Unit 6 Slides: *Creating* Sampling Distributions – Building Blocks for Inference

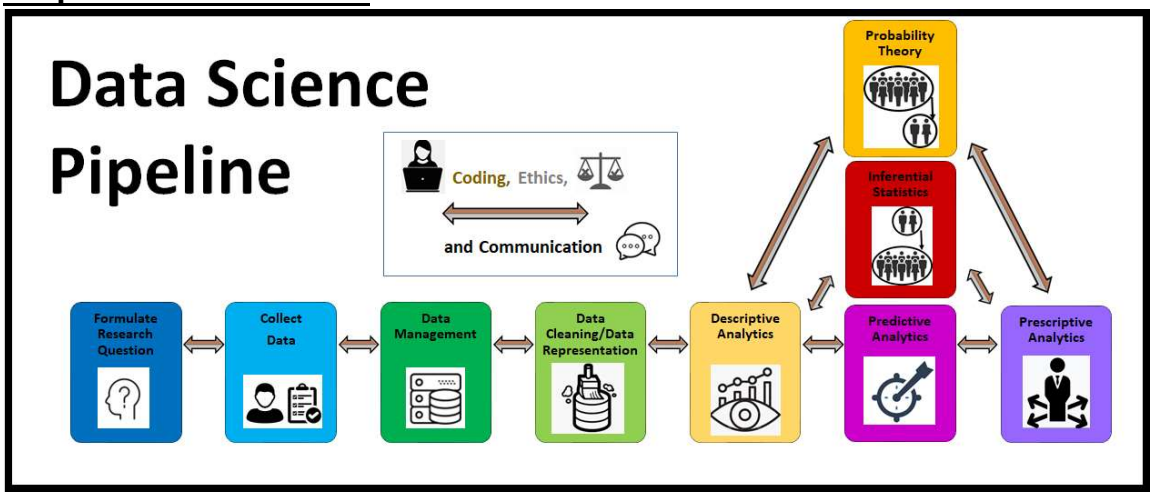
***Case Study*: UIUC Couse Enrollments – Create a Sampling Distribution of Sample Mean Enrollments**

- *We will consider the **enrollment numbers** of the 8 classes from our artificial UIUC course dataset to be a **population of numerical data.***
- *Suppose we collect many, many random samples (with replacement) from this population and calculate the **mean of each sample.** How will this **distribution of sample means** behave?*

***Case Study*: Coin Flip Outcomes – Create a Sampling Distribution of Sample Proportions of Heads**

- *We will consider the two **outcomes (heads or tails)** of a coin flip to be a **population of categorical data.***
- *Suppose we collect many, many random samples (with replacement) from this population and calculate the **proportion of heads in each sample.** How will this **distribution of sample proportion** behave?*

Purpose of this Lecture:



In this lecture we will cover the following topics.

1. **Notation: Summary statistics for Populations vs. Samples**
   - 1.1. Population mean
   - 1.2. Sample mean
   - 1.3. Population proportion
   - 1.4. Sample proportion
2. **Definition: Sampling Distribution**

2.1. What is a sampling distribution? How to create one?
2.2. Type of Sampling Distribution: Sampling Distribution of Sample Means
2.3. Type of Sampling Distribution: Sampling Distribution of Sample Proportions
2.4. What do we want to know about **sampling distributions** as the size of the samples (n) changes in order to **make an inference** about the *unknown* **population parameter**?

3. **Type of Sampling Distribution: Sampling Distribution of Sample Means – How to create one?**
4. **More about for loops**
5. **Case Study: Sampling Distribution of Sample Means – What happened to the mean, spread, and shape of the sampling distribution as we increased the size of the samples n?**
   5.1. Population and population mean
   5.2. Creating a sampling distribution of sample means in Python with samples of size n=10
   5.3. Creating a sampling distribution of sample means in Python with samples of size n=100
   5.4. Creating a sampling distribution of sample means in Python with samples of size n=400
   5.5. What happened to the mean, spread, and shape of the sampling distribution as we increased the size of the samples n?
6. **Type of Sampling Distribution: Sampling Distribution of Sample Proportions – How to create one?**
7. **Case Study: Sampling Distribution of Sample Proportions – What happened to the mean, spread, and shape of the sampling distribution as we increased the size of the samples n?**
   7.1. Creating a sampling distribution of sample proportions in Python with samples of size n=10
   7.2. Creating a sampling distribution of sample proportions in Python with samples of size n=100
   7.3. Creating a sampling distribution of sample proportions in Python with samples of size n=400
   7.4. What happened to the mean, spread, and shape of the sampling distribution as we increased the size of the samples n?

Additional resources:

- Section 4.1 in Diez, Barr, and Cetinkaya-Rundel, (2015), *OpenIntro Statistics* https://www.openintro.org/download.php?file=os3&redirect=/stat/textbook/os3.php

---

Color-Schema Note: The color-schema in these Unit 6 slides (from this point forward) does not relate to the data science pipeline color-schema.

We'll use this instead.

- **Population related terms**
- **Sample related terms**
- **Sample SIZE (n) related terms**
- **Sampling distribution related terms**

1. Notation: Summary Statistics for Populations vs. Samples

Populations vs. Samples

Populations are usually too large to collect completely, so the data and summary statistics we can collect about the populations is usually unknown. We can have different types of populations.

Population 1: Comprised of Numerical Data

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

What kind of **summary statistics (population parameters)** can we use to summarize this **numerical population** data?

If we take a **random sample of size *n*** what kind of summary statistic can we use to summarize this **numerical sample data?**

Population 2: Comprised of Categorical Data

	toss	value
0	heads	1
1	tails	0

What kind of **summary statistics (population parameters)** can we use to summarize this **categorical population** data?

If we take a **random sample of size *n*** what kind of summary statistic can we use to summarize this **categorical sample data?**

2. Definition: Sampling Distribution

2.1. What is a sampling distribution?

If we collect many, many **random samples** from a **population of data** (each drawn

_____), where each sample is of the _____, then the

_____ is the distribution of _____.

2.2. Type of Sampling Distribution: Sampling Distribution of Sample Means

Ex: The _____ is a **numerical** distribution of

_____ of random samples drawn from a population of numerical data with

replacement.

2.3. Type of Sampling Distribution: Sampling Distribution of Sample Proportions

Ex: The _____ is a **numerical** distribution of

_____ of random samples drawn from a population of categorical data with

replacement.

2.4. What do we want to know about sampling distributions as the size of the samples (n) changes in order to make an inference about the *unknown* population parameter?

In order to we make an **inference** about an unknown **population parameter,** there are usually three things we are interested in knowing about the corresponding sampling distributions:

1. _____
2. _____
3. _____

3. Type of Sampling Distribution: Sampling Distribution of Sample Means – *How to create one?*

Population of Numerical Data

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

Collect Many Random Samples (all of size n=10) drawn with replacement.

Random Sample of n=10 Course Enrollments *(drawn with replacement from population)*	**Random Sample of n=10 Course Enrollments** *(drawn with replacement from population)*	**Random Sample of n=10 Course Enrollments** *(drawn with replacement from population)*	**...**	**Random Sample of n=10 Course Enrollments** *(drawn with replacement from population)*
197	215	53	...	215
37	215	53	...	197
345	53	197	...	37
201	201	53	...	215
178	53	197	...	197
37	345	197	...	197
53	197	178	...	345
201	37	215	...	345
201	201	197	...	197
197	37	197	...	201

Sampling Distribution of Sample Means

Sample Means
164.7
155.4
153.7
...
214.6

Need to know the following to make an inference about *Unknown* Population Mean μ:

1. Mean of Sampling Distribution
2. Standard deviation of Sampling Distribution
3. Shape of Sampling Distribution

ALSO Need to know the following to make an inference about *Unknown* Population Mean μ:

1. Mean of Sampling Distribution *as size of samples (n) changes.*
2. Standard deviation of Sampling Distribution *as size of samples (n) changes.*
3. Shape of Sampling Distribution *as size of samples (n) changes.*

4. More about for loops

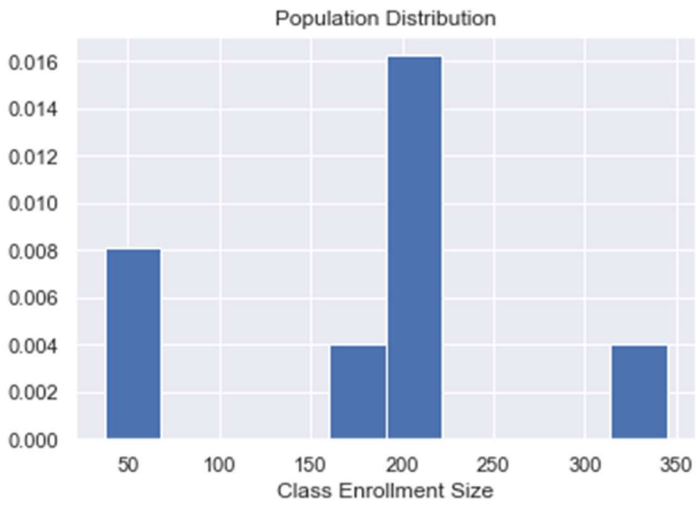
See Unit 6 notebook (section 4).

5. Type of Sampling Distribution: Sampling Distribution of Sample Means – What happens to the mean, standard deviation, and shape of the sampling distribution of sample means as we increase the size of the samples n?

5.1. First, what is the mean, standard deviation, and shape of the population distribution of enrollments?

See Unit 6 notebook (section 5.1) for code.

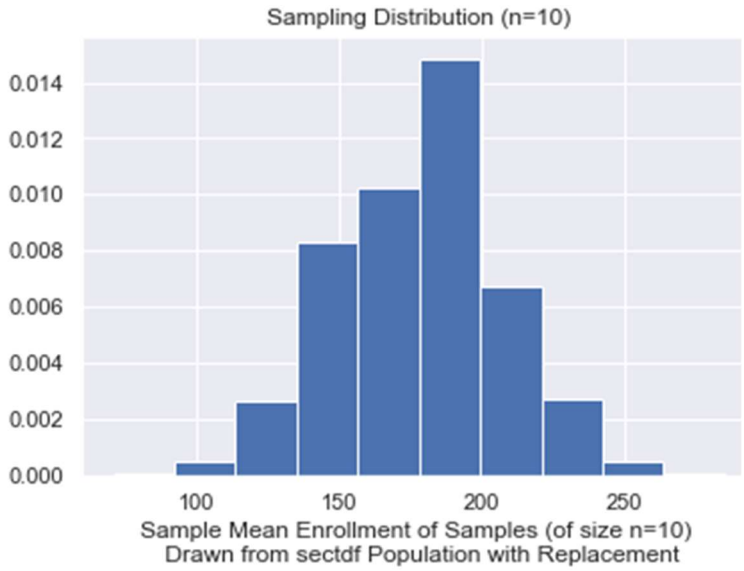
	course	section	enrolled
0	adv307	A	37
1	badm210	A	215
2	badm210	B	178
3	badm210	C	197
4	cs105	A	345
5	cs105	B	201
6	stat107	A	197
7	stat207	A	53



5.2. Create a sampling distribution of sample means drawn from samples of size n=10. What is the mean, standard deviation, and shape of the sample means in this sampling distribution?

See Unit 6 notebook (section 5.2) for code.

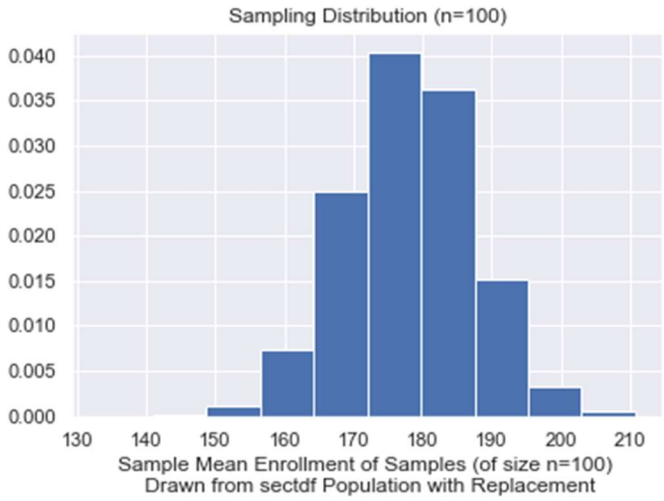
	enrolled
0	176.8
1	223.3
2	227.3
3	163.5
4	225.4
...	...
9995	141.2
9996	211.2
9997	215.0
9998	209.5
9999	162.3



5.3. Create a sampling distribution of sample means drawn from samples of size n=100. What is the mean, standard deviation, and shape of the sample means in this sampling distribution?

See Unit 6 notebook (section 5.3) for code.

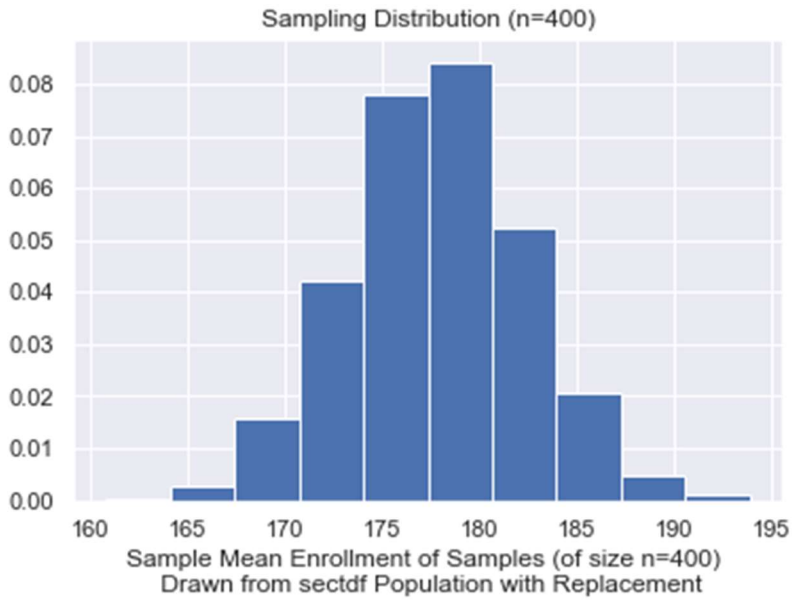
	enrolled
0	167.68
1	163.65
2	174.75
3	184.11
4	178.84
...	...
9995	158.56
9996	175.84
9997	185.53
9998	163.86
9999	188.51



5.4. Create a sampling distribution of sample means drawn from samples of size n=400. What is the mean, standard deviation, and shape of the sample means in this sampling distribution?

See Unit 6 notebook (section 5.4) for code.

	enrolled
0	175.2025
1	176.8450
2	184.4900
3	171.1325
4	171.7225
...	...
9995	179.1975
9996	183.8800
9997	171.6350
9998	182.2200
9999	184.0800



5.5. What did we observe about the sampling distribution of sample means as our sample size n increased?

In an upcoming unit, we will learn about the Central Limit Theorem, which will prove these observations.

6. Type of Sampling Distribution: Sampling Distribution of Sample Proportions – *How to create one?*

Population of Categorical Data

	toss	value
0	heads	1
1	tails	0

Collect Many Random Samples (all of size n=10) drawn with replacement.

Random Sample of n=10 Tosses *(drawn with replacement from population)*	Random Sample of n=10 Tosses *(drawn with replacement from population)*	Random Sample of n=10 Tosses *(drawn with replacement from population)*	…	Random Sample of n=10 Tosses *(drawn with replacement from population)*
1	1	1	…	0
0	1	0	…	0
1	1	1	…	0
0	1	0	…	0
1	1	0	…	1
0	1	0	…	0
0	1	1	…	0
1	1	1	…	0
0	1	0	…	1
1	0	0	…	1

Sampling Distribution of Sample Proportions

Sample Proportions
0.5
0.9
0.4
…
0.3

Need to know the following to make an inference about *Unknown* Population Proportion p:

1. Mean of Sampling Distribution
2. Standard deviation of Sampling Distribution
3. Shape of Sampling Distribution

ALSO Need to know the following to make an inference about *Unknown* Population Population Proportion p:

1. Mean of Sampling Distribution *as size of samples (n) changes.*
2. Standard deviation of Sampling Distribution *as size of samples (n) changes.*
3. Shape of Sampling Distribution *as size of samples (n) changes.*

7. Type of Sampling Distribution: Sampling Distribution of Sample Proportions – What happens to the mean, standard deviation, and shape of the sampling distribution of sample proportions as we increase the size of the samples n?

7.1. First, what is the proportion of heads of the population of coin outcomes?

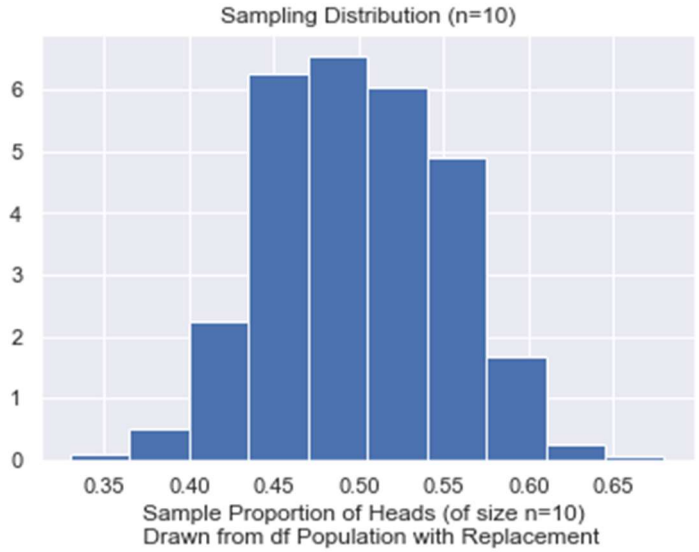
See Unit 6 notebook (section 7.1) for code.

	toss	value
0	heads	1
1	tails	0

7.2. Create a sampling distribution of sample proportions drawn from samples of size n=10. What is the mean, standard deviation, and shape of the sample proportions in this sampling distribution?

See Unit 6 notebook (section 7.2) for code.

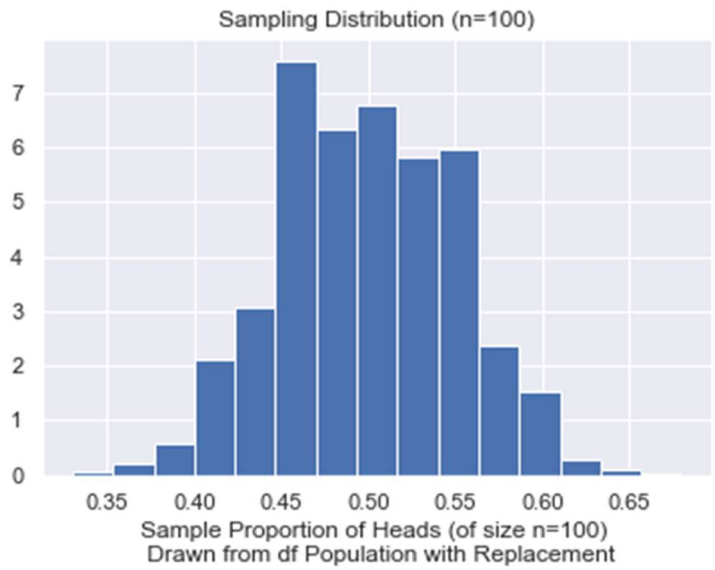
	value
0	0.2
1	0.6
2	0.5
3	0.7
4	0.5
...	...
9995	0.4
9996	0.3
9997	0.4
9998	0.3
9999	0.7



7.3. Create a sampling distribution of sample proportions drawn from samples of size n=100. What is the mean, standard deviation, and shape of the sample proportions in this sampling distribution?

See Unit 6 notebook (section 7.3) for code.

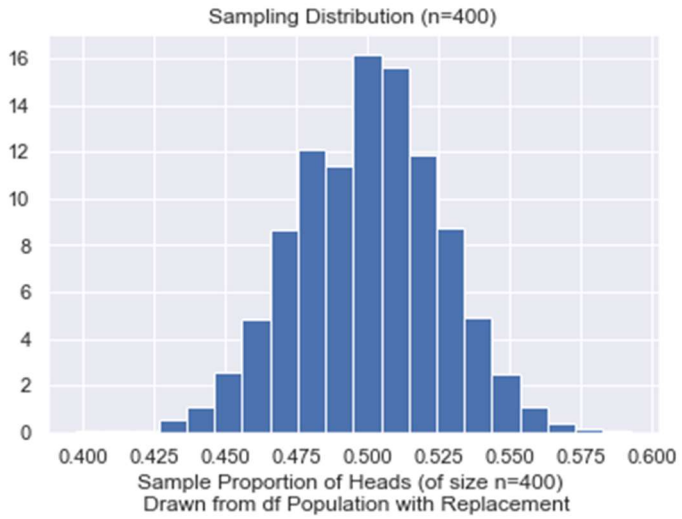
	value
0	0.52
1	0.49
2	0.52
3	0.51
4	0.52
...	...
9995	0.55
9996	0.51
9997	0.55
9998	0.51
9999	0.50



7.4. Create a sampling distribution of sample proportions drawn from samples of size n=400. What is the mean, standard deviation, and shape of the sample proportions in this sampling distribution?

See Unit 6 notebook (section 7.4) for code.

	value
0	0.4950
1	0.5625
2	0.5075
3	0.5175
4	0.4600
...	...
9995	0.4675
9996	0.5150
9997	0.5225
9998	0.4650
9999	0.4475



7.5. What did we observe about the sampling distribution of sample proportions as our sample size n increased?

In an upcoming unit, we will learn about the Central Limit Theorem, which will prove these observations.