Central Limit Theorem (CLT)

The Central Limit Theorem is one of the most fundamental concepts in statistics and probability theory that concerns the analysis of the behavior of a sequence of random variables when the number of observations tends to infinity.

<u>It states that, given the right conditions, the sum of a large number of independent random variables, each with any probability distribution, follows an approximately normal distribution, regardless of the shape of the original distribution of the random variables.</u>

For the Central Limit Theorem to be applicable, it is necessary that the summed random variables are independent and have the same distribution. Furthermore, the number of random variables added must be sufficiently large.

The theorem states that the sum of random variables approaches a normal distribution (Gaussian distribution) as the number of random variables added increases. This is true even if the starting random variables do not follow a normal distribution.

The Central Limit Theorem is a key principle that allows us to understand how the probability distribution of a sum of random variables evolves as the number of observations increases. It is an essential tool in statistics and provides an explanation of why the normal distribution is so prevalent in practical applications.

Central Limit Formula

The Central Limit Theorem is based on the concept of sampling distribution, which is the probability distribution of a statistic for a large number of samples taken from the population

The distribution of sample means is an example of a sampling distribution.

Mathematically, the Central Limit Theorem can be expressed as:

$$\lim_{n \to \infty} \frac{\overline{X} - \mu}{\sigma / \sqrt{n}}$$

Where:

$\overline{X}$ **(random variables)**: represent a sequence of independent and identically distributed random variables. $X_1, X_2, \ldots, X_n$.
Each X_i represents an individual random variable in the sequence.

μ **(mean)**: This parameter represent the mean of each individual random variable in the sequence $\overline{X}$.

σ^2 **(Variance)**: This parameter represents the variance of each individual random variable in the sequence $\overline{X}$. It denotes the measure of the dispersion of the random variable's values around its mean.

n **(sample size)**: This refers to the number of random variables or observations in the sample. The CLT asserts that as n increases, the distribution of the sample mean becomes increasingly closer to a normal distribution.

Central Limit Characteristics

- The CLT applies to a sequence of independent and identically distributed random variables.
  - These random variables must have finite mean (μ) and finite variance (σ^2)

- The random variables in the sequence should be independent of each other. This independence condition is fundamental for the CLT to work.

- As the sample size (n) increases, the distribution of the sample mean of these independent random variables approaches a normal distribution regardless of the original distribution of the individual random variables.

- <u>The theorem is particularly relevant when dealing with large sample sizes.</u> As the sample size increases, the distribution of the sample mean converges more closely to a normal distribution, regardless of the original distribution of the individual variables.

Central Limit Theorem Proof

The Central Limit Theorem has multiple versions and proofs, each based on different mathematical techniques.

The most commonly used proof involves characteristic functions.

Let's consider $X_1, X_2, \ldots, X_n$ as independent distributed random variables with mean μ and variance σ^2.

We're interested in the distribution of the standardized sample mean $\overline{X} = \frac{X_1+X_2+\ldots+X_n}{n}$ as n tends to infinity.

The characteristic function of a random variable X is denoted by $\phi_X(t)$, defined as $\phi_X(t) = E[e^{itX}]$

where t is a real number and i is the imaginary unit.

For the sum of n independent and distributed random variables $\sum_{i=1}^{n} X_i$, the characteristic function is the product of their individual characteristic functions due to independence:

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

At this point the characteristic function of the standardized sample mean must be examined:

$$\phi_{\overline{X}_n}(t) = E[e^{itX_n}] = E[e^{it(\frac{X_1+X_2+...+X_n}{n})}] = \phi_{\frac{1}{n}\sum_{i=1}^{n} X_i}(t) = (\phi_{X1}(\frac{t}{n}))^n$$

Using Taylor expansion, for n large enough, the characteristic function of $\overline{X_n}$ becomes:

$$\phi_{\overline{X}_n}(t) \approx (1 + \frac{it\mu}{n} - \frac{t^2\sigma^2}{2n^2} + ...)^n$$

As n approaches infinity, this expression tends toward the characteristic function of the standard normal distribution:

$$\lim_{n\to\infty} \phi_{\overline{X}_n}(t) = e^{-\frac{t^2}{2}}$$

The resulting characteristic function is that of a standard normal distribution, regardless of the original distribution of X_i.

Finally, according to the inversion theorem of characteristic functions, this convergence in characteristic functions implies the convergence in distribution. Thus, the standardized sample mean $\overline{X_n}$ converges in distribution to a standard normal distribution as n approaches infinity, which verifies the Central Limit Theorem.

References:

Central Limit Theorem - *Wikipedia*
https://en.wikipedia.org/wiki/Central_limit_theorem

Central Limit Theorem Proof - *Wikipedia, Youtube*
https://en.wikipedia.org/wiki/Central_limit_theorem
https://www.youtube.com/watch?v=I-1fDNvyUdM