Statistics for Data Science

Topics covered so far

1. Statistical Inference
    a. Distributions - Binomial, Uniform, Normal
    b. Sampling
    c. Central limit theorem
    d. Confidence Intervals
2. Hypothesis Testing
    a. Hypothesis Formulation
    b. One Tailed Test vs Two Tailed Test
    c. Type I and Type II Errors

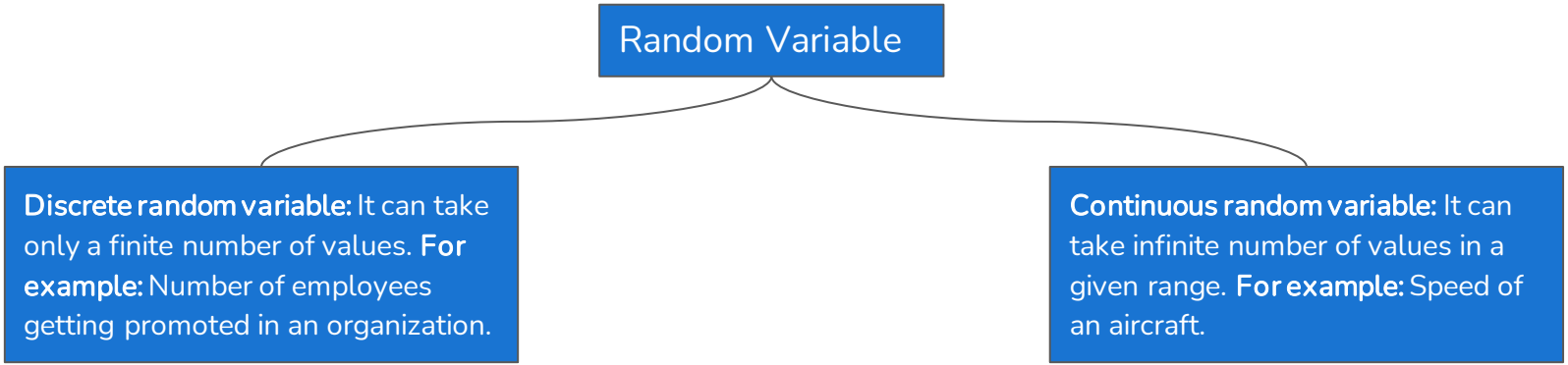
Gauge your Understanding

1. What is a random variable and how is it related to probability distribution?
2. What are some of the most commonly used distributions?
3. What is Central Limit Theorem (CLT) and when is it used?
4. What do you mean by estimations?

What is a Random Variable?

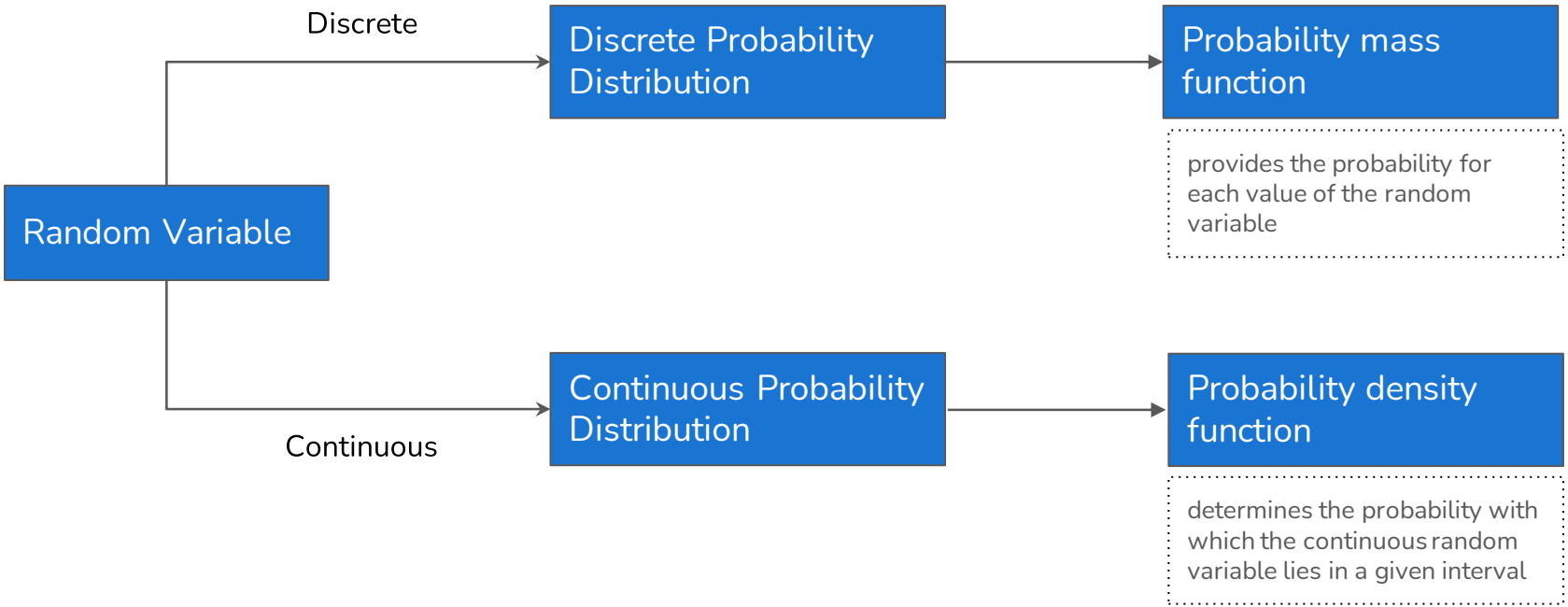
A random variable is a function that assigns a numerical value to each outcome of an experiment. It assumes different values with different probability. It is usually denoted by capital letter X and the probability associated with any particular value of X is denoted by P(X=x).

Example: Suppose that a fair coin is tossed twice and the possible outcome are {HH, HT, TH, TT}. Let X be the random variable representing the number of heads that can come up. So, X can take values from the set {2, 1, 0}.
The probability of two heads coming up is P(X=2) = ¼.



What is a Probability Distribution?

The probability distribution of a random variable describes the values that the random variable can take along with the probabilities of those values.



Distributions around us (commonly occuring)

Bernoulli	The outcome of tossing a fair coin
Binomial	The number of non-defective products in a production run
Uniform	The number of books sold weekly at a bookstore
Normal	IQ distribution of all the seven years old children in New York

Binomial Distribution

The binomial distribution is the probability distribution of the number of successes of an experiment that is conducted multiple times and has only two possible outcomes.

Example: Suppose you have purchased 10 lottery tickets and the possible outcomes are winning the lottery or not winning the lottery, then you can answer a question like what is the probability of winning 6 lottery tickets using binomial distribution.

The assumptions of Binomial distribution are as follows:

1. There are only two possible outcomes (success or failure) for each trial.
2. The number of trials is fixed.
3. The outcome of each trial is independent. In other words, none of the trials have an effect on the probability of the next trial.
4. The probability of success is exactly the same for each trial.

Note: In binomial distribution, if the number of trials for a given experiment is equal to 1, then it is called **Bernoulli distribution.**

Uniform Distribution

The **Uniform Distribution** is the probability distribution where all outcome are equal likely.

Discrete Uniform Distribution: Can take a finite number (*m*) of values and each value has equal probability of selection.

Example: Rolling a single die.

Continuous Uniform Distribution: Can take any value within a given range with equal probability.

Example: Weight gained by a person over next 2 months can be uniformly distributed between 2 to 5 Kg.

Normal Distribution

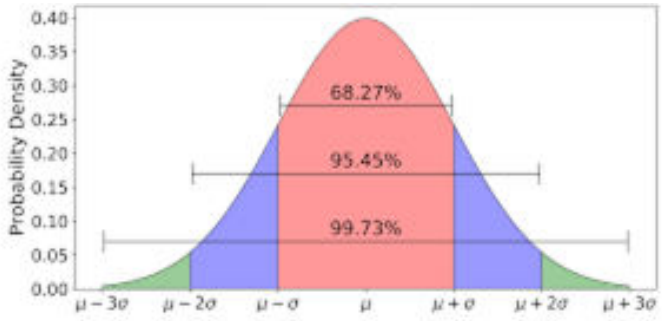
The normal distribution is a continuous probability distribution that is symmetric about the mean. It is also known as bell curve because the graph of its probability density function looks like a bell.

Example: The height of all adult males in a city

Properties:

- It has a zero skewness
- Mean = Median = Mode
- If mean = 0 and standard deviation = 1, then it is called a **standard normal distribution**

Empirical Rule



Sampling Distributions

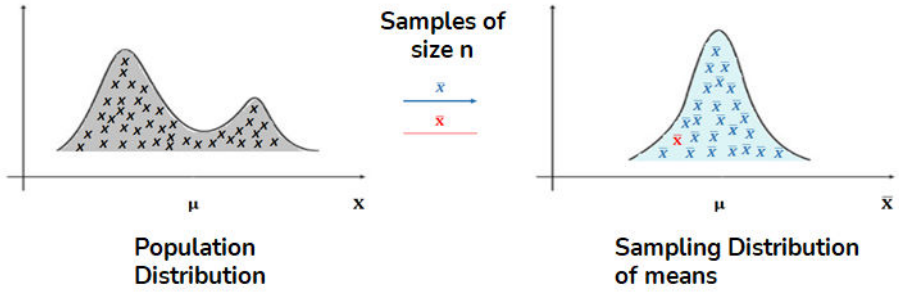
What is the need for sampling?

Given the limited resources and time, it is not always possible to study the population. That's why we choose a sample out of the population to make inference about the population.

Example: Suppose a new drug is manufactured and it needs is to be tested for the adverse side effects on a country's population. It is almost impossible to conduct a research study that involves everyone.

What are Sampling Distributions?

It is a distribution of a particular sample statistic obtained from all possible samples drawn from a specific population.



Central Limit Theorem

The sampling distribution of the sample means will approach normal distribution as the sample size gets bigger, no matter what the shape of the population distribution is.

Assumptions

- Data must be **randomly sampled**
- Sample values must be **independent** of each other
- Samples should come from the **same distribution**
- Sample size must be **sufficiently large (≥30)**

Let's see CLT in action by simulation - Link to external site

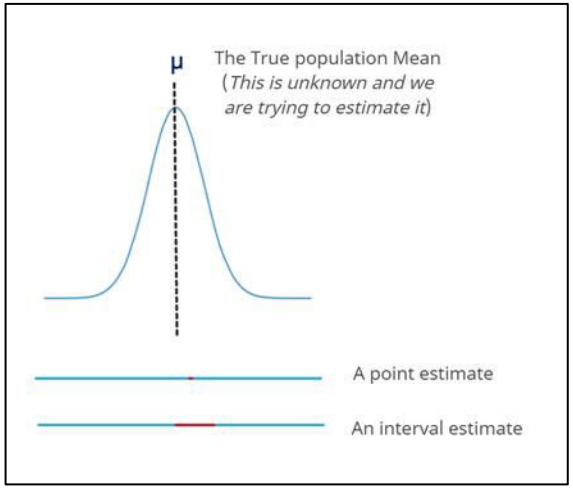
Estimations

Estimation

Make inferences about population parameter based on sample statistic

Point Estimation

Single value of a population parameter
Ex. Population mean as estimated from the sample mean is $20



Interval Estimation

The range of values within which the population parameter lies with some confidence
Ex. Population mean should lie in the range $15-$25, with 95% confidence

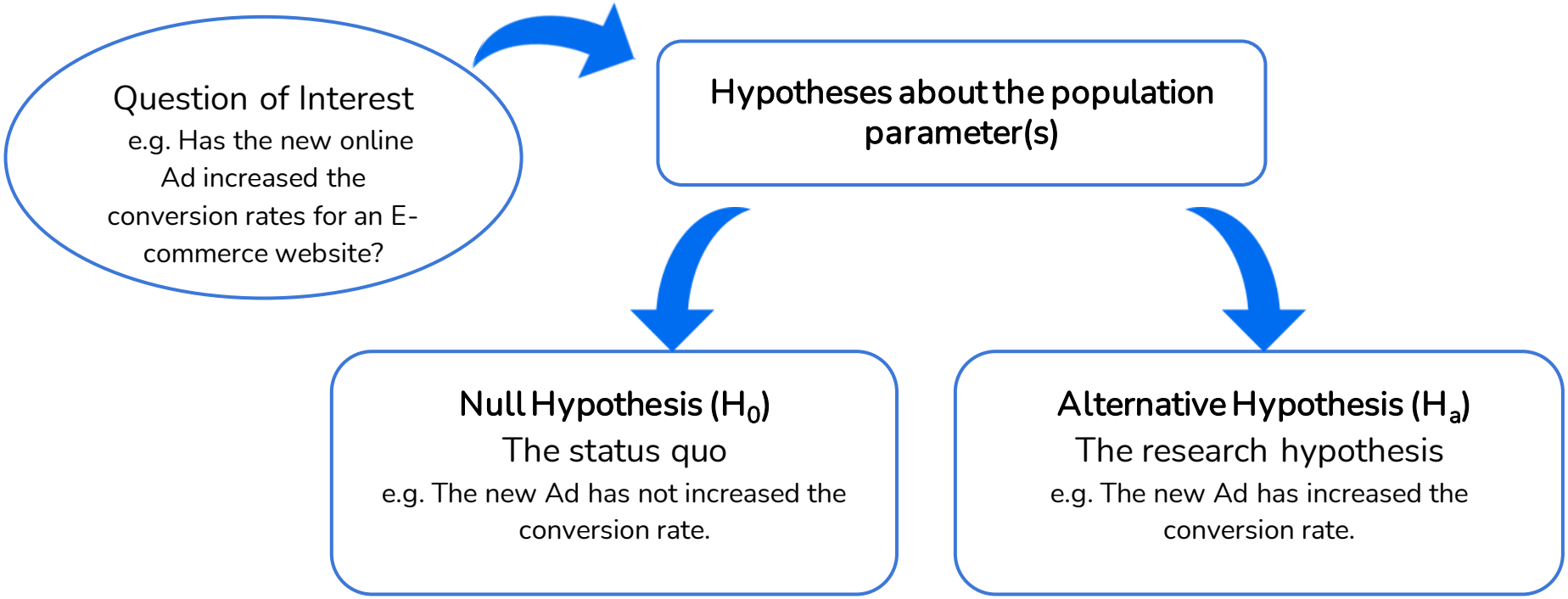
Case Study

Inferential Statistics

Gauge Your Understanding

1. What is hypothesis testing and what are different types of hypotheses?
2. What are some of the key terms involved in hypothesis testing?
3. What is the difference between one-tailed and two-tailed tests?
4. What are the steps to perform a hypothesis test?

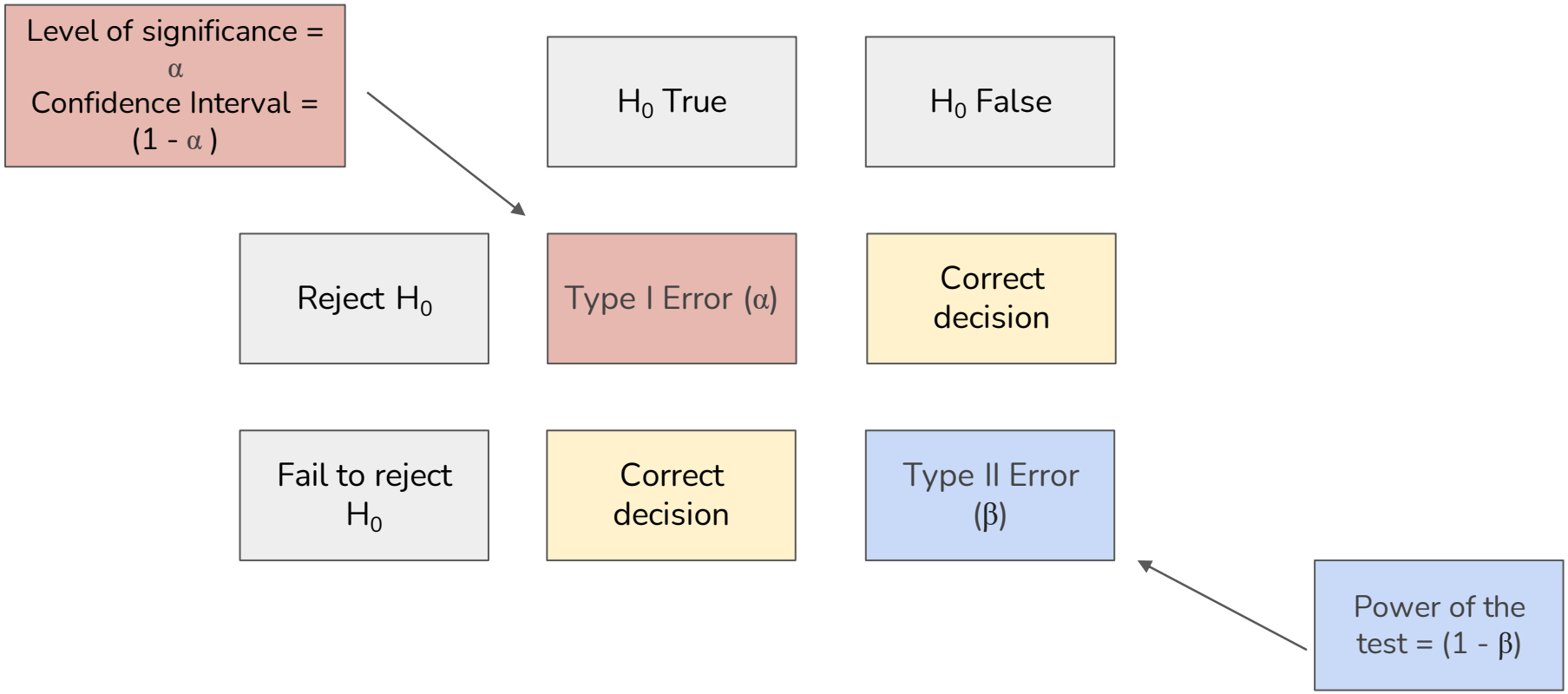
Introduction to Hypothesis Testing



Key terms in Hypothesis Testing

P-Value	• Probability of observing equal or more extreme results than the computed test statistic, under the null hypothesis. • The smaller the p-value, the stronger the evidence against the null hypothesis.
Level of Significance	• The significance level (denoted by α), is the probability of rejecting the null hypothesis when it is true. • It is a measure of the strength of the evidence that must be present in the sample data to reject the null hypothesis.
Acceptance or Rejection Region	• The total area under the distribution curve of the test statistic is partitioned into acceptance and rejection region • Reject the null hypothesis when the test statistic lies in the rejection region, else we fail to reject it
Types of Error	• There are two types of errors - Type I and Type II

Type I and Type II errors



Let's go through an example

Problem Statement: The store manager believes that the average waiting time for the customers at checkouts has become worse than 15 minutes. Formulate the hypothesis.

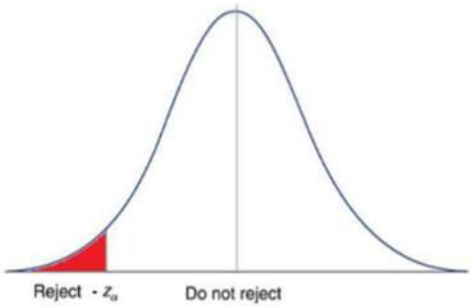
Null Hypothesis (H_0): The average waiting time at checkouts is less than equal to 15 minutes.

Alternate Hypothesis (H_a): The average waiting time at checkouts is more than 15 minutes.

Type I error (false positive): Reject Null hypothesis when it is indeed true. "The fact is that the average waiting time at checkout is less than equal to 15 minutes but the store manager has identified that it is more than 15 minutes".

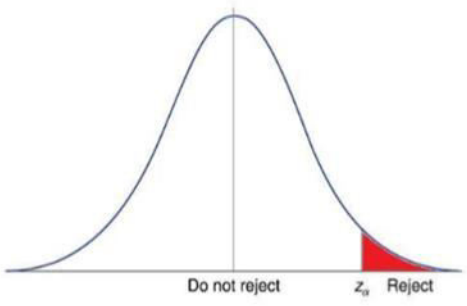
Type II error (false negative): Fail to reject Null hypothesis when it is indeed false. "The fact is that the average waiting time at checkout is more than 15 minutes but the store manager has identified that it is less than equal to 15 minutes".

One-tailed vs Two-tailed Test



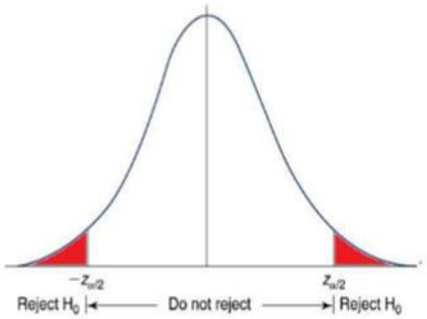
- Lower tail test.
- H_1: $\mu <$ .....

Reject H_0 if the value of test statistic is too small



- Upper tail test.
- H_1: $\mu >$ .....

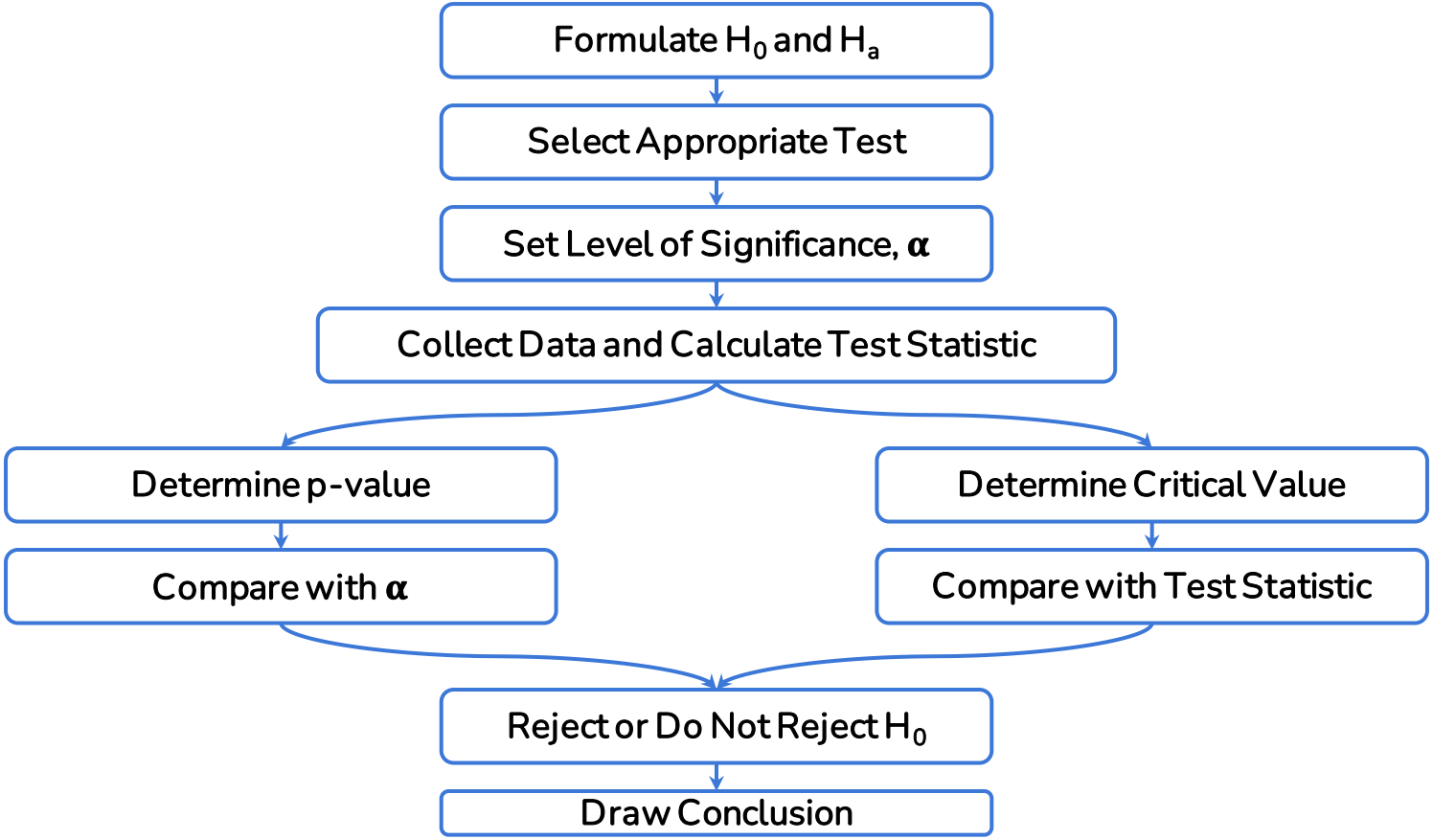
Reject H_0 if the value of test statistic is too large



- Two tail test.
- H_1: $\mu \neq$ .....

Reject H_0 if the value of test statistic is either too small or too large

Hypothesis Testing Steps



Case Study

Hypothesis Testing

Great Learning

Happy Learning !