<u>Unit 7 Slides</u>: Introduction to Random Variables – Building Blocks for Inference

***<u>Case Study</u>*: Weekly Hours Spent Watching Youtube**

- *How can we theoretically **model the distribution** of the number of hours adults spend watching Youtube each week?*
- *How can we **calculate the** probability that an adult watches a certain range of hours of Youtube each week?*

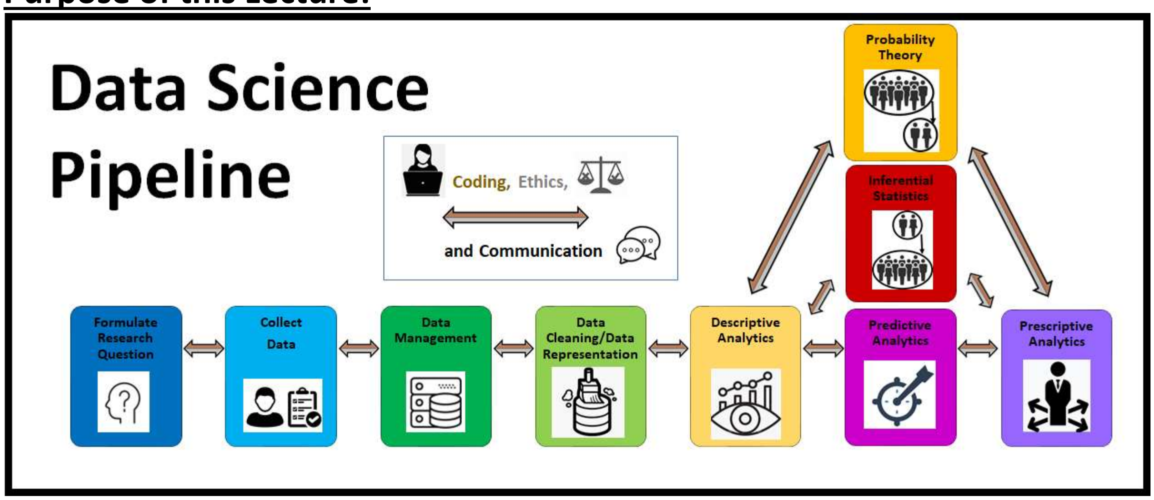
***<u>Case Study</u>*: Coin Flips**

- *What is the **average** number of coin flips it takes until we get a head?*

***<u>Case Study</u>*: Sample of Instagram Users**

- *How many adults in a random sample of 5 adults would we **expect** to use Instagram?*

<u>Purpose of this Lecture:</u>



In this lecture we will cover the following topics.

1. **<u>Goal of Unit 7:</u>**
2. **<u>General Probability Rules</u>**
3. **<u>Random Variable Definitions and Main Types</u>**
   - **Discrete** random variables

- **Continuous** random variables

4. **How to calculate the probability of events involving random variables**
    4.1. More definitions
    4.2. How to **calculate probabilities** involving discrete random variables *"from scratch."*
5. **How to identify if a random variable "fits the definition" of a well-known random variable.**
6. **Discrete Random Variables: Functions that calculate discrete random variable probabilities**
    6.1. Probability Mass Functions (in Python)
    6.2. Cumulative Distribution Functions (in Python)
7. **Examples of randomly generating values for a random variable**
    7.1. When we "don't know" if the random variable "fits the definition" of a well-known random variable (in Python).
    7.2. When we KNOW the random variable "fits the definition" of a well-known random variable. (in Python)
8. **Continuous Random Variables: Functions that calculate continuous random variable probabilities**
    8.1. Why do we not use probability mass functions for continuous random variables?
    8.2. Cumulative Distribution Functions and probability density functions
    8.3. Properties of cumulative distribution functions and probability density functions
    8.4. Calculating the probabilities of events involving random variables using cdf and pdf curves.
    8.5. Calculating the probabilities of events involving well-known random variables in Python.
9. **Calculating Summary Statistics of a Random Variable**
    9.1. By hand
    9.2. In Python
10. Coding:
    10.1. while loops

Additional resources:

- Section 2.1, 2.2, 2.4, and 2.5 in Diez, Barr, and Cetinkaya-Rundel, (2015), *OpenIntro Statistics* https://www.openintro.org/download.php?file=os3&redirect=/stat/textbook/os3.php
- https://www.w3schools.com/python/python_while_loops.asp
- https://www.tutorialspoint.com/scipy/scipy_stats.htm

1. Goals of Unit 7

- **Probability rules that work for any type of events.**
- **What is a random variable?**
- **What are some types of random variables and what are their properties?**
- **How do we calculate probabilities involving a random variable?**

Why do we need to know this?

2. General Probability Rules

Ex*: Suppose we know the following information about the people in this Zoom room today.*

- *The probability of randomly selecting a* ***junior*** *from the Zoom room is __________.*
- *The probability of randomly selecting a* ***statistics major*** *from the Zoom room is __________.*
- *The probability of randomly selecting a* ***junior statistics major*** *from the Zoom room is __________.*

Probability Notation: P(A) represents “probability of event A”.

Set of all Outcomes: If Ω is the set of all possible outcomes of a random experiment, then

$$P(\Omega) = 1.$$

Ex: What is the probability of randomly selecting a freshman, sophomore, junior, OR senior from this Zoom room?

Complementary Events: $P(\text{not A}) = 1 - P(A)$

Ex: What is the probability of randomly selecting someone from this Zoom room that is not a statistics major?

Mutually Exclusive (Disjoint) Events: Two events A and B are called mutually exclusive (or disjoint) if they cannot both happen at the same time, or in other words, $P(\text{A and B}) = 0$

Ex: Are the events of randomly selecting someone from this Zoom room who is both a statistics major AND a junior mutually exclusive?

Union or Events: $P(\text{A or B}) = P(A) + P(B) - P(A \text{ } and \text{ } B)$

Ex: What is the probability of randomly selecting someone from this Zoom room that is either a statistics major OR a junior?

Conditional Probability: The probability of event A happening GIVEN that we know that event B has happened is represented and calculated as follows:

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

Ex: What is the probability of randomly selecting a statistics major from the Zoom room GIVEN that the person is a junior?

Independent Events: If two events A and B are independent, then knowledge of how one of the events turned out does not help us predict the other event. Also the following three equations hold.

$$P(A \text{ and } B) = P(A)P(B)$$
$$P(A|B) = P(A)$$
$$P(B|A) = P(B)$$

Dependent Events: If two events A and B are dependent, then knowledge of how one of the events turned out does help us predict the other event. It's also the case that:

$$P(A \text{ and } B) \neq P(A)P(B)$$
$$P(A|B) \neq P(A)$$
$$P(B|A) \neq P(B)$$

Ex: Are the event of randomly selecting a Junior and the event of randomly selecting a statistics major from this Zoom room independent? Why or why not?

3. Random Variable Definitions and Main Types

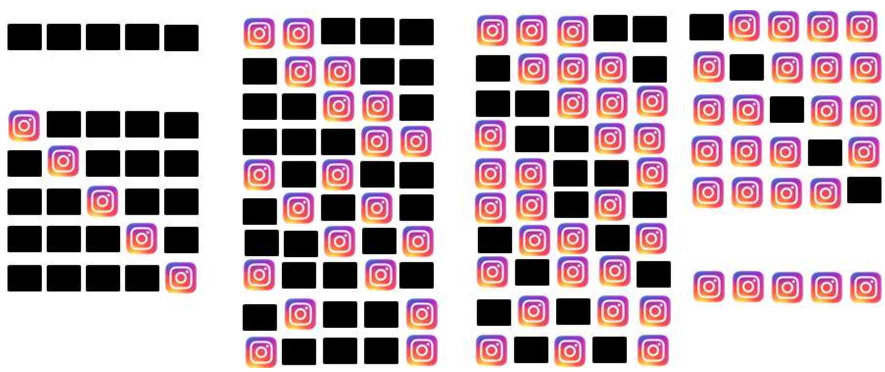
Definition

A **random variable** assigns some ___________________ to each simple event in a sample space.

Example:

About 35% of American adults use Instagram. We decide to collect a random sample of 5 American adults and ask if they use Instagram or not.

Sample Space



<u>Example of a Random Variable that Deals with this Experiment</u>: X = # of randomly selected American adults (out of 5) that are Instagram users.

<u>Ex:</u> Come up with another random variable, call it Y, that involves this experiment.

Main Types of Random Variables: Discrete vs. Continuous

- For **discrete random variables** there exists a way to write out every value the random variable can take on. There exist "gaps" in between the values that they can take on.

  Ex: Is the random variable X (from our Instagram experiment above) discrete? If so, list out every value that this random variable can take on.

- For **continuous random variables** there is no possible way to write out every possible value the random variable can take on.

Example of a Continuous Random Variable

We decide to randomly select an adult female from the population of all adult females.

Sample Space

... many more

Random Variable Associated with this Experiment

X = **height** of the randomly selected adult female from the population of ALL adult females.

Let's *try* to write out EVERY possible value that X could take on.

List of Events	Probabilities
X = 0"	0.00
...	...
X = 5'8"	
X = 5'9"	
...	...
...	...
...	...
...	...
X = 10"	

List of Events	Probabilities
X = 0"	0.00
...	...
X = 5'8"	
X = 5'8.5"	
X = 5'9"	
...	...
...	...
...	...
X = 10"	0.00

AGH!

Set of events is continuous.

Can't write them all out.

List of Events	Probabilities
X = 0"	0.00
...	...
X = 5'8"	
X = 5'8.53"	
X = 5'8.5"	
X = 5'9"	...
...	...
...	...
X = 10"	0.00

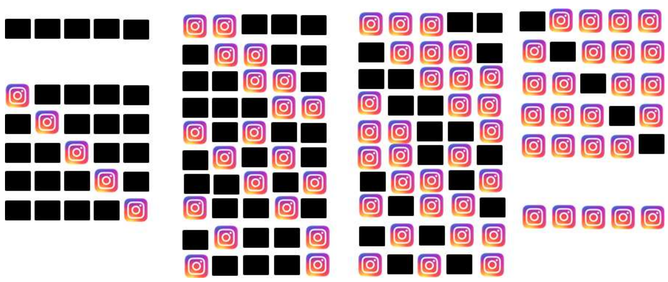
4. How to calculate the probability of events involving random variables.

Examples of Events Involving any Random Variable X

- Event: "$X = number$"
- Event: "$X < number$"
- Event: "$X > number$"
- Event: "$number_1 \leq X \leq number_2$"
- Event: "$number_1 < X < number_2$"
- Event: "$number_1 < X \leq number_2$"
- Event: "$number_1 \leq X < number_2$"

Ex 1: About 35% of American adults use Instagram. We decide to collect a random sample of 5 American adults and ask if they use Instagram or not. Let the random variable **X = the number of adults in the sample that use Instagram.**

a. Show all of the simple events that comprise the compound event "$X = 2$".



b. Can we use the uniform probability model rules to calculate $P(X = 2)$? Why or why not?

Two Common ways to Calculate Probabilities Involving a Random Variable

1. Does the random variable "fit the definition" of **random variable that is well-known**? If so, these well-known random variables come with formulas for calculating probabilities involving these random variables.

2. If not, we can try to calculate the probability involving this random variable "from scratch"

Calculating a Probability Involving a Random Variable "from Scratch"

Ex 2: Define a random variable **X = number of times you flip a coin until you get a head.**

- List a few simple events from the sample space of this experiment. Are there a finite number of events in this sample space?

- Calculate $P(X = 1)$.

- Suppose we're considering the event "$X=2$". Does the outcome of the first coin flip affect influence the probability that we will get a HEAD (instead of a TAIL) in the second coin flip?

- Calculate $P(X = 2)$.

- Calculate $P(X = 3)$.

- What is the probability that $P(X = k)$ for any k=1,2,3,…

5. How to identify if a random variable "fits the definition" of a well-known random variable.

One **Well Known Random Variable and it's Properties:**

Definition: A **geometric random variable** = the number of ____________________ trials of an experiment in it takes to get a _________________________________, where the probability of every trial being a____________________ is ***p.***

Probability Mass Function: Y is a random variable if and only if

$$\boldsymbol{p(k) = P(Y = k) = (1 - p)^{k-1} p} \quad \textbf{for k=1,2,3,...}$$

We'll learn about more well-known random variables in the next unit!

Ex: Define a random variable X = number of times you flip a coin until you get a head.

- Is X a geometric random variable?

    a. If so, what is considered a "trial" for X?

    b. Are the trials independent?

    c. What is considered a "success" for X?

    d. What is "*p*"*(ie. the probability a trial is a success)* for X?

Describing a Well-Known Random Variable:

Most Well-Known Types of Random Variables Have some **Additional Corresponding Parameter Values** that are important to it's definition.

Ex: For instance, geometric random variables always have a corresponding "p" parameter associated with them, which describes the probability that one of the independent trials is a "success".

We would use **notation shorthand** to describe the specific type of geometric random variable that X is (dictated by this p parameter) as follows.

6. Discrete Random Variables: Functions that Calculate Random Variable Probabilities

6.1. Probability Mass Functions

A **probability mass function (pmf)** is a function that gives the probability that a discrete random variable is *exactly* equal to some value.

Ie. $p(value) = P(Y = value)$

Ex: Define a random variable X = number of times you flip a coin until you get a head.

- Give the probability mass function of X.

- Go to Unit 7 notebook (section 6.1) to use the **.pmf()** function to calculate $P(X = 1). P(X = 2). and\ P(X = 3)$.

6.2. Cumulative Distribution Functions

A **cumulative distribution function (cdf)** is a function that gives the probability that a discrete random variable is *less than or equal to* some value.

Ie. $F(value) = P(Y \leq value)$

Ex: Define a random variable X = number of times you flip a coin until you get a head.

- Calculate $P(X \leq 2)$

- Go to Unit 7 notebook (section 6.2) to use the **.pmf()** function to calculate $P(X \leq 2)$

7. Examples of how to randomly generate values for a random variable.

7.1. Example of randomly generating values for a random variable

When we "don't know" if the random variable "fits the definition" of a well-known random variable.

Ex: Define a random variable **X = number of times you flip a coin until you get a head.**

- Go to Unit 7 notebook (section 6.1) to simulate this coin flip experiment (ie. keep flipping until you get a head). By doing so, we can randomly generate values for X.

7.2. Example of randomly generating values for a random variable

When we know that the random variable "fits the definition" of a well-known random variable.

- Go to Unit 7 notebook (section 5) to simulate this coin flip experiment (ie. keep flipping until you get a head). By doing so, we can randomly generate values for X. **This time we will use the helpful information that X is a random variable to do this.**

8. Continuous Random Variables: Functions that Calculate Random Variable Probabilities

8.1. Why do we not use probability mass functions P(Y=value) for continuous random variables?

Ex: X = height of the randomly selected adult female from the population of ALL adult females.

- What is $P(X = 5'8")$?

In general:

For any continuous random variable Y, $P(Y = any\ number)$=__________________, because…

8.2. Cumulative Density Function (cdf) and Probability Density Function (pdf)

A **cumulative distribution function (cdf)** is a function that gives the probability that a discrete random variable is *less than or equal to* some value.

Ie.$F(value) = P(X \leq value)$

- For a continuous random variable X we specifically define the **cumulative density function (cdf) of X, $\boldsymbol{F(value)}$** as

$$F(value) = P(X \leq value) = \int_{-\infty}^{value} f(x)dx$$

$$= \ area\ under\ the\ f(x)\ curve\ to\ the\ left\ of\ value$$

- We call $f(x)$ **the probability density function (pdf) of X**.

8.3. Properties of Cumulative Density Functions (cdf) and Probability Density Functions (pdf)

1. The total area under the pdf curve = _______________

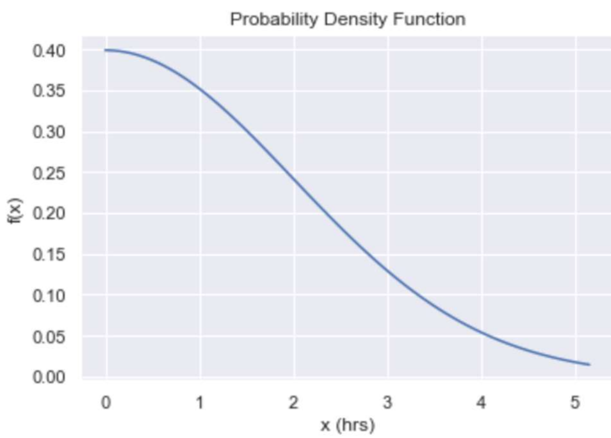
2. $P(X \leq value) = P(X < value)$

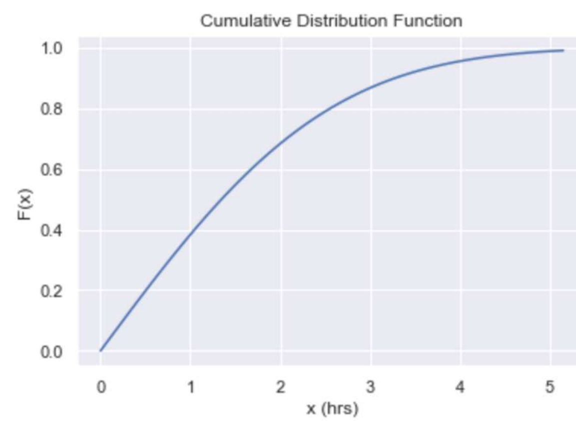
3. $P(X \geq value) = P(X > value)$

4. $P(a < X \leq b) = F(b) - F(a)$ (if a<b)

8.4 Calculating the probabilities of events involving random variables using pdf and cdf curves.

Ex: Let the random variable X = the number of hours a randomly selected adult spends watching Youtube each week. Suppose we know the pdf and the cdf for X shown below.





a. Use the pdf plot to *approximate* $P(\mathrm{X} \leq 2)$.

b. Use the pdf plot to *approximate* $P(\mathrm{X} > 2)$.

c. Use the cdf plot to determine $P(\mathrm{X} \leq 2)$.

d. Use the cdf plot to determine $P(\mathrm{X} > 2)$.

e. Use the cdf plot to determine $P(1 < \mathrm{X} \leq 3)$.

8.5 Calculating the probabilities of events involving <u>well-known</u> random variables in Python.

<u>Ex:</u> Suppose that after collecting data on the Youtube watching habits on a large sample of adults, researchers decided that the random variable **X = the number of hours a randomly selected adult spends watching Youtube each week closely** "fits the definition" of another well-known random variable called a **truncated normal random variable.**

A **truncated normal random variable** has four parameters that are associated with it:

- μ = mean of the random variable (had it not been truncated)
- σ = standard deviation of the random variable (had it not been truncated)
- a = lower bound of the random variable
- b = upper bound of the random variable

Suppose that the researchers specifically know the parameters associated with our X truncated random variable are μ=0, $\sigma = 2$, a=0, and b=20.

Using this information, go to Unit 7 notebook (section 8.5) to calculate the following.

a. $P(\mathrm{X} \leq 2)$

b. $P(\mathrm{X} > 2)$

9. Calculating Summary Statistics of a Random Variable

Ex: Suppose we conducted many, many coin flip experiments, where for each experiment we stopped after we flipped a head.

Population	
Experiment	**Number of Flips Until Stopping**
1	1
2	3
2	2
3	1
5	1
...	...
...	...
9999	2
10000	4

- What **percentage** of experiments would we expect in this population to have ended after 1 flip?

- What **percentage** of experiments would we expect in this population to have ended after 2 flips?

- What **percentage** of experiments would we expect in this population to have ended after 3 flips?

- What **percentage** of experiments would we expect in this population to have ended after at most two flips?

- What would we expect the **median** number of flips until stopping to be?

- What would we expect the **Q3** number of flips until stopping to be?

- What would we expect the **mean** number of flips until stopping to be? Describe how you would set up this equation.

- What would we expect the **standard deviation** number of flips until stopping to be? Describe how you would set up this equation.

9.1. Calculating a Summary Statistic of a Random Variable – "by hand"

Notation and Formal Definitions for Calculating Some Random Variable Summary Statistics by Hand

If X is a random variable, then we denote and calculate the following summary statistics of the random variable in the following way.

- **Median of a Random Variable X ("by hand"):**
    - If the value of *m* in which:
        - $P(X \leq m) = 0.5$ AND
        - $P(X \geq m) = 0.5$

- **qth Percentile of a Random Variable X ("by hand"):**
    - If q is a percentile, then the qth percentile of the random variable is the value of *m* in which:
        - $P(X \leq m) = q$ AND
        - $P(X \geq m) = q$

- **Mean of a Random Variable X ("by hand"):**
    - Also called the ________________________ value of X.
    - If X is a **discrete random variable**:
        - $\mu = E[X] = \sum_i x_i p(x_i) = x_1 P(X = x_1) + x_2 P(X = x_2) + x_3 P(X = x_3) + \cdots$
    - If X is a **continuous random variable**:
        - $\mu = E[X] = \int_{-\infty}^{\infty} x f(x) dx$

- **Mean of a Function of a Random Variable g(X) ("by hand"):**
    - Also called the ________________________ value of g(X).
    - If X is a **discrete random variable**:
        - $E[g(X)] = \sum_i f(g(x_i) p(x_i) = g(x_1) P(X = x_1) + g(x_2) P(X = x_2) + g(x_3) P(X = x_3) + \cdots$
    - If X is a **continuous random variable**:
        - $E[g(X)] = \int_{-\infty}^{\infty} g(x) f(x) dx$

- **Standard Deviation of a Random Variable X ("by hand"):**
    - If X is a **discrete random variable**:

- $\sigma = \sqrt{E[X-\mu]^2} = \sqrt{\sum_i (x_i-\mu)^2 p(x_i)} = \sqrt{(x_1-\mu)^2 P(X=x_1) + (x_2-\mu)^2 P(X=x_2) + (x_3-\mu)^2 P(X=x_3) + \cdots}$
- If X is a **continuous random variable**:
  - $\sigma = \sqrt{E[X-\mu]^2} = \sqrt{\int_{-\infty}^{\infty} (x-\mu)^2 f(x)dx}$
- **Variance of a Random Variable X ("by hand"):**
  - If X is a **discrete random variable**:
    - $\sigma^2 = E[X-\mu]^2 = \sum_i (x_i-\mu)^2 p(x_i) = (x_1-\mu)^2 P(X=x_1) + (x_2-\mu)^2 P(X=x_2) + (x_3-\mu)^2 P(X=x_3) + \cdots$
  - If X is a **continuous random variable**:
    - $\sigma^2 = E[X-\mu]^2 = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)dx$

Ex: Let X = # of adults in a random sample of size 5 that use Instagram. Below is the probability table detailing the probability for each possible value for X.

Random Variable X	P(X = #)
X=0	0.12
X=1	0.31
X=2	0.34
X=3	0.18
X=4	0.05
X=5	0.01

1. Find the mean of X. What does this represent?

2. Find the variance of X. What does this represent?

3. Find the standard deviation of X. What does this represent?

9.2. Calculating a Summary Statistic of a Random Variable – in Python

Go to the Unit 7 notebook (section 9.2) for how to calculate summary statistics of <u>well-known random variables</u> in Python.

10. <u>Coding</u>: while loops

Go to the Unit 7 notebook (section 7.1) for how to calculate summary statistics of <u>well-known random variables</u> in Python.