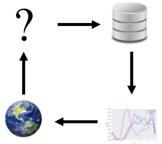
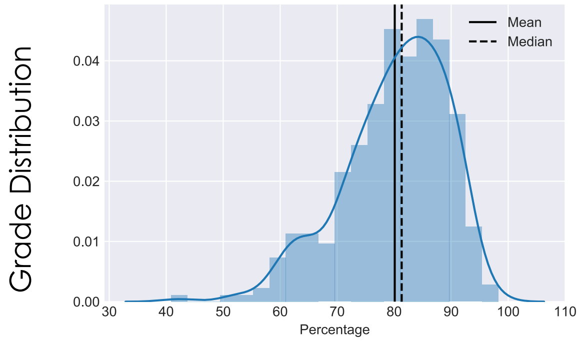
Data Science 100
Probability and Generalization

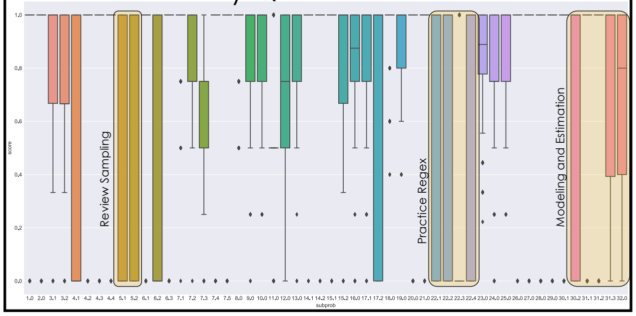
Slides by:
Joseph E. Gonzalez & Deb Nolan,
jegonzal@berkeley.edu
deborah_nolan@berkeley.edu



How was the Midterm?



Breakdown by Question

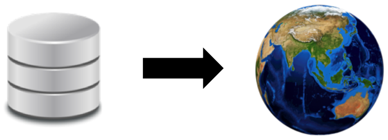


Recap: *Modeling and Estimation*

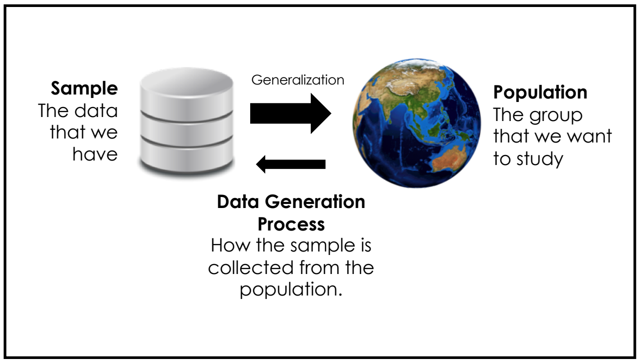
1. **Define the Model:** simplified representation of the world
2. **Define the Loss Function:** measures how well a particular instance of the model "fits" the data
3. **Minimize the Loss Function:** find the parameter values that minimize the loss **on the data**

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

Generalization



The focus of the next few lectures.



Sample
The data that we have

Generalization

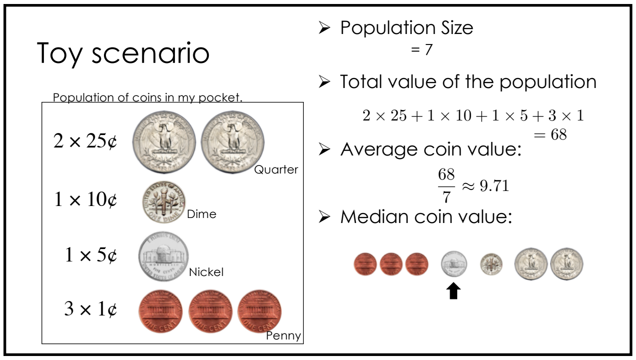
Population
The group that we want to study

Data Generation Process
How the sample is collected from the population.

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 of coins in my pocket.

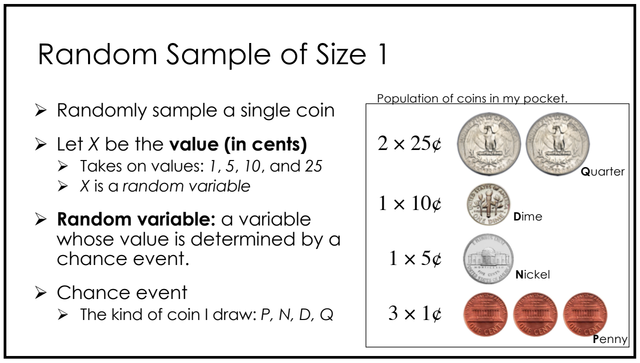
2×25¢ Quarter

1×10¢ Dime

1×5¢ Nickel

3×1¢ Penny

- 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*

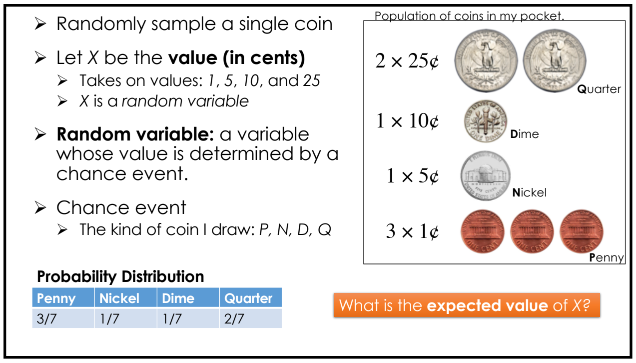
Population of coins in my pocket.

2×25¢ Quarter

1×10¢ Dime

1×5¢ Nickel

3×1¢ Penny



- 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*

Population of coins in my pocket.

2×25¢ Quarter

1×10¢ Dime

1×5¢ Nickel

3×1¢ Penny

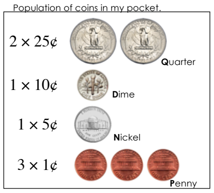
Probability Distribution

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

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

The Expected Value

Population of coins in my pocket.



Probability Distribution

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

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

- 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

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

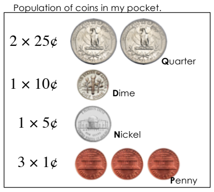
X_2

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

Joint Probability Distribution

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

Population of coins in my pocket.



Single Sample Prob. Distribution

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

Calculating the Expected Value

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

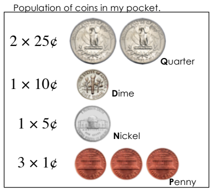
X_2

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

Joint Probability Distribution

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

Population of coins in my pocket.



Single Sample Prob. Distribution

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

Calculating the Expected Value

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

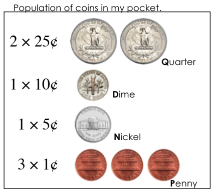
X_2

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

Joint Probability Distribution

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

Population of coins in my pocket.



Single Sample Prob. Distribution

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

Calculating the Expected Value

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

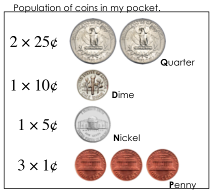
X_2

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

Joint Probability Distribution

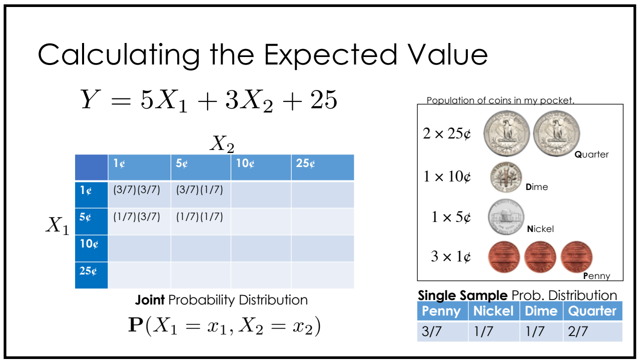
$$\mathbf{P}(X_1 = x_1, X_2 = x_2)$$

Population of coins in my pocket.



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

X_1

	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¢ **Q**uarter

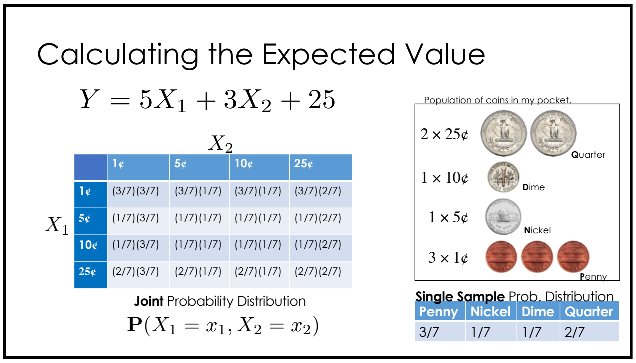
1 × 10¢ **D**ime

1 × 5¢ **N**ickel

3 × 1¢ **P**enny

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

X_1

	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¢ **Q**uarter

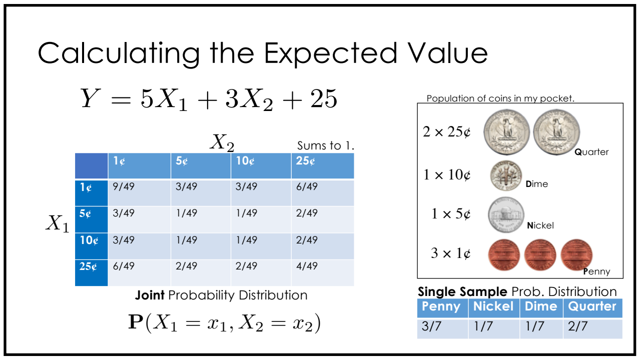
1 × 10¢ **D**ime

1 × 5¢ **N**ickel

3 × 1¢ **P**enny

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

Sums to 1.

X_1

	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¢ **Q**uarter

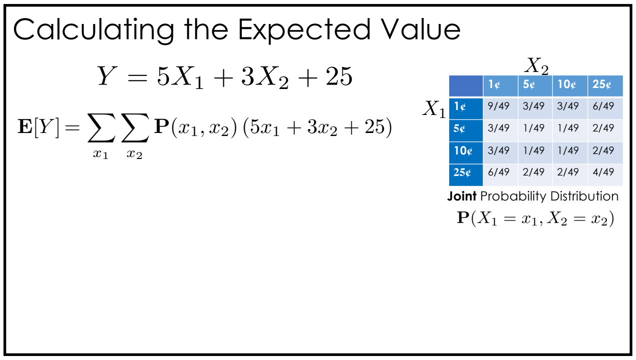
1 × 10¢ **D**ime

1 × 5¢ **N**ickel

3 × 1¢ **P**enny

Single Sample Prob. Distribution

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



Calculating the Expected Value

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

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

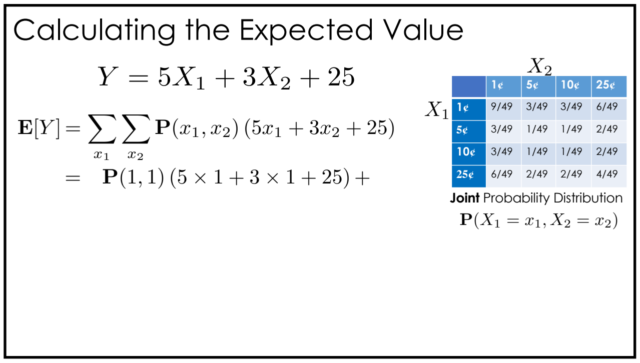
X_2

X_1

	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

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

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

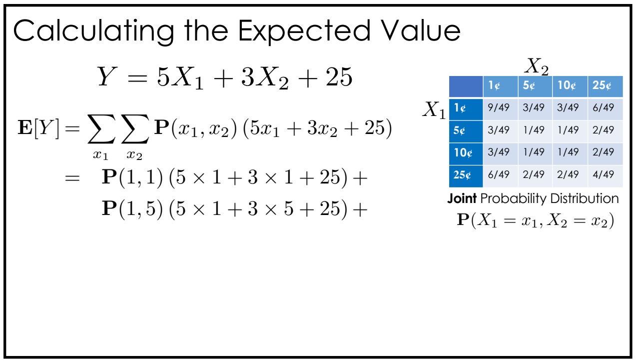
X_2

X_1

	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

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

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

X_2

X_1

	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

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

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

$$\begin{aligned} = \;& \mathbf{P}(1,1)\,(5\times 1 + 3\times 1 + 25) + \\ & \mathbf{P}(1,5)\,(5\times 1 + 3\times 5 + 25) + \\ & \mathbf{P}(1,10)\,(5\times 1 + 3\times 10 + 25) + \end{aligned}$$

X_2 (columns), X_1 (rows)

	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

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

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

$$\begin{aligned} = \;& \mathbf{P}(1,1)\,(5\times 1 + 3\times 1 + 25) + \\ & \mathbf{P}(1,5)\,(5\times 1 + 3\times 5 + 25) + \\ & \mathbf{P}(1,10)\,(5\times 1 + 3\times 10 + 25) + \\ & \mathbf{P}(1,25)\,(5\times 1 + 3\times 25 + 25) + \\ & \mathbf{P}(5,1)\,(5\times 5 + 3\times 1 + 25) + \\ & \ldots \end{aligned}$$

	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

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

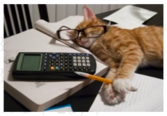
Calculating the Expected Value

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

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

$$\begin{aligned} = \;& 9/49\ (33) + \\ & 3/49\ (45) + \\ & 3/49\ (60) + \\ & 6/49\ (105) + \\ & 3/49\ (53) + \\ & \ldots \end{aligned}$$

This is exhausting …



	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

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

There is a better way!

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*?

$$\begin{aligned} \mathbf{E}[Y] &= \mathbf{E}[5X_1 + 3X_2 + 25] \\ &= \mathbf{E}[5X_1] + \mathbf{E}[3X_2] + \mathbf{E}[25] && \text{Linearity of expectation} \\ &= \mathbf{E}[5X_1] + \mathbf{E}[3X_2] + 25 && \text{Expectation of constant is the constant.} \\ &= 5\mathbf{E}[X_1] + 3\mathbf{E}[X_2] + 25 && \text{Linearity of expectation} \end{aligned}$$

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*?

$$\mathbf{E}[Y] = 5\underbrace{\mathbf{E}[X_1]}_{\frac{68}{7}} + 3\underbrace{\mathbf{E}[X_2]}_{\frac{68}{7}} + 25 \approx 102.71$$

- 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 without replacement

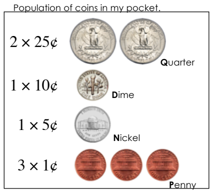
X_2 (columns), X_1 (rows)

	1¢	5¢	10¢	25¢
1¢				
5¢				
10¢				
25¢				

Joint Probability Distribution

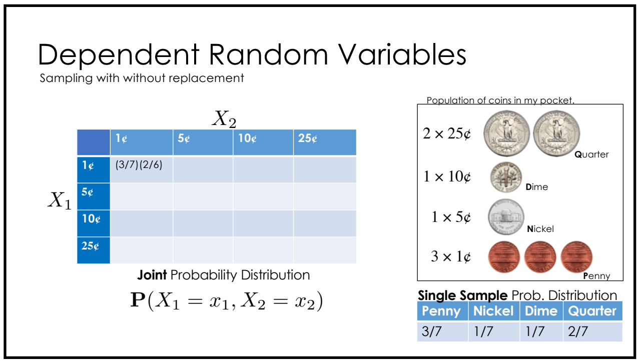
$$\mathbf{P}(X_1 = x_1, X_2 = x_2)$$

Population of coins in my pocket.



Single Sample Prob. Distribution

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



Dependent Random Variables

Sampling with without replacement

X_2

X_1	1¢	5¢	10¢	25¢
1¢	(3/7)(2/6)			
5¢				
10¢				
25¢				

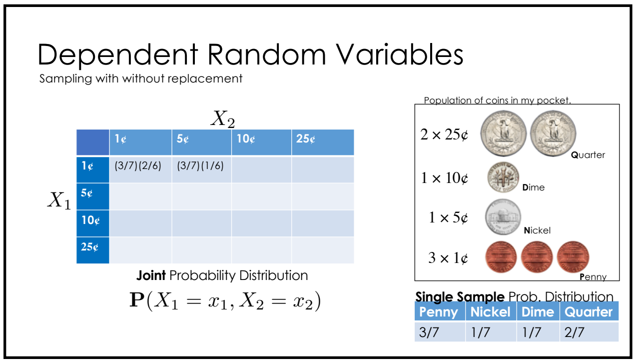
Joint Probability Distribution

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

Population of coins in my pocket.

Single Sample Prob. Distribution

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



Dependent Random Variables

Sampling with without replacement

X_2

X_1	1¢	5¢	10¢	25¢
1¢	(3/7)(2/6)	(3/7)(1/6)		
5¢				
10¢				
25¢				

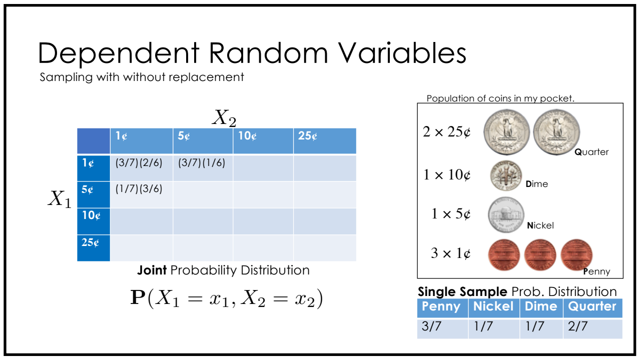
Joint Probability Distribution

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

Population of coins in my pocket.

Single Sample Prob. Distribution

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



Dependent Random Variables

Sampling with without replacement

X_2

X_1	1¢	5¢	10¢	25¢
1¢	(3/7)(2/6)	(3/7)(1/6)		
5¢	(1/7)(3/6)			
10¢				
25¢				

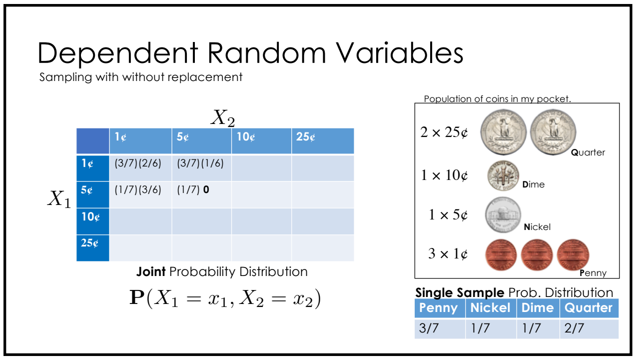
Joint Probability Distribution

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

Population of coins in my pocket.

Single Sample Prob. Distribution

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



Dependent Random Variables

Sampling with without replacement

X_2

X_1	1¢	5¢	10¢	25¢
1¢	(3/7)(2/6)	(3/7)(1/6)		
5¢	(1/7)(3/6)	(1/7) **0**		
10¢				
25¢				

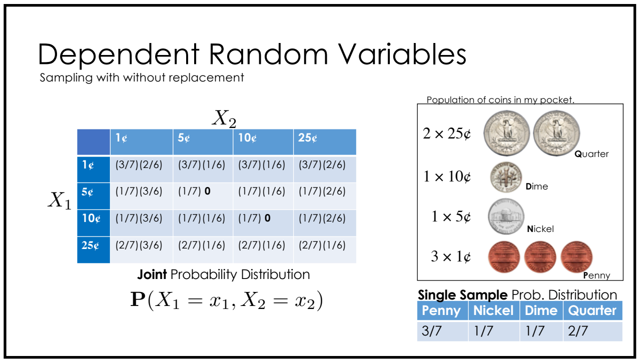
Joint Probability Distribution

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

Population of coins in my pocket.

Single Sample Prob. Distribution

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



Dependent Random Variables

Sampling with without replacement

X_2

X_1	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)

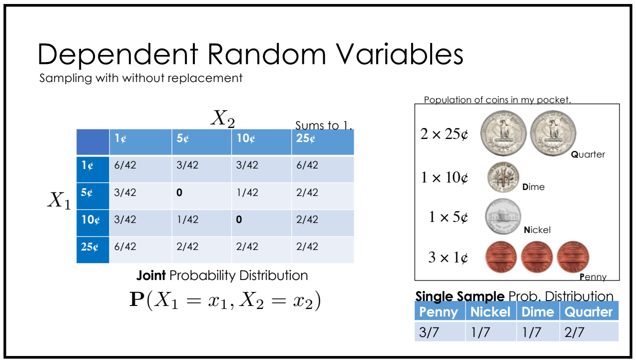
Joint Probability Distribution

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

Population of coins in my pocket.

Single Sample Prob. Distribution

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



Dependent Random Variables

Sampling with without replacement

X_2

Sums to 1.

X_1	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

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

Population of coins in my pocket.

Single Sample Prob. Distribution

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

Dependent Random Variables

Sampling without 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

X_1	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

$$= \frac{719}{7} \approx 102.71$$

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

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

$$= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} \mathbf{P}(x,y)ax + \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} \mathbf{P}(x,y)by + \underbrace{\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} \mathbf{P}(x,y)}_{\text{Sums to 1}}c$$

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

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

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

Using the above identity:

$$\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} \mathbf{P}(x,y)ax = \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} \mathbf{P}(y\,|\,x)\mathbf{P}(x)ax$$

Factoring out the terms that do not depend on y

$$= a\sum_{x\in\mathcal{X}} \mathbf{P}(x)x \underbrace{\sum_{y\in\mathcal{Y}} \mathbf{P}(y\,|\,x)}_{\text{Sums to 1}} = a\sum_{x\in\mathcal{X}} \mathbf{P}(x)x$$

$$= a\mathbf{E}[x]$$

Proving Linearity of Expectation

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

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 $\mathbf{E}[XY] \overset{?}{=} \mathbf{E}[X]\mathbf{E}[Y]$

Summary | Expected Value and Linearity of Expectation

- Expected Value

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

- Linearity of Expectation

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

  - 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

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

- Useful Identity:

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

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

- Useful Identity:

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

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

Linearity of expectation $= \mathbf{E}\left[X^2\right] - \mathbf{E}\left[2X\overbrace{\mathbf{E}\left[X\right]}^{\text{constant}}\right] + \mathbf{E}\left[\underbrace{\mathbf{E}\left[X\right]^2}_{\text{constant}}\right]$

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

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

The Variance

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

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

- Properties of Variance:

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

  - If X and Y are independent:

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

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

- Properties of Variance:

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

  - If X and Y are independent:

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

- Standard Deviation (easier to interpret units)

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

  - Useful identity

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

Covariance

- The covariance describes how to variables vary jointly

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

- Basic properties of the covariance

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

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

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

Correlation

Covariance

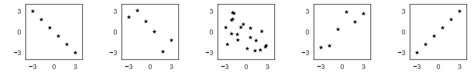
$$\mathbf{Cov}[X,Y] = \mathbf{E}\left[(X - \mathbf{E}[X])(Y - \mathbf{E}[Y])\right]$$
$$= \mathbf{E}\left[XY\right] - \mathbf{E}[X]\mathbf{E}[Y]$$

- The units of covariance can be difficult to reason about
- **Correlation** is the "normalized" covariance

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

  - A number between -1 and 1

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



Practice Distributions

Binary Random Variable (Bernoulli)

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

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

- Characterized by probability *p*

Value	1	0
Chance	p	1-p

- Expected Value:

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

- Variance

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

Bernoulli PMF

Value	1	0
Chance	*p*	*1-p*

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

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

What is the value of the following in terms of p

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

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

Binary Random Variable (Bernoulli)

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

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

- Characterized by probability *p*

Value	1	0
Chance	p	1-p

- Expected Value:

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

- Variance

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

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

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

- 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

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

- 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:

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

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

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

Linearity of expectation

$$= \frac{1}{n}\sum_{i=1}^{n} \mu = \mu$$

Let μ be the expected value for all X_i

$$= p^*$$

For the shishito peppers setting we have $\mu = p^*$

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

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

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

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

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

$$\textbf{Bias} \; = \mathbf{E}\left[\bar{X}\right] - \mu = 0$$

Sample Mean is a Random Variable

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

- Expected Value:

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

- Variance:

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

- Variance:

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

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

- 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:

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

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

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

For shishto 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:

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

- Variance:

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

---

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

- Expected Value:

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

- Variance:

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

- Standard Error:

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

---

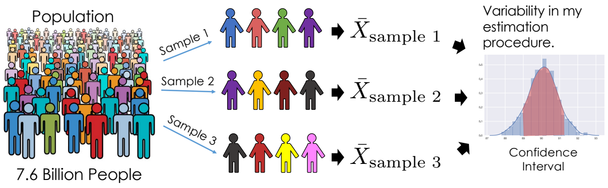
$\bar{X}$ has a probability mass function

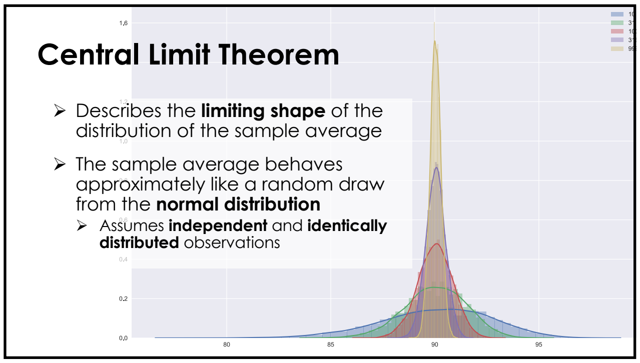
ALSO KNOW AS A

SAMPLING DISTRIBUTION

The Distribution of an Estimator

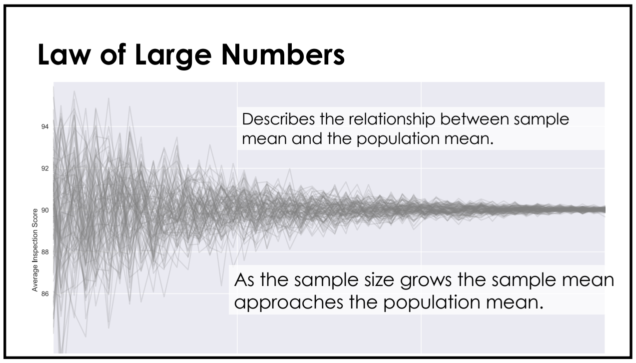
- Resampling the population to estimate the sample distribution.





Central Limit Theorem

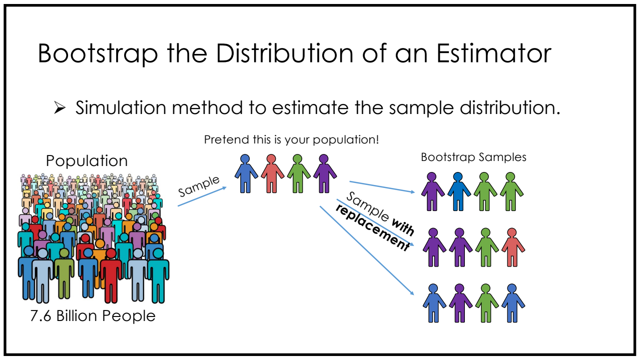
- Describes the **limiting shape** of the distribution of the sample average
- The sample average behaves approximately like a random draw from the **normal distribution**
  - Assumes **independent** and **identically distributed** observations



Law of Large Numbers

Describes the relationship between sample mean and the population mean.

As the sample size grows the sample mean approaches the population mean.



Bootstrap the Distribution of an Estimator

- Simulation method to estimate the sample distribution.

Population

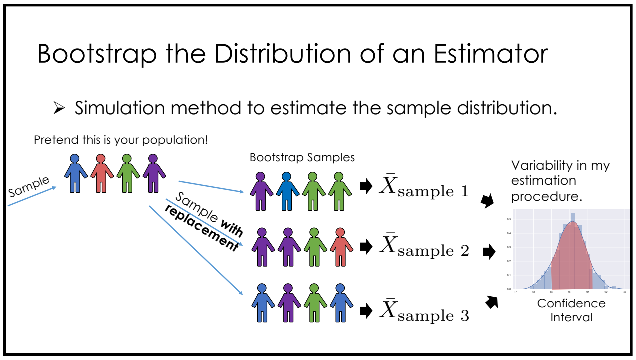
Pretend this is your population!

Bootstrap Samples

Sample

Sample **with replacement**

7.6 Billion People



Bootstrap the Distribution of an Estimator

- Simulation method to estimate the sample distribution.

Pretend this is your population!

Bootstrap Samples

Sample

Sample **with replacement**

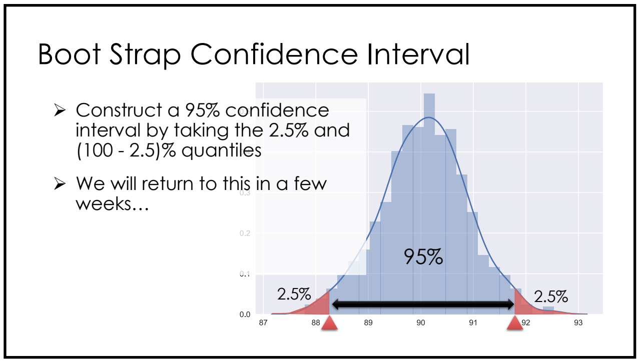
$\bar{X}_{\text{sample 1}}$

$\bar{X}_{\text{sample 2}}$

$\bar{X}_{\text{sample 3}}$

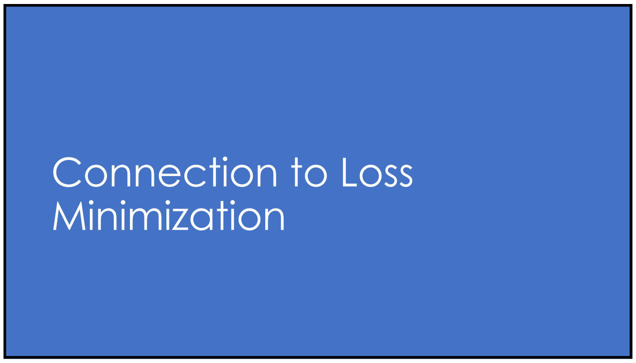
Variability in my estimation procedure.

Confidence Interval



Boot Strap Confidence Interval

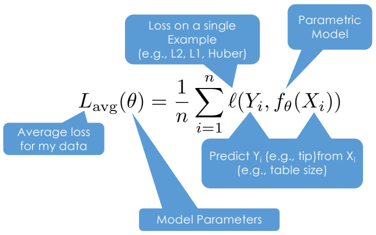
- Construct a 95% confidence interval by taking the 2.5% and (100 - 2.5)% quantiles
- We will return to this in a few weeks…



Connection to Loss Minimization

The Sample Loss

- Recall earlier that we used the average loss



Loss on a single Example (e.g., L2, L1, Huber)

Parametric Model

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

Average loss for my data

Predict Y_i (e.g., tip)from X_i (e.g., table size)

Model Parameters

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 related the **expected loss** to the **sample loss**

Risk and the **Expected** Loss

Average Sample Loss

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

- We can define the expected loss as:

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

- 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:

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

- 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:

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

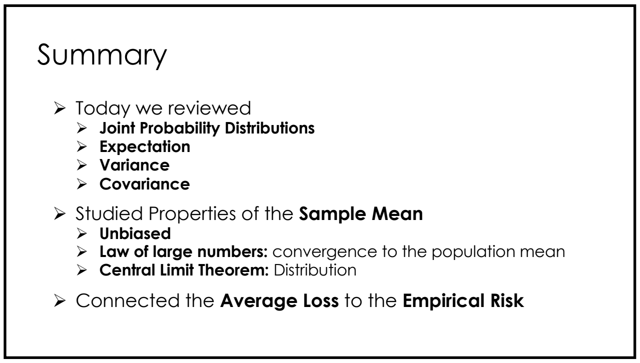
- The **empirical risk** approximates the true risk

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

- This is just the average loss from before:

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

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



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**