<u>Unit 8 Slides</u>: Common Types of Random Variables and How to use Them – Building Blocks for Inference

<u>Case Study</u>: Heights

- ***Height*** *tend can often be modeled as a* ***normal random variable.*** *How can we use this fact to calculate the probability a randomly selected person is a certain height?*

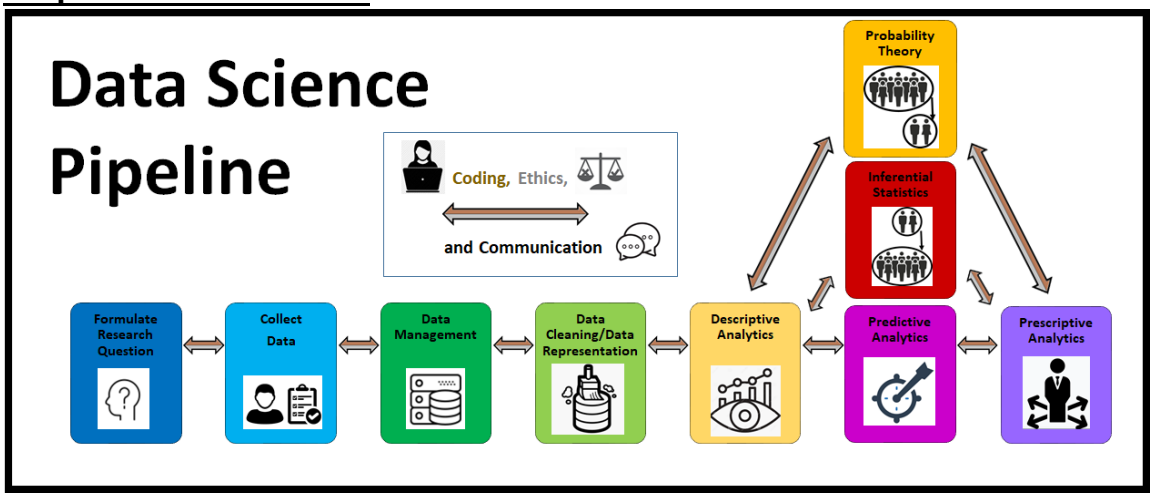
<u>Case Study</u>: Coin Flips

- *How can we model the outcomes of a sample of coin flip experiments (stop flipping when you get a head)?*

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

- *If we randomly sampled many, many adults, what would we expect the* ***average, variance, and standard deviation*** *number to use Instagram would be?*

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



In this lecture we will cover the following topics.

1. <u>A Well-Known Type of Discrete Random Variable</u>: **Bernoulli Random Variable**
   1.1. Definition of a Bernoulli random variable

1.2. Mean and Standard deviation of a Bernoulli random variable
2. A Well-Known Type of Continuous Random Variable: **Normal Random Variable**
   2.1. Definition
   2.2. Mean, Variance, and Standard Deviation
   2.3. Other Properties
   2.4. How do you KNOW when a distribution is approximately normal?
   2.5. Calculating probabilities involving normal random variables – In Python
   2.6. Calculating percentiles involving normal random variables – In Python
3. Z-scores
   3.1. Definitions
   3.2. Relationship between random variable and Z-score of random variable.
4. A Well-Known Type of Continuous Random Variable: Standard Normal Random Variable
   4.1. Definition
   4.2. Relationship between the z-score of a NORMAL random variable and the standard normal random variable.
   4.3. Calculating probabilities involving standard normal random variables – In Python
   4.4. Calculating percentiles involving standard normal random variables – In Python
   4.5. Using standard normal random variables to help solve problems involving normal random variables. – In Python
5. Representing a sample statistic with multiple random variables
   5.1. Sample statistics in the coin flip experiment in Python

Additional resources:

Section 3.1-3.3 in Diez, Barr, and Cetinkaya-Rundel, (2015), *OpenIntro Statistics* https://www.openintro.org/download.php?file=os3&redirect=/stat/textbook/os3.php

1. A well-known type of discrete random variable: Bernoulli Random Variable

***Bernoulli Random Variable*:**

Definition: A **bernoulli random variable X** corresponds to a simple experiment in which there are only two outcomes (that we observe).

- Outcome 1: "Success" (ie. outcome that _____).
  - If the outcome is a success, then X = _____.
  - The probability of a success is known to be **p.**
- Outcome 2: "Failure" (ie. outcome that _____).
  - If the outcome is a failure, then X = _____.
  - The probability of a failure is known to be **1-p.**

Short-Hand: _____

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

$$\boldsymbol{p(1) = P(X = 1) = p}$$

$$\boldsymbol{p(0) = P(X = 0) = 1 - p}$$

Ex: Suppose we know that 35% of adults use Instagram. We randomly select an adult and see if they use Instagram.

a. Set up a random variable, X, that models the outcome of this experiment. Is this random variable a Bernoulli random variable?

b. Calculate the expected value (ie. mean) of X, $E[X]$.

c. Calculate the variance X, $V[X]$.

d. Calculate the standard deviation of X, ie. $SD[X]$.

e. What does $E[X]$, $V[X]$, and $SD[X]$ represent in this case?

Experiment	Outcome	X
Experiment 1	randomly sampled adult uses Insta	1
Experiment 2	randomly sampled adult doesn't use Insta	0
Experiment 3	randomly sampled adult doesn't use Insta	0
Experiment 4	randomly sampled adult uses Insta	1
...	...	...
...	...	...
Experiment 10,000	randomly sampled adult doesn't use Insta	0

***Bernoulli Random Variable*:**

Mean, Variance, and Standard Deviation: If X is a **Bernoulli random variable** then

$$\boldsymbol{E[X] = p}$$

$$\boldsymbol{V[X] = p(1-p)}$$

$$\boldsymbol{SD[X] = \sqrt{p(1-p)}}$$

2. A well-known type of continous random variable: Normal Random Variable

2.1. Normal Random Variable: Definition

***Normal Random Variable*:**

Definition: A continuous random variable is said to be a **normal random variable** if it has the following probability density function (pdf).

Short-Hand: _____

Probability Density Function: X is a **normal random variable** if and only if

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-u)^2}{2\sigma^2}}, for -\infty < x < \infty$$

2.2. Normal Random Variable: Mean, Variance, and Standard Deviation

***Normal Random Variable*:**

Mean, Variance, and Standard Deviation: If X is a **normal random variable**

$$\boldsymbol{E[X] = \mu}$$

$$\boldsymbol{V[X] = \sigma^2}$$

$$\boldsymbol{SD[X] = \sigma}$$

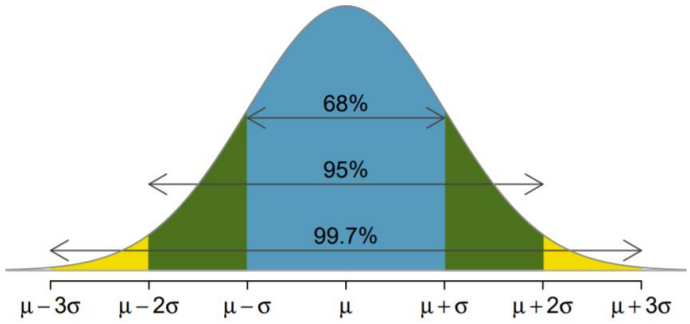
	X
Experiment 1 Randomly sample a woman from population (with replacement)	69"
Experiment 2 Randomly sample a woman from population (with replacement)	62"
Experiment 3 Randomly sample a woman from population (with replacement)	64"
...	...
Experiment 50,000 Randomly sample a woman from population (with replacement)	70"

2.3. Normal Random Variable: Other Properties

Normal Random Variable:

Properties:

- Pdf Shape: _____.
- **Pdf follows the "68-95-99.7 Rule"**
  - A distribution that is approximately normal will have:
    - _____ of it's observations $\mu \pm \sigma$ (ie. within one standard deviation away from the mean),
    - _____ of it's observations $\mu \pm 2\sigma$ (ie. within two standard deviations away from the mean),
    - _____ of it's observations $\mu \pm 3\sigma$ (ie. within three standard deviation away from the mean.)



Ex: The average height of a woman in the U.S. is about 64" with a standard deviation of 2.5". Assume that the heights of women in the U.S. has a normal distribution.

a. What percent of adult females in the U.S are taller than 69"?

b. Can we use the 68-95-99.7 rule to approximate percent of adult females in the U.S are taller than 70”? Why or why not?

2.4. Normal Random Variable: How do you KNOW when a distribution is approximately normal?

- **Some simpler ways to *approximate*.**
  - **Common examples**
    - _____
  - **Does the distribution have the following specifications?**
    - _____
    - _____
    - _____
- **Theory (ie. prove it)!**
  - **Central Limit Theorem (later Unit)** will prove to us that **sampling distributions** are approximately normal ***when certain conditions are met.***
- **More advanced ways to *approximate* (out of scope)**
  - **qq-plots**
  - **Chi-Squared Test for Normality**

2.5. Normal Random Variable: Calculating probabilities involving normal random variables.

Cumulative Distribution Function (cdf) for Normal Random Variables

$$F(\boldsymbol{x}) = P(X \leq \boldsymbol{x})$$

$$= \int_{-\infty}^{\boldsymbol{x}} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(y-u)^2}{2\sigma^2}} dy$$

$$= \text{area under the } f(y) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(y-u)^2}{2\sigma^2}} \text{ curve to the left of } \boldsymbol{x}$$

This is impossible to calculate by hand, so we must use computer/calculator functions to approximate it as best we can.

Go to the unit 8 notebook (section 2.5) to learn how to use Python functions to help answer the following questions below.

Ex: The average height of a woman in the U.S. is about 64" with a standard deviation of 2.5". Assume that the heights of women in the U.S. has a normal distribution.

a. What percentage of women in the U.S. are **at most** 70"? What percentage of women in the U.S. are **shorter than** 70"? If we randomly selected a woman in the U.S. what is the probability that her height is less than 70"?

b. If we randomly selected a woman in the U.S. what is the probability her height is **at least** 70”?

c. If we randomly selected a woman in the U.S. what is the probability her height is **between** 60” and 70”?

2.6. Normal Random Variable: Calculating percentiles involving normal random variables.

Go to the unit 8 notebook (section 2.6) to learn how to use Python functions to help answer the following questions below.

Ex: The average height of a woman in the U.S. is about 64" with a standard deviation of 2.5". Assume that the heights of women in the U.S. has a normal distribution.

d. How tall does a woman in the U.S. have to be in order to be in the top 10% of women's heights?

3. z-scores

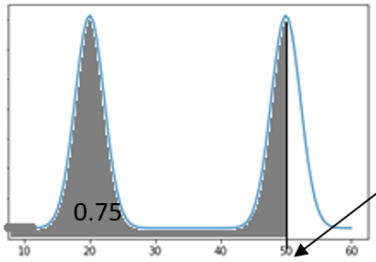
Ex:

Population: babies and teenagers
Distribution: heights (in)

- Mean = 35”
- Standard deviation = 5”

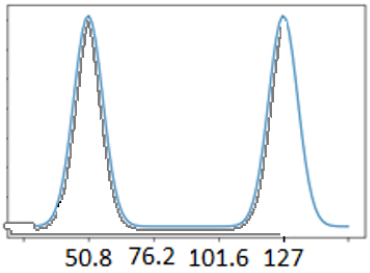
Joe is taller than 75% of the population when measured in inches. If we converted *all* heights to centimeters, would Joe still be taller than 75% of the population?

Probability Density Function



Heights (in)

Probability Density Function



Heights (cm)

3.1. z-score: Definitions

Z-score of an Observation

The **z-score** of an observation x from a population with a given mean and standard deviation is:

$$z - score\ of\ x = \frac{x - mean\ of\ population}{standard\ deviation\ of\ population}$$

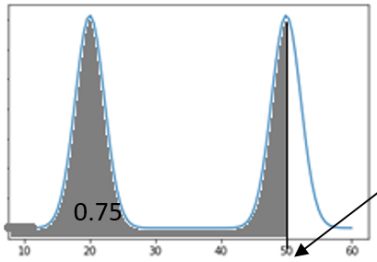
Ex:

Population: babies and teenagers
Distribution: heights (in)

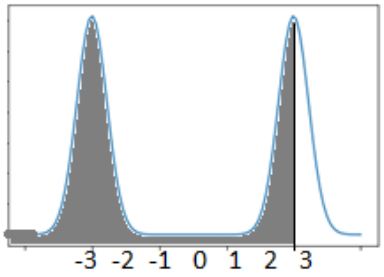
- Mean = 35”
- Standard deviation = 5”

Joe is taller than 75% of the population when measured in inches. If we converted *all* heights to z-scores, would Joe’s z-score be larger than the z-scores of 75% of the population?

Probability Density Function



Heights (in)



Z-scores of the heights

Suppose Joe is 50” tall. What is the z-score of his height?

Z-score of a Random Variable X

The **z-score** of a random variable X is:

$$z - score\ of\ X = Z = \frac{X - E[X]}{SD[X]}$$

3.2. Z-score: Relationship between a Random Variable X and the z-score of X

In general:

If X is a random variable and $Z = \frac{X - E[X]}{SD[X]}$, then

$$P(X \leq x) = P(Z \leq \frac{x - E[X]}{SD[X]})$$

Ex: Joe 50" tall and is a member of a population that has an average height of 35" and a standard deviation height of 5".

4. A well-known type of continuous random variable: Standard Normal Random Variable

4.1. Standard Normal Random Variable: Definition

***Standard Normal Random Variable*:**

Definition: A **normal random variable** is said to be a **standard normal random variable** if it has:

- mean = _____
- standard deviation = _____

Short-Hand: _____

Probability Density Function: X is a **standard normal random variable** if and only if

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}, for -\infty < x < \infty$$

4.2. Relationship between the z-score of a normal random variables and a standard normal random variable

If X is a normal random variable and $Z = \frac{X - E[X]}{SD[X]}$

then $Z = \frac{X - E[X]}{SD[X]}$ is a _____.

$$P(X \leq x) = P(Z \leq \frac{x - E[X]}{SD[X]})$$

Ex: Suppose X is a normal random variable and $Z = \frac{X - E[X]}{SD[X]}$.

a. What is $E[Z]$?

b. What is $SD[Z]$?

4.3. Calculating probabilities involving standard normal random variables – in Python.

Go to the unit 8 notebook (section 4.3) to learn how to use Python functions to help answer the following questions below.

Ex: What is the probability that a standard normal random variable is between -1.96 and 1.96?

4.4. Calculating percentiles involving standard normal random variables – in Python.

Go to the unit 8 notebook (section 4.4) to learn how to use Python functions to help answer the following questions below.

Ex: What is the value in the standard normal distribution that is greater than 20% of observations?

4.5 Using standard normal random variables to help solve problems involving normal random variables – in Python.

Go to the unit 8 notebook (section 4.5) to learn how to use Python functions to help answer the following questions below.

Ex: The average time it took for all seniors at a local high school to finish a race was 8 minutes. The finishing times followed a normal distribution.

You ran the race in 7 minutes and your time was in the top 20% (ie. your time was higher than 20% of the participants). What was the standard deviation of the finishing times?

5. Representing a Sample Statistic with Multiple Random Variables

Ex: Suppose we have defined the following experiment and random variable as follows.

Experiment: Randomly select an adult female from a population.

A Random Variable for this Experiment: X = height of the *single* randomly selected woman.

Now suppose we set up a new experiment.

New Experiment: Randomly selecting (with replacement) n=3 adult females from the population.

How could we define a random variable that represents the mean height of the three women in the sample?

- Experiment 1: Randomly select an adult female from a population.
  - A Random Variable for this Experiment: X_1 = height of the randomly selected woman.
- Experiment 2: Randomly select an adult female from a population.
  - A Random Variable for this Experiment: X_2 = height of the randomly selected woman.
- Experiment 3: Randomly select an adult female from a population.
  - A Random Variable for this Experiment: X_3 = height of the randomly selected woman.

A Random Variable for this New Experiment:

Sample Statistics

Mathematicatical notation using random variables

Remembering that X is a random variable representing one draw from the population distribution, we represent the sample data as a collection of n *independent* random draws from the population:

$$\text{Sample:} \quad X_1, X_2, X_3, \ldots, X_n.$$

Using this notation the sample statistics analogous to the population parameters above can be described mathematically and operationally as follows. In each case the generic form of a pandas function call is given. Substitute the name of data frame for 'df' and the name of target variable for 'x'.

- **Sample mean:**

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{X_1 + X_2 + X_3 + \cdots + X_n}{n}$$

- **Sample variance:**

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

- **Standard deviation:**

$$S = \sqrt{S^2}$$

- **Proportion ≤ 2:**

$$\hat{p} = \frac{1}{n}\sum_{i=1}^{n} 1\{X_i \leq 2\} = \frac{\#\{X_i \leq 2\}}{n}$$

- **Median:** Split the sorted data in half

1. Represent sorted data as $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$
2. Middle value if n is odd:

$$M = X_{\left(\frac{n+1}{2}\right)}$$

3. Average of two middle values if n is even:

$$M = \frac{X_{\left(\frac{n}{2}\right)} + X_{\left(\frac{n}{2}+1\right)}}{2}$$

5.1. Sample Statistics of the Coin Flip Experiment in Python

Ex: Consider our coin toss experiment from Unit 7. This is where we keep flipping a coin until we get a head.

We let $X = number\ of\ flips\ until\ stopping$.

We also learned that $X \sim Geom(p = 0.5)$.

Suppose we decided to repeat this coin flip experiment multiple 20 times, we can represent

$$X_i = number\ of\ flips\ in\ the\ ith\ experiment\ until\ stopping$$

Go to the unit 8 notebook (section 5.1) to generate a random value for each X_i. Use these values to generate a random values for:

- $\bar{X} = \frac{X_1 + X_2 + \cdots + X_{20}}{20}$: the mean number of flips until stopping from the 20 experiments
- M: the median number of flips until stopping from the 20 experiments
- S: the standard deviation number of flips until stopping from the 20 experiments
- P_2: the proportion of the 20 experiments in which the number of flips (until stopping) was at most 2.

Also use these values to generate a random sample distribution.