**7 Most Asked Questions on Central Limit Theorem**

Understanding the Importance of the Central Limit Theorem



Photo by [Michael Olsen](https://unsplash.com/@mganeolsen?utm_source=medium&utm_medium=referral) on [Unsplash](https://unsplash.com/?utm_source=medium&utm_medium=referral" \t "_blank)

In this post, we will unpack one of the most important theorems in statistics: The **Central Limit Theorem**. I will walk you through various aspects of this theorem and discuss why it lays the foundation for many statistical applications.

**#1: What is the Central Limit Theorem?**

*The Central Limit Theorem states that, given a sufficiently large sample size, the****sampling distribution****of the****sample mean****will approximate a****normal distribution****regardless of the data distribution in the population.*

If this definition sounds abstract and confusing to you, don’t worry. This post will break down this complex definition into more digestible pieces.

**#2: What is Sampling Distribution?**

Unlike the **sample distribution** (or data distribution) which describes the frequency of all possible values in the population, **sampling distribution** describes the distribution of a **sample statistic**(e.g., the sample mean) over many samples from the same population (i.e., **repeated sampling**).

For example, we collect 100 data points from the population and compute the sample mean. We repeat this process 1000 times and have 1000 different sample means. The distribution of these 1000 sample means is called the sampling distribution of the sample mean.

**#3: What are the conditions of the Central Limit Theorem?**

In order for the Central Limit Theorem to work, we need to make sure the following 3 conditions are met.

1. The sample size is **sufficiently large**.
2. The samples are **independent and identically distributed (IID) random variables**.
3. The population distribution has **finite variance**.

**#4: Does Central Limit Theorem work if the population distribution is NOT normal? Yes!**

In a population, the underlying data could follow different kinds of distributions, e.g., normal, left-skewed, right-skewed, uniform distribution, etc.

Regardless of the population distribution, the sampling distribution of the sample mean will approximate a normal distribution as long as the sample size is sufficiently large. This is a powerful theorem in statistics.

**#5: How large does the sample size need to be for the normal approximation to occur?**

*Keep in mind that the “sample size” discussed in the context of the Central Limit Theorem is not just the sample size of ONE sample, it applies to ALL samples in the sampling distribution (i.e., all samples need to have an identical and sufficiently large number of observations from the same population).*

The larger the sample size, the more closely the sampling distribution of the sample mean will follow a normal distribution.

Typically, we consider a sample size of**30** to be sufficiently large.

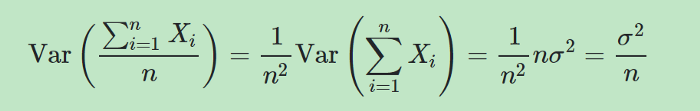
* **If the sample size is less than 30**, the central limit theorem doesn’t work anymore. The sampling distribution of the sample mean will follow normal distribution only if the population distribution is also normal.
* **If the sample size is greater than 30**, the central limit theorem applies and the sampling distribution will follow normal distribution regardless of the population distribution. However, a strongly skewed distribution requires a larger sample size.

**#6: How does sample size affect the sampling distribution of the mean?**

As the sample size increases,

* the sampling distribution will converge on a normal distribution.
* the mean of the sampling distribution will converge to the population mean
* and the standard deviation of the sample distribution (aka, the spread, [the standard error of the mean](https://towardsdatascience.com/standard-deviation-vs-standard-error-whats-the-difference-ae969f48adef" \t "_blank), SEM), which equals σ/√n, will be smaller.

This can be easily proved with a bit of math.



https://miro.medium.com/max/700/0*SN4SgB7aY29YG1nS.png

*As the sample size (n) increases, the denominator of the formula becomes bigger, then the SEM becomes****smaller****, the sampling distribution becomes****tighter****, and the****more******precise****the sample mean can be used to estimate the population mean.*

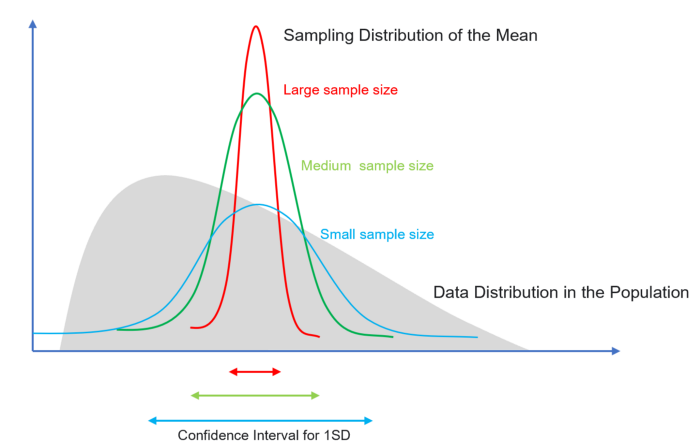


Image by author

**#7: Why Central Limit Theorem is Important?**

1. **The normality assumption**:

In real-world data, it is common to have outliers, skewness, and asymmetry. Many statistical practices, such as, hypothesis tests, confidence intervals, and t-tests are based on the normality assumption. The use of appropriate sample size and the Central Limit Theorem helps us to get around the problem of non-normality in the data.

2. **The precision of estimates**:

In practice, we typically only have ONE set of random samples. We often use statistical inference to estimate population parameters (e.g., population mean) using sample statistics (e.g., the sample mean). The sample mean is the Best linear unbiased estimator (BLUE) of the population mean. However, it is not good enough to use the point estimate (e.g., the sample mean) to infer the population mean because it will almost always be off the mark.

If we were to take random samples over and over again and compute the mean of all these sample means. It will be a very good estimate of the population mean. The sampling distribution of the sample mean also allows us to quantify the variability, the difference between the sample mean and the population mean.

*Fortunately, we don’t need repeated sampling to estimate the sampling distribution of the sample mean. The Central Limit Theorem allows us to do so based on just ONE set of random samples if we were to take random samples over and over again.*

Moreover, the property of the Central Limit Theorem tells us, as the sample size increases, our estimate of the population mean will be more precise with a smaller variability.

**Conclusion:**

The Central Limit Theorem justifies the normality assumption for statistical inference when your data is not normally distributed and tells us to increase the precision of the estimate of a population mean with a larger sample size.