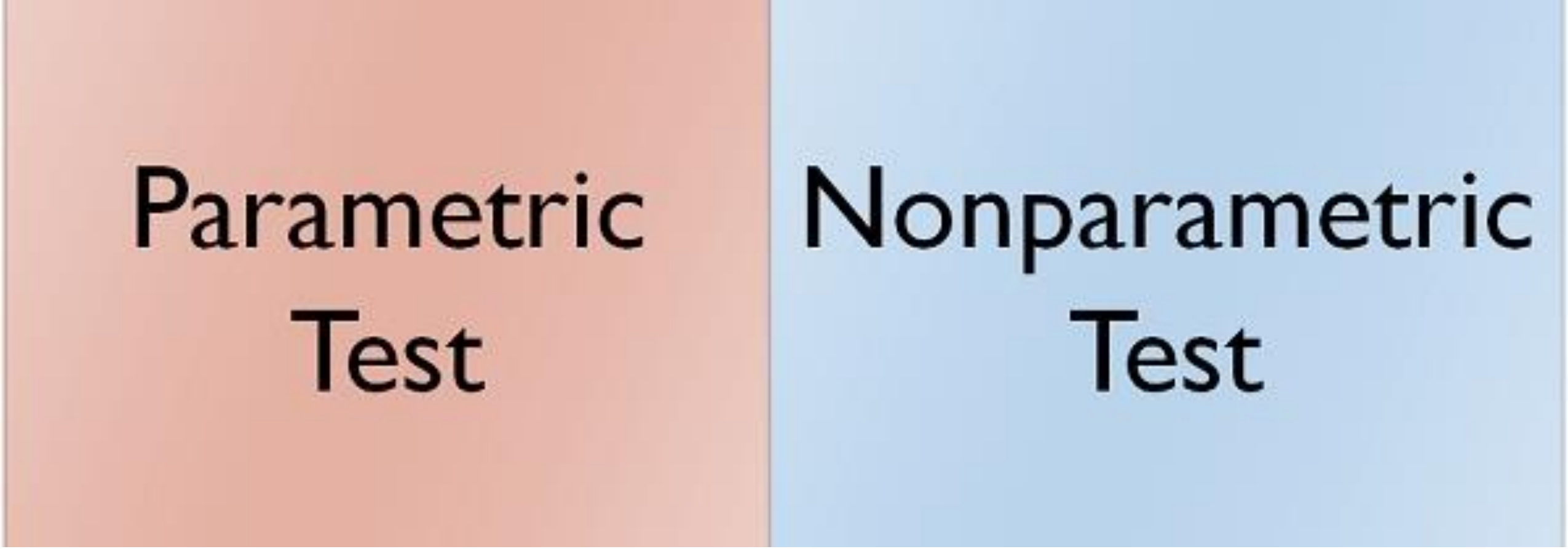


Alok Yadav

https://www.linkedin.com/in/alokyadavonline/

A "Statistic" is a measure of a property of a Sample, for example, the Sample Mean or Sample Standard Deviation.

The corresponding term for a Population or Process is "Parameter".

The most commonly used statistical tests are "Parametric", that is, they require that one or more Parameters meet certain conditions or "assumptions".

Most frequently, the assumption is that the Distribution of the Population or Process is roughly **Normal**.

Roughly equal Variance is also a common assumption.

If these conditions are not met, the Parametric test cannot be used, and a Nonparametric test must be used instead. This table shows the Nonparametric test that can be used in place of several common Parametric tests.

Parametric versus Non Parametric test.

Parametric test	Non parametric test
Specific assumptions are made regarding the population	No specific assumptions are made regarding the population
Parametric test is powerful if it is exists	Not powerful like parametric test
Test statistics based on distribution	Test statistics is arbitrary

One Categorical Predictor

- **So you want to explain a numeric outcome variable with one categorical predictor.**
- ***Question:*** **How many groups does your categorical variable have?**

Two Groups

Example: You want to determine if cats and dogs sleep a different amount.

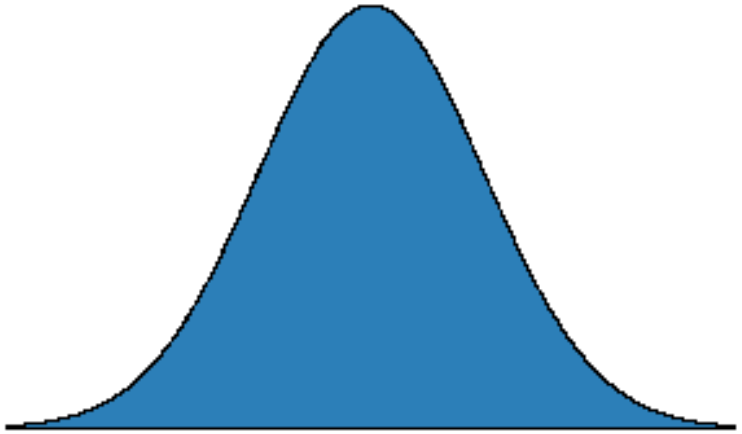
More than two Groups

Example: You want to see if math, biology, and English majors spend a different amount on textbooks.

Testing a Claim

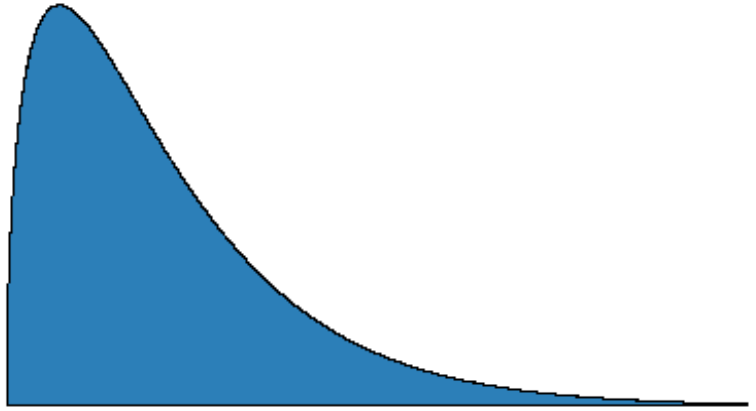
- **So you want to test if the mean of a numeric variable from a single population equals some claimed value.**
- ***Question:*** **What is the shape of your variable's distribution?**

Normal



A normal distribution is the classic bell curve. It is symmetric, with most cases falling near the mean and fewer cases falling away from the mean in either tail.

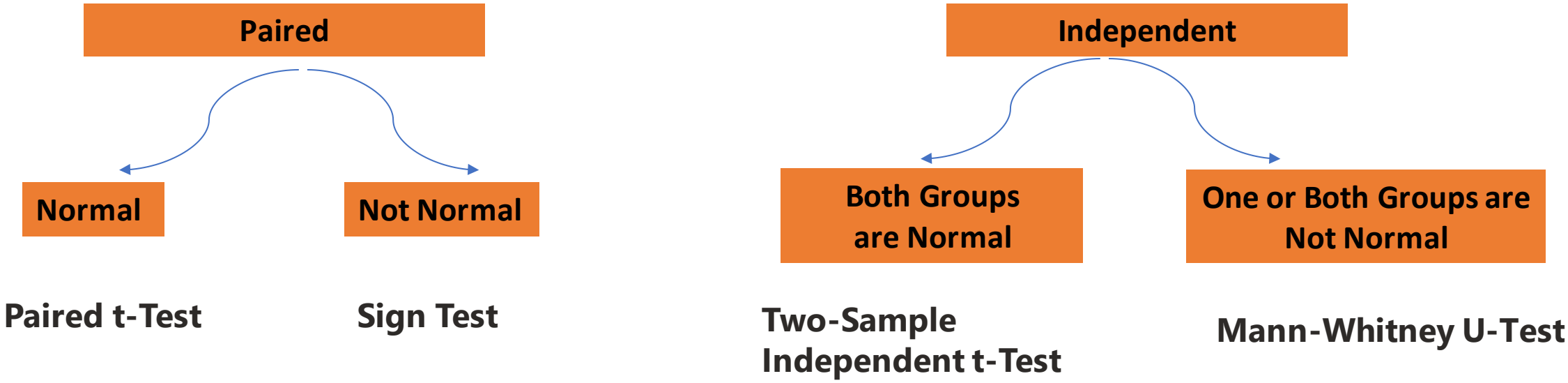
Not Normal



If the variable's distribution is skewed (has a tail in either direction), has more than one mode (peak), or contains outliers, then it will fail the normality assumption of many common statistical methods.

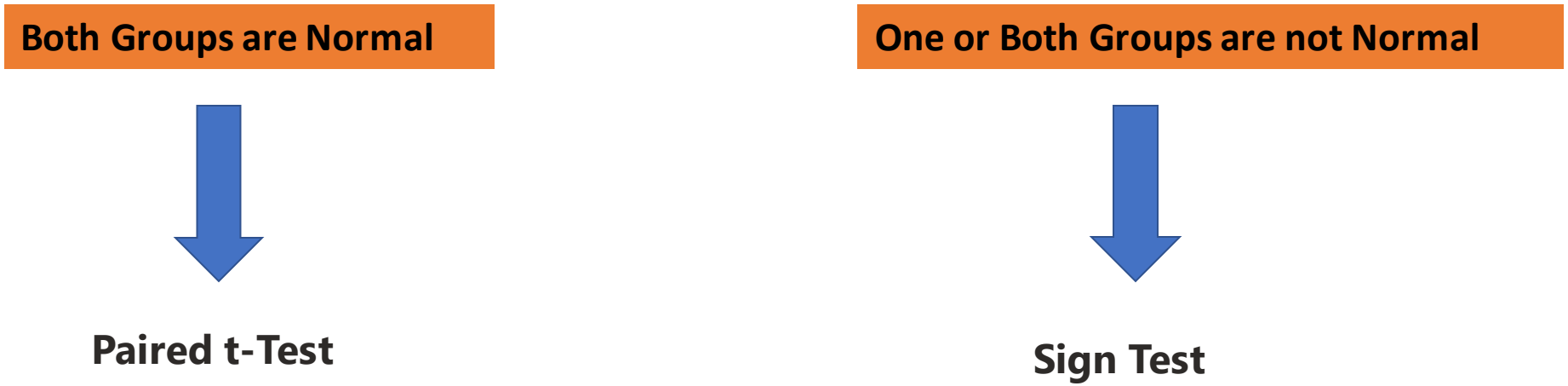
Two Groups

- **So you want to test if the mean of a numeric outcome differs across two groups.**
- ***Question:* Are your measurements paired or independent?**



Independent Groups

- So you want to test if the mean of a numeric outcome differs between two independent groups.
- *Question:* Are the distributions for both groups normal?

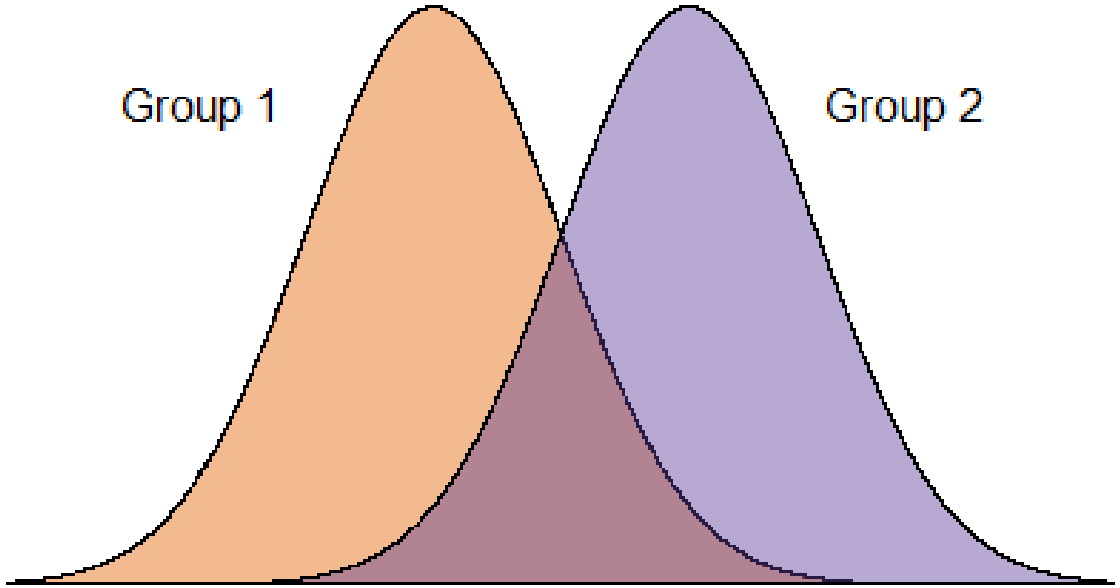


Independent Groups

Normal

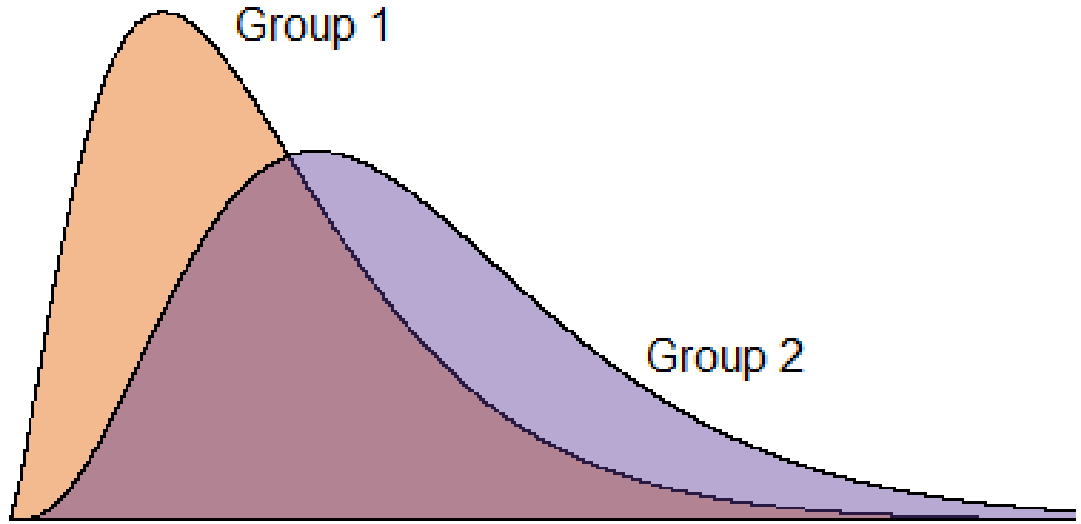
Not Normal

Both Groups are Normal



A normal distribution is the classic bell curve.
It is symmetric, with most cases falling near the mean and fewer cases falling away from the mean in either tail.

One or Both Groups are not Normal



If the variable's distribution is skewed, has more than one mode (peak), or contains outliers, then it will fail the normality assumption of many parametric statistical methods.

Independent Groups

- **So you want to test if the mean of a numeric outcome differs between two independent groups.**
- ***Question:* Are the distributions for both groups normal?**

Both Groups are Normal

Two-Sample Independent t-Test

One or Both Groups are not Normal

Mann-Whitney U-Test

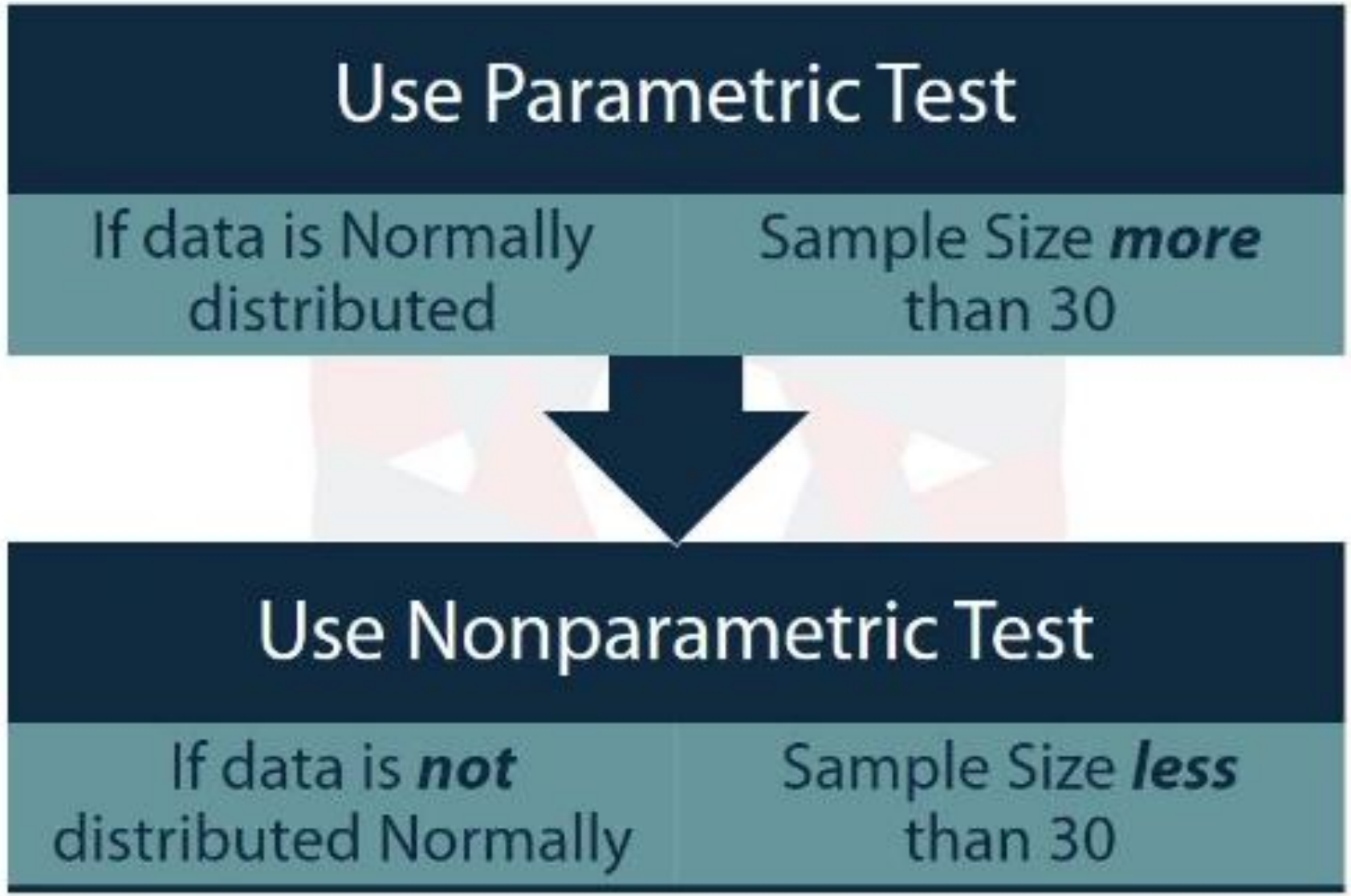
More than two Groups

Kruskal-Wallis Test

- The non-parametric analogue for a one-way ANOVA test is the Kruskal-Wallis test.
- . Remember that a non-parametric test is used when the distribution is either highly skewed or we are comparing ordinal or rank ordered data.
- Example of a skewed distribution 1 2 3 4 5 6

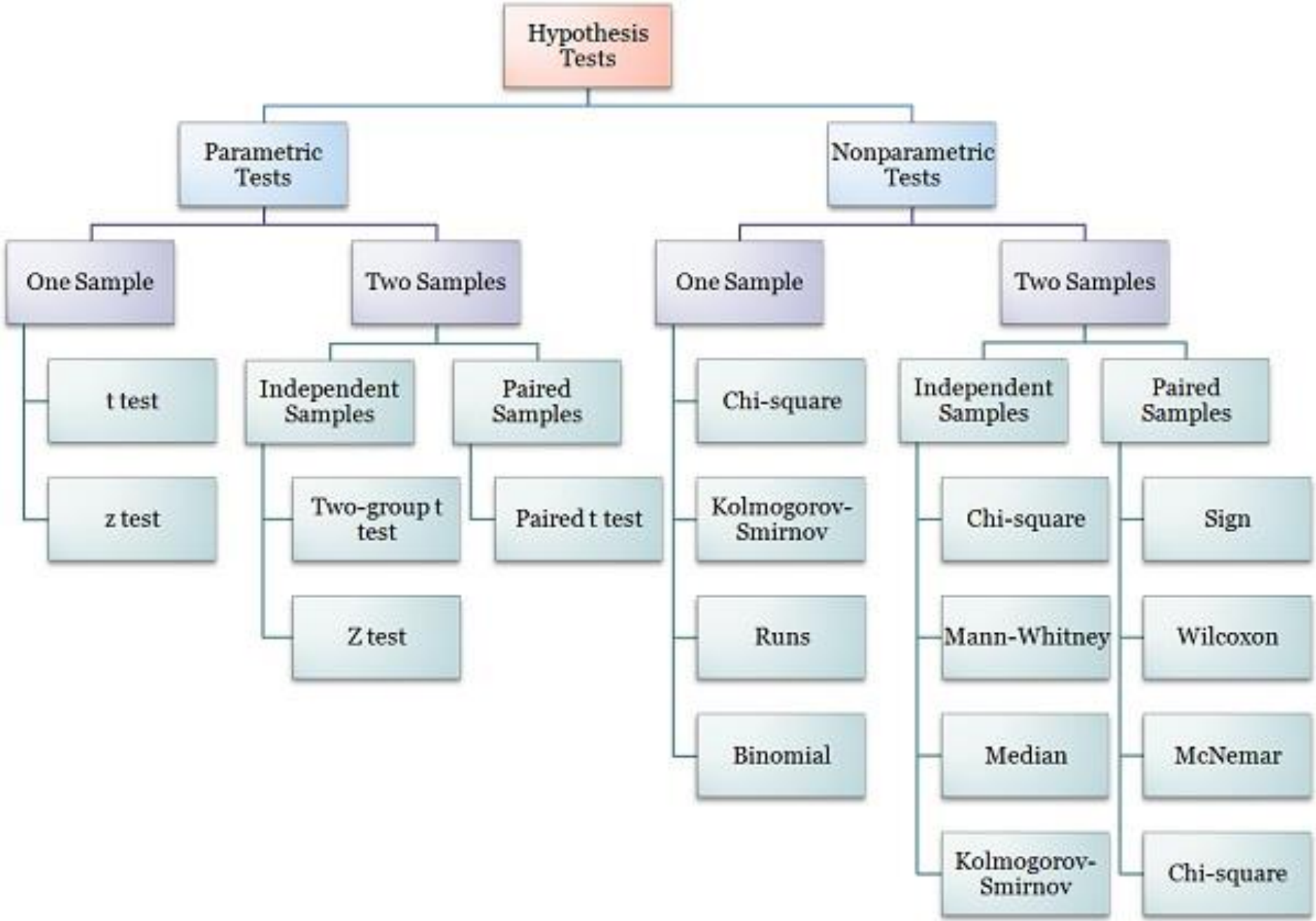
Parametric Tests: When do you use which?

ANOVA	comparing the means of (more than two) samples
Chi-Square Test	testing 'goodness of fit' to an assumed distribution
F-test	comparing variances of two samples
t-test	comparing the mean to a value, or the means of two samples
z-test(u-test)	as t-test but for large samples


Use Parametric Test
If data is Normally distributed
Sample Size more than 30
Use Nonparametric Test
If data is not distributed Normally
Sample Size less than 30

BASIS FOR COMPARISON	PARAMETRIC TEST	NONPARAMETRIC TEST
Meaning	A statistical test, in which specific assumptions are made about the population parameter is known as parametric test.	A statistical test used in the case of non-metric independent variables, is called non-parametric test.
Basis of test statistic	Distribution	Arbitrary
Measurement level	Interval or ratio	Nominal or ordinal
Measure of central tendency	Mean	Median
Information about population	Completely known	Unavailable
Applicability	Variables	Variables and Attributes
Correlation test	Pearson	Spearman

Nonparametric Test	What it does	Parametric Counterpart
Wilcoxon Signed Rank	Compares 1 Median to a specified value	z-test, 1-Sample t-test
	Compares 2 Dependent (Paired) Medians	Paired (Dependent) Samples t-test
Mann-Whitney	Compares 2 Independent Medians	2 (Independent) Samples t-test
Kruskal-Wallis	Compares 3 or more Medians, 1 Variable	1-way ANOVA
Friedman	Compares 3 or more Medians, 2 Variables	2-way ANOVA
Chi-Square Test of Independence	Tests 2 Categorical Variables for Independence (lack of Association)	none


Hypothesis Tests
Parametric Tests
Nonparametric Tests
One Sample
Two Samples
One Sample
Two Samples
t test
z test
Independent Samples
Paired Samples
Two-group t test
Z test
Paired t test
Chi-square
Kolmogorov-Smirnov
Runs
Binomial
Independent Samples
Paired Samples
Chi-square
Mann-Whitney
Median
Kolmogorov-Smirnov
Sign
Wilcoxon
McNemar
Chi-square

Which Non Parametric Tests To Apply When:

While dealing with hypothesis testing, we come across situations where nothing can be assumed about the population distribution, or when the data is not present in representable numerical form (ordinal or nominal data).

In such situations, the basic assumptions for the parametric tests don't hold true and non parametric tests are used.

Nonparametric tests take into account fewer assumptions as compared to parametric tests.

They don't assume anything beforehand about the probability distribution of the population and hence are referred to as distribution-free tests.

They are readily comprehensible and easy to use.

The hypotheses which can be tested using nonparametric tests are:

- Testing whether two independent samples come from identical populations
- Testing whether the samples are drawn from populations having an identical median
- Testing the randomness of one sample or two samples
- To test if a sample comes from specified theoretical distribution

- The Mann-Whitney U test is the nonparametric equivalent of the two sample t-test.
- While the t-test makes an assumption about the distribution of a population (i.e. that the sample came from a t-distributed population), the Mann Whitney U Test makes no such assumption.

Null Hypothesis for the Test:

- The test compares two populations.
- The null hypothesis for the test is that *the probability is 50% that a randomly drawn member of the first population will exceed a member of the second population.*

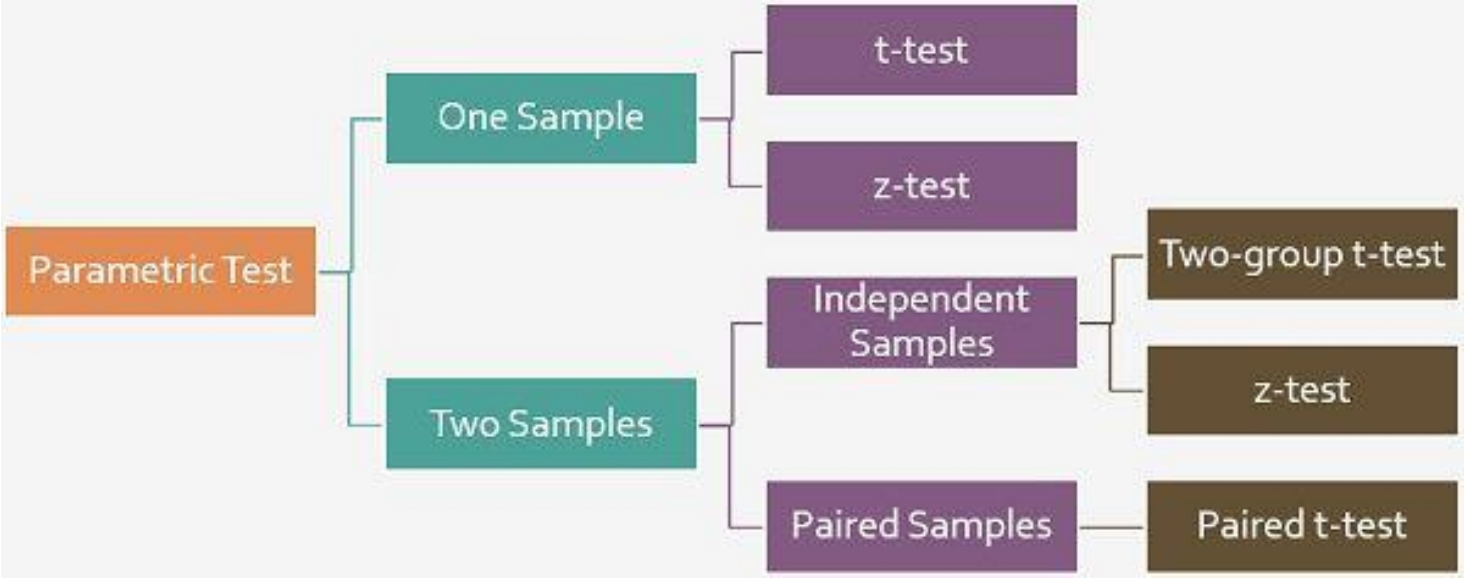
Alternate Hypothesis:

- An alternate null hypothesis is that the two samples come from the same population (i.e. that they both have the same median).

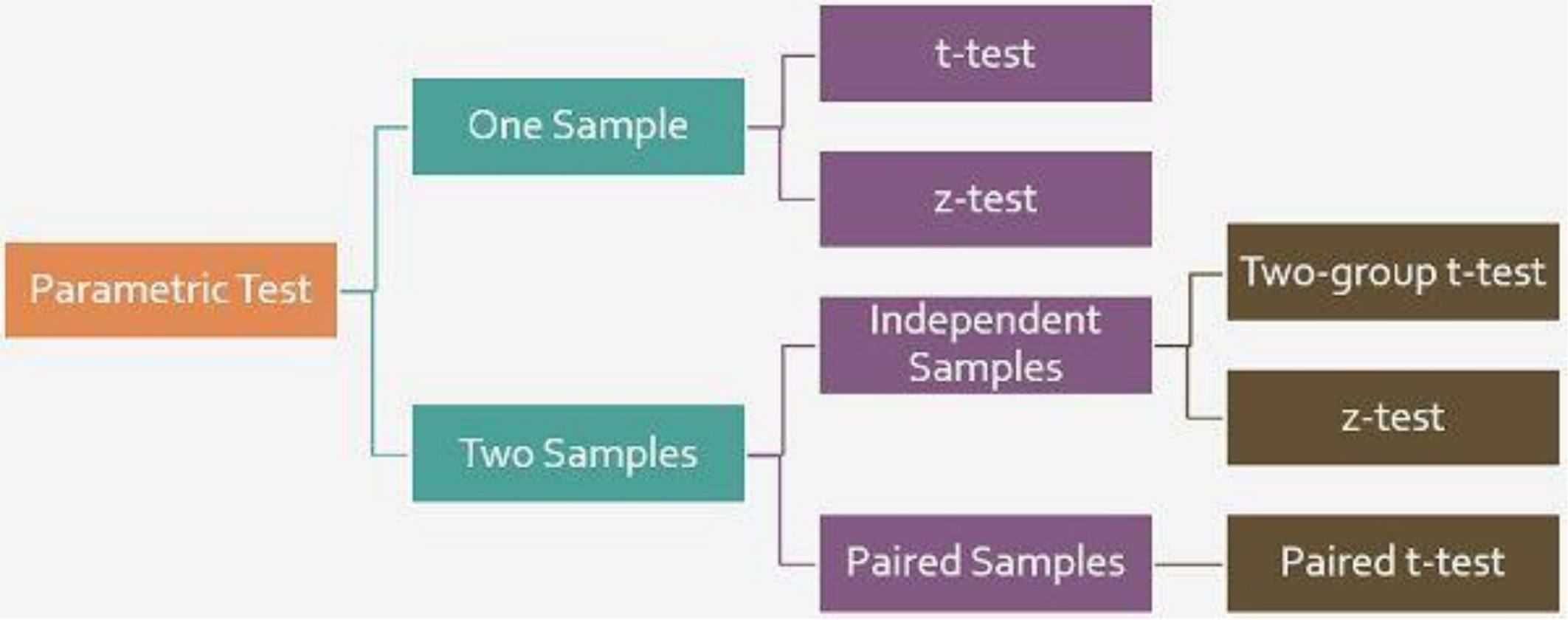
Difference Between T-test and Z-test:

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.

On the other hand, **Z-test** is also a univariate test that is based on standard normal distribution.



Difference Between T-test and Z-test



Comparison Chart

BASIS FOR COMPARISON	T-TEST	Z-TEST
Meaning	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.	Z-test implies a hypothesis test which ascertains if the means of two datasets are different from each other when variance is given.
Based on	Student-t distribution	Normal distribution
Population variance	Unknown	Known
Sample Size	Small	Large

- **Assumptions of T-test**:
- 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.
- Sample values are to be taken and recorded accurately.

Definition of T-test

- A t-test is a hypothesis test used by the researcher to compare population means for a variable, classified into two categories depending on the less-than interval variable.
- More precisely, a t-test is used to examine how the means taken from two independent samples differ.
- 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.

- **Assumptions of T-test**:
- 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.
- Sample values are to be taken and recorded accurately.

The test statistic is:

$$\text{T-test} = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

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

- **Paired t-test**: A statistical test applied when the two samples are dependent and paired observations are taken.

- Definition of Z-test
- Z-test refers to a univariate statistical analysis used to test the hypothesis that proportions from two independent samples differ greatly.
- It determines to what extent a data point is away from its mean of the data set, in standard deviation.
- The researcher adopts z-test, when the population variance is known, in essence, when there is a large sample size, sample variance is deemed to be approximately equal to the population variance. In this way, it is assumed to be known, despite the fact that only sample data is available and so normal test can be applied.

Assumptions of Z-test:

- All sample observations are independent
- Sample size should be more than 30.
- Distribution of Z is normal, with a mean zero and variance 1.
- The test statistic is:
-

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

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

Key Differences Between T-test and Z-test

- The difference between t-test and z-test can be drawn clearly on the following grounds:
- The t-test can be understood as a statistical test which is used to compare and analyse whether the means of the two population is different from one another or not when the standard deviation is not known.
- As against, Z-test is a parametric test, which is applied when the standard deviation is known, to determine, if the means of the two datasets differ from each other.
- The t-test is based on Student's t-distribution. On the contrary, z-test relies on the assumption that the distribution of sample means is normal. Both student's t-distribution and normal distribution appear alike, as both are symmetrical and bell-shaped. However, they differ in the sense that in a t-distribution, there is less space in the centre and more in the tails.
- One of the important conditions for adopting t-test is that population variance is unknown. Conversely, population variance should be known or assumed to be known in case of a z-test.
- Z-test is used to when the sample size is large, i.e. $n > 30$, and t-test is appropriate when the size of the sample is small, in the sense that $n < 30$.

- Conclusion
- By and large, t-test and z-test are almost similar tests, but the conditions for their application is different, meaning that t-test is appropriate when the size of the sample is not more than 30 units. However, if it is more than 30 units, z-test must be performed. Similarly, there are other conditions, which makes it clear that which test is to be performed in a given situation.

Comparison Chart

BASIS FOR COMPARISON	T-TEST	ANOVA
Meaning	T-test is a hypothesis test that is used to compare the means of two populations.	ANOVA is a statistical technique that is used to compare the means of more than two populations.
Test statistic	$(\bar{x}-\mu)/(s/\sqrt{n})$	Between Sample Variance/Within Sample Variance

Key Differences Between T-test and ANOVA

The significant differences between T-test and ANOVA are discussed in detail in the following points:

1. A hypothesis test that is used to compare the means of two populations is called t-test. A statistical technique that is used to compare the means of more than two populations is known as Analysis of Variance or ANOVA.
2. Test Statistic for T-test is:

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

Test Statistic for ANOVA is:

$$F = \frac{s_b^2}{s_w^2}$$

		Truth (unknown)	
		H_0 is True	**H_0 is False**
Test Outcome (our decision)	**Reject H_0**	Error (Type I)	Correct Decision
	Fail to Reject H_0	Correct Decision	Error (Type II)