**Central Limit Theorem: -**

It states, “the average of sample means will be equal to population mean” <-> “Distribution of sample means approximates a normal distribution as sample size gets larger”.

Or

In other words when we add up the means from all the samples and average it, then it will be equal to the mean of population.

Similar to that if we you find the average of standard deviation of all the samples, then it will be equal to the standard deviation of the population.

A point to note or remember is, all samples should be of same size say “N”. More accuracy can be obtained when the sample size is more.

**Steps: -**

1. Assume there is a population of N numbers.
2. Take different samples from the population each of same size n (here n should be same)
3. Take mean of all samples. This will be equal or nearby to the population mean.
4. Plot the mean of all samples
5. It will be a Normal Distribution.

As a general rule, sample sizes equal to or greater than 30 are deemed sufficient for the CLT to hold, meaning that the distribution of the sample means is fairly normally distributed. Therefore, the more samples one takes, the more the graphed results take the shape of a normal distribution.

Central Limit Theorem is a Projection and not a Prediction

Central Limit Theorem can be applied in finance with following scenario: -

* Assume an Investor wishes to analyze the overall return of stock index worth of 1000 equities.
* It’s generally Impossible to calculate the return of all 1000 equities.
* Hence the Best Practice is to take random samples from those 1000 equities across various sectors and calculate the average returns.

Assume,

* Population size = 1000
* Number of Samples = 50 (Each of size 5)
* The mean of all these 50 samples will vary from sample to sample
* This is called as Standard error =>





Few conditions are: -

● “**Randomization**”: Each sample should represent a random sample from the population, or at least follow the population distribution.

● “**10% Rule**”: The sample size must not be bigger than 10% of the entire population. Large Enough Sample Size

● Sample size n should be large enough so that

**Inferential Statistics**

Inferential Statistics allows us to make predictions from the data. Or Inferencing a population with samples in hand is called as Inferential Statistics.

With inferential statistics, you take data from [samples](https://www.statisticshowto.datasciencecentral.com/sample/)and make generalizations about a [population](https://www.statisticshowto.datasciencecentral.com/what-is-a-population/).

**Example: -**

you might stand in a mall and ask a sample of 100 people if they like shopping at [Sears](http://www.sears.com/). You could make a [bar chart](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/descriptive-statistics/bar-chart-bar-graph-examples/) of yes or no answers (that would be [descriptive statistics](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/descriptive-statistics/)) or you could use your research (and inferential statistics) to reason that around 75-80% of the population (all shoppers in all malls) like shopping at Sears.

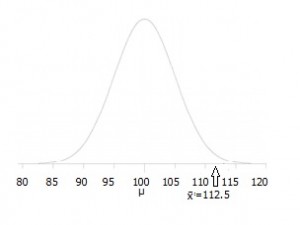
There are two main areas of inferential statistics:

1. **Estimating parameters**. This means taking a [statistic](https://www.statisticshowto.datasciencecentral.com/statistic/)from your sample data (for example the [sample mean](https://www.statisticshowto.datasciencecentral.com/sample-mean/)) and using it to say something about a population parameter (i.e. the population mean).
2. [**Hypothesis tests**](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/hypothesis-testing/). This is where you can use sample data to answer research questions. For example, you might be interested in knowing if a new cancer drug is effective. Or if breakfast helps children perform better in schools.

**Application / Implementation: -**

Let’s say you have some sample data about a potential new cancer drug,

With inferential statistics you take that sample data from a small number of people and try to determine if the data can predict whether the drug will work for everyone (i.e. the population). There are various ways you can do this, from calculating a [z-score](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/z-score/) (z-scores are a way to show where your data would lie in a [normal distribution](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/normal-distributions/) to [post-hoc](https://www.statisticshowto.datasciencecentral.com/post-hoc/) (advanced) testing.

[](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2014/10/hypothesis-testing-example.jpg)

*A hypothesis test can show where your data is placed on a distribution like this one.*

**Confidence Intervals: -**

The min and max range is called as Confidence Interval. For confidence interval we consider only Mean and not Median.

[Confidence interval](https://stattrek.com/Help/Glossary.aspx?Target=Confidence%20interval) describes the amount of uncertainty associated with a sample estimate of a population [parameter](https://stattrek.com/Help/Glossary.aspx?Target=Parameter). A confidence interval gives an estimated range of values which is likely to include an unknown population parameter, the estimated range being calculated from a given set of sample data.

A confidence interval is a range of values, derived from [sample](https://statisticsbyjim.com/glossary/sample/) [statistics](https://statisticsbyjim.com/glossary/statistics/), which is likely to contain the value of an unknown [population](https://statisticsbyjim.com/glossary/population/) [parameter](https://statisticsbyjim.com/glossary/parameter/). Because of their random nature, it is unlikely that two samples from a given population will yield identical confidence intervals. But, if you repeat your sample many times, a certain percentage of the resulting confidence intervals will contain the unknown population parameter. The percentage of these confidence intervals that contain the parameter is the confidence level of the interval.

The purpose of taking a random sample from a lot or population and computing a statistic, such as the mean from the data, is to approximate the mean of the population. How well the sample statistic estimates the underlying population value is always an issue. A confidence interval addresses this issue because it provides a range of values which is likely to contain the population parameter of interest.

Confidence intervals are constructed at a confidence level, such as 95 %, selected by the user. What does this mean? It means that if the same population is sampled on numerous occasions and interval estimates are made on each occasion, the resulting intervals would bracket the true population parameter in approximately 95 % of the cases. A confidence stated at a 1−α level can be thought of as the inverse of a significance level, α.

 A confidence interval calculates the probability that a population parameter will fall between two set values.

 Confidence intervals measure the degree of uncertainty or certainty in a sampling method.

 Most often, confidence intervals reflect confidence levels of 95% or 99%.

**Confidence Levels: -**

A **confidence level** refers to the percentage of all possible samples that can be expected to include the true population parameter. For example, suppose all possible samples were selected from the same population, and a confidence interval were computed for each sample. A 95% confidence level implies that 95% of the confidence intervals would include the true population parameter.

When a poll is reported in the media, a confidence level is often included in the results. For example, a survey might report a 95 percent confidence level. But what exactly does this mean? At first glance you might think that it means it’s 95 percent accurate. That’s close to the truth, but like many things in statistics, it’s actually a little more defined.

**Real Life Example**

**Example:** [A recent article on Rasmussen Reports](http://www.rasmussenreports.com/public_content/politics/current_events/healthcare/health_care_law) states that “**38% of Likely U.S. Voters now say their health insurance coverage has changed because of Obamacare**”. If you scroll down to the bottom of the article, you’ll see this line: “The [margin of sampling error](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/hypothesis-testing/margin-of-error/)is +/- 3 percentage points with a 95% level of confidence.”

It’s impractical to survey all 300 million+ U.S. residents, so it’s impossible to know exactly how many people would actually respond “yes my health insurance has changed.” We take a [sample](https://www.statisticshowto.datasciencecentral.com/sample/)(say, 2,000 people) and, using good statistical techniques like [simple random sampling](https://www.statisticshowto.datasciencecentral.com/simple-random-sample/), take our “best guess” at what that actual figure is (we call that unknown figure a [population](https://www.statisticshowto.datasciencecentral.com/what-is-a-population/)[parameter](https://www.statisticshowto.datasciencecentral.com/what-is-a-parameter-statisticshowto/)). What a 95 percent confidence level is saying is that if the poll or survey were repeated over and over again, the results would match the results from the actual population 95 percent of the time.

**What about “+/- 3 percentage points”?**

The **width** of the [confidence interval](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/confidence-interval/) tells us more about how certain (or uncertain) we are about the true figure in the population. This width is stated as a plus or minus (in this case,+/- 3) and is called the confidence interval. When the interval and confidence level are put together, you get a spread of percentage. In this case, you would expect the results to be 35 (38-3) to 41 (35+3) percent, 95% of the time.

### Factors that Affect Confidence Intervals (CI)

* **Population size:** this does not usually affect the CI but can be a factor if you are working with small and known groups of people.
* [**Sample Size**](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/find-sample-size/)**:** the smaller your sample, the less likely it is you can be confident the results reflect the true population [parameter](https://www.statisticshowto.datasciencecentral.com/what-is-a-parameter-statisticshowto/).
* **Percentage:** Extreme answers come with better [accuracy](https://www.statisticshowto.datasciencecentral.com/accuracy-and-precision/). For example, if 99 percent of voters are for gay marriage, the chances of error are small. However, if 49.9 percent of voters are “for” and 50.1 percent are “against” then the chances of error are bigger.

### 0% and 100% Confidence Level

A 0% confidence level means you have **no faith at all**that if you repeated the survey that you would get the same results. A 100% confidence level means there is**no doubt at all**that if you repeated the survey you would get the same results. In reality, you would never publish the results from a survey where you had no confidence at all that your statistics were accurate (you would probably repeat the survey with better techniques). A 100% confidence level doesn’t exist in statistics, unless you surveyed an entire population — and even then you probably couldn’t be 100 percent sure that your survey wasn’t open to some kind or error or bias.

**Parameter**: - Anything we estimate or calculate for population.

**Statistics**: - Anything we estimate or calculate for sample.

**Margin of Error: -**

A **margin of error** tells you **how many percentage points your results will differ**from the real population value.

For example, a 95% [confidence interval](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/confidence-interval/) with a 4 percent margin of error means that your statistic will be within 4 percentage points of the real population value 95% of the time.

More technically, the **margin of error**is the [range](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/statistics-definitions/range-statistics/) of values below and above the [sample statistic](https://www.statisticshowto.datasciencecentral.com/sample-statistic-definition-examples/) in a [confidence interval](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/confidence-interval/). The confidence interval is a way to show what the [uncertainty](https://www.statisticshowto.datasciencecentral.com/uncertainty-in-statistics/) is with a certain [statistic](https://www.statisticshowto.datasciencecentral.com/statistic/)(i.e. from a poll or survey).

For example, a poll might state that there is a 98% confidence interval of 4.88 and 5.26. That means if the poll is repeated using the same techniques, 98% of the time the true population parameter ([parameter vs. statistic](https://www.statisticshowto.datasciencecentral.com/how-to-tell-the-difference-between-a-statistic-and-a-parameter/)) will fall within the interval estimates (i.e. between 4.88 and 5.26) 98% of the time.

The margin of error can be calculated in two ways, depending on whether you have [parameters](https://www.statisticshowto.datasciencecentral.com/what-is-a-parameter-statisticshowto/)from a population or [statistics](https://www.statisticshowto.datasciencecentral.com/statistic/)from a sample:

1. Margin of error = Critical value x [Standard deviation](https://www.statisticshowto.datasciencecentral.com/relative-standard-deviation/) for the population.
2. Margin of error = Critical value x [Standard error](https://www.statisticshowto.datasciencecentral.com/what-is-the-standard-error-of-a-sample/) of the sample.

It’s a max expected difference between the true population parameter and a sample estimate of that population.

**Hypothesis Test: -**

It helps in validating the projections.

There are only two hypotheses. They are,

1. NULL Hypothesis. Denoted to H0.
2. ALTERNATE Hypotheses. Denoted by HA.

Generally, all starts with NULL Hypothesis. It’s a default Hypothesis. (Assumption is everything will happen correct).

Hypothesis testing is an act in statistics whereby an analyst [tests](https://www.investopedia.com/terms/w/wilcoxon-test.asp) an assumption regarding a population parameter. The methodology employed by the analyst depends on the nature of the data used and the reason for the analysis.

Hypothesis testing is used to assess the plausibility of a hypothesis by using sample data. Such data may come from a larger population, or from a data-generating process.

We can interpret data by assuming a specific structure our outcome and use statistical methods to confirm or reject the assumption. The assumption is called a hypothesis and the statistical tests used for this purpose are called statistical hypothesis tests.

Whenever we want to make claims about the distribution of data or whether one set of results are different from another set of results in applied machine learning, we must rely on statistical hypothesis tests.

* We always Accept or Reject only the NULL Hypothesis, and this means we never talk about the acceptance or rejection of alternate Hypothesis.

**Critical Region: -**

A critical region, also known as the rejection region, is a set of values for the test statistic for which the null hypothesis is rejected. i.e. if the observed test statistic is in the critical region then we reject the null hypothesis and accept the alternative hypothesis.

Three types of hypothesis are,

* Left Tail
* Right Tail
* Two Tails

**Steps: -**

* State NULL and Alternate Hypothesis
* Choose Level of Significance
* Find Critical Values
* Find Test Statistics
* Draw the Conclusion

Four types of Tests are :-

* Z Test
* T test
* Chi-Square Test
* F test

**Z Test: -**

A z-test is a statistical test used to determine whether two population means are different when the variances are known, and the sample size is large.

* A z-test is a statistical test to determine whether two population means are different when the variances are known and the sample size is large.
* It can be used to test hypotheses in which the z-test follows a normal distribution.
* A z-statistic, or z-score, is a number representing the result from the z-test.
* Z-tests are closely related to t-tests, but t-tests are best performed when an experiment has a small sample size.
* Also, t-tests assume the standard deviation is unknown, while z-tests assume it is known.
* **Z Test is for Mean.**

**Example: -**

Assume an investor wishes to test whether the average daily return of a stock is greater than 1%. A simple random sample of 50 returns is calculated and has an average of 2%. Assume the standard deviation of the returns is 2.5%. Therefore, the null hypothesis is when the average, or mean, is equal to 3%.

Conversely, the alternative hypothesis is whether the mean return is greater than 3%. Assume an alpha of 0.05% is selected with a [two-tailed test](https://www.investopedia.com/terms/t/two-tailed-test.asp). Consequently, there is 0.025% of the samples in each tail, and the alpha has a critical value of 1.96 or -1.96. If the value of z is greater than 1.96 or less than -1.96, the null hypothesis is rejected.

Few Conditions to perform Z Test: -

1. Sample size should be greater than 30
2. Population Standard Deviation should be known.

**T test: -**

A t-test is a type of inferential [statistic](https://www.investopedia.com/terms/s/statistics.asp) used to determine if there is a significant difference between the means of two groups, which may be related in certain features. It is mostly used when the data sets, like the data set recorded as the outcome from flipping a coin 100 times, would follow a normal distribution and may have unknown variances.

A t-test is used as a hypothesis testing tool, which allows testing of an [assumption](https://www.investopedia.com/ask/answers/073115/what-assumptions-are-made-when-conducting-ttest.asp) applicable to a population.

A t-test looks at the t-statistic, the [t-distribution](https://www.investopedia.com/terms/t/tdistribution.asp) values, and the degrees of freedom to determine the statistical significance. To conduct a test with three or more means, one must use an [analysis of variance](https://www.investopedia.com/terms/a/anova.asp).

* A t-test is a type of inferential statistic used to determine if there is a significant difference between the means of two groups, which may be related in certain features.
* The t-test is one of many tests used for the purpose of [hypothesis testing](https://www.investopedia.com/terms/h/hypothesistesting.asp) in statistics.
* Calculating a t-test requires three key data values. They include the difference between the mean values from each data set (called the mean difference), the standard deviation of each group, and the number of data values of each group.
* There are several different types of t-test that can be performed depending on the data and type of analysis required.

**Conditions: -**

1. Sample size should be less than 30.
2. Population Standard Deviation is not known.

## T-Test Assumptions

1. The first assumption made regarding t-tests concerns the scale of measurement. The assumption for a t-test is that the scale of measurement applied to the data collected follows a continuous or ordinal scale, such as the scores for an IQ test.
2. The second assumption made is that of a simple random sample, that the data is collected from a representative, randomly selected portion of the total population.
3. The third assumption is the data, when plotted, results in a normal distribution, bell-shaped distribution curve.
4. The final assumption is the homogeneity of variance. Homogeneous, or equal, variance exists when the standard deviations of samples are approximately equal.

**Degree of Freedom: -**

Degrees of freedom are the number of independent values that a statistical analysis can [estimate](https://statisticsbyjim.com/glossary/estimator/). You can also think of it as the number of values that are free to vary as you estimate [parameters](https://statisticsbyjim.com/glossary/parameter/). I know, it’s starting to sound a bit murky!

Degrees of freedom encompasses the notion that the amount of independent information you have limits the number of parameters that you can estimate. Typically, the degrees of freedom equal your [sample](https://statisticsbyjim.com/glossary/sample/) size minus the number of parameters you need to calculate during an analysis. It is usually a positive whole number.

Degrees of freedom is a combination of how much data you have and how many parameters you need to estimate. It indicates how much independent information goes into a [parameter](https://statisticsbyjim.com/glossary/parameter/) estimate. In this vein, it’s easy to see that you want a lot of information to go into parameter [estimates](https://statisticsbyjim.com/glossary/estimator/) to obtain more precise estimates and more powerful hypothesis tests. So, you want many degrees of freedom!