

# How do we represent data?

Jeffrey Leek, Assistant Professor of Biostatistics  
Johns Hopkins Bloomberg School of Public Health

# How do we write about data?

- Each data point is usually represented by a capital letter.
  - $H$  for height,  $W$  for weight.
- If there are more than one data point of the same type we use subscripts.
  - $H_1, H_2, H_3$  for three different people's heights.
- Sometimes it is more compact to write  $X_1$  for height and  $X_2$  for weight.
- Then we need another subscript for the individual data point
  - $X_{11}$  for the height of the first person.
- $Y$  represents general outcomes and  $X$  general covariates.
- In this course we will try to use informative letters when possible.

# Randomness

- Variables like  $X$  and  $Y$  are called *random variables* because we expect them to be *random* in some way.
- In general, randomness is a hard thing to define
- In this class a variable may be random because
  - It represents an incompletely measured variable
  - It represents a sample drawn from a population using a random mechanism.
- Once we are talking about a specific value of a variable we have observed it isn't random anymore, we write these values with lower case letters  $x, y$ , etc.
- We write  $X = x$  or  $X = 1$  to indicate we have observed a specific value  $x$  or 1.

# Randomness and measurement

- A coin flip is commonly considered random
- But it can be modeled by deterministic equations
  - Dynamical bias in the coin toss ([Diaconis, Holmes and Montgomery SIAM Review 2007](#))
  - Modeled the tossing as a dynamical system
  - Showed that a coin is more likely to land on the side it started
  - Did experiments that demonstrated it was a 51% chance
- Some have taken it a bit farther making [predictable coin flipping machines](#) based on [physical properties](#).

# Distributions

- In statistical modeling, random variables like  $X$  are assumed to be samples from a *distribution*
- A distribution tells us the possible values of  $X$  and the probabilities for each value.
- Probability is the chance something will happen and is abbreviated  $Pr$
- The probabilities must all be between 0 and 1.
- The probabilities must add up to 1.
- An example:
  - Let's flip a coin and allow  $X$  to represent whether it is heads or tails
  - $X = 1$  if it is heads and  $X = 0$  if it is tails
  - We expect that about 50% of the time it will be heads.
  - The distribution can then be written  $Pr(X = 1)=0.5$  and  $Pr(X = 0)=0.5$

# Continuous versus discrete distributions

- *discrete* distributions specify probabilities for discrete values
  - Qualitative variables are discrete
  - So are variables that take on all values 0,1,2,3...
- *continuous* distributions specify probabilities for ranges of values
  - Quantitative variables are often assumed to be continuous
  - But we might only see specific values

# Parameters

- Distributions are defined by a set of fixed values called *parameters*.
- *parameters* are sometimes represented by Greek letters like  $\mu$ ,  $\sigma$ ,  $\tau$ .
- Distributions are written as letters with the parameters in parentheses like  $N(\mu, \sigma)$  or  $Poisson(\lambda)$ .
- $X \sim N(\mu, \sigma)$  means that  $X$  has the  $N(\mu, \sigma)$  distribution.

# The three most important parameters

- If  $X$  is a random variable, the mean of that random variable is written  $E[X]$ 
  - Stands for expected value
  - Measures the "center" of a distribution
- The variance of that random variable is written  $Var[X]$ 
  - Measures how "spread out" a distribution is
  - Measurement is in  $(\text{units of } X)^2$
- The standard deviation is written  $SD[X] = \sqrt{Var[X]}$ 
  - Also measures how "spread out" a distribution is
  - Measurement is in units of  $X$



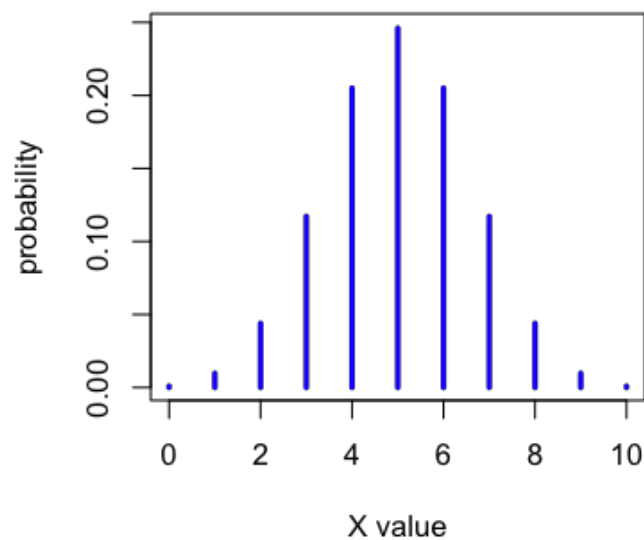
# Conditioning

- The variables  $X$  are considered to be random
- The parameters are considered to be fixed values
- Sometimes we want to talk about a case where one of the random variables is fixed
- To indicate what is fixed, we *condition* using the symbol " $|$ "
  - $X|\mu$  means that  $X$  is a random variable with fixed parameter  $\mu$
  - $Y|X = 2$  means  $Y$  is the random variable  $Y$  when  $X$  is fixed at 2.

# Example distribution: Binomial

Binomial distribution:  $Bin(n, p)$

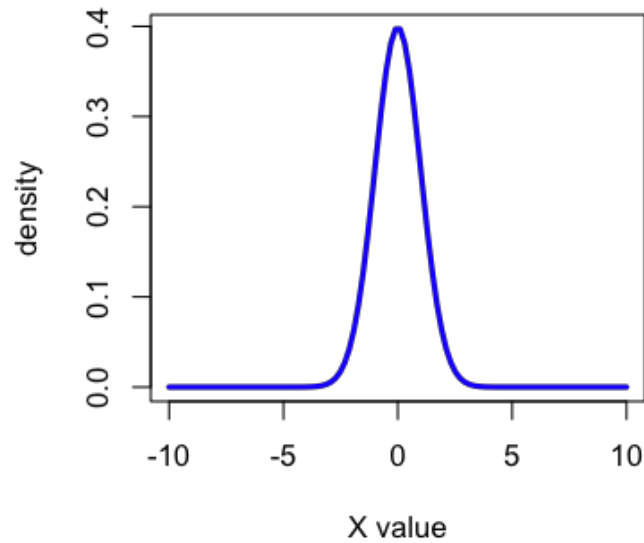
- $X \sim Bin(10, 0.5)$



# Example distribution: Normal

Normal Distribution:  $N(\mu, \sigma)$

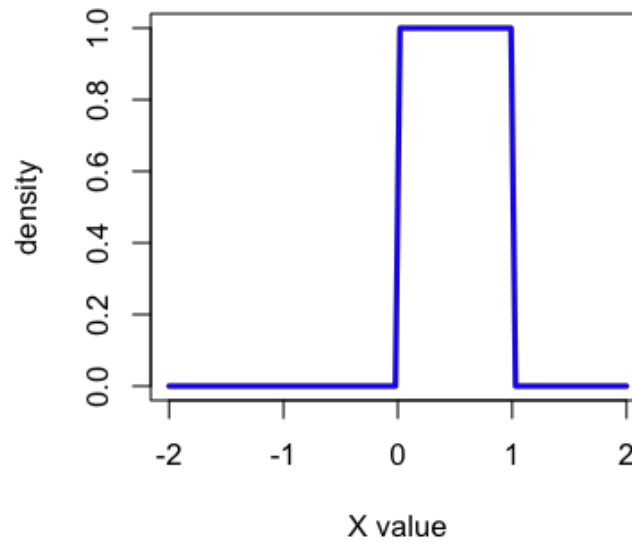
- $X \sim N(0, 1)$



# Example distribution: Uniform

Uniform distribution:  $U(\alpha, \beta)$

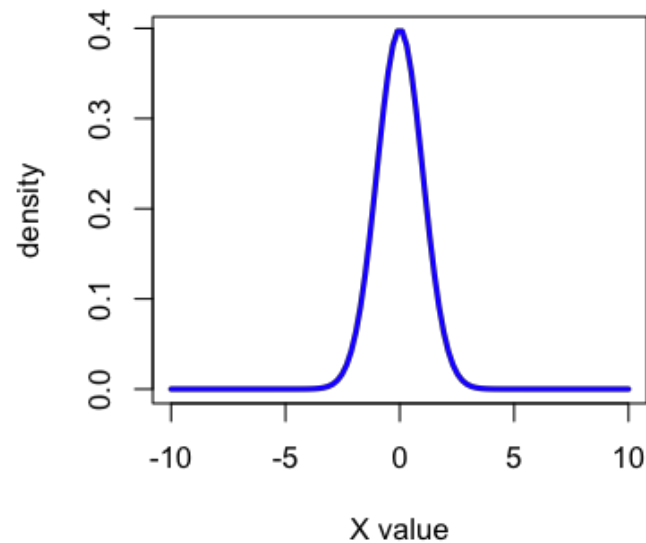
- $X \sim U(0, 1)$



# Changing parameters

**Normal Distribution:**  $N(\mu, \sigma)$

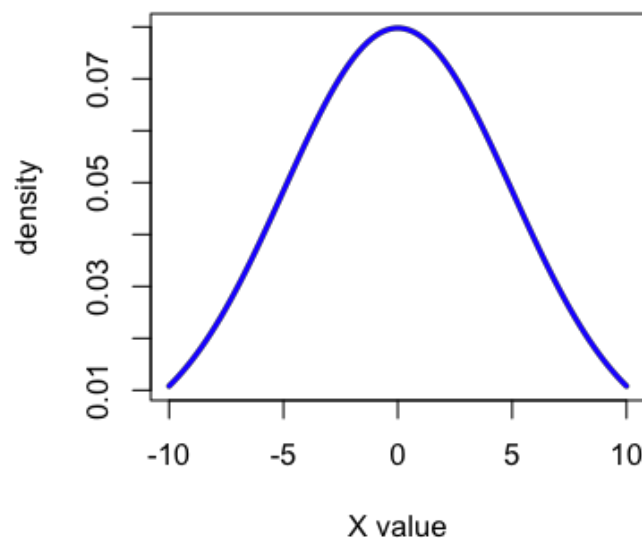
- $X \sim N(0, 1), E[X] = \mu = 0, Var[X] = \sigma^2 = 1$



# Changing parameters: the variance

**Normal Distribution:**  $N(\mu, \sigma)$

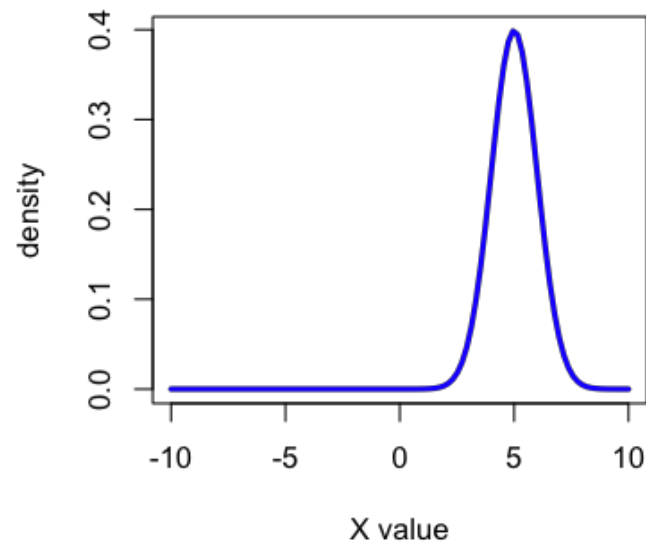
- $X \sim N(0, 5), E[X] = \mu = 0, Var[X] = \sigma^2 = 25$



# Changing parameters: the mean

**Normal Distribution:**  $N(\mu, \sigma)$

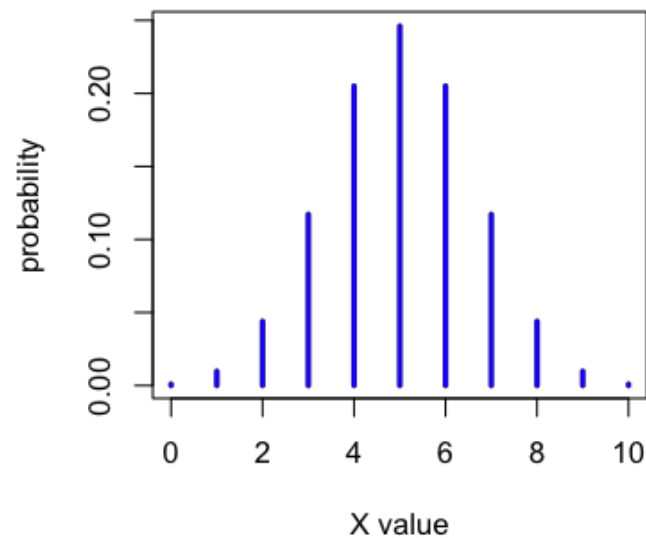
- $X \sim N(5, 1), E[X] = \mu = 5, Var[X] = \sigma^2 = 1$



# Example distribution: Binomial

**Binomial distribution:**  $Bin(n, p)$

- $X \sim Bin(10, 0.5)$ ,  $E[X] = n \times p = 5$ ,  $Var[X] = n \times p \times (1 - p) = 2.5$

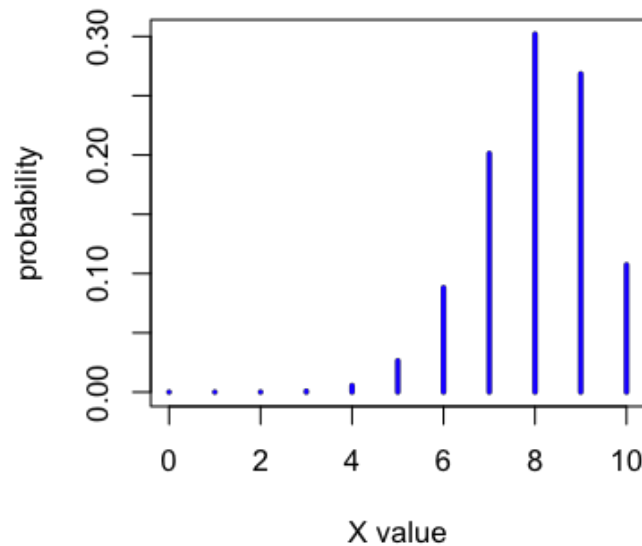




# Changing parameters: both mean and variance

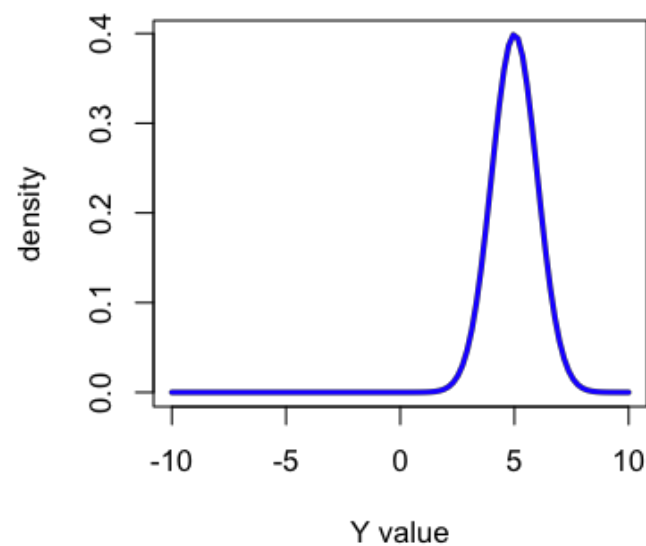
**Binomial distribution:**  $Bin(n, p)$

- $X \sim Bin(10, 0.8)$ ,  $E[X] = n \times p = 8$ ,  $Var[X] = n \times p \times (1 - p) = 1.6$



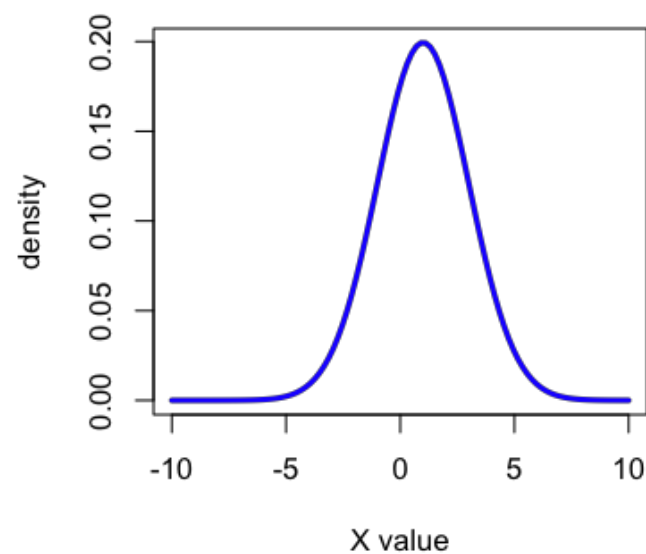
# Conditioning

- Suppose  $Y \sim N(X, 1)$  and  $X \sim N(0, 1)$ , then the distribution of  $Y|X = 5$  is



# Conditioning

- Suppose  $Y \sim N(X, 1)$  and  $X \sim N(0, 1)$ , then the distribution of  $Y$  is



[http://en.wikipedia.org/wiki/Law\\_of\\_total\\_variance](http://en.wikipedia.org/wiki/Law_of_total_variance)

[http://en.wikipedia.org/wiki/Law\\_of\\_total\\_expectation](http://en.wikipedia.org/wiki/Law_of_total_expectation)

# Learning more about a specific distribution

The screenshot shows the Wikipedia page for the Poisson distribution. The page title is "Poisson distribution" and it is part of the "en.wikipedia.org/wiki/Poisson\_distribution" page. The page includes a sidebar with navigation links, a main content area with a definition and examples, and two plots on the right: a Probability mass function (PMF) plot and a Cumulative distribution function (CDF) plot.

**Wikipedia Page Structure:**

- Header:** "W Poisson distribution - Wikipe" and "en.wikipedia.org/wiki/Poisson\_distribution".
- Navigation:** "Article", "Talk", "Read", "Edit", "View history", "Search".
- Left Sidebar:**
  - WIKIPEDIA The Free Encyclopedia
  - Main page, Contents, Featured content, Current events, Random article, Donate to Wikipedia, Wikimedia Shop
  - Interaction: Help, About Wikipedia, Community portal, Recent changes, Contact Wikipedia
  - Toolbox, Print/export
  - Languages: العربية, Български, Català, Česky, Deutsch, Ελληνικά, Español, Euskara, فارسی, Français, 한국어, Bahasa Indonesia, Italiano, עברית, Lietuvių
- Main Content:**

**Poisson distribution**

From Wikipedia, the free encyclopedia

In **probability theory** and **statistics**, the **Poisson distribution** (pronounced [pwas5]) is a **discrete probability distribution** that expresses the probability of a given number of events occurring in a fixed interval of time and/or space if these events occur with a known average rate and **independently** of the time since the last event.<sup>[1]</sup> The Poisson distribution can also be used for the number of events in other specified intervals such as distance, area or volume.

For instance, suppose someone typically gets 4 pieces of mail per day on average. There will be, however, a certain spread: sometimes a little more, sometimes a little less, once in a while nothing at all.<sup>[2]</sup> Given only the average rate, for a certain period of observation (pieces of mail per day, phonecalls per hour, etc.), and assuming that the process, or mix of processes, that produce the event flow are essentially random, the Poisson distribution specifies how likely it is that the count will be 3, or 5, or 11, or any other number, during one period of observation. That is, it predicts the degree of spread around a known average rate of occurrence.<sup>[2]</sup>

The section: [Derivation of the Poisson distribution](#) shows the relation with the formal definition.

Historical background of the Poisson distribution has been described by Gullberg (1997).<sup>[3]</sup>

**Contents** [hide]

  - History
  - Definition
  - Properties
    - 3.1 Mean
    - 3.2 Median
    - 3.3 Higher moments
    - 3.4 Other properties
  - Related distributions
  - Occurrence
    - 5.1 Derivation of Poisson distribution — The law of rare events
    - 5.2 Multi-dimensional Poisson process
    - 5.3 Other applications in science
  - Generating Poisson-distributed random variables
  - Parameter estimation
    - 7.1 Maximum likelihood
- Right Side Plots:**

**Poisson**

**Probability mass function**

The horizontal axis is the index  $k$ , the number of occurrences. The function is only defined at integer values of  $k$ . The connecting lines are only guides for the eye.

**Cumulative distribution function**

The horizontal axis is the index  $k$ , the number of occurrences. The function is only defined at integer values of  $k$ . The connecting lines are only guides for the eye.

[http://en.wikipedia.org/wiki/Poisson\\_distribution](http://en.wikipedia.org/wiki/Poisson_distribution)

# Learning more about representing data

OpenIntro

SUBJECTS CONTRIBUTE RIGHTS BLOG LOGIN

Statistics

OVERVIEW TEXTBOOK SUPPLEMENTS LABS LINKS

**SECOND EDITION**

OpenIntro Statistics is a free textbook for introductory statistics. We've spent thousands of hours to make this textbook ready to compete on any stage. The book can be downloaded for free as a PDF or purchased on Amazon.com for \$9.94 (get 2-day shipping with a free student trial of Amazon Prime).

Download Second Edition

Click links to download individual chapters, appendices, or the textbook's source files.

FRONT MATERIAL OF BOOK

REVIEW OR SUBMIT TYPOS

CHAPTER 1	CHAPTER 6
CHAPTER 2	CHAPTER 7
CHAPTER 3	CHAPTER 8
CHAPTER 4	APPENDICES
CHAPTER 5	SOURCE

READ ONLINE REVIEWS

**FIRST EDITION**

For information about the differences between the First and Second Edition, please see [this blog post](#). The First Edition can be downloaded for free as a PDF or purchased on Amazon.com for \$9.02.

Click links to download individual chapters, appendices, or the textbook's source files.

FIRST EDITION, FULL TEXT

FRONT MATERIAL OF BOOK

REVIEW OR SUBMIT TYPOS

CHAPTER 1	CHAPTER 6
CHAPTER 2	CHAPTER 7
CHAPTER 3	CHAPTER 8
CHAPTER 4	APPENDICES
CHAPTER 5	SOURCE

READ ONLINE REVIEWS

**DATA SETS**

Data sets used in the textbook are included in the zipped file below. Each data set is saved as a tab-delimited text file so it can be easily loaded into any statistical software. These data sets can also be found in the OpenIntro R packages ([openintro](#), [Oldata](#)).

Data Sets

**PROBABILITY TABLES**

Normal, t, and chi-square probability distribution tables in a printer-friendly PDF. Or if you'd rather customize the tables, download the LaTeX source.

Printer-friendly PDF

Download the LaTeX source

<http://www.openintro.org/stat/textbook.php>