**Central Limit Theorem, Hypothesis testing, T-distribution: what, why, and how?**

[[Nikita Volzhin](https://medium.com/@volzhinnv?source=post_page-----f531a2c395bd--------------------------------)](https://medium.com/@volzhinnv?source=post_page-----f531a2c395bd--------------------------------)

[Nikita Volzhin](https://medium.com/@volzhinnv?source=post_page-----f531a2c395bd--------------------------------)

·

[Follow](https://medium.com/m/signin?actionUrl=https%3A%2F%2Fmedium.com%2F_%2Fsubscribe%2Fuser%2F4e81753aeef4&operation=register&redirect=https%3A%2F%2Fmedium.com%2F%40volzhinnv%2Fcentral-limit-theorem-hypothesis-testing-t-distribution-what-why-and-how-f531a2c395bd&user=Nikita+Volzhin&userId=4e81753aeef4&source=post_page-4e81753aeef4----f531a2c395bd---------------------post_header-----------)

7 min read

·

Nov 25

124

1

**Intro**

Recently, I decided to revise my knowledge of statistics and watch an online course with formulas and all this stuff which should have been boring. Nonetheless, I was so surprised and thrilled because I finally understood the concepts, meaning, and purposes of all these statistical things I studied in university and could not grasp them. Most importantly now I know why I need them and how to use them in real life in my data scientist career. I got so filled with all this mathematical and analytical magic that I want to share it with you to help anyone who struggles with this now (and to egoistically make it stuck in my mind even better:)). In this article, I will explain the concepts regarding the **Central Limit Theorem, hypothesis testing, and t-distribution**, which I completely could not get in university.

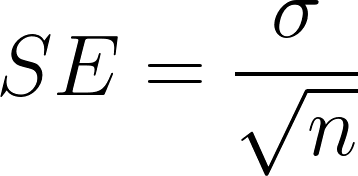
**NB:** I assume that my reader is already familiar with **normal distribution** (sometimes referred to as Laplacian or Gaussian distribution) and normalization to **z-score**. If you feel not comfortable with these two topics yet, there are a bunch of great articles on Medium. Enjoy! Now, let’s begin.

**Central Limit Theorem (CLT)**

That’s a really cool thing on which all inferential statistics is based. Imagine you have a population (ANY population with ANY distribution) and you start drawing one by one samples from this population of a fixed size of **n** elements in each. As a true statistician, you measure a certain characteristic of elements (that is **x**), and then you measure the **mean** (**x̄**) of this characteristic in each sample.

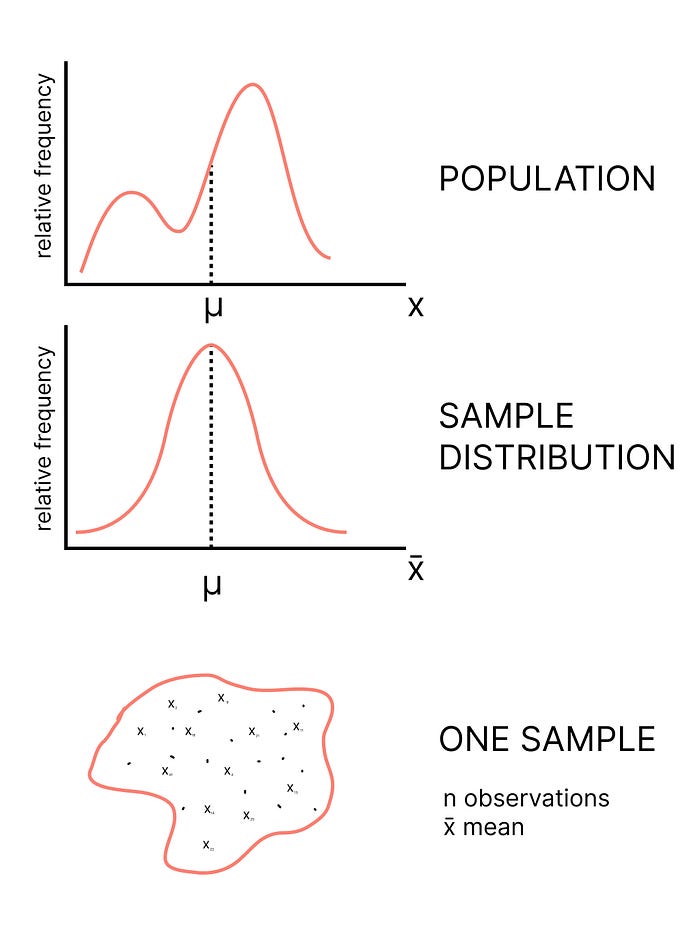
After measuring quite some sample means you have a distribution of these means (that is called **sample distribution**). So the theorem is about this sample distribution. It claims:

* **The mean of this distribution is the same as the population mean (μ)**
* That sample distribution is **NORMAL**
* The standard deviation of this distribution (that is **standard error**) is



Standard error formula (σ is the standard deviation of the population)

I drew the picture below to better illustrate the theorem.



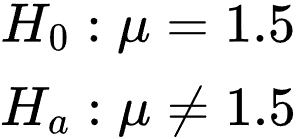
**NB**: there could be a population with **arbitrary** distribution, but sample distribution will always be **normal**.

These three facts are true when **n** is large (approaches infinity). But how large on practice? During statistics courses, it is said that if n is **larger than 30**, this all is approximately true. Moreover in this case, if population distribution is normal, we can assume that sample and population standard deviations are equal (**σ = s**).

Why that is cool? Because we can take a **sample** from a population and make a conclusion about the whole **population** mean based on the sample’s mean (remember that the sample distribution mean and population mean should be equal). More about this in the next section!

**Hypothesis testing**

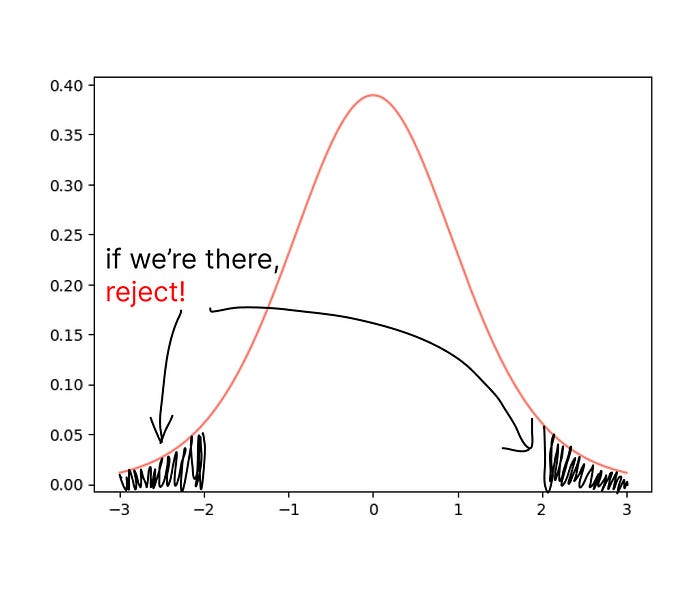
Imagine. My grandmother has a farm with sheep. She claims that her sheep are longer than average sheep (because she raised them with love;)). Let’s check it! First, we need a hypothesis. In statistics to prove something we need to prove that the opposite is false. The opposite hypothesis, which we want to disprove is called a **zero hypothesis (H0)**, the one we want to prove is an **alternative hypothesis (Ha)**. I know exactly that a usual (NOT my grandma’s) sheep is 1.5 meters in length with let’s say standard deviation of 0.2 meters. So hypotheses are:



* *Zero hypothesis:*Grandma’s sheep are usual with average length of 1.5 meters.
* *Alternative hypothesis*: Grandma’s sheep are NOT usual; their average length is NOT 1.5 meters.

Then, according to CLT, if we start measuring random samples of 40 usual sheep, their length averages will be normally distributed with a 1.5-meter mean and standard error (that is the standard deviation of these length averages) of 0.2/sqrt(40) ≈ 0.032 meters. Knowing this we know everything about this distribution and we can normalize the length average of any sheep sample. Let’s do it with grandmother’s sample!

I will randomly select 40 sheep from the grandmother’s herd and measure their length. Let’s say I got the result of 1.6 meters. If after normalization this value falls significantly far away from the mean of the normalized sample distribution, we can reject the zero hypothesis, and say that the grandma’s sheep are longer. Otherwise, we do not have enough evidence to reject the zero hypothesis. I illustrated this in the picture below.



But how do we know what is significantly far away? It depends on how confident we want to be when rejecting the hypothesis. Very often 95% **confidence interval** is used. This means that the probability that our sample mean falls to the far regions that I shaded on the graph above is 5% (that is referred to as **alpha**), which is 2.5% for the left-hand shaded region and 2.5% for the right-hand one. There are special z-score tables that can tell you what acceptable z-scores are for this confidence level (for 95% that is [-1.96, 1.96]).

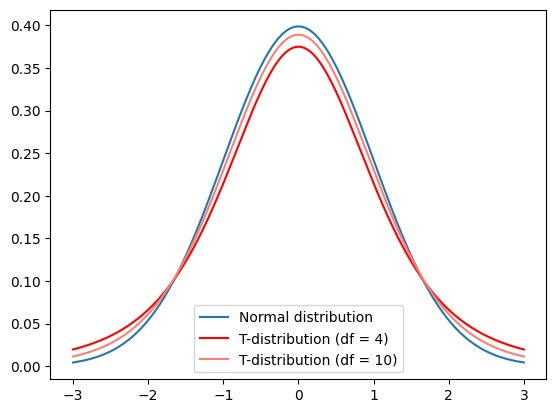
I let you calculate the z-score and make a decision about whether the grandma sheep are special on your own. But what if we do not know the population variance or samples are too small? Let’s see in the next section!

**T-distribution**

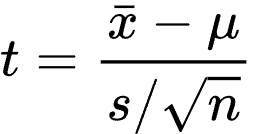
A normal distribution is very idealistic. For it to be exactly accurate we must know the population variance and mean, while the sample size must be large. When these conditions do not hold**t-distribution** (also called **Student’s t-distribution**) will help. It is a very powerful tool and, in most cases, in practice not normal but t-distribution is used.

What is magnificent about it, for t-distribution we must know only **the number of degrees of freedom (df)**. What is **df**? That is a bit complex term, but in simple words that is the number of observations we need to calculate a statistic (e.g. mean) such that if we know the value of the statistic and the value of exactly **df** observations, the rest of the observations are obvious to us. For instance, if I know that the mean of 3 observations is 4, and two of them are 2 and 5, the third one is obviously 5. Hence, I need to know only two of them to know for sure what is the third one. Often, **df = n — 1** (**n** — number of observations, so here df is the sample size minus one).

With an increase in the number of df t-distribution approaches normal, that is why when df is larger than 30, we can approximate it as normal.



How to proceed with it? Exactly the same way as with normal! First, we normalize, but this time we use not z-score but **t-score**:



All good. It is exactly like z-score but instead of population variance(which we don’t know often), we use sample variance. After calculating this we go to the appropriate table for t-distribution, check the probability, and compare it with our alpha score. But you don’t have to do it, because now there are tons of amazing sites that will construct the t-distribution (and not only) and let you check the probability for a specific t-score. One of such I’m leaving below.

**[Distribution Calculator](https://gallery.shinyapps.io/dist_calc/?source=post_page-----f531a2c395bd--------------------------------" \t "_blank)**

[Edit description](https://gallery.shinyapps.io/dist_calc/?source=post_page-----f531a2c395bd--------------------------------" \t "_blank)

[gallery.shinyapps.io](https://gallery.shinyapps.io/dist_calc/?source=post_page-----f531a2c395bd--------------------------------" \t "_blank)

**NB**: for t-distribution we need our population to be normally distributed. At least for small sample sizes. Otherwise, we have a bunch of different options such as *the Mann–Whitney U-test* but I will not talk about it in this article. However, for large sample sizes, our population does not have to be normally distributed, as mentioned in *Miroslav Tushev’s article*:

**[Does my sample have to be normally distributed for a t-test?](https://miroslavtushev.medium.com/does-my-sample-have-to-be-normally-distributed-for-a-t-test-7ee91aaaca2a?source=post_page-----f531a2c395bd--------------------------------" \t "_blank)**

[While reading up on t-test’s normality assumption I came across a lot of conflicting information. Most of the resources…](https://miroslavtushev.medium.com/does-my-sample-have-to-be-normally-distributed-for-a-t-test-7ee91aaaca2a?source=post_page-----f531a2c395bd--------------------------------" \t "_blank)

[miroslavtushev.medium.com](https://miroslavtushev.medium.com/does-my-sample-have-to-be-normally-distributed-for-a-t-test-7ee91aaaca2a?source=post_page-----f531a2c395bd--------------------------------" \t "_blank)

So practically, t-distribution allows us to draw quite accurate conclusions about a population based **only on sample values**. This tool is actually cool, and it is used in medical tests, AB tests, and so on.

**Conclusion**

I know I have not mentioned a lot (P-value, comparing Two Population Means, one-side confidence bound, etc. etc.) But if one understands the concepts of this article, what I did not mention will be simple to grasp too. I hope this helped anyone at least a bit better understand these concepts or see how and why we need these things. There are many more cool statistical things that were not clear to me in university that I understand now and want to share. So, I will publish more articles in the future! Thank you for reading this article and I am happy to see any suggestions and ideas in the comments!

***P.S. Love data, science, and data science!***