Central Limit Theorem (CLT)

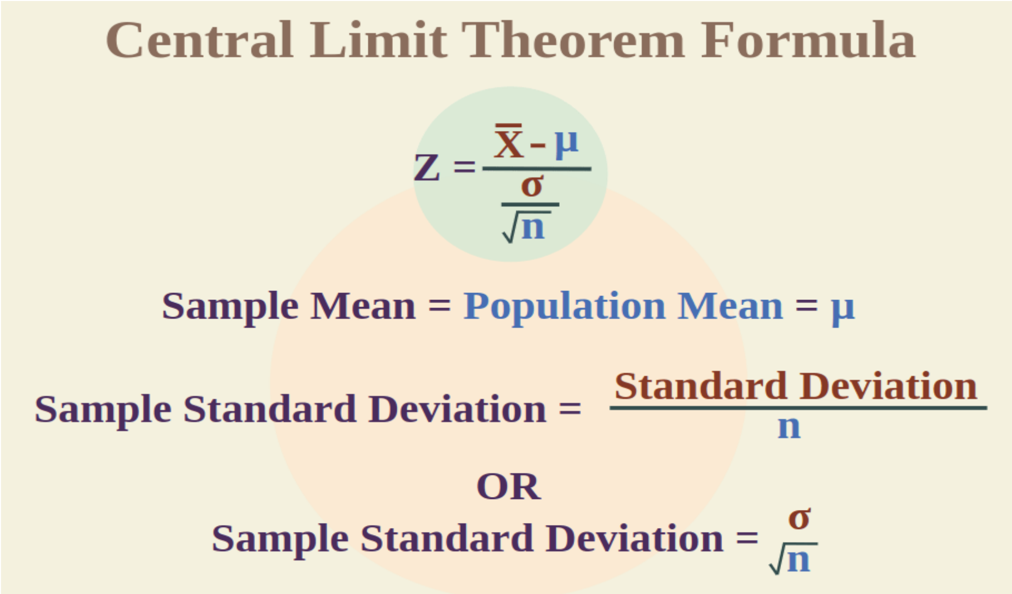
Introduction

The Central Limit Theorem in Statistics states that as the sample size increases and its variance is finite, then the distribution of the sample mean approaches normal distribution irrespective of the shape of the population distribution.

The central limit theorem posits that the distribution of sample means will invariably conform to a normal distribution provided the sample size is sufficiently large. This holds regardless of the underlying distribution of the population, be it normal, Poisson, binomial, or any alternative distribution.

Formula

Z-score: **Z-Score in statistics** is a measurement of how many standard deviations away a data point is from the mean of a distribution. A z-score of 0 indicates that the data point's score is the same as the mean score. A positive z-score indicates that the data point is above average, while a negative z-score indicates that the data point is below average.



Data

We have generated two random data set using the in-built python functions. These are two different non-normal distributions (exponential and uniform). Our aim is to apply Central Limit theorem and study at what point (sample size) the CLT holds true for this dataset., i.e. at what sample size the distribution of sample means approaches a normal distribution. For reference the histogram of raw Exponential and Uniform data is given below:

Exponential distribution:

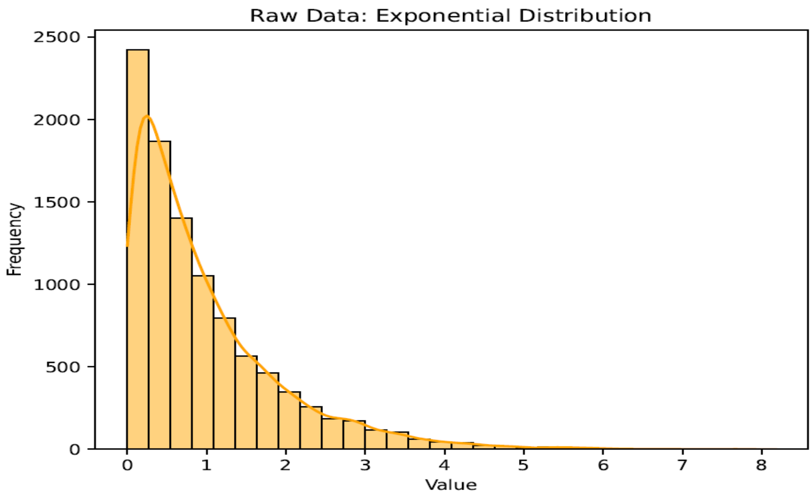


Figure 1: Histogram of raw exponential distribution data.

Uniform distribution:

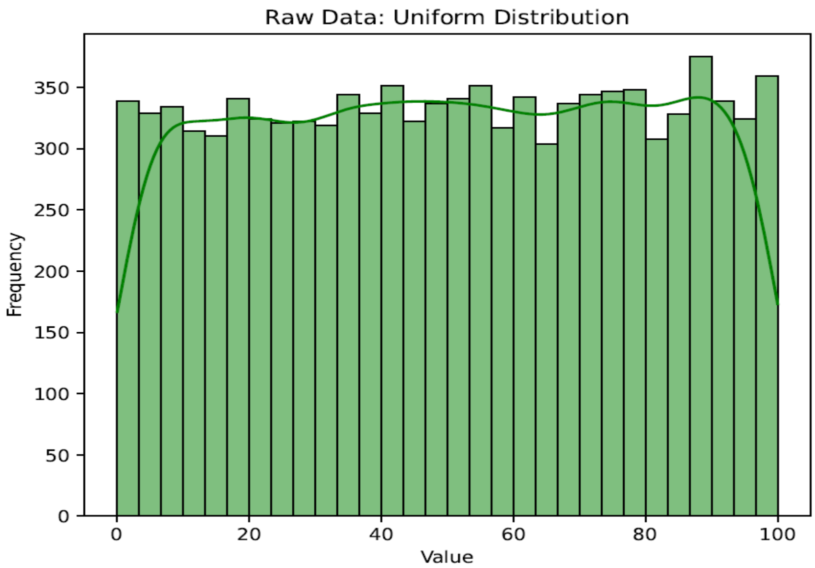


Figure 2: Histogram of raw uniform distribution data.

Q-Q plot of exponential vs uniform distribution:

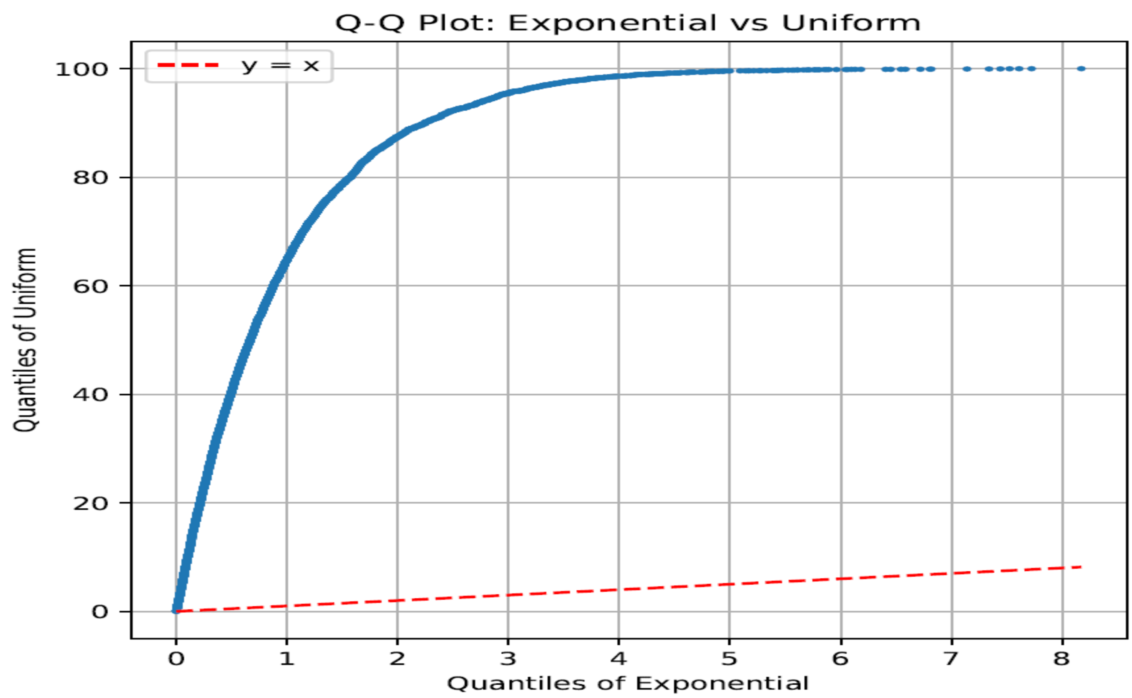


Figure 3: Q-Q plot comparing exponential and uniform distributions.

Method:

We will use python to plot graphs and visualization to determine whether the Central Limit Theorem holds true.
Steps:

1. Plot the distribution of sample mean for a given sample size (n = 10, 30, 100)
2. Plot the Q-Q plot of distribution of sample means and the normal distribution
3. Observe how closely it follows the line y=x

Exponential distribution

- With sample size n = 10

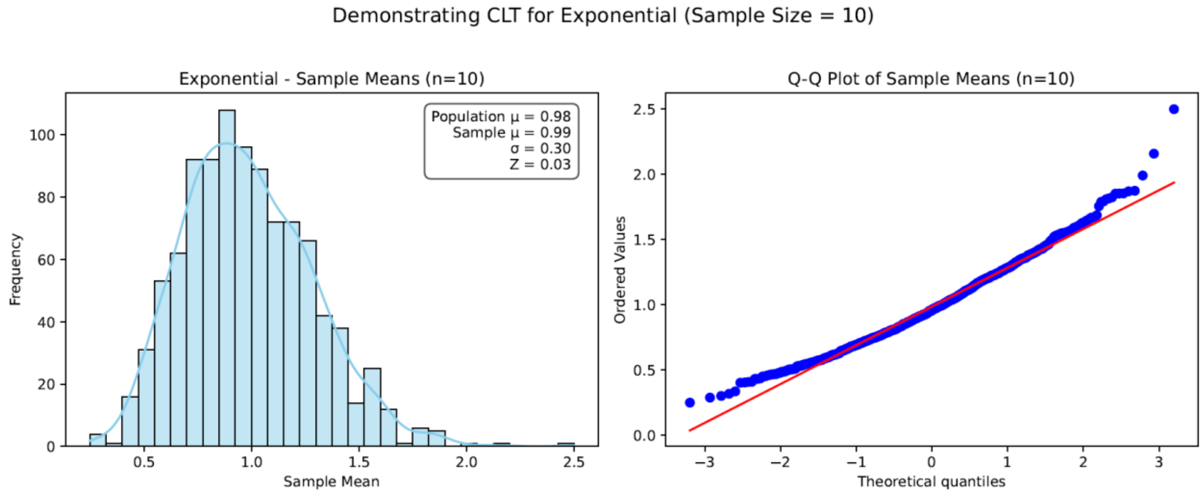


Figure 4: sample distribution plot and Q-Q plot for sample size =10

- With sample size n = 30

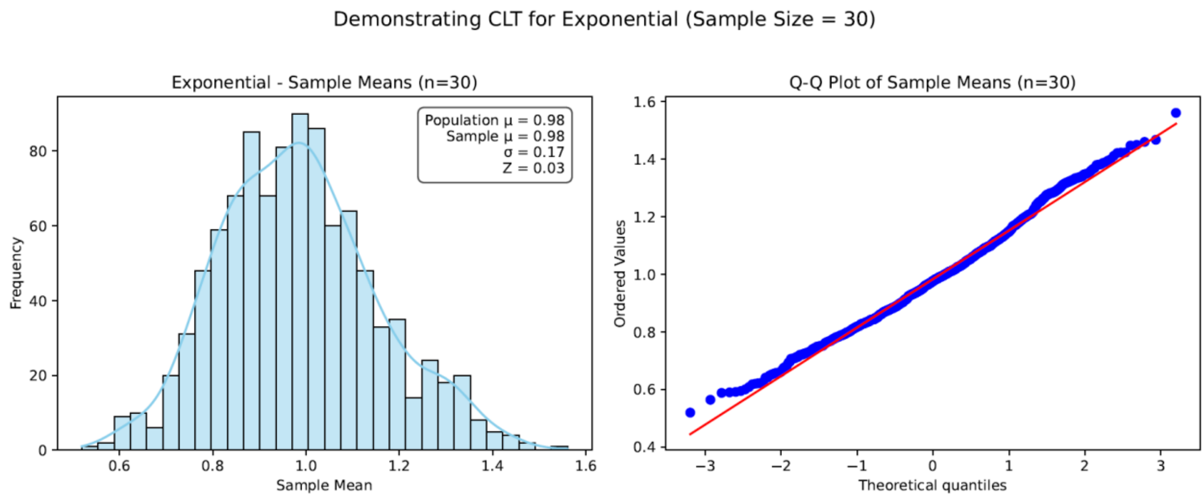


Figure 5: sample distribution plot and Q-Q plot for sample size =30

- With sample size n = 100

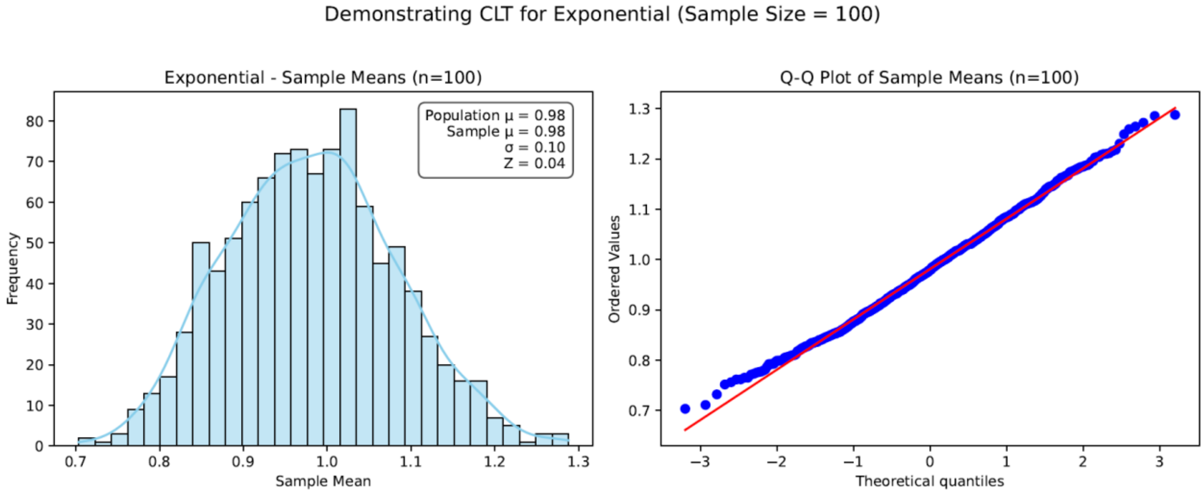


Figure 6: sample distribution plot and Q-Q plot for sample size =100

Uniform distribution

- With sample size n = 10

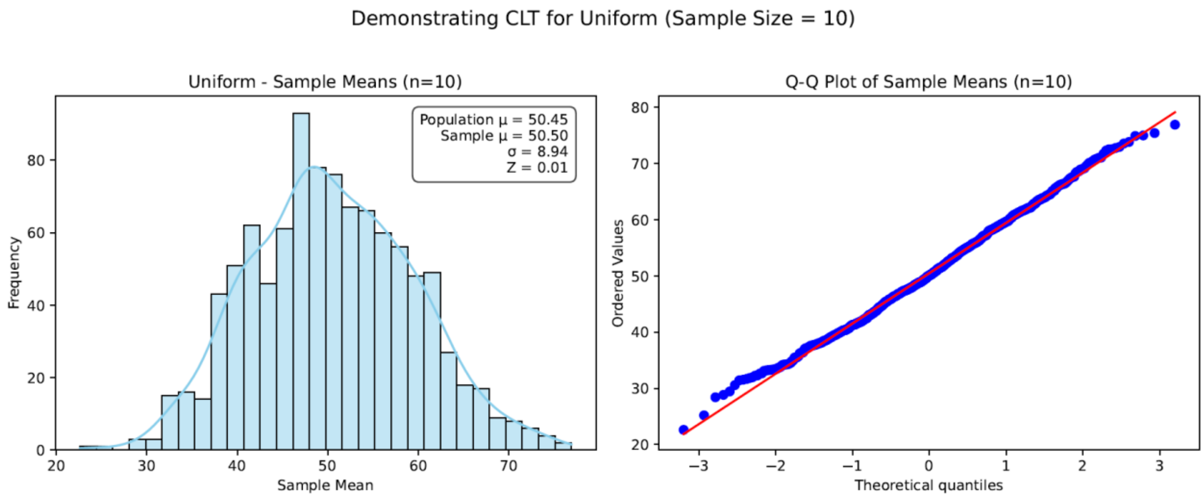


Figure 7: sample distribution plot and Q-Q plot for sample size =10

- With sample size n = 30

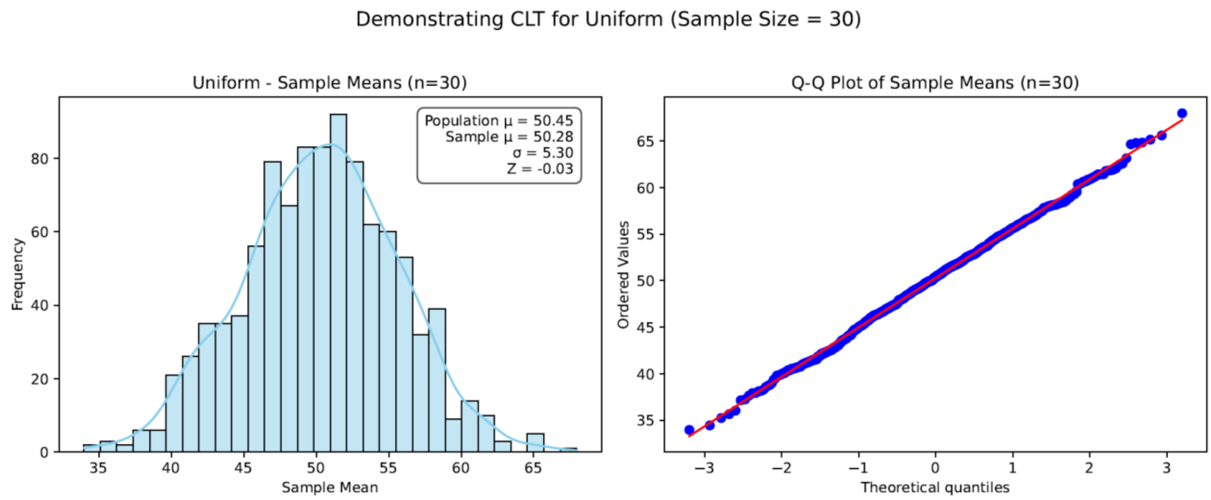


Figure 8: sample distribution plot and Q-Q plot for sample size =30

- With sample size n = 100

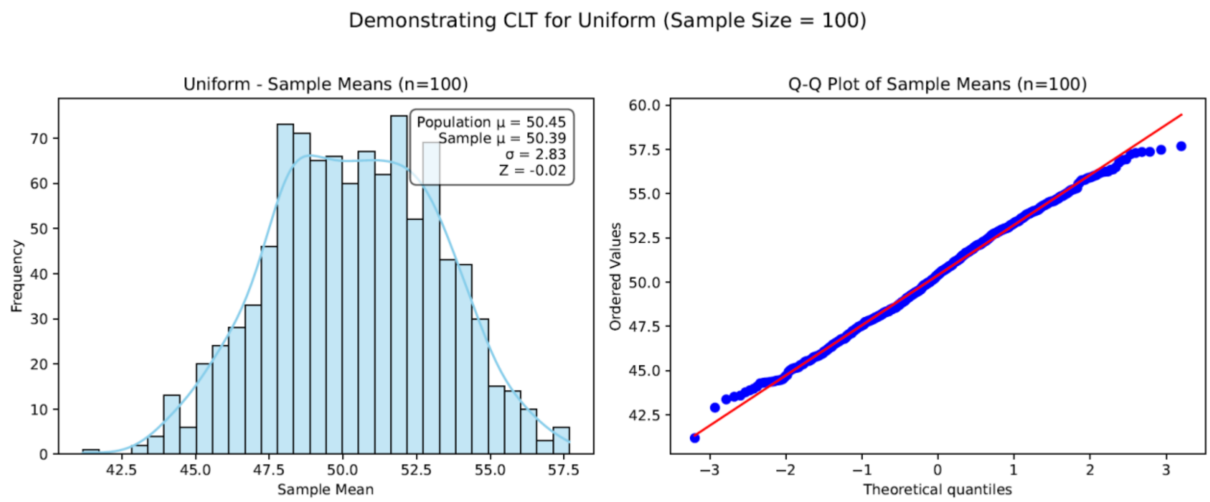


Figure 9: sample distribution plot and Q-Q plot for sample size =100

Conclusion:

We can observe that the Q-Q plot fit more closely to the y=x line as we increase the sample size. That means it is closely resembling a normal distribution. Hence, we can conclude, that with the increase in sample size the distribution, irrespective of its original shape moves towards a normal distribution shape.

References:

GeeksforGeeks. (n.d.). *Central Limit Theorem*. GeeksforGeeks. https://www.geeksforgeeks.org/central-limit-theorem/

Wikipedia contributors. (n.d.). *Central limit theorem*. *Wikipedia*. Retrieved April 22, 2025, from https://en.wikipedia.org/wiki/Central_limit_theorem

GeeksforGeeks. (n.d.). *Z-score in statistics*. GeeksforGeeks. https://www.geeksforgeeks.org/z-score-in-statistics/

GeeksforGeeks. (n.d.). *Quantile-Quantile plot*. GeeksforGeeks. https://www.geeksforgeeks.org/quantile-quantile-plots/

OpenAI. (2023). *ChatGPT* (Mar 23 version) [Large language model]. https://chat.openai.com/

SciPy Community. (2020). *SciPy v1.6.0 Reference Guide: scipy.stats*. SciPy. https://docs.scipy.org/doc/scipy/reference/stats.html

Appendix:

- **Figure 1**: Histogram of raw exponential distribution data.
- **Figure 2**: Histogram of raw uniform distribution data.
- **Figure 3**: Q-Q plot comparing exponential and uniform distributions.
- **Figures 4-9**: Histograms and Q-Q plots of sample means for n = 10, 30, 100 from both exponential and uniform datasets.

Submitted by:
Aakash Vashist (100004661)
Abhishek Negi (100004670)