.

For example, you might be interested in $E(Y|X)$. (Why?)

If $Y$ follows a Bernoulli distribution, that means $E(Y) = p$.

And if we condition on $X$, then: $E(Y|X) = p(X)$.

$\hookrightarrow$ The parameter is a function of the independent variable $X$.

# Introducing an Independent Variable

The fact that $p$ is a function of the independent variable $X$ doesn't tell us *how* it is a function of $X$.

The **link function** specifies precisely how $p$ changes with $X$.

In many (possibly most) situations with a binary dependent variable $Y$, we typically assume a linear link function:

$$p(X) = E(Y|X) = \alpha + \beta X$$

This is the **linear regression model for DGPs**.

There are other link functions that are not linear.

$\hookrightarrow$ For example: **logit** and **probit**.

# Introducing an Independent Variable

What if we want to model a dependent variable $Y$ as a continuous random variable and not a binary random variable?

You do exactly the same thing (condition a parameter of $Y$ on $X$), except the parameter you now care about is $\mu$:

$$\mu(X) = E(Y|X) = \alpha + \beta X$$

When we calculated regression lines in Topic 3, we were *assuming* a linear relationship and estimating the values for $\alpha$ and $\beta$.

We called those values *a* and *b* to emphasize they were estimates from a real life dataset. We now use Greek letters because we're thinking about a probability distribution of a hypothetical relationship in a DGP.

# Introducing an Independent Variable

You will learn a lot more about this in POL 213.

Suffice to say: one way to statistically model the relationship between a binary dependent variable $Y$ and an independent variable $X$ is using the linear regression model for DGPs.