EDUBRIDGE INDIA

Group Project- HYPOTHESIS TESTING

Meet the Group

LISET

Group MEMBER1

AYISHA

GROUP MEMBER2

Introduction

The ability to estimate population parameters or to test hypotheses about population parameters using sample statistics is one of the main applications of statistics. Whether estimating parameters or testing hypotheses about parameters, both are a part of Inferential Statistics consists of taking a random sample from a group or body (the population), analyzing data from the sample, and reaching conclusions about the population using the sample data

Background

Hypothesis Testing - seeks to validate a supposition based on limited evidence,inferred using a sample from the population.
for eg: By how much does this new drug delay relapse

Estimation Testing- seeks to validate a supposition based on a limited evidence, inferred using a sample from a population
for eg: Does this new drug delay relapse?

Inferential statistics - A set of statistical methods and techniques for infering the characteristics of a population when only a sample is given

Goals

Our First Goal

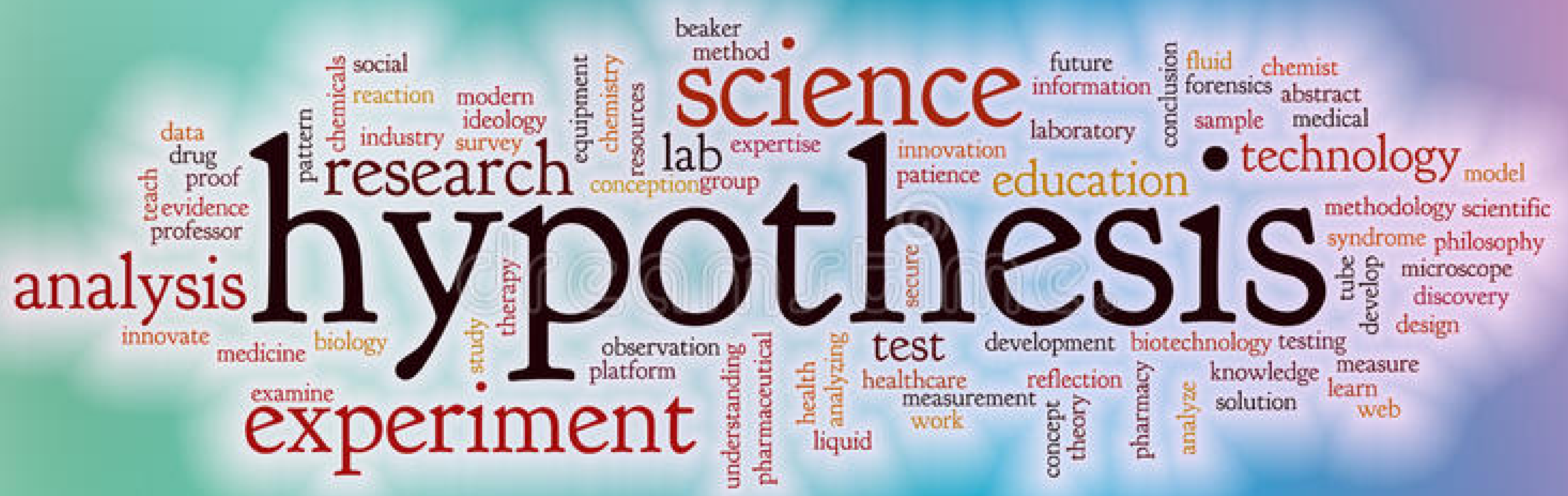
To know what actually Hypothesis means

Our Second Goal

Steps and Types of Hypothesis testing .

Our Third Goal

Practical view of Hypothesis with one of the tests.- t-test



The hypothesis is a definite statement or assertion about the population parameters or equivalently about the probability distribution characterizing a population which we want to verify on the basis of the information available from a sample.

Main Concepts of Hypothesis

01 Null hypothesis

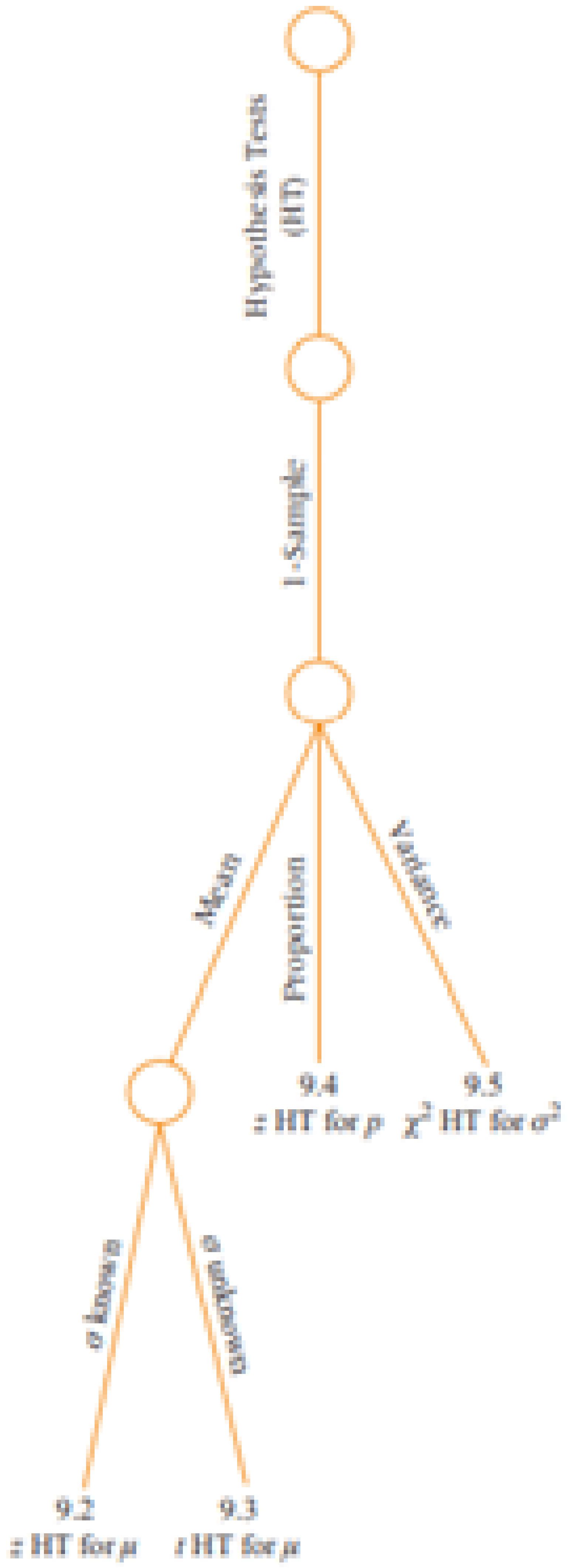
"The hypothesis of no difference ". The neutral or non committal attitude of the statistician before the sample observations are taken is the keynote of the null hypothesis Ho

02 Alternate hypothesis

"The hypothesis of difference". the hypothesis which a researcher wants to accept by rejecting the null hypothesis. In many researches, it is the research hypothesis H1

Hypothesis testing

A two-action decision problem after the experimental sample values have been obtained, the two actions being the acceptance or rejection of the hypothesis under consideration.



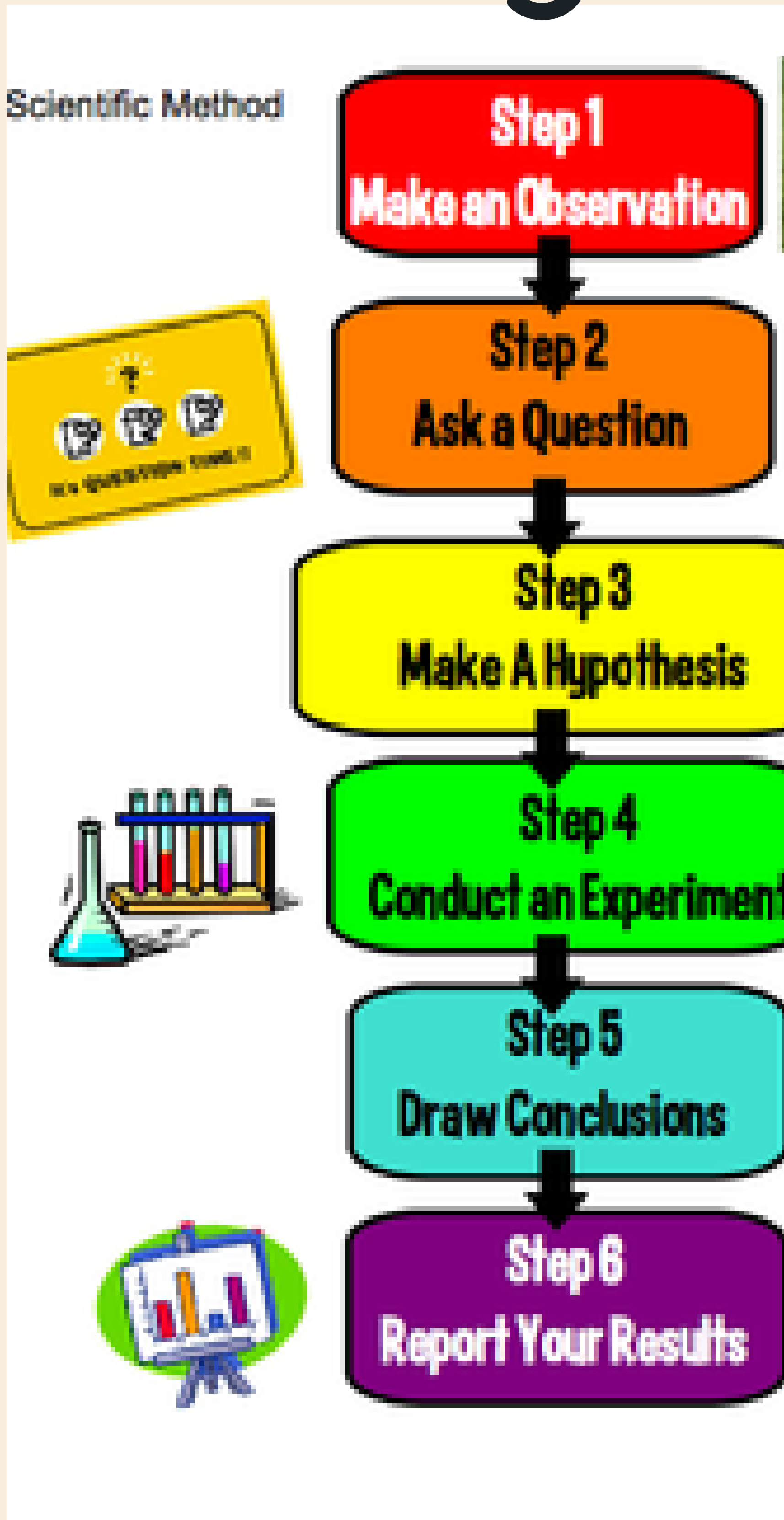
steps involved in hypothesis testing

01 develop a research hypothesis that can be tested mathematically.

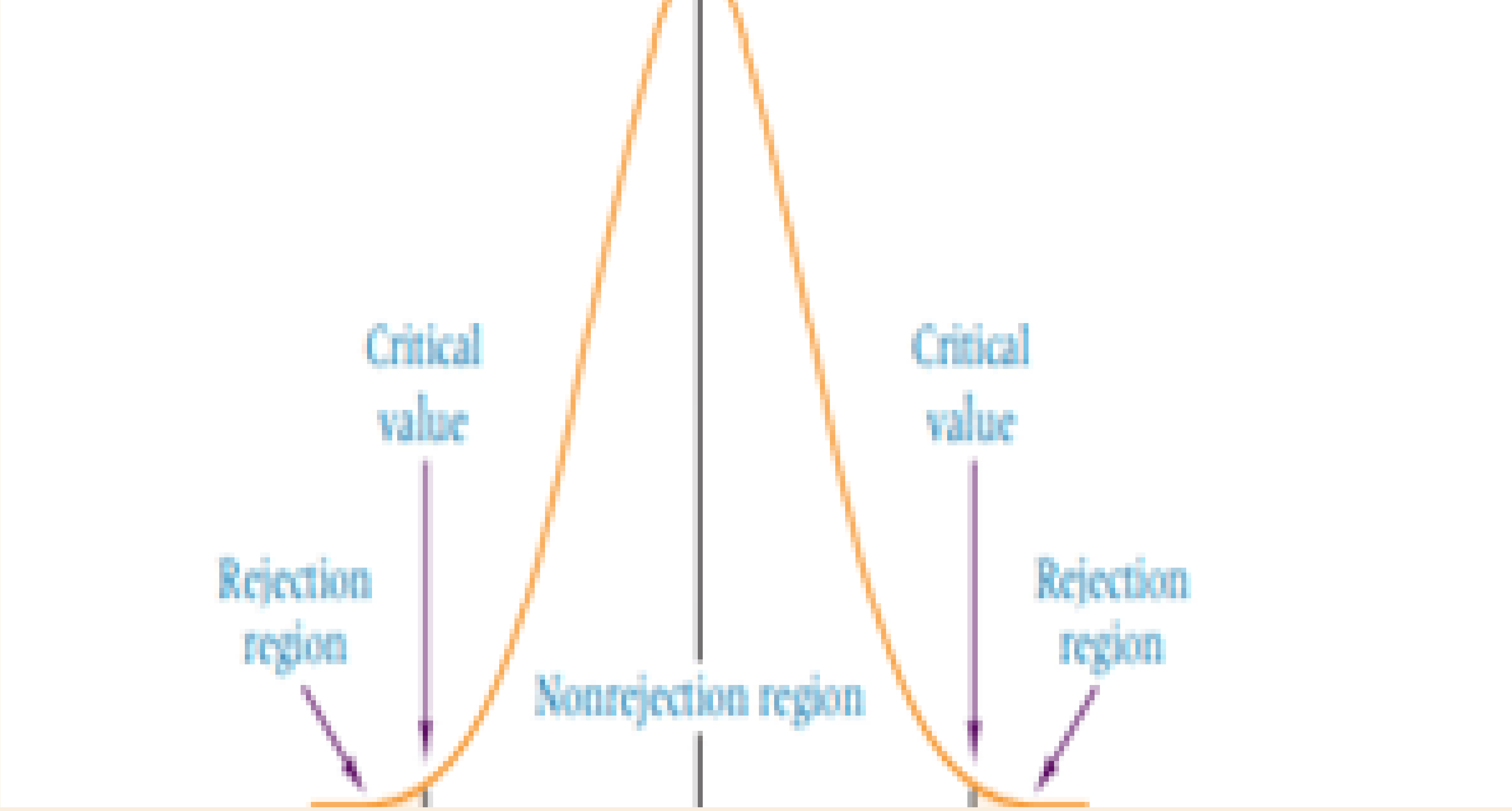
02 formally state the null and alternative hypothesis

03 Decide on the appropriate statistical test, and do the calculations

04 Make your decisions



Test statistic



01 choose a statistic

02 divide the range of possible values for the test statistic into two parts depending upon confidence interval

a The acceptance region

b The critical or rejection region

03 if the value of test statistic is in the acceptance region we accept Ho otherwise reject Ho

Types of errors

01

Type I error

Rejecting a null hypothesis Ho when it is actually true

02

Type II error

Accepting a null hypothesis when it is false

	Null true	Null false
Fail to reject null	Correct decision	Type II error (β)
Reject null	Type I error (α)	Correct decision (power)

Types of Hypothesis testing

-One Sample Tests

- Z-statistic

when population variance is known

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

- t-statistic

When population parameter is known

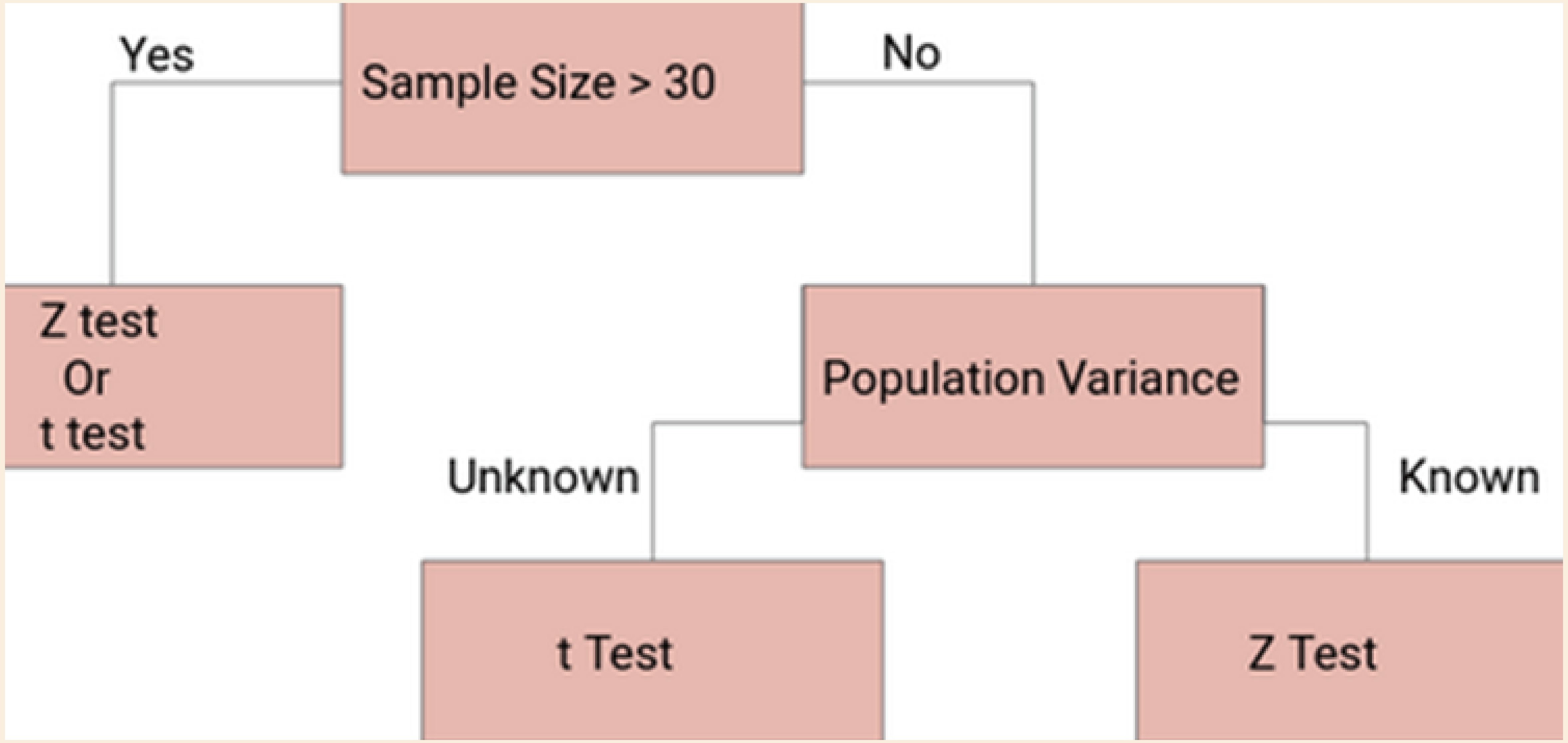
$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

- chisquare- statistic

Non-Parametric test

$$\chi^2 = \frac{ns^2}{\sigma_0^2}$$

Decide about the statistic



Z- Test

Z Test Statistics Formula

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

where x = any value from the population

μ = population mean

σ = population standard deviation

- Z-test is a kind of hypothesis test which ascertains if the averages of the 2 datasets are different from each other when standard deviation or variance is given.The Sample size is large.Normal Distribution for Z, with an average zero and variance = 1.Based on Normal distribution.

Steps for z-Test

1. State the assumptions/conditions (Random Selection, large sample size or normally distributed and the σ is known.

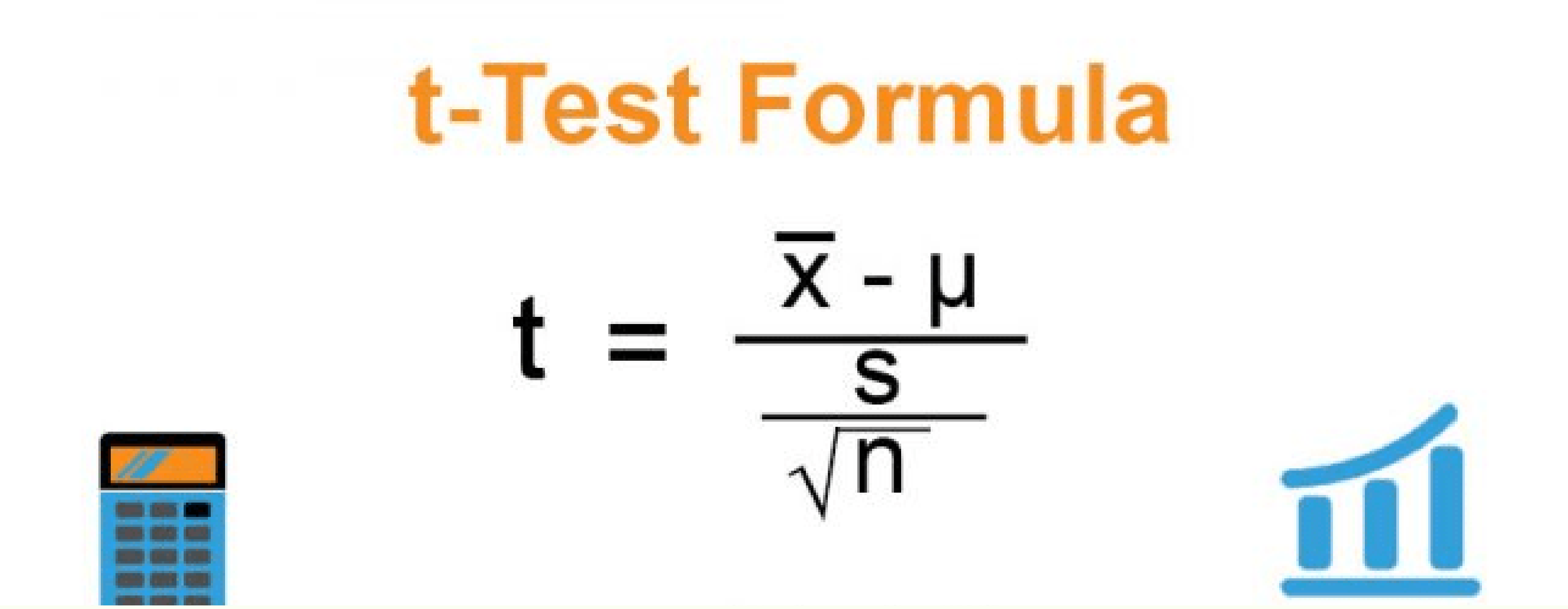
2. State the null and alternative hypothesis and identify the claim. (H_0; H_a, claim)

3. Specify the level of significance. (α)

3. Find the z-score and the p-value. $z = \dfrac{\bar{x} - \mu}{\sigma / \sqrt{n}}$

4. Draw a conclusion about the null hypothesis and the claim.

t-Test



$\bar{x}$: sample mean

μ : population mean

s : sample standard deviation

n : sample size

- The t-test can be referred to as a kind of parametric test that is applied to an identity, how the averages of 2 sets of data differ from each other when the standard deviation or variance is not given.Here the Sample Size is small.Sample values are to be recorded and taken accurately.Based on Student-t distribution.

Types of t-tests

There are three main types of t-test:

·An Independent Samples t-test compares the means for two groups.

·A Paired sample t-test compares means from the same group at different times (say, one year apart).

·A One sample t-test tests the mean of a single group against a known mean.

T-test Steps

1. State the hypothesis
2. State the level of risk (<0.05)
3. Select the test statistics
4. Compute the value of the test statistics
5. Use the appropriate table of critical values
6. Compare the value obtained with the critical value
7. If the value obtained is greater than the critical value, the null hypothesis is rejected.
8. If the value obtained is less than the critical value, the null hypothesis is the most logical explanation(Salkind,2014).

1 Identify the Hypothesis. H_1: $\mu \leq 76"$

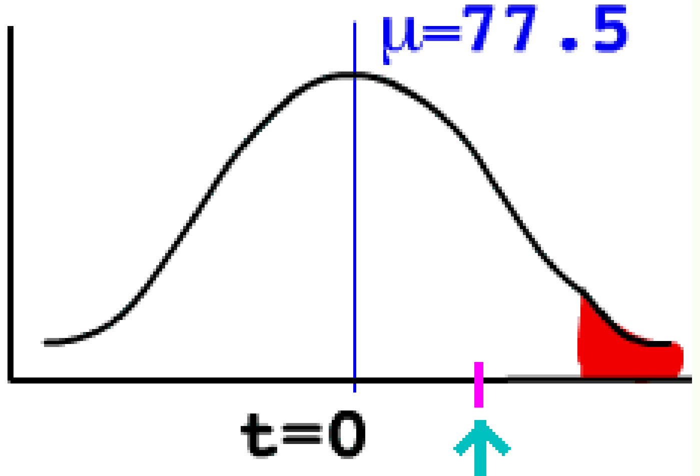
2 State the Alternative H. H_2: $\mu > 76"$

3 Let H with = be H_o, the test hypothesis. H_o: $\mu \leq 76"$ H_A: $\mu > 76"$

4 Relevant stat and distribution? Population Mean, T-Distribution

5 Set α. $0.05;\ 14^o = 1.761$

6 Graph the distribution.



$H_o \leq 76"$

$$t = \frac{77.5 - 76}{5.6/\sqrt{15}}$$

$$t = 1.03740$$

Within acceptable parameters

Chi-Square (χ^2) Formula

$$\chi^2 = \frac{ns^2}{\sigma_o^2}$$

χ^2 -Test

σ = population standard deviation

s= sample mean

Chi Square (χ^2) test is one of the simplest and most commonly used non-parametric tests of significance. Generally the t-test is meant for taking decisions about the population, the chi square test is used to draw inferences about the population dispersion, mainly variance. This test is conducted when we want to test if the given normal population has a specified variance $\sigma^2=\sigma_o^2$. The Chi Square test for variance is generally a right tailed test.

A manufacturing process is expected to produce goods with a specified weight with variance less than 5 units. A random sample of 10 was found to have variance 6.2 units. Is there reason to suspect that the process variance has increased. (α=0.05)

Solution

Given σ_o^2=5, n=10, s^2=6.2

Here we are testing

H_o: σ^2=5 against H_1: σ^2>5

Given α=0.05. The beat critical region is w= $\chi^2>\chi\alpha^2$

From χ^2 table $\chi\alpha^2$ for 9 df and probability α=0.05 is 16.92.

The test statistic is

$$\chi^2=ns^2/\sigma_o^2 = 10*6.2/5 = 12.4<16.92$$

Since the calculated value of χ^2 lies in the acceptance region. H_o is accepted. That is, there is no reason to suspect that the process variance has increased

Practical view

```
import pandas as pd
import numpy as np
import seaborn as sns
import scipy.stats as stats
import math
import matplotlib.pyplot as plt
import warnings
warnings.filterwarnings('ignore')
```

```
student_performance=pd.read_csv("D:\MY STUDY MATERIALS\SOME DATAS FOR ANALYSIS\StudentsPerformance.csv")
```

```
student_performance
```

	gender	race/ethnicity	parental level of education	lunch	test preparation course	math score	reading score	writing score
0	female	group B	bachelor's degree	standard	none	72	72	74
1	female	group C	some college	standard	completed	69	90	88
2	female	group B	master's degree	standard	none	90	95	93
3	male	group A	associate's degree	free/reduced	none	47	57	44
4	male	group C	some college	standard	none	76	78	75
...	...	...	...	...	...	...	...	...
995	female	group E	master's degree	standard	completed	88	99	95
996	male	group C	high school	free/reduced	none	62	55	55
997	female	group C	high school	free/reduced	completed	59	71	65
998	female	group D	some college	standard	completed	68	78	77
999	female	group D	some college	free/reduced	none	77	86	86

```
Average_mathscore=student_performance['math score'].mean()
Average_mathscore
```

```
66.089
```

```
Average_mathscore=student_performance['math score']
```

```
student_performance.shape
```

```
(1000, 8)
```

```
samp_students=student_performance.sample(100)
samp_students
```

	gender	race/ethnicity	parental level of education	lunch	test preparation course	math score	reading score	writing score
665	female	group C	some high school	free/reduced	completed	50	60	60
456	female	group D	bachelor's degree	standard	none	79	89	89
676	female	group E	some college	standard	completed	73	78	76
969	female	group B	bachelor's degree	standard	none	75	84	80
515	female	group C	some high school	standard	completed	76	87	85
...	...	...	...	...	...	...	...	...
936	male	group A	associate's degree	standard	none	67	57	53
197	male	group E	high school	free/reduced	none	55	56	51
915	female	group E	some college	standard	none	68	70	66
945	female	group C	associate's degree	standard	none	54	61	58
42	female	group B	associate's degree	standard	none	53	58	65

In [8]:
```
samp_Average_mathscore=samp_students['math score'].mean()
samp_Average_mathscore
```
Out[8]:
```
65.84
```
In [9]:
```
samp_Average_mathscore=samp_students['math score']
```

keeping the hypothesis

- Ho: sample mean = population mean
- H1: sample mean <population mean
- confidence level 0.05

In [10]:
```
stats.ttest_1samp(a=samp_Average_mathscore,popmean=Average_mathscore.mean())
```
Out[10]:
```
Ttest_1sampResult(statistic=-0.18796372149008772, pvalue=0.8512898207299664)
```

This test statistic tells us how the sample mean deviates from the null hypothesis. if the t-statistic lies outside the quantiles of the t-distributon corresponding to our confidence level and degrees of freedom,we reject the null hypothesis.

To check the quantiles and df

we take df = n-1,n= number of samples

To check for the lower quantile

```
In [12]: stats.t.ppf(q=0.025,df=99)
Out[12]: -1.9842169515086832
```

To check for the upper quantile

```
In [13]: stats.t.ppf(q=0.975,df=99)
Out[13]: 1.9842169515086827
```

lower quantile and upper quantile will be same with opposite sign

since the t statistic lies within the range we accept the null hypothesis so p value should be less than 0.5

```
sigma=samp_Average_mathscore.std()/math.sqrt(100)
T_Test=stats.t.interval(0.95,df=99,loc=samp_Average_mathscore.mean(),scale=sigma)
```

```
T_Test
```

```
(64.71105683204202, 70.50894316795798)
```

The values of mean lies between the given range which indicates we accept the null hypothesis .

Thank You!

Do you have any questions for us?
We will be happy to clear the doubts.