The Central Limit Theorem (CLT): meaning, proof and simulations

Pietro Colaguori

November 2023

1 Meaning

The Central Limit Theorem (CLT) is a fundamental concept in probability and statistics that describes the distribution of the sum (or average) of a large number of independent, identically distributed random variables. The theorem is particularly powerful because it allows us to make certain probabilistic statements about the sum or average of a sample even when we don't know the exact distribution of the underlying population.

Let $X_1, X_2, ..., X_n$ be a sequence of independent and identically distributed (i.i.d.) random variables with mean μ and standard deviation σ. As n grows the distribution of the standardized sum or average $\frac{\sum_{i=1}^{n}(X_i - n\mu)}{\sigma\sqrt{n}}$ converges to a standard normal distribution, i.e. a normal distribution with mean 0 and standard deviation 1.

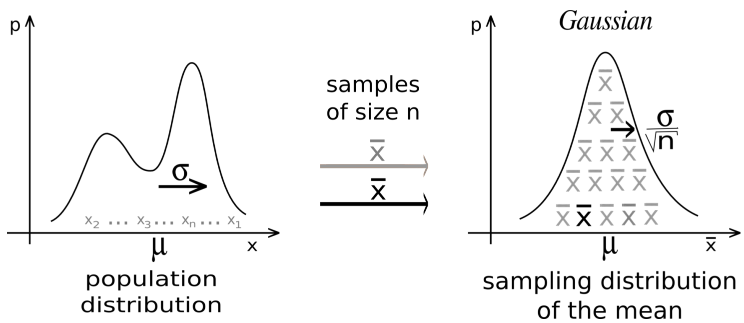

Figure 1: Illustration of the CLT.

Mathematically we can express the CLT as follows:

$$\lim_{n\to\infty} P\left(\frac{\sum_{i=1}^{n}(X_i - n\mu)}{\sigma\sqrt{n}} \leq x\right) = \Phi(x)$$

where $\Phi(x)$ is the cumulative distribution function (CDF) of the standard normal distribution.

We can make some additional practical observations about the CLT:

- Sample size matters: it is important to know that the CLT works well for sample sizes n that can be considered big, otherwise, if the sample size is too small other considerations come into play and the behavious may not be as expected.
- Independence assumption: the random variables in the sample should be independent and the sample size should be small, compared to the size of the population.
- Applicability on averages: the CLT is mostly applied to sample means, but it can also be applied to other statistics, e.g. sample sums.

The Central Limit Theorem is a cornerstone of statistical theory, providing a bridge between the behavior of individual observations and the behavior of sample statistics, making statistical inference more practical and widely applicable.

2 Proof

Suppose $X_1, X_2, ..., X_n$ are i.i.d. random variables with mean 0 and variance σ^2. Then consider the sum $S_n = \sum_{i=1}^{n} X_i$. We can compute the standardized sum as

$$Z_n = \frac{S_n - n\mu}{\sigma\sqrt{n}}$$

To analyze the distribution of Z_n we can use the characteristic function, which is just the Fourier transform of the probability distribution. Then we can emply a Taylor expansion of the characteristic function arond 0. At this point we notice that the distribution of Z_n converges to the standard normal distribution as $n \to \infty$. By the Continous Mapping Theorem the convergence of characteristic functions implies the convergence in distribution. Therefore, Z_n converges in distribution to a standard normal random variable when $n \to \infty$. $\square$

3 Simulation

We can use the following Python script to simulate the Central Limit Theorem (CLT). The idea is to create a histogram based on the means of the values sampled at random using the *random* function.

```python
import streamlit as st
import numpy as np
import matplotlib.pyplot as plt
import random

# Function to simulate the central limit theorem
def simulate_clt(sample_size, num_samples):
    means = []
    for _ in range(num_samples):
        sample = [random.random() for _ in
            range(sample_size)]
        sample_mean = np.mean(sample)
        means.append(sample_mean)
    return means

# Get user input for sample size and number of samples
sample_size = st.number_input("Enter sample size:",
    min_value=1, value=10)
num_samples = st.number_input("Enter number of
    samples:", min_value=1, value=100)

# Function to update the simulation based on user
    input
def update_simulation():
    means = simulate_clt(sample_size, num_samples)
    plt.hist(means, bins='auto')
    plt.xlabel("Sample Mean")
    plt.ylabel("Frequency")
    plt.title("Central Limit Theorem Simulation")
    st.pyplot()

# Button to update the simulation
if st.button("Update"):
    update_simulation()

# Initial simulation
update_simulation()
```

Upon running this script, a web page will be opened, allowing the user to select the sample size and the number of samples. Below I show the results I obtained using always a sample size of 10 but at first with a number of samples of 100 and then with a number of samples of 1000.

We can clearly observe how the means of the samples follow a standard

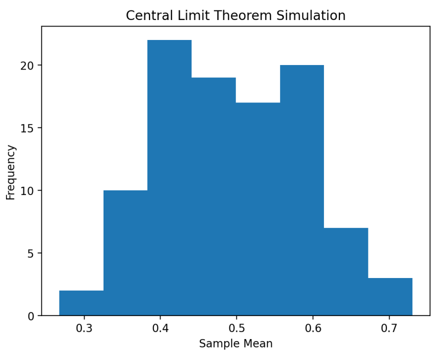

Figure 2: With 100 samples

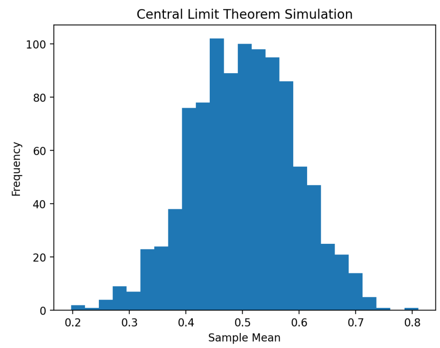

Figure 3: With 1000 samples

normal distribution.