Central Limit Theorem Presentation

Explanation of the Central Limit Theorem

- The Central Limit Theorem is an important statistical concept that says that if you select sufficient random samples from a population the distribution of the sample means will be approximately normal and centered at the population mean.
- As the size of the random samples increases, the resulting sampling distribution of the means will become narrower.
- This result is remarkable since it holds regardless of the shape of the underlying population distribution.

Set Up the Example

- Let's say that we are curious about how many pieces are in a lego set, on average.
- How do we find out?
- Well we have lego set data scraped from the Lego website. Therefore, we have access to the entire population of Lego sets available and we can calculate the average amount of pieces directly using Python package pandas which has a function called 'describe' that gives us statistics about a set of data. We see that the mean is 493 and the standard deviation is 825.
- Since we have access to the population, we are also able to see the distribution of the data. Using the Python package matplotlib, we can create a histogram of the data and plot a vertical line at the mean.
- We want the x-axis to be the different values for the number of pieces. The y-axis is the density corresponding to the different x-values.
- We can see that this distribution is heavily right skewed, with a long tail.

Explanation of how we will use the CLT

- What if we did not have access to this information about the population?
- Let's say the Lego website is down for example and as a result we cannot get the data we need.
- Instead, we can invoke the Central Limit Theorem to estimate the population mean by taking repeated samples from the population and finding the distribution of the sample means. We know that even though the underlying population is skewed, our sampling distributions of the means should approach normal as we increase the sizes of our samples. We can model this scenario using Python.

Explanation of Functions

- First, we can create a function called 'get_sample_means' that takes 200 random samples from the population given some user specified size for each of the samples.

- Using a for loop, we can use 'np.random.choice' to randomly select a sample from the population and store the means of those samples in a list.
- Next, we create a function called 'plot_histogram' that plots the sampling distribution of the means using the "matplotlib" library. It also uses the Scipy package to plot a normal distribution with the sample mean and standard deviation. Finally, we plot a vertical line at the population mean for comparison.

Simulation

- We simulate the Central Limit theorem using our two previously defined functions. We choose 4 different sample sizes (1, 30, 50 and 100).
  This simulation shows that as we increase the size of the samples we take, the distribution of the sample means becomes more normal and its variance decreases. Also, as the sample size increases, the density of means around the population means increases. Thus we have demonstrated the Central Limit Theorem.