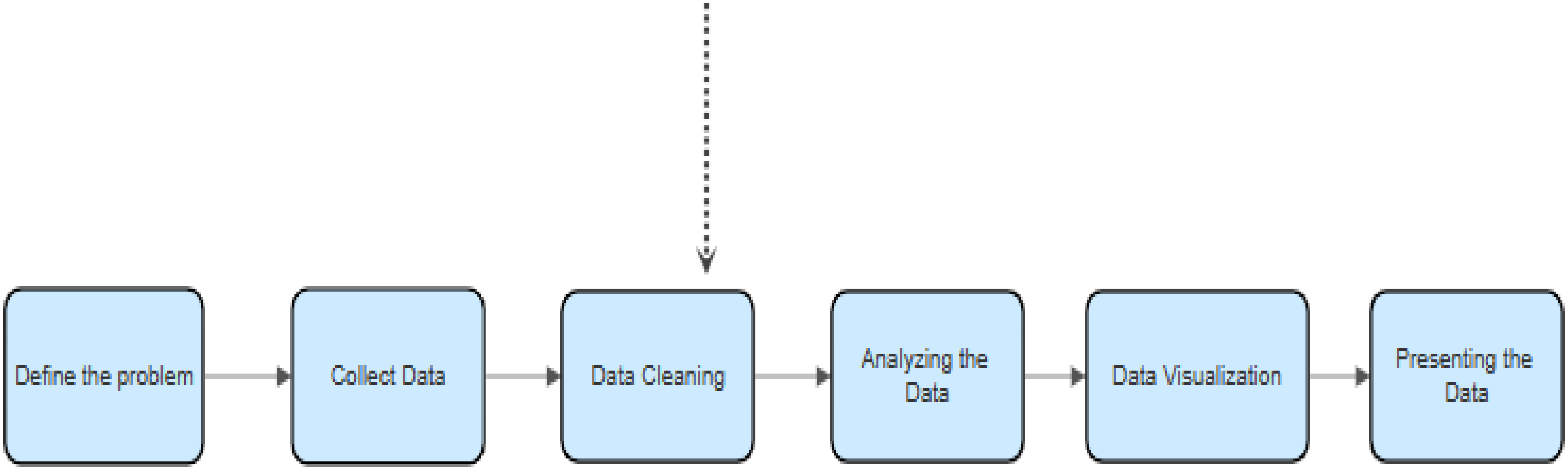
Data Analytics

Data Analysis

- Data is Everywhere, in sheets, in social media platforms, in product reviews and feedback, everywhere. In this latest information age it's created at blinding speeds and, when data is analyzed correctly, can be a company's most valuable asset. "To grow your business even to grow in your life, sometimes all you need to do is Analysis!"
- For data analysis start with a clear objective, gather relevant data, clean and preprocess, use statistical methods or models, interpret results, and communicate findings effectively for informed decision-making.

Six Steps of Data Analysis Process



Define the problem

Collect Data

Data Cleaning

Analyzing the Data

Data Visualization

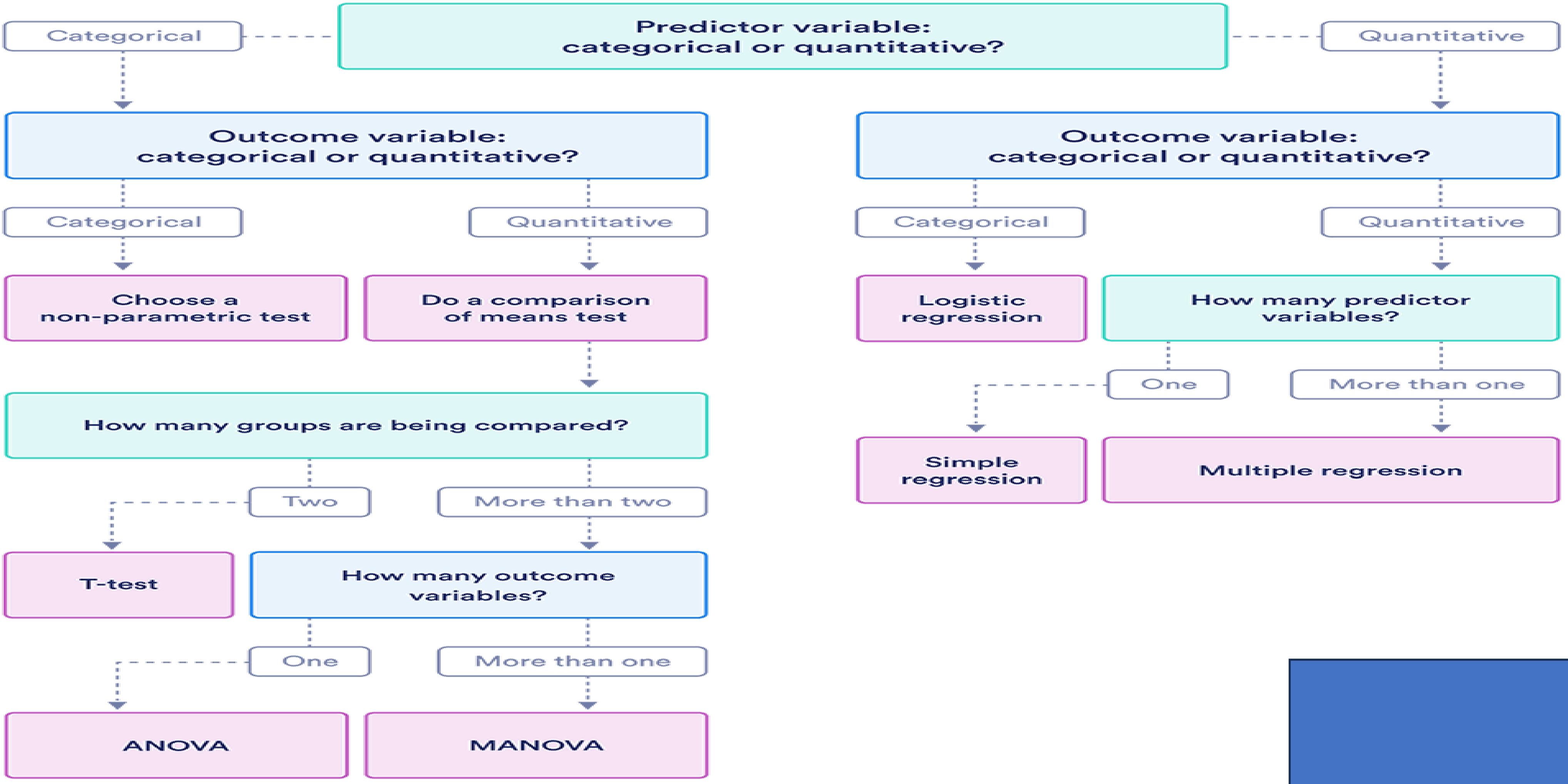
Presenting the Data

Statistical Tests

- Statistical Tests are conducted to test the hypothesis and to find the inferences about the population.
- For that, samples are selected & various tests are performed on them to find the inference about the population under study.
- These are of two types: Parametric Tests & Non-Parametric Tests.

Choosing a statistical test

This flowchart helps you choose among parametric tests



Predictor variable: categorical or quantitative?

Categorical

Outcome variable: categorical or quantitative?

Categorical

Choose a non-parametric test

Quantitative

Do a comparison of means test

How many groups are being compared?

Two

T-test

More than two

How many outcome variables?

One

ANOVA

More than one

MANOVA

Quantitative

Outcome variable: categorical or quantitative?

Categorical

Logistic regression

Quantitative

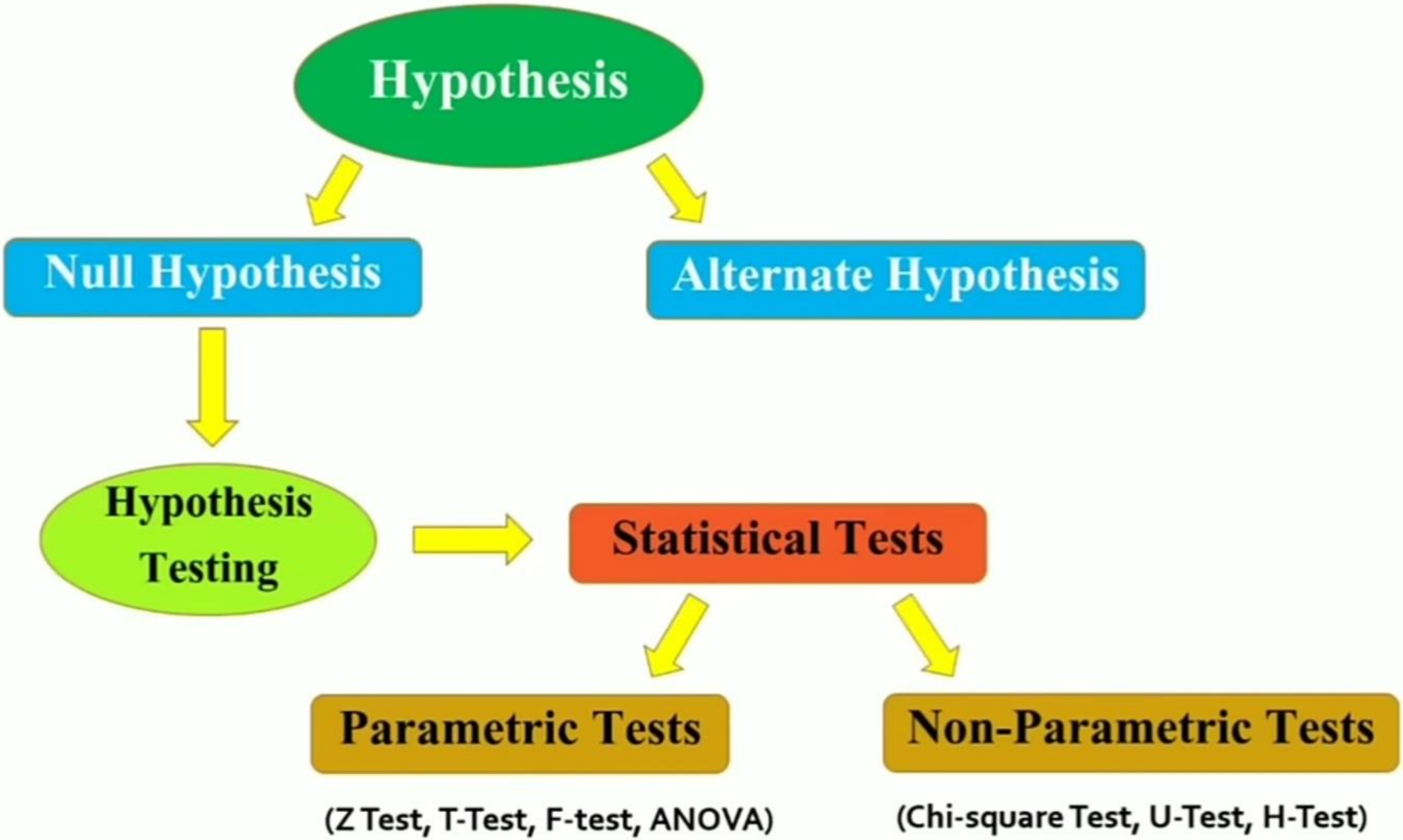
How many predictor variables?

One

Simple regression

More than one

Multiple regression



Hypothesis

Null Hypothesis

Alternate Hypothesis

Hypothesis Testing

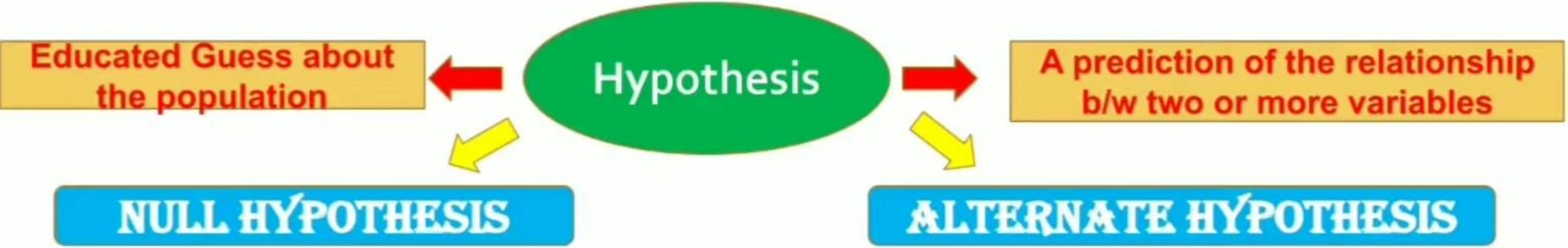
Statistical Tests

Parametric Tests

(Z Test, T-Test, F-test, ANOVA)

Non-Parametric Tests

(Chi-square Test, U-Test, H-Test)



Educated Guess about the population

Hypothesis

A prediction of the relationship b/w two or more variables

NULL HYPOTHESIS

□ A statement which states that there is no relationship b/w the variables.

□ For example, increase in number of cancer patients is not due to the increase in the pollution level.

□ It is denoted by Ho.

□ It is an exact opposite of what an investigator predicts or expects.

ALTERNATE HYPOTHESIS

□ A statement which states that there is relationship b/w the variables.

□ For example, Cancer patients are increasing due to increase in the pollution level.

□ It is denoted by H1 or Ha.

□ It is an exactly what an investigator predicts or expects.

Parametric Tests

- Parameters tests are applied under the circumstances where the population is normally distributed or is assumed to be normally distributed.
- Parameters like mean, standard deviation etc. are used.
- For example, T-test, Z-Test, F-test, ANOVA, Pearson's Coefficient correlation.
- These are applied where the data is quantitative.
- These are applied where the scale of measurement is either an interval or a ratio scale.

Parametric Tests

- T-Test
- Z-Test
- F-Test
- ANOVA

T-Test

- It is a parametric test of hypothesis testing based on Student's T distribution.
- It was developed by William Sealy Gosset.
- It is essentially, testing the significance of the difference of the mean values when the sample size is small (i.e. less than 30) & when population standard deviation is not available.
- It assumes:

✓ Population distribution is normal, and

✓ Samples are random & independent.

✓ Sample size is small.

✓ Population standard deviation is not known.

- Mann-Whitney 'U' Test is a non-parametric counterpart of T-test.

Z-Test

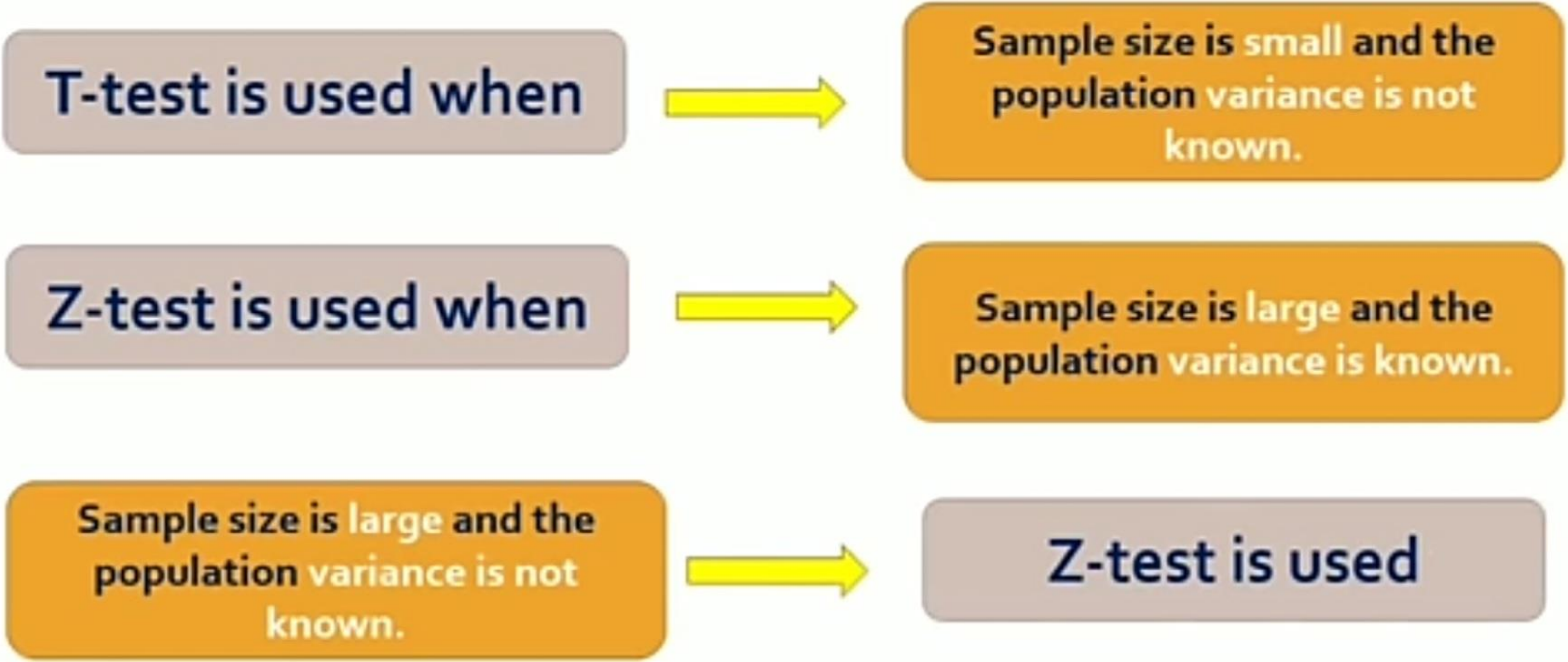
- It is a parametric test of hypothesis testing.
- It is used to determine whether the means are different when the population variance is known & the sample size is large (i.e. greater than 30).
- It assumes:

✓ Population distribution is normal, and

✓ Samples are random & independent.

✓ Sample size is large.

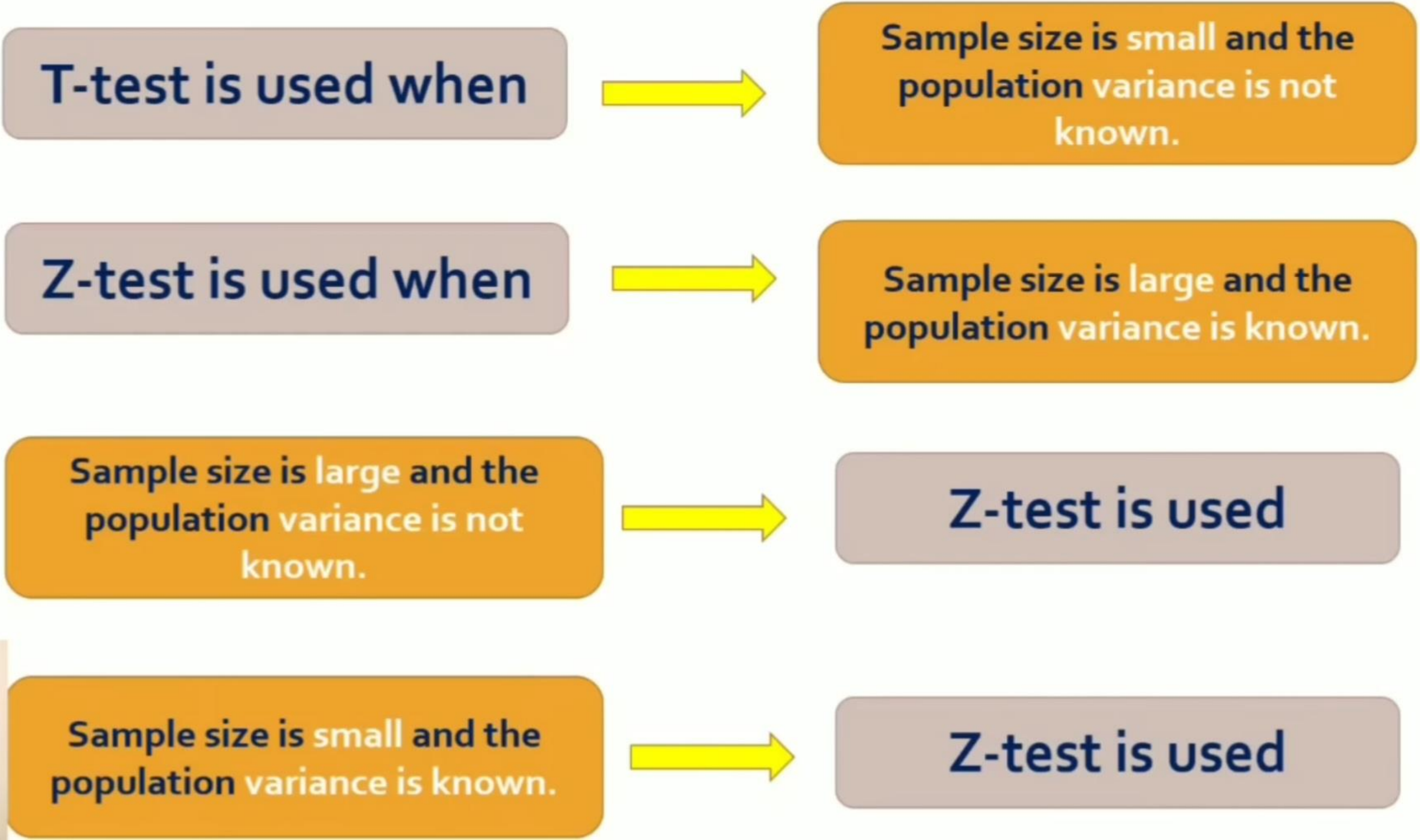
✓ Population standard deviation is known.



T-test is used when → Sample size is small and the population variance is not known.

Z-test is used when → Sample size is large and the population variance is known.

Sample size is large and the population variance is not known. → Z-test is used



T-test is used when → Sample size is small and the population variance is not known.

Z-test is used when → Sample size is large and the population variance is known.

Sample size is large and the population variance is not known. → Z-test is used

Sample size is small and the population variance is known. → Z-test is used

A T-Test can be a

One Sample T-Test

To compare sample mean with that of the population mean.

$$t = \frac{\overline{x} - \mu_0}{s/\sqrt{n}}$$

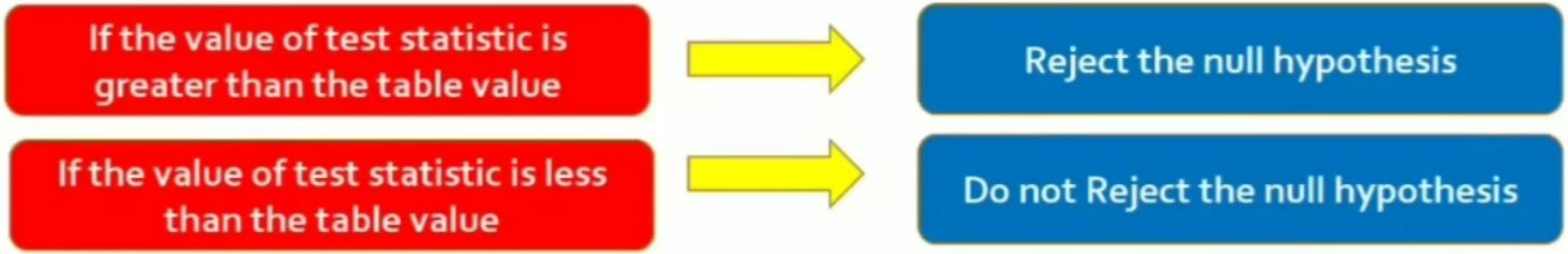
$\bar{x}$ is the sample mean
s is sample standard deviation
n is sample size
μ is the population mean

Two Sample T-Test

To compare means of two different samples.

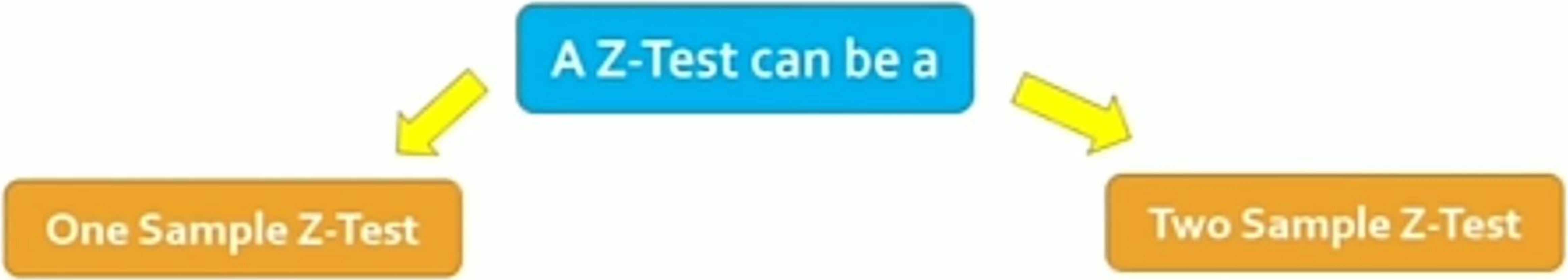
$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

$\bar{X}_1$ is the sample mean of Ist group
$\bar{X}_2$ is the sample mean of 2nd group
s1 is sample 1 standard deviation
S2 is sample 2 standard deviation
n is sample size



If the value of test statistic is greater than the table value → Reject the null hypothesis

If the value of test statistic is less than the table value → Do not Reject the null hypothesis



A Z-Test can be a

One Sample Z-Test

Two Sample Z-Test

To compare sample mean with that of the population mean.

$$\text{Z-test} = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

$\bar{x}$ is the sample mean
σ is population standard deviation
n is sample size
μ is the population mean

To compare means of two different samples.

$$z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

$\bar{X}_1$ is the sample mean of Ist group
$\bar{X}_2$ is the sample mean of 2^{nd} group
σ_1 is population 1 standard deviation
σ_2 is population 2 standard deviation
n is sample size

F-Test

- It is a parametric test of hypothesis testing based on Snedecor F-distribution.
- F-test is named after its test statistic, F, which was named in the honour of Sir Ronald Fisher.
- It is a test for the null hypothesis that two normal populations have the same variance.
- An F-test is regarded as a comparison of equality of sample variances.
- F-statistic is simply a ratio of two variances.
- It is calculated as:

$$F = \frac{s_1^2}{s_2^2}$$

$$where\ s^2 = \frac{\sum(x-\bar{x})^2}{n-1}$$

- By changing the variance in the ratio, F-test becomes a very flexible test. It can then be used to:

✓ Test the overall significance for a regression model

✓ To compare the fits of different models and

✓ To test the equality of means.

- It assumes:

✓ Population distribution is normal, and

✓ Samples are drawn randomly & independently.

ANOVA

- Also called as (Analysis of Variance),it is a parametric test of hypothesis testing.
- It was developed by Ronald Fisher, also referred to as Fisher's ANOVA.
- It is an extension of T-Test & Z-test.
- It is used to test the significance of the differences of the mean values among more than two sample groups.
- It uses F-test to statistically test the equality of means & the relative variance between them.
- It assumes:

✓ Population distribution is normal, and

✓ Samples are random & independent.

✓ Homogeneity of sample variance.

- One-way ANOVA & Two-way ANOVA are its types.

F-statistic =
variance between the sample means /
variance within the samples

Non-Parametric Tests

- Chi-square test
- Mann-Whitney Test (U-Test)
- Kruskal-Wallis Test (H-Test)

Non-Parametric Tests

- Non- Parametric tests are applied under the circumstances where the population is not normally distributed (skewed distribution) or is not assumed to be normally distributed.
- Where parametric tests cannot be applied, then non-parametric tests come into play.
- These tests are also called as Distribution-free tests.
- Parameters like mean, standard deviation etc. are not used.
- For example, Chi-square test, U-Test (Mann Whitney Test), H-test (Kruskal Wallis Test), Spearman’s Rank correlation test.
- These are applied where the data is qualitative.
- These are applied where the scale of measurement is either an ordinal or a nominal scale.

Chi-Square Test

- It is a non-parametric test of hypothesis testing.
- As a non-parametric test, chi-square can be used (i) as a test of goodness of fit and (ii) as a test of independence of two variables.
- It helps in assessing the goodness of fit b/w a set of observed values & those expected theoretically.
- It makes comparison between the expected frequencies & the observed frequencies.
- Greater the difference, greater is the value of Chi-square.
- If there is no difference between the expected & observed frequencies, then the value of Chi-square is zero.
- It is also called as the 'Goodness of Fit Test' which determines whether a particular distributions fits the observed data or not.
- It is calculated as:

$$X^2 = \Sigma(O - E)^{2/}\mathbf{E}$$

- Chi-square is also used to test for the independence of two variables.

- Conditions for Chi-square Test:
- Observations recorded and used are collected on a random basis.
- All the items in the sample must be independent.
- No group should contain very few items, say less than 10.
- The overall number of items must also be reasonably large. It should normally be at least 50, howsoever small the number of groups may be.

- Chi-square as a parametric test is used as a test for population variance on the basis of sample variance.
- If we take each one of a collection of sample variances, divide them by the known population variance and multiply these quotients by (n – 1), where n means the number of items in the sample, we get the value of chi-square.
- It is calculated as:

$$\chi^2 = \frac{\sigma_s^2}{\sigma_p^2}(n-1)$$

Mann-Whitney U-Test

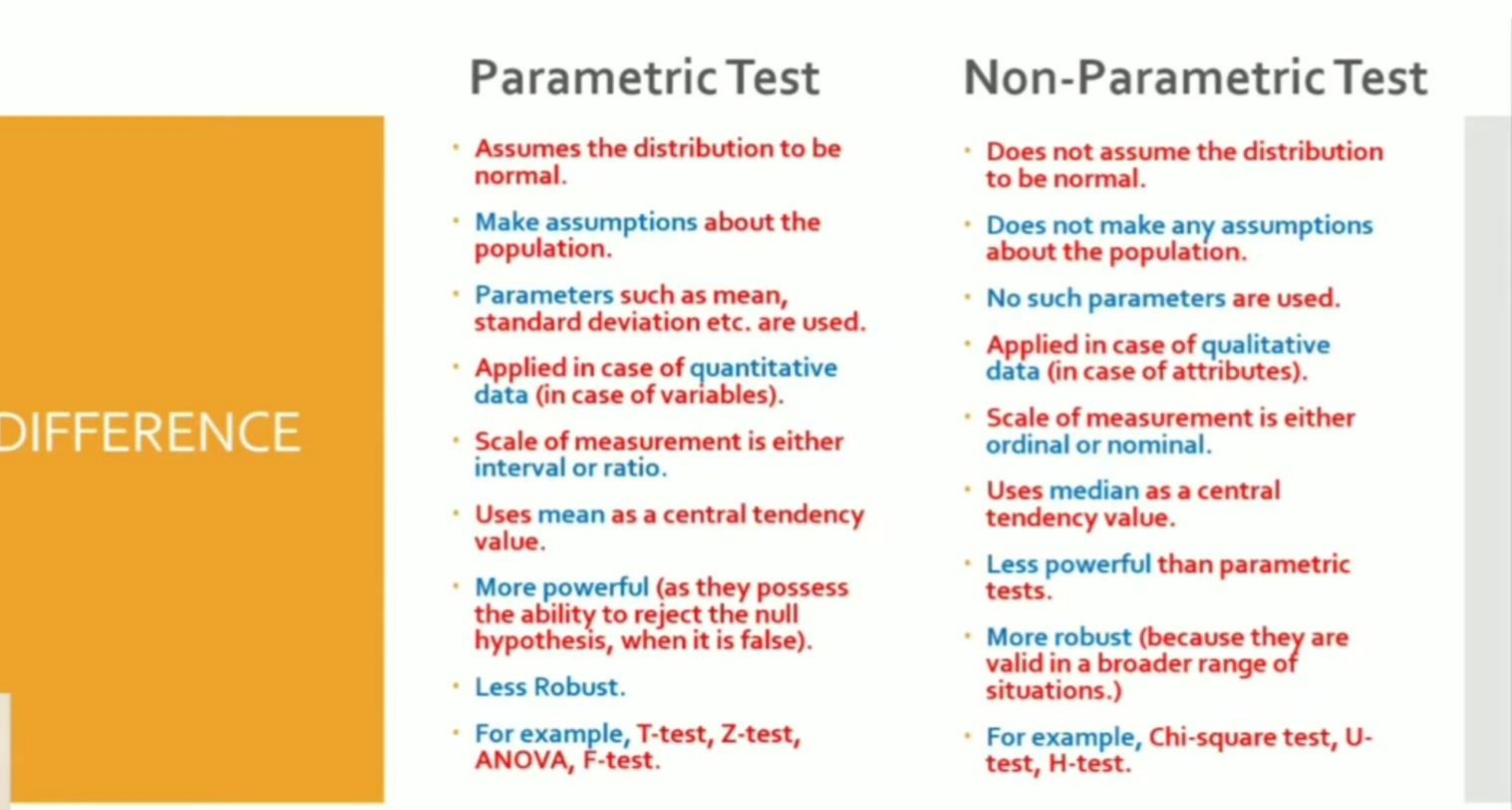
- It is a non-parametric test of hypothesis testing.
- This test is used to investigate whether two independent samples were selected from population having the same distribution.
- It is a true non-parametric counterpart of T-test & gives the most accurate estimates of significance especially when sample sizes are small & the population is not normally distributed.
- It is based on the comparison of every observation in the first sample with every observation in the other sample.
- The test statistic used here is ‘U’.
- Maximum value of ‘U’ is ‘n1*n2’ and minimum value is zero.
- It is also known as:

✓ Mann-Whitney Wilcoxon Test

✓ Mann-Whitney Wilcoxon Rank Sum Test

Kruskal-Wallis H-Test

- It is a non-parametric test of hypothesis testing.
- This test is used for comparing two or more independent samples of equal or different sample sizes.
- It extends the Mann-Whitney U-Test which is used for comparing only two groups.
- One-Way ANOVA is the parametric equivalent of this test. And that's why, it's also called as 'One-way ANOVA on ranks'.
- It uses ranks instead of actual data.
- It does not assume the population to be normally distributed.
- The test statistic used here is 'H'.



DIFFERENCE

Parametric Test	Non-Parametric Test
Assumes the distribution to be normal.	Does not assume the distribution to be normal.
Make assumptions about the population.	Does not make any assumptions about the population.
Parameters such as mean, standard deviation etc. are used.	No such parameters are used.
Applied in case of quantitative data (in case of variables).	Applied in case of qualitative data (in case of attributes).
Scale of measurement is either interval or ratio.	Scale of measurement is either ordinal or nominal.
Uses mean as a central tendency value.	Uses median as a central tendency value.
More powerful (as they possess the ability to reject the null hypothesis, when it is false).	Less powerful than parametric tests.
Less Robust.	More robust (because they are valid in a broader range of situations.)
For example, T-test, Z-test, ANOVA, F-test.	For example, Chi-square test, U-test, H-test.

Advantage of Parametric Test

More powerful.

Disadvantage of Parametric Test

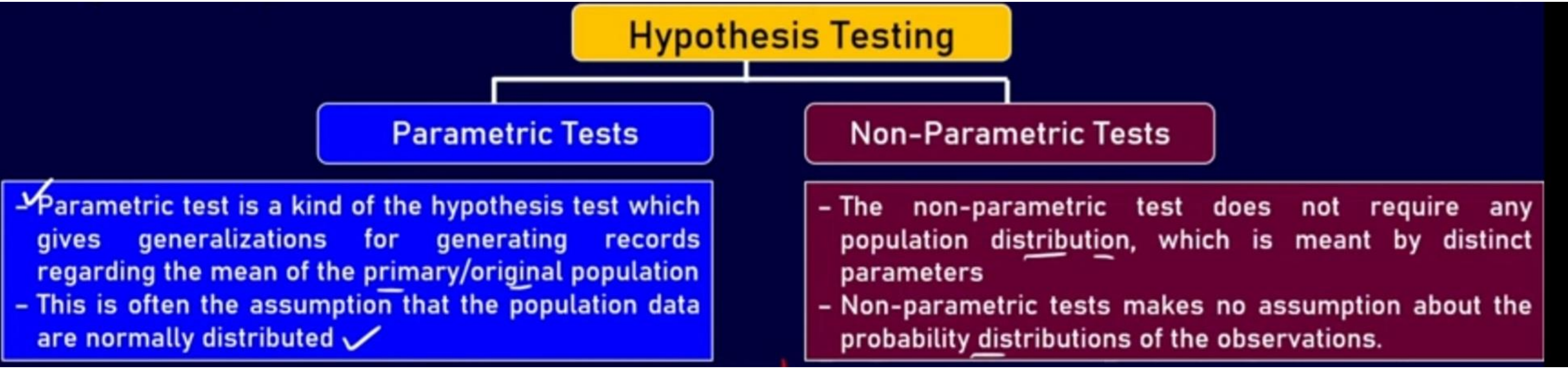
Less Robust.

Advantage of Non- Parametric Test

More Robust.

Disadvantage of Non- Parametric Test

Less powerful.



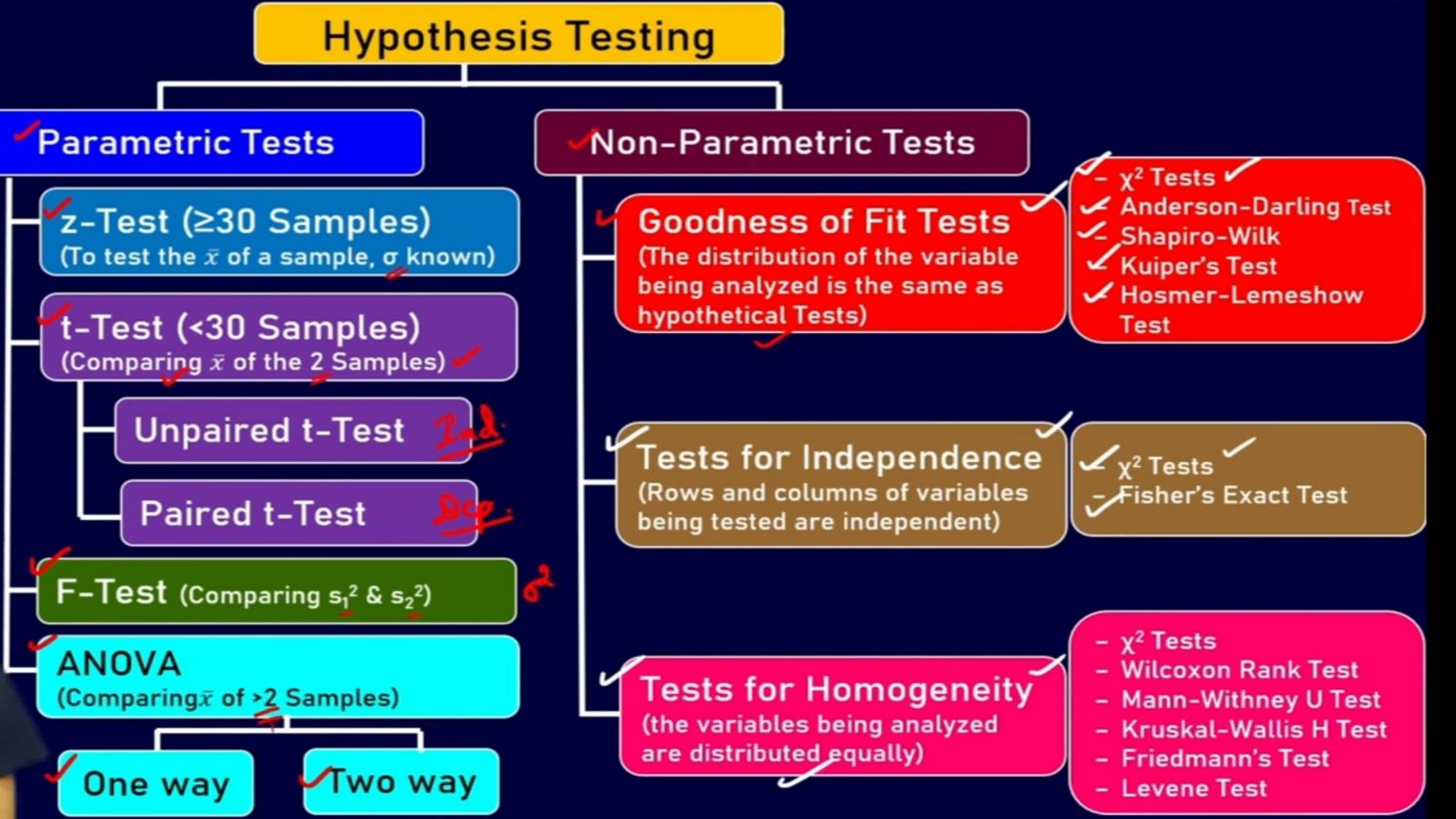
Hypothesis Testing

Parametric Tests

- ✓Parametric test is a kind of the hypothesis test which gives generalizations for generating records regarding the mean of the primary/original population
- This is often the assumption that the population data are normally distributed ✓

Non-Parametric Tests

- The non-parametric test does not require any population distribution, which is meant by distinct parameters
- Non-parametric tests makes no assumption about the probability distributions of the observations.



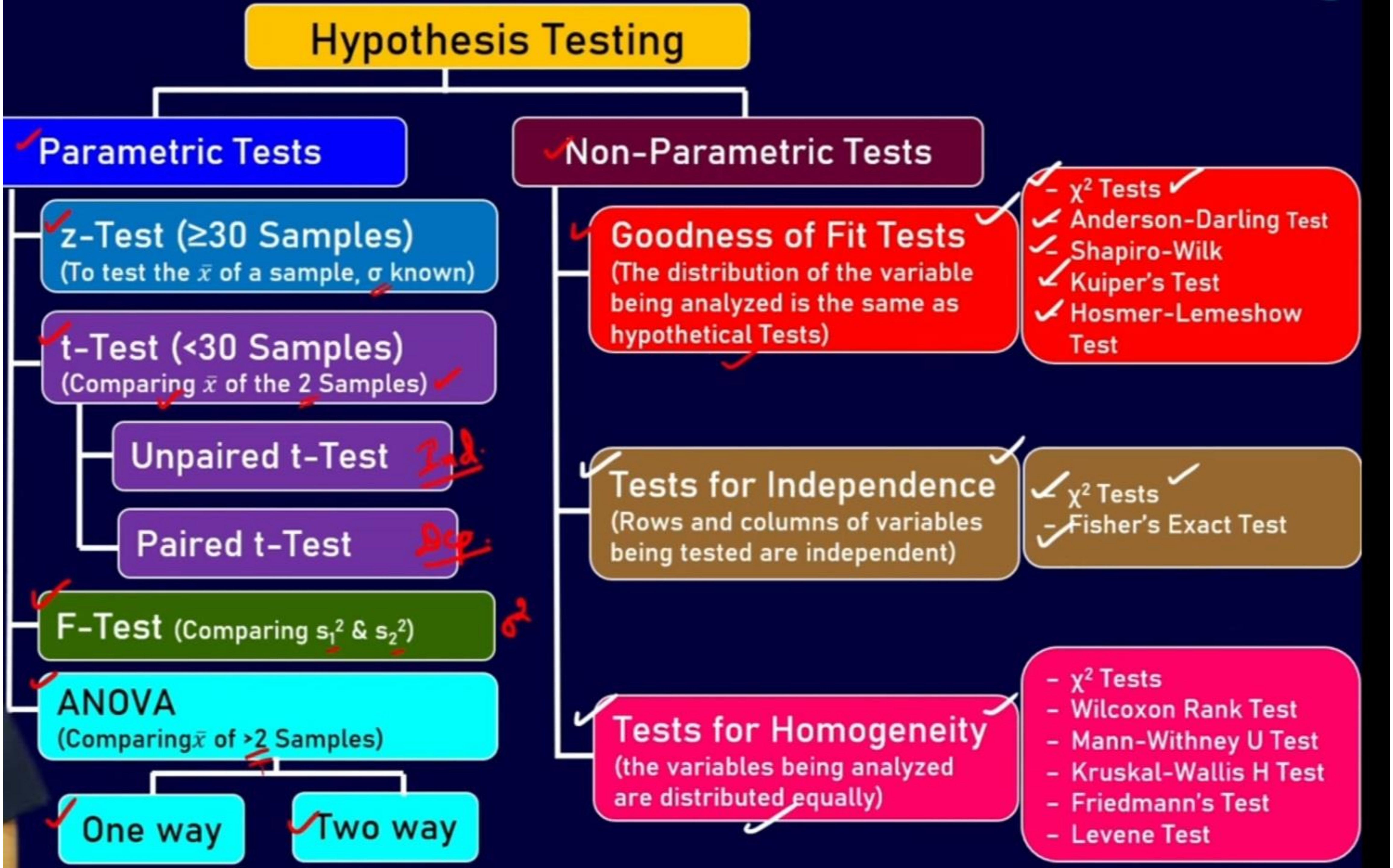
Hypothesis Testing

Parametric Tests

- z-Test ($\geq$30 Samples) (To test the $\bar{x}$ of a sample, σ known)
- t-Test (<30 Samples) (Comparing $\bar{x}$ of the 2 Samples)
  - Unpaired t-Test Ind.
  - Paired t-Test Dep.
- F-Test (Comparing s_1^2 & s_2^2) σ^2
- ANOVA (Comparing $\bar{x}$ of >2 Samples)
  - One way
  - Two way

Non-Parametric Tests

- Goodness of Fit Tests (The distribution of the variable being analyzed is the same as hypothetical Tests)
  - χ^2 Tests
  - Anderson-Darling Test
  - Shapiro-Wilk
  - Kuiper's Test
  - Hosmer-Lemeshow Test
- Tests for Independence (Rows and columns of variables being tested are independent)
  - χ^2 Tests
  - Fisher's Exact Test
- Tests for Homogeneity (the variables being analyzed are distributed equally)
  - χ^2 Tests
  - Wilcoxon Rank Test
  - Mann-Withney U Test
  - Kruskal-Wallis H Test
  - Friedmann's Test
  - Levene Test



Hypothesis Testing

Parametric Tests

- z-Test (≥30 Samples)
  (To test the $\bar{x}$ of a sample, σ known)
- t-Test (<30 Samples)
  (Comparing $\bar{x}$ of the 2 Samples)
  - Unpaired t-Test Ind.
  - Paired t-Test Dep.
- F-Test (Comparing s_1^2 & s_2^2) σ^2
- ANOVA
  (Comparing $\bar{x}$ of >2 Samples)
  - One way
  - Two way

Non-Parametric Tests

- Goodness of Fit Tests
  (The distribution of the variable being analyzed is the same as hypothetical Tests)
  - χ^2 Tests
  - Anderson-Darling Test
  - Shapiro-Wilk
  - Kuiper's Test
  - Hosmer-Lemeshow Test
- Tests for Independence
  (Rows and columns of variables being tested are independent)
  - χ^2 Tests
  - Fisher's Exact Test
- Tests for Homogeneity
  (the variables being analyzed are distributed equally)
  - χ^2 Tests
  - Wilcoxon Rank Test
  - Mann-Withney U Test
  - Kruskal-Wallis H Test
  - Friedmann's Test
  - Levene Test