M5-06: Finding Confidence Intervals in Python

Part of the "Polling, Confidence Intervals, and the Normal Distribution" Learning Badge

Video Walkthrough: https://discovery.cs.illinois.edu/m5-06/

Confidence Intervals with the GPA Dataset

A study that we conducted at The University of Illinois was to find the average class size by sampling the courses offered at Illinois.

Experiment #1: You and Three Friends

If you participate in this study, **you and three of your friends** would find the total size of each of the five classes you are enrolled in the current semester. We can simulate a possible set of results through using the GPA dataset:

`df = pd.read_csv("gpa.csv")`
`sample = df.sample(n=20)`
`sample`
	Displays the sample DataFrame, which consists of 20 courses sampled from the GPA dataset.

`sample["Students"].mean()`	39.8
`sample["Students"].std()`	6.8

The central limit theorem tells that the sum or average of a distribution will be normal, so we can model this as a normal distribution in Python:

Python:	
Description:	A normal distribution with mean=39.8 and std=6.8.

Given a distribution, Python can find the range of a given level of confidence. For example:

Python:	`D.interval(0.68)`
Description:	Returns the 68% confidence interval.
Output:	

Puzzle #1: What is the 95% confidence interval?

M5-06: Finding Confidence Intervals in Python

Part of the "Polling, Confidence Intervals, and the Normal Distribution" Learning Badge
Video Walkthrough: https://discovery.cs.illinois.edu/m5-06/

Experiment #2: One Hundred Courses Surveyed

Twenty of your other friends now got together and wanted to predict the total class size at Illinois using the data you collected. We can simulate this by taking **100** random rows for the GPA dataset using **`df.sample(n=100)`** and finding some basic statistics on the sample:

`df = pd.read_csv("gpa.csv")`
`sample = df.sample(n=100)`
`sample`
	Displays the sample DataFrame, which consists of 20 courses sampled from the GPA dataset.

`sample["Students"].mean()`	50.91
`sample["Students"].std()`	7.1

Puzzle #2: What is the confidence we have that the true average class size is 50.91 ± 7.1 students?

Analysis: When the two groups meet up and compare the results, a shocking result was found:

- With n=20, the _______ confidence interval was ±_______ students.
- With n=100, the _______ confidence interval was ±_________ students. **(!!)**

This result is unexpected -- a larger sample should provide a smaller confidence interval!

Puzzle #3: What are the **three** factors that determine the confidence interface?

Puzzle #4: What are possible reasons that may explain why the interval grew larger between our **n=20** sample and our **n=100** sample?

Puzzle #5: What is the actual average class size, based on the GPA dataset?