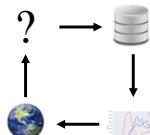


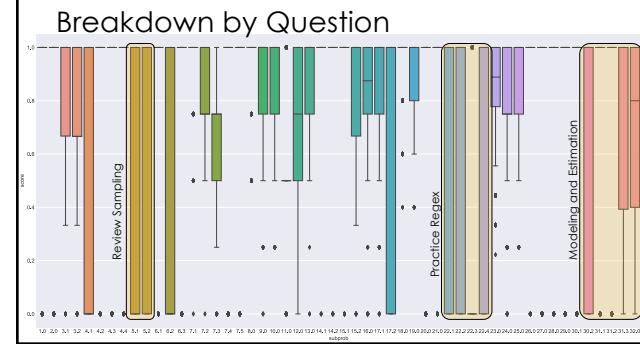
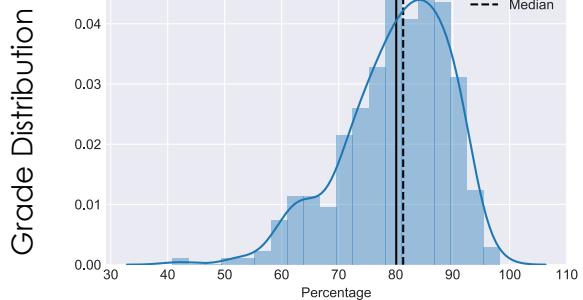
Data Science 100

Probability and Generalization

Slides by:
Joseph E. Gonzalez & Deb Nolan,
jegonzal@berkeley.edu
deborah.nolan@berkeley.edu



How was the Midterm?

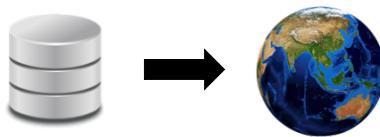


Recap: Modeling and Estimation

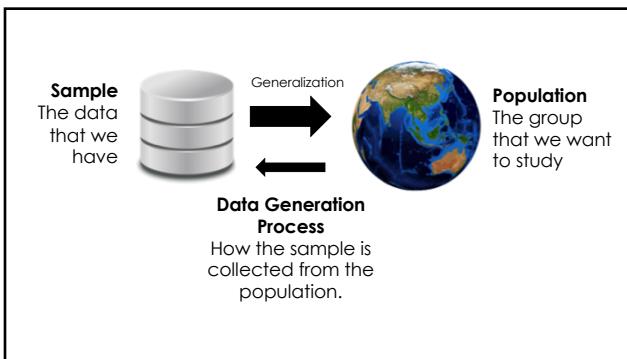
1. **Define the Model:** simplified representation of the world
2. **Define the Loss Function:** measures how well a particular instance of the model "fits" the data
3. **Minimize the Loss Function:** find the parameter values that minimize the loss **on the data**

What does a model that fits the data have to do with **the world?**

Generalization



The focus of the next few lectures.

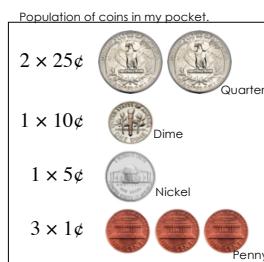


What we will do:

1. Examine a Population
2. Study a data generation process
 - a. **Simulation** for insight
 - b. **Theory** for proof
3. Draw conclusions from a sample
 - a. Theory to connect to population
 - b. Bootstrap to go beyond theory

Review Probability Concepts

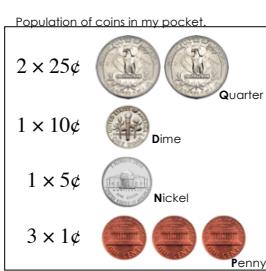
Toy scenario



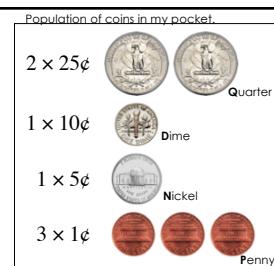
- Population Size = 7
- Total value of the population $2 \times 25 + 1 \times 10 + 1 \times 5 + 3 \times 1 = 68$
- Average coin value: $\frac{68}{7} \approx 9.71$
- Median coin value:

Random Sample of Size 1

- Randomly sample a single coin
- Let X be the **value (in cents)**
 - Takes on values: 1, 5, 10, and 25
 - X is a random variable
- **Random variable:** a variable whose value is determined by a chance event.
- Chance event
 - The kind of coin I draw: P, N, D, Q



- Randomly sample a single coin
- Let X be the **value (in cents)**
 - Takes on values: 1, 5, 10, and 25
 - X is a random variable
- **Random variable:** a variable whose value is determined by a chance event.
- Chance event
 - The kind of coin I draw: P, N, D, Q



Probability Distribution

Penny	Nickel	Dime	Quarter
3/7	1/7	1/7	2/7

What is the **expected value** of X ?

The Expected Value

$$\mathbb{E}[X] = \sum_{x \in \mathcal{X}} x \mathbf{P}(x)$$

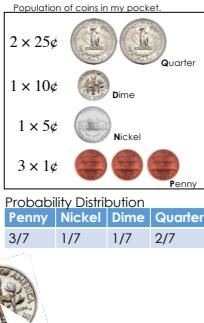
➤ Computing expectations:

$$1 \frac{3}{7} + 5 \frac{1}{7} + 10 \frac{1}{7} + 25 \frac{2}{7} = \frac{68}{7} \approx 9.71$$

➤ So the expected value is 9.71...

➤ Have you ever seen a 9.71 coin?

➤ Is this a problem?



Sampling Twice (Sample size 2)

- Suppose I sample two coins *with replacement*
- **With replacement:** put the coin back in pocket after sampling
- Let X_1 and X_2 be the first and second coin values.

- A friend gives me 4 more X_1 and 2 more X_2 and a quarter

➤ I define a new random variable:

$$Y = 5X_1 + 3X_2 + 25$$

➤ What is the value of Y ?

➤ Random

➤ What is the expected value of Y ?

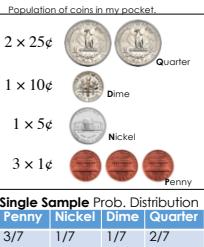
Calculating the Expected Value

$$Y = 5X_1 + 3X_2 + 25$$

		X_2			
		1¢	5¢	10¢	25¢
X_1	1¢				
	5¢				
	10¢				
	25¢				

Joint Probability Distribution

$$\mathbf{P}(X_1 = x_1, X_2 = x_2)$$



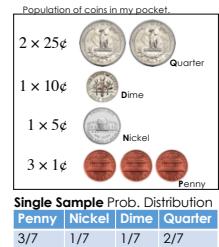
Calculating the Expected Value

$$Y = 5X_1 + 3X_2 + 25$$

		X_2			
		1¢	5¢	10¢	25¢
X_1	1¢	(3/7)(3/7)	(3/7)(1/7)		
	5¢				
	10¢				
	25¢				

Joint Probability Distribution

$$\mathbf{P}(X_1 = x_1, X_2 = x_2)$$



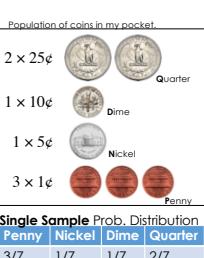
Calculating the Expected Value

$$Y = 5X_1 + 3X_2 + 25$$

		X_2			
		1¢	5¢	10¢	25¢
X_1	1¢	(3/7)(3/7)	(3/7)(1/7)		
	5¢				
	10¢				
	25¢				

Joint Probability Distribution

$$\mathbf{P}(X_1 = x_1, X_2 = x_2)$$



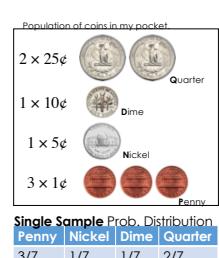
Calculating the Expected Value

$$Y = 5X_1 + 3X_2 + 25$$

		X_2			
		1¢	5¢	10¢	25¢
X_1	1¢	(3/7)(3/7)	(3/7)(1/7)		
	5¢				
	10¢				
	25¢				

Joint Probability Distribution

$$\mathbf{P}(X_1 = x_1, X_2 = x_2)$$



Calculating the Expected Value

$$Y = 5X_1 + 3X_2 + 25$$

	X_2			
	1¢	5¢	10¢	25¢
1¢	(3/7)(3/7)	(3/7)(1/7)		
5¢	(1/7)(3/7)	(1/7)(1/7)		
10¢				
25¢				

Joint Probability Distribution

$$\mathbf{P}(X_1 = x_1, X_2 = x_2)$$

Population of coins in my pocket.				
	2 × 25¢	1 × 10¢	1 × 5¢	3 × 1¢
Single Sample Prob. Distribution	Penny	Nickel	Dime	Quarter
	3/7	1/7	1/7	2/7

Calculating the Expected Value

$$Y = 5X_1 + 3X_2 + 25$$

	X_2			
	1¢	5¢	10¢	25¢
1¢	(3/7)(3/7)	(3/7)(1/7)	(3/7)(1/7)	(3/7)(2/7)
5¢	(1/7)(3/7)	(1/7)(1/7)	(1/7)(1/7)	(1/7)(2/7)
10¢	(1/7)(3/7)	(1/7)(1/7)	(1/7)(1/7)	(1/7)(2/7)
25¢	(2/7)(3/7)	(2/7)(1/7)	(2/7)(1/7)	(2/7)(2/7)

Joint Probability Distribution

$$\mathbf{P}(X_1 = x_1, X_2 = x_2)$$

Population of coins in my pocket.				
	2 × 25¢	1 × 10¢	1 × 5¢	3 × 1¢
Single Sample Prob. Distribution	Penny	Nickel	Dime	Quarter
	3/7	1/7	1/7	2/7

Calculating the Expected Value

$$Y = 5X_1 + 3X_2 + 25$$

	X_2			
	1¢	5¢	10¢	25¢
1¢	9/49	3/49	3/49	6/49
5¢	3/49	1/49	1/49	2/49
10¢	3/49	1/49	1/49	2/49
25¢	6/49	2/49	2/49	4/49

Joint Probability Distribution

$$\mathbf{P}(X_1 = x_1, X_2 = x_2)$$

Population of coins in my pocket.				
	2 × 25¢	1 × 10¢	1 × 5¢	3 × 1¢
Single Sample Prob. Distribution	Penny	Nickel	Dime	Quarter
	3/7	1/7	1/7	2/7

Calculating the Expected Value

$$Y = 5X_1 + 3X_2 + 25$$

$$\mathbf{E}[Y] = \sum_{x_1} \sum_{x_2} \mathbf{P}(x_1, x_2) (5x_1 + 3x_2 + 25)$$

	X_2			
	1¢	5¢	10¢	25¢
1¢	9/49	3/49	3/49	6/49
5¢	3/49	1/49	1/49	2/49
10¢	3/49	1/49	1/49	2/49
25¢	6/49	2/49	2/49	4/49

Joint Probability Distribution
 $\mathbf{P}(X_1 = x_1, X_2 = x_2)$

Calculating the Expected Value

$$Y = 5X_1 + 3X_2 + 25$$

$$\mathbf{E}[Y] = \sum_{x_1} \sum_{x_2} \mathbf{P}(x_1, x_2) (5x_1 + 3x_2 + 25)$$

$$= \mathbf{P}(1, 1) (5 \times 1 + 3 \times 1 + 25) +$$

	X_2			
	1¢	5¢	10¢	25¢
1¢	9/49	3/49	3/49	6/49
5¢	3/49	1/49	1/49	2/49
10¢	3/49	1/49	1/49	2/49
25¢	6/49	2/49	2/49	4/49

Joint Probability Distribution
 $\mathbf{P}(X_1 = x_1, X_2 = x_2)$

Calculating the Expected Value

$$Y = 5X_1 + 3X_2 + 25$$

$$\mathbf{E}[Y] = \sum_{x_1} \sum_{x_2} \mathbf{P}(x_1, x_2) (5x_1 + 3x_2 + 25)$$

$$= \mathbf{P}(1, 1) (5 \times 1 + 3 \times 1 + 25) +$$

	X_2			
	1¢	5¢	10¢	25¢
1¢	9/49	3/49	3/49	6/49
5¢	3/49	1/49	1/49	2/49
10¢	3/49	1/49	1/49	2/49
25¢	6/49	2/49	2/49	4/49

Joint Probability Distribution
 $\mathbf{P}(X_1 = x_1, X_2 = x_2)$

Calculating the Expected Value

$$Y = 5X_1 + 3X_2 + 25$$

$$\begin{aligned} \mathbf{E}[Y] &= \sum_{x_1} \sum_{x_2} \mathbf{P}(x_1, x_2) (5x_1 + 3x_2 + 25) \\ &= \mathbf{P}(1,1) (5 \times 1 + 3 \times 1 + 25) + \\ &\quad \mathbf{P}(1,5) (5 \times 1 + 3 \times 5 + 25) + \\ &\quad \mathbf{P}(1,10) (5 \times 1 + 3 \times 10 + 25) + \end{aligned}$$

		X_2			
		1€	5€	10€	25€
X_1	1€	9/49	3/49	3/49	6/49
	5€	3/49	1/49	1/49	2/49
	10€	3/49	1/49	1/49	2/49
	25€	6/49	2/49	2/49	4/49

Joint Probability Distribution
 $\mathbf{P}(X_1 = x_1, X_2 = x_2)$

Calculating the Expected Value

$$Y = 5X_1 + 3X_2 + 25$$

$$\begin{aligned} \mathbf{E}[Y] &= \sum_{x_1} \sum_{x_2} \mathbf{P}(x_1, x_2) (5x_1 + 3x_2 + 25) \\ &= \mathbf{P}(1,1) (5 \times 1 + 3 \times 1 + 25) + \\ &\quad \mathbf{P}(1,5) (5 \times 1 + 3 \times 5 + 25) + \\ &\quad \mathbf{P}(1,10) (5 \times 1 + 3 \times 10 + 25) + \\ &\quad \mathbf{P}(1,25) (5 \times 1 + 3 \times 25 + 25) + \\ &\quad \mathbf{P}(5,1) (5 \times 5 + 3 \times 1 + 25) + \end{aligned}$$

		X_2			
		1€	5€	10€	25€
X_1	1€	9/49	3/49	3/49	6/49
	5€	3/49	1/49	1/49	2/49
	10€	3/49	1/49	1/49	2/49
	25€	6/49	2/49	2/49	4/49

Joint Probability Distribution
 $\mathbf{P}(X_1 = x_1, X_2 = x_2)$

Calculating the Expected Value

$$Y = 5X_1 + 3X_2 + 25$$

$$\begin{aligned} \mathbf{E}[Y] &= \sum_{x_1} \sum_{x_2} \mathbf{P}(x_1, x_2) (5x_1 + 3x_2 + 25) \\ &= \frac{9}{49} (33) + 1 + 3 \quad \text{This is exhausting ...} \\ &\quad \frac{3}{49} (45) + 1 + 3 \\ &\quad \frac{3}{49} (60) + 1 + 3 \\ &\quad \frac{6}{49} (105) + 1 + 3 \\ &\quad \frac{3}{49} (53) + 1 + 3 \end{aligned}$$



		X_2			
		1€	5€	10€	25€
X_1	1€	9/49	3/49	3/49	6/49
	5€	3/49	1/49	1/49	2/49
	10€	3/49	1/49	1/49	2/49
	25€	6/49	2/49	2/49	4/49

Joint Probability Distribution
 $\mathbf{P}(X_1 = x_1, X_2 = x_2)$

There is a better way!

Lowercase Letters
are Constants
(not Random)

$$\mathbf{E}[aX + Y + b] = a\mathbf{E}[X] + \mathbf{E}[Y] + b$$

➤ What is the expected value of Y ?

$$\mathbf{E}[Y] = \mathbf{E}[5X_1 + 3X_2 + 25]$$

$$= \mathbf{E}[5X_1] + \mathbf{E}[3X_2] + \mathbf{E}[25] \quad \text{Linearity of expectation}$$

$$= \mathbf{E}[5X_1] + \mathbf{E}[3X_2] + 25 \quad \text{Expectation of constant is the constant.}$$

$$= 5\mathbf{E}[X_1] + 3\mathbf{E}[X_2] + 25 \quad \text{Linearity of expectation}$$

Linearity of Expectation

Lowercase Letters
are Constants
(not Random)

$$\mathbf{E}[aX + Y + b] = a\mathbf{E}[X] + \mathbf{E}[Y] + b$$

➤ What is the expected value of Y ?

$$\mathbf{E}[Y] = 5\mathbf{E}[X_1] + 3\mathbf{E}[X_2] + 25 \approx 102.71$$

$$\frac{68}{7}$$

$$\frac{68}{7}$$

➤ What if X_1 and X_2 were sampled **without replacement**?

- Can $X_1 = X_2 = 5$?
- Can I sample two dimes

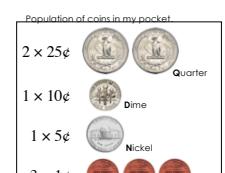


Dependent Random Variables

Sampling with replacement

		X_2			
		1€	5€	10€	25€
X_1	1€				
	5€				
	10€				
	25€				

Joint Probability Distribution
 $\mathbf{P}(X_1 = x_1, X_2 = x_2)$



Population of coins in my pocket
 2 × 25€ Quarter
 1 × 10€ Dime
 1 × 5€ Nickel
 3 × 1€ Penny

Single Sample Prob. Distribution
 Penny Nickel Dime Quarter
 3/7 1/7 1/7 2/7

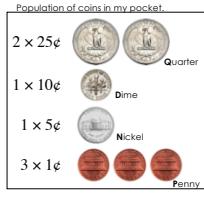
Dependent Random Variables

Sampling with replacement

	X_2			
	1¢	5¢	10¢	25¢
1¢	(3/7)(2/6)			
5¢				
10¢				
25¢				

Joint Probability Distribution

$$\mathbf{P}(X_1 = x_1, X_2 = x_2)$$



Single Sample Prob. Distribution

Penny Nickel Dime Quarter

3/7 1/7 1/7 2/7

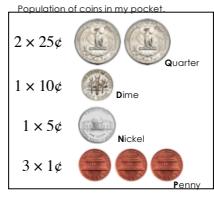
Dependent Random Variables

Sampling with replacement

	X_2			
	1¢	5¢	10¢	25¢
1¢	(3/7)(2/6)	(3/7)(1/6)		
5¢				
10¢				
25¢				

Joint Probability Distribution

$$\mathbf{P}(X_1 = x_1, X_2 = x_2)$$



Single Sample Prob. Distribution

Penny Nickel Dime Quarter

3/7 1/7 1/7 2/7

Dependent Random Variables

Sampling with replacement

	X_2			
	1¢	5¢	10¢	25¢
1¢	(3/7)(2/6)	(3/7)(1/6)		
5¢	(1/7)(3/6)	(1/7) 0	(1/7)(1/6)	(1/7)(2/6)
10¢	(1/7)(3/6)	(1/7)(1/6)	(1/7) 0	(1/7)(2/6)
25¢	(2/7)(3/6)	(2/7)(1/6)	(2/7)(1/6)	(2/7)(1/6)

Joint Probability Distribution

$$\mathbf{P}(X_1 = x_1, X_2 = x_2)$$



Single Sample Prob. Distribution

Penny Nickel Dime Quarter

3/7 1/7 1/7 2/7

Dependent Random Variables

Sampling with replacement

	X_2			
	1¢	5¢	10¢	25¢
1¢	(3/7)(2/6)	(3/7)(1/6)		
5¢	(1/7)(3/6)	(1/7) 0	(1/7)(1/6)	(1/7)(2/6)
10¢	(1/7)(3/6)	(1/7)(1/6)	(1/7) 0	(1/7)(2/6)
25¢	(2/7)(3/6)	(2/7)(1/6)	(2/7)(1/6)	(2/7)(1/6)

Joint Probability Distribution

$$\mathbf{P}(X_1 = x_1, X_2 = x_2)$$



Single Sample Prob. Distribution

Penny Nickel Dime Quarter

3/7 1/7 1/7 2/7

Dependent Random Variables

Sampling with replacement

	X_2			
	1¢	5¢	10¢	25¢
1¢	(3/7)(2/6)	(3/7)(1/6)	(3/7)(1/6)	(3/7)(2/6)
5¢	(1/7)(3/6)	(1/7) 0	(1/7)(1/6)	(1/7)(2/6)
10¢	(1/7)(3/6)	(1/7)(1/6)	(1/7) 0	(1/7)(2/6)
25¢	(2/7)(3/6)	(2/7)(1/6)	(2/7)(1/6)	(2/7)(1/6)

Joint Probability Distribution

$$\mathbf{P}(X_1 = x_1, X_2 = x_2)$$



Single Sample Prob. Distribution

Penny Nickel Dime Quarter

3/7 1/7 1/7 2/7

Dependent Random Variables

Sampling with replacement

	X_2			
	1¢	5¢	10¢	25¢
1¢	6/42	3/42	3/42	6/42
5¢	3/42	0	1/42	2/42
10¢	3/42	1/42	0	2/42
25¢	6/42	2/42	2/42	2/42

Joint Probability Distribution

$$\mathbf{P}(X_1 = x_1, X_2 = x_2)$$



Single Sample Prob. Distribution

Penny Nickel Dime Quarter

3/7 1/7 1/7 2/7

Dependent Random Variables

Sampling with replacement

$$Y = 5X_1 + 3X_2 + 25$$

$$\mathbf{E}[Y] = \sum_{x_1} \sum_{x_2} \mathbf{P}(x_1, x_2) (5x_1 + 3x_2 + 25)$$

	X_2	1€	5€	10€	25€
1€	6/42	3/42	3/42	6/42	
5€	3/42	0	1/42	2/42	
10€	3/42	1/42	0	2/42	
25€	6/42	2/42	2/42	2/42	

Joint Probability Distribution $\mathbf{P}(X_1 = x_1, X_2 = x_2)$



$$= (6/42) 33 + (3/42) 45 + (3/42) 60 + (6/42) 105 + (3/42) 53 + (0) 65 + (1/42) 80 + (2/42) 125 + (3/42) 78 + (1/42) 90 + (0) 105 + (2/42) 150 + (6/42) 153 + (2/42) 165 + (2/42) 180 + (2/42) 225$$

$$= \frac{719}{7} \approx 102.71 \quad \text{We have seen this before!}$$

$$\mathbf{E}[Y] = 5\mathbf{E}[X_1] + 3\mathbf{E}[X_2] + 25 \approx 102.71$$

Summary | Expected Value and Linearity of Expectation

➤ Expected Value

$$\mathbf{E}[X] = \sum_{x \in \mathcal{X}} x \mathbf{P}(x)$$

➤ Linearity of Expectation

$$\mathbf{E}[aX + Y + b] = a\mathbf{E}[X] + \mathbf{E}[Y] + b$$

➤ independence **not** required

➤ Proof?

Proving Linearity of Expectation

$$\begin{aligned} \mathbf{E}[aX + bY + c] &= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} \mathbf{P}(x, y)(ax + by + c) \\ &= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} \mathbf{P}(x, y)ax + \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} \mathbf{P}(x, y)by + \underbrace{\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} \mathbf{P}(x, y)c}_{\text{Sums to 1}} \\ &= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} \mathbf{P}(x, y)ax + \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} \mathbf{P}(x, y)by + c \end{aligned}$$

$$\mathbf{E}[aX + bY + c] = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} \mathbf{P}(x, y)ax + \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} \mathbf{P}(x, y)by + c$$

$$\boxed{\text{Conditional Defn. } \mathbf{P}(x, y) = \mathbf{P}(x|y)\mathbf{P}(y) = \mathbf{P}(y|x)\mathbf{P}(x)}$$

Using the above identity:

$$\begin{aligned} \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} \mathbf{P}(x, y)ax &= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} \mathbf{P}(y|x)\mathbf{P}(x)ax \\ &\stackrel{\text{Factoring out the terms that do not depend on } y}{=} a \sum_{x \in \mathcal{X}} \mathbf{P}(x)x \sum_{y \in \mathcal{Y}} \mathbf{P}(y|x) \\ &\stackrel{\text{Sums to 1}}{=} a \sum_{x \in \mathcal{X}} \mathbf{P}(x)x \\ &= a\mathbf{E}[x] \end{aligned}$$

Proving Linearity of Expectation

$$\mathbf{E}[aX + bY + c] = a\mathbf{E}[x] \mathbf{P}(x, y)ax + \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} \mathbf{P}(x, y)by + c$$

The remainder of the proof is left as an exercise.



Summary | Expected Value and Linearity of Expectation

➤ Expected Value

$$\mathbf{E}[X] = \sum_{x \in \mathcal{X}} x \mathbf{P}(x)$$

➤ Linearity of Expectation

$$\mathbf{E}[aX + Y + b] = a\mathbf{E}[X] + \mathbf{E}[Y] + b$$

➤ independence **not** required

➤ What about $\mathbf{E}[XY] \stackrel{?}{=} \mathbf{E}[X]\mathbf{E}[Y]$

Summary | Expected Value and Linearity of Expectation

- Expected Value

$$\mathbf{E}[X] = \sum_{x \in \mathcal{X}} x \mathbf{P}(x)$$

- Linearity of Expectation

$$\mathbf{E}[aX + Y + b] = a\mathbf{E}[X] + \mathbf{E}[Y] + b$$

- independence **not** required

- If X and Y are **independent** then $\mathbf{E}[XY] = \mathbf{E}[X]\mathbf{E}[Y]$

Characterizing Random Variables

- **Probability Mass Function (PMF): Discrete Distribution**

➤ The probability a variable will take on a particular value

- **Probability Density Function (PDF): Continuous Distributions**

➤ Not covered ... here there be dragons

- **Expectation**

➤ The average value the variable takes (the mean)

- **Variance**

➤ The spread of the variable about the mean

The Variance

$$\mathbf{Var}[X] = \mathbf{E}[(X - \mathbf{E}[X])^2] = \sum_{x \in \mathcal{X}} (x - \mathbf{E}[X])^2 \mathbf{P}(x)$$

- Useful Identity:

$$\mathbf{Var}[X] = \mathbf{E}[(X - \mathbf{E}[X])^2]$$

Expanding the square = $\mathbf{E}[X^2 - 2X\mathbf{E}[X] + \mathbf{E}[X]^2]$

- Useful Identity:

$$\mathbf{Var}[X] = \mathbf{E}[(X - \mathbf{E}[X])^2]$$

Expanding the square = $\mathbf{E}[X^2 - 2X\mathbf{E}[X] + \mathbf{E}[X]^2]$

Linearity of expectation = $\mathbf{E}[X^2] - \mathbf{E}[2X\mathbf{E}[X]] + \mathbf{E}[\mathbf{E}[X]^2]$

Linearity of expectation = $\mathbf{E}[X^2] - 2\mathbf{E}[X]\mathbf{E}[X] + \mathbf{E}[X]^2$

Algebra = $\mathbf{E}[X^2] - \mathbf{E}[X]^2$

The Variance

$$\begin{aligned} \mathbf{Var}[X] &= \mathbf{E}[(X - \mathbf{E}[X])^2] = \sum_{x \in \mathcal{X}} (x - \mathbf{E}[X])^2 \mathbf{P}(x) \\ &= \mathbf{E}[X^2] - \mathbf{E}[X]^2 \end{aligned}$$

- Properties of Variance:

$$\mathbf{Var}[aX + b] = a^2 \mathbf{Var}[X] + 0$$

- If X and Y are independent:

$$\mathbf{Var}[X + Y] = \mathbf{Var}[X] + \mathbf{Var}[Y]$$

$$= \mathbf{E}[X^2] - \mathbf{E}[X]^2$$

- Properties of Variance:

$$\mathbf{Var}[aX + b] = a^2 \mathbf{Var}[X] + 0$$

- If X and Y are independent:

$$\mathbf{Var}[X + Y] = \mathbf{Var}[X] + \mathbf{Var}[Y]$$

- Standard Deviation (easier to interpret units)

$$\mathbf{SD}[X] = \sqrt{\mathbf{Var}[X]}$$

- Useful identity

$$\mathbf{SD}[aX + b] = |a| \mathbf{SD}[X]$$

Covariance

- The covariance describes how two variables vary jointly

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

- Basic properties of the covariance

$$\text{Cov}[aX + u, bY + v] = ab\text{Cov}[X, Y]$$

- If X and Y are **independent** then: $\mathbf{E}[XY] = \mathbf{E}[X]\mathbf{E}[Y]$

$$\text{Cov}[X, Y] = 0$$

Correlation

Covariance
$\text{Cov}[X, Y] = \mathbf{E}[(X - \mathbf{E}[X])(Y - \mathbf{E}[Y])]$ $= \mathbf{E}[XY] - \mathbf{E}[X]\mathbf{E}[Y]$

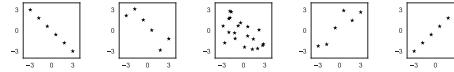
- The units of covariance can be difficult to reason about

- Correlation** is the "normalized" covariance

$$\rho_{X,Y} = \text{Corr}[X, Y] = \frac{\text{Cov}[X, Y]}{\sqrt{\text{Var}[X]}\sqrt{\text{Var}[Y]}} = \frac{\text{Cov}[X, Y]}{\text{SD}[X]\text{SD}[Y]}$$

- A number between -1 and 1

$$\rho_{X,Y} = -1 \iff \rho_{X,Y} \approx 0 \iff \rho_{X,Y} = 1$$



Practice Distributions

Bernoulli PMF		
Value	1	0
Chance	p	$1-p$

<http://bit.ly/ds100-sp18-var>

$$\text{Var}[X] = \mathbf{E}[(X - \mathbf{E}[X])^2] = \mathbf{E}[X^2] - \mathbf{E}[X]^2$$

$$\mathbf{E}[X] = \sum_{x \in \mathcal{X}} x \mathbf{P}(x)$$

What is the value of the following in terms of p

$$\mathbf{E}[X] =$$

$$\text{Var}[X] =$$

Binary Random Variable (Bernoulli)

- Takes on two values (e.g., {0,1}, {heads, tails}...)

$$X \sim \text{Bernoulli}(p)$$

- Characterized by probability p

Value	1	0
Chance	p	$1-p$

- Expected Value:

$$\mathbf{E}[X] =$$

- Variance

$$\text{Var}[X] =$$

Binary Random Variable (Bernoulli)

- Takes on two values (e.g., {0,1}, {heads, tails}...)

$$X \sim \text{Bernoulli}(p)$$

- Characterized by probability p

Value	1	0
Chance	p	$1-p$

- Expected Value:

$$\mathbf{E}[X] = 1 * p + 0 * (1 - p) = p$$

- Variance

$$\text{Var}[X] = (1 - p)^2 * p + (0 - p)^2(1 - p) = p(1 - p)$$

Another Example



- I like to eat shishito peppers
- Usually they are not too spicy ...
 - but occasionally you get unlucky (or lucky)
- Suppose we **sample n peppers** at random from the **population of all shishito peppers**
 - can we do this in practice?
 - Difficult! Maybe cluster sample farms?
 - What can our sample tell us about the population?



Formalizing the Shishito Peppers

- **Population:** all shishito peppers
- **Generation Process:** simple random sample
- **Sample:** we have a sample of n shishito peppers
- **Random Variables:** we define a set of n random variables

$$X_1, X_2, \dots, X_n \sim \text{Bernoulli}(p^*)$$

- Where $X_i = 1$ if the i^{th} pepper is spicy and 0 otherwise.

Population Parameter
(We don't know it.)
Remember star is for the universe.

- **Random Variables:** we define a set of n random variables

$$X_1, X_2, \dots, X_n \sim \text{Bernoulli}(p^*)$$

- Where $X_i = 1$ if the i^{th} pepper is spicy and 0 otherwise.

Population Parameter
(We don't know it.)
Remember star is for the universe.

- **Sample Mean:** Is a random variable

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$$

- **Expected Value** of the sample mean:

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i \quad | \quad X_1, X_2, \dots, X_n \sim \text{Bernoulli}(p^*)$$

- **Expected Value** of the sample mean:

$$\begin{aligned} \mathbf{E}[\bar{X}] &= \mathbf{E}\left[\frac{1}{n} \sum_{i=1}^n X_i\right] &= \frac{1}{n} \sum_{i=1}^n \mathbf{E}[X_i] \\ &= \frac{1}{n} \sum_{i=1}^n \mu = \mu &\text{Let } \mu \text{ be the expected value for all } X_i \\ &= p^* &\text{For the shishito peppers setting we have } \mu = p^* \end{aligned}$$

The expected value of the **sample mean** is the **population mean!**

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i \quad | \quad X_1, X_2, \dots, X_n \sim \text{Bernoulli}(p^*)$$

- **Expected Value** of the sample mean:

$$\mathbf{E}[\bar{X}] = \frac{1}{n} \sum_{i=1}^n \mu = \mu$$

- The **sample mean** is an **unbiased estimator** of the population mean

$$\text{Bias} = \mathbf{E}[\bar{X}] - \mu = 0$$

Sample Mean is a Random Variable

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$$

- **Expected Value:**

$$\mathbf{E}[\bar{X}] = \mathbf{E}\left[\frac{1}{n} \sum_{i=1}^n X_i\right] = \frac{1}{n} \sum_{i=1}^n \mu = \mu$$

- **Variance:**

$$\mathbf{Var}[\bar{X}] = \mathbf{Var}\left[\frac{1}{n} \sum_{i=1}^n X_i\right]$$

➤ Variance:

$$\text{Var} [\bar{X}] = \text{Var} \left[\frac{1}{n} \sum_{i=1}^n X_i \right] = \frac{1}{n^2} \text{Var} \left[\sum_{i=1}^n X_i \right] \text{Property of the Variance}$$

$$\text{If the } X_i \text{ are independent!} = \frac{1}{n^2} \sum_{i=1}^n \text{Var} [X_i]$$

➤ In the shishito peppers example are the X_i independent?

➤ Depends on the sampling strategy

➤ Random with replacement (after tasting) → Yes!



➤ Random **without** replacement → No!
➤ Correction factor is small for large populations

➤ Variance:

$$\text{Var} [\bar{X}] = \text{Var} \left[\frac{1}{n} \sum_{i=1}^n X_i \right] = \frac{1}{n^2} \text{Var} \left[\sum_{i=1}^n X_i \right] \text{Property of the Variance}$$

$$\text{If the } X_i \text{ are independent!} = \frac{1}{n^2} \sum_{i=1}^n \text{Var} [X_i]$$

$$\text{Define the variance of } X_i \text{ as } \sigma^2 = \frac{1}{n^2} \sum_{i=1}^n \sigma^2 = \frac{\sigma^2}{n}$$

$$\text{For shishito peppers with replacement} = \frac{p^*(1-p^*)}{n}$$

The variance of the sample mean decreases at a rate of one over the sample size

Summary of Sample Mean Statistics

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$$

➤ Expected Value:

$$\mathbf{E} [\bar{X}] = \mathbf{E} \left[\frac{1}{n} \sum_{i=1}^n X_i \right] = \frac{1}{n} \sum_{i=1}^n \mu = \mu$$

➤ Variance:

$$\text{Var} [\bar{X}] = \text{Var} \left[\frac{1}{n} \sum_{i=1}^n X_i \right] = \frac{\sigma^2}{n} \text{ Assuming } X_i \text{ are independent}$$

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$$

➤ Expected Value:

$$\mathbf{E} [\bar{X}] = \mathbf{E} \left[\frac{1}{n} \sum_{i=1}^n X_i \right] = \frac{1}{n} \sum_{i=1}^n \mu = \mu$$

➤ Variance:

$$\text{Var} [\bar{X}] = \text{Var} \left[\frac{1}{n} \sum_{i=1}^n X_i \right] = \frac{\sigma^2}{n} \text{ Assuming } X_i \text{ are independent}$$

➤ Standard Error:

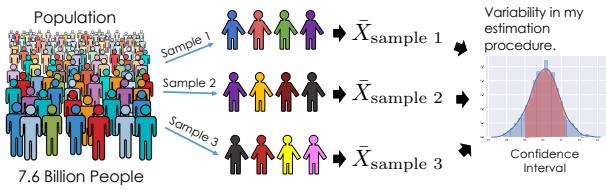
$$\text{SE} (\bar{X}) = \sqrt{\text{Var} [\bar{X}]} = \frac{\sigma}{\sqrt{n}} \text{ Square root law}$$

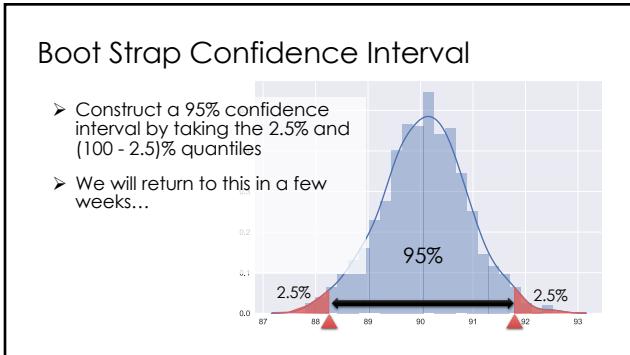
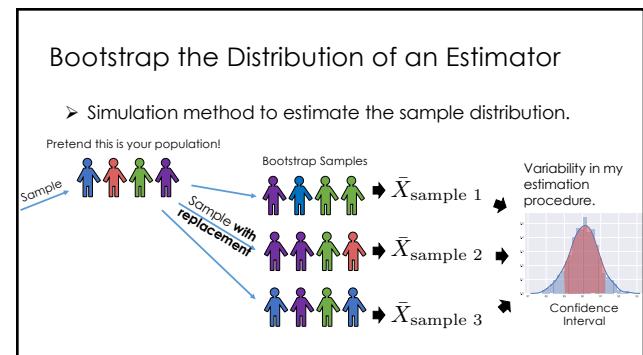
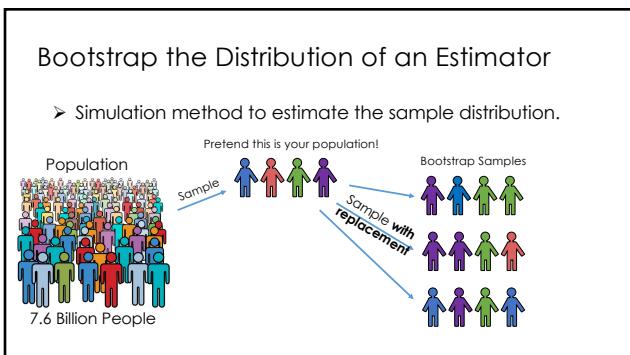
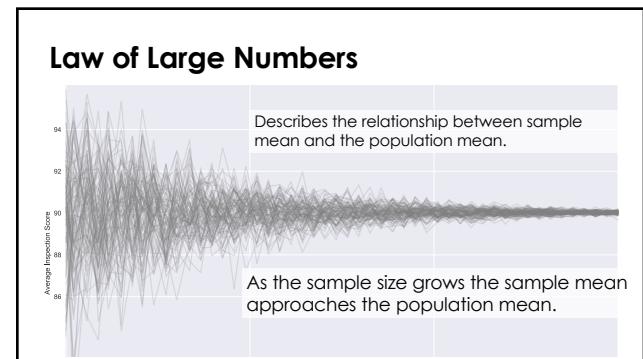
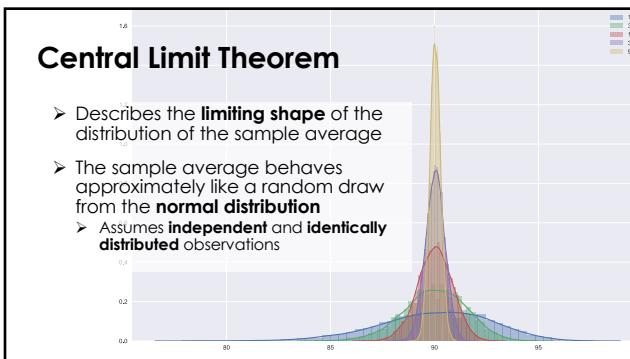
\bar{X} has a probability mass function

ALSO KNOW AS A
SAMPLING DISTRIBUTION

The Distribution of an Estimator

➤ Resampling the population to estimate the sample distribution.





Connection to Loss Minimization

The Sample Loss

- Recall earlier that we used the average loss

$$L_{\text{avg}}(\theta) = \frac{1}{n} \sum_{i=1}^n \ell(Y_i, f_\theta(X_i))$$

The Sample Loss

- Recall earlier that we used the average loss

$$L_{\text{avg}}(\theta) = \frac{1}{n} \sum_{i=1}^n \ell(Y_i, f_\theta(X_i))$$

Notice that this is really a **sample loss**

It is a **random variable** (depends on X_i and Y_i)

How does it relate to the population?

We will answer this question precisely for the squared loss in the next lecture using **bias** and **variance**

Today we will relate the **expected loss** to the **sample loss**

Risk and the Expected Loss

- We can define the expected loss as:

$$R(\theta) = \mathbf{E}[\ell(Y, f_\theta(X))]$$

- This is called the **risk**

It is the risk associated with the choice of θ
Not a random variable

- Given access to the joint probability of X and Y we can rewrite the **risk** as:

$$R(\theta) = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} \ell(y, f_\theta(x)) \mathbf{P}(x, y)$$

$$\text{Average Sample Loss}$$

$$L_{\text{avg}}(\theta) = \frac{1}{n} \sum_{i=1}^n \ell(Y_i, f_\theta(X_i))$$

Given access to the **joint probability** of X and Y we can rewrite the **risk** as:

$$R(\theta) = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} \ell(y, f_\theta(x)) \mathbf{P}(x, y)$$

A natural objective would be to minimize the risk

$$\hat{\theta} = \arg \min_{\theta} \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} \ell(y, f_\theta(x)) \mathbf{P}(x, y)$$

Unfortunately, we don't have the joint prob. $\mathbf{P}(x, y)$

We can approximate $\mathbf{P}(x, y)$ with our samples.

- Given access to the **joint probability** of X and Y we can rewrite the **risk** as:

$$R(\theta) = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} \ell(y, f_\theta(x)) \mathbf{P}(x, y)$$

- The **empirical risk** approximates the true risk

$$R(\theta) \approx \hat{R}(\theta) = \sum_{i=1}^n \ell(Y_i, f_\theta(X_i)) \frac{1}{n}$$

Where the X_i and Y_i are drawn from the joint probability (a random sample)

$$(X_i, Y_i) \sim \mathbf{P}(x, y)$$

- Given access to the **joint probability** of X and Y we can rewrite the **risk** as:

$$R(\theta) = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} \ell(y, f_\theta(x)) \mathbf{P}(x, y)$$

- The **empirical risk** approximates the true risk

$$R(\theta) \approx \hat{R}(\theta) = \sum_{i=1}^n \ell(Y_i, f_\theta(X_i)) \frac{1}{n}$$

This is just the average loss from before:

$$L_{\text{avg}}(\theta) = \frac{1}{n} \sum_{i=1}^n \ell(Y_i, f_\theta(X_i))$$

Assuming:
 $(X_i, Y_i) \sim \mathbf{P}(x, y)$

Summary

- Today we reviewed
 - **Joint Probability Distributions**
 - **Expectation**
 - **Variance**
 - **Covariance**
- Studied Properties of the **Sample Mean**
 - **Unbiased**
 - **Law of large numbers:** convergence to the population mean
 - **Central Limit Theorem:** Distribution
- Connected the **Average Loss** to the **Empirical Risk**