Central Limit Theorm: A Basic Approach

Navin Babu

November 24, 2020

1 Introduction

Sample is basically a collection of data from a large population set. It represents the population and is used to study the nature of the population dataset.

For example, let's say that that there are around 100,000 companies traded in the Indian markets. Studying all the 100,000 companies is a tedious and time consuming process. Moreover, data on the companies will be huge and mostly difficult to find. So, we can take a sample of 100 companies that are randomly selected and study the sample set, which helps us in studying the population as a whole. Population is studied on the basis of the sample data and the result is extrapolated.

1.1 Normal Distribution

Prior to understanding the Central Limit Theorem, it is vital to understand the properties of a normal distribution. In probability theory, a normal distribution is a non-uniform continuous distribution for any random variables with different probabilities and different outcomes. It is a non-uniform, non-cumulative and a continuous distribution with mean '0' and standard deviation '1'. It is also commonly referred to as a Gaussian distribution (after mathematician Carl Friedrich Gauss) or a bell curve.

Now, the probability density function of a normal distribution is as follows:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{(\mu - x)^2}{2\sigma^2}}$$

.

When the μ is '0' and σ^2 is 1, then such a distribution is called a Standard Normal Distribution. It is denoted by 'Z'.

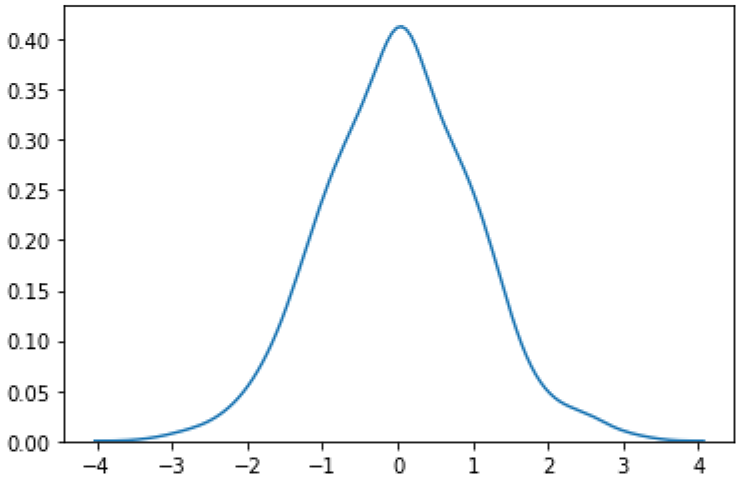

Figure 1: Normal Distribution

Properties:

1. Many continuous random variables are approximately randomly distributed
2. The distribution of discrete random variables can be approximated by a normal distribution
3. Two parameters: μ (Mean) and σ^2 (Variance)
4. It has Skewness = 1 and Kurtosis = 3
5. A normal distribution is symmetrically distributed and ranges from $-\infty$ to ∞

Figure 1 is a normal distribution with mean value '0' and standard deviation '1'. Such a distribution is called a Standard Normal Distribution, where we use a Z-table to calculate the probability under the curve. In this distribution, the probability reduces as we move away from the mean and a linear combination of two normal distributions is also a Normal distribution.

1.2 Population Parameter vs Sample Statistic

Based on the sample data, we study the nature of the population and extrapolate the results. The below table shows the representation of the estimators:

Estimator	**Population Parameter**	**Sample Statistic**
Mean	μ	$\overline{x}$
Standard Deviation	σ	S_x
No of Observation	N	n

Generally, $S_x < \sigma_x$. In other words, the sample standard deviation is a biased estimator of population parameter as it underestimates the population variance. Hence, it needs to be corrected to be an unbiased estimator of the population data.

Now, we know that $\sigma_x = \sqrt{\frac{\sum(x-\mu)^2}{n}}$. Now to make this an unbiased sample statistic, we deduct '1' from the 'n'. Therefore, $S_x = \sqrt{\frac{\sum(x-\overline{x})^2}{n-1}}$.

The sample standard deviation underestimates the population standard deviation. A sample set with a larger 'n' represents the population data better than the sample set with a smaller 'n'. So, if 'n' is larger, then there is less estimation error and thereby a smaller correction is needed. Whereas, if 'n' is smaller, then there is an higher estimation error and therefore a larger correction is required.

If n is large, then n-1 has a smaller impact, and the sample statistic moves a bit closer to the population parameter. And, when the n is smaller, then n-1 has a larger impact and the population parameter moves higher and gets closer to the population parameter.

<u>We can summarize as follows</u>:

1. If 'n' is larger - small correction needed - 'n-1' has a small percentage impact
   $\therefore S_x \uparrow$ (little bit) to get closer to the population parameter (σ_x)

2. If 'n' is smaller - large correction needed - 'n-1' has a large percentage impact
   $\therefore S_x \uparrow$ (more) to get closer to the population parameter (σ_x)

2 Central Limit Theorem

For any distribution of data, if we take a sample size of n $\geq$ 30 and calculate the mean ($\overline{x}'s$) for such samples, then the distribution of $\overline{x}'s$ is approximately normal. Let us say that we have data for 50,000 companies, which is the population data. Out of this 50,000 companies, we take a simple random sample of companies, where each sample has a minimum of 30 companies (sampling with replacement). This SRS process is repeated again and again for about 50 times and a sample mean for each individual sample set is calculated. Distribution of the sample mean ($\overline{x}'s$) is known as sampling distribution.

Figure 1 is a standard normal distribution with mean = 0 and standard deviation = 1. It is a distribution of all the sample mean's and is known as the Sampling distribution.

<u>There are three points to remember under Central Limit Therorem</u>:

1. Average of $\overline{x}'s = \mu$

2. The standard deviation of $\overline{x}'s$ is known as Standard Error (SE).

3. We can find out the sampling error of the mean by deducting the population average (μ) from the sample's average ($\overline{x}$)

3 Formulas

1. E[$\overline{x}$'s] $= \mu$
2. Standard Error $\sigma_{\overline{x}} = \frac{\sigma_x}{\sqrt{n}}$
3. Sampling Error of Mean $= \overline{x} - \mu$

Note: In case, the sample standard deviation (S_x) is not given, then we can use the population SD to calculate the standard error (SE).

4 Chebyshev's Inequality

The Weak Law of Large Numbers states that if the given random variable is independent and identically distributed (i.i.d), then as the sample size grows, the sample statistic tends towards the population parameter.

For any $\epsilon > 0$ and $0 < \delta < 1$, then

$$\lim_{n \to \infty} [|\overline{x} - \mu| < \epsilon] \geq 1 - \delta$$

and,

$$n > \frac{\sigma^2}{\epsilon^2 \delta}$$

5 A Good Estimator

The following are the desirable properties of a good estimator:

1. Unbiased: E(Sample Statistic) = Population Parameter
2. Efficient: Estimator with lowest standard error (SE) is desired
3. Consistency: If n $\uparrow$, then SE $\downarrow$. This ensures accuracy, as 'n' $\to \infty$, then SE $\to 0$. If not, then the estimator is not consistent