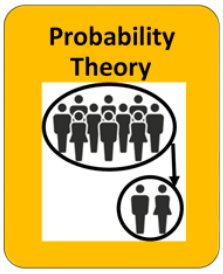
Unit 4: Sampling and Probability

***Case Study*: UIUC Couse Data and Playing Cards**

We will learn:

- Basics of Probability



Summary of Concepts:

- **How is probability used in data science pipeline**
  - Relationship between inferential statistics and probability theory
- **Probability Definitions**
  - Experiment
  - Simple and Compound Events
  - Sample Space
  - Uniform Probability Distributions
  - Probability (two definitions)
  - Law of Large Numbers
- **Types of Sampling**
  - With replacement
  - Without replacement
- **Event Types**
  - Independent events
  - Dependent events
- **Calculating certain types of probabilities**
  - When calculating the probability of a *single event* (simple or compound) and the sample space events are equally likely.
    - *Using permutation counting equations*
    - *Using combination counting equations*
  - When calculating the probability of a *multiple events* (simple or compound) and the two events are *independent*.
  - When calculating the probability of a *multiple events* (simple or compound) and the two events are *dependent*.
- **Python Coding Useful for Sampling**
  - DATAFRAME**.sample()**
  - DATAFRAME**.repeat()**
- **General Python Coding**
  - **for** loops
  - **+=** operator
  - **if then** statements
  - pd**.Series()**
  - **all()** function

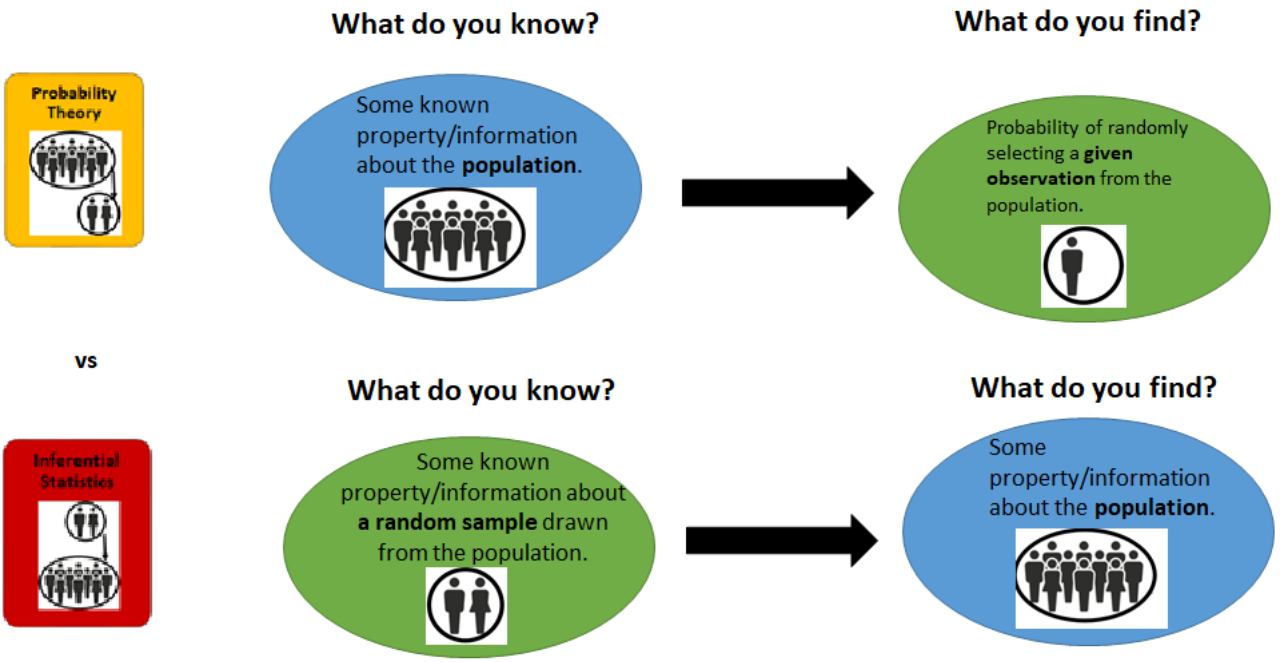
How is probability used in data science?

Definitions

Population: large body of data that is our target of interest

Sample: a subset of data selected from the population

Observation: a single "row" of data selected from the population (ie. sample of size n=1).



Probability Definitions

An _____ is the process by which an **observation** is made. *(Most often we are interested in observations made from completely **uncontrollable situations.)***

Ex:

	course	section	enrolled
0	cs105	B	345
1	cs105	A	201
2	stat107	A	197
3	stat207	A	53
4	badm210	A	215
5	badm210	C	197
6	badm210	B	178
7	adv307	A	37

When an experiment is performed it can result in one or more outcomes, which are called _____.

Ex:

A _____ event can be decomposed into several events, while a _____ event cannot.

The _____ (S) associated with an experiment is the set consisting of all possible simple events (ie. sample points).

Types of Random Sampling

We can think of collecting a random sample (of n observations) from a population as conducting an experiment where we observe which observations were selected.

- **Randomly Sampling _____** from a population means that observations can only be selected and put the sample _____. Once we randomly select an observation, we do not "put the observation back" into the population.

- **Randomly Sampling _____** from a population means that observations CAN show up in the sample multiple times. Once we randomly select an observation, we "put it back into the population" where it *can* be randomly selected again.

Ex: Suppose we conducted two experiments. First we randomly selected a sample of size n=2 from our population of 7 courses in the dataframe with replacement. Then we randomly selected a sample of size n=2 from our population of 7 courses in the dataframe without replacement. Which of the two sample spaces below corresponds to the random sampling experiment conducted a.) with replacement and b.) without replacement?

	course	section	enrolled
0	cs105	B	345
1	cs105	A	201
2	stat107	A	197
3	stat207	A	53
4	badm210	A	215
5	badm210	C	197
6	badm210	B	178
7	adv307	A	37

{0, 0}	{2, 0}	{4, 0}	{6, 0}
{0, 1}	{2, 1}	{4, 1}	{6, 1}
{0, 2}	{2, 2}	{4, 2}	{6, 2}
{0, 3}	{2, 3}	{4, 3}	{6, 3}
{0, 4}	{2, 4}	{4, 4}	{6, 4}
{0, 5}	{2, 5}	{4, 5}	{6, 5}
{0, 6}	{2, 6}	{4, 6}	{6, 6}
{0, 7}	{2, 7}	{4, 7}	{6, 7}
{1, 0}	{3, 0}	{5, 0}	{7, 0}
{1, 1}	{3, 1}	{5, 1}	{7, 1}
{1, 2}	{3, 2}	{5, 2}	{7, 2}
{1, 3}	{3, 3}	{5, 3}	{7, 3}
{1, 4}	{3, 4}	{5, 4}	{7, 4}
{1, 5}	{3, 5}	{5, 5}	{7, 5}
{1, 6}	{3, 6}	{5, 6}	{7, 6}
{1, 7}	{3, 7}	{5, 7}	{7, 7}

	{2, 0}	{4, 0}	{6, 0}
{0, 1}	{2, 1}	{4, 1}	{6, 1}
{0, 2}		{4, 2}	{6, 2}
{0, 3}	{2, 3}	{4, 3}	{6, 3}
{0, 4}	{2, 4}		{6, 4}
{0, 5}	{2, 5}	{4, 5}	{6, 5}
{0, 6}	{2, 6}	{4, 6}	
{0, 7}	{2, 7}	{4, 7}	{6, 7}
{1, 0}	{3, 0}	{5, 0}	{7, 0}
	{3, 1}	{5, 1}	{7, 1}
{1, 2}	{3, 2}	{5, 2}	{7, 2}
{1, 3}		{5, 3}	{7, 3}
{1, 4}	{3, 4}	{5, 4}	{7, 4}
{1, 5}	{3, 5}		{7, 5}
{1, 6}	{3, 6}	{5, 6}	{7, 6}
{1, 7}	{3, 7}	{5, 7}	

Two Definitions of Probability

	course	section	enrolled
0	cs105	B	345
1	cs105	A	201
2	stat107	A	197
3	stat207	A	53
4	badm210	A	215
5	badm210	C	197
6	badm210	B	178
7	adv307	A	37

- **Probability of an event**: measure of one's belief in the occurrence of this event
- **Probability of an event**: the _____ of times the outcome would occur if we observed the random process _____ of times.

Ex 1: Let's say we collected a random sample of size n=1000 (with replacement).

a.) What proportion of our sample would you expect to be STAT207?

b.) What proportion of our sample would you expect to be a statistics class?

Ex 2: Let's say we collected a random sample of size n=1,000,000 (with replacement).

a.) What proportion of our sample would you expect to be STAT207?

b.) What proportion of our sample would you expect to be a statistics class?

c.) Would you expect the ACTUAL proportions found in example 2 to be closer, farther away, or the same distance to to our predictions than the ACTUAL proportions found in example 1?

Calculating Certain Types of Probabilities

Compound events from uniform probability distributions

In the original data frame from which we sampled there were 8 rows each of which had the same chance of being selected. In this case we say that the row probabilities are all the same, and thus **uniform.**

If we select one row at random, this implies that each row has a 1/8 = 0.125 chance of being selected.
Uniform probability rule: If we make a random draw from a set of nn possible choices, and each choice has *the same probability of selection*, then each outcome has probability 1/n of occurring.

Notice that, in our example, the course selections themselves are *not* uniformly distributed because they appear in different numbers of rows. Instead their probabilities are given by the proportions of times they appear in the original data frame. Courses that appear only once in the list have probability 1/8 of being selected. Courses that appear more than once have higher probabilities of selection. This observation leads to our second rule about uniform probability distributions.

Rule for calculating event probabilities from uniform probability distributions: If an event of interest includes k of the n possible choices in a random draw from a set, then the probability of the event is k/n.

True or False

Suppose we flip a coin twice. The probability of getting at least one head is ¾.

True or False

Suppose we drive down a street with two traffic lights, where the probability of getting a green for each light is 0.6. The probability of getting at least one green on the road is ¾.

Calculating Certain Types of Probabilities

Independent Events or Random Sampling with Replacement

- Two events are _____ if the outcome of one event does not influence the outcome of the other event (and vice versa).
- If events A and B are _____, then P(A and B) = _____.

- The outcomes of two observations randomly sampled from a population with replacement are _____.

Ex: What is the probability of randomly selecting two statistics courses from our dataset (population) *with replacement?*

	course	section	enrolled
0	cs105	B	345
1	cs105	A	201
2	stat107	A	197
3	stat207	A	53
4	badm210	A	215
5	badm210	C	197
6	badm210	B	178
7	adv307	A	37

{0, 0}	{2, 0}	{4, 0}	{6, 0}
{0, 1}	{2, 1}	{4, 1}	{6, 1}
{0, 2}	{2, 2}	{4, 2}	{6, 2}
{0, 3}	{2, 3}	{4, 3}	{6, 3}
{0, 4}	{2, 4}	{4, 4}	{6, 4}
{0, 5}	{2, 5}	{4, 5}	{6, 5}
{0, 6}	{2, 6}	{4, 6}	{6, 6}
{0, 7}	{2, 7}	{4, 7}	{6, 7}
{1, 0}	{3, 0}	{5, 0}	{7, 0}
{1, 1}	{3, 1}	{5, 1}	{7, 1}
{1, 2}	{3, 2}	{5, 2}	{7, 2}
{1, 3}	{3, 3}	{5, 3}	{7, 3}
{1, 4}	{3, 4}	{5, 4}	{7, 4}
{1, 5}	{3, 5}	{5, 5}	{7, 5}
{1, 6}	{3, 6}	{5, 6}	{7, 6}
{1, 7}	{3, 7}	{5, 7}	{7, 7}

Calculating Certain Types of Probabilities

Independent Events or Random Sampling without Replacement

- Two events are _____ if the outcome of one event DOES influence the outcome of the other event (or vice versa).
- If events A and B are _____, then P(A and B) ≠ _____.

- The outcomes of two observations randomly sampled from a *small* population without replacement are _____.

Ex: What is the probability of randomly selecting two statistics courses from our dataset (population) *without replacement?*

Sample Space (where Order Matters)

	{2, 0}	{4, 0}	{6, 0}
{0, 1}	{2, 1}	{4, 1}	{6, 1}
{0, 2}		{4, 2}	{6, 2}
{0, 3}	{2, 3}	{4, 3}	{6, 3}
{0, 4}	{2, 4}		{6, 4}
{0, 5}	{2, 5}	{4, 5}	{6, 5}
{0, 6}	{2, 6}	{4, 6}	
{0, 7}	{2, 7}	{4, 7}	{6, 7}
{1, 0}	{3, 0}	{5, 0}	{7, 0}
	{3, 1}	{5, 1}	{7, 1}
{1, 2}	{3, 2}	{5, 2}	{7, 2}
{1, 3}		{5, 3}	{7, 3}
{1, 4}	{3, 4}	{5, 4}	{7, 4}
{1, 5}	{3, 5}		{7, 5}
{1, 6}	{3, 6}	{5, 6}	{7, 6}
{1, 7}	{3, 7}	{5, 7}	

Sample Space (where Order Doesn't Matter)

{0, 1}			
{0, 2}			
{0, 3}	{2, 3}		
{0, 4}	{2, 4}		
{0, 5}	{2, 5}	{4, 5}	
{0, 6}	{2, 6}	{4, 6}	
{0, 7}	{2, 7}	{4, 7}	{6, 7}
{1, 2}			
{1, 3}			
{1, 4}	{3, 4}		
{1, 5}	{3, 5}		
{1, 6}	{3, 6}	{5, 6}	
{1, 7}	{3, 7}	{5, 7}	

	course	section	enrolled
0	cs105	B	345
1	cs105	A	201
2	stat107	A	197
3	stat207	A	53
4	badm210	A	215
5	badm210	C	197
6	badm210	B	178
7	adv307	A	37

Combinatorics Equations to Help Us Count

1. **Number of ways to select a set of *k* items <u>with replacement</u> from a population of size n and <u>order/pattern/sequence matters</u>.**

<u>Ex:</u>

{0, 0}	{2, 0}	{4, 0}	{6, 0}
{0, 1}	{2, 1}	{4, 1}	{6, 1}
{0, 2}	{2, 2}	{4, 2}	{6, 2}
{0, 3}	{2, 3}	{4, 3}	{6, 3}
{0, 4}	{2, 4}	{4, 4}	{6, 4}
{0, 5}	{2, 5}	{4, 5}	{6, 5}
{0, 6}	{2, 6}	{4, 6}	{6, 6}
{0, 7}	{2, 7}	{4, 7}	{6, 7}
{1, 0}	{3, 0}	{5, 0}	{7, 0}
{1, 1}	{3, 1}	{5, 1}	{7, 1}
{1, 2}	{3, 2}	{5, 2}	{7, 2}
{1, 3}	{3, 3}	{5, 3}	{7, 3}
{1, 4}	{3, 4}	{5, 4}	{7, 4}
{1, 5}	{3, 5}	{5, 5}	{7, 5}
{1, 6}	{3, 6}	{5, 6}	{7, 6}
{1, 7}	{3, 7}	{5, 7}	{7, 7}

2. **Number of ways to select a set of *k* items <u>without replacement</u> from a population of size n and <u>order/pattern/sequence matters</u>.**

<u>Ex:</u>

	{2, 0}	{4, 0}	{6, 0}
{0, 1}	{2, 1}	{4, 1}	{6, 1}
{0, 2}		{4, 2}	{6, 2}
{0, 3}	{2, 3}	{4, 3}	{6, 3}
{0, 4}	{2, 4}		{6, 4}
{0, 5}	{2, 5}	{4, 5}	{6, 5}
{0, 6}	{2, 6}	{4, 6}	
{0, 7}	{2, 7}	{4, 7}	{6, 7}
{1, 0}	{3, 0}	{5, 0}	{7, 0}
	{3, 1}	{5, 1}	{7, 1}
{1, 2}	{3, 2}	{5, 2}	{7, 2}
{1, 3}		{5, 3}	{7, 3}
{1, 4}	{3, 4}	{5, 4}	{7, 4}
{1, 5}	{3, 5}		{7, 5}
{1, 6}	{3, 6}	{5, 6}	{7, 6}
{1, 7}	{3, 7}	{5, 7}	

3. **Number of ways to select a set of *k* items <u>without replacement</u> from a population of size n and <u>order/pattern/sequence DOESN'T MATTER</u>.**

<u>Ex:</u>

{0, 1}			
{0, 2}			
{0, 3}	{2, 3}		
{0, 4}	{2, 4}		
{0, 5}	{2, 5}	{4, 5}	
{0, 6}	{2, 6}	{4, 6}	
{0, 7}	{2, 7}	{4, 7}	{6, 7}
{1, 2}			
{1, 3}			
{1, 4}	{3, 4}		
{1, 5}	{3, 5}		
{1, 6}	{3, 6}	{5, 6}	
{1, 7}	{3, 7}	{5, 7}	

Calculating Certain Types of Probabilities

Compound events from uniform probability distributions AND using combinatorics equation to help count the numerator and denominator of the proportion.

Ex: What's the probability of randomly drawing a full house?

***Ex*: We toss a coin 10 times. One possible sequence of heads and tails with 6 tails is THTTTHTTHH. How many possible sequences of heads and tails are there with exactly 6 tails?**