Demonstration of Central Limit Theorem Using Simulations

Jemish Naliyapara (Roll No.: 18114)

Instructor: Dr. Ankur Raina

Under the course
ECS301: Principles of Communications

1 Introduction

The Central Limit Theorem, or CLT for short, has a wide range of uses, particularly in the field of probability theory. The CLT has been proved by numerous mathematicians in a variety of situations, leading to the development of numerous variants of the theorem. The concept of the CLT mainly extended by Abraham de Moivre [1], Pierre Simon Laplace[2] and Aleksandr Lyapunov[3]. The more general proof of the central limit theorem is given by Lindeberg[4] and Levy in his independent work.

The Central Limit Theorem is a powerful statistical theorem that claims that with a sufficiently large sample size n, a normal distribution will arise regardless of the initial distribution. The Central Limit Theorem is used in many applications, such as hypothesis testing, confidence intervals, and estimation, to make fair assumptions about the population, which is sometimes difficult to do when the population is not normally distributed and the shape of the distribution is uncertain.

In the field of mathematics, the Central Limit Theorem has had and continues to have a significant influence. The theorem was enlarged such that it could be applied to topology, analysis, and many other areas of mathematics in addition to probability theory.

2 Theory

In the fields of probability and statics theory, the Central Limit Theorem (CLT) is a crucial and extensive subject. Let's first understand some definations related to CLT [5].

- A **population** is the complete set of observations with which we are concerned.
- A collection of favourable outcomes that may be repeated is an **experiment**.
- A **sample** is a subset of the population.

2.1 Central Limit Theorem

Numerous mathematicians have worked on the Central Limit Theorem and its proof over the years, and as a result, the theorem can be expressed in a wide variety of ways. Throughout the report, we will follow to the formulation of the central limit theorem described by the Lindeberg-Levy Theorem.

The central limit theorem (CLT) says that as sample size increases, the distribution of sample means approaches a normal distribution, independent of the distribution of the population. The samples is independent and identically distributed. If all random variables are independent of one another and have the same probability distribution, then the sequence of random variables is said to be **independent and identically distributed**.

A set of n independent random variables (r.v.s), $X_1, X_2, X_3, ..., X_n$ with the given distribution and the expectation and variance values $E(X_i) = \mu$ and $Var(X_i) = \sigma^2$, respectively, constitute a random sample of size n from the given distribution.

We can calculate the sample mean, expectation value of sample mean and variance of sample mean as follow:
Sample Mean:

$$\bar{X}_n = \frac{X_1 + X_2 + X_3 + ... + X_n}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{1}$$

Expectation Value of Sample Mean:

$$E[\bar{X}_n] = E\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n} E[X_i] = \frac{1}{n}(n\mu) = \mu \tag{2}$$

Variance of Sample Mean:

$$Var[\bar{X}_n] = Var\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n^2}\sum_{i=1}^{n} Var[X_i] = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n} \tag{3}$$

According to the Central Limit Theorem (CLT), a sufficiently large random sample from an essentially arbitrary distribution has a mean that roughly follow a normal distribution. In our case, Sample mean follow the normal distribution $N(\mu, \frac{\sigma}{\sqrt{n}})$.

Given a large number of sample sizes n, the central limit theorem implies that the sample mean converges to the population mean. Similarly, the population standard deviation can be calculated using the sample mean standard deviation. As sample size (n) rises, the standard deviation of the sample mean distribution converges to zero.

We can test these assertions using the simulation described in the following section.

3 Approach in Simulation

I used the Python programming language to simulate the central limit theorem in Google Colab. I took different approach then shown in [6] and [7], which is described in following steps.

Step:1
First generate population set using random number generation. As shown in Figure (1), population is made up of random numbers taking from -60 to 60 using the random python module. Here, the population size (n_population) is 10000. The random.randint() generates randomly distributed integer numbers from -60 to 60.The distribution of population is shown in figure (3). The mean of population (μ) is 0.404 and standard deviation of population (σ) is 34.645.

```
import numpy as np
import matplotlib.pyplot as plt
import random
```

```
n_population=10000  #size of population
A=np.zeros(n_population)  # A define as population
for i in range(n_population):
    A[i]=random.randint(-60,60)   #randomly generated numbers from -60 to 60

A_mean=np.mean(A)
A_median=np.median(A)
A_std=np.std(A)

print("Mean:",A_mean)
print("Median:",A_median)
print("Standard Deviation:",A_std)
```

Figure 1: Population set is generate using random number generate from -60 to 60, and get population mean and standard deviation.

Step:2
In second step, I randomly choose numbers from population to generate sample as shown in the figure (2). We can able to study simulations by changing size of the sample (n_sample). n2 represent number of sample used to generate distribution of mean of the sample. Here, 1000 samples are used to generate sample mean distribution. The detailed studies of the mean and standard deviation of these distribution is discussed in the next section. From this studies, I am able to prove the theoretical results of the central limit theorem.

```
n_sample=5000 #sample size
n2=1000  #Number of sampling
A_bar=np.zeros(n2)
for i in range(n2):
    a=[]
    for j in range(n_sample):
        a.append(random.choice(A)) #Sample generated randomly from population(A)
    A_bar[i]=np.mean(a)
#print(A_bar)
x_bar2=np.unique(A_bar)
#print(x_bar2)
n3=np.size(x_bar2)
Freq_A_bar=np.zeros(n3)
for i in range(n3):
    for j in range(n2):
        if A_bar[j]==x_bar2[i]:
            Freq_A_bar[i]+=1
A_bar_mean=round(np.mean(A_bar),3)
A_bar_std=round(np.std(A_bar),3)
```

Figure 2: The distribution of sample mean generated by this code.

4 Simulation Results

As discussed in the previous section, the randomly distributed integer numbers from -60 to 60 is created using python code. The population distribution is shown in figure (3). The mean of population (μ) is 0.404 and standard deviation of population (σ) is 34.645.

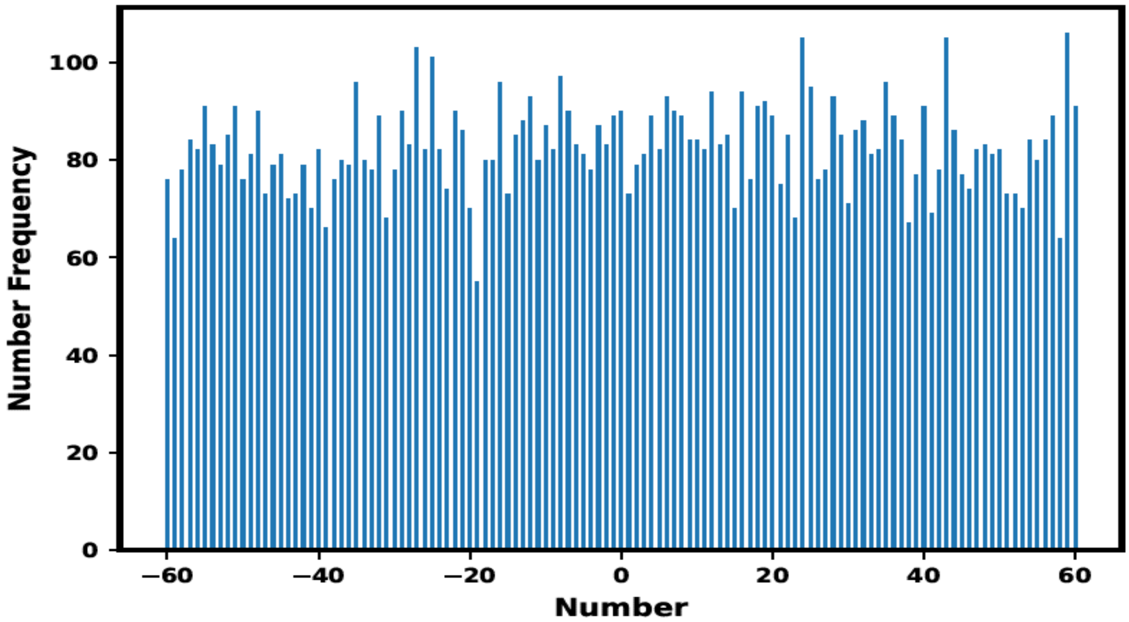
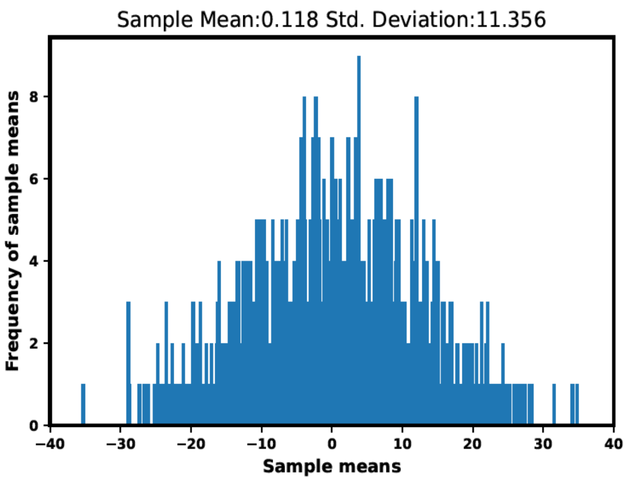


Figure 3: The randomly distributed population generated by the python code.

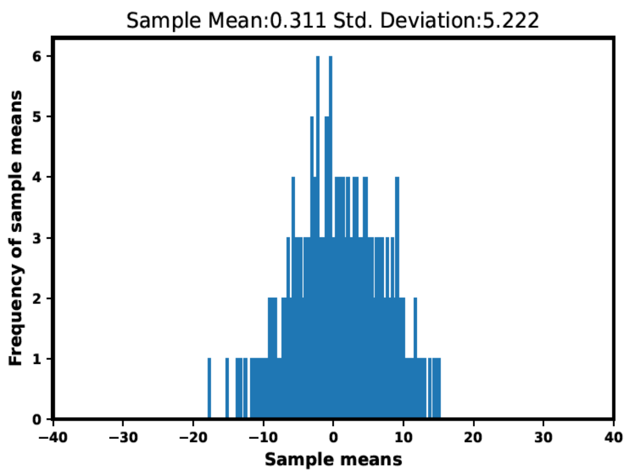
In real life problem, We can not able to find mean and standard deviation of the randomly distributed population. For that, we use central limit theorem to find the statistical parameter like mean and standard deviation of the population distribution. As CLT states, the mean of samples is converge to population means, as we increase size of sample. The simulations of this statement are shown below.

4.1 Simulations

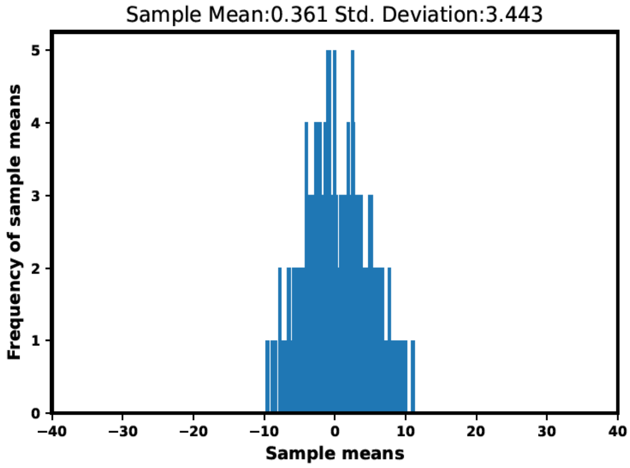
The distributions of sample mean with different sample size is given below:



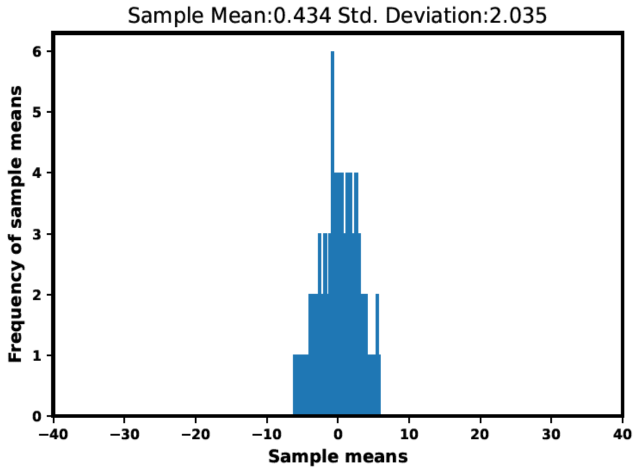
(a) Sample Size = 10



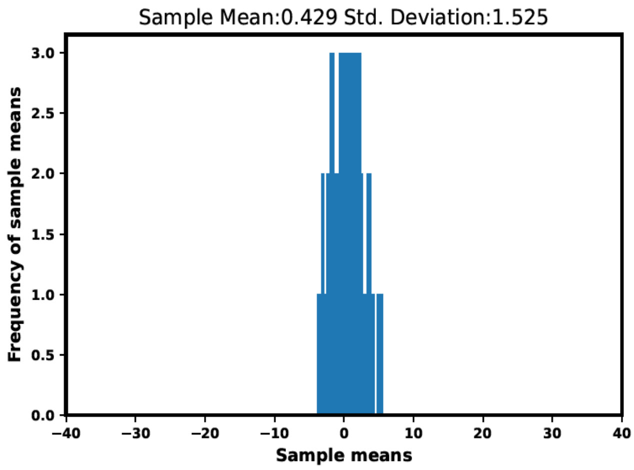
(b) Sample Size = 50



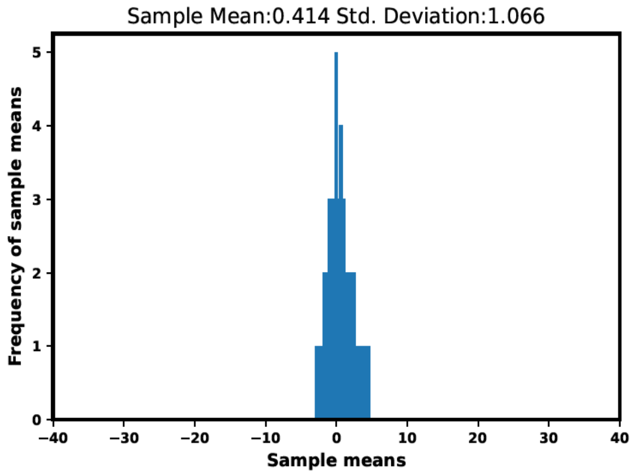
(c) Sample Size = 100



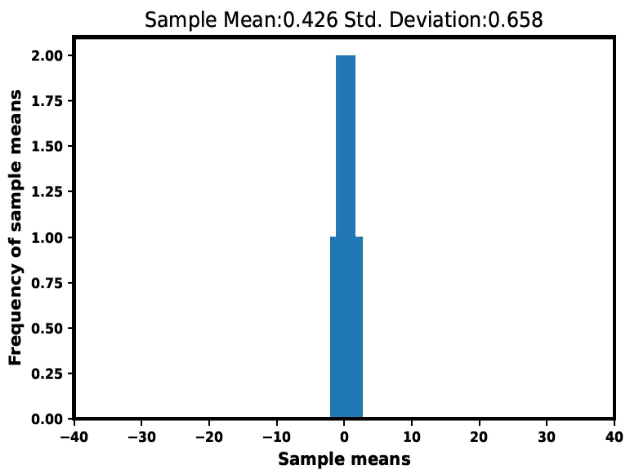
(d) Sample Size = 300



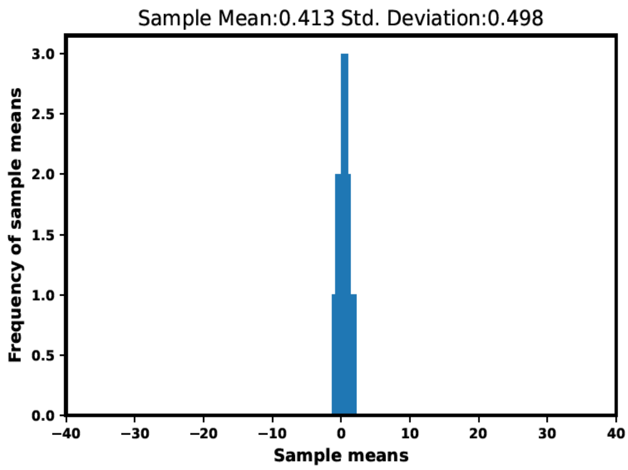
(e) Sample Size = 500



(f) Sample Size = 1000



(g) Sample Size = 3000



(h) Sample Size = 5000

Figure 4: The sample mean distributions with different sample size is shown in figures. The mean ($\bar{x}$) of the distribution and Sample standard deviation (σ_{sample}) are also displayed on the top of the figures.

5 Discussion

To prove the statements of central limit theorem, I took 29 different sample size and plot means and standard deviation of the sample mean distribution as follows;

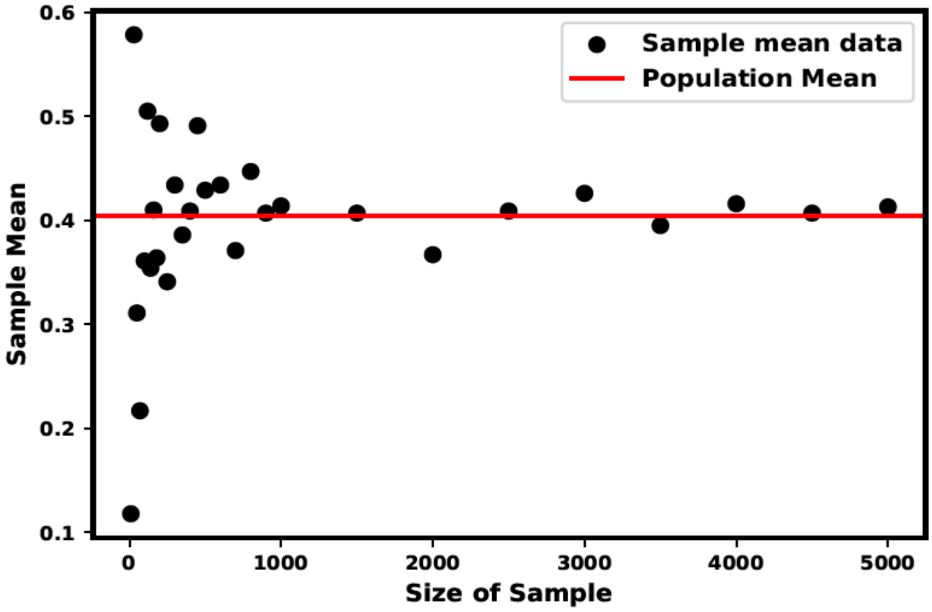


Figure 5: Sample mean converges to population mean with increase in sample size.

As discussed in section 2, the central limit theorem stats that the sample mean converges to the population mean for large number of sample size.

$$E[\bar{x}_n] \to \mu \qquad for\ large\ sample\ size \tag{4}$$

The proof of this statement can be shown by figure (5). As we increase sample size, the data points which present sample data are converges to the population mean (Red line) in the figure (5). Here, we know the population mean of random distributed numbers. But, in real life, it is difficult to find population by its distribution. So, The population mean can be found by taking mean of large number of sample from the population. Similarly, standard deviation of population can be found from sample mean distribution as shown in the below figure,

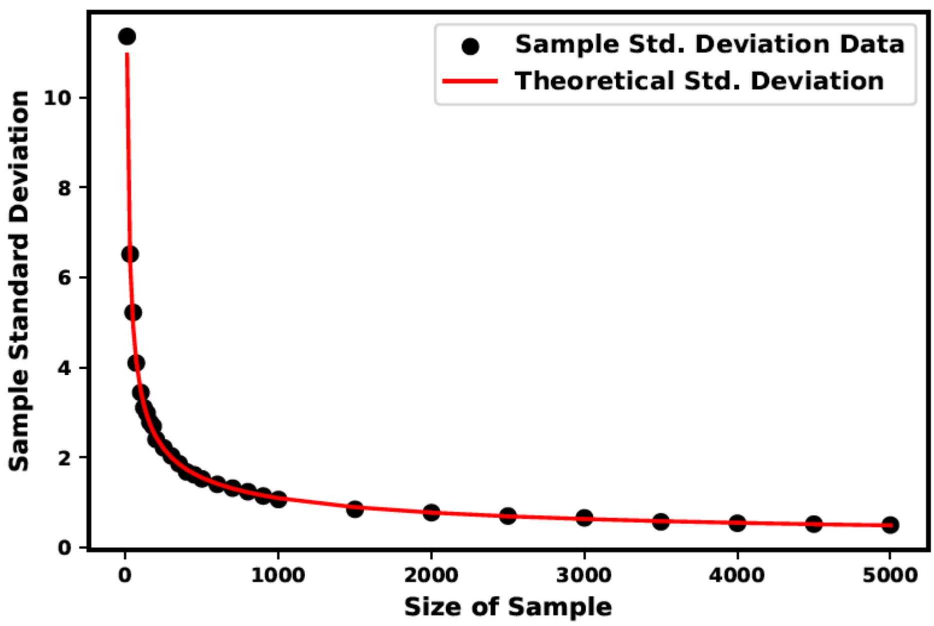


Figure 6: Standard deviation of the sample mean distribution approach to zero as sample size increase.

We can find the standard deviation of the population from standard deviation of the sample mean distribution as discussed in section 2. Theoretically, the standard deviation of sample mean distribution converges to zero as sample size increase.

$$\sigma_{sample} = \frac{\sigma_{population}}{\sqrt{n}} \to 0 \qquad for\ large\ sample\ size \tag{5}$$

Figure (6) compares the standard deviation of sample mean distribution from the simulation and the theoretical standard deviation given in equation (5). The simulation results (black data points) agreed with theoretical calculation (red lines). Collectively, from figure (5) and figure (6), we can say that the simulation given in the section 3 is perfectly demonstrate the statements of the central limit theorem.

6 Real Life Application

The central limit theorem has numerous statistical and practical applications, as follows:

- **The Central Limit Theorem in Finance:**
  As discussed in [8], The CLT is helpful for analysing returns of a single stock or larger indices because the analysis is straightforward and the relevant financial data can be generated very easily. Therefore, the CLT is used by investors of all stripes to evaluate stock returns, build portfolios, and control risk.
- **Election Polls:**
  Many statistical company predict result of election poll using the central limit theorem. They take samples from the population and predict results within statstical confidence interval.
- **Economics:**
  Economic data such as per capita income and expenditure are based on the central limit theorem. The central limit theorem approaches per capita income by using sample mean or average income of the specific location.

We concluded by talking about the central limit theorem by Lindeberg and Levy. The report also discusses the simulation method for proving the central limit theorem. Additionally, we contrasted the simulation results with the claims made by the theorem. These simulation findings perfectly match the calculations made in theory. The practical application of the central limit theorem is discussed in this report.

References

[1] Abraham de Moivre, *The Doctrine of Chances* (Book Link)

[2] Pierre Simon Laplace, *Theorie Analytique des Probabilités* (Book Link)

[3] Aleksandr Lyapunov, Chapter:5 The Way Toward Morden Probability in book *History of the Central Limit Theorem* (Book Link)

[4] Lindeberg's Conditions (Link)

[5] Walpole, Ronald E, Raymond H. Myers, Sharon L. Myers, and Keying Ye. *Probability & Statistics For Engineers & Scientists*. Prentice Hall. 2012.

[6] Python – Central Limit Theorem, *geeksforgeeks.com*(Link)

[7] Central Limit Theorem Explained with Python Code, *towarddatascience.com*(Link)

[8] Central Limit Theorem in Finance, *investopedia.com*(Link)