Central Limit Theorem

Chiara Iannicelli

November 2023

1 Introduction

The central limit theorem (CLT) states that the distribution of a sample variable approximates a normal distribution (i.e., a "bell curve") as the sample size becomes larger, assuming that all samples are identical in size, and regardless of the population's actual distribution shape.

Put another way, CLT is a statistical premise that, given a sufficiently large sample size from a population with a finite level of variance, the mean of all sampled variables from the same population will be approximately equal to the mean of the whole population. Furthermore, these samples approximate a normal distribution, with their variances being approximately equal to the variance of the population as the sample size gets larger, according to the law of large numbers.

2 Key component

The central limit theorem is comprised of several key characteristics. These characteristics largely revolve around samples, sample sizes, and the population of data.

- **Sampling is successive.** This means some sample units are common with sample units selected on previous occasions.
- **Sampling is random.** All samples must be selected at random so that they have the same statistical possibility of being selected.
- **Samples should be independent.** The selections or results from one sample should have no bearing on future samples or other sample results
- **Samples should be limited.** It's often cited that a sample should be no more than 10% of a population if sampling is done without replacement. In general, larger population sizes warrant the use of larger sample sizes.

- **Sample size is increasing.** The central limit theorem is relevant as more samples are selected.

3 Graphic example

A population follows a Poisson distribution (left image). If we take 10,000 samples from the population, each with a sample size of 50, the sample means follow a normal distribution, as predicted by the central limit theorem (right image).

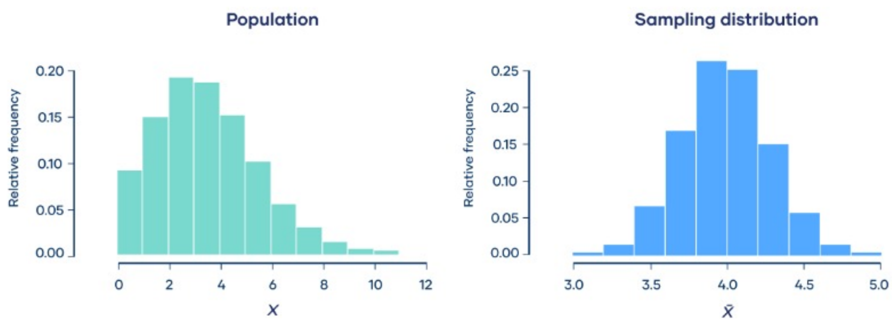

Figure 1: CLT example

4 Central Limit Theorem Formula

The shape of the sampling distribution of the mean can be determined without repeatedly sampling a population. The parameters are based on the population:

- The mean ($\mu_{\bar{x}}$) of the sampling distribution equals the mean of the population(μ).
- The standard deviation ($\sigma_{\bar{x}}$) of the sampling distribution is the population standard deviation (σ) divided by the sqare root of the sample size ($\sqrt{n}$).

Notation:

$$\bar{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

Where:

- $\bar{X}$ is the sampling distribution of the sample means.
- $\sim$ means "follows the distribution".
- N is the normal distribution.

- μ is the mean of the population.
- σ is the standard deviation of the population.
- n is the sample size.

5 Formal Definition

Let's put a formal definition to CLT:

Given a dataset with unknown distribution (it could be uniform, binomial or completely random), the sample means will approximate the normal distribution.

These samples should be sufficient in size. The distribution of sample means, calculated from repeated sampling, will tend to normality as the size of your samples gets larger.

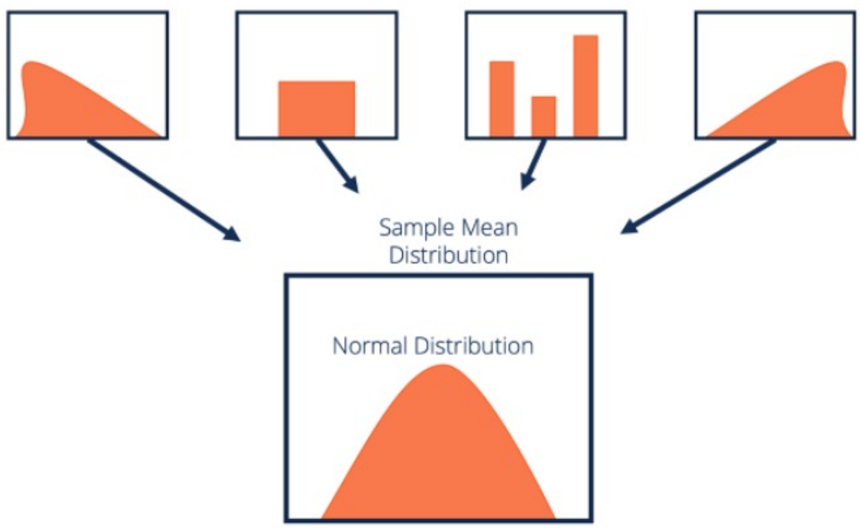

Figure 2: Tendens of normal distrbution

The central limit theorem has a wide variety of applications in many fields and can be used with python and its libraries like numpy, pandas, and matplotlib.