Central Limit Theorem

Giacomo Babudri

December 2023

1 Introduction

The central limit theorem relies on the concept of a sampling distribution, which is the probability distribution of a statistic for a large number of samples taken from a population.

The central limit theorem says that the sampling distribution of the mean will always be normally distributed, as long as the sample size is large enough. Regardless of whether the population has a normal, Poisson, binomial, or any other distribution, the sampling distribution of the mean will be normal.

A normal distribution is a symmetrical, bell-shaped distribution, with increasingly fewer observations the further from the center of the distribution.

1.1 Statement of the theorem

The CLT asserts that the distribution of the sum (or average) of a large number of independent, identically distributed random variables approaches a normal distribution, regardless of the original distribution of the variables. Mathematically, it can be expressed as:

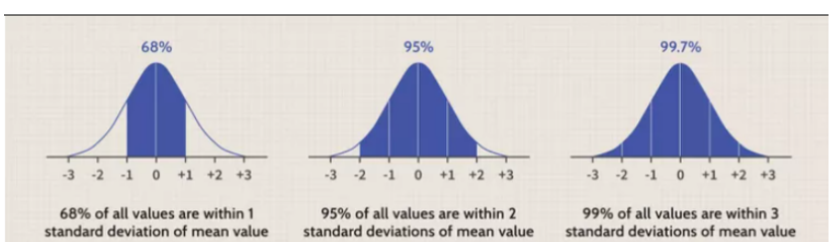

Figure 1: Example

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

Where:

- X_i are independent and identically distributed random variables,
- μ is the mean of X_i,
- σ is the standard deviation of X_i,
- n is the sample size,
- $\Phi(x)$ is the cumulative distribution function of the standard normal distribution.

2 Proof

Premise: Consider a sequence of independent and identically distributed random variables $X_1, X_2, \ldots, X_n$ with a mean μ and variance σ^2.

Proof:

Consider the characteristic function $\phi_{X_i}(t) = E[e^{itX_i}]$ of the individual random variables X_i. The characteristic function of the sum of independent random variables is the product of their individual characteristic functions:

$$\phi_{\sum_{i=1}^{n} X_i}(t) = \prod_{i=1}^{n} \phi_{X_i}(t)$$

Define the standardized sample mean as:

$$Z_n = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}$$

At this point it's necessary to show that the characteristic function of the standardized sample mean converges pointwise to the characteristic function of the standard normal distribution as n approaches infinity:

$$\lim_{n\to\infty} \phi_{\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}}(t) = e^{-\frac{t^2}{2}}$$

Apply Levy's Continuity Theorem, which states that if the characteristic functions of a sequence of random variables converge pointwise to the characteristic function of another random variable, then the distributions of the random variables converge weakly to the distribution of that random variable.

Conclusion:

Conclude that the distribution of the standardized sample mean converges to the standard normal distribution as n approaches infinity.

3 Pratical Example

Biology Biologists use the central limit theorem whenever they use data from a sample of organisms to draw conclusions about the overall population of organisms.

For example a biologist may measure the height of 30 randomly selected plants and then use the sample mean height to estimate the population mean height.

If the biologist finds that the sample mean height of the 30 plants is 10.3 inches, then her best guess for the population mean height will also be 10.3 inches.

Economy Economists often use the central limit theorem when using sample data to draw conclusions about a population.

For example an economist may collect a simple random sample of 50 individuals in a town and use the average annual income of the individuals in the sample to estimate the average annual income of individuals in the entire town.

If the economist finds that the average annual income of the individuals in the sample is 58,000, then her best guess for the true average annual income of individuals in the entire town will be 58,000.

Agricultural Agricultural scientists use the central limit theorem whenever they use data from samples to draw conclusions about a larger population.

For example, an agricultural scientist may test a new fertilizer on 15 different fields and measure the average crop yield of each field.

If it's found that the average field produces 400 pounds of wheat, then the best guess for the average crop yield for all fields will also be 400 pounds.

4 Simulation

During the course was asked to simulate a system with this features: "M systems are subject to a series of N attacks. On the x-axis, we indicate the attacks and on the Y-axis we simulate the accumulation of a "security score" (-1, 1), where the score is -1 if the system is penetrated and 1 if the system was successfully "shielded" or protected. Simulate the score "trajectories" for all systems, assuming, for simplicity, a constant penetration probability p at each attack". The execution of this exercise, with the right input represents a clear example of how CLT works. For this reason let's try this simulation with two different input and then analyze the relatives output.

First case: N = 500, M = 20, p = 0.5.

For a large number of systems and attacks i can notice in the output a clear normal distribution; in particular it's evident the bell shape along all the representations, indicating the normal distribution. The histogram it's less useful, but still gives as more informations about the simulation.

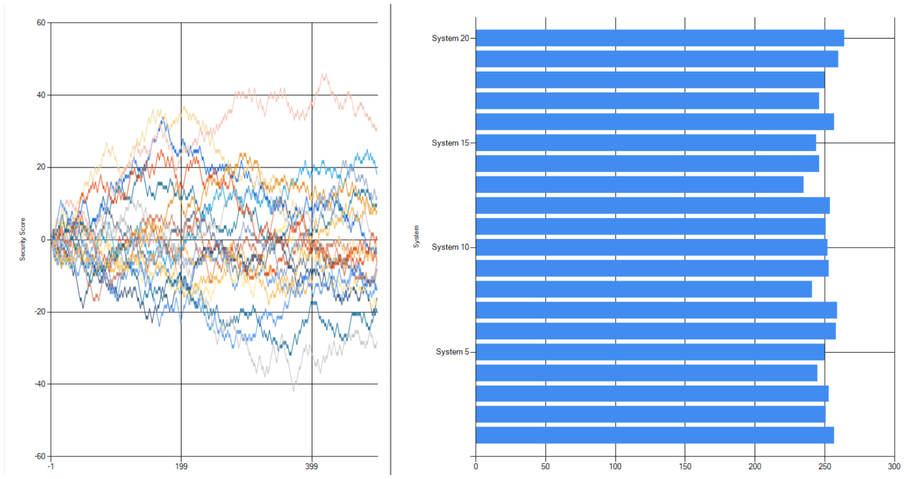

Figure 2: Simulation number one

Second case: N = 50, M = 5, p = 0.5.

It's clear that the distribution, it's far from the normal one and it's much more unpredictable. This bring us the conclusion that the Central Limit Theorem works just as we expect to.

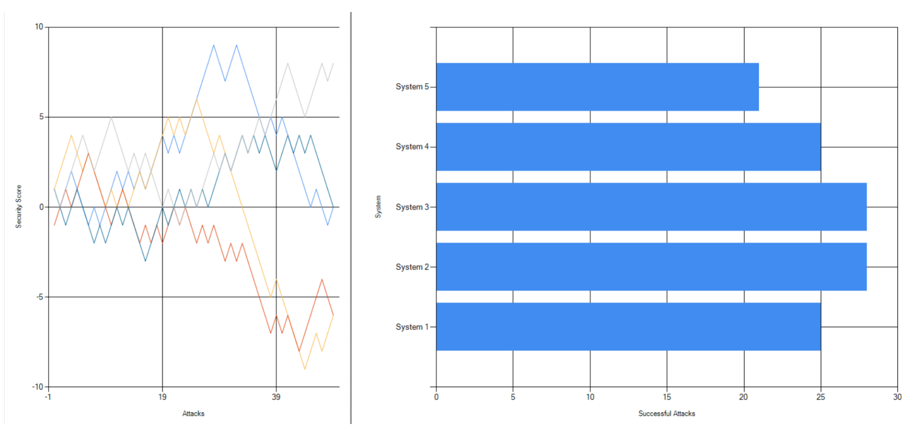

Figure 3: Simulation number two

The execution it's realized through this piece of code written in c:

```
1 riferimento
private int[][] AttacksLanded(int M, int N, double p)
{

    Random random = new Random();
    int[][] Systems = new int[M][];
    for (int i = 0; i < M; i++)
    {
        int[] attacks = new int[N];
        for (int j = 0; j < N; j++)
        {

            double att = random.NextDouble();
            if (att < p)
            {
                attacks[j] = 1;
            }
            else
            {
                attacks[j] = -1;
            }
        }
        Systems[i] = attacks;

    }
    return Systems;

}
```

Figure 4: Code of the simulation