Soc 500: Applied Social Statistics
Week 1: Introduction and Probability

Brandon Stewart[1]

Princeton

September 3, 2024

[1] Many of these slides are adapted from slides by Matt Blackwell and Adam Glynn with contributions from Justin Grimmer, Erin Hartman, and Matt Salganik. Illustrations by Shay O'Brien.

Where We've Been and Where We're Going...

- Last Week
  - preparing for the semester
- This Week
  - course structure
  - core ideas
  - introduction to probability
  - three big ideas in probability
  - introduction to random variables
- Next Week
  - characterizing distributions
  - joint distributions
- Long Run
  - probability $\rightarrow$ inference $\rightarrow$ regression $\rightarrow$ causal inference

Welcome and Introductions

- I
  - . . . use he/him pronouns
  - . . . am an Associate Professor in Sociology
  - . . . am trained in political science and statistics
  - . . . do research in statistical text analysis and causal inference
  - . . . love doing collaborative research
  - . . . talk very quickly
- Your Preceptors
  - Sofia Avila
  - Christina Pao

Overview

- Goal: train you in statistical thinking.
- Fundamentally a graduate course for sociologists, but also useful for research in other fields, policy evaluation, industry etc.
- Difficult course but with many resources to support you.
- When we are done you will be able to teach yourself many things
- Syllabus is a useful resource including philosophy of the class.

Specific Goals

- critically read and reason about quantitative social science using linear regression techniques.
- conduct, interpret, and communicate results from analysis using multiple regression.
- explain the limitations of observational data for making causal claims and distinguish between identification and estimation.
- understand the logic and assumptions of several modern designs for making causal claims.
- write clean, reusable, and reliable `R` code in `tidyverse` style.
- feel empowered working with data

Why R?

- It will give you super powers (but not at first)
- It is free and open source
- It is the *de facto* standard in many applied statistical fields

Artwork by @allison_horst

Mathematical Prerequisites

- No formal pre-requisites.
- Balance of rigor and intuition.
  - no rigor for rigor's sake.
  - we will tell you *why* you need the math, but also feel free to ask.
  - course focus on how to reason about statistics, not just memorize guidelines.
- They key to success is pushing on what you don't know even when doing so is uncomfortable.
- This class is not about innate statistical aptitude, it is about effort.
- We all come from very different backgrounds. Please have patience with yourself and with others.

Ways to Learn

- Lectures
  learn broad topics
- Precept
  learn data analysis skills, get targeted help on assignments
- Problem Sets
  reinforce understanding of material, practice
- Ed
  ask questions of us and your classmates
- Office Hours
  ask us even more questions

Problem Sets

- Schedule (due Thursday at 8PM eastern)
- Grading and solutions
- Collaboration policy
- You may find these difficult. Start early and seek help!
- Initial sense of disconnect (coding & math)

Problem sets are the most important part of the class!

Ways to Learn

- Lectures and Precept
- Daily Feedback
- Readings and Slides
- Ed
- Office Hours
- Problem Sets
- Instructor Office Hours
- Final Exam Prep
- External Consulting

Your Job: work hard and get help when you need it!

Staying in Touch

Can we go over Bayes' Rule again?

A Note on Reading

- Think of the lecture slides as primary reading.
- If you want material to read, come talk to me about recommendations.
- Suggested Books (more in the syllabus!):
  - Angrist and Pischke. 2008. *Mostly Harmless Econometrics*
  - Aronow and Miller. 2019. *Foundations of Agnostic Statistics*
  - Blitzstein and Hwang. 2019. *Introduction to Probability: Second Edition*
- A somewhat obvious tip: don't skip the math!

Advice from Prior Generations

- Ask questions if you don't know what's going on!
- Investing a considerable amount of time in getting familiar with R and its various tools will pay off in the long run!
- Go over the lecture slides each week. This can be hard when you feel like you're treading water and just staying afloat, but I wish I had done this regularly.
- It's challenging but very doable and rewarding if you put the time in. There are plenty of resources to take advantage of for help.
- I found it helpful to read through the lecture slides again after I had opened the problem set. It made it easier to create connections between what we went through and how to do it.
- Go over your psets and the pset solutions the moment they are graded as a habit and figure out what you don't know.

Outline of Topics

Outline in reverse order:

- Causal Inference with Regression:
  inferring counterfactual effect given association using regression.
- Regression:
  estimate association.
- Observational and Causal Inference:
  estimating things we don't know from data.
- Probability:
  learning what data we would expect if we did know the truth.

Probability $\rightarrow$ Inference $\rightarrow$ Regression $\rightarrow$ Causal Inference

Attribution and Thanks

- My philosophy on teaching: don't reinvent the wheel customize, refine, improve.
- Huge thanks to those who have provided slides particularly: Matt Blackwell, Adam Glynn, Justin Grimmer, Jens Hainmueller, Erin Hartman, Kevin Quinn
- Also thanks to those who have discussed with me at length including Dalton Conley, Chad Hazlett, Gary King, Kosuke Imai, Matt Salganik and Teppei Yamamoto.
- Previous generations of preceptors who have made important contributions to the course: Clark Bernier, Emily Cantrell, Elisha Cohen, Ian Lundberg, Simone Zhang, Alex Kindel, Ziyao Tian, Shay O'Brien, Emily Cantrell, Alejandro Schugurensky, Max Fineman, and Angela Li.
- Shay O'Brien for many hand-drawn illustrations.

Welcome To Class!

Be sure to read the syllabus for more details.

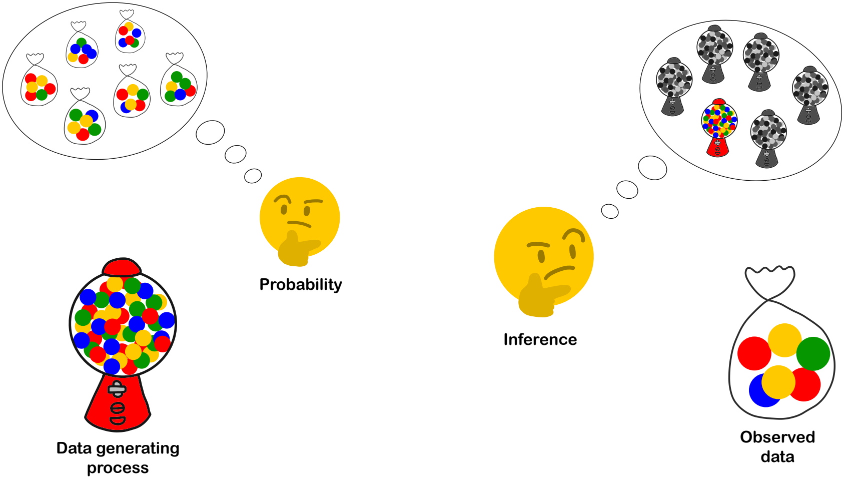
What is Statistics?

- Branch of mathematics studying collection and analysis of data
- The name statistic comes from the word state
- The arc of developments in statistics
  1) an applied scholar has a problem
  2) they solve the problem by inventing a specific method
  3) statisticians generalize and export the best of these methods
- Relatively recent field (started at end of 19th century)
- Goal: principled guesses with known properties based on stated assumptions.
- In practice, an essential part of research, policy making, political campaigns, selling people things...

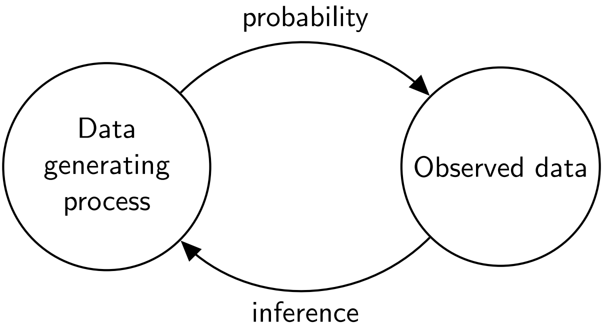
Why study probability?

It enables inference.

In Picture Form



In Picture Form



Example: The Lady Tasting Tea

- The Story Setup
  (lady discerning about tea)
- The Experiment
  (perform a taste test)
- The Hypothetical
  (count possibilities)
- The Result
  (boom she was right)

Tea-Tasting Distribution

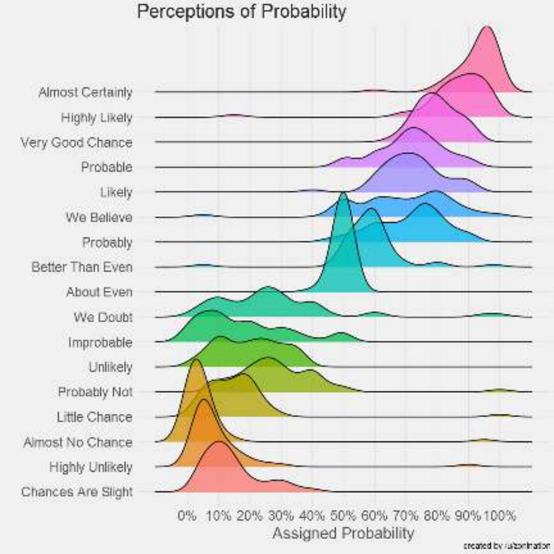
Success count	Permutations of selection	Number of permutations
0	oooo	$1 \times 1 = 1$
1	ooox, ooxo, oxoo, xooo	$4 \times 4 = 16$
2	ooxx, oxox, oxxo, xoxo, xxoo, xoox	$6 \times 6 = 36$
3	oxxx, xoxx, xxox, xxxo	$4 \times 4 = 16$
4	xxxx	$1 \times 1 = 1$
Total		70

This became the Fisher Exact Test.

A Note on Fisher and the History of Statistics

- The statistician in that story was Sir Ronald Fisher, arguably the most influential statistician of the 20th century.
- Besides founding key areas of statistics, Fisher was also one of the founders of population genetics.
- He was also a eugenicist and a racist.
- Statistics has been used intermittently as a force for progress and a force against progress.

From 'Probably' to Probability



Why Probability?

- Helps us envision hypotheticals
- Describes uncertainty in how the data is generated
- Estimates probability that something will happen
- Thus: we need to know how probability gives rise to data

Intuitive Definition of Probability

While there are several interpretations of what probability is, most modern (post 1935 or so) researchers agree on an axiomatic definition of probability.

3 Axioms (Intuitive Version):

1. The probability of any particular event must be non-negative.
2. The probability of anything occurring among all possible events must be 1.
3. The probability of one of many mutually exclusive events happening is the sum of the individual probabilities.

All the rules of probability can be derived from these axioms.
To state them formally, we first need some definitions.

Sample Spaces

To define probability we need to define the set of possible outcomes.

The sample space is the set of all possible outcomes, and is often written as Ω.

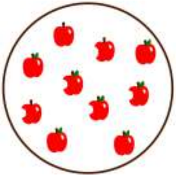
For example, if we flip a coin twice, there are four possible outcomes,

$$\Omega = \big\{\{heads, heads\}, \{heads, tails\}, \{tails, heads\}, \{tails, tails\}\big\}$$

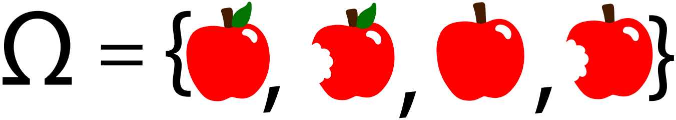
Thus the table in Lady Tasting Tea was defining the sample space.
(Note we defined illogical guesses to be prob$= 0$)

A Running Visual Metaphor

Imagine that we sample one apple from a bag.
Looking in the bag we see:



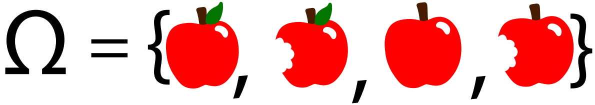
The sample space is:



Events

Events are subsets of the sample space.

For example, if



then

and

are both **events.**

Events Are a Kind of Set

Sets are collections of things, in this case collections of outcomes

One way to define an event is to describe the common property that all of the outcomes share. We write this as

$$\{\omega | \omega \text{ satisfies Property they share}\},$$

Example:

If $A = \{\omega | \omega \text{ has a leaf}\}$:

 $\in A$,  $\in A$,  $\notin A$,  $\notin A$

Complement

A complement of event A, denoted A^c, is also a set.

A^c, is everything else not in A.

and

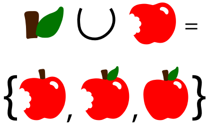
are **complements.**

$A^c = \{\omega \in \Omega | \omega \notin A\}$.

Important complement: $\Omega^c = \emptyset$, where $\emptyset$ is the empty set.

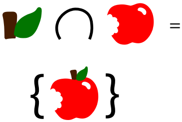
Unions and Intersections

The union of two events, A and B is the event that A or B occurs:



$$A \cup B = \{\omega | \omega \in A \text{ or } \omega \in B\}.$$

The intersection of two events, A and B is the event that both A and B occur:



$$A \cap B = \{\omega | \omega \in A \text{ and } \omega \in B\}.$$

Operations on Events

We say that two events A and B are disjoint or mutually exclusive if they don't share any elements or that $A \cap B = \emptyset$.

An event and its complement A and A^c are by definition disjoint.

Sample spaces can have infinite events where we will often write the different events using subscripts of the same letter: $A_1, A_2, \ldots A_\infty$ (e.g. imagine an event that was the count of some object)

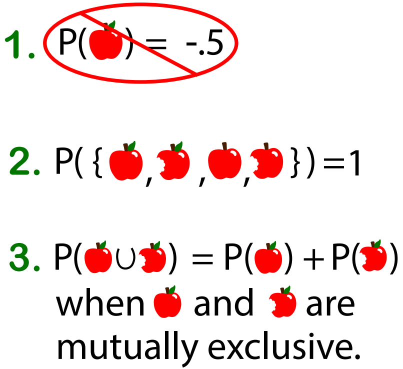
Probability Function

A probability function $P(\cdot)$ is a function defined over all subsets of a sample space Ω that satisfies the following three axioms:

1) $P(A) \geq 0$ for all A in the set of all events. nonnegativity

2) $P(\Omega) = 1$ normalization

3) if events $A_1, A_2, \ldots$ are mutually exclusive then $P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$. additivity



All the rules of probability can be derived from these axioms. (See Blitzstein & Hwang, Def 1.6.1.)

Three Big Ideas

Marginal, joint, and conditional probabilities

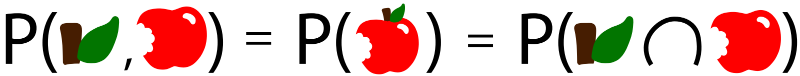
Bayes' rule

Independence

Marginal and Joint Probability

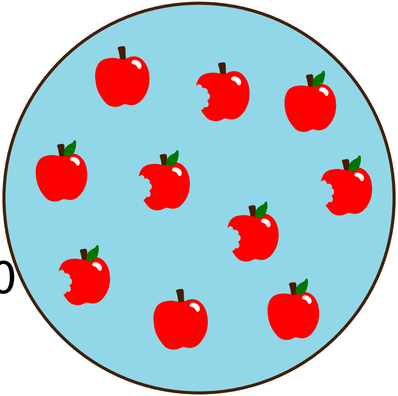
So far we have only considered situations where we are interested in the probability of a single event A occurring. We've denoted this $P(A)$. $P(A)$ is sometimes called a marginal probability.

Suppose we are now in a situation where we would like to express the probability that an event A and an event B occur. This quantity is written as $P(A \cap B)$, $P(B \cap A)$, $P(A, B)$, or $P(B, A)$ and is the joint probability of A and B.



P() = 7/10

P(, ) = 4/10



Conditional Probability

The "soul of statistics"
If $P(A) > 0$ then the probability of B conditional on A is

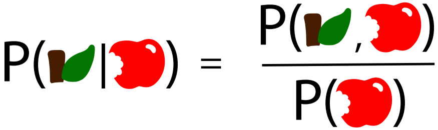
$$P(B|A) = \frac{P(A, B)}{P(A)}$$

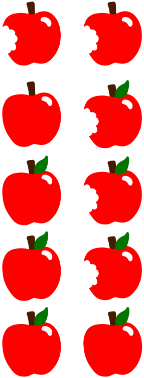
This implies that

$$P(A, B) = P(A)P(B|A) = P(B)P(A|B)$$

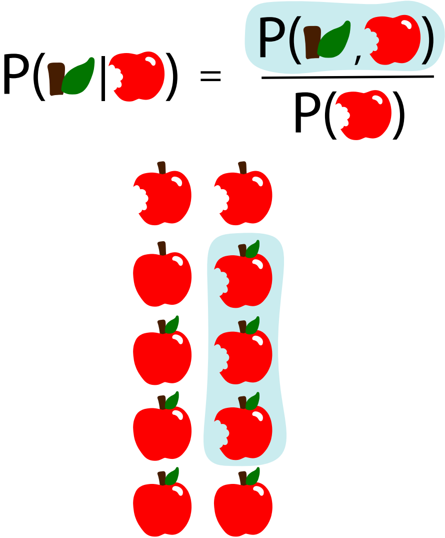
Hopefully this second formulation is intuitive!

Conditional Probability: A Visual Example

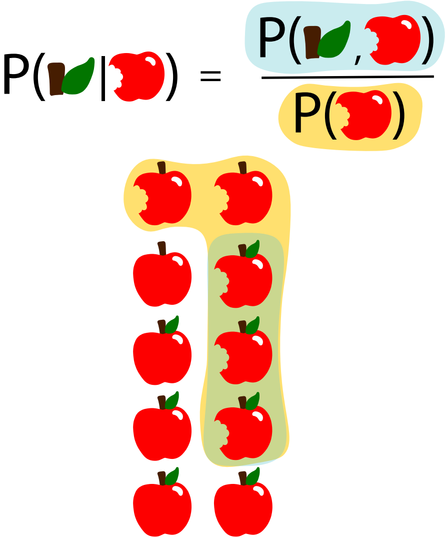




Conditional Probability: A Visual Example



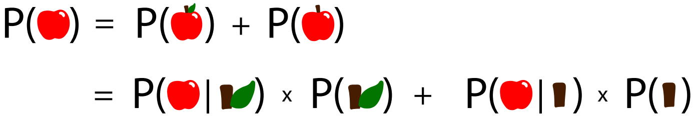
Conditional Probability: A Visual Example



Law of Total Probability (LTP)

With 2 Events:

$$\begin{aligned} P(B) &= P(B, A) + P(B, A^c) \\ &= P(B|A)P(A) + P(B|A^c)P(A^c) \end{aligned}$$



In general, if $\{A_i : i = 1, 2, 3, \dots\}$ forms a partition of the sample space, then

$$\begin{aligned} P(B) &= \sum_i P(B, A_i) \\ &= \sum_i P(B|A_i)P(A_i) \end{aligned}$$

Example: Voter Mobilization

Suppose that we want to know what the probability of voting is $P(\text{vote})$ after a campaign which mobilizes (contacts) some voters. We know the following:

- $P(\text{vote}|\text{mobilized}) = 0.75$
- $P(\text{vote}|\text{not mobilized}) = 0.15$
- $P(\text{mobilized}) = 0.6$ and so $P(\text{not mobilized}) = 0.4$

Note that mobilization partitions the data. Everyone is either mobilized or not. Thus, we can apply the LTP:

$$\begin{aligned} P(\text{vote}) =& P(\text{vote}|\text{mobilized})P(\text{mobilized}) + \\ & P(\text{vote}|\text{not mobilized})P(\text{not mobilized}) \\ =& 0.75 \times 0.6 + 0.15 \times 0.4 \\ =& .51 \end{aligned}$$

Engaged and Disengaged Respondents

PNAS RESEARCH ARTICLE POLITICAL SCIENCES OPEN ACCESS

Current research overstates American support for political violence

[illegible]

Edited by [illegible], University of Chicago, Chicago, IL; received September [illegible]; accepted February [illegible] by Editorial Board Member [illegible]

Political scientists, pundits, and citizens worry that America is entering a new period of violent partisan conflict. Provocative survey data show that a large share of Americans (between 8% and 40%) support politically motivated violence. Yet, despite media attention, political violence is rare, amounting to a little more than 1% of violent hate crimes in the United States. We reconcile these seemingly conflicting facts with four large survey experiments ($n = 4,904$), demonstrating that self-reported attitudes on political violence are biased upward because of respondent disengagement and survey questions that allow multiple interpretations of political violence. Addressing question wording and respondent disengagement, we find that the median of existing estimates of support for partisan violence is nearly 6 times larger than the median of our estimates (18.5% versus 2.9%). Critically, we show the prior estimates overstate support for political violence because of random responding by disengaged respondents. Respondent disengagement also inflates the relationship between support for violence and previously identified correlates by a factor of 4. Partial identification bounds imply that, under generous assumptions, support for violence among engaged and disengaged respondents is, at most, 6.86%. Finally, nearly all respondents support criminally charging suspects who commit acts of political violence. These findings suggest that, although recent acts of political violence dominate the news, they do not portend a new era of violent conflict.

political violence | affective polarization | democratic norms

Significance

Recent political events show that members of extreme political groups support partisan violence, and survey evidence supposedly shows widespread public support. We show, however, that after accounting for survey-based measurement error, support for partisan violence is far more limited. Prior estimates overstate support for political violence because of random responding by disengaged respondents and because of a reliance on hypothetical questions about violence in general instead of questions on specific acts of political violence. These same issues also cause the magnitude of the relationship between previously identified correlates and partisan violence to be overstated. As policy makers consider interventions designed to dampen support for violence, our results provide critical information about the magnitude of the problem.

Provocative recent work (1–3)—cited in PNAS (4, 5), *The American Journal of Political Science* (6), 60 other articles and books, and 40 news articles that, together, have garnered over 2,281,193 Twitter engagements—asserts that large segments of the American population now support politically motivated violence. These studies report that up to 44% of Americans would endorse hypothetical violence in some undetermined future event (1–3, 7). This survey work fits within a media landscape that regularly raises the specter of political violence. Since 2016, we counted 7,863 mentions of political violence on news television, more than 630 news stories about political violence, and over 10 million tweets on the topic of the January 6 riot alone (see *SI Appendix*, section 1 for details for all counts in this paragraph). Political violence, however, remains exceedingly rare in the United States, amounting to 48 incidents (8) in 2019 (the most recent year for which data are available) compared to 1,526 incidents of nonpolitical violent hate crimes (9) and 1,203,808 total violent crimes (10) documented by the Department of Justice.

Three Big Ideas

Marginal, joint, and conditional probabilities

Bayes' rule

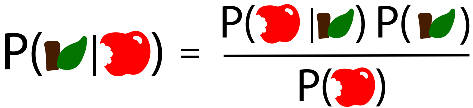
Independence

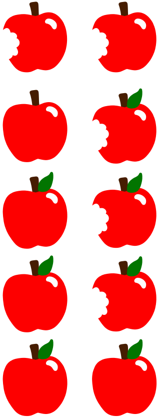
Bayes' Rule

- Often we have information about $P(B|A)$, but want $P(A|B)$.
- When this happens, always think: Bayes' rule
- Bayes' rule: if $P(B) > 0$, then:

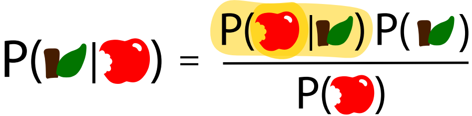
$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

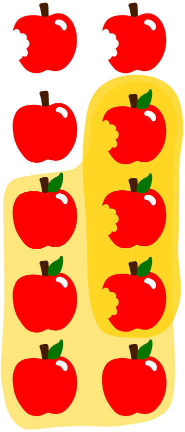
Bayes' Rule Mechanics



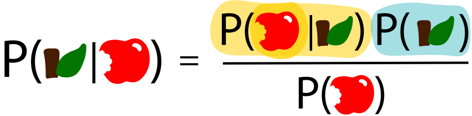


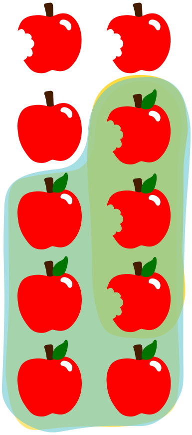
Bayes' Rule Mechanics



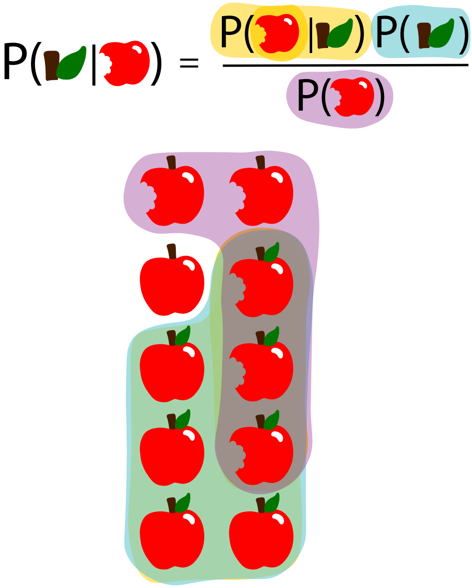


Bayes' Rule Mechanics





Bayes' Rule Mechanics



Example: Race and Names

What the Demolition of Public Housing Teaches Us about the Impact of Racial Threat on Political Behavior

Ryan D. Enos Harvard University

How does the context in which a person lives affect his or her political behavior? I exploit an event in which demographic context was exogenously changed, leading to a significant change in voters' behavior and demonstrating that voters react strongly to changes in an outgroup population. Between 2000 and 2004, the reconstruction of public housing in Chicago caused the displacement of over 25,000 African Americans, many of whom had previously lived in close proximity to white voters. After the removal of their African American neighbors, the white voters' turnout dropped by over 10 percentage points. Consistent with psychological theories of racial threat, their change in behavior was a function of the size and proximity of the outgroup population. Proximity was also related to increased voting for conservative candidates. These findings strongly suggest that racial threat occurs because of attitude change rather than selection.

Example: Race and Names

- Note that the Census collects information on the distribution of names by race.
- For example, Washington is the most common last name among African-Americans in America:
  - $P(\text{AfAm}) = 0.132$
  - $P(\text{not AfAm}) = 1 - P(\text{AfAm}) = .868$
  - $P(\text{Washington}|\text{AfAm}) = 0.00378$
  - $P(\text{Washington}|\text{not AfAm}) = 0.000061$
- We can now use Bayes' Rule

$$P(\text{AfAm}|\text{Wash}) = \frac{P(\text{Wash}|\text{AfAm})P(\text{AfAm})}{P(\text{Wash})}$$

Example: Race and Names

Note we don't have the probability of the name Washington. Remember that we can calculate it from the LTP since the sets African-American and not African-American partition the sample space:

$$
\begin{aligned}
P(\text{AfAm}|\text{Wash}) &= \frac{P(\text{Wash}|\text{AfAm})P(\text{AfAm})}{P(\text{Wash})} \\
&= \frac{P(\text{Wash}|\text{AfAm})P(\text{AfAm})}{P(\text{Wash}|\text{AfAm})P(\text{AfAm}) + P(\text{Wash}|\text{not AfAm})P(\text{not AfAm})} \\
&= \frac{0.132 \times 0.00378}{0.132 \times 0.00378 + .868 \times 0.000061} \\
&\approx 0.9
\end{aligned}
$$

Three Big Ideas

Marginal, joint, and conditional probabilities

Bayes' rule

Independence

Independence

Intuitive Definition

Events A and B are independent if knowing whether A occurred provides no information about whether B occurred.
i.e. a formalization of what it means for two events to be unrelated.

Formal Definition

$$P(A,B) = P(A)P(B) \implies A \perp\!\!\!\perp B$$

With all the usual > 0 restrictions, this implies

- $P(A|B) = P(A)$
- $P(B|A) = P(B)$

Conditional Independence

$$P(A,B|C) = P(A|C)P(B|C) \implies A \perp\!\!\!\perp B|C$$

Independence is a massively important concept in statistics.

Independence, the Heroic Assumption

Deploy with Caution!

But Really, Why Probability?

- Envision hypothetical data given some assumed by intrinsicly uncertain system.
- Set us up to do inference (i.e. run that backwards!).
- A mathematical model that hopefully corresponds well to real behavior $\rightarrow$ sometimes this correspondence is direct and other times it is more like a loose analogy.
- It can feel frustrating to start with probability if you came to class wanting to analyze data, but it is the single most important piece of infrastructure.

This Week in Review

- Course logistics
- Core ideas in statistics
- Foundations of probability
- Three big probability concepts

Going Deeper:

Blitzstein, Joseph K., and Hwang, Jessica. (2019). *Introduction to Probability: Second Edition*. CRC Press. `http://stat110.net/`

Next week: random variables!

i.e. how we link more complex real-world sample spaces to numerical recipes that we can reuse.

The New York Times

Comey and McCabe, Who Infuriated Trump, Both Faced Intensive I.R.S. Audits

The former F.B.I. director and his deputy, both of whom former President Donald J. Trump wanted prosecuted, were selected for a rare audit program that the tax agency says is random.

"Maybe it's a coincidence or maybe somebody misused the I.R.S. to get at a political enemy," said James B. Comey, the former F.B.I. director. "Given the role Trump wants to continue to play in our country, we should know the answer to that question." Jared Soares for The New York Times

Example: Ballot Order

Evidence suggests that candidates gain a small advantage from ballot order.

As a response, in 2008 New Hampshire chose a letter from the alphabet and then listed the candidates in alphabetical order starting with that letter.

We can use probability to assess the “fairness” of this process.

We will do this by introducing a random variable X to be Barack Obama’s position on the 2008 New Hampshire primary ballot.

What is a Random Variable?

Intuition: functions that map outcomes to numbers.

Formal: X is a function that maps the sample space to the real numbers.

Imagine two coin flips

$$\Omega = \big\{\{heads, heads\}, \{heads, tails\}, \{tails, heads\}, \{tails, tails\}\big\}$$

we could define a random variable $X(\omega)$ to be the function that returns the number of heads for each element (ω) of the sample space (Ω).

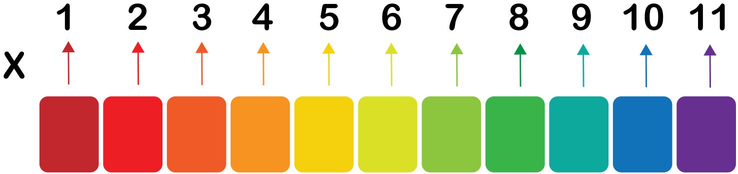
- $X(\{heads, heads\}) = 2$
- $X(\{heads, tails\}) = 1$
- $X(\{tails, heads\}) = 1$
- $X(\{tails, tails\}) = 0$

We will generally suppress the function notation and just refer to X.

A Visual Example

A Visual Example

A Visual Example



A Brief Note on Notation

- We almost always use capital roman letters for the "name" of the random variable such as X.
- We refer to a fixed value with a lower case letter x.
- So we might write $P(X = x)$ to be the probability that the number of heads we observe is equal to some fixed value x.
- We will sometimes write out the mapping from the sample space to the random variable. For example,

$$X = \begin{cases} 1 \text{ if heads} \\ 0 \text{ if tails} \end{cases}$$

- Other times the sample space is already numeric so its more obvious (e.g. how many minutes until the train arrives).

Quick FAQ

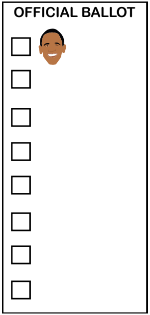
- Why have random variables at all?
  it makes the math easier, even across very different sample spaces.
- Why are they random variables?
  they are realizations of a stochastic process (i.e. randomness in the outcome, not the mapping).
- Is it really easier this way? It seems hard.
  random variables are about bridging the abstract math and the concrete world. that can be hard, but it is super important and better than the alternative!

NH Ballot Order Example

Candidates:

- Joe Biden
- Hillary Clinton
- Chris Dodd
- John Edwards
- Mike Gravel
- Dennis Kucinich
- Barack Obama
- Bill Richardson

$$X = \begin{cases} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \\ 8 \end{cases}$$



A,B,C,D,E,F,G,H,I,J,K,L,M,N,O,P,Q,R,S,T,U,V,W,X,Y,Z

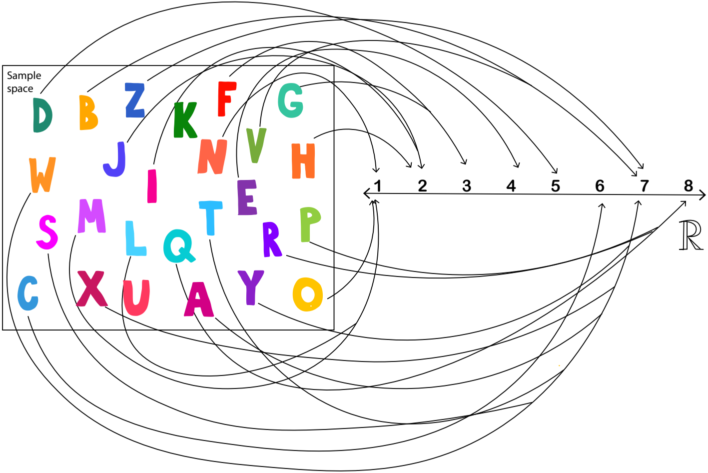
X is a random variable indicating Obama's position on the ballot. Highlighted letters are those leading to a given ballot position. Highlighted individual is first.

Discrete Distributions

- The distribution of a random variable specifies the probability of all events associated with that random variable.
- For discrete distributions, the random variable X takes on a finite, or a countably infinite number of values.
- A common shorthand is to think of discrete random variables taking on distinct values.
- A probability mass function (PMF) and a cumulative distribution function (CDF) are two common ways to define the probability distribution for a discrete random variable.
- Probability mass functions provide a compact way to represent information about how likely various outcomes are.

Where do Distributions Come From?

The probabilities associated with each realization of the random variables come from the underlying stochastic realization of the sample space.

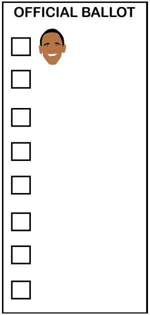


Example: New Hampshire

Candidates:

- Joe Biden
- Hillary Clinton
- Chris Dodd
- John Edwards
- Mike Gravel
- Dennis Kucinich
- Barack Obama
- Bill Richardson

$$p_X(x) = \begin{cases} 4/26 & x = 1 \\ 4/26 & x = 2 \\ 2/26 & x = 3 \\ 1/26 & x = 4 \\ 1/26 & x = 5 \\ 1/26 & x = 6 \\ 10/26 & x = 7 \\ 3/26 & x = 8 \end{cases}$$



A,B,C,D,E,F,G,H,I,J,K,L,M,N,O,P,Q,R,S,T,U,V,W,X,Y,Z

Probability of the random variable equaling a number is just the probability of the underlying event (subset of the sample space).

Discrete Probability Mass Functions

Definition (Probability Mass Function)

The probability mass function (PMF) of a discrete random variable X is the function p_X given by,

$$p_X(x) = P(X = x)$$

Understanding the Notation:

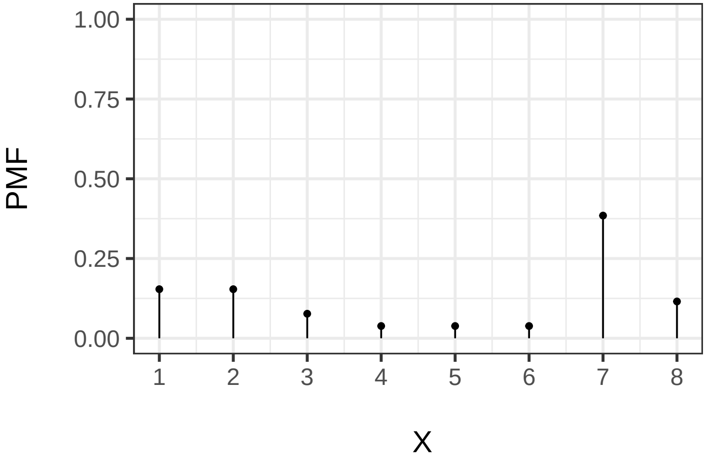
- $X = x$ is defining an event.

  More formally we might say, $\{X = x\}$ is shorthand for $\{\omega \in \Omega : X(\omega) = x\}$ which can be read as the set of realizations ω in the sample space Ω such that the function $X(\omega)$ returns the fixed value x.
- Later we will drop the subscript when it is clear from context.

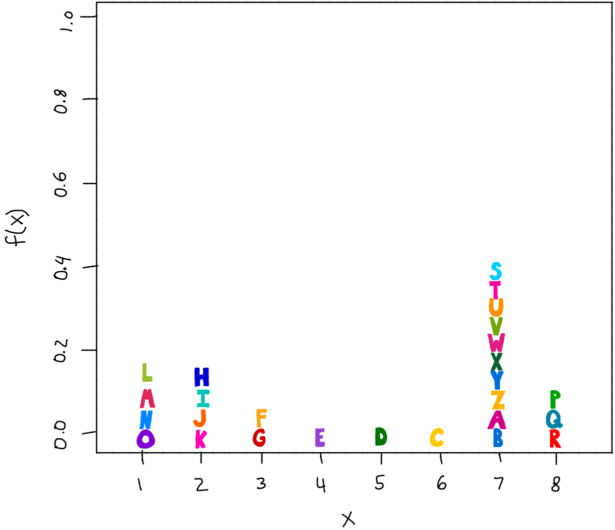
Three key properties:

- this will always be non-negative.
- the support of X is the set of values where the PMF is non-zero.
- $\sum_x p_X(x) = 1$.

NH Obama Ballot Position PMF Plot



NH Obama Ballot Position PMF Plot



Cumulative Distribution Function

Definition (Cumulative Distribution Function)

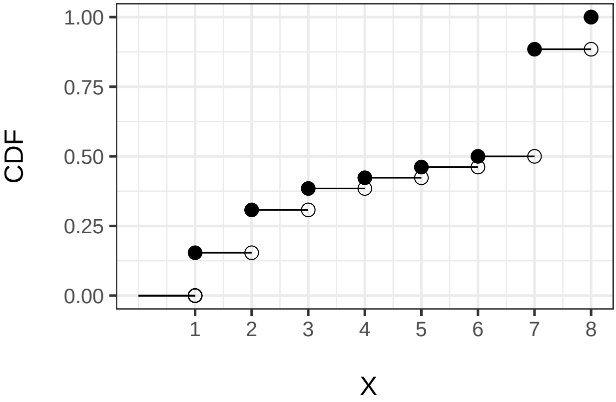
The cumulative distribution function (CDF) of a random variable X is the function F_X given by,

$$F_X(x) = P(X \leq x)$$

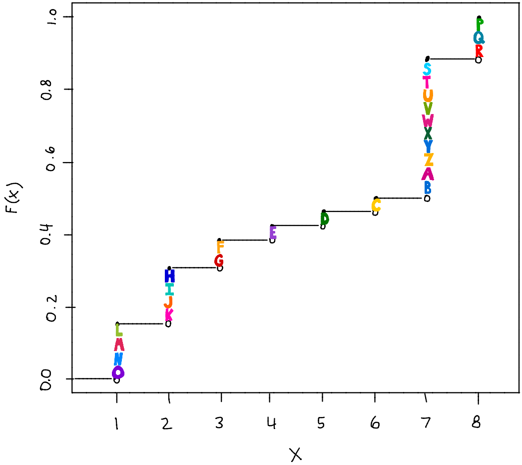
Key properties:

- non-decreasing
- right-continuous
- converges to 0 and 1 in the limits

NH Obama Ballot Position CDF Plot



NH Obama Ballot Position CDF Plot



Example Discrete Distributions

- A major advantage of random variables is that they often have a distribution with a known form (that comes with known results!)
  - Bernoulli distribution: Let X be a binary variable with $P(X=1)=\pi$ and, thus, $P(X=0)=1-\pi$, where $\pi \in [0,1]$. It has PMF:

$$p_X(x) = \pi^x (1-\pi)^{1-x} \qquad \text{for } x \in \{0,1\}.$$

  - Discrete Uniform distribution: Let X be a random variable that puts equal probability on each value that X can take:

$$p_X(x) = \begin{cases} 1/k & \text{for } x = 1, \dots, k \\ 0 & \text{otherwise} \end{cases}$$

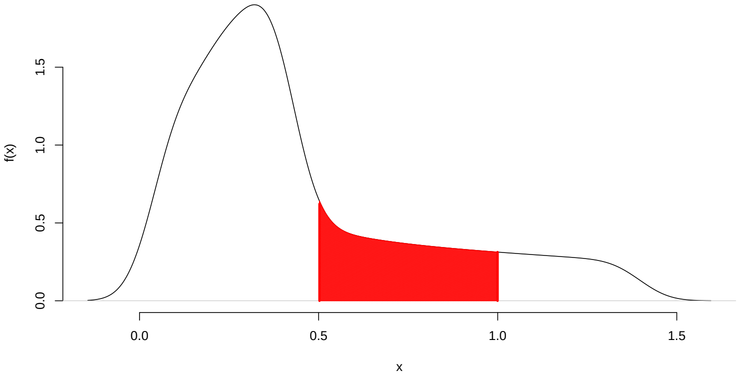
- We can summarize these distributions with one number

Continuous Distributions

- Continuous random variables take on an uncountably infinite number of values.
- This is often a useful approximation when variables take many values.
- A probability density function (PDF) and a cumulative distribution function (CDF) are two common ways to define the distribution for a continuous random variable.
- They are similar to the discrete case with a few subtle differences.

Calculus Review: Integration

Suppose we have some function $f(x)$



What is the area under $f(x)$ between $\frac{1}{2}$ and 1?

Area under curve $= \int_{1/2}^{1} f(x)dx = F(1) - F(1/2)$

Continuous Random Variable

A continuous random variable has a continuous cumulative distribution function (CDF) which, as in the discrete case, defines the probability that $P(X \leq x)$.

Definition (Continuous Distribution)

A random variable has a continuous distribution if its CDF is differentiable. We also allow there to be endpoints (or finitely many points) where the CDF is continuous but not differentiable, as long as the CDF is differentiable everywhere else. (Blizstein and Hwang Definition 5.1.1)

Probability Density Function (PDF)

The probability density function is the analog of the probability mass function for discrete random variables.

Definition (Probability density function)

For a continuous random variable X with CDF F_X, the probability density function of X is the derivative f of the CDF, given by $f_X(x) = \frac{d}{dx}F_X(x)$

Key Properties:

- non-negative
- integrates to 1. $\int_{-\infty}^{\infty} f_X(x)dx = 1$
- for any measurable set of real numbers B,

$$P(X \in B) = \int_B f_X(x)dx$$

Defining the CDF in terms of the PDF

Definition (CDF of a Continuous Random Variable)

For a continuous random variable X define its cumulative distribution function $F_X(x)$ as,

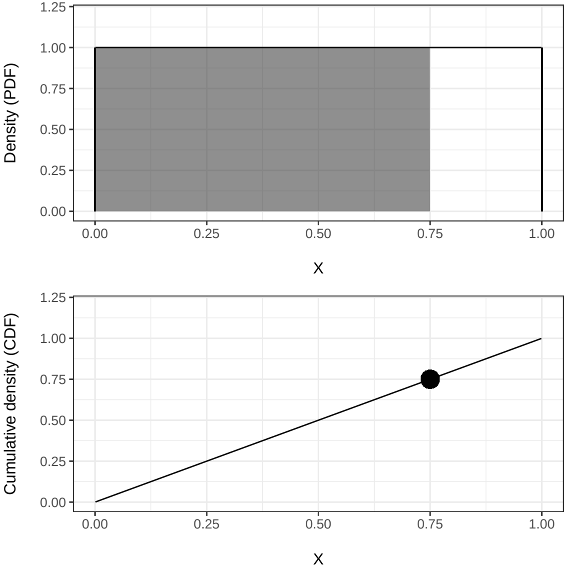
$$F_X(t) = P(X \leq t) = \int_{-\infty}^{t} f_X(x)dx$$

A Visual Example

Imagine you choose a number completely at random between 0 and 1 with all equally sized sets of values being equally likely. This is a standard uniform distribution which has the CDF,

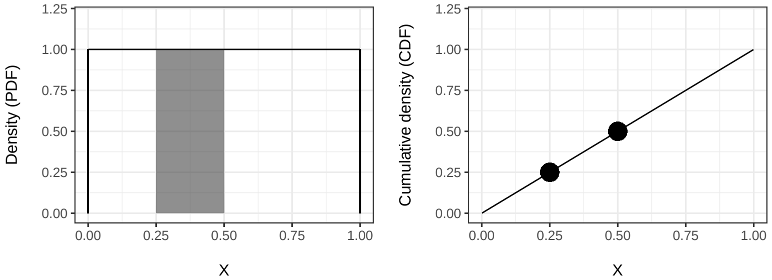
$$F_X(x) = x$$

with support over $[0, 1]$.



Density (PDF)
1.25
1.00
0.75
0.50
0.25
0.00
0.00
0.25
0.50
0.75
1.00
X
Cumulative density (CDF)
1.25
1.00
0.75
0.50
0.25
0.00
0.00
0.25
0.50
0.75
1.00
X

What is the probability that the number is between 0.25 and 0.5?



$$F_X(.5) - F_X(.25) = .25$$

The Core PMF/PDF Difference

The probability mass function provides the probability of a set of outcomes by summing over the probability mass function evaluated at each of those outcomes.

The probability density function for continuous variables provides the probability of a set of outcomes by integrating over a range of values.

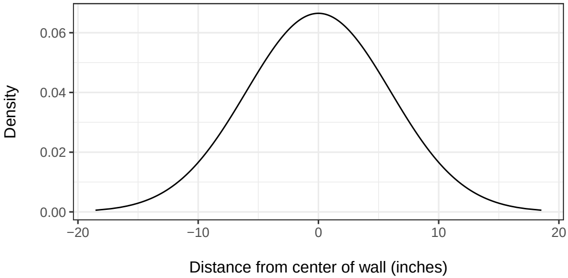
This means (perhaps counterintuitively) that a probability density function:

- can return a value greater than 1
- assigns the probability of any exact value is zero.

Let's explain!

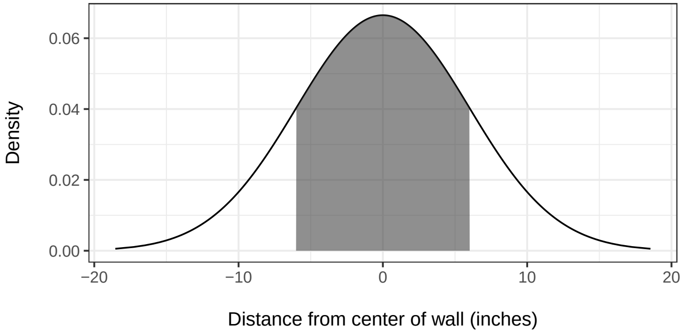
A Numerical Example

Let's suppose we have someone throwing darts and we measure how far they are from the center of the wall in inches. In this case, perhaps the darts will be distributed with the following PDF.



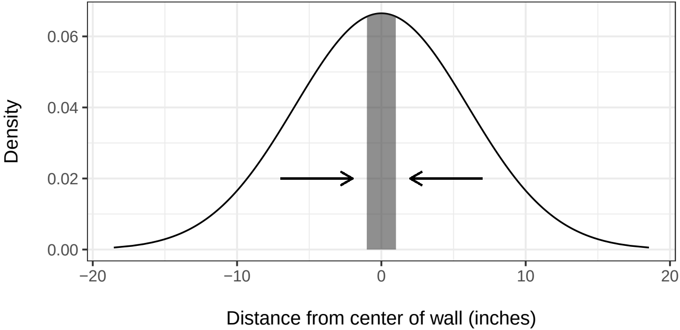
A Numerical Example

How would we calculate the probability that a dart lands within 6 inches of the center of the wall?



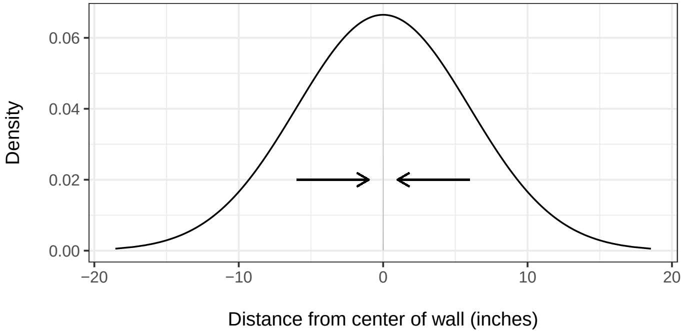
$$\begin{aligned} P(X \in (-6, 6)) &= \int_{-6}^{6} f_X(x)dx \\ &= F_X(6) - F_X(-6) \\ &= P(X < 6) - P(X < -6) \\ &= 0.683 \end{aligned}$$

One inch?



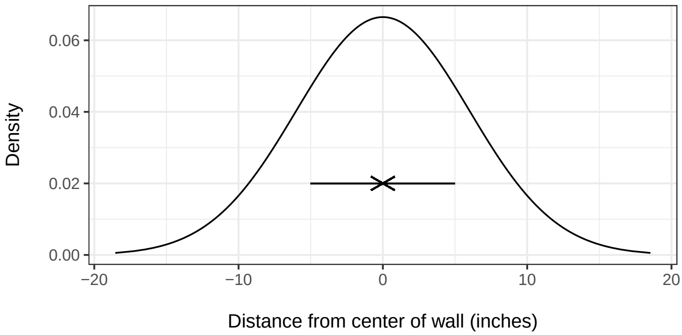
$$P(X \in (-1, 1)) = 0.0664135$$

1/100th of an inch?



$$P(X \in (-.01, .01)) = 0.0006649037$$

A perfect bullseye?



$$P(X = 0) = 0$$

The probability that a continuous variable takes on a discrete value is 0!
Why?
Because the width of the range we are calculating is zero, the area is zero.

The Practical Upshot

- The cumulative distribution function (CDF) has the same interpretation between discrete and continuous.
- The probability mass function (PMF) is for discrete variables and returns a probability.
- The probability density function (PDF) is for continuous variables and provides a probability when integrated over a subset of the support.
- Reconciling the continuous/discrete divide is the purview of measure theory which is a layer deeper than we are going to go in this class.
- As with discrete random variables there are common families of distributions.

References

- Enos, Ryan D. "What the demolition of public housing teaches us about the impact of racial threat on political behavior." *American Journal of Political Science* (2015).
- Salsburg, David. *The Lady Tasting Tea: How Statistics Revolutionized Science in the Twentieth Century* (2002).