1. **Binomial Distribution Formula**

URL: <https://byjus.com/maths/binomial-distribution/>

**P(x:n,p) = nCx px (1-p)n-x**

**Or**

**P(x:n,p) = nCx px (q)n-x**

p = Probability of Success in a single experiment

q = Probability of Failure in a single experiment = 1 – p

**Binomial Distribution Mean and Variance**

For a binomial distribution, the mean, variance and standard deviation for the given number of success are represented using the formulas

Mean, **μ = np**

Variance, **σ2 = npq**

Standard Deviation **σ= √(npq)**

Where p is the probability of success, q is the probability of failure, where q = 1-p

**Suppose a new cancer treatment has been discovered, claiming to increase the one year survival rate for pancreatic cancer to 40%. In other words, the probability that a patient suffering from pancreatic cancer would survive for at least one year after receiving this treatment is 40%.**

**Suppose a  hospital is planning to use this treatment for its pancreatic cancer patients.**

1. **What is the probability that 2 of the 10 packets tested would turn out to be defective?**

Given that each packet has a 5% probability of being defective (p=0.05p = 0.05p=0.05), we use the binomial probability formula:

****

First, compute the binomial coefficient:



Now, compute the probability terms:

(0.05)2=0.0025(0.05)^2 = 0.0025(0.05)2=0.0025 (0.95)8(0.95)^8(0.95)8

Let's calculate the final probability.

The probability that exactly 2 out of the 10 tested packets are defective is **0.0746** or **7.46%**. ​

1. **What is the probability that, after testing these 10 packets, not more than 2 packets would turn out to be defective?**

This will be cumulative frequency of probability

Having no defective + 1 defective + 2 defective

**P(X<=2) = P(X=0) + P(X=1) + P(X=2)**

**= 10C0(0.95)10+ 10C1(0.05)1(0.95)9+ 10C2(0.05)2(0.95)8**

**= 0.5987 + 0.3151 + 0.0746 = 0.9884, or 98.84%.**

We already calculated P(X=2)=0.0746P(X = 2) = 0.0746P(X=2)=0.0746.

Now, let's compute P(X=0)P(X = 0)P(X=0) and P(X=1)P(X = 1)P(X=1) and sum them up.

The probability that not more than 2 packets are defective (i.e., at most 2 defective packets) is 0.9885 or **98.85%.**

1. **Let’s define X as the number of packets found to be defective after the 10 packets have been tested. What will be the expected value of X?**

**(Hint: You can Use Excel for the calculations for this part. Use "COMBIN(n,r)" in Excel )**

If X is defined as the number of pasta packets found to be defective after testing 10 packets, then X would follow a binomial distribution with n = 10 and p = 0.05. Now, you want to find the expected value,

which would be equal to

EV(X) = 0\*P(X=0) + 1\*P(X=1) + 2\*P(X=2) + 3\*P(X=3) + 4\*P(X=4) + 5\*P(X=5) + 6\*P(X=6) + 7\*P(X=7) + 8\*P(X=8) + 9\*P(X=9) + 10\*P(X=10).

Using a spreadsheet to calculate the individual probabilities, you can find that the expected value EV(X) = 0.5

**Suppose a new cancer treatment has been discovered, claiming to increase the one year survival rate for pancreatic cancer to 40%. In other words, the probability that a patient suffering from pancreatic cancer would survive for at least one year after receiving this treatment is 40%.**

**Suppose a  hospital is planning to use this treatment for its pancreatic cancer patients.**

**Graded Questions-Graded Quiz-14280076**

**What is the probability that exactly 4 of these patients would survive the first year after receiving this treatment?**

We can model this situation using a binomial distribution, where:

* n=10n = 10n=10 (total number of patients),
* k=4k = 4k=4 (number of survivors),
* p=0.40p = 0.40p=0.40 (probability of a single patient surviving),
* 1−p=0.601 - p = 0.601−p=0.60 (probability of a single patient not surviving).

The binomial probability formula is:



Substituting the values:



We'll now compute this probability.

The probability that exactly 4 out of the 10 patients will survive the first year after receiving this treatment is 0.2508 or 25.08%.

**What is the probability that the number of patients that survive the first year after receiving the treatment would not be more than 2?**

We need to find the probability that not more than 2 patients survive, i.e., P(X≤2) which is:

P(X≤2) = P(X=0) + P(X=1) + P(X=2)

= P(X = 0) + P(X = 1) + P(X = 2)

Using the binomial probability formula:

****

We will compute P(X=0)P(X = 0)P(X=0), P(X=1)P(X = 1)P(X=1), and P(X=2)P(X = 2)P(X=2), then sum them up.

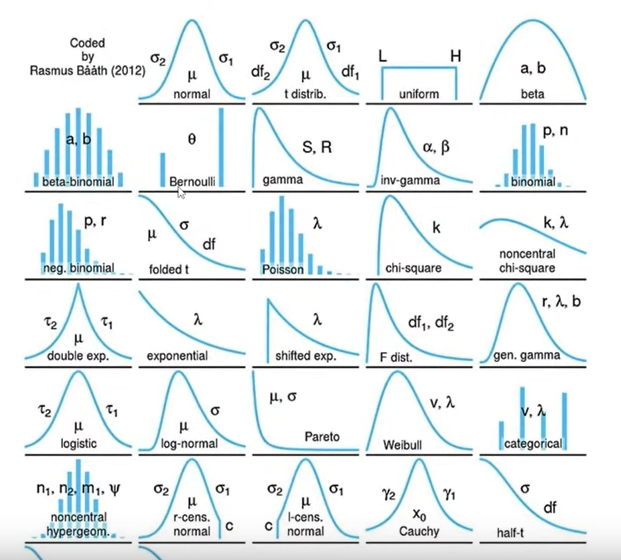
The probability that not more than 2 patients survive the first year after receiving the treatment is 0.1673 or 16.73%

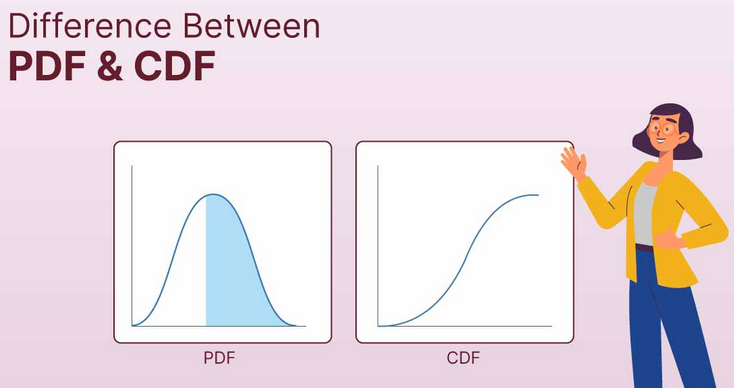
1. **Cumulative Distribution Function**

**URL**: <https://byjus.com/maths/cumulative-distribution-function/>

The Cumulative Distribution Function (CDF), of a real-valued random variable X, evaluated at x, is the probability function that X will take a value less than or equal to x. It is used to describe the [probability distribution of random variables](https://byjus.com/maths/probability-distribution/) in a table. And with the help of these data, we can easily create a CDF plot in an excel sheet

**Continuous Probability Distributions**

****

****

**What is a PDF?**

**PDF stands for** called **Probability Distribution Function**. It is two types.

1. When the [Probability Distribution Function](https://www.geeksforgeeks.org/probability-density-function)deals with **continuous random variables,** then it is called[**Probability Density Function**](https://www.geeksforgeeks.org/probability-density-function). It is a smooth curve that shows how likely different outcomes are within a range of values.

**For example,** consider the temperature in a city on a given day. The PDF could show the likelihood of **temperatures falling within certain ranges, like between 70°F and 80°F**.

PDF does not give the probability of specific values, but rather the probability of the variable falling within a small interval around a particular value. The area under the PDF curve for a range of values represents the probability of the variable falling within that range. To find the probability of a single value, it requires to calculate the integral of the PDF at that point, which means finding the area under the curve at that specific value.

1. When the [Probability Distribution Function](https://www.geeksforgeeks.org/probability-density-function)deals with **discrete random variables,** then it is called **Probability Mass Function (PMF).**

Example rolling dice.

* The probability assigned to each value must be non-negative.
* The sum of probabilities assigned to all possible values must equal to 1.

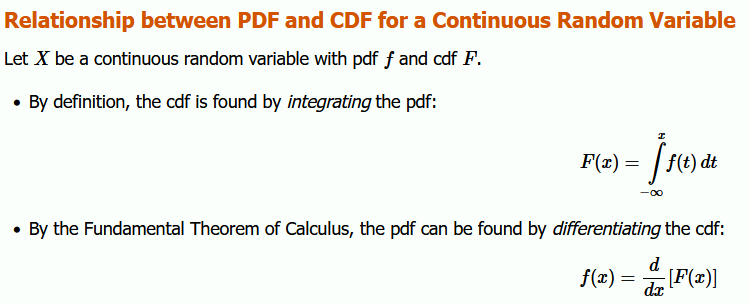
**What is a CDF?**

**CDF stands for cumulative distribution function**. The CDF complements the both types Probability Distribution Function (**PDF and PMF**) and provides a cumulative view of the probabilities linked to a random variable. Unlike the smooth curve of the PDF, the CDF appears as a step function, jumping at specific values. It shows the probability that a random variable will be less than or equal to a given value.

Starting from 0 for negative values, the CDF gradually **increases to 1** as the value of the random variable increases. For discrete random variables, the CDF rises in steps, corresponding to the probabilities of each possible outcome. With continuous random variables, it increases smoothly and reflects the combined probabilities across different intervals

Relation between PDF and CDF:

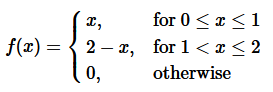
URL: <https://stats.libretexts.org/Courses/Saint_Mary%27s_College_Notre_Dame/MATH_345__-_Probability_(Kuter)/4%3A_Continuous_Random_Variables/4.1%3A_Probability_Density_Functions_(PDFs)_and_Cumulative_Distribution_Functions_(CDFs)_for_Continuous_Random_Variables>



**Example 4.1.1**

Let the randoxm variable *X*

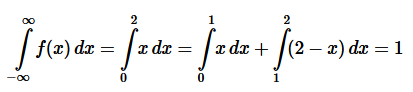
denote the time a person waits for an elevator to arrive. Suppose the longest one would need to wait for the elevator is 2 minutes, so that the possible values of *X* (in minutes) are given by the interval [0,2]. A possible **pdf** for *X* is given by

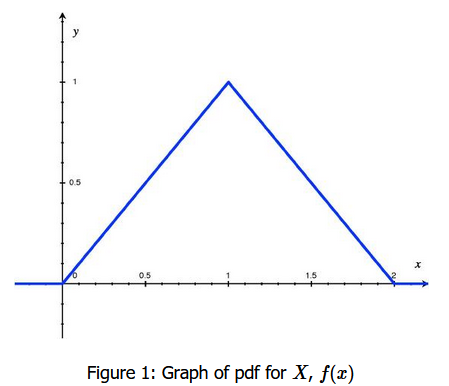


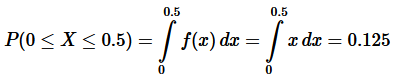
The graph of *f* is given below, and we verify that *f*

satisfies the first three conditions in Definition 4.1.1:

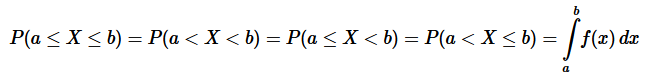
1. From the graph, it is clear that *f*(*x*)≥0 for all *x*∈R
2. Since there are no holes, jumps, asymptotes, we see that *f*(*x*) is (piecewise) continuous.
3. Finally we compute:





So, if we wish to calculate the probability that a person waits less than 30 seconds (or 0.5 minutes) for the elevator to arrive, then we calculate the following probability using the pdf and the fourth property in Definition 4.1.1:  


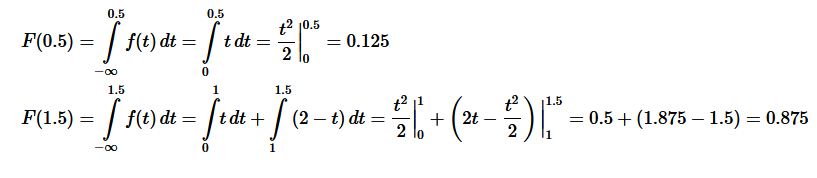
***probability for a continuous random variable is given by areas under pdf's***, then, since there is no area in a line, there is no probability assigned to a random variable taking on a single value. This does not mean that a continuous random variable will never equal a single value, only that we do not assign any probability to single values for the random variable. For this reason, we only talk about the probability of a continuous random variable taking a value in an INTERVAL, not at a point. And whether or not the endpoints of the interval are included does not affect the probability. In fact, the following probabilities are all equal:



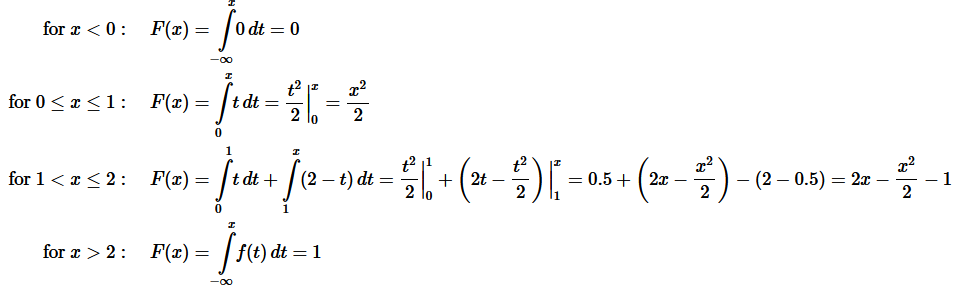
**Example 4.1.2**

**Continuing in the context of** [**Example 4.1.1**](https://stats.libretexts.org/Courses/Saint_Mary%27s_College_Notre_Dame/MATH_345__-_Probability_(Kuter)/4%3A_Continuous_Random_Variables/4.1%3A_Probability_Density_Functions_(PDFs)_and_Cumulative_Distribution_Functions_(CDFs)_for_Continuous_Random_Variables#Example_.5C(.5CPageIndex.7B1.7D.5C))**, we find the corresponding cdf.**

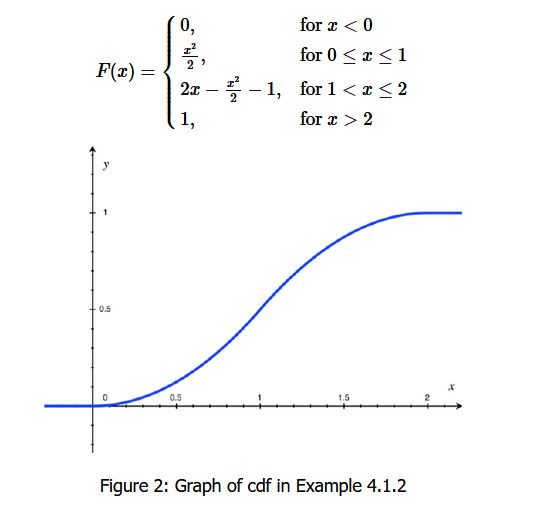
First, let's find the **cdf** at two possible values of *X*, *x*=0.5 and *x*=1.5:



Now we find *F*(*x*) more generally, working over the intervals that *f*(*x*) has different formulas:



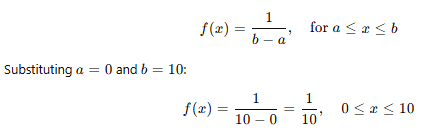
Putting this altogether, we write *F* as a piecewise function and Figure 2 gives its graph:



**Example 4.1.3**

In a uniform PDF, all the possible values have the same probability density. The figure below shows such a uniform PDF, where the possible values are 0 to 10.

Probability Density Function (PDF) For a uniform distribution, the PDF is given by:



This means that for any value within the range [0, 10], the probability density is **0.1**.

h

1. 10

Since all possible values are between 0 and 10, the area under the curve between 0 and 10 is equal to 1.

Clearly, this area is the area of a rectangle with length 10 and unknown height h. Hence, you can say that 10 \* h = 1, which gives us h = 0.1. So, the value of the PDF for all values between 0 and 10 is 0.1.

**Example 4.1.4** CDF for .5?

It will be the sum of area till x=.5. Now for 1 the probability = area= .5 \* .1=.05

**Normal Distribution Z value: Standardized Normal Variable**

****

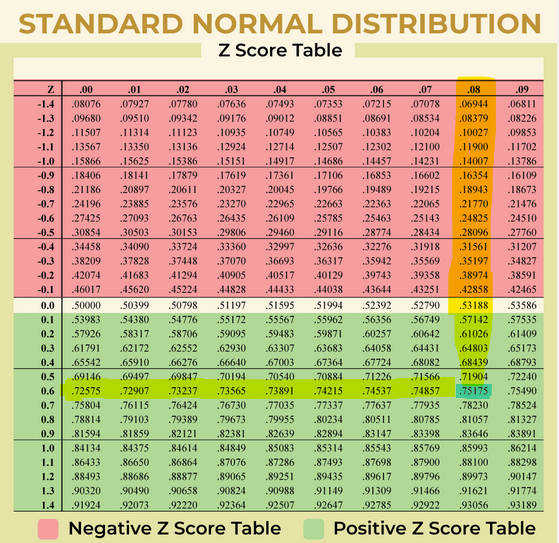
Z = (X-μ) / σ, it is the measurement of the distance of point in the normal distribution curve, from mean position μ, with respect to standard deviation σ.

|  |  |
| --- | --- |
|  |  |
|  |  |

All data that is normally distributed follows the **1-2-3 rule**. This rule states that there is a -

1. **68%** probability of the variable lying **within 1 standard deviation** of the mean
2. **95%** probability of the variable lying **within 2 standard deviations** of the mean
3. **99.7%** probability of the variable lying **within 3 standard deviations** of the mean

Z Score Table

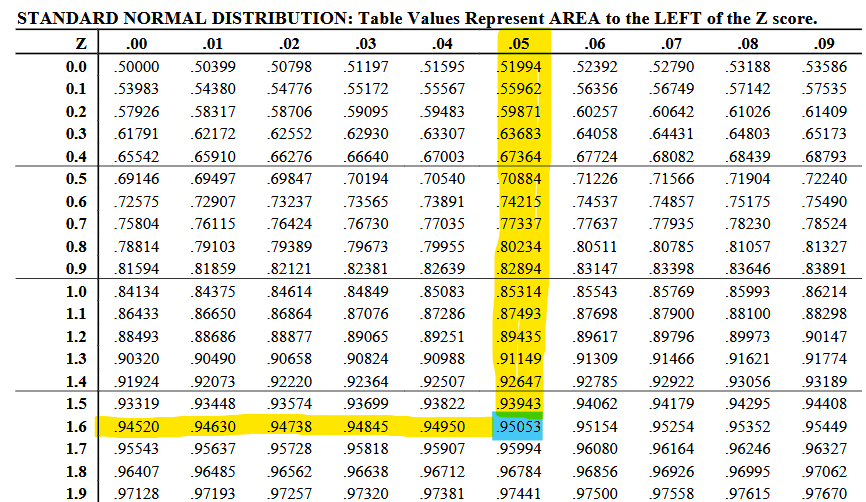


So, we can see from above table that **0.68 Z score** = **.75175, which is the cumulative probability.**

**Example 4.1.4**

Suppose mean μ = 35 and SD σ = 5, what will be P(X < 43.25) ?

So, Z = (43.25 – 35) / 5 = 1.65



Now from above table cumulative probability for Z value 1.65 is

**Example 4.1.5**

What is the probability X, lying between 44.8 and 25.2? mean μ = 35 and SD σ = 5

P(25.2 < X < 44.8)

Z Value of 44.8 = (44.8 – 35) / 5 = 1.96

Z Value of 25.2 = (25.2-35) / 5 = - 1.96

So P (-1.96 Z < 1.96) = P(1.96) – P(-1.96) = + .97500 **-** .02500 = **0.95, note that lower range we are subtracting as both are cumulative probability from 0**

[**Distributions**](https://charts.upgrad.com/dist-sd/index.html)

As you can see, the value of σ is an indicator of how wide the graph is. This will be true for any graph, not just the normal distribution. A **low**value of σ means that the graph is **narrow**, while a **high** value implies that the graph is **wider**. This will happen because the wider graph will clearly have more values away from the mean, resulting in a high standard deviation.

Again, there are some more probability distributions that are commonly seen among continuous random variables. They are not covered in this course, but if you want to go through some of them, you can use the links below -

1. [Exponential Distribution](https://online.stat.psu.edu/stat414/lesson/15/15.1)
2. [Gamma Distribution](https://online.stat.psu.edu/stat414/lesson/15/15.4)
3. [Chi-Squared Distribution](https://online.stat.psu.edu/stat414/lesson/15/15.8)

Let’s say you **work as an analyst** at a **pharma company** which manufactures an antipyretic drug (tablet form) with **paracetamol** as the active ingredient. The amount of paracetamol specified by the drug regulatory authorities is **500 mg** with a **permissible error** of **10%**. Anything below 450 mg would be a quality issue for your company since the drug will be ineffective, while above 550 mg would be a serious regulatory issue.

Mean 410 , sd 20

**Example 4.1.6**

Now, the company’s QC (Quality Control) department comes and selects a tablet at random from Batch Z2. It is interested in finding if the paracetamol level is above 450 mg or not.

What is the probability that the tablet selected by QC has a paracetamol level above 450 mg?

Let’s define X as the amount of paracetamol in the selected tablet. Now, X is a normally distributed random variable, with mean μ = 510 mg and standard deviation σ = 20 mg. Now, you have to find the probability of X being more than 450, i.e. P(X>450). Converting this to Z, you get P(X>450) = P(Z>{450-510}/20) = P(Z>-3) = 1 - P(Z<-3) = 0.9987, or 99.87%.

**Example 4.1.7**

Now, let’s say that QC decides to sample one more tablet. This time, it selects a tablet from Batch Y4. Based on previous knowledge, you know that Batch Y4 has a mean paracetamol level of 505 mg, and its standard deviation is 25 mg. This time, QC wants to check both the upper limit and the lower limit for the paracetamol level.

What is the probability that the tablet selected by QC has a paracetamol level between 450 mg and 550 mg?

Mean 505, sd 25

Z score for 450 = (450-505) / 25 = **-2.2**

Z score for 550 = (550-505) / 25 = **1.8**

**Probability = P(Z < 4) – P(Z < -2.2) = .96407- .01390 = 95%**

Graded Questions-Graded Quiz-14280142

Cumulative probability for 4.5 = area (5+2.5) \* .1 = .75

Cumulative probability for -1.5 = area (5-1.5) \* .1 = .35

So probability of range between -1.5 and 4.5 is .4

Graded Questions-Graded Quiz-14280147

The **normal distribution**, aka the **Gaussian distribution**, was discovered by **Carl Friedrich Gauss** in 1809. Gauss was trying to create a probability distribution for **astronomical errors**. Astronomical errors are the errors that were made by astronomers while observing phenomena such as distances in space.

For example, Gauss found that an astronomer trying to estimate the distance between Earth and Uranus always makes an error. This **error** is **normally distributed**, with **µ = 0 km** and **σ = 1,000 km**.

Based on the information above, what is the probability of the astronomer overestimating the distance by 2,330 km or more?

Z = (2330 – 0) / 1000 = 2.33, Cumulative probability id .9901 ( for distance 0 to 2330)

So probability for 2330 and more is 1 - .9901= 1%

Hence, what is the probability that the astronomer under- or over-estimates the distance by less than 500 km?

So P (-500 < Z < 500)

Z for -500 = -500/1000 = -.5

Z for 500 = 500/2000 = .5

P(Z< 500) – P(Z < -500) = .6915 - .3085= 38.3%

**Central Limit Theorem (CLT)**

**1. Sample size and normality**

**The larger the sample size, the more closely the sampling distribution will follow a** [**normal distribution**](https://www.scribbr.com/statistics/normal-distribution/)**.**

**When the sample size is small, the sampling distribution of the mean is sometimes non-normal. That’s because the central limit theorem only holds true when the sample size is “sufficiently large.”**

**By convention, we consider a sample size of 30 to be “sufficiently large.”**

* **When *n* < 30, the central limit theorem doesn’t apply. The sampling distribution will follow a similar distribution to the population. Therefore, the sampling distribution will only be normal if the population is normal.**
* **When *n* ≥ 30, the central limit theorem applies. The sampling distribution will approximately follow a normal distribution.**

**2. Sample size and standard deviations**

**The sample size affects the standard deviation of the sampling distribution. Standard deviation is a measure of the** [**variability**](https://www.scribbr.com/statistics/variability/) **or spread of the distribution (i.e., how wide or narrow it is).**

* **When *n* is low, the standard deviation is high. There’s a lot of spread in the samples’ means because they aren’t precise estimates of the population’s mean.**
* **When *n* is high, the** [**standard deviation**](https://www.scribbr.com/statistics/standard-deviation/) **is low. There’s not much spread in the samples’ means because they’re precise estimates of the population’s mean.**

**Conditions of the central limit theorem**

**The central limit theorem states that the sampling distribution of the mean will always follow a** [**normal distribution**](https://www.scribbr.com/statistics/normal-distribution/) **under the following conditions:**

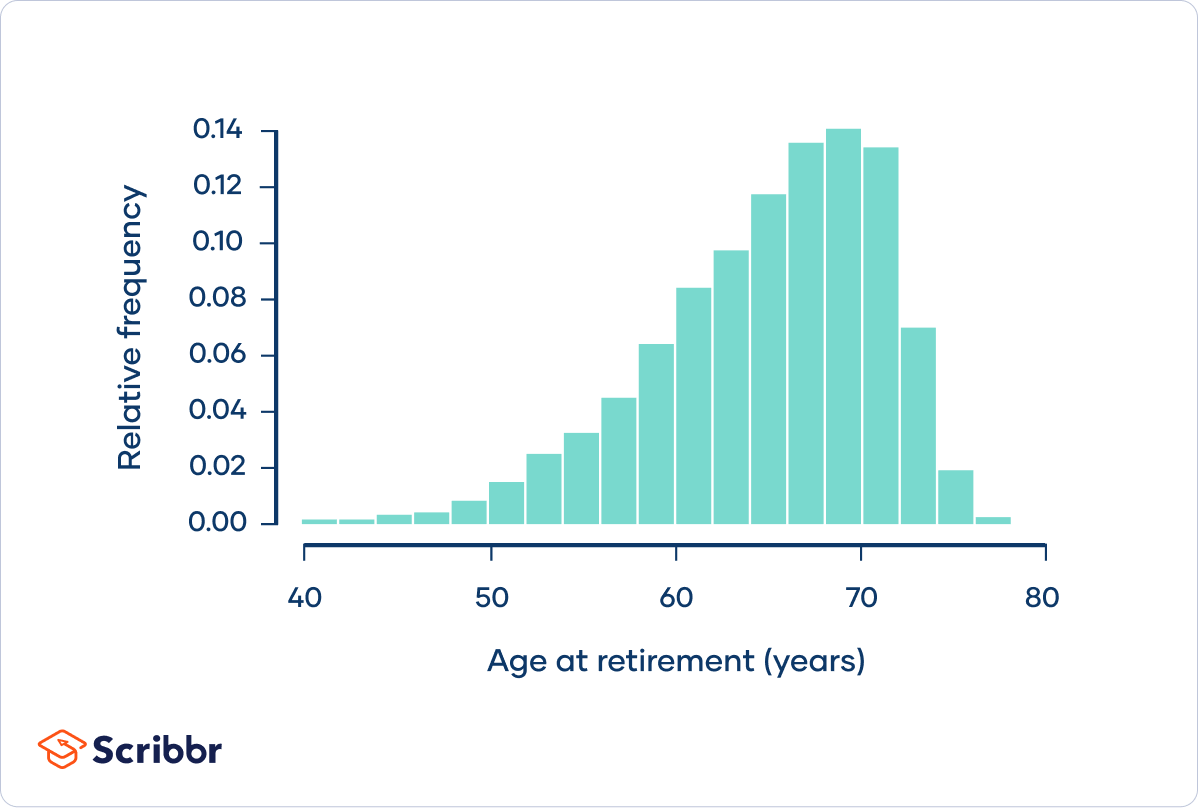
1. **The sample size is sufficiently large. This condition is usually met if the sample size is *n* ≥ 30.**
2. **The samples are independent and identically distributed (i.i.d.) random variables. This condition is usually met if the** [**sampling is random**](https://www.scribbr.com/methodology/simple-random-sampling/)**.**
3. **The population’s distribution has finite** [**variance**](https://www.scribbr.com/statistics/variance/)**. Central limit theorem doesn’t apply to distributions with infinite variance, such as the Cauchy distribution. Most distributions have finite variance.**

## Central limit theorem examples

Applying the central limit theorem to real distributions may help you to better understand how it works.

### Continuous distribution

Suppose that you’re interested in the age that people retire in the United States. The [**population**](https://www.scribbr.com/methodology/population-vs-sample/) is all retired Americans, and the distribution of the population might look something like this:



Age at retirement follows a [left-skewed](https://www.scribbr.com/statistics/skewness/#left-skew) distribution. Most people retire within about five years of the mean retirement age of 65 years. However, there’s a “long tail” of people who retire much younger, such as at 50 or even 40 years old. The population has a standard deviation of 6 years.

Imagine that you take a small **sample** of the population. You randomly select five retirees and ask them what age they retired.

Example: Central limit theorem; sample of n = 5

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 68 | 73 | 70 | 62 | 63 |

The mean of the sample is an [estimate](https://www.scribbr.com/statistics/parameter-vs-statistic/#estimating-parameters-from-statistics) of the population mean. It might not be a very precise estimate, since the sample size is only 5.

Example: Central limit theorem; mean of a small samplemean = (68 + 73 + 70 + 62 + 63) / 5

mean = 67.2 years

Suppose that you repeat this procedure 10 times, taking samples of five retirees, and calculating the mean of each sample. This is a **sampling distribution of the mean**.

Example: Central limit theorem; means of 10 small samples

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 60.8 | 57.8 | 62.2 | 68.6 | 67.4 | 67.8 | 68.3 | 65.6 | 66.5 | 62.1 |

If you repeat the procedure many more times, a histogram of the sample means will look something like this:

Although this sampling distribution is more normally distributed than the population, it still has a bit of a [left skew](https://www.scribbr.com/statistics/normal-distribution/).

Notice also that the spread of the sampling distribution is less than the spread of the population.

The **central limit theorem** says that the sampling distribution of the mean will always follow a normal distribution when the sample size is sufficiently large. This sampling distribution of the mean isn’t normally distributed because its sample size isn’t sufficiently large.

**Types of Distributions**

There are 3 main types of sampling distributions are:

* Sampling Distribution of Mean
* Sampling Distribution of Proportion
* T-Distribution
* **Sample Mean Distribution:** Sampling distribution of the sample mean of normally distributed random numbers. With increasing sample size, the sampling distribution becomes more and more centralized.

It focuses on calculating the mean or rather the average of every sample group chosen from the population and plotting the data points. The graph shows a normal distribution where the center is the mean of the sampling distribution, which represents the mean of the entire population.

We take many random samples of a given size n from a population with mean µ and standard deviation σ. Some sample means will be above the population mean µ and some will be below, making up the sampling distribution.

For any population with mean µ and standard deviation σ:

Mean, or center of the sampling distribution of x̄, is equal to the population mean, µ.

µx−​ = µ

There is no tendency for a sample mean to fall systematically above or below µ, even if the distribution of the raw data is skewed. Thus, the mean of the sampling distribution is an unbiased estimate of the population mean µ.

**Standard deviation of the sampling distribution(Standard error)**

Standard deviation of the sampling distribution measures how much the sample statistic varies from sample to sample. It is smaller than the standard deviation of the population by a factor of √n. Averages are less variable than individual observations.

The [standard deviation](https://en.wikipedia.org/wiki/Standard_deviation) of the sampling distribution of a [statistic](https://en.wikipedia.org/wiki/Statistic) is referred to as the [standard error](https://en.wikipedia.org/wiki/Standard_error_(statistics)) of that quantity. For the case where the statistic is the sample mean, and samples are uncorrelated, the standard error **SE** is:

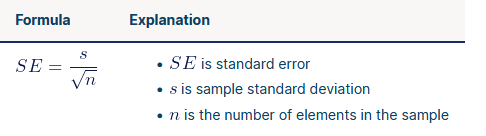
**Standard Error = Standard Deviation / √(Sample Size)**



where **σ is the standard deviation of the population distribution (SE)** of that quantity and **n is the sample size** (number of items in the sample).

**When population parameters are unknown**

When the population standard deviation is unknown, you can use the below formula to only estimate standard error. This formula takes the sample standard deviation as a [point estimate](https://www.scribbr.com/frequently-asked-questions/point-estimate-vs-interval-estimate/) for the population standard deviation.

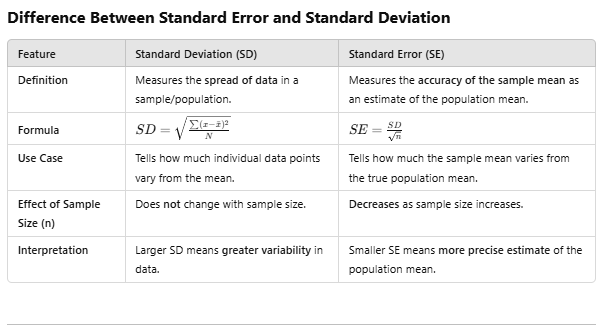


An important implication of this formula is that the sample size must be quadrupled (multiplied by 4) to achieve half (1/2) the measurement error. When designing statistical studies where cost is a factor, this may have a role in understanding cost–benefit tradeoffs.

For the case where the statistic is the sample total, and samples are uncorrelated, the standard error is: σ Σ x = σ n

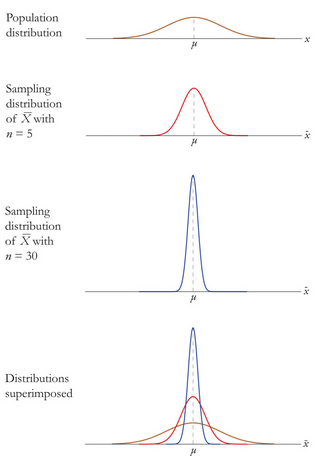


where, again, σ is the standard deviation of the population distribution of that quantity and n is the sample size (number of items in the sample).



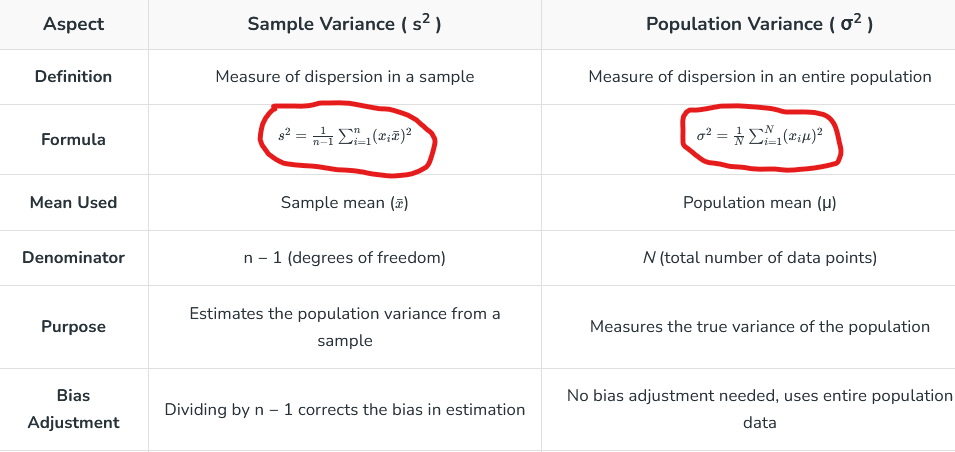
So, there are two important properties of a sampling distribution of the mean:

1. **Sampling distribution’s mean(** μX¯ )= **Population mean**(μ)
2. Sampling distribution’s standard deviation (**Standard error**) = σ/(n)1/2, where σ is the population’s standard deviation and n is the sample size
3. **For n > 30**, the sampling distribution becomes a **normal distribution**.



Central Limit Theorem for Sample Means,





Similarly, let’s say you made a sampling distribution for Y, the proportion of people that voted for INC.

The mean of this sampling distribution is μ¯X = 0.50 and the standard error is equal to 0.048.

You have to find the standard deviation for all people or, in other words, the population standard deviation σ. Using CLT, you can say that (σ/√n) = SE. Hence, (σ/√100) = 0.048, which gives σ = .48 or 48%

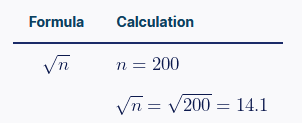
**Example:** Standard error vs standard deviation In a random sample of 200 students, the **mean math SAT score is 550**. In this case, the sample is the **200** students, while the population is all test takers in the region.

The **standard deviation** of the math scores is **180**. This number reflects on average how much each score differs from the sample mean score of 550.

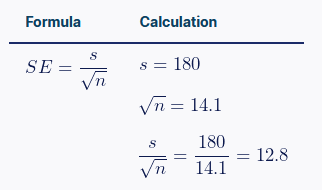
The standard error of the math scores, on the other hand, tells you how much the sample mean score of 550 differs from other sample mean scores, in samples of equal size, in the population of all test takers in the region.

**Example:** Using the standard error formula to estimate the standard error for math SAT scores, you follow two steps.

First, find the square root of your sample size (*n*).



Next, divide the sample standard deviation by the number you found in step one.



The standard error of math SAT scores is 12.8.

**Margin Of Error**

The margin of error is a statistic expressing an amount of random sampling error in a survey’s results. It asserts a likelihood that the result from a sample is close to the number one would get if the whole population had been queried. In simple words, the margin of error is the product of critical value and the standard deviation.

The margin of error is denoted by **E** and the formula is given as,

**Margin of Error = (critical value)(standard error)**



where,

n= sample size

σ= Population Standard Deviation

z = z score

**Confidence Interval**

**Confidence Interval = sample mean ± (critical value)(standard error)**

**Question:**A random sample of 30 students has average yearly earnings of 2450 and a standard deviation of 587. Find the margin of error if c = 0.95?

**Solution:**

Given  
n=30, Standard Deviation= 587

σ = 587

c = 0.95

At 95% level of confidence z = 1.96

Margin of error = z × σ/√n

= 1.96 × 587/√30

= 1.96 × 107.12

= 209.96

= 210 (approx)

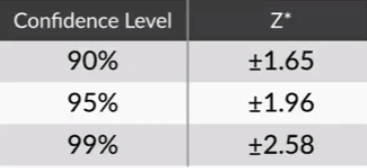
1 First, take a sample of size n.

2 Then, find the mean X and standard deviation S of this sample.

 3 Now, you can say that for a y% confidence level, the confidence interval for the population mean μ is given by



However, as you may have seen in the video above, **you cannot finish step 3 without the CLT**. The CLT lets you assume that the sample mean would be normally distributed, with mean μ



Let’s say you **work as a data analyst** at a **pharma company** which manufactures an antipyretic drug (tablet form) with **paracetamol** as the active ingredient. The amount of paracetamol specified by the drug regulatory authorities is **500 mg** with an **allowed error of 10%**. Anything below 450 mg would be a quality issue for your company since the drug would become ineffective, while anything above 550 mg would be a serious regulatory issue.

There are 10 identical manufacturing lines in the plant, each of which produces approximately 10,000 tablets per hour.

You want to take some samples, measure the amount of paracetamol, and test if the manufacturing process is running successfully. You have the resources and time to take a **sample** **of 100 tablets** and measure the paracetamol content in each.

For the 100 tablets sampled by you, you find that the **mean paracetamol content is 530 mg** and the s**tandard deviation is 100 mg**.

Now, you want to know what the average content is for all the tablets in the plant. You are thinking of reporting the average as a **confidence interval**, for which you are **95% confident**.

With this information, answer the questions given below.

**Sampling**

These four types of sampling methods are —

1. **Random Sampling:** In this method, people in the sample are selected randomly. This is similar to randomly pulling names out of a hat.

Example: Suppose you want to find out the average internet usage per person in India. You just put the names of all the Indians in a hat and pull out 100 names at random, and then calculate the average internet usage of these 100 Indians.

1. **Stratified Sampling:** Here, people are divided into subgroups and then selected randomly from those subgroups. But this is done in such a way that the final sample has the same proportions of these subgroups as the population.

Example: Again, suppose you want to find out the average internet usage per person in India. Note that 70% of Indians live in rural areas, and 30% live in urban areas. So, you would put the names of all the rural Indians in hat A and the names of all the urban Indians in hat B. Then, you’d pull 70 names out of hat A and 30 names out of hat B. Now, again, you’d have a sample of 100 Indians, but this time, your sample would be more representative of the population as its rural and urban proportions would be the same as that of the population.

1. **Volunteer Sampling:** Here, your sample is composed of people who want to volunteer for the survey.

Example: Suppose that once more, you want to find out the average internet usage per person in India. You could ask people to take an online survey, which asks them how often/much they use the internet. You could ask the same question through a telephonic survey.

The good thing about this type of sampling is that it looks unbiased and random because the survey participants are selected at random through the medium (internet, telephone) itself. There is no human interference. However, the medium will also bring in some bias. For example, an internet survey is more likely to include people who have high internet usage, whereas a telephone survey is a little more likely to have a balanced representation of heavy internet users and people who use the internet infrequently.

1. **Opportunity Sampling:** In this method, the people around and close to the surveyor form their sample space.

Example: This time, when you want to find out the average internet usage per person in India, you just ask 100 people around you about their internet usage.

Clearly, this sampling method has the potential to become extremely biased. The only good thing here, probably, is that this is a relatively convenient sampling method.

**So, there are four typical cases in which sampling is generally used:**

1. **Market research:** Suppose your company wants to launch a product that depends on people having a decent internet connection, such as Hotstar, Netflix, etc. For such a product, you need to first understand what the potential market size is. For this, you need to conduct a survey with some people and based on their data, infer parameters such as the average data usage, the willingness to adopt new technologies, etc. for the entire population.
2. **Marketing campaign efficacy**: Suppose you work for a company such as Hotstar, Netflix, etc. You want more and more people to move from your competitors’ platforms to your platform. You are planning to do this through a marketing campaign. But how should this marketing campaign be structured? How much should its budget be? What should the strategy used (free membership for a week/lower membership fees for a few weeks/etc.) be? You can use your past marketing campaigns’ data and your knowledge of sampling techniques to make these decisions.
3. **Pilot testing**: Again, let’s go to the Hotstar and Netflix example. Suppose you’ve done all the market research required, and you’ve developed the product. Now, before putting your product out there, you might want to give it a trial run. For this, you can perform what is called a pilot test. What this means is that instead of giving your product a full-fledged launch, you can just launch it partially for a few people. These people can test your product and help you decide whether it is good enough for the full launch.
4. **Quality control**: This is more of a manufacturing-centred application. Let’s say your company produces 10 million smartphones every year. This means that around 30,000 phones are produced every day. In such a situation, QA (quality assurance) becomes a function of utmost importance. Since it is difficult to check 30,000 phones every day, your company would just “sample” a few and then make decisions based on those samples.

**Hypothesis Testing**

**Hypothesis testing:**

Hypothesis testing determines if the considered hypothesis is a valid one by drawing conclusions from the results of testing methods.

**Steps for hypothesis testing:**

**Hypothesis testing involves five steps:**

1. Stating the null and alternate hypothesis
2. Specifying the level of significance (α)
3. Finding the critical values
4. Selecting a test statistic
5. Drawing the conclusions

**1. Types of hypothesis:**

**There are two types of hypotheses:**

**A. Null hypothesis (H0):**

* Null hypothesis is the actual claim from the problem.
* This is the hypothesis that should be tested, to know whether the considered hypothesis is valid or not.
* If the considered hypothesis is not valid or failed, then we state that the null hypothesis is rejected.
* Else if, the considered hypothesis is valid, we must not state that we accept the null hypothesis. We must state that we failed to reject the null hypothesis.
* In null hypothesis, there is always an equality relation between the variables. You’ll find symbols like =, ≤ and ≥.

**B. Alternate hypothesis (H1):**

* This is the alternative hypothesis for the null hypothesis.
* Unlike null hypothesis, alternate hypothesis is not tested.
* Alternate hypothesis is considered when the null hypothesis is rejected.
* In this hypothesis, there will be no equality relation between the variables. You will find symbols like <, > and ≠.

**Example1:  
A company has claimed that the overall sales of its total products by the end of the month are at least 5000.**

**Answer:**In this example,  
Null hypothesis: total products ≥ 5000  
Alternate hypothesis: total products < 5000

Example2:  
A company has claimed that the overall sales of its total products by the end of the month are more than 5000

Answer:  
In this example,  
Null hypothesis: total products ≤ 5000  
Alternate hypothesis: total products > 5000

**2. Level of significance (α):**

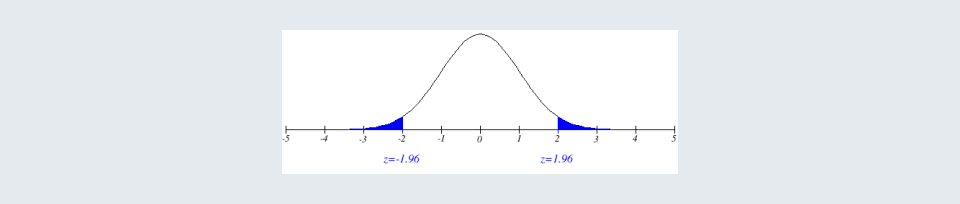
* Level of significance determines the probability of rejecting the null hypothesis.
* Level of significance must be chosen carefully. If the level of significance is high, then there are more chances that we reject the null hypothesis. Therefore, the probability for a type 1 error (discussed below) is high.
* If the level of significance considered is too low, then the probability of rejecting the null hypothesis is low. So, the probability for a type 2 error (discussed below) increases.
* α =5% is most widely considered as the level of significance because it balances perfectly between type 1 and 2 errors.  
  Assuming α=5%, leaves us confidence interval with 95%.
* This means out of all the randomly selected samples, for 95% of the sample intervals the true population mean lies in the interval and for the remaining 5% of the samples the true population mean would be outside the confidence interval.

Note:

Finally, there will be only one population mean value known as true population value. In most of the cases, we will never know the actual true population mean value.

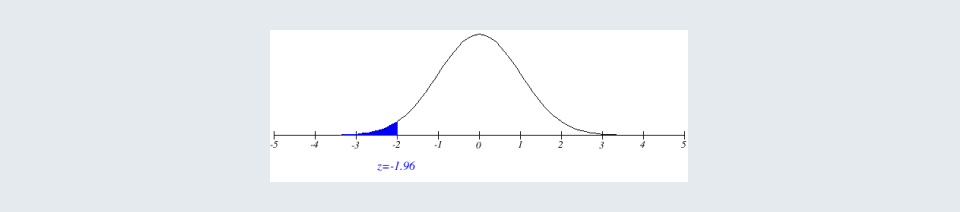
**3. Finding the critical values:**

* Now we must find the region where the null hypothesis is rejected and the region where we fail to reject the null hypothesis.
* If the alternate hypothesis is of the form,

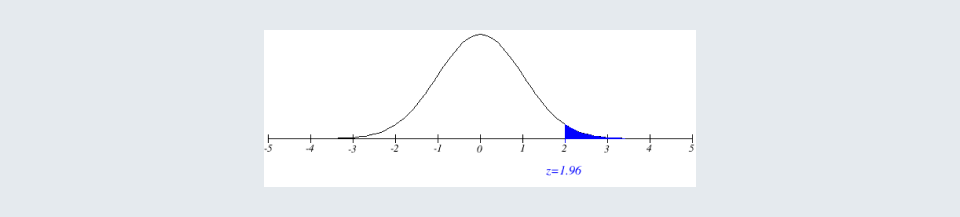
****

**H1 ≠, then it is a two tailed test  
H1 > or H1 <, then it is a one tailed test**

**Two tailed,**Let α = 5% or 0.05  
Here α is divided by 2, to serve as lower and upper boundary. I.e. αu = 0.025 (upper) same for µl (lower)

****

**Left tail,**Let α =5% or 0.05

****

**To do so, we have 2 methods to reject Null Hypothesis:**

**a. Critical value method:**

* Zc is considered as the critical value, Zc = (1 – α).
* Find the upper and lower critical value using,

UCV (upper boundary) = µ + (Zc \* σ x̅)  
LCV (lower boundary) = µ - (Zc \* σ x̅)  
µ   Population mean  
Zc   critical value  
σ x̅   standard deviation of sample

**Critical Value Approach:**

* **Step 1:** Define the null and alternative hypotheses, and choose a significance level (alpha, typically 0.05).
* **Step 2:** Calculate the test statistic based on your sample data.
* **Step 3:** Determine the critical value(s) from the appropriate statistical distribution (e.g., t-distribution, z-distribution) based on the significance level and the type of test (one-tailed or two-tailed).
* **Step 4:** Compare the calculated test statistic to the critical value(s).

**Decision:**

* If the test statistic falls within the rejection region (beyond the critical value), reject the null hypothesis.
* If the test statistic falls outside the rejection region, fail to reject the null hypothesis

**The critical value approach**

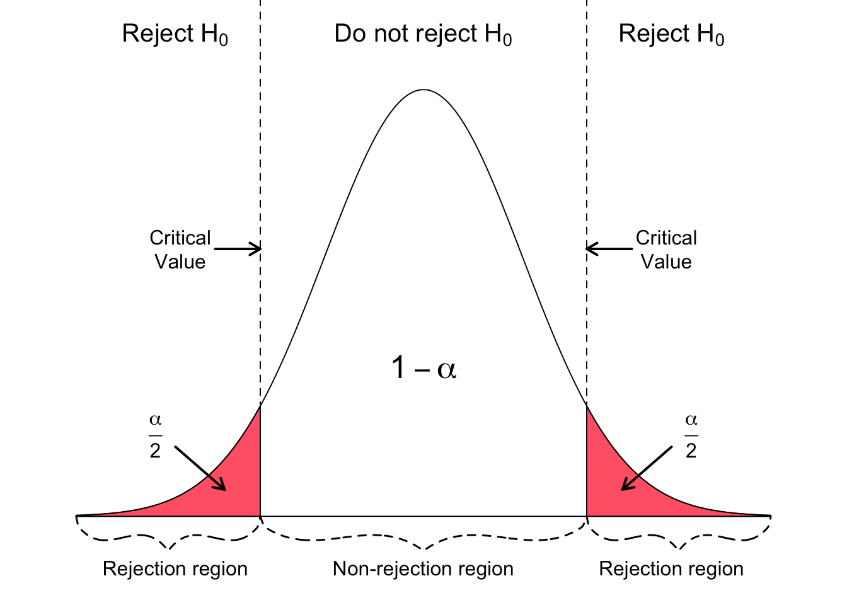
By applying the ***critical value approach***, it is determined whether or not the observed test statistic is more extreme than a defined critical value. Therefore, the observed test statistic (calculated based on sample data) is compared to the critical value (a kind of cutoff value). The null hypothesis is rejected if the test statistic is more extreme than the critical value. The null hypothesis is not rejected if the test statistic is not as extreme as the critical value. The critical value is computed based on the given significance level α

and the type of probability distribution of the idealized model. The critical value divides the area under the probability distribution curve in **rejection region(s)** and **non-rejection region**.

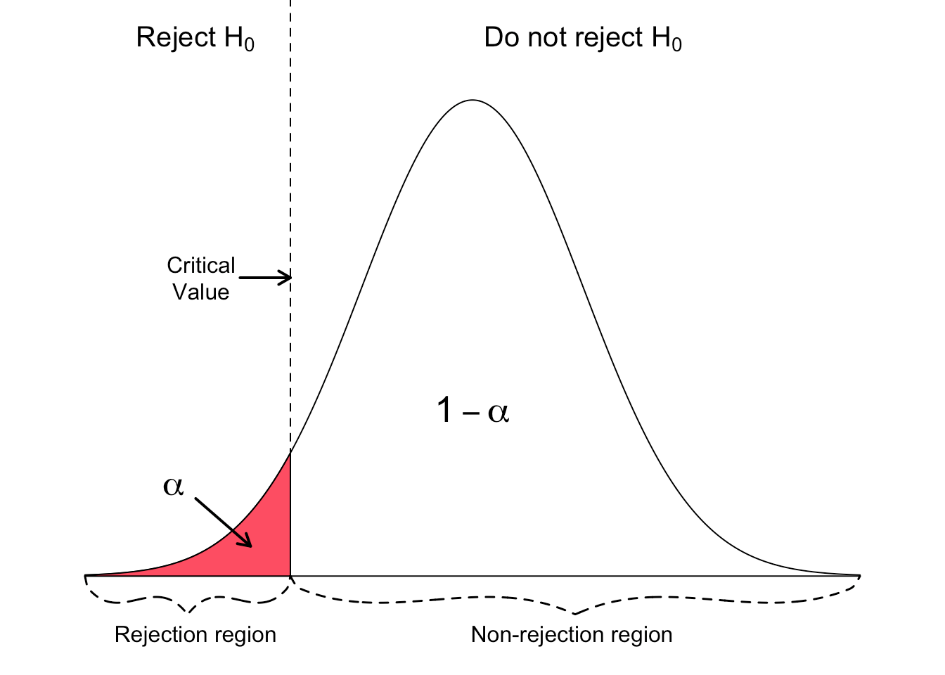
The following three figures show a right-tailed, left-tailed, and two-sided test. The idealized model in the figures, and thus H0

, is described by a bell-shaped normal distribution curve.

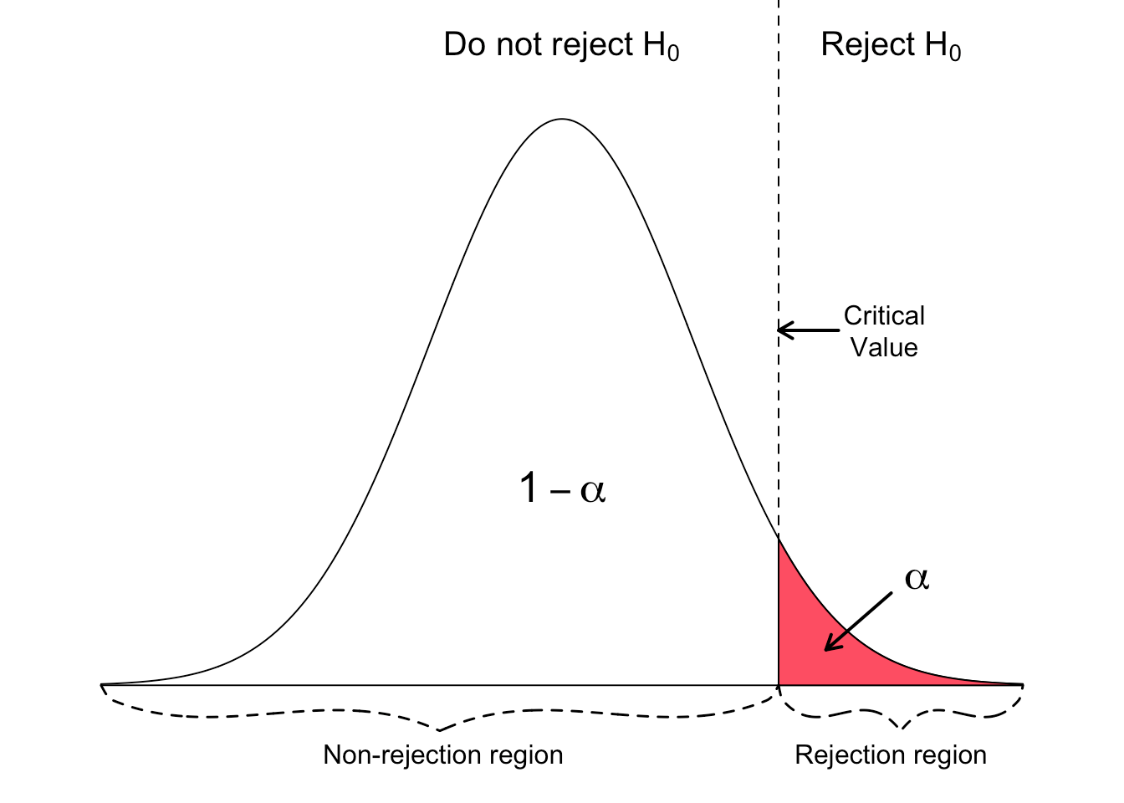
In a **two-sided** test, the null hypothesis is rejected if the test statistic is too small or too large. Thus, the rejection region for such a test consists of two parts: one on the left and one on the right.



The null hypothesis is rejected for a **left-tailed test** if the test statistic is too small. Thus, the rejection region for such a test consists of one part left from the centre.



The null hypothesis is rejected for a **right-tailed test** if the test statistic is too large. Thus, the rejection region for such a test consists of one part right from the centre.



**b. P value method:**

This method is the most widely used one. **It works only with Z-test.**

* Find the **Z score**.
* Find the p value for relative Z score.
* If the problem is a one-tailed test then, p value remains the same.
* If the problem is a two-tailed test then we must consider both upper and lower boundary, so multiply the obtained valued with 2.

**P-Value Approach:**

* **Step 1:** Define the null and alternative hypotheses, and choose a significance level (alpha, typically 0.05).
* **Step 2:** Calculate the **Z test** statistic based on your sample data.
* **Step 3:** Determine the p-value, which is the probability of observing a test statistic as extreme as, or more extreme than, the one calculated, assuming the null hypothesis is true.
* **Step 4:** Compare the p-value to the significance level (alpha).

**Decision:**

* If the p-value is less than or equal to alpha, reject the null hypothesis.
* If the p-value is greater than alpha, fail to reject the null hypothesis

## The p-value approach

For the **p**-value approach\*, the likelihood (p\*-value) of the numerical value of the test statistic is compared to the specified significance level (α) of the hypothesis test.

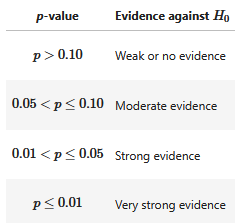
The p-value corresponds to the probability of observing sample data at least as extreme as the actually obtained test statistic. Small p-values provide evidence against the null hypothesis. The smaller (closer to 0) the p-value, the stronger is the evidence against the null hypothesis.

The null hypothesis is rejected if the p-value is less than or equal to the specified significance level α. Otherwise, the null hypothesis is not rejected.



Consequently, by knowing the p-value, any desired significance level may be assessed. For example, if the p-value of a hypothesis test is 0.01, the null hypothesis can be rejected at any significance level larger than or equal to 0.01. It is not rejected at any significance level smaller than 0.01. Thus, the p-value is commonly used to evaluate the strength of the evidence against the null hypothesis without reference to the significance level.

The following table provides guidelines for using the p-value to assess the evidence against the null hypothesis:



**Key Differences and Similarities:**

**Similarities:**

Both approaches aim to determine whether the observed data provides sufficient evidence to reject the null hypothesis.

**Differences:**

* **Method:** The critical value approach uses a fixed threshold (critical value) for comparison, while the p-value approach uses the probability of observing the data.
* **Interpretation:** The critical value approach focuses on whether the test statistic is extreme enough, while the p-value approach focuses on how likely the observed data is, assuming the null hypothesis is true.

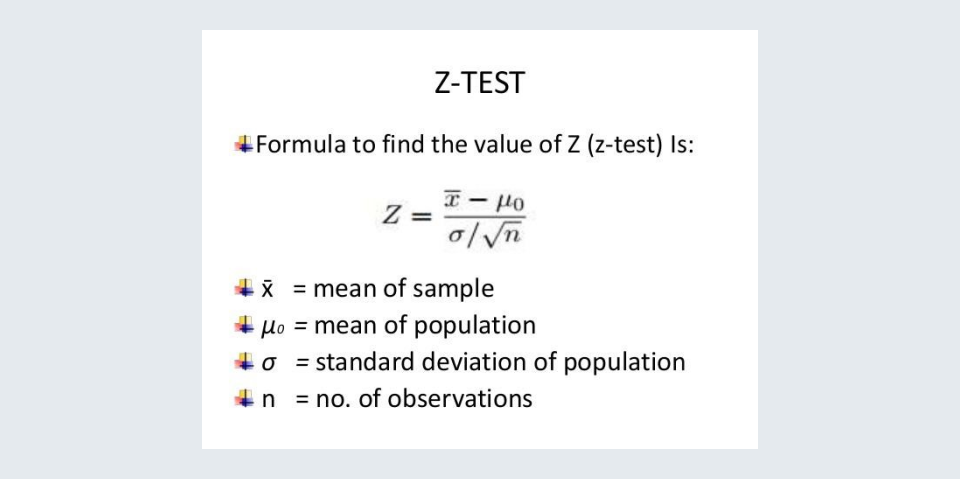
**4. Selecting a test statistic:**

* **There are various statistical methods. Here we will be learning about the first two methods, i.e. Z-test and T-test.**

1. **Z -test method**
2. **T test method**
3. **Chi-square method etc.**

**Z -test method**

* Formula for calculating the Z score is,

****

**Types of errors:**

There are two types of errors that may occur in hypothesis testing:

* **Type 1 (α):** 
  + Type 1 error occurs when we reject the null hypothesis, even when the null hypothesis is true or valid.
  + A type 1 error is represented with alpha (α).
* **Type 2 (β):** 
  + Type 2 error occurs when we fail to reject the null hypothesis, when the null hypothesis is false or invalid.
  + A type 2 error is represented with beta (β).
  + In some cases, having a type 2 error is more dangerous when compared to type 1.

**Example1:  
A pharmaceutical company is producing a medicine to cure cancer and the company claims that the medicine is effective.**

**Answer:**In this example,  
Null hypothesis: medicine produced is effective.  
Alternate hypothesis: medicine produced is not effective.

Type 1 (α): medicine is effective, but we rejected the null hypothesis.  
Type 2 (β): medicine is not effective, but we failed to reject the null hypothesis

From the above example, type 1 error does not affect people. Whereas type 2 may show some side effects or even lead to death.

This is the reason having a type 2 error is more dangerous when compared to type 1 error.

**Example1:  
An insurance company is reviewing its current policy rates. When originally setting the rates they believed that the average claim amount will be maximum Rs1,80,000. They are concerned that the true mean is actually higher than this, because they could potentially lose a lot of money. They randomly selected 40 claims and found the sample mean to be Rs1,95,000. Assuming that the standard deviation of claims is Rs50,000 and set (level of significance) α= 0.05 or 5%, test to see if the insurance company should be concerned or not.**

**Answer:**Given,  
Population standard deviation = 50000  
Mean of the sample data = 195000  
Sample size = 40

Step 1: consider the null and alternate hypothesis

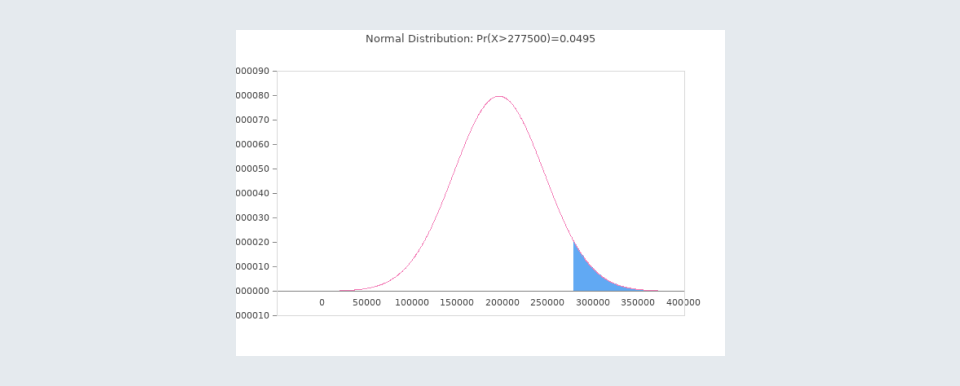
* Null hypothesis (H0) = µ ≤ 180000
* Alternate hypothesis (H1) = µ > 180000
* From the above, we can say that it is a right tailed test.

Step 2: consider the level of significance (α) and calculate the critical value

* Significance level = 5%
* Critical value = Z value of (1 – 0.05 = 0.950) = 1.65
* Common critical values obtained for relative level of significance (α),

|  |  |  |  |
| --- | --- | --- | --- |
|  | 1% | 5% | 10% |
| Critical value (1 tailed) | 2.33 | 1.65 | 1.28 |
| Critical value (2 tailed) | 2.58 | 1.96 | 1.65 |

* Upper critical value = 195000 + (1.65 \* 50000) = 277500

****

Step 3: select suitable test method and evaluate the value

* Here we choose the Z- test method. Z value = 1.897 (using Z test formula)

Note:  
If you want to learn about how to find the Z value using Z table, have a look at this video.

Step 4: compare the obtained Z value with critical value

* Z value(1.897) > critical value(1.65)

Step 5: drawing conclusions

* As the obtained Z value lies in the critical region, we reject the null hypothesis.

**Graded Question**

**Cadbury states that the average weight of one of its chocolate products ‘Dairy Milk Silk’ is 60 g. As an analyst on the internal Quality Assurance team, you would like to test whether, at the 2% significance level, the average weight is 60 g or not. A sample of 100 chocolates is collected and the sample mean size is calculated to be 62.6 g. The standard deviation, as calculated from the sample, is 10.7 g.**

**Answer the following questions in order to draw a conclusion from the test.**

**Quiz-14280411**

**Suppose you conduct a hypothesis test and observe that the values of the sample mean and sample standard deviation when n = 25 do not lead to the rejection of the null hypothesis. You calculate the p-value as 0.0667. What would happen to the p-value if you observe the same sample mean and sample standard deviation for a larger sample size, say greater than 50?**

Effect of Increasing Sample Size on the p-Value

We are given:

* Sample size n=25n = 25n=25
* p-value = 0.0667
* The same sample mean and sample standard deviation are observed when the sample size increases to n>50n > 50n>50

Understanding the Impact of a Larger Sample Size

1. **Standard Error Decreases**

SE=snSE = \frac{s}{\sqrt{n}}SE=n​s​

Since nnn increases, the denominator n\sqrt{n}n​ increases, making the Standard Error (SE) smaller.

1. Test Statistic (Z or t) Increases

Z=Xˉ−μ0SEZ = \frac{\bar{X} - \mu\_0}{SE}Z=SEXˉ−μ0​​

Since SE decreases, the Z (or t) value increases (moves further away from 0).

1. **p-Value Decreases**
   * A larger test statistic (Z or t) means the tail probability (p-value) gets smaller.
   * Since the p-value represents the probability of observing a sample mean at least as extreme as the one obtained, a larger sample provides more evidence against H0H\_0H0​.
   * As a result, the p-value will decrease when nnn increases, making it more likely to reject H0H\_0H0​.

Final Answer

The **p-value will decrease if the sample size increases** while keeping the sample mean and standard deviation constant.

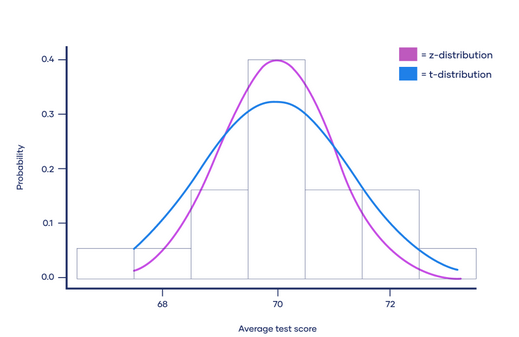
**What is a *t*-distribution?**

The *t*-distribution is a type of normal distribution that is used for smaller sample sizes. Normally-distributed data form a bell shape when plotted on a graph, with more observations near the mean and fewer observations in the tails.

The *t*-distribution is used when data are *approximately* normally distributed, which means the data follow a bell shape but the population variance is unknown. The variance in a *t*-distribution is estimated based on the [degrees of freedom](https://www.scribbr.com/statistics/degrees-of-freedom/) of the data set (total number of observations minus 1).

It is a more conservative form of the [**standard normal distribution**](https://www.scribbr.com/statistics/standard-normal-distribution/), also known as the *z*-distribution. This means that it gives a lower probability to the center and a higher probability to the tails than the standard normal distribution.

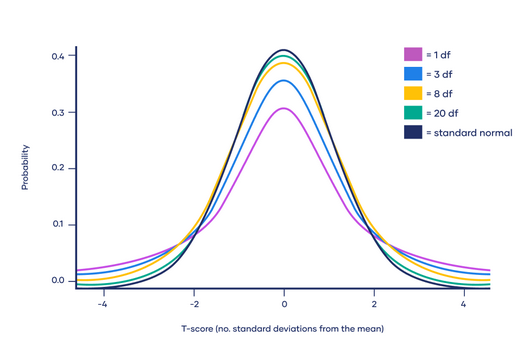
If you measure the average test score from a sample of only 20 students, you should use the *t*-distribution to estimate the confidence interval around the mean. If you use the *z*-distribution, your confidence interval will be artificially precise.



As the [**degrees of freedom**](https://www.scribbr.com/statistics/degrees-of-freedom/) (total number of observations minus 1) increases, the *t*-distribution will get closer and closer to matching the standard normal distribution, a.k.a. the *z*-distribution, until they are almost identical.

Above 30 degrees of freedom, the *t*-distribution roughly matches the *z*-distribution. Therefore, the *z*-distribution can be used in place of the *t*-distribution with large sample sizes.

The *z*-distribution is preferable over the *t*-distribution when it comes to making statistical estimates because it has a known variance. It can make more precise estimates than the *t*-distribution, whose variance is approximated using the degrees of freedom of the data.



**Significance of the t-Distribution**

1. **Degrees of Freedom and Tail Heaviness:**  
   The t-distribution [degrees of freedom](https://www.geeksforgeeks.org/degrees-of-freedom/) influence tail heaviness, with smaller values yielding heavier tails. Higher degrees of freedom make the t-distribution more akin to a standard normal distribution (mean 0, standard deviation 1), shaping its spread.
2. **Small Sample Size:**  
   The t-distribution is vital for small sample sizes, offering a precise probability distribution for [statistical inferences](https://www.geeksforgeeks.org/statistical-inference/#:~:text=Statistical%20inference%20is%20the%20process,on%20data%20from%20a%20sample.) on population parameters, especially the mean. This is crucial when the population standard deviation is unknown and must be estimated from the sample.
3. **t-Score Calculation for Inference:**  
   In situations where the standard deviation of the population is not known, the t-score (T) is calculated to make inferences about the population mean.The distinction between s and σ (population standard deviation) and the utilization of (n – 1) degrees of freedom delineate the characteristics of the t-distribution.
4. **Comparison with Z-Score and Normal Distribution:**  
   Unlike the z-score, which employs the population standard deviation, the t-score uses the estimated standard deviation from the sample. This results in a t-distribution with (n – 1) degrees of freedom, emphasizing the t-distribution’s role in handling uncertainty when estimating the population standard deviation, especially in small sample sizes.

**Interpretation of t-Distribution**

A confidence interval for the mean is a statistical range computed from the data, designed to encompass a plausible “population” mean. This interval is expressed as xˉ±t∗s/(n)xˉ±t∗s/(

​n), *t* represents a critical value obtained from the t-distribution.

Suppose we are investigating the mean study time for an exam by collecting data from a sample of 20 students. To establish a 90% confidence interval for the population mean study time using the above formula.

Let us say xˉxˉ = 4 hours, s= 1.5 hours and n =20. The critical t-value is obtained for a 90% confidence interval with 19 degrees of freedom. Assuming a critical t-value of 1.729(calculated using the t table or online calculator), the calculation results in a 90% confidence interval for the average study time, such as between 3.58 hours and 4.42 hours. This utilization of the t-distribution addresses the uncertainty linked to estimating the population mean from a sample, especially in cases where the population standard deviation is unknown.

**Properties of the t-Distribution**

* The variable in t-distribution ranges from -∞ to +∞ (**-∞ < t < +∞**).
* t- distribution will be symmetric like the normal distribution if the power of t is even in the probability density function(pdf).
* For large values of ν(i.e. increased sample size n); the t-distribution tends to a standard normal distribution. This implies that for different ν values, the shape of t-distribution also differs.
* The t-distribution is less peaked than the normal distribution at the center and higher peaked in the tails. From the above diagram, one can observe that the red and green curves are less peaked at the center but higher peaked at the tails than the blue curve.
* The value of y(peak height) attains highest at μ = 0 as one can observe the same in the above diagram.
* The mean of the distribution is equal to 0 for ν > 1 where ν = degrees of freedom, otherwise undefined.
* The median and mode of the distribution is equal to 0.
* The variance is equal to *ν / ν-2* for ν > 2 and ∞ for 2 < ν ≤ 4 otherwise undefined.

***T*-distribution and *t*-scores**

A *t*-score is the number of [standard deviations](https://www.scribbr.com/statistics/standard-deviation/) from the mean in a *t*-distribution. You can typically look up a *t*-score in a [*t*-table](https://sphweb.bumc.bu.edu/otlt/MPH-Modules/PH717-QuantCore/PH717-Module6-RandomError/PH717-Module6-RandomError11.html), or by using an online *t*-score calculator.

In statistics, *t*-scores are primarily used to find two things:

1. The upper and lower bounds of a confidence interval when the data are approximately normally distributed.
2. The *p*-value of the test statistic for *t*-tests and regression tests.

***T*-scores and confidence intervals**

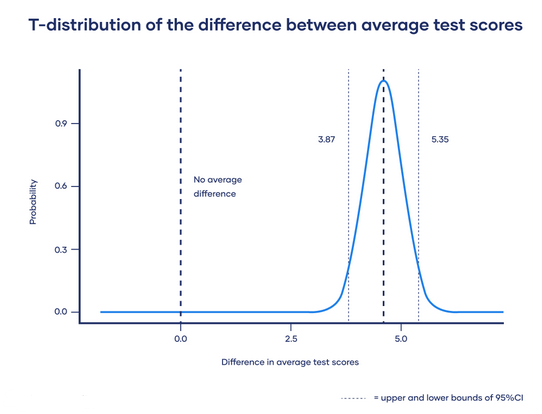
Confidence intervals use *t*-scores to calculate the upper and lower bounds of the prediction interval. The *t*-score used to generate the upper and lower bounds is also known as the **critical value** of *t*, or *t*\*.

**Example of a confidence interval**

You have sampled 20 students from two different classes to estimate the mean standardized test scores and want to know if there is a difference between the two groups.

Using a two-tailed *t*-test, you generate an estimate of the difference between the two classes and a confidence interval around that estimate. From the *t*-test you find the difference in average score between class 1 and class 2 is 4.61, with a 95% confidence interval of 3.87 to 5.35.

Because the confidence interval does not cross zero, and is in fact quite far from zero, it is unlikely that this difference in test scores could have occurred under the [null hypothesis](https://www.scribbr.com/statistics/null-and-alternative-hypotheses/) of no difference between groups.



***T*-scores and *p*-values**

Statistical tests generate a [test statistic](https://www.scribbr.com/statistics/test-statistic/) showing how far from the null hypothesis of the statistical test your data is. They then calculate a [*p*-value](https://www.scribbr.com/statistics/p-value/) that describes the likelihood of your data occurring if the null hypothesis were true.

The test statistic for [*t*-tests](https://www.scribbr.com/statistics/t-test/) and [regression tests](https://www.scribbr.com/statistics/simple-linear-regression/) is the *t*-score. While most statistical programs will automatically calculate the corresponding *p*-value for the *t*-score, you can also look up the values in a *t*-table, using your degrees of freedom and *t*-score to find the *p*-value.

The *t*-score which generates a *p*-value below your threshold for [statistical significance](https://www.scribbr.com/statistics/statistical-significance/) is known as the critical value of *t*, or *t*\*.

**Limitations of Using a T-Distribution**

* **Sensitivity to Departure from Normality:** The t-distribution assumes normality in the underlying population. When data deviates significantly from a normal distribution, reliance on the t-distribution may introduce inaccuracies in statistical inferences.
* **Limited Applicability for Large Samples:** As sample sizes increase, the t-distribution converges to the normal distribution. Therefore, for sufficiently large samples and known population standard deviation, the normal distribution is more appropriate, and using the t-distribution may not offer additional benefits.
* **Impact of Outliers and Small Sample Sizes:** The t-distribution can be sensitive to outliers, and its tails can be influenced by small sample sizes. Outliers may distort results, and in cases where the sample size is very small, the t-distribution may have heavier tails, affecting the accuracy of inferences.
* **Requires Random Sampling:** The assumptions underlying the t-distribution, such as random sampling and independence of observations, need to be met for valid results. If these assumptions are violated, the accuracy of inferences drawn from the t-distribution may be compromised.

**T- Distribution Applications**

1. [**Testing for the Hypothesis**](https://www.geeksforgeeks.org/understanding-hypothesis-testing/) **of the Population Mean:**T-distributions are commonly used in hypothesis tests regarding the population mean. This involves assessing whether a sample mean is significantly different from a hypothesized population mean.
2. **Testing for the Hypothesis of the Difference Between Two Means:**T-tests can be employed to examine if there is a significant difference between the means of two independent samples. This can be done under the assumption of equal variances or when variances are unequal.In scenarios where samples are not independent, such as paired or dependent samples, t-tests can be used to assess the significance of the mean difference between related observations.
3. **Testing for the Hypothesis about the** [**Coefficient of Correlation:**](https://www.geeksforgeeks.org/karl-pearsons-coefficient-of-correlation-methods-and-examples/)T-distributions play a role in hypothesis testing related to correlation coefficients. This includes situations where the population correlation coefficient is assumed to be zero (*ρ*=0) or when testing for a non-zero correlation coefficient (*ρ*≠0).

**Difference Between T-Distribution and Normal Distribution**

| **T-Distribution** | **Normal Distribution** |
| --- | --- |
| T-Distribution is defined by  its degree of freedom which itself depends upon the sample size | Normal distribution is defined by its mean and standard deviation |
| T- distribution is used when the sample size is small | Normal distribution is used when we have large no data points in the dataset |
| It has a heavier tail than normal distribution which means more data points are away from the mean of the distribution | Normal distribution has a lighter tail than T-distribution which means more data points lie near the mean of the distribution |
| We use T-distribution in hypothesis testing when the standard variation of the population is unknown | Normal distribution is used when the standard deviation is known |
| T-Distribution has a larger range of critical values as compared to the normal distribution as this distribution has heavier tails | Normal distribution has a smaller range as compared to t-distribution |

We can also use Python to implement t-distribution for hypothesis testing the article regarding this could be found [here](https://www.geeksforgeeks.org/python-students-t-distribution-in-statistics/).

**Conclusions**

The t-distribution serves as a vital tool in statistics, particularly when estimating the significance of population parameters with small sample sizes or unknown variations. While sharing the bell-shaped and symmetric characteristics of the normal distribution, the t-distribution distinguishes itself with heavier tails, introducing a higher likelihood of extreme values. Understanding its properties and applications is essential for accurate statistical inference in scenarios where the assumptions of normality and known population standard deviation are not met.

**When to Use the t-Distribution?**

Student’s t Distribution is used when :

* The sample size is 30 or less than 30.
* The population standard deviation(σ) is unknown.
* The population distribution must be unimodal and skewed.

**T test method**

In Z-test method problem is it depends on

* 1. Population Mean (P value / Critical Value)
  2. Population SD (P value / Critical Value)

Which are not available in real problem.

Formula for calculating the T score is,  
t = (x̅ - µ0) / (s /√n)  
x̅   sample mean  
µ0 mean of second sample / Population mean?  
s   standard deviation of sample  
n   sample size

**5. Drawing the conclusions:**

* **With the critical value method, if the Z score or T score lies in the upper or lower critical region, we can reject the null hypothesis.**
* **With the critical value method, if the Z score or T score does not lie in the upper or lower critical region, we fail to reject the null hypothesis.**
* **With the p value method, if the p value is ≤ level of significance, we can reject the null hypothesis.**

**With the p value method, if the p value is > level of significance, we fail to reject the null hypothesis.**

**1-sample t-test: testing the value of a population mean**

**To test, if the population mean of data is likely to be equal to a given value**

**scipy.stats.ttest\_1samp()**

**stats.ttest\_1samp(data['column'], x)**

**#where x is the mean value you want to test**

**2-sample t-test: testing for difference across populations**

**scipy.stats.ttest\_ind()**

**stats.ttest\_ind(column\_1,column\_2)**

**Paired tests: repeated measurements on the same individuals**

**stats.ttest\_rel()**

**stats.ttest\_rel(column\_1,column\_2)**

**So what did you learn in this session?**

1. **T-distribution:**
   * **A T-distribution is used whenever the standard deviation of the population is unknown**
   * **The degrees of freedom of a T-distribution is equal to sample size n - 1**
   * **For sample size ≥ 30, the T-distribution becomes the same as the normal distribution**
   * **The output values and results of both t-test and z-test are same for sample size ≥ 30**
2. **Two-sample mean test - paired:**
   * **It is used when your sample observations are from the same individual or object**
   * **During this test, you are testing the same subject twice**
3. **Two-sample mean test - unpaired:**
   * **During this test, you are not testing the same subject twice**
   * **It is used when your sample observations are independent**
4. **Two-sample proportion test:**
   * **It is used when your sample observations are categorical, with two categories**
   * **It could be True/False, 1/0, Yes/No, Male/Female, Success/Failure, etc.**
5. **A/B Testing:**
   * **A/B testing is a direct industry application of the two-sample proportion test**
   * **It is a widely used process in digital companies in the ecommerce, manufacturing and advertising domains**
   * **It provides a way to test two different versions of the same element and see which one performs better**

**You can download the lecture notes for the module in the next section. The lecture notes include a summary of the entire module.**

[**Lecture Notes - Hypothesis Testing**](https://kh3-ls-storage.s3.us-east-1.amazonaws.com/UPGrad/Lecture%20Notes%20-%20Hypothesis%20Testing.pdf)

**Note : The number of stores (page number 6 or page which contains Figure 5) should be 36 instead of 25.**