**Hypothesis Testing**

**URL for reference : https://jovian.com/sharathdeepu9751/inferential-statistics-notes-and-graded-questions#C67**

**Module 1 : Concepts of Hypothesis Testing - I**

# Introduction:-

Welcome to the module on 'Hypothesis Testing'.

In the previous modules, you learnt exploratory data analysis and inferential statistics.

## In this session

The statistical analyses learnt in Inferential Statistics enable you to try to make inferences about population mean from the sample data when you have no idea of population mean. However, sometimes you have some starting assumption about the population mean and you want to confirm those assumptions using the sample data. It is here that **hypothesis testing** comes into the picture. We will cover the basic concepts of hypothesis testing in this session, which are as follows:

- Types of hypotheses
- Types of tests
- Decision criteria
- Critical value method of hypothesis testing

This session covers the **concepts of Hypothesis Testing from the theory perspective**, since that is very important while performing Hypothesis Testing in industry using tools like Python, Excel etc. The demonstration of Hypothesis Testing on Excel has been done in the last session of this module.

## Understanding Hypothesis Testing:-

In the last two modules, you learned about the following topics:

- Exploratory data analysis: Exploring data for insights and patterns
- Inferential statistics: Making inferences about the population using the sample data

Now, these methods help you formulate a basic idea or conclusion about the

population. Such assumptions are called “hypotheses”. But how do you really confirm these conclusions or hypotheses? Let’s see.

Let’s understand the **basic difference between inferential statistics and hypothesis testing**.

**Inferential statistics** is used to find some population parameter (mostly population mean) 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 parameter (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 the population parameter is true or not.

Both these modules have a few similar concepts, so don’t confuse terminology used in hypothesis testing with inferential statistics.

Let’s get started by understanding the basics of hypothesis testing.

Hypothesis Testing starts with the formulation of these two hypotheses:

- **Null hypothesis** ($H_0$): The status quo (i.e. the prevailing belief about a population, it assumes that the status quo is holds true.)
- **Alternate/Research hypothesis** ($H_1$): The challenge to the status quo

Now, having got a brief idea about what hypothesis testing is, in the next segment, we will look at its different aspects in detail, starting with the formulation of the null and alternate hypotheses.

## Null and Alternate Hypotheses:-

The first step of hypothesis testing is the formulation of the null and alternate hypotheses for a given situation. Let’s learn how to do this through different examples.

You have seen examples where you can write the null hypothesis (or status quo) easily from the claim statement, like in the last question - Flipkart claimed that its total valuation in December 2016 was $14 billion.

But in some instances, if your claim statement has words like “at least”, “at most”, “less than”, or “greater than”, **you cannot formulate the null hypothesis just from the claim statement** (because it’s not necessary that the **claim is always about the status quo**).

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 <

For example:

**Situation 1:** Flipkart claimed that its total valuation in December 2016 was at least $14 billion. Here, the claim contains ≥ sign (i.e. the at least sign), so **the null hypothesis is the original claim**.

The hypothesis in this case can be formulated as:


**Figure 1 - Hypotheses for Situation 1**

**Situation 2:** Flipkart claimed that its total valuation in December 2016 was greater than $14 billion. Here, the claim contains > sign (i.e. the 'more than' sign), so **the null hypothesis is the complement of the original claim**. The hypothesis in this case can be formulated as:

The hypothesis in this case can be formulated as:


**Figure 2 - Hypotheses for Situation 2**

Now, answer the following questions to consolidate your learning about the formulation of null and alternate hypotheses.

To summarize this, you cannot decide the status quo or formulate the null hypotheses from the claim statement, you need to take care of signs in writing the null hypothesis. Null Hypothesis never contains ≠ or > or < signs. It always has to be formulated using = or ≤ or ≥ signs.

Before you go ahead and look at some more examples of formulating null and alternate hypothesis, let us hear from Ankit Jain about how he used hypothesis testing during his time at Facebook.

In the next segment, we will look at how decisions are made on whether to accept or reject a hypothesis.

## Making a Decision:-

Once you have formulated the null and alternate hypotheses, let's turn our attention to the most important step of hypothesis testing — **making the decision to either reject or fail to reject the null hypothesis** — through an interesting example of a friend playing archery.

So, you learnt about what critical values are and how your decision to reject or fail to reject the null hypothesis is based on the critical values and the position of the sample mean on the distribution.

Let's learn more about the critical region and understand how the position of the critical region changes with the different types of null and alternate hypotheses.

The formulation of the null and alternate hypotheses determines the type of the test and the position of the critical regions in the normal distribution.

You can tell the type of the test and the position of the critical region on the basis of 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

In the next segment, we will look at how to find the critical values for the critical region in the distribution and make the final decision of rejecting or failing to reject the null hypothesis

## Critical Value Method:-

Now, let's learn how to find the critical values for the critical region in the distribution and make the final decision of rejecting or failing to reject the null hypothesis.

*(**Note**: In the video below, the graph showing the distribution of average sales*

*data at 1:06 incorrectly displays 370.6 as the sample mean instead of 370.16. Also, it would be*

$$\sigma_{\bar{x}} = 15 \text{ instead of } \sigma = 15 \text{at } 3{:}41)$$

Before you proceed with finding the Zc and finally the critical values, let's revise the steps performed in this method till now.

1. First, you define a new quantity called α, which is also known as the significance level for the test. It refers to the proportion of the sample mean lying in the critical region. For this test, α is taken as 0.05 (or 5%).
2. Then, you calculate the cumulative probability of UCV from the value of α, which is further used to find the z-critical value (Zc) for UCV.

Attempt the following questions before you go ahead and learn the remaining steps in this method.

After formulating the hypothesis, the steps you have to follow to **make a decision** using **the critical value method** are as follows:

1. Calculate the value of Zċ from the given value of α (significance level). Take it as 5% if not specified in the problem.
2. Calculate the critical values (UCV and LCV) from the value of Zċ.
3. Make the decision on the basis of the value of the sample mean x with respect to the critical values (UCV AND LCV).

You can download the z-table from the attachment below. It will be useful in the subsequent questions.

**Let's solve the following problem stepwise** to consolidate your learning on how to make a decision about any hypothesis.

A manufacturer claims that the average life of its product is 36 months. An auditor selects a sample of 49 units of the product, and calculates the average life to be 34.5 months. The population standard deviation is 4 months. Test the manufacturer's claim at 3% significance level using the critical value method.

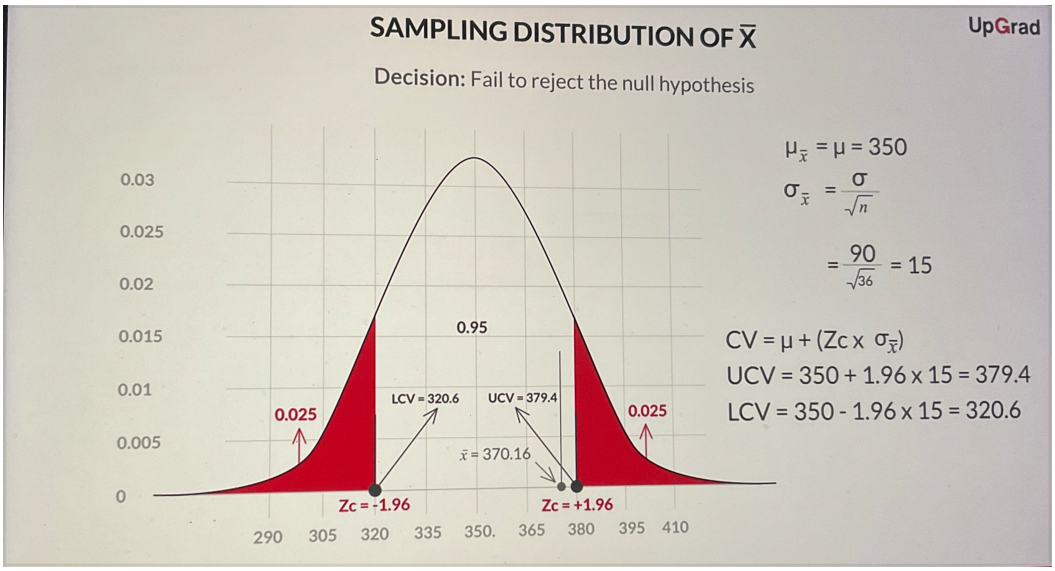
First, you need to **formulate the hypotheses** for this two-tailed test, which would be:

$$H_0: \mu = 36 \text{ months and } H_1: \mu \neq 36 \text{ months}$$

Now, you need to follow the three steps to **find the critical values and make a decision**.

Try out the three-step process by answering the following questions.

In the next segment, we will look at certain examples of critical value method




## Critical Value Method - Examples:-

You have learnt how to perform the three steps of the critical value method with the help of the AC sales problem as well as the above product lifecycle comprehension problem, which was a two-tailed test. But what would happen if it were a one-tailed test? Let's watch the video below to understand.

Attempt the following questions to complete this one-tailed test.

You can download the z-table from the attachment below. It will be useful in the subsequent questions.

Government regulatory bodies have specified that the maximum permissible amount of lead in any food product is 2.5 parts per million or 2.5 ppm. Let's say you are an analyst working at the food regulatory body of India FSSAI. Suppose you take 100 random samples of Sunshine from the market and have them tested for the amount of lead. The mean lead content turns out to be 2.6 ppm with a standard deviation of 0.6.

One thing you can notice here is that the standard deviation of the sample is given as 0.6, instead of the population's standard deviation. In such a case, you can approximate the population's standard deviation to the sample's standard deviation, which is 0.6 in this case.

Answer the following questions in order to find out if a regulatory alarm should be raised against Sunshine or not, at 3% significance level.

You can look at the solution of this comprehension from this video.

## Summary:-

So what did you learn in this session?

1. Hypothesis — a claim or an assumption that you make about one or more population parameters
2. Types of hypotheses:
    - **Null hypothesis** ($H_0$) - Makes an assumption about the status quo
      - Always contains the symbols '=', '≤' or '≥'
    - **Alternate hypothesis** ($H_1$) - Challenges and complements the null hypothesis
      - Always contains the symbols '≠', '<' or '>'
3. Types of tests:
    - **Two-tailed test** - The critical region lies on both sides of the distribution
      - The alternate hypothesis contains the ≠ sign
    - **Lower-tailed test** - The critical region lies on the left side of the distribution
      - The alternate hypothesis contains the < sign
    - **Upper-tailed test** - The critical region lies on the right side of the distribution
      - The alternate hypothesis contains the > sign
4. Making a decision - Critical value method:
    - Calculate the value of Zc from the given value of α (significance level)
    - Calculate the critical values (UCV and LCV) from the value of Zc
    - Make the decision based on the value of the sample mean $\bar{x}$ with respect to the critical values (UCV AND LCV)

**Module 2 : Concepts of Hypothesis Testing - II**

# Introduction:-

Welcome to the module on 'Concepts of Hypothesis Testing- II'.

In the previous module, you learnt types of hypotheses, types of tests, decision

criteria and critical value method for hypothesis testing.

## In this session

The statistical analyses learnt in Inferential Statistics enable you to try to make inferences about population mean from the sample data when you have no idea of population mean. However, sometimes you have some starting assumption about the population mean and you want to confirm those assumptions using the sample data. It is here that **hypothesis testing** comes into the picture. We will cover the basic concepts of hypothesis testing in this session, which are as follows:

- The p-value method
- Types of errors

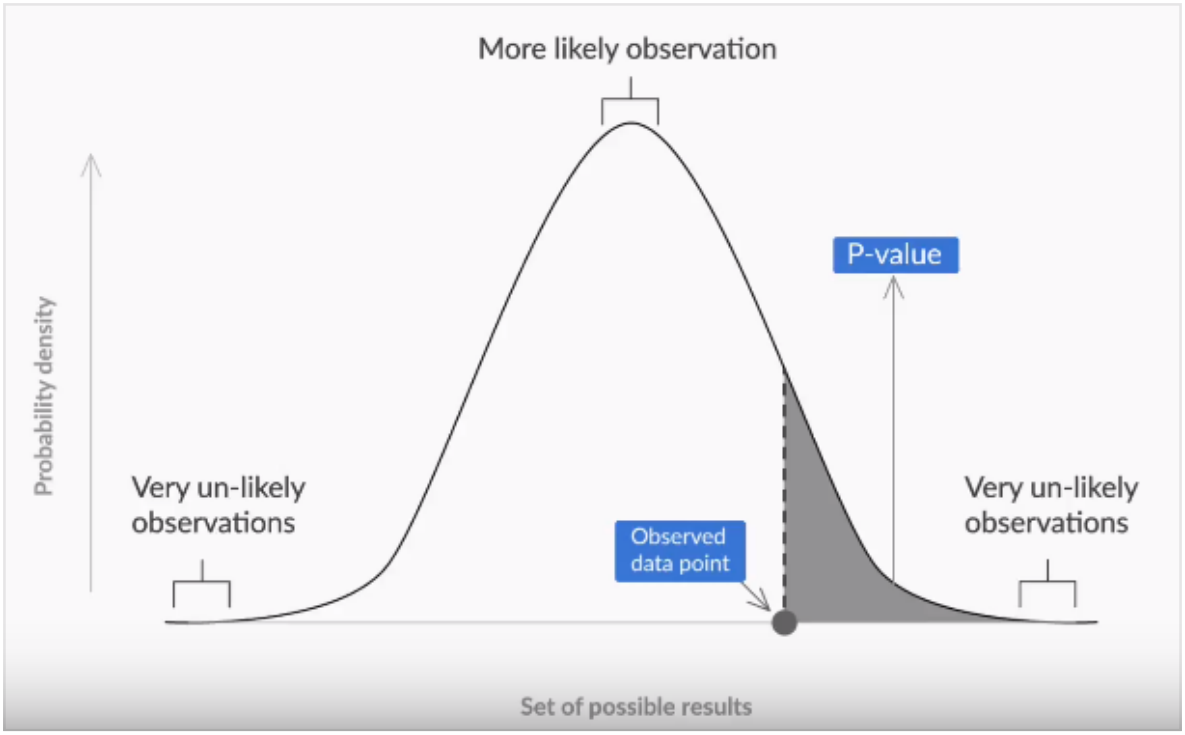
This session covers the **concepts of Hypothesis Testing from the theory perspective**, since that is very important while performing Hypothesis Testing in industry using tools like Python, Excel etc. The demonstration of Hypothesis Testing on Excel has been done in the last session of this module.

## The p-value Method:-

Let’s get started with the p-value method of making a decision.

Prof. Tricha has defined **p-value** as the **probability that the null hypothesis** will not be rejected. This statement is not the technical (or formal) definition of p-value; it is used for better understanding of the p-value.

The higher the p-value, the higher is the probability of failing to reject a null hypothesis. And the lower the p-value, the higher is the probability of the null hypothesis being rejected.




After formulating the null and alternate hypotheses, the steps to follow in order to **make a decision** using the **p-value method** are as follows:

1. Calculate the value of the z-score for the sample mean point on the distribution.
2. Calculate the p-value from the cumulative probability for the given z-score using the z-table.
3. Make a decision on the basis of the p-value (multiply it by 2 for a two-tailed test) with respect to the given value of α (significance value).

To find the correct p-value from the z-score, find the **cumulative probability** first, by simply looking at the z-table, which gives you the area under the curve till that point.

**Situation 1:** The sample mean is on the right side of the distribution mean (the z-score is positive).

**Example:** z-score for sample point = + 3.02

| z | .00 | .01 | .02 | .03 | .04 | .05 | .06 | .07 | .08 | .09 |
|---|---|---|---|---|---|---|---|---|---|---|
| 2.8 | .9974 | .9975 | .9976 | .9977 | .9977 | .9978 | .9979 | .9979 | .9980 | .9981 |
| 2.9 | .9981 | .9982 | .9982 | .9983 | .9984 | .9984 | .9985 | .9985 | .9986 | .9986 |
| 3.0 | .9987 | .9987 | .9987 | .9988 | .9988 | .9989 | .9989 | .9989 | .9990 | .9990 |
| 3.1 | .9990 | .9991 | .9991 | .9991 | .9992 | .9992 | .9992 | .9992 | .9993 | .9993 |
| 3.2 | .9993 | .9993 | .9994 | .9994 | .9994 | .9994 | .9994 | .9995 | .9995 | .9995 |
| 3.3 | .9995 | .9995 | .9995 | .9996 | .9996 | .9996 | .9996 | .9996 | .9996 | .9997 |
| 3.4 | .9997 | .9997 | .9997 | .9997 | .9997 | .9997 | .9997 | .9997 | .9997 | .9998 |

Cumulative probability of the sample point = 0.9987

For a one-tailed test: p = 1 - 0.9987 = 0.0013
For a two-tailed test: p = 2 (1 - 0.9987) = 2 * 0.0013 = 0.0026

**Situation 2:** The sample mean is on the left side of the distribution mean (the z-score is negative).

**Example:** The z-score for the sample point = -3.02

| z | .00 | .01 | .02 | .03 | .04 | .05 | .06 | .07 | .08 | .09 |
|---|---|---|---|---|---|---|---|---|---|---|
| −3.4 | .0003 | .0003 | .0003 | .0003 | .0003 | .0003 | .0003 | .0003 | .0003 | .0002 |
| −3.3 | .0005 | .0005 | .0005 | .0004 | .0004 | .0004 | .0004 | .0004 | .0004 | .0003 |
| −3.2 | .0007 | .0007 | .0006 | .0006 | .0006 | .0006 | .0006 | .0005 | .0005 | .0005 |
| −3.1 | .0010 | .0009 | .0009 | .0009 | .0008 | .0008 | .0008 | .0008 | .0007 | .0007 |
| −3.0 | .0013 | .0013 | .0013 | .0012 | .0012 | .0011 | .0011 | .0011 | .0010 | .0010 |
| −2.9 | .0019 | .0018 | .0018 | .0017 | .0016 | .0016 | .0015 | .0015 | .0014 | .0014 |
| −2.8 | .0026 | .0025 | .0024 | .0023 | .0023 | .0022 | .0021 | .0021 | .0020 | .0019 |
| −2.7 | .0035 | .0034 | .0033 | .0032 | .0031 | .0030 | .0029 | .0028 | .0027 | .0026 |

Cumulative probability of the sample point = 0.0013

For a one-tailed test: p = 0.0013
For a two-tailed test: p = 2 * 0.0013 = 0.0026

You can download the z-table from the attachment below. It will be useful in the subsequent questions.

**Let's solve the following problem stepwise** to consolidate your learning on how to make a decision about any hypothesis using the p-value method.

You are working as a data analyst at an auditing firm. A manufacturer claims that the average life of its product is 36 months. An auditor selects a sample of 49 units of the product and calculates the average life to be 34.5 months. The population standard deviation is 4 months. Test the manufacturer's claim at a 3% significance level using the p-value method.

First, **formulate the hypothesis** for this two-tailed test, which would be:

$H_0$: μ = 36 months and $H_1$: μ ≠ 36 months

Now, you need to follow the three steps to **find the p-value and make a decision**.

Try out the three-step process by answering the following questions.

You learnt how to perform the three steps of the p-value method through the AC sales problem as well as the product life cycle comprehension problem given above.

## The p-value Method: Examples:-

### Comprehension

Let's revisit an example we looked at earlier.

Let's say you work at a pharmaceutical company that manufactures an antipyretic drug in tablet form, with paracetamol as the active ingredient. An antipyretic drug reduces fever. The amount of paracetamol deemed safe by the drug regulatory authorities is 500 mg. If the value of paracetamol is too low, it will make the drug ineffective and become a quality issue for your company. On the other hand, a value that is too high would become a serious regulatory issue.

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

Your task is to take a few samples, measure the amount of paracetamol in them, and test the hypothesis that the manufacturing process is running successfully, i.e., the paracetamol content is within regulation. You have the time and resources to take about 900 sample tablets and measure the paracetamol content in each.

Upon sampling 900 tablets, you get an average content of 510 mg with a standard deviation of 110. What does the test suggest if you set the significance level at 5%? Should you be happy with the manufacturing process, or should you ask the production team to alter the process? Is it a regulatory alarm or a quality issue?

**Solve the following questions** in order to find the answers to the questions stated above.

One thing you can notice here is that the standard deviation of the sample of 900 is given as 110 instead of the population standard deviation. In such a case, you can **assume the population standard deviation to be the same as**

**the sample standard deviation, which is 110 in this case**.

Here's another exercise set to consolidate your learning.

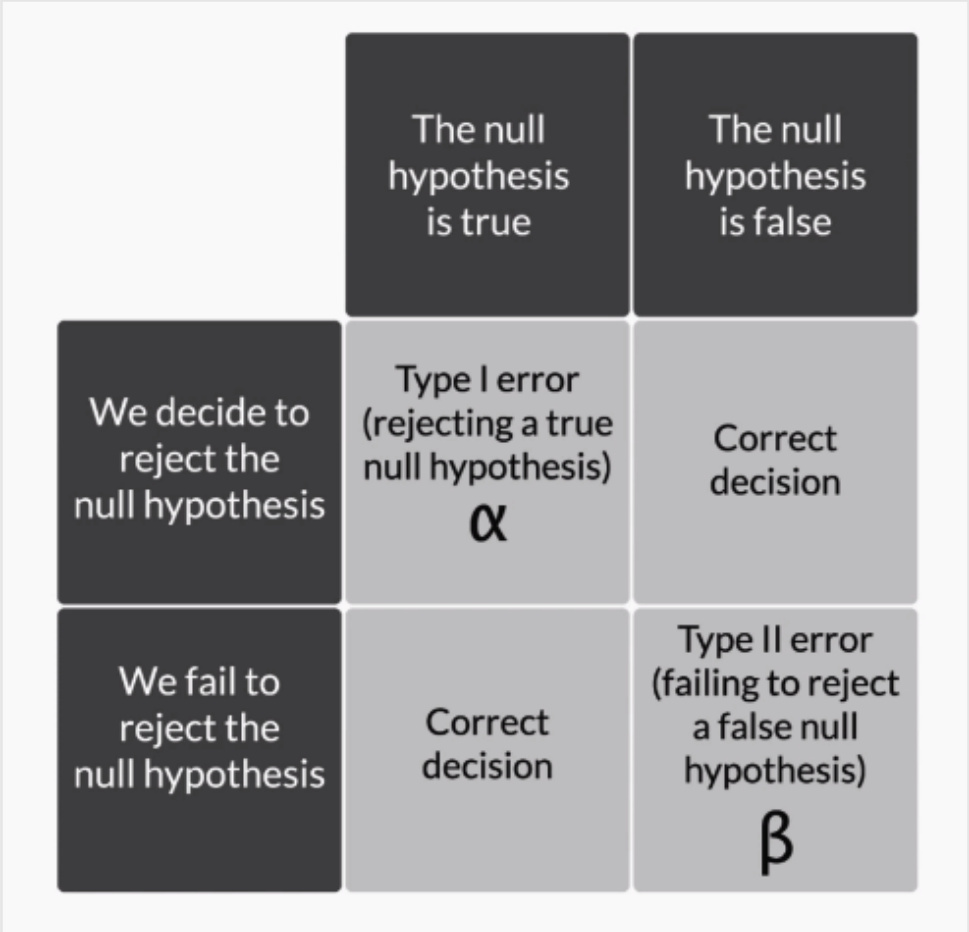
A nationwide survey claimed that the unemployment rate of a country is at least 8%. However, the government claimed that the survey was wrong and the unemployment rate is less than that. The government asked about 36 people, and the unemployment rate came out to be 7%. The population standard deviation is 3%.

Before we move on to the last topic of this session, let's hear from Kalpana on how hypothesis testing can be very useful in making business decisions in the industry.

## Types of Errors:-

While doing hypothesis testing, there is always a possibility of making the wrong decision about your hypothesis; such instances are referred to as 'errors'. Let’s learn about the different types of errors in hypothesis testing.

There are two types of errors that you might make in the hypothesis testing process: type-I error and type-II error.



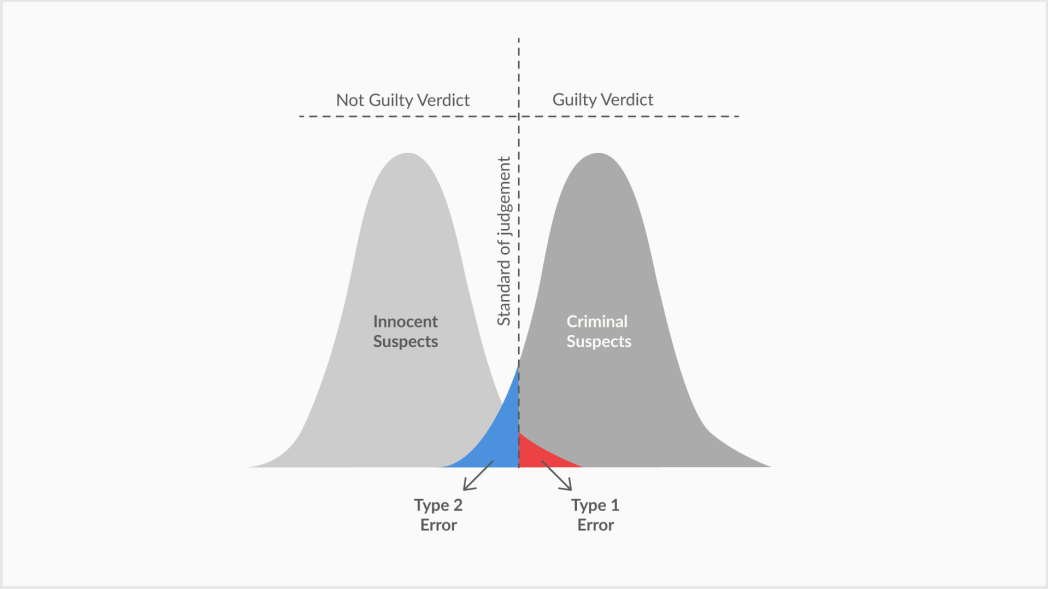

A **type I-error**, represented by α, occurs when you reject a true null hypothesis.

A **type-II error**, represented by β, occurs when you fail to reject a false null hypothesis.

The power of any hypothesis test is defined by 1 - β. The power of the test or the calculation of β is beyond the scope of this course. You can study more about the power of a test at this link.

If you go back to the analogy of the **criminal trial example**, you would find that the probability of making a type-I error is more if the jury convicts the accused even on less substantial evidence. The probability of a type-I error can be reduced if the jury follows more stringent criteria to convict an accused party.

However, reducing the probability of a type-I error may increase the probability of making a type-II error. If the jury becomes very liberal in acquitting people on trial, there is a higher probability of an actual criminal walking free.




Let us summarize our learnings of hypothesis testing in the next segment.

## Summary:-

So, what did you learn in this session?

1. Making a decision using the p-value method, which involves the following steps:
- Calculate the value of the Z-score for the sample mean point of the distribution.
- Calculate the p-value from the cumulative probability of the given Z-score using the Z-table.
- Make a decision on the basis of the p-value with respect to the given value of α (significance level).

2. Types of errors:
- **Type-I error:** This occurs when you reject a true null hypothesis. Its probability is represented by α.
- **Type-II error:** This occurs when you fail to reject a false null hypothesis. Its probability is represented by β.

Let's quickly recap the first two sessions in this video.

In the next segment, we have provided the graded questions, answer the questions based on the concepts learnt in this session.

## Module 3 : Industry Demonstration I

## Introduction:-

In the previous sessions, you learnt about the basic concepts of hypothesis testing and learnt how to statistically test the inferences made from inferential statistics or the insights generated during EDA. You learnt how to formulate the hypotheses and then make a decision through either the critical value method or the p-value method.

### In this session

You will learn how hypothesis testing is used in the industry and how the basic concepts learnt in the previous session can help you in solving industry problems. Let's hear from Rahim as he introduces this session

### Business Understanding:-

In this entire demonstration, you'll be looking at an important business metric used in the e-commerce industry, that is the **Click-through Rate** popularly known as **CTR**. On this metric, you'll be making claims and hypotheses and then test them using the concepts that you learnt in the previous two sessions. Before getting into the hypothesis testing part, let's understand what CTR means, and how specifically, you are going to use it in this demonstration.

Thus as you saw, the Clickthrough Rate is a pretty generic yet important term in the online marketing industry. In plain terms, it measures the action rate of some entity - may it be a banner ad, a home page of some website and so on. Now let's understand how the Click-Through Rate or CTR can be defined in the search context

So in the case of online search, the CTR is the proportion of searches that were successful. When you're trying to compute the Search CTR, the relevant formula is given as -

$$Search\ CTR = \frac{Total\ number\ of\ successful\ searches}{Total\ number\ of\ searches}$$

Now in the next segment, you'll be introduced to the problem statement.

### Problem Statement:-

Let's get formally introduced to the industry problem statement that we are

going to solve in this entire demonstration.

Thus, you're a data analyst at Flipkart and the product manager comes to you and claims the following - **" The Search CTR has increased from 35% to 40%"** and you'll be trying to test this claim using the concepts of hypothesis testing.

Let's dissect this problem statement a bit further which will help us in the hypothesis formulation.

To summarise the video above, the product manager is claiming that the average population Search CTR is 40%. That means out of 100 such searches on Flipkart's website, around 40 searches will result in conversions. Using this, you'll be formulating the hypotheses in the next segment.

## Hypothesis Formulation:-

In this segment, you'll learn how to formulate the null and the alternate hypothesis from the given claim.

> **Note: In the above the SME mentions $H_A$ which is nothing but $H_1$ as we have learned so far.**

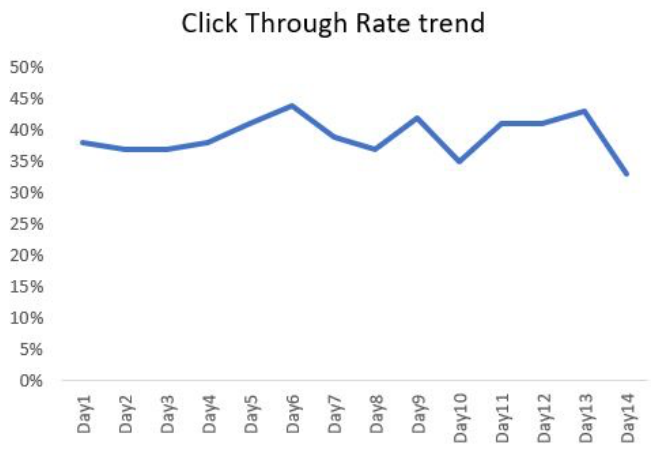
So, the given claim would yield the following null and alternate hypotheses

$$H_0 : p = 0.4$$
$$H_A : p \neq 0.4$$

On the basis of the above hypothesis statements, we'll be testing the significance of the results obtained from the sample. Now with that in place, let's go and check what data we've collected.

As explained, the dataset in the given experiment is summarised in the following table:

| | 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 |




Now you must be thinking, we can go ahead and conduct the hypothesis test directly on the given data. But think about it for a while. Which data exactly are you going to test the hypotheses on? Is it the CTR of one day, one week or multiple days? Since the data is varying a lot across all the days, how do we ensure that the sample is a **random representative** one? Let's go to the next segment and find it out.

**Additional References**

- Check this link to understand why it is necessary to take a random representative sample while doing a hypothesis test.

## Choosing the Representative Sample:-

As mentioned earlier, you need to choose a representative sample before conducting the hypothesis test. Let's take a look at the following video and understand how to do the same.

According to the given claim, the population mean for the given Search CTR = 40% or 0.4. Now the sample can show variation and may not be exactly equal to this value. It is only that on an average when you compute the mean CTR for several days, will the search CTR average out to 0.4.

Now once, you have understood this concept, you need to know which distribution you need to look at. And that will help you get at the representative sample.

Thus, as you saw, you need to focus on the "**sample mean**" for the entire set of the samples. You already know that the sampling distribution, or the distribution of the sample means, which is normal due to the Central Limit Theorem will help us out in this hypothesis testing procedure. With that in mind,

let's go ahead and compute the respective values for the sample mean.

Now, we have finally computed the sample parameters, or the sample mean which came out to be **0.39** with the sample size as **7400**. The computation for this is pretty straightforward, i.e. all you did was find the total number of searches for all the 14 days and computed the average search CTR. You can check the calculations in the following image

| | 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**

Now in the next segment, you'll be computing the necessary test statistic and then make the necessary inferences.

## Computing the Test-Statistic:-

Now that we have the required sample parameters, we need to compute the test statistic for this given sample. For this, we need the sampling distribution's standard deviation as well.
Let's say you're also given the **population variance as 0.24** (If you want to know how the population variance is calculated, check the explanation given at the end of the page).

Now let's go ahead and compute the sampling distribution's standard deviation

Note that in the above video, **you're already given the population variance of 0.24.** From this and the given sample size, you need to compute the sampling distribution standard deviation.
Recall, from the Central Limit Theorem, the formula for Sample Distribution's standard deviation comes

out to be $\frac{\sigma}{\sqrt{n}}$ where $\sigma$ is the population standard deviation and $n$ is the sample size.

Now, given that $\sigma^2 = 0.24$ and $n = 7400$. Thus the sampling distribution's standard deviation is computed as

$$\frac{\sigma}{\sqrt{n}} = \sqrt{\frac{0.24}{7400}} = 0.00569 \approx 0.006$$

Thus as you calculated, the sampling distribution's standard deviation, given from Central Limit Theorem comes out to be **0.006**

Now let's go ahead and calculate the test statistic for this sample, which you know will be its Z-score.

As, you know, the Z-score can be calculated by the following formula $\frac{X-\mu}{\sigma/\sqrt{n}}$ . Now for the given problem, we have the parameters as follows

$$Sample\ mean(X) = 0.39$$
$$Population\ Mean(\mu) = 0.40$$
$$Sampling\ Distribution\ Standard\ Deviation(\frac{\sigma}{\sqrt{n}}) = 0.005695 \approx 0.006$$

Substituting the values, we get $Z_s = \frac{0.39-0.40}{0.005695} = \frac{-0.01}{0.005695} = -1.756$. This is the test statistic for our sample data.

In the upcoming segment, you'll be taking a look at another parameter that'll help you to find the critical region using this test statistic and finally enable you to make the decision.

## Additional Notes

- To compute the population variance when you're dealing with proportions, we have the formula **p(1-p),** where p is the proportion. Since in the given population, our null hypothesis states that the mean proportion CTR or  p = 0.4, therefore our population variance can be computed as 0.4*(1-0.4) = 0.4*0.6 = 0.24. To understand how this formula is derived, you can check this link

## Finding the Critical Region:-

Now that we got the test-statistic from our sample data, the second crucial parameter that we need is the level of significance or alpha (α). Let's understand this in the next video.

So the level of significance or alpha (α) is the maximum permissible limit to which the sample can deviate and for the given test we took it as 0.05. Using this, we can find the critical region for the given test conditions, let's hear it from Rahim.

As you saw in the video, for finding the critical region, you need to use the value of alpha (α) as well as the given null hypothesis. Since it is a two-tailed test, the critical regions would lie on both the sides of the distribution. With the given level of significance at 0.05, the critical region is between -1.960 and +1.960 . So how did we arrive at this value? Since it is a two-tailed test the probability value will be as follows:

$$p - value(Area\ under\ the\ curve) = 1.0 - (\alpha/2) = 1 - 0.025 = 0.975$$

The corresponding Z value for 0.975 is +1.960. Since normal distribution curves are symmetrical along the axis both the left and right side value will be -1.960 and +1.960.

Thus now, that we have found the critical region, it becomes easier for us to make the final decision based on the given hypothesis. Let's get to that in the next segment.

## Making the Decision:-

Now, that you've found the critical region, you need to check where your test-statistic lies and based on that make the decision. Let's check that out in the next video.

So, since the given Z-statistic is -1.756 which is between -1.960 and +1.960 (critical region), we can't reject the null hypothesis. Therefore, you can say that you didn't find enough statistical evidence to disprove the product manager's claim. Let's go ahead and summarise all that you've learnt in this session so far.

Now, the above procedure follows the critical value method to do the hypothesis tests. As you know, you also have the **p-value approach** to do the same as well. Let's go ahead to the next segment and check that out as well.

## Using p-value approach:-

In this segment, you will be doing the same hypothesis, using an alternative procedure, i.e. the **p-value** approach. This is the most common way of finding the significance of test results in the industry. Technically, the p-value is the probability of observing values which are similar to or more extreme than the observed value given that the null hypothesis is true.  This metric also enables us to determine whether the given test statistic gives us sufficient evidence to reject or fail to reject the null hypothesis. Let's go ahead and see it in action in the next video.

So, when you computed the p-value, you sort of followed an opposite route to the original critical value route. First, you computed the probability corresponding to the given Z-score = -1.756, which came out to be 0.0395. Since this is a two-tailed test, you multiplied this value with 2 and obtained the p-value as 0.079. Now, since this value is greater than the given level of significance, you fail to reject the null hypothesis.

To summarise, the p-value approach also gives us the same result as the critical value approach, which is that either both will reject the null hypothesis or fail to reject the null hypothesis.

In the upcoming segment, we'll be making a slight twist to the null hypothesis and then observe how the given hypothesis statement changes.

## Changing the Hypothesis:-

In an earlier segment, you devised the null and alternate hypothesis as follows

$$H_0 : p = 0.4$$
$$H_a : p \neq 0.4$$

Now, let's change the hypothesis statement slightly. Let's say the claim was made that the **Search CTR is actually equal to or more than 40%**. How would that change the hypothesis? And does the decision change in this case? Let's take a look at the next video to find out.

[**Note**: The Prof says "greater than 40%" instead of "equal to or greater than" 40% at **0:09.** The alternate hypothesis is considered less than 40% therefore the null hypothesis should be greater than equal to 40%.]

Thus as you saw, the new null and alternate hypotheses came out to be

$$H_o : p \geq 0.4$$
$$H_a : p < 0.4$$

This changed the entire hypothesis test from a two-tailed test to a one-tailed test. Therefore, when you computed the p-value you only took into consideration one tail and that value came out to be 0.0395. In the critical value terms as well, the necessary critical region only consists of 0.0395 area. Since it is less than the value of 0.05, you need to reject the null hypothesis in this case.

With that, you have come to the end of the hypothesis testing demonstration. Now go ahead and attempt the graded questions in the next segment. Best of Luck!

**Module 4 : Industry Demonstration II**

# Introduction:-

In the previous sessions, you learnt about the basic concepts of hypothesis testing and learnt how to statistically test the inferences made from inferential statistics or the insights generated during EDA. You learnt how to formulate the hypotheses and then make a decision through either the critical value method or the p-value method.

## In this session

You will learn how hypothesis testing is used in the industry and how the basic concepts learnt in the previous session form an important foundation of useful industry concepts such as A/B testing.

We will cover the following topics in this session:

- T distribution
- Two-sample mean test
- Two-sample proportion test
- A/B testing
- Industry relevance

**Module 5 : Statistics Interview Practice**