Central Limit Theorem

What is the Central Limit Theorem (CLT) ?

- If **independent random samples** of size n are selected from a population with mean μ, then the **distribution of the samples means** will approach a **normal distribution** centered at the true population mean μ.

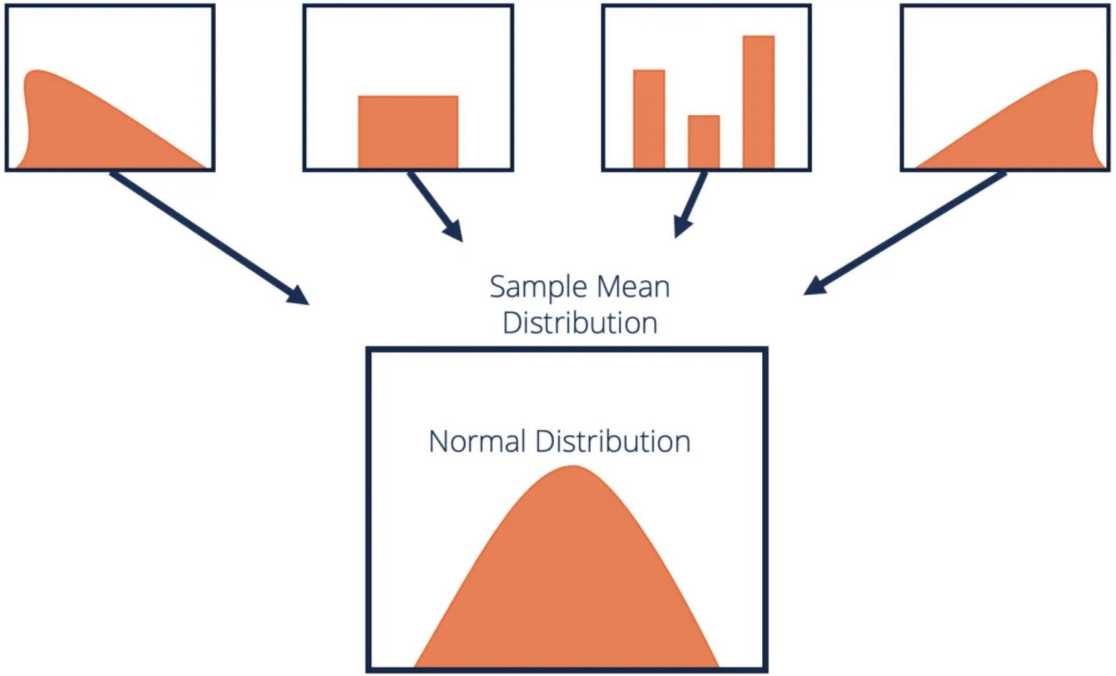
- Underlying population distribution distribution does not matter

Random Sample Requirements

- In general, **n ≥ 30** is sufficiently large
  - Larger sample size means the sampling distribution will be **more normal** and **narrower**
- Observations are:
  - **Independent**
  - **Identically distributed**
- Mean of the sample $\bar{x}$ is an estimate of population mean μ

What is the average number of pieces in a lego set?

Underlying Population Distribution Doesn't Matter!



10+
Ages

340
Pieces

162
VIP Points

40477
Item

Is 340 pieces above or below average?

Take Random Samples and Calculate Sample Means

$\bar{X}_1 = 290$

$\bar{X}_2 = 1054$

$\bar{X}_3 = 529$

. . .

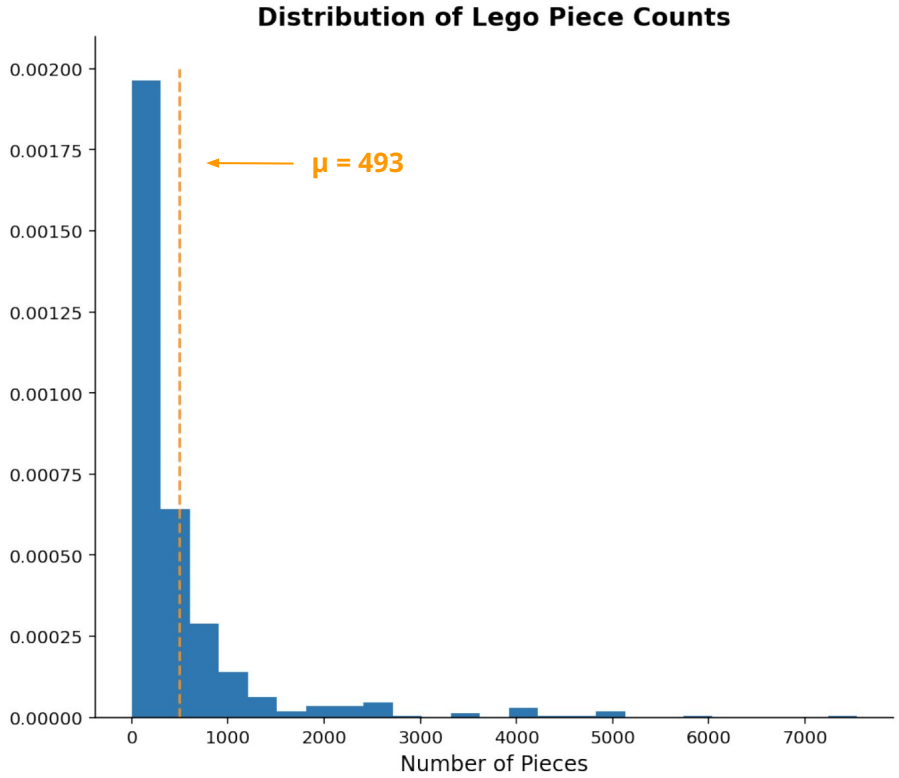
Distribution of the Sample Means

$\bar{X}_1 = 290$

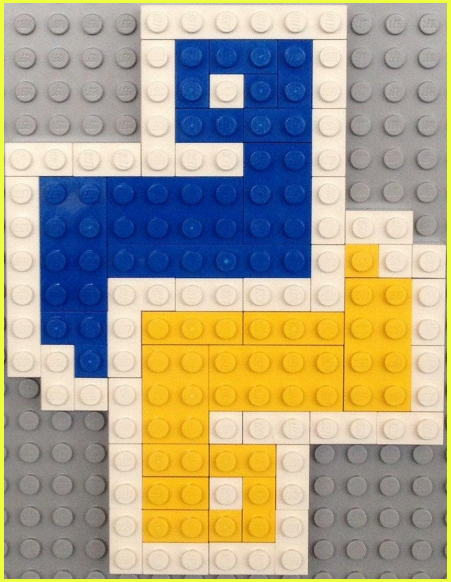
$\mu_{sample} = 49$ $\bar{X}_3 = 529$ ulation

$\bar{X}_2 = 1054$

Number of Pieces - Underlying Population Distribution



Python Simulation



Simulation Overview (1 / 2)

Step 1: Randomly draw sample of given size from dataset (population) with replacement

Step 2 : Calculate mean of sample

Step 3: Repeat above process 100 times to get distribution of sample means

```
def get_sample_means(sample_size):
    """
    This function takes 100 random samples from the population of
    Lego Piece Counts and returns a list of the sample means.
    --------------------
    Input parameters:
    sample_size: integer specifying sample size
    """

    piece_count = lego['piece_count']
    n = 100
    sample_means = []

    # Take 100 random samples from the population and store the sample means.
    for i in range(n):
        sample = np.random.choice(piece_count, sample_size, replace=True)
        sample_means.append(np.mean(sample))
    return sample_means
```

Simulation Overview (2 / 2)

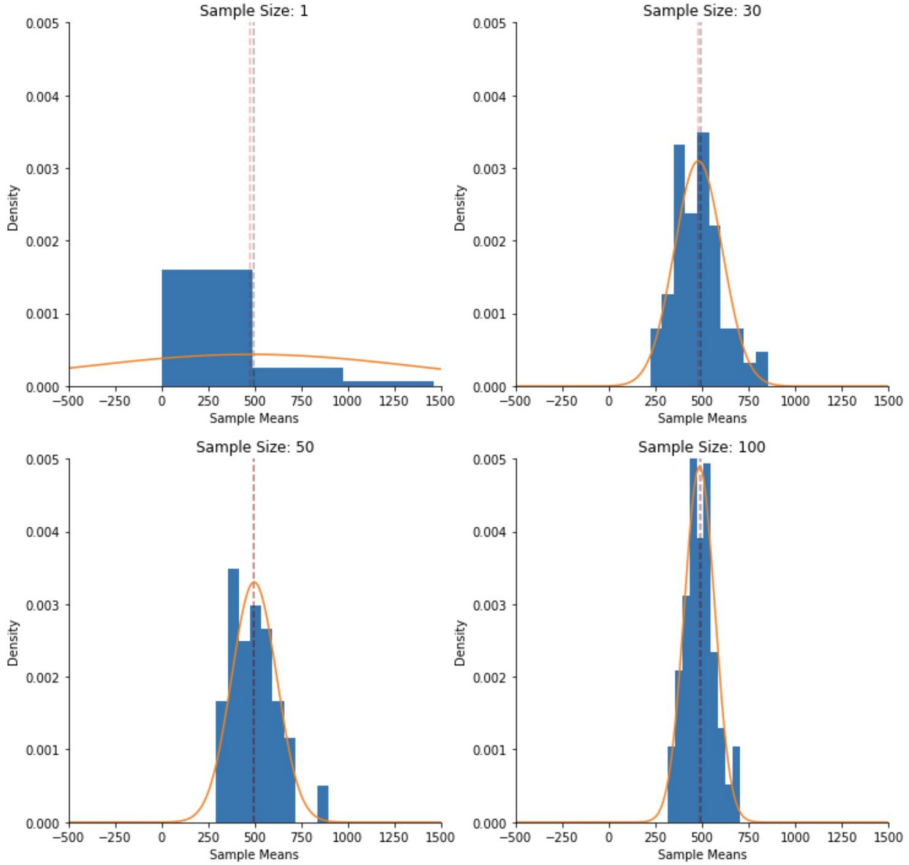
Step 4 : Get sample means distribution for four different sample sizes

Step 5 : Plot histogram of sample distributions

Step 6 : Add ideal normal distribution, population mean for reference.

```
def plot_histogram(sample_means_ls):
    """
    This function returns plots the sample means distibution for four
    different sample sizes, specified by the user.
    --------------------
    Input parameters:
    sample_means_ls: list of lists containing sample means.
    """

    # Define a grid to plot 4 graphs.
    fig, axes = plt.subplots(2, 2, figsize=(12, 12))
    for idx, ax in enumerate(axes.flatten(), 0):
        # Find the sample statistics.
        mean = np.mean(sample_means_ls[idx])
        std = np.std(sample_means_ls[idx])

        # Plot a density histogram for each sample.
        ax.hist(sample_means_ls[idx], density=True)

        # Format the histrogram labels, range, and borders.
        ax.set_xlim(-500, 1500)
        ax.set_ylim(0, 0.005)
        ax.spines['right'].set_visible(False)
        ax.spines['top'].set_visible(False)
        x_axis = range(-1000, 2000)
        ax.set_ylabel("Density")
        ax.set_xlabel("Sample Means")

        # Plot a normal distribution with sample mean and standard deviation.
        ax.plot(x_axis, norm.pdf(x_axis, mean, std))
        ax.set_title(f'Sample Size: {sample_size_ls[idx]}')

        # Plot a verticle line at population mean for comparison
        pop_mean = np.mean(lego['piece_count'])
        ax.vlines(x=pop_mean, colors='black', ymin=0, ymax=0.005,
                  linestyles='--', alpha=0.3)
        ax.vlines(x=mean, colors='red', ymin=0, ymax=0.005,
                  linestyles='--', alpha=0.3)
    plt.show()
```

Sample mean distribution with varying sample sizes



Additional Notes on the CLT

Confidence Intervals and the Central Limit Theorem

- **Confidence intervals** are essentially statements of probability, the long-run frequency
- To calculate probability, we need to know the probability distribution
- **Central Limit Theorem** tells us if our samples are sufficiently large, we can **assume** the sampling distribution of the means is Normal

Central Limit Theorem vs Law of Large Numbers

- CLT and Law of Large Numbers are commonly confused
- As the size of a sample is increased
  - A **single** sample mean converges to the common expectation of the random variables in the sequence
- Observations are:
  - Independent
  - Identically distributed

Q & A