SAPIENZA
UNIVERSITÀ DI ROMA

Master's Degree in Cybersecurity

The Central Limit Theorem (CLT)

Author:
Riccardo Tuzzolino

University ID:
1954109

Academic Year 2023/24

Contents

Chapter 1

Introduction

In the realm of statistical theory, the Central Limit Theorem (CLT) stands as a cornerstone, offering profound insights into the behavior of sample means. Developed in the early 18th century, the theorem has become a fundamental concept in probability and statistics, casting its influence across various scientific disciplines and applications.

At its core, the Central Limit Theorem addresses the behavior of the sum or average of a large number of independent, identically distributed random variables. It asserts that, regardless of the underlying distribution of the original variables, the distribution of the sum (or average) tends to converge towards a normal distribution as the sample size increases. This striking property facilitates the application of Gaussian statistics (statistics based on the normal distribution) to a wide array of real-world scenarios, providing a powerful tool for statistical inference.

This thesis aims to explore the intricacies of the Central Limit Theorem, examining its mathematical foundations, applications, limitations and potential extensions, underlining its role in shaping the landscape of inferential statistics.

Chapter 2

Central Limit Theorem

In order to understand what the theorem states we first need to understand some basic key concepts.

In general, to analyze random experiments, we usually focus on some numerical aspects of the experiment. These numerical values give some information about the outcome of the random experiment. They are called *random variables.*

A **random variable** X is a function from the sample space Ω to the real numbers:

$$X : \Omega \to \mathbb{R}$$

In essence, a random variable is a real-valued function that assigns a numerical value to each possible outcome of the random experiment.

The random variables $X_1, X_2, ..., X_n$ are said to form a (simple) **random sample** of size n if:

1. The X_i's are **independent** random variables (knowledge of the value of one variable gives no information about the value of the other and vice versa)
2. Every X_i has the **same probability distribution** (which means they have the same mean μ and the same variance σ^2)

We say that these X_i's are <u>independent</u> and <u>identically</u> distributed (i.i.d.).

Suppose that $X_1, X_2, ..., X_n$ are i.i.d. random variables with expected value $E(X_i) = \mu$ and variance $Var(X_i) = \sigma^2$. Then, the **sample mean**

$$\bar{X} = \frac{X_1 + X_2 + ... + X_n}{n}$$

has mean $E(\bar{X}) = \mu$ and variance $Var(\bar{X}) = \frac{\sigma^2}{n}$.

This can easily be proven:

- Mean of Sample Mean

$$E[\bar{X}] = E\Big[\frac{X_1 + X_2 + ... + X_n}{n}\Big] = \frac{1}{n}[E(X_1) + E(X_2) + ... + E(X_n)] =$$

$$\frac{1}{n}[\mu + \mu + ... + \mu] = \frac{1}{n}[n\mu] = \mu.$$

- Variance of Sample Mean

$$Var(\bar{X}) = Var\Big(\frac{X_1 + X_2 + ... + X_n}{n}\Big) = Var\Big(\frac{1}{n}X_1 + \frac{1}{n}X_2 + ... + \frac{1}{n}X_n\Big) =$$

$$\frac{1}{n^2}Var(X_1) + \frac{1}{n^2}Var(X_2) + ... + \frac{1}{n^2}Var(X_n) = \frac{1}{n^2}[\sigma^2 + \sigma^2 + ... + \sigma^2] =$$

$$\frac{1}{n^2}[n\sigma^2] = \frac{\sigma^2}{n}.$$

  The result shows that as the sample size n increases, the variance of the sample mean decreases.

The central limit theorem has several variants. In its common form, the random variables must be independent and identically distributed (i.i.d.). This requirement can be weakened; convergence of the mean to the normal distribution also occurs for non-identical distributions or for non-independent observations if they comply with certain conditions.

Here, we state a version of the CLT that applies to independent and identically distributed (i.i.d.) random variables.

The Central Limit Theorem (CLT)

Let $X_1, X_2, ..., X_n$ be i.i.d. random variables (a random sample of n independent observations from a population) with expected value $E(X_i) = \mu$ and variance $Var(X_i) = \sigma^2$. Then, if n is **large enough** (typically $n \geq 30$), the sample mean $\bar{X}$ (which is itself a random variable) has approximately a normal distribution with mean μ and variance $\frac{\sigma^2}{n}$:

$$\bar{X} \sim N\Big(\mu, \frac{\sigma^2}{n}\Big)$$

Note This result holds regardless of the shape of the X distribution (i.e. the X_i's don't have to be normally distributed!). So, the theorem is a key concept in probability theory because **it implies that probabilistic and statistical methods that work for normal distributions can be applicable to many problems involving other types of distributions**.

In other words, suppose that a large sample of observations is obtained, each observation being randomly produced in a way that does not depend on the values of the other observations, and that the average (arithmetic mean) of the observed values is computed. If this procedure is performed many times, resulting in a collection of observed averages, the central limit theorem says that if the sample size was large enough, the probability distribution of these averages will closely approximate a normal distribution.

The importance of the central limit theorem stems from the fact that, in many real applications, a certain random variable of interest is a sum of a large number of independent random variables. In these situations, we are often able to use the CLT to justify using the normal distribution. Examples of such random variables are found in almost every discipline. Here are a few:

- Laboratory measurement errors are usually modeled by normal random variables: the measurement error in a scientific experiment can be thought of as the sum of a number of underlying perturbations and errors of small magnitude.
- In communication and signal processing, Gaussian noise is the most frequently used model for noise.
- In finance, the percentage changes in the prices of some assets are sometimes modeled by normal random variables.
- When we do random sampling from a population to obtain statistical knowledge about the population, we often model the resulting quantity as a normal random variable.

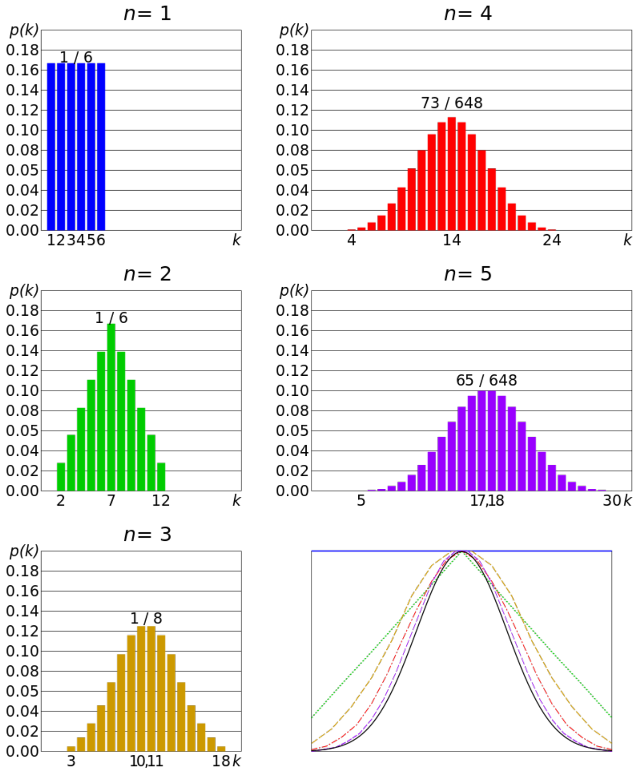
The CLT is also very useful in the sense that it can simplify our computations significantly. If we have a problem in which we are interested in a sum of one thousand i.i.d. random variables, it might be extremely difficult, if not impossible, to find the distribution of the sum by direct calculation. Using the CLT we can immediately write the distribution, if we know the mean and variance of the X_i's.

Another question that comes to mind is how large n should be so that we can use the normal approximation. The answer generally depends on the distribution of the X_i's. So, the accuracy of the approximation for a particular n depends on the shape of the original underlying distribution being sampled. Nevertheless, as a rule of thumb it is often stated that if n is larger than or equal to 30, then the normal approximation is very good.

Chapter 3

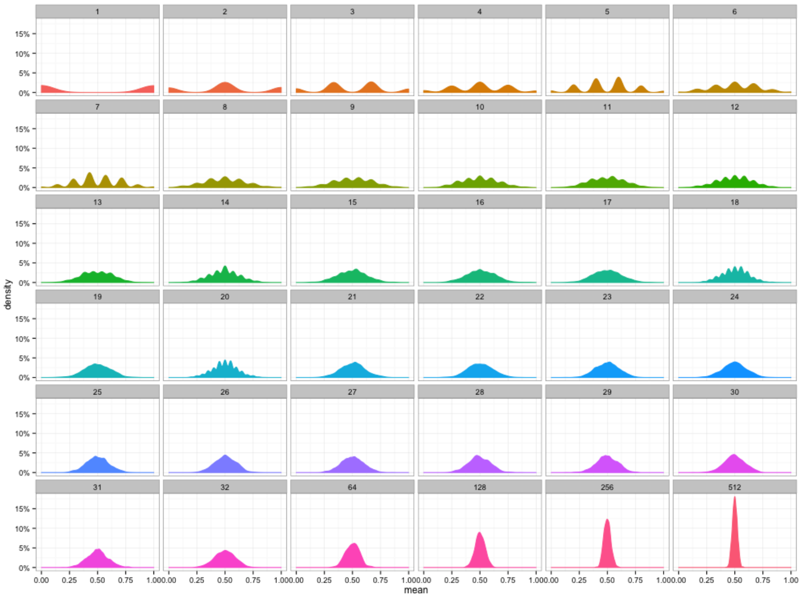
Applications and examples

A simple example of the central limit theorem is rolling many identical, unbiased dice. The distribution of the sum (or average) of the rolled numbers will be well approximated by a normal distribution.



The picture above shows a comparison of probability density functions $p(k)$ for the sum of n fair 6-sided dice to show their convergence to a normal distribution with increasing n, in accordance to the central limit theorem. In the bottom-right graph, smoothed profiles of the previous graphs are rescaled, superimposed and compared with a normal distribution (black curve).

Let's consider another simulation, using the binomial distribution, in which random 0s and 1s were generated, and then their means calculated for sample sizes ranging from 1 to 512. Looking at the picture we can observe that as the sample size increases the tails become thinner and the distribution becomes more concentrated around the mean.



Since real-world quantities are often the balanced sum of many unobserved random events, the central limit theorem also provides a partial explanation for the prevalence of the normal probability distribution. It also justifies the approximation of large-sample statistics to the normal distribution in controlled experiments.

To sum up, in a lot of situations where we use statistics, the ultimate goal is to identify the characteristics of a *population*. **The Central Limit Theorem is an approximation we can use when the population we are studying is so big, it would take a long time to gather data about each individual that's part of it**.

Population is the group of individuals that we are studying. And even though they are referred to as individuals, the elements that make a population don't need to be people.

Depending on the problem, it will be extremely hard to gather data for the entire population. So, in statistical terms, we collect *samples* from the population, and by combining the information from the samples we can draw conclusions about the population.

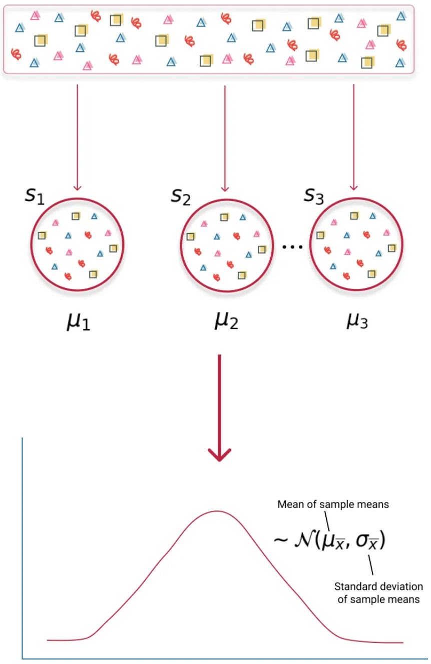
A good sample must be:

- Representative of the population,
- Big enough to draw conclusions from, which in statistics is a sample size greater or equal to 30.
- Picked at random, so you're not biased towards certain characteristics in the population.

A representative sample must showcase all the different characteristics of the population.

The Central Limit Theorem helps us balance the time and cost of collecting all the data we need to draw conclusions about the population.

Remembering the definition of Central Limit Theorem for sample means: *When we collect a sufficiently large sample of n independent observations from a population with mean n and standard deviation σ, the sampling distribution of the sample means will be nearly normal with mean $= \mu$ and standard error $= \sigma/\sqrt{n}$.*


s_1
s_2
s_3
μ_1
μ_2
μ_3
Mean of sample means
$\sim \mathcal{N}(\mu_{\bar{x}}, \sigma_{\bar{x}})$
Standard deviation of sample means

Visualizing the Central Limit Theorem. Taking samples from the population, getting their mean and creating the sample means distribution.

Chapter 4

Simulations

A simulation of the Central Limit Theorem has been implemented during the course, in particular in Exercise 3.

Exercise 3: *"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."*

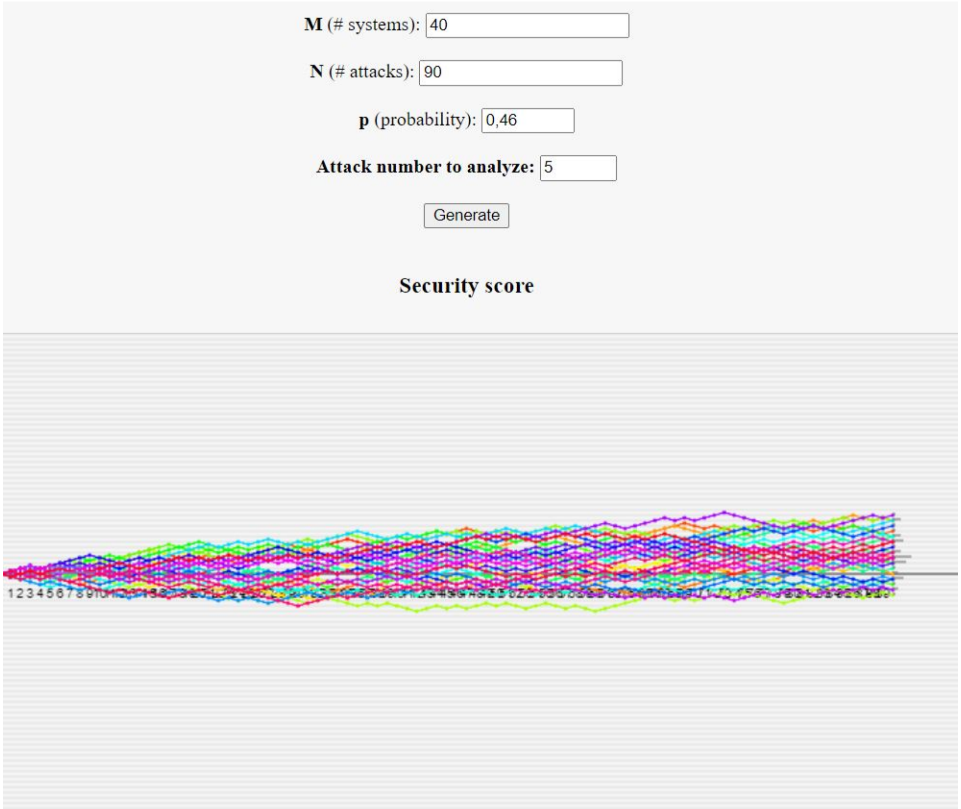
The Central Limit Theorem (CLT) states that the sum of a large number of independent and identically distributed random variables tends towards a normal distribution, regardless of the shape of the original distribution.

In the context of this exercise, each attack on a system can be considered as a random variable that takes on the values -1 (penetration) or 1 (successful shield), with probabilities p and 1-p respectively. These random variables are independent (the result of one attack does not influence another) and identically distributed (the probabilities remain constant for each attack).

If we simulate this scenario for a large number of attacks (N), the CLT would suggest that **the distribution of the sum of the security scores** (which could range from -N to N) **would approximate a normal distribution as N becomes large**. This is because the sum of the security scores is essentially the sum of N independent and identically distributed random variables.

So, **if we plot the distribution of the sum of the security scores over many simulations, we would expect it to look increasingly like a bell curve as N increases, according to the CLT**.

The score trajectories have been simulated for $M = 40$ systems, each subjected to a series of $N = 90$ attacks, assuming a penetration probability $p = 0.46$ at each attack. The output of the simulation is shown in the following picture.



M (# systems): 40

N (# attacks): 90

p (probability): 0,46

Attack number to analyze: 5

Simulation of Exercise 3

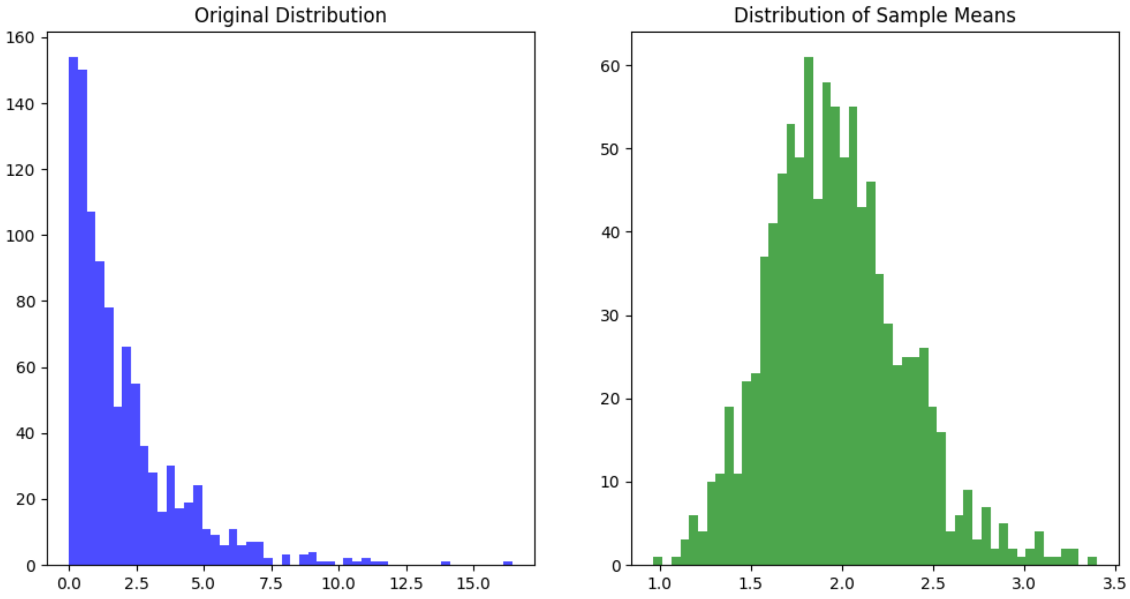
Another simulation can be implemented using Python. Consider this simple Python code that simulates the central limit theorem by generating a population with a specified distribution (exponential) and then taking random samples from that population to observe how the distribution of sample means evolves.:

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

# Set the parameters for the simulation
population_size = 1000  # Size of the population
sample_size = 30        # Size of each sample
num_samples = 1000      # Number of samples to generate

# Generate a population with a non-normal distribution (e.g., exponential distribution)
population = np.random.exponential(scale=2, size=population_size)

# Simulate the central limit theorem by taking the mean of random samples
sample_means = [np.mean(np.random.choice(population, size=sample_size)) for _ in range(num_samples)]

# Plot the original distribution and the distribution of sample means
plt.figure(figsize=(12, 6))
```

```

# Plot the original distribution
plt.subplot(1, 2, 1)
plt.hist(population, bins=50, color='blue', alpha=0.7)
plt.title('Original Distribution')

# Plot the distribution of sample means
plt.subplot(1, 2, 2)
plt.hist(sample_means, bins=50, color='green', alpha=0.7)
plt.title('Distribution of Sample Means')

plt.show()
```

In this example, I'm using an exponential distribution for the population, but it can be replaced with any other distribution. The key point is to observe how the distribution of sample means approaches a normal distribution as the sample size increases.

By running the code, we can observe the transformation from a non-normal distribution in the original population to a distribution of sample means that tends to be more normal as the sample size increases, in accordance with the central limit theorem. The histograms provide a visual comparison between the original distribution and the distribution of sample means. In particular, the left subplot shows the original distribution, and the right subplot shows the distribution of sample means.



Simulation in python

Bibliography

[1] https://www.colorado.edu/amath/sites/default/files/attached-files/lesson6_clt_0.pdf

[2] https://ocw.mit.edu/courses/15-063-communicating-with-data-summer-2003/9ec5cf038a5cb8795279df96fb0a19d7_lecture10.pdf

[3] https://www.probabilitycourse.com/chapter7/7_1_2_central_limit_theorem.php

[4] https://en.wikipedia.org/wiki/Central_limit_theorem

[5] https://online.stat.psu.edu/stat414/lesson/24/24.4

[6] https://towardsdatascience.com/central-limit-theorem