LESSON 10

Intro to Random Variables

CSCI 3022

Course Logistics: Your Fourth Week At A Glance

Mon 2/3	Tues 2/4	Wed 2/5	Thurs 2/6	Fri 2/7
Attend & Participate in Class		Attend & Participate in Class	HW 4 Due 11:59pm via Gradescope	In Class Quiz 3 (beginning of class): Scope: Lessons 1-6; HW 2 and HW 3 Attend & Participate in Class
Quiz 2 feedback/ grades posted			HW 3 feedback/ grades posted	HW 5 released (8am)

The Core Probability Toolkit

The Law of Total Probability

$$P(E) = P(E \text{ and } F) + P(E \text{ and } F^{C})$$

$$P(E) = P(E|F)\,P(F) + P(E|F^{C})\,P(F^{C})$$

$$P(E) = \sum_{i=1}^{n} P(E \text{ and } B_i)$$

$$= \sum_{i=1}^{n} P(E|B_i)\,P(B_i)$$

Bayes' Theorem

$$P(B|E) = \frac{P(E|B) \cdot P(B)}{P(E)}$$

$$P(B|E) = \frac{P(E|B) \cdot P(B)}{P(E|B) \cdot P(B) + P(E|B^{C}) \cdot P(B^{C})}$$

Definition of Conditional Probability

$$P(E|F) = \frac{P(E \text{ and } F)}{P(F)}$$

Axiom 1: $0 \leq P(E) \leq 1$

Axiom 2: $P(S) = 1$

$$P(E^{C}) = 1 - P(E)$$

De Morgan's Laws

$(A \text{ or } B)^C = A^C \text{ and } B^C$

$(A \text{ and } B)^C = A^C \text{ or } B^C$

Axiom 3: If E and F are mutually exclusive, then $P(E \text{ or } F) = P(E) + P(F)$

Otherwise, use Inclusion-Exclusion:

$$P(E \text{ or } F) = P(E) + P(F) - P(E \text{ and } F)$$

Multiplication Rule

$$P(E \text{ and } F) = P(E|F) \cdot P(F)$$

$$= P(F|E) \cdot P(E)$$

Independence

$$P(E|F) = P(E)$$

$$P(E \text{ and } F) = P(E)\,P(F)$$

Road Map

- More Applications of Bayes (lesson 8 slide 46)
- Lesson 9: Independent Events (video assignment in HW 5)

This Lesson:

Intro to Discrete RV

Independent RV

Random Variables

Learning Objectives:

Define random variables.

Explain the difference between random variables and events

Define a discrete random variable in terms of its probability mass function

Use tables, histograms and/or closed-form functions to represent PMFs

Use PMFs to calculate probabilities

Simulate discrete random variables using Python

State the mathematical definition of what it means for 2 random variables to be independent.

Determine whether 2 discrete RV are independent using the mathematical definition

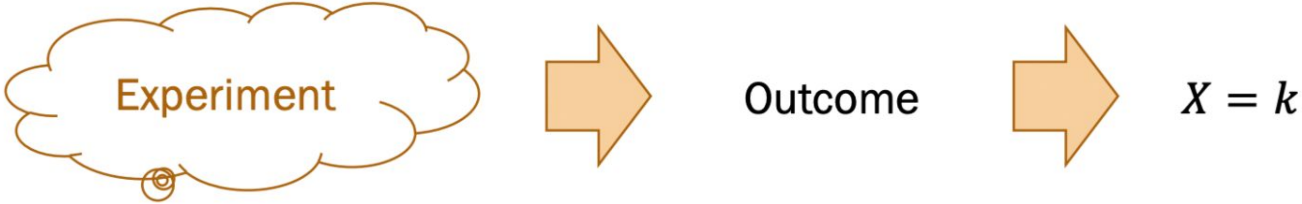
- Introduction to Random Variables
- Discrete Random Variables
    - Plotting Histograms of Probability Mass Functions (PMF)
    - Simulating Distributions of Discrete Random Variables
- Independent Random Variables
    - IID RV

Random Variables

- Introduction to Random Variables

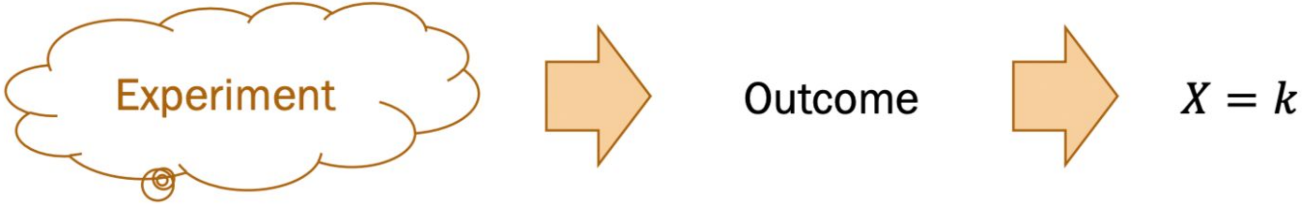
Random Variables

A random variable is a real-valued function defined on a sample space.



Random Variables

A random variable is a real-valued function defined on a sample space.



Example:

3 coins are flipped.
Let $X = \#$ of heads.
X is a random variable.

1. What is the value of X for the outcomes:
   - (T,T,T)?
   - (H,H,T)?
2. What is the event (set of outcomes) where $X = 2$?
3. What is $P(X = 2)$?

Random Variables vs Events

Random variables are NOT events!

It is confusing that random variables and events use the same notation.

- Random variables $\neq$ events.
- We can define an event to be a particular assignment of a random variable, or more generally, in terms of a random variable.

Random variables are NOT events!

It is confusing that random variables and events use the same notation.

- Random variables $\neq$ events.
- We can define an event to be a particular assignment of a random variable, or more generally, in terms of a random variable.

Example:

3 coins are flipped.
Let $X = \#$ of heads.
X is a random variable.

$$X = 2$$

event

$$P(X = 2)$$

probability
(number b/t 0 and 1)

Random variables are NOT events!

It is confusing that random variables and events use the same notation.

- Random variables ≠ events.
- We can define an event to be a particular assignment of a random variable, or more generally, in terms of a random variable.

Example:

3 coins are flipped.
Let X = # of heads.
X is a random variable.

$X = x$	Set of outcomes	$P(X = k)$
$X = \mathbf{0}$	{(T, T, T)}	1/8
$X = \mathbf{1}$	{(H, T, T), (T, H, T), (T, T, H)}	3/8
$X = \mathbf{2}$	{(H, H, T), (H, T, H), (T, H, H)}	3/8
$X = \mathbf{3}$	{(H, H, H)}	1/8
$X \geq 4$	{ }	0

Random Variables & Samples

Suppose we draw a random sample of size n from a population.

A **random variable** is a numerical function of a sample.

sample was drawn at random   value depends on how the sample came out

- Often denoted with uppercase "variable-like" letters (e.g. X, Y).
- Domain (input): all random samples of size n
- Range (output) also called **Support**

- **Definition:** The support of a random variable X is defined as the set of numbers that are possible values of the random variable.

Random Variables & Samples

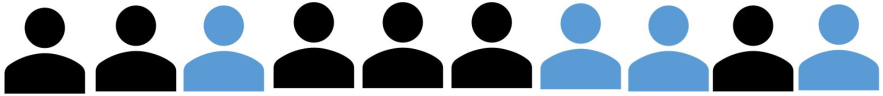
Suppose we draw a random sample of size n from a population.

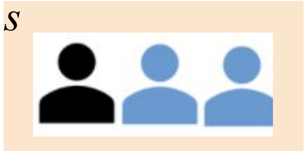
A **random variable** is a **numerical function** of a sample.

sample was drawn at random   value depends on how the sample came out

- Often denoted with uppercase "variable-like" letters (e.g. X, Y).
- Also known as a sample statistic, or **statistic**. (next lecture).
- Domain (input): all random samples of size n
- Range (output) also called **Support:** some subset of the number line

Suppose you draw a random sample s of size 3 from the following population:

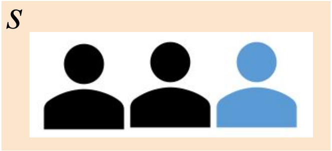




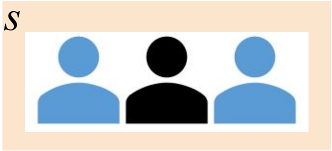
$X(s) = 2$

Define X = **# of blue people**.

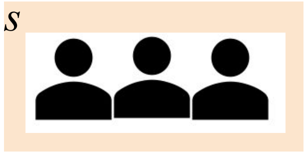
X is a random variable!



$X(s) = 1$



$X(s) = 2$



$X(s) = 0$

Discrete Random Variables

- Introduction to Random Variables
- **Discrete Random Variables**

DISCRETE Random Variables

A random variable X is discrete if it can take on countably many values.

- $X = x$, where $x \in \{x_1, x_2, x_3, \dots\}$

Ex). Which of the following would typically be considered ***discrete*** random variables? Select all that apply.

A). The number of people who check out at a grocery line in a given hour.

B). The finish times of randomly chosen runners from the Bolder Boulder 10K.

C). The number of games played in the best of 7 NBA playoffs.

D). The weight of dogs taken from a random sample around Boulder.

E). The volume of water in randomly chosen Colorado lakes.

[Terminology] Distribution of Discrete Random Variable

The **distribution** of a **DISCRETE** random variable X, is called a **Probability Mass Function (PMF).** It's a description of how the total probability of 100% is split over all the possible values of X.

A distribution fully defines a random variable.

$$P(X = k)$$

The probability that discrete random variable X takes on the value k.

$$\sum_{all\ k} P(X = k) = 1$$

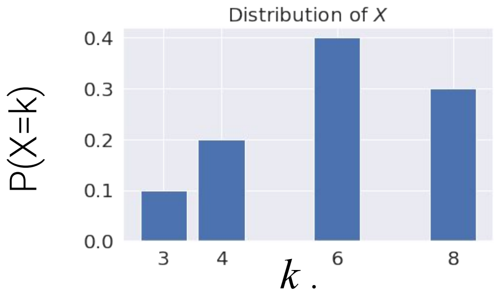
Probabilities must sum to 1.

We can represent a discrete distribution (i.e. the PMF of the random variable) using:

a). **A table**

k	P(X = k)
3	0.1
4	0.2
6	0.4
8	0.3

b) . **A histogram**



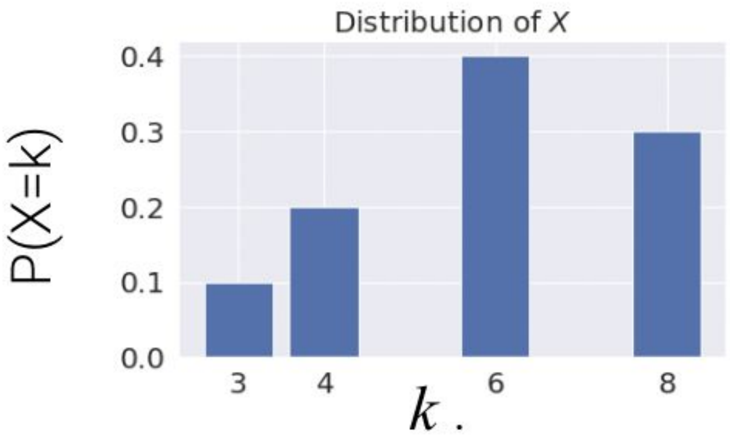
c). (Sometimes) A closed-form function

Understanding Discrete Random Variables

Compute the following probabilities for the random variable X.

1. $P(X = 4) =$
2. $P(X < 6) =$
3. $P(X \leq 6) =$
4. $P(X = 7) =$
5. $P(X \leq 8) =$

k	P(X = k)
3	0.1
4	0.2
6	0.4
8	0.3

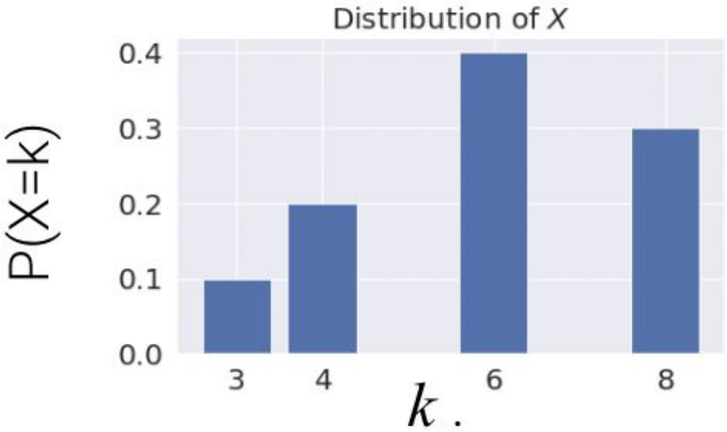


Understanding Discrete Random Variables

Compute the following probabilities for the random variable X.

1. $P(X = 4) =$ 0.2
2. $P(X < 6) =$ $0.1 + 0.2 = 0.3$
3. $P(X \leq 6) =$ $0.1 + 0.2 + 0.4 = 0.7$
4. $P(X = 7) =$ 0
5. $P(X \leq 8) =$ 1

k	P(X = k)
3	0.1
4	0.2
6	0.4
8	0.3



A Whole New World with Random Variables

Event-driven probability

- Relate only binary events
  - Either something happens (E)
  - or it doesn't happen (E^C)
- Can only report probability
- Lots of combinatorics

Random Variables

- Link multiple similar events together ($X = 1,\ X = 2, \ldots, X = 6$)
- Can compute statistics: report the "average" outcome
- Once we have the PMF (for discrete RVs), we can do regular math

Visualizing Distributions of Discrete Random Variables

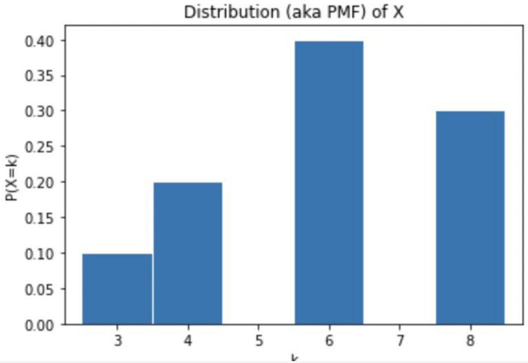
- Plotting Histograms of Probability Mass Functions (PMF)
- Simulating Distributions of Discrete Random Variables

Probability vs Empirical Distributions

Probability (aka Population or Theoretical) Distribution/PMF Function

- All possible values it can take
- The probability it takes each value
  - Often challenging to calculate analytically (the math may not be possible…)

Recall the discrete Random Variable X from the last lecture. Here is the PMF of X:

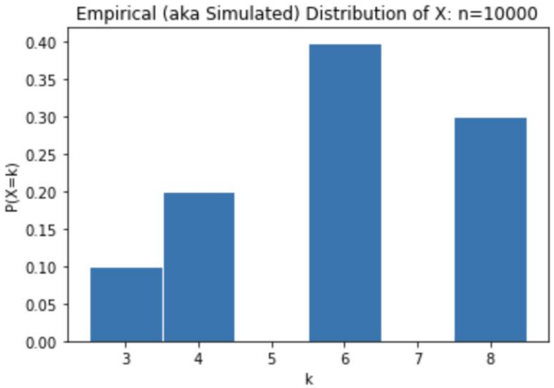



```
x_pmf = pd.Series([0.1, 0.2, 0, 0.4, 0, 0.3], index=[3, 4, 5, 6, 7, 8])
```

```
x_pmf.plot.bar(rot=0,width=1, ec='white')
```

Empirical (aka Simulated or Sample) Distribution:

- Based on random samples (or simulations)
- Observations can be from **repetitions of an experiment or random samples from a population**
  - All observed values
  - The proportion of times each value appears




```
n=10000
sim_data=np.random.choice([3,4,6,8], p=[0.1, 0.2, 0.4, 0.3], size=n)
```

```
plt.hist(sim_data, density=True, bins=np.arange(2.5, 9.5), ec='white')
```

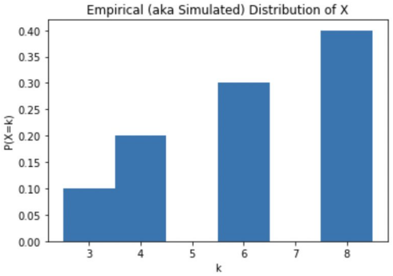
Law of Averages / Law of Large Numbers

If a chance experiment is **repeated many times**,
independently and under the **same conditions**,
then the **Empirical (Sample) Distribution** gets closer to the Theoretical **Probability Distribution.**

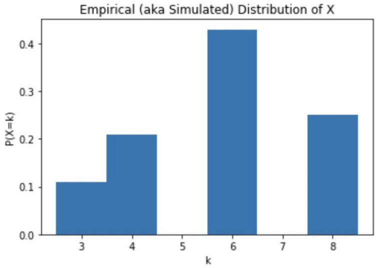
Ex:

```
sim_data=np.random.choice([3,4,6,8], p=[0.1, 0.2, 0.4, 0.3], size=n)
```

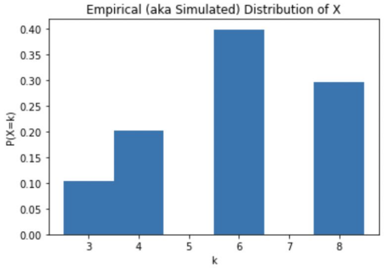
n=10

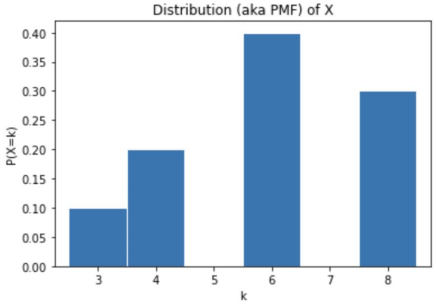


n=100



n=10000





Empirical Probability Distributions

Theoretical Probability Distribution (PMF)

Simulating Distributions

- Any discrete random quantity has a probability distribution:
  - All possible values it can take
  - The probability it takes each value
    - Often challenging to calculate analytically (the math may not be possible…)

- When simulating independent repeated draws, it has an empirical distribution:
  - All observed values it took
  - The proportion of times it took each value

- After many independent draws, the **empirical distribution** looks more and more like the ***probability distribution***

Jupyter NB Demo

Learning Objectives:

- Determine whether 2 discrete RV are independent
- Define IID

Independent RV

Independent RV

Independent Random Variables

Independent events E and F $\Leftrightarrow$ $P(EF) = P(E)P(F)$, $P(E|F) = P(E)$

Recall the definition of independent events E and F:

Two discrete random variables X and Y are independent if:

for all x, y:

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

or $P(X=x \mid Y=y) = \boxed{}$ or $P(Y=y \mid X=x) = \boxed{}$

- Intuitively: knowing value of X tells us nothing about the distribution of Y (and vice versa)
- If two variables are not independent, they are called dependent.

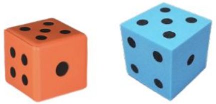
Ex: Testing RV for Independence

Two discrete random variables X and Y are independent if:

for all x, y:
$$P(X = x, Y = y) = P(X = x)P(Y = y)$$

Let: D_1 and D_2 be the outcomes of two rolls
$S = D_1 + D_2$, the sum of two rolls

- Each roll of a 6-sided die is an independent trial.
- Random variables D_1 and D_2 are independent.



Are events $D_1 = 1$ and $S = 7$ independent?

$D_1 = 1$: $\{(1,1), (1,2), (1,3), (1,4), (1,5), (1,6)\}$

$S = 7$: $\{(1,6), (2,5), (3,4), (4,3), (5,2), (6,1)\}$

$$P(D_1 = 1) = \frac{6}{36} = \frac{1}{6}$$

$$P(S = 7) = \frac{6}{36} = \frac{1}{6}$$

$$P(D_1 = 1, S = 7) = \frac{1}{36}$$

✅ independent

Are events $D_1 = 1$ and $S = 5$ independent?

$D_1 = 1$: $\{(1,1), (1,2), (1,3), (1,4), (1,5), (1,6)\}$
$S = 5$: $\{(1,4), (2,3), (3,2), (4,1)\}$

$$P(D_1 = 1) = \frac{6}{36} = \frac{1}{6}$$

$$P(S = 5) = \frac{4}{36} = \frac{1}{9}$$

$$P(D_1 = 1, S = 5) = \frac{1}{36}$$

❌ dependent

Are RANDOM VARIABLES D_1 and S independent?

❌ dependent

All events $(X = x, Y = y)$ must be independent for X, Y to be independent RVs.

Independence of Multiple Discrete Random Variables:

Recall independence of n events $E_1, E_2, \ldots, E_n$:

for $r = 1, \ldots, n$:

for every subset $E_1, E_2, \ldots, E_r$:

$$P(E_1, E_2, \ldots, E_r) = P(E_1)P(E_2)\cdots P(E_r)$$

We have independence of n discrete random variables $X_1, X_2, \ldots, X_n$ if

for $r = 1, \ldots, n$:

for all subsets $x_1, x_2, \ldots, x_r$:

$$P(X = x_1, X = x_2, \ldots, X_r = x_r) = \prod_{i=1}^{r} P(X_i = x_i)$$