

Data Analytics

Course Taught at IIFT

Session 3: Sampling Distributions and Hypothesis Testing

Dr. Tanujit Chakraborty

www.ctanujit.org

Today's Topics.....

- Basics of sampling distribution
- Central limit theorem and its Applications
- Standard Sampling distributions
- Principle of Statistical Inference (SI)
- Hypotheses testing procedures
- Case Study 1: Coffee Sale
- Case Study 2: Machine Testing
- Summary

Basic terminologies

Some basic terminology which are closely associated to the above-mentioned tasks are reproduced below.

- **Population:** A **population** consists of the totality of the observation, with which we are concerned.
- **Sample:** A sample is a subset of a population.
- **Random variable:** A random variable is a function that associates a real number with each element in the sample.
- **Statistic:** Any function of the random variable constituting random sample is called a statistic.
- **Statistical inference:** It is an analysis basically concerned with generalization and prediction.

Statistical Inference

There are two facts, which are key to statistical inference.

1. Population parameters are fixed number whose values are usually **unknown**.
 2. Sample statistics are known values for any given sample, but **vary from sample to sample**, even taken from the same population.
- In fact, it is unlikely for any two samples drawn independently, producing identical values of sample **statistics**.
 - In other words, the **variability of sample statistics** is always present and must be accounted for in any inferential procedure.
 - This variability is called **sampling variation**.

Note:

A sample statistic is random variable and like any other random variable, a sample statistic has a probability distribution.

Sampling Distribution

More precisely, sampling distributions are probability distributions and used to describe the variability of sample statistics.

Definition : Sampling distribution

The sampling distribution of a statistics is the probability distribution of that statistic.

- The probability distribution of sample mean (hereafter, will be denoted as \bar{X}) is called the sampling distribution of the mean (also, referred to as the distribution of sample mean).
- Like \bar{X} , we call sampling distribution of variance (denoted as S^2).
- Using the values of \bar{X} and S^2 for different random samples of a population, we are to make inference on the parameters μ and σ^2 (of the population).

Sampling Distribution

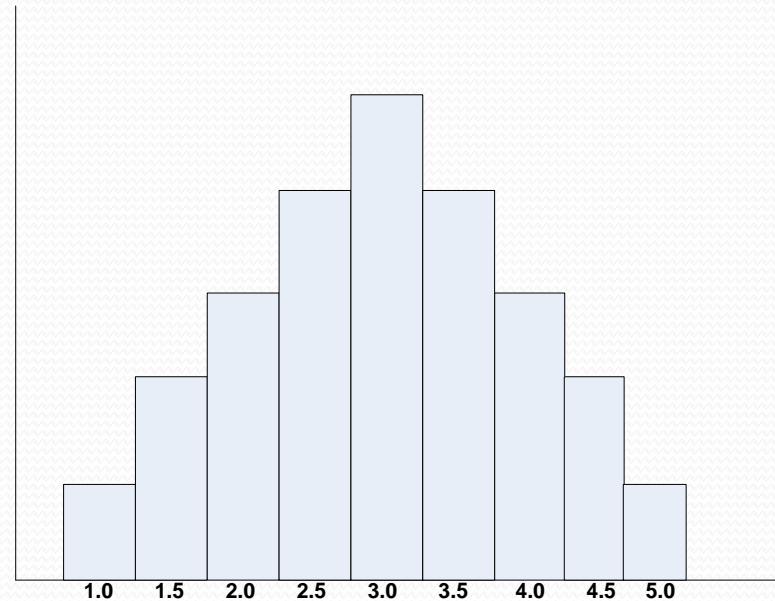
Example 1: Consider five identical balls numbered and weighting as 1, 2, 3, 4 and 5. Consider an experiment consisting of drawing two balls, replacing the first before drawing the second, and then computing the mean of the values of the two balls. Following table lists all possible samples and their mean.

Sample (X)	Mean (\bar{X})	Sample (X)	Mean (\bar{X})	Sample (X)	Mean (\bar{X})
[1,1]	1.0	[2,4]	3.0	[4,2]	3.0
[1,2]	1.5	[2,5]	3.5	[4,3]	3.5
[1,3]	2.0	[3,1]	2.0	[4,4]	4.0
[1,4]	2.5	[3,2]	2.5	[4,5]	4.5
[1,5]	3.0	[3,3]	3.0	[5,1]	3.0
[2,1]	1.5	[3,4]	3.5	[5,2]	3.5
[2,2]	2.0	[3,5]	4.0	[5,3]	4.0
[2,3]	2.5	[4,1]	2.5	[5,4]	4.5
				[5,5]	5.0

Sampling Distribution

Sampling distribution of means

\bar{X}	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0
$f(\bar{X})$	$\frac{1}{25}$	$\frac{2}{25}$	$\frac{3}{25}$	$\frac{4}{25}$	$\frac{5}{25}$	$\frac{4}{25}$	$\frac{3}{25}$	$\frac{2}{25}$	$\frac{1}{25}$



Issues with Sampling Distribution

1. In practical situation, for a large population, it is infeasible to have all possible samples and hence probability distribution of **sample statistics**.
2. The sampling distribution of a statistic depends on
 - the size of the population
 - the size of the samples and
 - the method of choosing the samples.



Theorem on Sampling Distribution

Theorem 1: Sampling distribution of mean and variance

The sampling distribution of a random sample of size n drawn from a population with mean μ and variance σ^2 will have mean $\bar{X} = \mu$ and variance

$$S^2 = \frac{\sigma^2}{n}$$

Example 2: With reference to data in Example 1

For the population, $\mu = \frac{1+2+3+4+5}{5} = 3$

$$\sigma^2 = \frac{(25-1)}{12} = 2$$

Applying the theorem, we have $\bar{X} = 3$ and $S^2 = 1$

Hence, the theorem is verified!

Central Limit Theorem

Theorem 1 is an amazing result and in fact, also verified that if we sampling from a population with unknown distribution, the sampling distribution of \bar{X} will still be approximately normal with mean μ and variance $\frac{\sigma^2}{n}$ **provided that the sample size is large.**

This further, can be established with the famous “central limit theorem”, which is stated below.

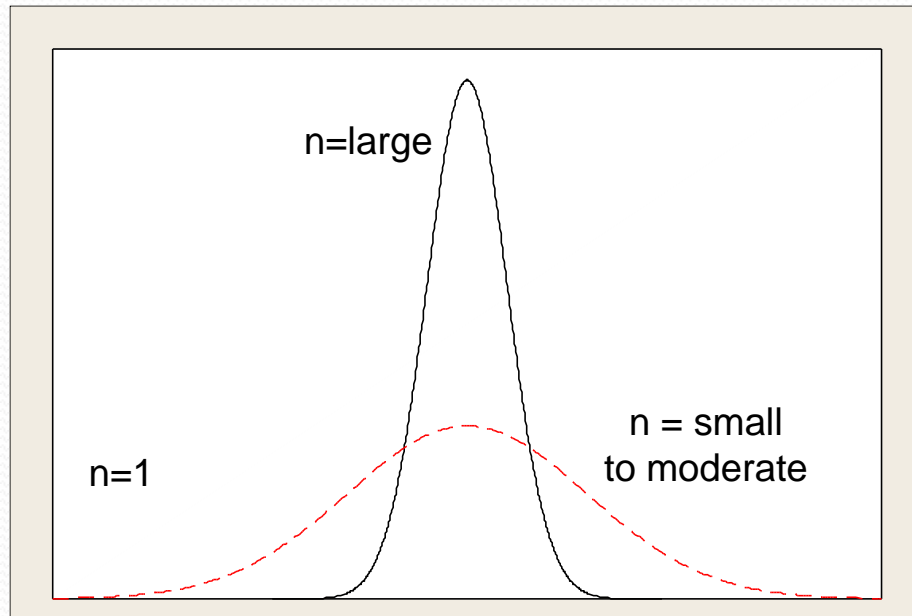
Theorem 2: Central Limit Theorem

If random samples each of size n are taken from any distribution with mean μ and variance σ^2 , the sample mean \bar{X} will have a distribution approximately normal with mean μ and variance $\frac{\sigma^2}{n}$.

The approximation becomes better as n increases.

Applicability of Central Limit Theorem

- The normal approximation of \bar{X} will generally be good if $n \geq 30$
- The sample size $n = 30$ is, hence, a guideline for the central limit theorem.
- The normality on the distribution of \bar{X} becomes more accurate as n grows larger.



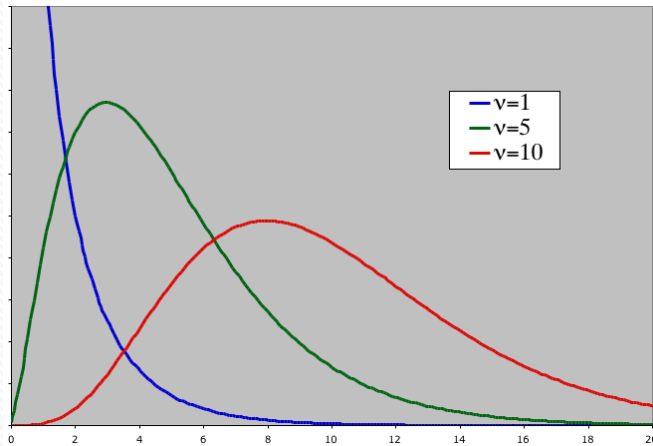
- One very important application of the **Central Limit Theorem** is the determination of reasonable values of the population mean μ and variance σ^2 .
- For standard normal distribution, we have the z-transformation

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

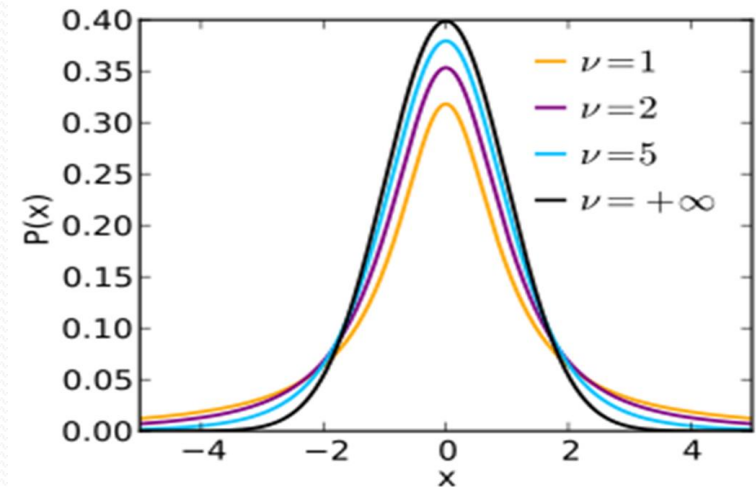
Standard Sampling Distributions

- Apart from the standard normal distribution to describe sampling distribution, there are some other quite different sampling, which are extensively referred in the study of statistical inference.
 - χ^2 : Describes the distribution of variance.
 - t : Describes the distribution of normally distributed random variable standardized by an estimate of the standard deviation.
 - F : Describes the distribution of the ratio of two variables.

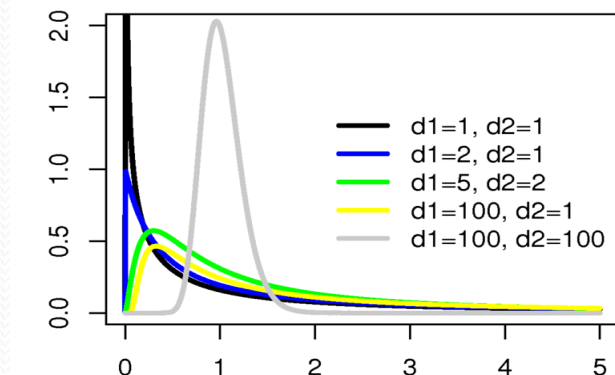
Standard Sampling Distributions



χ^2 - (Chi-Square) distribution curve



t- distribution curve



F - distribution curve

The χ^2 Distribution

A common use of the χ^2 distribution is to describe the distribution of the sample variance.

Definition 1: χ^2 distribution

If x_1, x_2, \dots, x_n are independent random variables having identical normal distribution with mean μ and variance σ^2 , then the random variable

$$Y = \sum_{i=1}^n \left(\frac{x_i - \mu}{\sigma} \right)^2$$

has a Chi squared distribution with $n-1$ degrees of freedom.

Note: To **calculate degrees of freedom**, subtract the number of relations from the number of observations. For **determining the degrees of freedom** for a sample variance, you need to subtract one (1) from the number of observations, n .

The χ^2 Distribution

Note: Each of the n independent random variable $\left(\frac{x_i - \mu}{\sigma}\right)^2, i = 1, 2, 3, \dots, n$ has Chi-squared distribution with 1 degree of freedom.

Now we can derive χ^2 - distribution for sample variance.

We can write

$$\begin{aligned}\sum_{i=1}^n (x_i - \mu)^2 &= \sum_{i=1}^n [(x_i - \bar{x}) + (\bar{x} - \mu)]^2 \\ &= \sum_{i=1}^n (x_i - \bar{x})^2 + n. (\bar{x} - \mu)^2\end{aligned}$$

or,	$\frac{1}{\sigma^2} \sum (x_i - \mu)^2$	$=$	$\frac{(n-1)S^2}{\sigma^2}$	$+$	$\frac{(\bar{x} - \mu)^2}{\sigma^2/n}$
	Chi-square distribution with n-degree		Chi-square distribution with (n-1) degree of freedom		Chi-square distribution with 1 degree of freedom [= Z^2]

The χ^2 Distribution

Definition 2: χ^2 -distribution for Sampling Variance

If S^2 is the variance of a random sample of size n taken from a normal population having the variance σ^2 , then the statistics

$\chi^2 = \frac{(n-1)S^2}{\sigma^2} = \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{\sigma} \right)^2$ has a chi-squared distribution with $v = n - 1$ degrees of freedom.

This way χ^2 - distribution is used to describe the sampling distribution of S^2 .

Chi-Squared Distribution

Definition 3: Chi-squared distribution

The continuous random variable x has a Chi-squared distribution with ν degrees of freedom, is given by

$$f(x: \nu) = \begin{cases} \frac{1}{2^{\frac{\nu}{2}} \Gamma(\nu/2)} x^{\nu/2-1} e^{-\frac{x}{2}}, & x > 0 \\ 0 & \text{Otherwise} \end{cases}$$

where ν is a positive integer.

- The Chi-squared distribution plays an important role in statistical inference .
- The mean and variance of Chi-squared distribution are:

$$\mu = \nu \quad \text{and} \quad \sigma^2 = 2\nu$$

The *t* Distribution

- **The *t* Distribution**

1. To know the sampling distribution of mean we make use of Central Limit Theorem with $Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$
2. This require the **known value of σ** a priori.
3. However, in many situation, σ is certainly no more reasonable than the knowledge of the population mean μ .
4. In such situation, only measure of the standard deviation available may be the sample standard deviation S .
5. It is natural then to substitute S for σ . The problem is that the resulting statistics is not normally distributed!
6. The *t* distribution is to alleviate this problem. This distribution is called ***student's t*** or simply *t – distribution*.

The *t* Distribution

Definition: *t* –distribution

The *t* –distribution with ν degrees of freedom actually takes the form

$$t(\nu) = \frac{Z}{\sqrt{\frac{\chi^2(\nu)}{\nu}}}$$

where Z is a standard normal random variable, and $\chi^2(\nu)$ is χ^2 random variable with ν degrees of freedom.

The t Distribution

Corollary: Let X_1, X_2, \dots, X_n be independent random variables that are all normal with mean μ and standard deviation σ .

$$\text{Let } \bar{X} = \frac{1}{n} \sum_{i=1}^n X_i \text{ and } S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$

Using this definition, we can develop the sampling distribution of the sample mean when the population variance, σ^2 is unknown.

That is,

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \text{ has the standard normal distribution.}$$

$$\chi^2 = \frac{(n-1)S^2}{\sigma^2} \text{ has the } \chi^2 \text{ distribution with } (n-1) \text{ degrees of freedom.}$$

$$\text{Thus, } T = \frac{\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}}{\sqrt{\frac{(n-1)S^2/\sigma^2}{n-1}}} \quad \text{or}$$

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

This is the t – *distribution* with $(n-1)$ degrees of freedom.

The F Distribution

Definition : F distribution

The statistics F is defined to be the ratio of two independent Chi-Squared random variables, each divided by its number of degrees of freedom. Hence,

$$F(v_1, v_2) = \frac{\chi^2(v_1)/v_1}{\chi^2(v_2)/v_2}$$

Corollary : Recall that $\chi^2 = \frac{(n-1)S^2}{\sigma^2}$ is the Chi-squared distribution with $(n - 1)$ degrees of freedom.

Therefore, if we assume that we have sample of size n_1 from a population with variance σ_1^2 and an independent sample of size n_2 from another population with variance σ_2^2 , then the statistics

$$F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2}$$

Note: The F distribution finds enormous applications in comparing sample variances.

Basic Inferential Approaches

Approach 1: Hypothesis testing

- We conduct **test on hypothesis**.
 - We hypothesize that one (or more) parameter(s) has (have) some specific value(s) or relationship.
- Make our decision about the parameter(s) based on one (or more) sample statistic(s)
- Accuracy of the decision is expressed as the probability that the **decision is incorrect**.

Approach 2: Confidence interval measurement

- We estimate one (or more) parameter(s) using sample statistics.
 - This estimation usually done in the form of an interval.
- Accuracy of the decision is expressed as the **level of confidence** we have in the interval.

Hypothesis Testing



Statistical inference



Null hypothesis



Sample



Alternative hypothesis

A **hypothesis** about the value of a population parameter is an **assertion** about its value.

Hypothesis Testing

- A useful analogy for statistical decision making is the decision process used by a jury in a legal proceeding.
- In a criminal trial, the null hypothesis is that the defendant is “not guilty” (“not” in the null hypothesis).
- The alternative is, of course, that the defendant is guilty. The “benefit of the doubt” goes to the null hypothesis (i.e., the evidence must be evaluated by beginning with the assumption that the defendant is not guilty).
- The “burden of proof” is on the prosecution to demonstrate that the assumption of not guilty is incorrect (i.e., evidence must be gathered to support the claim of guilty. The defendant need not prove innocence).
- Jurors must only vote for guilty if they are convinced “beyond reasonable doubt” that the defendant is guilty. Basically, the decision process for a juror goes something like this.

Hypothesis Testing

- Assume the defendant is not guilty. Evaluate the evidence in light of this assumption. Could all the evidence presented simply be a coincidence? (That is, could the degree that the evidence points to guilt of the defendant just be explained as coincidence?)
- What is the chance of that? If that chance is small, then vote for guilty. Otherwise don't vote for guilty. How small is a small chance? Whatever the juror interprets as “beyond reasonable doubt.”
- In statistical hypothesis testing, the principles of the trial carry through quite similarly. The null hypothesis gets the “benefit of the doubt.”
- The “burden of proof” is on the alternative hypothesis. The experimenter must have a test statistic (evidence) which convinces us that the alternative is true rather than the null hypothesis.
- The same question must be asked as we make a statistical decision as that asked above by the juror.

Statistical Hypothesis

- If the hypothesis is stated in terms of population parameters (such as mean and variance), the hypothesis is called **statistical hypothesis**.
- Data from a sample (which may be an experiment) are used to test the validity of the hypothesis.
- A procedure that enables us to agree (or disagree) with the statistical hypothesis is called a **test of the hypothesis**.

Example :

1. To determine whether the wages of men and women are equal.
2. A product in the market is of standard quality.
3. Whether a particular medicine is effective to cure a disease.

The Hypotheses

- The main purpose of statistical hypothesis testing is to choose between two competing hypotheses.

Example : One hypothesis might claim that wages of men and women are equal, while the **alternative** might claim that men make more than women.

- Hypothesis testing start by making a set of two statements about the parameter(s) in question.
- The hypothesis actually to be tested is usually given the symbol H_0 and is commonly referred as the **null hypothesis**.
- The other hypothesis, which is assumed to be true when null hypothesis is false, is referred as the **alternate hypothesis** and is often symbolized by H_1
- The two hypotheses are **exclusive** and **exhaustive**.

The Hypotheses

Example:

Ministry of Human Resource Development (MHRD), Government of India takes an initiative to improve the country's human resources and hence set up **23 IIT's** in the country.

To measure the engineering aptitudes of graduates, MHRD conducts GATE examination for a mark of 1000 in every year. A sample of 300 students who gave GATE examination in 2020 were collected and the mean is observed as 220.

In this context, statistical hypothesis testing is to determine the mean mark of the all GATE-2020 examinee.

The two hypotheses in this context are:

$$H_0: \mu = 220$$

$$H_1: \mu < 220$$

The Hypotheses

Note:

1. As null hypothesis, we could choose $H_0: \mu \leq 220$ or $H_0: \mu \geq 220$
2. It is customary to always have the null hypothesis with an equal sign.
3. As an alternative hypothesis there are many options available with us.

Examples:

- I. $H_1: \mu > 220$
 - II. $H_1: \mu < 220$
 - III. $H_1: \mu \neq 220$
-
4. The two hypothesis should be chosen in such a way that they are **exclusive** and **exhaustive**.
 - One or other must be true, but they cannot both be true.

The Hypotheses

One-tailed test

- A statistical test in which the alternative hypothesis specifies that the population parameter lies entirely above or below the value specified in H_0 is called a one-sided (or one-tailed) test.

Example.

$$H_0: \mu = 100$$

$$H_1: \mu > 100$$

Two-tailed test

- An alternative hypothesis that specifies that the parameter can lie on their sides of the value specified by H_0 is called a two-sided (or two-tailed) test.

Example.

$$H_0: \mu = 100$$

$$H_1: \mu \neq 100$$

Hypothesis Testing Procedures

The following **five steps** are followed when testing hypothesis

1. Specify H_0 and H_1 , the null and alternate hypothesis, and an **acceptable level of α** .
2. Determine an appropriate sample-based test statistics and the **rejection region** for the specified H_0 .
3. Collect the sample data and calculate the test statistics.
4. Make a decision to either reject or fail to reject H_0 .
5. Interpret the result in common language suitable for practitioners.

Hypothesis Testing Procedure

- In summary, we have to choose between H_0 and H_1
- The standard procedure is to assume H_0 is true.
(Just we presume innocent until proven guilty)
- Using statistical test, we try to determine whether there is sufficient evidence to declare H_0 false.
- We reject H_0 only when the **chance is small** that H_0 is true.
- The procedure is based on probability theory, that is, there is a chance that we can **make errors**.

Errors in Hypothesis Testing

In hypothesis testing, there are two types of errors.

Type I error: A type I error occurs when we incorrectly reject H_0 (i.e., we reject the null hypothesis, when H_0 is true).

Type II error: A type II error occurs when we incorrectly fail to reject H_0 (i.e., we accept H_0 when it is not true).

Decision	Observation	
	H_0 is true	H_0 is false
H_0 is accepted	Decision is correct	Type II error
H_0 is rejected	Type I error	Decision is correct

Probabilities of Making Errors

Type I error calculation

α : denotes the probability of making a Type I error

$$\alpha = \mathbf{P}(\text{Rejecting } H_0 | H_0 \text{ is true})$$

Type II error calculation

β : denotes the probability of making a Type II error

$$\beta = \mathbf{P}(\text{Accepting } H_0 | H_0 \text{ is false})$$

Note:

- α and β are not independent of each other as one increases, the other decreases
- When the sample size increases, both to decrease since sampling error is reduced.
- In general, we focus on Type I error, but Type II error is also important, particularly when sample size is small.

Calculating α

Assuming that we have the results of random sample. Hence, we use the characteristics of sampling distribution to calculate the probabilities of making either Type I or Type II error.

Example :

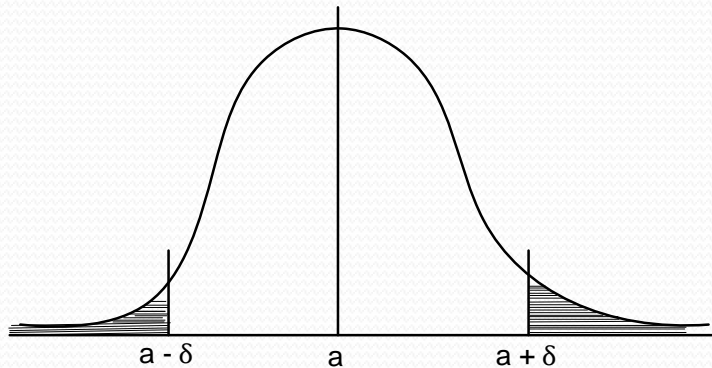
Suppose, two hypotheses in a statistical testing are:

$$H_0: \mu = a$$

$$H_1: \mu \neq a$$

Also, assume that for a given sample, population obeys normal distribution. A threshold limit say $a \pm \delta$ is used to say that **they are significantly different from a**.

Calculating α



Here, shaded region implies the probability that, $\bar{X} < a - \delta$ or $\bar{X} > a + \delta$

Thus the null hypothesis is to be rejected if the mean value is less than $a - \delta$ or greater than $a + \delta$.

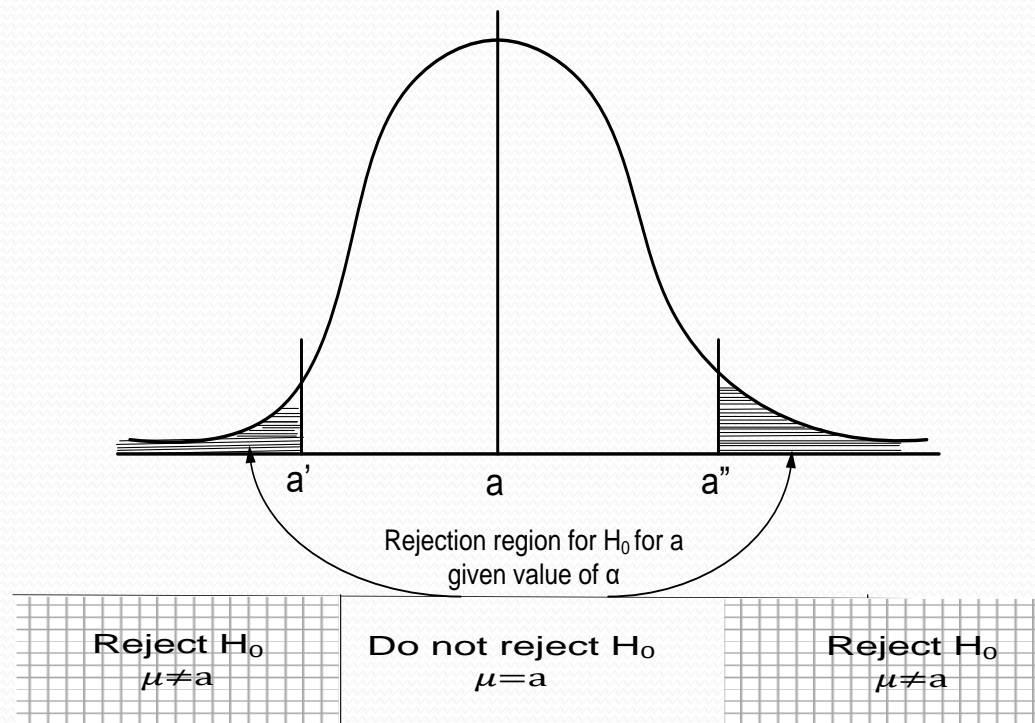
If \bar{X} denotes the sample mean, then the Type I error is

$$\alpha = P(\bar{X} < a - \delta \text{ or } \bar{X} > a + \delta, \quad \text{when } \mu = a, \quad \text{i.e., } H_0 \text{ is true})$$

The Rejection Region

The rejection region comprises of value of the test statistics for which

1. The probability when the null hypothesis is true is less than or equal to the specified α .
2. Probability when H_1 is true are greater than they are under H_0 .



Two-Tailed Test

For two-tailed hypothesis test, hypotheses take the form

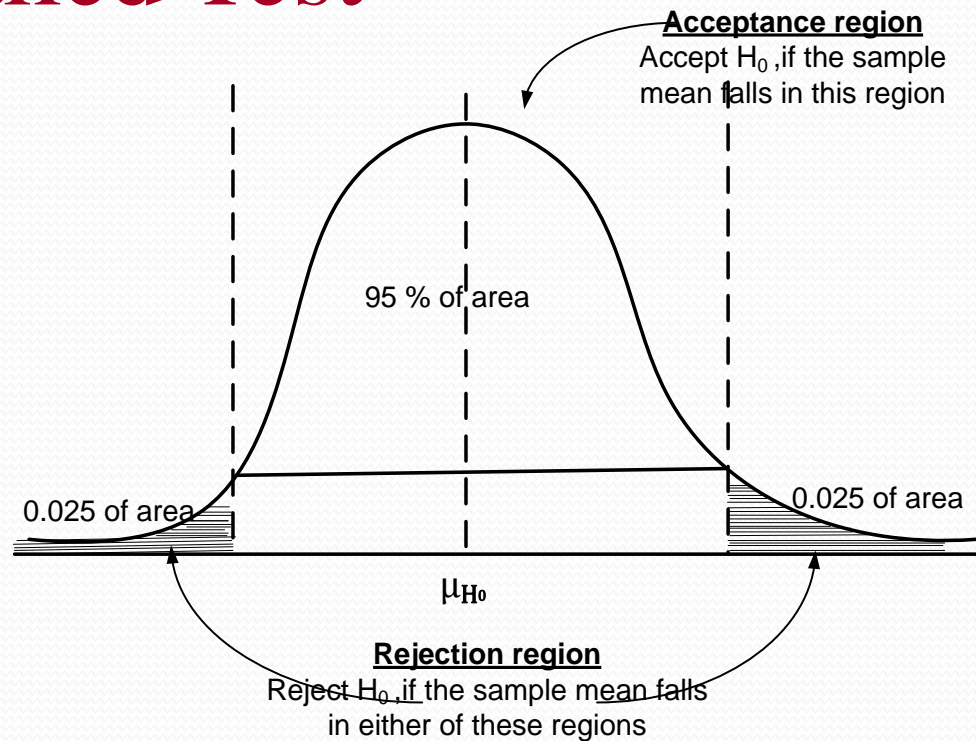
$$H_0: \mu = \mu_{H_0}$$

$$H_1: \mu \neq \mu_{H_0}$$

In other words, to reject a null hypothesis, sample mean $\mu > \mu_{H_0}$ or $\mu < \mu_{H_0}$ under a given α .

Thus, in a two-tailed test, there are two rejection regions (also known as critical region), one on each tail of the sampling distribution curve.

Two-Tailed Test



Acceptance and rejection regions in case of a two-tailed test with 5% significance level.

One-Tailed Test

A one-tailed test would be used when we are to test, say, whether the population mean is either lower or higher than the hypothesis test value.

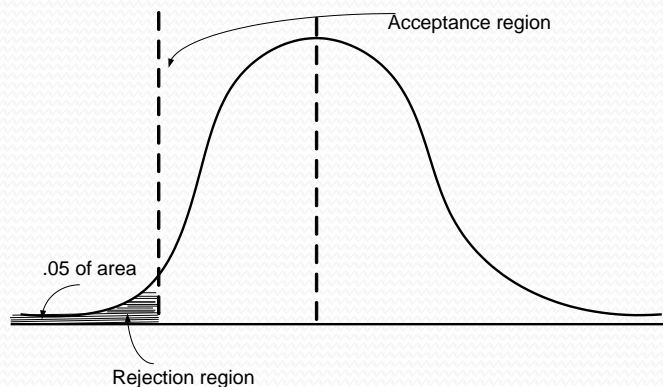
Symbolically,

$$H_0: \mu = \mu_{H_0}$$

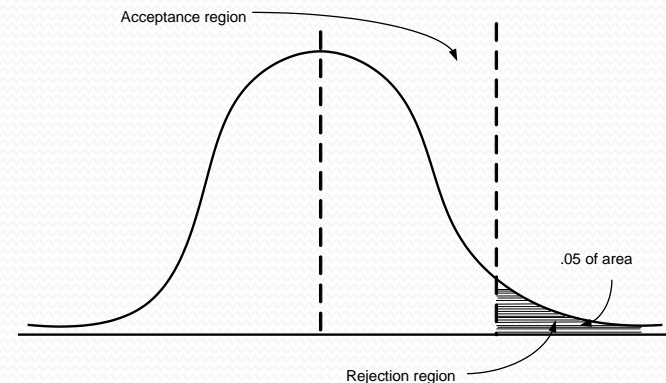
$$H_1: \mu < \mu_{H_0}$$

$$[or \mu > \mu_{H_0}]$$

Wherein there is one rejection region only on the left-tail (or right-tail).



Left – tailed test



Right – tailed test

Example : Calculating α

Consider the two hypotheses are

The null hypothesis is

$$H_0: \mu = 8$$

The alternative hypothesis is

$$H_1: \mu \neq 8$$

Assume that given a sample of size 16 and standard deviation is 0.2 and sample follows normal distribution.

Example : Calculating α

We can decide the rejection region as follows.

Suppose, the null hypothesis is to be rejected if the mean value is less than 7.9 or greater than 8.1. If \bar{X} is the sample mean, then the probability of Type I error is

$$\alpha = P(\bar{X} < 7.9 \text{ or } \bar{X} > 8.1, \text{ when } \mu = 8)$$

Given σ , the standard deviation of the sample is 0.2 and that the distribution follows **normal distribution**.

Thus,

$$P(\bar{X} < 7.9) = P\left[Z = \frac{7.9 - 8}{0.2/\sqrt{16}}\right] = P[Z < -2.0] = 0.0228$$

and

$$P(\bar{X} > 8.1) = P\left[Z = \frac{8.1 - 8}{0.2/\sqrt{16}}\right] = P[Z > 2.0] = 0.0228$$

Hence, $\alpha = 0.0228 + 0.0228 = 0.0456$

Example : Calculating α and β

There are two identically appearing boxes of chocolates. Box A contains 60 red and 40 black chocolates whereas box B contains 40 red and 60 black chocolates. There is no label on the either box. One box is placed on the table. We are to test the hypothesis that “Box B is on the table”.

To test the hypothesis an experiment is planned, which is as follows:

- Draw at random five chocolates from the box.
- We replace each chocolates before selecting a new one.
- The number of red chocolates in an experiment is considered as the **sample statistics**.

Note: Since each draw is independent to each other, we can assume the sample distribution follows binomial probability distribution.

Example : Calculating α

Let us express the population parameter as p = the number of red chocolates in Box B.

The hypotheses of the problem can be stated as:

$$H_0: p = 0.4 \quad // \text{ Box B is on the table}$$

$$H_1: p = 0.6 \quad // \text{ Box A is on the table}$$

Calculating α :

In this example, the null hypothesis (H_0) specifies that the probability of drawing a red chocolate is 0.4. This means that, lower proportion of red chocolates in observations (*i.e., sample*) favors the null hypothesis. In other words, **drawing all red chocolates** provides **sufficient evidence to reject the null hypothesis**. Then, the probability of making a *Type I* error is the probability of getting five red chocolates in a sample of five from Box B. That is,

$$\alpha = P(X = 5 \text{ when } p = 0.4)$$

Using the binomial distribution

$$\begin{aligned} &= \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x} \text{ where } n = 5, x = 5 \\ &= (0.4)^5 = 0.01024 \end{aligned}$$

Thus, the probability of rejecting a true null hypothesis is ≈ 0.01 . That is, there is approximately 1 in 100 chance that the box B will be mislabeled as box A.

Example : Calculating β

The *Type II* error occurs if we fail to reject the null hypothesis when it is not true. For the current illustration, such a situation occurs, if Box A is on the table but we did not get the five red chocolates required to reject the hypothesis that Box B is on the table.

The probability of *Type II* error is then the probability of getting four or fewer red chocolates in a sample of five from Box A.

That is,

$$\beta = P(X \leq 4) \quad \text{when } p = 0.6$$

Using the probability rule:

$$P(X \leq 4) + P(X = 5) = 1$$

$$\text{That is, } P(X \leq 4) = 1 - P(X = 5)$$

$$\text{Now, } P(X = 5) = (0.6)^5$$

$$\begin{aligned} \text{Hence, } \beta &= 1 - (0.6)^5 \\ &= 1 - 0.07776 = 0.92224 \end{aligned}$$

That is, the probability of making *Type II* error is over 92%. This means that, if Box A is on the table, the probability that we will be unable to detect it is 0.92.

Case Study 1: Coffee Sale

A coffee vendor nearby Howrah railway station has been having average sales of 500 cups per day. Because of the development of a bus stand nearby, it expects to increase its sales. During the first 12 days, after the inauguration of the bus stand, the daily sales were as under:

550 570 490 615 505 580 570 460 600 580 530 526

On the basis of this sample information, can we conclude that the sales of coffee have increased?

Consider 5% level of confidence.



Hypothesis Testing : 5 Steps

The following **five steps** are followed when testing hypothesis

1. Specify H_0 and H_1 , the null and alternate hypothesis, and an **acceptable level of α** .
2. Determine an appropriate sample-based test statistics and the **rejection region** for the specified H_0 .
3. Collect the sample data and calculate the test statistics.
4. Make a decision to either reject or fail to reject H_0 .
5. Interpret the result in common language suitable for practitioner.

Case Study 1: Step 1

Step 1: Specification of hypothesis and acceptable level of α

Let us consider the hypotheses for the given problem as follows.

$$H_0: \mu = 500 \text{ cups per day}$$

The null hypothesis that sales average 500 cups per day and they have not increased.

$$H_1: \mu > 500$$

The alternative hypothesis is that the sales have increased.

Given the acceptance level of $\alpha = 0.05$ (*i. e.*, 5% level of significance)

Case Study 1: Step 2

Step 2: Sample-based test statistics and the rejection region for specified H_0

Given the sample as

550 570 490 615 505 580 570 460 580 530 526

Since the sample size is small and the population standard deviation is not known, we shall use t – $test$ assuming normal population. The test statistics t is

$$t = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

To find \bar{X} and S , we make the following computations.

$$\bar{X} = \frac{\sum X_i}{n} = \frac{6576}{12} = 548$$

Case Study 1: Step 2

<i>Sample #</i>	X_i	$X_i - \bar{X}$	$(X_i - \bar{X})^2$
1	550	2	4
2	570	22	484
3	490	-58	3364
4	615	67	4489
5	505	-43	1849
6	580	32	1024
7	570	22	484
8	460	-88	7744
9	600	52	2704
10	580	32	1024
11	530	-18	324
12	526	-22	484
$n = 12$	$\sum X_i = 6576$		$\sum (X_i - \bar{X})^2 = 23978$

Case Study 1: Step 2

$$S = \sqrt{\frac{\sum (X_i - \bar{X})^2}{n - 1}} = \sqrt{\frac{23978}{12 - 1}} = 46.68$$

$$\text{Hence, } t = \frac{\bar{X} - \mu}{S/\sqrt{n}} = \frac{48}{46.68/\sqrt{12}} = \frac{48}{13.49} = 3.558$$

Note:

Statistical table for t-distributions gives a t -value given n , the degrees of freedom and α , the level of significance and vice-versa.

Case Study 1: Step 3

Step 3: Collect the sample data and calculate the test statistics

$$\text{Degree of freedom} = n - 1 = 12 - 1 = 11$$

As H_1 is one-tailed, we shall determine the rejection region applying one-tailed in the right tail because H_1 is more than type) at 5% level of significance.

Using table of t – *distribution* for 11 degrees of freedom and with 5% level of significance,

$$R: t > 1.796$$

Case Study 1: Step 4

Step 4: Make a decision to either reject or fail to reject H_0

The observed value of $t = 3.558$ which is in the rejection region and thus H_0 is rejected at 5% level of significance.

Case Study 1: Step 5

Step 5: Final comment and interpret the result

We can conclude that the sample data indicate that coffee sales have increased.

Case Study 2: Machine Testing

A medicine production company packages medicine in a tube of 8 ml. In maintaining the control of the amount of medicine in tubes, they use a machine. To monitor this control a sample of 16 tubes is taken from the production line at random time interval and their contents are measured precisely. The mean amount of medicine in these 16 tubes will be used to test the hypothesis that the machine is indeed working properly.



Case Study 2: Step 1

Step 1: Specification of hypothesis and acceptable level of α

The hypotheses are given in terms of the population mean of medicine per tube.

The null hypothesis is

$$H_0: \mu = 8$$

The alternative hypothesis is

$$H_1: \mu \neq 8$$

We assume α , the significance level in our hypothesis testing ≈ 0.05 .

(This signifies the probability that the machine needs to be adjusted less than 5%).

Case Study 2: Step 2

Step 2: Sample-based test statistics and the rejection region for specified H_0

Rejection region: Given $\alpha = 0.05$, which gives $|Z| > 1.96$ (obtained from standard normal calculation for $n(Z: 0,1) = 0.025$ for a rejection region with two-tailed test).

Case Study 2: Step 3

Step 3: Collect the sample data and calculate the test statistics

Sample results: $n = 16$, $\bar{x} = 7.89$, $\sigma = 0.2$

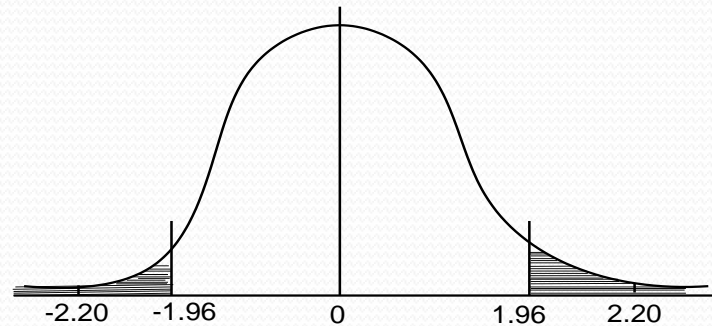
With the sample, the test statistics is

$$Z = \frac{7.89 - 8}{0.2 / \sqrt{16}} = -2.20$$

Hence, $|Z| = 2.20$

Case Study 2: Step 4

Step 4: Make a decision to either reject or fail to reject H_0



Since $Z > 1.96$, we reject H_0

Case Study 2: Step 5

Step 5: Final comment and interpret the result

We conclude $\mu \neq 8$ and recommend that the machine be adjusted.

Case Study 2: Alternative Test

Suppose that in our initial setup of hypothesis test, if we choose $\alpha = 0.01$ instead of 0.05, then the test can be summarized as:

1. $H_0: \mu = 8, H_1: \mu \neq 8 \quad \alpha = 0.01$
2. Reject H_0 if $Z > 2.576$
3. Sample result $n=16, \sigma = 0.2, \bar{X}=7.89, Z = \frac{7.89-8}{0.2/\sqrt{16}} = -2.20, |Z| = 2.20$
4. $|Z| < 2.20$, we fail to reject $H_0=8$
5. We do not recommend that the machine be readjusted.

Hypothesis Testing Strategies

- The hypothesis testing determines the validity of an assumption (technically described as null hypothesis), with a view to choose between two conflicting hypothesis about the value of a **population** parameter.
- There are two types of tests of hypotheses
 - ✓ Non-parametric tests (also called distribution-free test of hypotheses)
 - ✓ Parametric tests (also called standard test of hypotheses).

Parametric Tests : Applications

- Usually assume certain properties of the population from which we draw samples.
 - Observation come from a normal population
 - Sample size is small
 - Population parameters like mean, variance, etc. are hold good.
 - Requires measurement equivalent to interval scaled data.

Important Parametric Tests

The widely used sampling distribution for parametric tests are

- $Z - test$
- $t - test$
- $\chi^2 - test$
- $F - test$

Parametric Tests

Note:

All these tests are based on the assumption of normality (i.e., the source of data is considered to be normally distributed).

Z – test: This is most frequently test in statistical analysis.

- It is based on the normal probability distribution.
- Used for judging the significance of several statistical measures particularly the mean.
- It is used even when *binomial distribution* or *t – distribution* is applicable with a condition that such a distribution tends to normal distribution when n becomes large.
- Typically it is used for comparing the mean of a sample to some hypothesized mean for the population in case of large sample, or when **population variance** is known.

Parametric Tests : t-test

t – test: It is based on the t-distribution.

- It is considered an appropriate test for judging the significance of a sample mean or for judging the significance of difference between the means of two samples in case of
 - small sample(s)
 - **population variance is not known** (in this case, we use the variance of the sample as an estimate of the population variance)

Parametric Tests : χ^2 -test

χ^2 – *test*: It is based on Chi-squared distribution.

- It is used for comparing a sample variance to a theoretical population variance.

Parametric Tests : F -test

F – *test*: It is based on F-distribution.

- It is used to compare the variance of two independent samples.
- This test is also used in the context of analysis of variance (ANOVA) for judging the significance of more than two sample means.

Hypothesis Testing : Assumptions

Case 1: Normal population, population infinite, sample size may be large or small, variance of the population is known.

$$z = \frac{\bar{X} - \mu_{H_0}}{\sigma/\sqrt{n}}$$

Case 2: Population normal, population **finite**, sample size may large or small.....variance is known.

$$z = \frac{\bar{X} - \mu_{H_0}}{\sigma/\sqrt{n}[\sqrt{(N-n)/(N-1)}]}$$

Case 3: Population normal, population infinite, **sample size is small** and variance of the **population is unknown**.

$$t = \frac{\bar{X} - \mu_{H_0}}{s/\sqrt{n}} \quad \text{with degree of freedom} = (n - 1)$$

and

$$s = \sqrt{\frac{\sum(X_i - \bar{X})^2}{(n-1)}}$$

Case 4: Population finite

$$t = \frac{\bar{X} - \mu_{H_0}}{\sigma/\sqrt{n}[\sqrt{(N-n)/(N-1)}]} \quad \text{with degree of freedom} = (n - 1)$$

Note: If variance of population (σ) is known, replace S by σ . Population normal, population infinite, **sample size is small** and variance of the **population is unknown**.

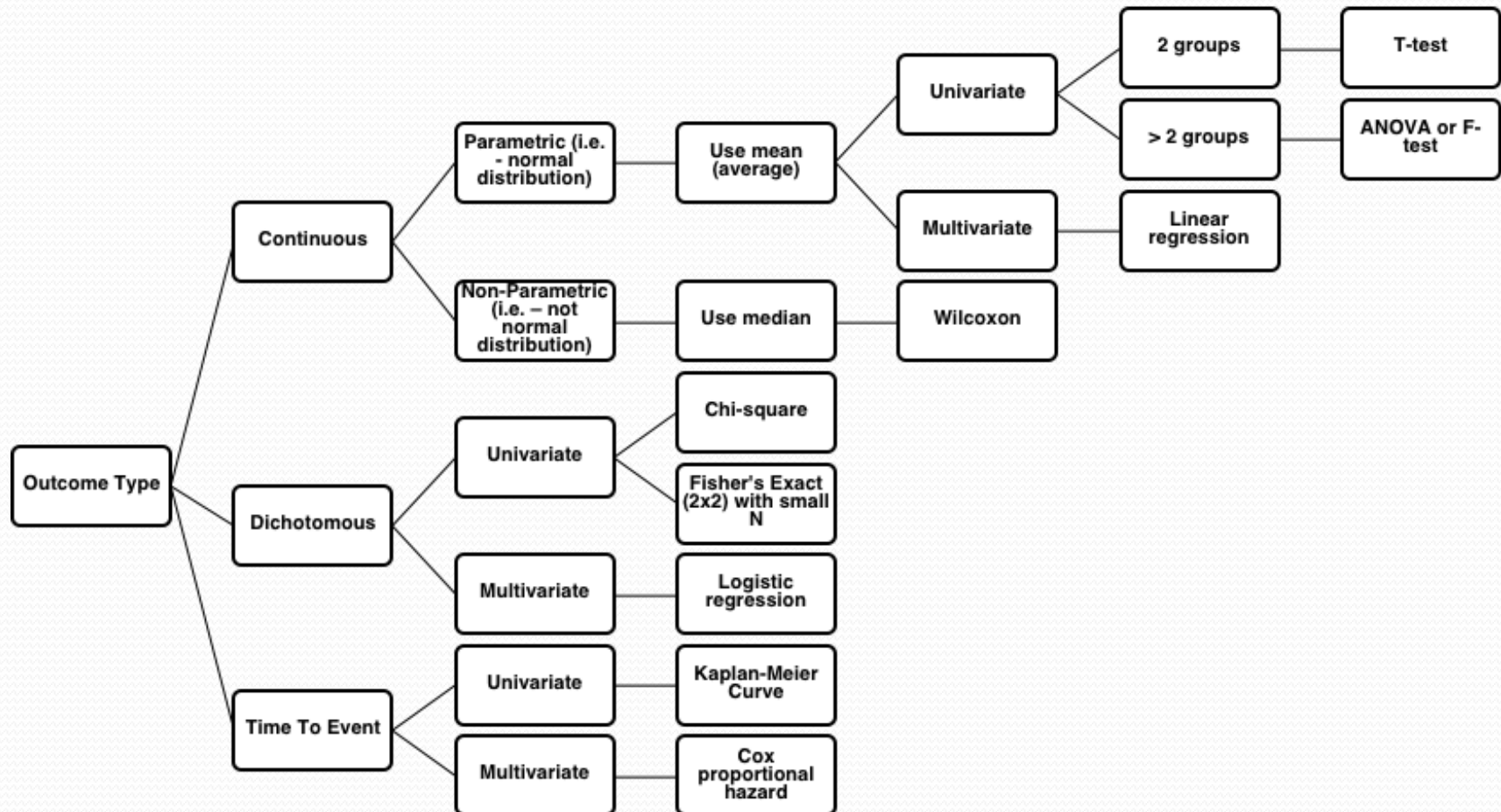
Hypothesis Testing : Non-Parametric Test

- *Non-Parametric tests*
 - ✓ Does not under any assumption
 - ✓ Assumes only nominal or ordinal data

Note: Non-parametric tests need entire population (or very large sample size)

Cheatsheet :

STATISTICAL TESTS CHEAT SHEET





Any question?

You may also send your question(s) at ctanujit@gmail.com