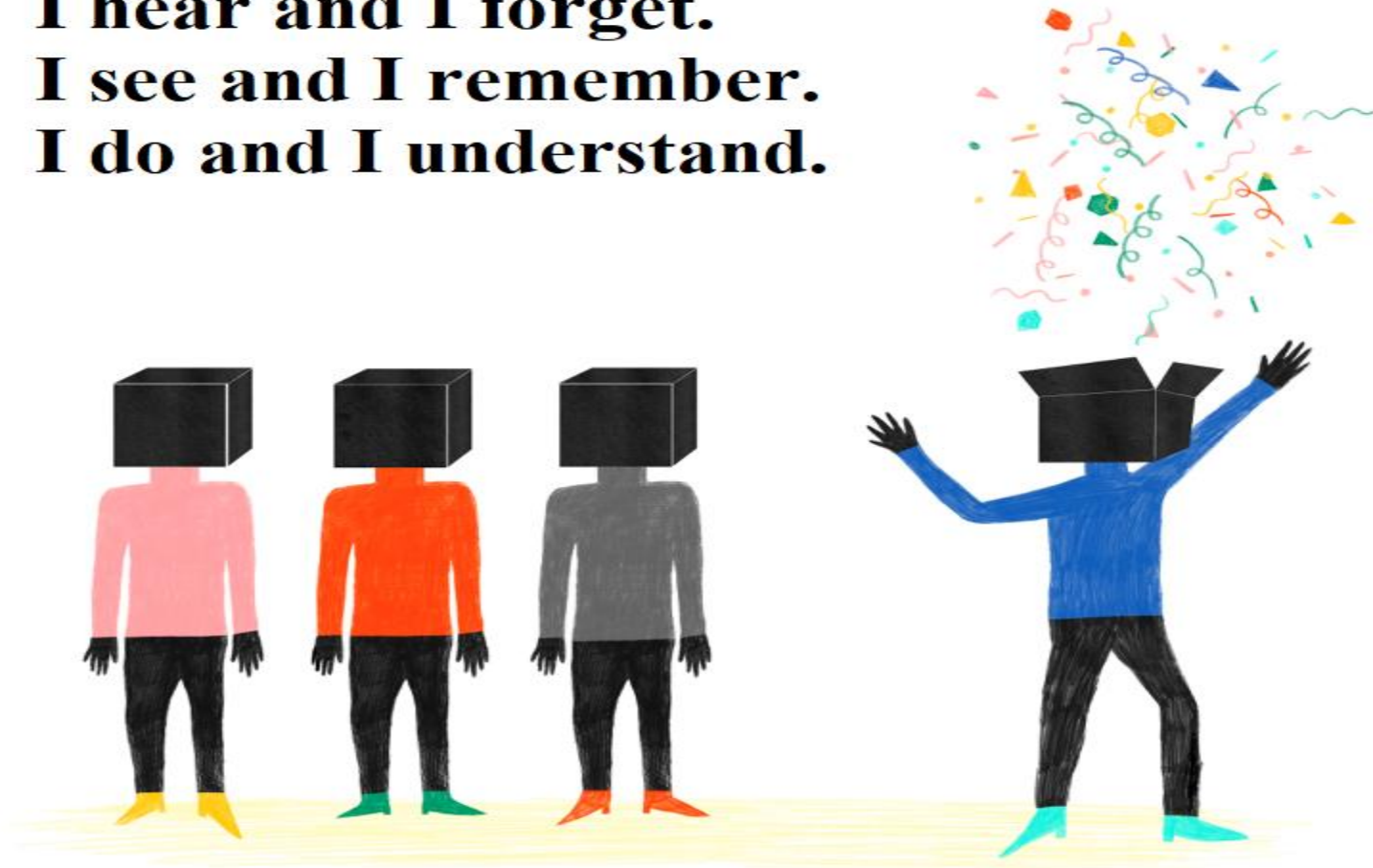


**I hear and I forget.
I see and I remember.
I do and I understand.**



Chandan Verma

Corporate Trainer(Machine Learning,AI,Cloud Computing,IOT)

www.facebook.com/verma.chandan.070

Hypothesis



Hypothesis Testing

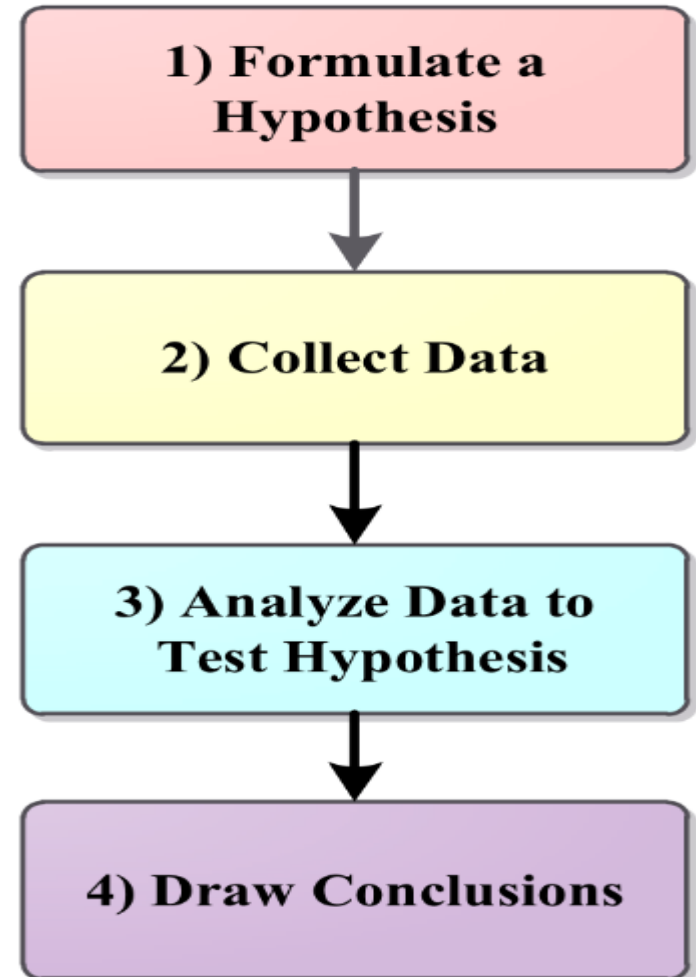
Hypothesis Testing: A systematic way to select samples from a group or population with the intent of making a determination about the expected behavior of the entire group.

A **statistical hypothesis** is an assumption about a population parameter. This assumption may or may not be true. **Hypothesis testing** refers to the formal procedures used by statisticians to accept or reject statistical hypotheses.

A hypothesis is similar to a *theory*

If you believe something might be true but don't yet have definitive proof, it is considered a theory until that proof is provided.

Turning theories into accepted statements of fact is the basis of the scientific method



Hypothesis testing procedures

In simple terms, a hypothesis refers to a supposition which is to be accepted or rejected. There are two hypothesis testing procedures, i.e. parametric test and non-parametric test

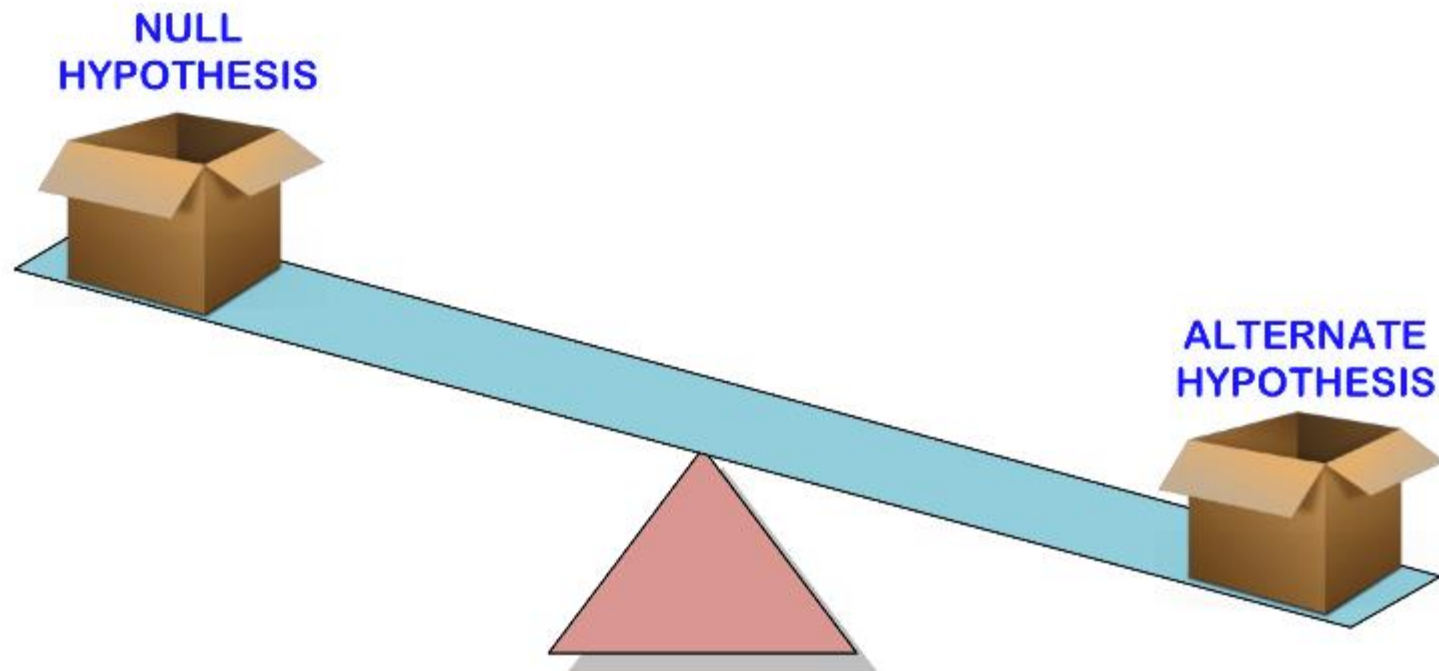
wherein the parametric test is based on the fact that the variables are measured on an **interval scale**, whereas in the non-parametric test, the same is assumed to be measured on an **ordinal scale**.

State the Hypothesis — Null & Alternative

Null hypothesis: Null hypothesis is a statistical hypothesis that assumes that the observation is due to a chance factor. Null hypothesis is denoted by; $H_0: \mu_1 = \mu_2$, which shows that there is no difference between the two population means.

Alternative hypothesis: Contrary to the null hypothesis, the alternative hypothesis shows that observations are the result of a real effect.

Hypothesis-Testing



Criminal Trial Analogy

Our criminal justice system assumes "**the defendant is innocent until proven guilty.**" That is, our initial assumption is that the defendant is innocent.

- H_0 : Defendant is not guilty (innocent)
- H_A : Defendant is guilty

In statistics, we always **assume the null hypothesis is true**. That is, the null hypothesis is always our initial assumption.

The prosecution team then collects evidence — such as finger prints, blood spots, hair samples, carpet fibers, shoe prints, ransom notes, and handwriting samples — with the hopes of finding "sufficient evidence" to make the assumption of innocence refutable.

Errors

		Truth	
		Not Guilty	Guilty
Jury Decision	Not Guilty	OK	ERROR
	Guilty	ERROR	OK

		Truth	
		Null Hypothesis	Alternative Hypothesis
Decision	Do not Reject Null	OK	Type II Error
	Reject Null	Type I Error	OK

Errors

- **Type 1 error:** Null Hypothesis was correct but the analysis proved it wrong.
- **Type 2 error:** Null Hypothesis was wrong but the analysis couldn't prove that it was wrong

Type I Error : The null hypothesis is rejected when it is true.

Type II Error : The null hypothesis is not rejected when it is false.

There is always a chance of making one of these errors. But, a good scientific study will minimize the chance of doing so!

Idea In Nutshell

- Make an assumption about a concept or data
- Collect information to test the assumption
- Verify if the assumption is right
- State your hypothesis
- Based on the available evidence (data), deciding whether to reject or not reject the initial assumption.

Decision

The jury then makes a decision based on the available evidence:

- If the jury finds sufficient evidence — beyond a reasonable doubt — to make the assumption of innocence refutable, the jury **rejects the null hypothesis** and deems the defendant guilty. We behave as if the defendant is guilty.
- If there is insufficient evidence, then the jury **does not reject the null hypothesis**. We behave as if the defendant is innocent.

In statistics, we always make one of two decisions. We either "reject the null hypothesis" or we "fail to reject the null hypothesis."

P value

p-values are often reported whenever we perform a statistical significance test (like t-test, chi-square test etc)

The p-value reported is used to make a decision on whether the null hypothesis being tested can be rejected or not.

p value can be interpreted as the probability that the null hypothesis is correct. So higher the p value, more is the probability of the H_0 to be accepted.

P value

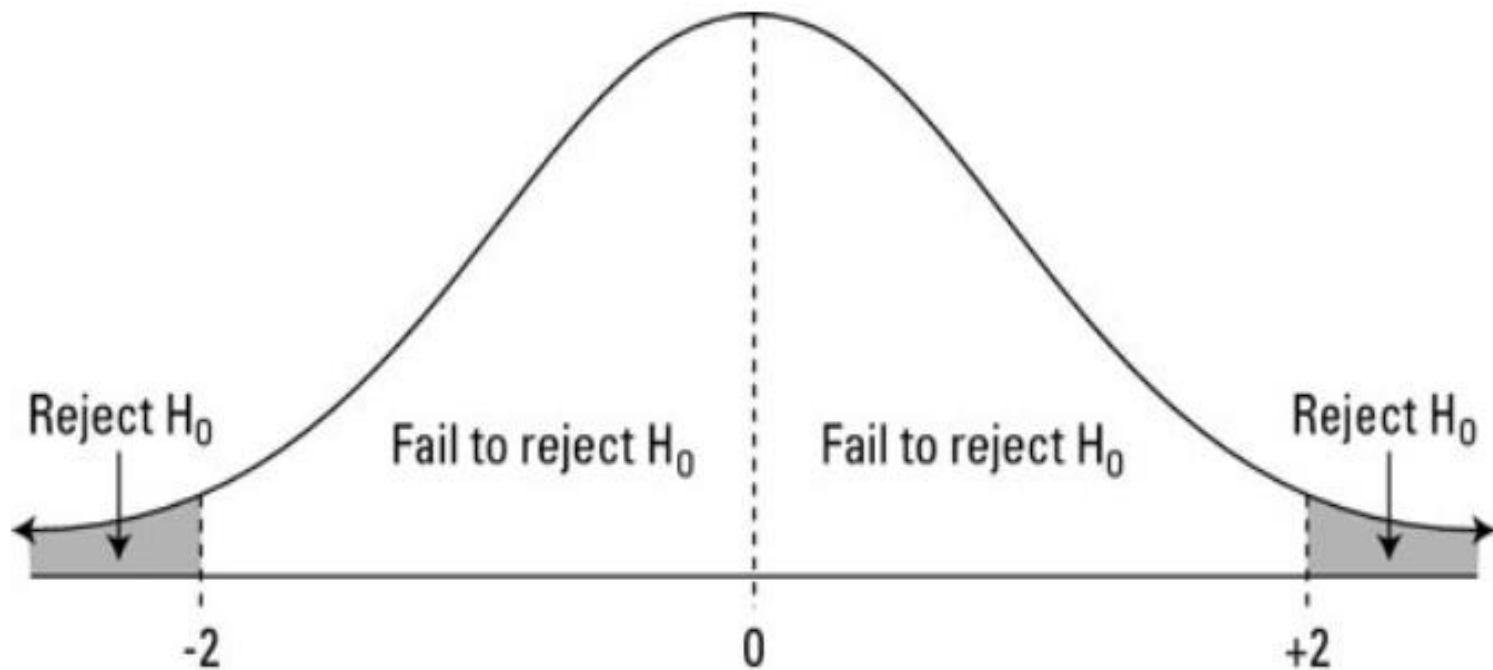
The P Value basically helps to answer the question: **‘Does the data really represent the observed effect?’**.

The P Value is **the probability of seeing the effect(E) when the null hypothesis is true.**

$$P \text{ Value} = P(E \mid H_0)$$

when the p-value is low enough, we reject the null hypothesis and conclude the observed effect holds

P value



a: not-equal-to."/>

Decisions for H_a : not-equal-to.

Hypothesis Testing (P-Value Approach)

Specify the null and alternative hypotheses.

Using the sample data and assuming the null hypothesis is true, calculate the value of the test statistic. Again, to conduct the hypothesis test for the population mean μ , we use the t -statistic t^* which follows a t -distribution with $n - 1$ degrees of freedom.

$$t^* = \frac{\bar{x} - \mu}{s / \sqrt{n}}$$

Using the known distribution of the test statistic, calculate the **P-value**: "If the null hypothesis is true, what is the probability that we'd observe a more extreme test statistic in the direction of the alternative hypothesis than we did?"

Level of significance

Refers to the degree of significance in which we accept or reject the null-hypothesis. 100% accuracy is not possible for accepting or rejecting a hypothesis, so we therefore select a level of significance that is usually 5%.

This is normally denoted with alpha(maths symbol α) and generally it is 0.05 or 5% , which means your output should be 95% confident to give similar kind of result in each sample.

Level of significance

p value > level of significance, Null hypothesis is Accepted

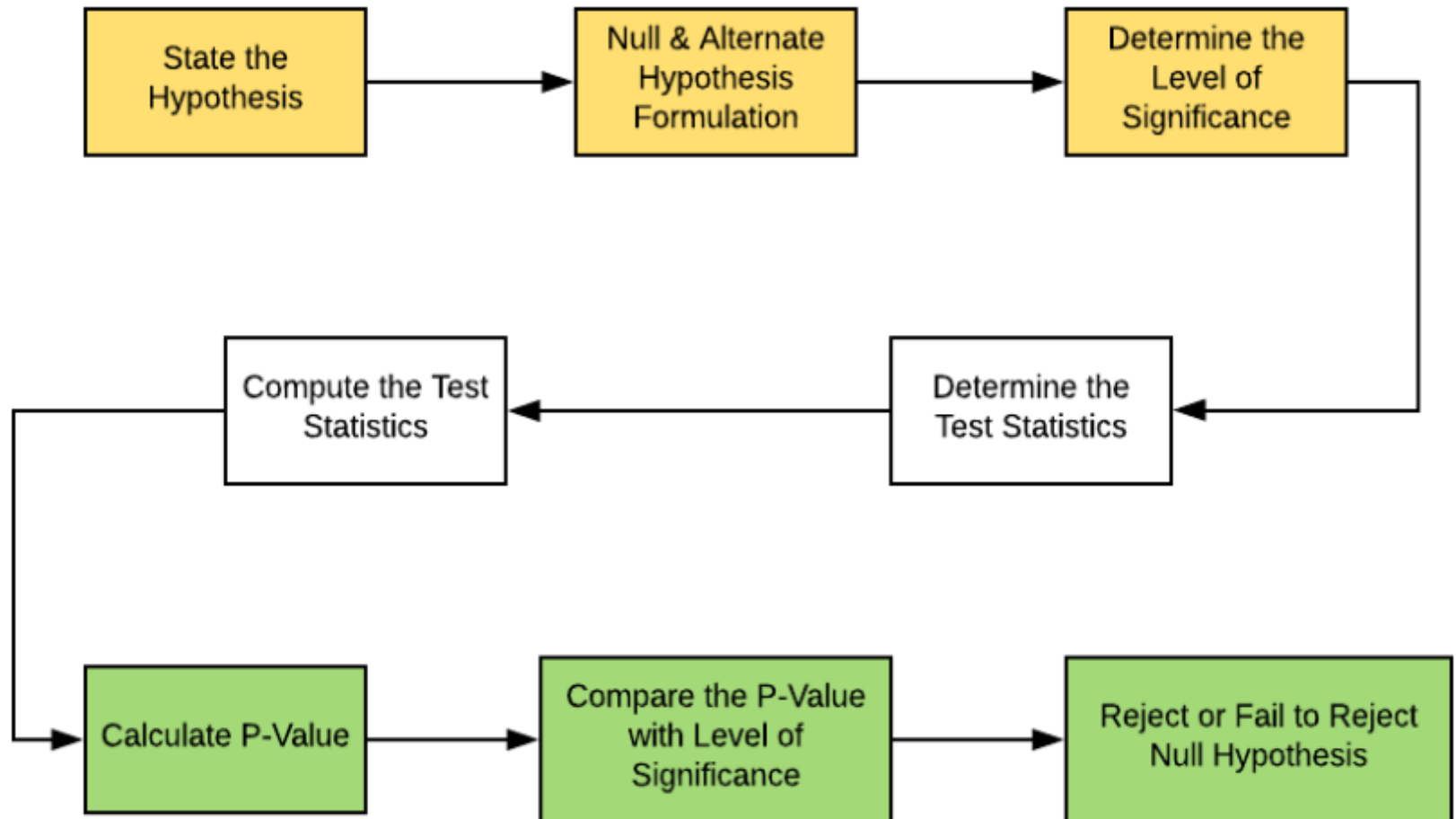
p value < level of significance, Null hypothesis is Rejected

Test

One tailed test :- A test of a statistical hypothesis , where the region of rejection is on only one side of the sampling distribution , is called a one-tailed test.

Two-tailed test :- A two-tailed test is a statistical test in which the critical area of a distribution is two-sided and tests whether a sample is greater than or less than a certain range of values. If the sample being tested falls into either of the critical areas, the alternative hypothesis is accepted instead of the null hypothesis

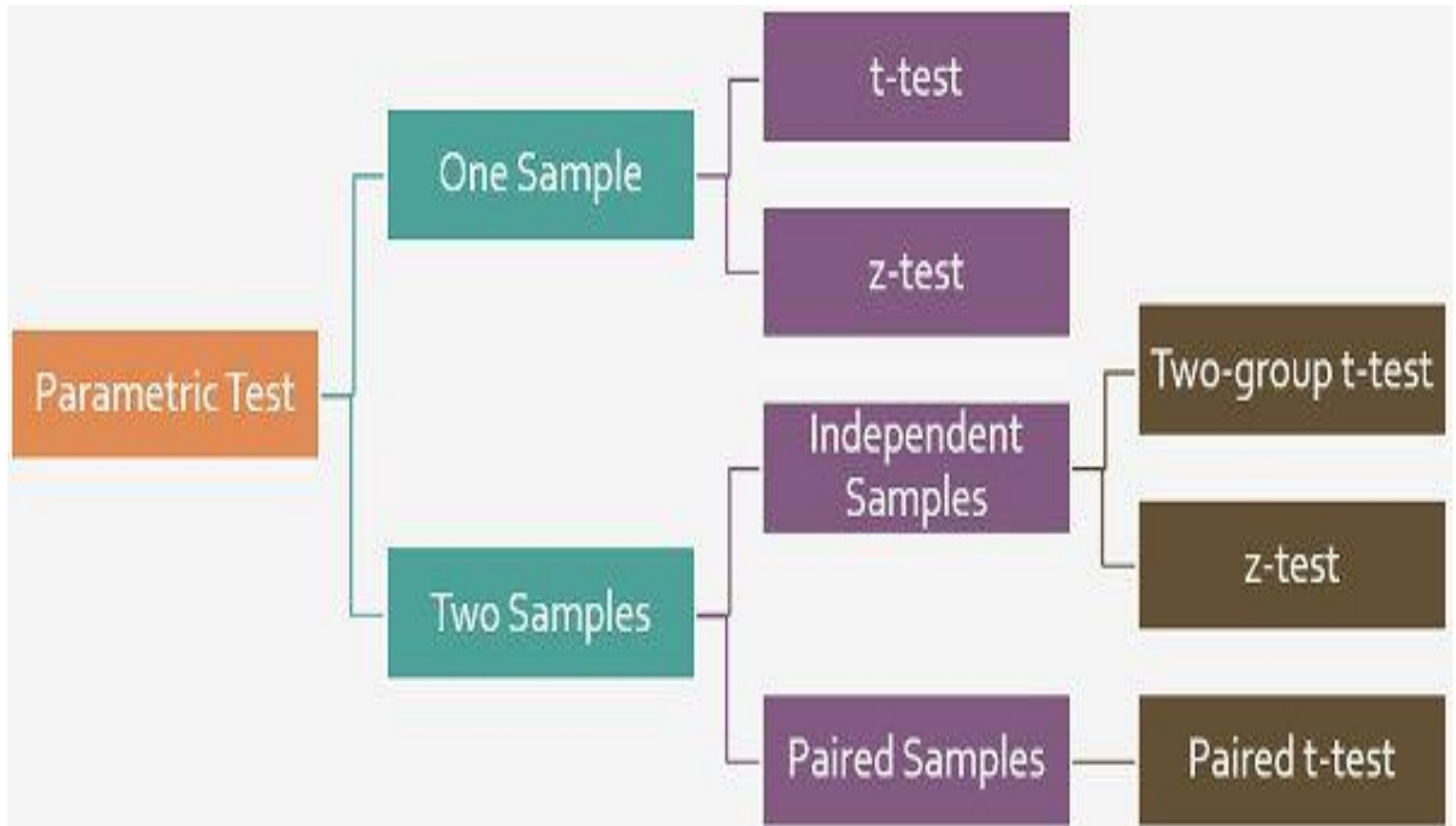
Workflow of Hypothesis Testing



Test

- 1. T-test**
- 2. Z-test**
- 3. F-test**
- 4. Chi-square**
- 5. ANOVA**

T-test and Z-test



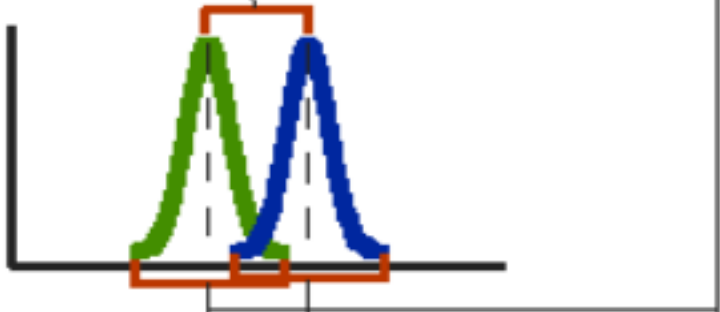
T-test

T-test refers to a type of parametric test that is applied to identify, how the means of two sets of data differ from one another when variance is not given.

T-test follows t-distribution, which is appropriate when the sample size is small, and the population standard deviation is not known. The shape of a t-distribution is highly affected by the degree of freedom. The degree of freedom implies the number of independent observations in a given set of observations.

$$t = \frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{\left\{ \frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} \right\}}}$$

t-score

$$\begin{aligned} \frac{\text{signal}}{\text{noise}} &= \frac{\text{difference between group means}}{\text{variability of groups}} \\ &= \frac{\bar{X}_T - \bar{X}_C}{\text{SE}(\bar{X}_T - \bar{X}_C)} \\ &= \text{t-value} \end{aligned}$$


The diagram shows two overlapping normal distributions, one green and one blue, on a coordinate system. A horizontal line with a double-headed arrow spans the distance between the peaks of the two distributions. A vertical line with an arrow points from this horizontal line up to the 'difference between group means' part of the formula. Another vertical line with an arrow points from the width of the distributions (representing variability) up to the 'variability of groups' part of the formula. The label 't-value' is placed next to the third line of the equation.

Assumptions of T-test

T-test refers to a univariate hypothesis test based on t-statistic, wherein the mean is known, and population variance is approximated from the sample.

- All data points are independent.
- The sample size is small. Generally, a sample size exceeding 30 sample units is regarded as large, otherwise small but that should not be less than 5, to apply t-test.
- To test the Hypothesis that correlation coefficient in Population is Zero.

t score

The t score is a ratio between the difference between two groups and the difference within the groups. The larger the t score, the more difference there is between groups.

- A large t-score tells you that the groups are different.
- A small t-score tells you that the groups are similar.

One sample t-test

The One Sample t Test determines whether the sample mean is statistically different from a known or hypothesised population mean. The One Sample t Test is a parametric test.

Example :- you have 10 ages and you are checking whether avg age is 30 or not

Two sampled /Independent T-test

The Independent **Samples t Test** or 2-sample t-test compares the means of two independent groups in order to determine whether there is statistical evidence that the associated population means are significantly different.

Example : is there any association between week1 and week2

Paired sampled t-test

The paired sample t-test is also called dependent sample t-test. It's an univariate test that tests for a significant difference between 2 related variables

H0 :- means difference between two sample is 0

H1:- mean difference between two sample is not 0

Types of t-tests

1. An Independent Samples t-test compares the means for two groups.
2. A Paired sample t-test compares means from the same group at different times (say, one year apart).
3. A One sample t-test tests the mean of a single group against a known mean.

Confidence Interval

How far is far enough depends on the standard error. The standard error is represented as $SE(\beta_1)$ in case of β_1 .

The standard error tells us how much our estimate differs from the actual value. In case of estimating the mean of a population.

$$SE = \frac{\sigma}{\sqrt{n}}$$

Where n is the sample size while σ is the standard deviation of the sample.

Z-TEST

A Z test is a statistical hypothesis test which is best used when the population is normally distributed with known variance and population size greater than 30. As per central limit theorem as the sample size grows and number of data points get more than 30, the samples are considered to be normally distributed. Because of this whenever sample size gets bigger than 30, we assume data is normally distributed and we can use Z-test.

A z-test can be used to determine whether two population means are different when the variances are known and the sample size is large.

Z scores

$$z = \frac{x - \mu}{\sigma}$$

μ = Mean

σ = Standard Deviation

$$\text{Standard Error} = \frac{\sigma_x}{\sqrt{N}}$$

$$Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$$

Select Appropriate Statistics: T vs Z vs CHI vs F

- Is data frequency known? If it is known then use chi squares test.
- Is data variance known? If the answer is Yes then use Z statistics.
- otherwise use Student T statistics.

Hypothesis Testing Examples