Hypothesis Testing

When we perform an analysis on a population sample — the analysis could be **descriptive**, **inferential**, or **exploratory** in nature we get certain information from which we can make claims about the entire population. This kind of claim or assumption is called a hypothesis.

Difference between Inferential Statistics & Hypothesis Testing?

Inferential statistics is used to find the mean of a population parameter when you have no initial number to start with. So, you start with the sampling activity and find out the sample mean. Then, you estimate the population mean from the sample mean using the confidence interval.

Hypothesis testing is used to confirm your conclusion (or hypothesis) about the population mean (which you know from EDA or your intuition). Through hypothesis testing, you can determine whether there is enough evidence to conclude if the hypothesis about a population parameter is true or not.

Hypothesis Testing starts with the formulation of these two hypotheses:

- **Null hypothesis** (H_0): The status quo
- **Alternate hypothesis** (H_1): The challenge to the status quo

These two hypotheses are always opposing

Example

- We can take a criminal trial example

In our criminal trial example, the defendant was considered innocent. So, the **null hypothesis claims that he is innocent**, just like he was before the murder charge. **Null Hypothesis is denoted by H_0**

The alternate hypothesis, or research hypothesis as it is also called, is the claim that opposes the null hypothesis. If you were the prosecutor in the trial, your claim would be that the defendant is guilty, and you would try to prove this. So, the alternate hypothesis is an assumption that competes with the null hypothesis. **Alternate Hypothesis is denoted by H_1**

Outcome

- If the **defendant is found guilty then the jury rejects the null hypothesis and is in favor of the null hypothesis.**
- If the jury acquits the defendant, it means that there is not enough evidence to support the alternate hypothesis.

You can use the following rule to formulate the null and alternate hypotheses:

- The null hypothesis always has the following signs: = OR ≤ OR ≥
- The alternate hypothesis always has the following signs: ≠ OR > OR <

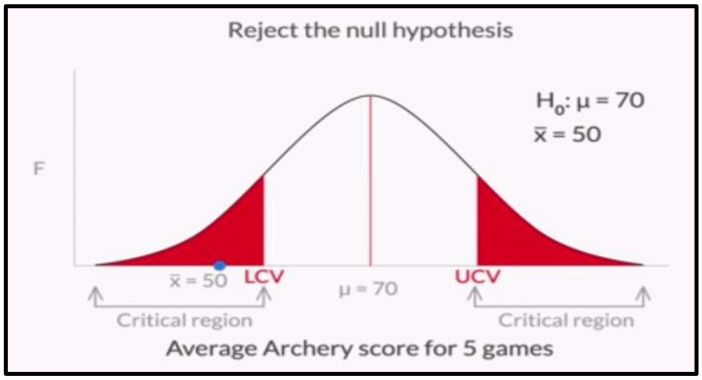
Making the decision to either reject or fail to reject the null hypothesis

Example: I am playing a game of Archery with one of my friend she tells me that here avg. score is 70. But over 5 games of archery let’s say

- Her score comes out to be 20. I would say she is bragging that her avg. score is 70
- Here score comes out to be 65. I would say she is not lying.

My null and alternate hypothesis would be :

H_0: $\mu = 70$
H_1: $\mu \neq 70$
Let's say avg. score x is 50
Population mean is 70 in this case because this the value about which we have hypothesis.



If a value falls **below or above LCV or UCV** then we say that Apurva's mean score is not 70. If the score lies below the LCV or above the UCV then we would **reject the null hypothesis**.

If a value falls between the LCV or UCV region that is the **acceptance region** then we would **fail to reject the null hypothesis**.

Every problem scenario would be different that means the critical region is not necessary that it would be present on both sides. This is actually determined using the '**sign**' in the alternate hypothesis.

$\neq$ in H_1 → Two-tailed test → Rejection region on both sides of distribution

< in H_1 → Lower-tailed test → Rejection region on left side of distribution

> in H_1 → Upper-tailed test → Rejection region on right side of distribution

<u>Critical Value Method</u>

<u>AC Sales Example</u>

H_0: $\mu = 350$ i.e., mean demand of AC units per month per store during summer
H_1: $\mu \neq 350$
Sample mean x=370.6 (of 36 stores=n)
Standard Deviation (SD) of sample $\sigma_{x(bar)}$ = (Pop. SD)/ √36 i.e., 90/6 = 15
α=5% this is the **level of significanc**e that means in 5% of the cases we reject the null hypothesis

- Area under critical region is 5% (alpha value) (two tailed test)
- Area under acceptance region is 95%

Z_c = 1-0.025=0.975 (Area till critical point)

So, in critical value method z-score is calculated for critical points .

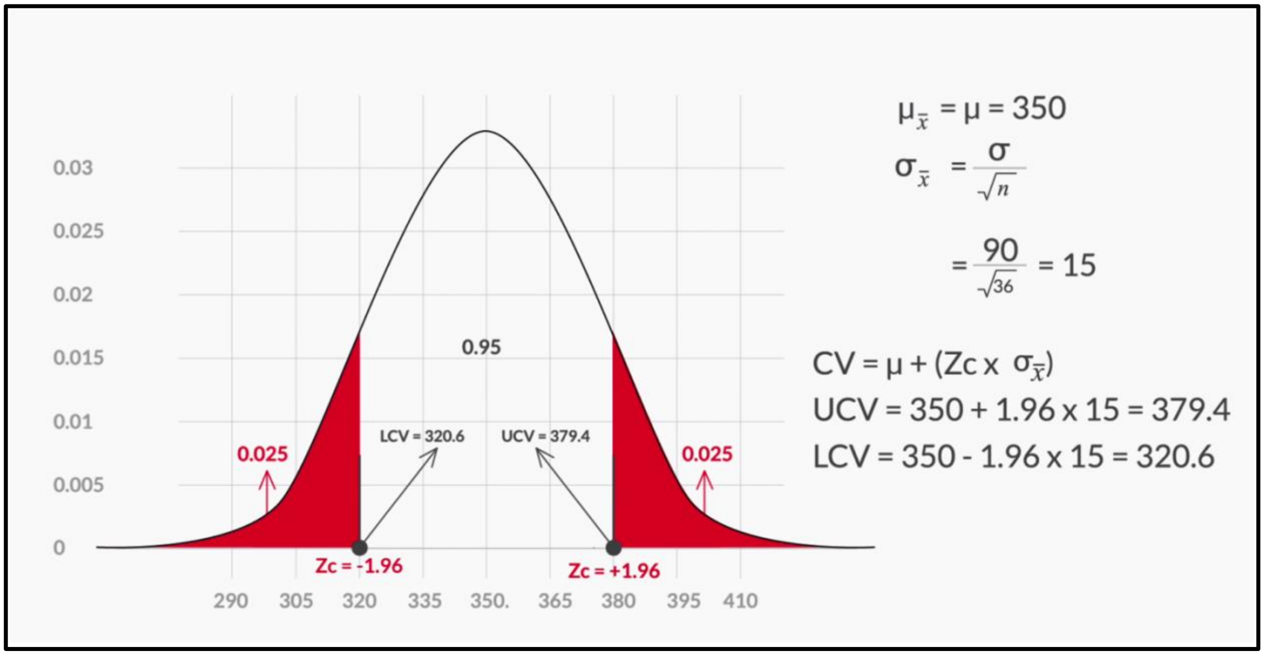
We will now find the z-score for 0.975 from Z-table, **it is 1.96**

We will now calculate the critical values:

CV= $\mu+(Zc * \sigma_{\bar{x}})$

UCV=350+(1.96*15)=379.4
LCV=350-(1.96*15)=320.6

Sample mean is 370.6 it lies in the acceptance region not in the critical region that means we fail to reject the null hypothesis.



2ND Example

H_0: μ <= 350 i.e., mean demand of AC units per month per store during summer
H_1: $\mu > 350$
sample mean $\bar{x}$ = 370.6
n=36
Pop. SD (σ)=90
sample SD ($\sigma_{\bar{x}}$)=90/√36 = 15

- This will be a one-tailed test (critical region-right side)
- α=5%
- **Z**$_c$=1-0.05=0.95 (one-tailed test)
- We will check for 0.95 in z-score table and the z-score comes out to be **1.645**

UCV=350+(1.645*15)=374.67

Since, 370.6 is less than 374.67 that means we fail to reject the null hypothesis.

P-Value Method

- It has the **probability** that null hypothesis will not be rejected
- We will 1st calculate the **z-score** of the sample mean($\bar{x}$) using the formula:
  - $(\bar{x} - \mu_{\bar{x}}) / \sigma_{\bar{x}}$
- If the z-score comes out to be **negative** and it's a **two tailed test** then **p-value** = 2*z-table value
- If the z-score comes out to be **negative** and it's a **one tailed test** then **p-value** = z-table value

AC sales Example

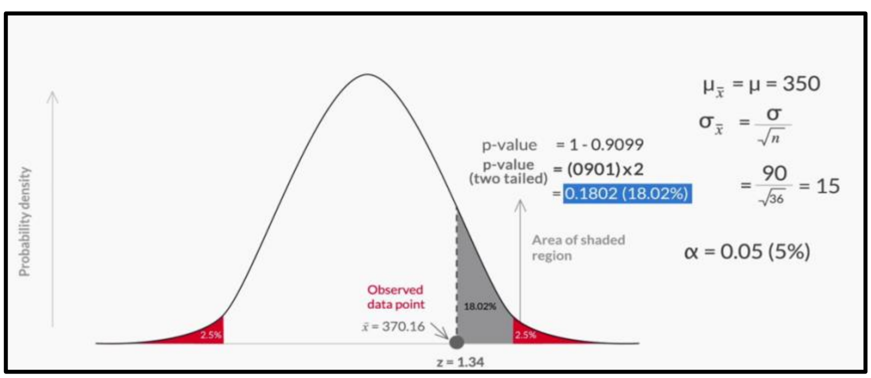
H_0: $\mu = 350$ i.e., mean demand of AC units per month per store during summer
H_1: $\mu \neq 350$
sample mean $\bar{x} = 370.16$ from n=36 stores
z-score = (370.16-350)/15 =1.34

- **So, in p-value z-score is calculated for sample mean**

If the p-value is less than the significance level, we reject the null hypothesis.

*The p-value is equivalent to the probability of the null hypothesis not being rejected. So, the smaller the p-value, the farther is the sample mean from the hypothesized population mean, which indicates more evidence in support of the sample mean lying in the critical region, and the alternative hypothesis is accepted***.**

p-value = 1-0.9099
p-value = 2*(0.0901) = 0.1802 **(**two tailed test-critical region on both sides**)**
α=5% = 0.05
➔ **p-value > α**
So, we fail to reject the null hypothesis



Errors in Hypothesis Testing

- **Type 1 error** - Rejecting null hypothesis when it is actually true – **α (**0.05 or 0.01**)**
  - If the defendant is innocent and is still convicted by jury
- **Type 2 error** - Fail to reject the null hypothesis when it is false – ß
  - If the defendant is guilty and the jury acquits him.

- **Example**
    - The amount of medication in a heart pill is same as the amount required to cure a patient.
    - We perform hypothesis testing and fail to reject the null hypothesis but later it turns out that the medication is way to, high and people are dying from over medication.

- Now in the above example making a type 2 error is a serious offence.
- Type 1 and Type 2 error are inversely related to each other
- If we want both the errors to be small then we need to take a much larger sample size.

CTR-Click Through Rate

- It is a very important metric in online marketing.
- **Example**- If there's a banner present on the platform how many people would click on it.
- **Formula**:
    - **Search CTR=** Total no. of successful searches/ Total no. of searches

Example:
A company claims that their search CTR has increased from 35% to 40%

Hypothesis
H_0: $\mu = 0.4$
H_1: $\mu \neq 0.4$
n=7400 (samples)
sample mean = total no. of clicks / total no. of searches = 2885/7400 = 0.39
population variance = 0.24 (given)= σ^2
sample sd= $\sigma\sqrt{n}$ = $\sqrt{(0.24/7400)}$= 0.006

	CTR	Searches	Clicks
Day1	38%	500	190
Day2	37%	500	185
Day3	37%	500	185
Day4	38%	600	228
Day5	41%	500	205
Day6	44%	500	220
Day7	39%	800	312
Day8	37%	500	185
Day9	42%	500	210
Day10	35%	500	175
Day11	41%	500	205
Day12	41%	500	205
Day13	43%	500	215
Day14	33%	500	165
	Total	**7400**	**2885**

So for the 14 days
Average Search CTR = (Total Number of Clicks)/(Total Number of Searches)
= 2885/7400 = 0.39

Therefore
Sample Mean = 0.39
Sample Size ('n') = 7400

$\mathbf{Z_s} = (\bar{x} - \mu_{\bar{x}}) / \sigma_{\bar{x}}$
= 0.39-0.4/ 0.006
= -1.756
α=5% = 0.05 (two tailed test-0.025% critical region on both sides)

A.) Critical Value method

Z_c = 1-0.025=0.975 (area till critical point)

We will now calculate the critical values:

CV= μ+(Zc * $\sigma_{\bar{x}}$)

UCV=0.4+(1.96*0.006)=0.41176

LCV=0.4-(1.96*0.006)= 0.38824

Since, sample mean 0.39 lies in the acceptance region we fail to reject the null hypothesis

B.) p-value method

p-value = 2*z-table value

= 2*(0.0401)

= 0.08

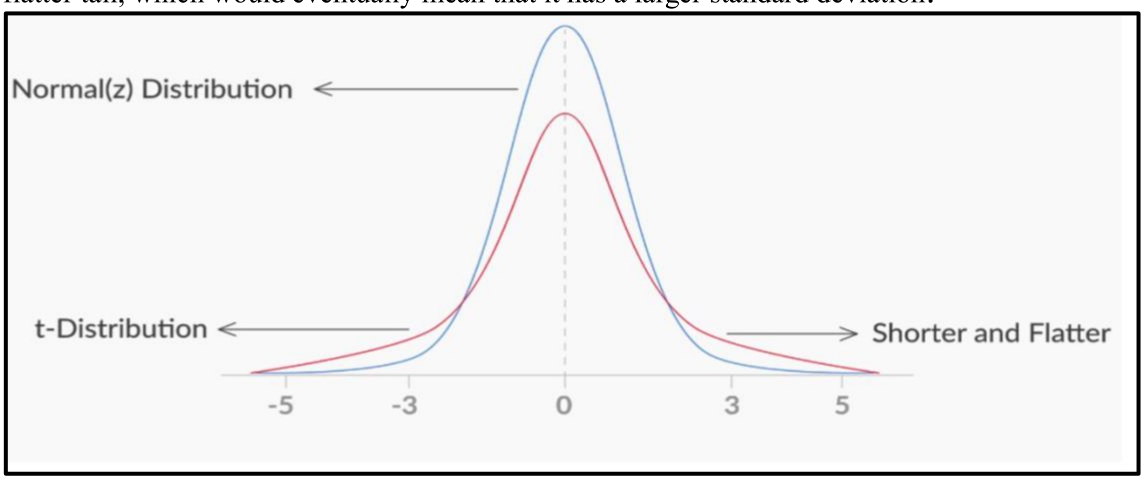
p-value > α then I fail to reject the null hypothesis

If the above case was a one-tailed test with critical region on one side then we don't need to multiply it by 2.

➔p-value=0.04 and our **α=5%=0.05** we reject the null hypothesis **(p-value< α)**

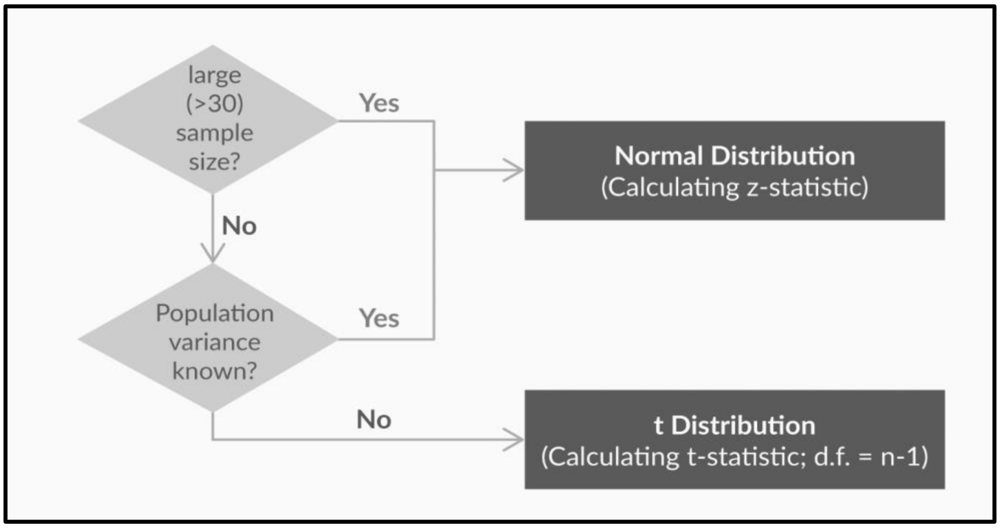
T-Distribution

A T-distribution (or Student T distribution) is similar to the normal distribution in many cases; for example, it is symmetrical about its central tendency. However, it is shorter than the normal distribution and has a flatter tail, which would eventually mean that it has a larger standard deviation.



At a sample size beyond 30, the t-distribution becomes approximately equal to the normal distribution. Each t-distribution is distinguished by what statisticians call degrees of freedom, which are related to the sample size of the data set. If your sample size is n, the degrees of freedom for the corresponding t-distribution is n -1. For example, if your sample size is 10, you use a t-distribution with 10 -1 or 9 degrees of freedom, denoted t9.

Smaller sample sizes have flatter t-distributions than larger sample sizes. And as you may expect, the larger the sample size is, and the larger the degree of freedom, the more the t-distribution looks like a standard normal distribution or the Z-distribution.



Whenever the standard deviation of the population is known, you have to use z-distribution, irrespective of the value of the sample size (N).

The t-table contains values of Zc for a given degree of freedom and value of α (significance level). Zc, in this case, can also be called as t-statistic (critical).

Two-sample proportion test

Two-sample proportion test 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.

For example, if we are comparing the effectiveness of two drugs, you would define the desired outcome of the drug as the success. So, you would take a sample of patients who consumed the new drug and record the number of successes and compare it with successes in another sample who consumed the standard drug.

AB Testing

A/B testing is a direct industry application of the two-sample proportion test sample.

While developing an e-commerce website, there could be different opinions about the choices of various elements, such as the shape of buttons, the text on the call-to-action buttons, the color of various UI elements, the copy on the website, or numerous other such things.

Often, the choice of these elements is very subjective and is difficult to predict which option would perform better. To resolve such conflicts, you can use A/B testing. A/B testing provides a way for you to test two different versions of the same element and see which one performs better.

Process:

1. **Identify the Objective**: Determine what you want to improve or optimize. This could be anything from increasing conversions on a sales page to improving engagement with an email newsletter.

2. **Create Variations**: Develop two or more versions of the asset you want to test. One version (the control or A) remains unchanged, while the other version (the variant or B) includes one or more changes that you believe may improve performance.

3. **Collect Data**: Allow the test to run for a sufficient amount of time to collect a statistically significant amount of data. This ensures that the results are reliable and not simply due to chance.

4. **Analysis**: Compare the performance of the control and variant versions using the predefined metric(s). Statistical analysis is often used to determine whether any differences observed are statistically significant or simply due to random variation.