STATISTICAL ESTIMATION THEORY

Population: A collection of all well-defined objects under study is called population.

Example: Suppose we want to study the economic conditions of primary teachers in Maharashtra, then the group of all primary teachers in the state of Maharashtra is a population.

Sample: A well defined finite subset of the population is called a sample.

Example: Suppose we want to study the economic conditions of primary teachers in the state of Maharashtra, then the few primary teachers (set of few teachers) in the state of Maharashtra forms a sample.

Parameter: An unknown constant of a population that summarises or describes an aspect of the population (such as a mean or a standard deviation) is called parameter. Let *f (x,* θ*)* be the pdf of a random variable “X‟ having an unknown constant θ.

Statistic: Any function of a sample value (observed value) is called a statistic. The sample statistic is constants but it differ from sample to sample.

Sampling distribution: The probability distribution of the sample statistic is called a sampling distribution.

Parameter space: The set of all admissible values of a parameter of the distribution is called parameter space. It is denoted by Θ.

Example: $X \sim$ Normal $(\mu, \sigma\ 2\ )$

$\Theta = \{(\mu, \sigma\ 2\ ) / -\infty<\mu 0\}$

Estimator: Let $x_1, x_2,\ x_2,\ .....\ x_n$ be a sample of size n taken from a distribution having pdf f (x, θ) where $\theta \in \Theta$ is an unknown parameter. A function $T = T\ (x_1, x_2,\ x_2,\ .....\ x_n\ )$ which maps sample space (S) to parameter space Θ is called an estimator. In other words, If a statistic $T = T(x_1, x_2,\ x_2,\ .....\ x_n)$ is used to estimate θ, and its value belongs to parameter space then it is said to be an estimator of θ.

Estimate: A particular value of an estimator corresponding to the given sample values is called an estimate of the population parameter.

Requirements of good and reliable estimators:

1. Unbiasedness: An estimator is called unbiased if it is expected to draw. Such a value from the sample which is equal to the population parameter being estimated.

2. Consistency: An estimator with smaller variance is called efficient.

3. Efficiency : An estimator is said to be consistent if its expected value gets closer and closer to the parameter being estimated as the sample size increases.

4. Sufficiency: An estimator is said to be sufficient, if it conveys all information the sample can furnish for the estimation of the parameter being estimated.

Standard Error :

In sampling we are considering measure for both population and sample at the same time. The standard deviation of the statistic is termed as S*tandard Error* (SE) just to differentiate it from the standard deviation of the population. Also, we know that the statistic value changes from sample to sample. So the sampling distribution measures these deviations, say of the sample means from the mean of all samples. Such an error is called *sampling error* or *standard error*. Students should remember that SE is nothing but the standard deviation of the sample taken for investigation.

Central limit Theorem :

This is a very important theorem in sampling distribution theory. In practice not all population show a normal distribution. The *Central limit theorem* state that "When a random sample is drawn from a population which is not normally distributed then as the sample size is increased, the standard deviation of the sample mean is approximately normally distributed with mean equal to the mean of the population and standard deviation equal to $\sigma/\sqrt{n}$, where σ is the standard deviation of the population".

If the size of the sample is more than 30 it is said to be large, in general.

Types of Hypothesis

Null Hypothesis

The hypothesis which is to be tested is called *Null Hypothesis*. It is denoted by H_0. The null hypothesis assumes that there is no difference between the statistic and the parameter. In case of the measure being the mean, if the mean of the sample is μ_0 and the population mean is μ then we write the null hypothesis as follows:

H_0: $\mu = \mu_0$ or H_0: $\mu > \mu_0$ or H_0: $\mu < \mu_0$

The latter two types will be discussed a little later in *One Tailed Tests.*

Alternative Hypothesis

The alternate (or opposite) of the null hypothesis is called *Alternative Hypothesis*. It is denoted by H_1.If H_0 is true then H_1 is false and vice versa.

1. If H_0: $\mu = \mu_0$, then H_1: $\mu \neq \mu_0$
2. If H_0: $\mu > \mu_0$, then H_1: $\mu \leq \mu_0$
3. If H_0: $\mu < \mu_0$, then H_1: $\mu \geq \mu_0$

Errors in making a decision

Consider a case when a college has to give a scholarship to all students whose performance has increased in the second term compared with the first term. If the college administration assumes that there is no difference in the performance (null hypothesis) then what are the decisions taken for a particular student selected at random? There are four possibilities:

1. The student's performance has improved and is given scholarship: Correct decision!
2. The student's performance has not improved but is given a scholarship. Good Luck!
3. The student's performance has not improved and is not given the scholarship. Correct decision.
4. The student's performance has not improved but is given the scholarship. Bad decision!

The remarks given are mine and need not be compared with any statistical conclusions. But roughly we come across some mistakes or errors in making decision. The decision number 2 and 4 are the wrong decisions or errors. In testing a hypothesis these errors are named as Type I and Type II error.

Type I Error

The error made when the null hypothesis is rejected when it is true is called *Type I error*.

Level Of Significance:

The probability that a Type I error is made is called as *level of significance*. It is denoted by α. It is the risk taken by the decision making body of making a Type I error. In majority of cases we assume a 5% level of significance for testing a hypothesis. This means that the probability rejecting the null hypothesis when it is true is only 0.05. In other words, the probability of making a correct decision is 95%. There are situations where more accuracy or less accuracy is required. In such cases the level of significance is taken as 1% or 10%.

Type II Error

The error made when the null hypothesis is accepted when it is not true is called *Type II error*. The probability of making such an error is denoted by β

It is for the decision maker to decide which type of error he wants to avoid. It is not possible to control both the errors at the same time, as a decrease in one type leads to an increase in the probability of the other type of error.

Power of a test:

The *power of a test* is the probability that a Type II error is avoided. The probability of making a Type II error is β, so the formula to find the power of a test is as follows:

Power of a test = 1 – β

The two types of errors can be tabulated as shown below:

	H_0 is accepted	H_0 is rejected
H_0 is true	*Correct decision*	**Type I Error**
H_0 is false	**Type II Error**	*Correct decision*

Critical Region

In testing a hypothesis we assume a normal distribution of the random variable. For drawing Statistical inferences, the standard normal variate *z* is used. We know from the chapter on probability distribution, that $z = \frac{X-\mu}{SE}$, where μ is the population mean, S.E. is the standard deviation of the sample and μ_0 is the sample mean.

From the normal table, we observe that area between $z = -1.96$ and $z = 1.96$ is 95 % of the total area. (The sum of the values is 0.4750 + 0.4750 = 0.95). Similarly, the area between $z = -2.58$ and $z = 2.58$ is 99.02% of the total area and the area between $z = -1.64$ and

$z = 1.64$ is 90% of the total area. The knowledge of these areas leads us to the probabilities of the confidence level which determines the level of significance. Thus, for a 5%level of significance we know now that the z value should be between –1.96 and +1.96.These limits are called as *confidence limits*. The region between these confidence limits is called the *region of acceptance*. The region beyond these limits is the region of rejection called as the *critical region*.

If the value of z is beyond these confidence limits *i.e.* $z >–1.96$ or $z > 1.96$ then we conclude that it is not only due to sampling fluctuations but some more serious reasons. This is because the probability of such fluctuations is only 0.05, which is very small. Thus, we say that the difference between the statistic and parameter is highly significant. As a result the null hypothesis is rejected.

If the value of z_α is in the critical region, *i.e.* $–1.96 \leq z \leq 1.96$, then we say that the difference is observed due to some sampling fluctuations and is not significant. As a result the null hypothesis is accepted.

The same rule is followed for 1% and 10% level of significance. The confidence limits for the three levels of significance are as shown below:

1%	5%	10%
$–2.58 \leq z \leq 2.58$	$–1.96 \leq z \leq 1.96$	$– 1.64 \leq z \leq 1.64$

Steps to Test a Hypothesis

The steps to test a hypothesis are as follows:

1. To decide the Null Hypothesis: It is the first step to determine, our assumption about the population parameter, which is to be tested on the basis of a sample statistic. There are three null hypotheses: $\mu = \mu_0$ or $\mu > \mu_0$ or $\mu < \mu_0$.
2. To decide the level of significance: This is the probability of making a Type I error that is of rejecting the true null hypothesis. Generally we take 5% level of significance. For quality testing of a drug 1% level of significance is considered, while for pre-poll and exit poll surveys a 10% level of significance may be selected. It is left to the concerned authority's discretion to select the level of confidence based on his requirement.
3. Critical Region: Once the level of significance is decided the critical region follows immediately. It is the region of acceptance of the null hypothesis.
4. Test Statistic: After the confidence limits are decided now we require a statistical test to analyse the sample statistic. This is called as test statistic.

Since we are going to study about large sample with normal distribution the S.N.V. z is used as the test statistic.

5. Decision Making: The last and important step for testing a hypothesis is making a decision based on the result of the test statistic. The conclusion is either to accept H_0 or to reject H_0. There are two tests for making this decision depending upon what is our H_1? We have seen before that there are three possibilities for H_1, so there are the following three tests:

a) **Two Tailed Test**: If we observe the standard normal curve we see that due to symmetry of the curve, it is moving infinitely on the sides of the mean. These two ends are called *tails* of the curve. If the null hypothesis is that $\mu = \mu_0$, then we use the two tailed test as the alternative hypothesis is $\mu \neq \mu_0$ which means either $\mu < \mu_0$ or $\mu > \mu_0$. Thus we have to check the statistic for both the tails.
For 5% level of significance, if $|z_{\text{cal}}| < 1.96$, H_0 is accepted and if $|z_{\text{cal}}| > 1.96$ H_0 is rejected.

b) **Left Tailed Test**: If our null hypothesis is H_0: $\mu = \mu_0$, and H_1: $\mu < \mu_0$. Here the area of rejection is to the left of the normal curve, hence the test is called *left tailed test*.
For 5% level of significance, if $z_{\text{cal}} \leq -1.64$, then H_0 is rejected. The value -1.64 corresponds to the 47.5% area to the left of the normal curve.

c) **Right Tailed Test**: If our null hypothesis is H_0: $\mu = \mu_0$, and H_1: $\mu > \mu_0$. Here the area of rejection is to the right of the normal curve, hence the test is called *right tailed test*.
For 5% level of significance, if $z_{\text{cal}} \geq 1.64$, then H_0 is rejected. The value 1.64 corresponds to the 47.5% area to the right of the normal curve.

The following will make the idea more clear:

Large Sample Tests

Now we shall see some examples using the z-test, where a large sample ($n > 30$) is taken. In problems where the sample size is small (*i.e.* $n < 30$), another test called the t-test or the student's t – test is used. We will confine ourselves to the first test.

Example 1:

A random sample of 100 bundles gives a mean of 8.5 tons and standard deviation 4 tons. Can the sample be regarded as drawn from a population with mean 7 tons? Test this at 5% level of significance.

Ans: Given: $\mu = 7$, $X = 8.5$, SE = 4 and $n = 100$

If the standard deviation of the population is not known, the sample standard deviation is to be taken $\therefore \sigma = SE = 4$

Null Hypothesis: H_0: $\mu = 7$ Alternative Hypothesis: H_1: $\mu \neq X$

Now, $z = \frac{X-\mu}{\sigma/\sqrt{n}} = \frac{8.5-7}{4/\sqrt{100}} = -3.75$

$\therefore \; |z_{cal}| = 3.75$

At 5% level of significance the value of $z_\alpha = 1.96$

It is observed that $|z_{cal}| > z_\alpha$. Thus, the null hypothesis is rejected.

Thus, the sample cannot be regarded as being taken from the population with mean 7 tons

Example 2:

A machine produces copper plates of thickness 2cm with standard deviation of 0.4 cm. A sample of 50 copper plates is selected at random. The average thickness of the sample is 2.04cm. Test the hypothesis that the machine is performing in a normal way, at 5% level of significance.

Ans: Given: $\mu = 2$, $X = 2.04$, $\sigma = 0.4$ and $n = 50$

Let H_0: $\mu = X$ and H_1: $\mu \neq X$

Now, $z = \frac{X-\mu}{\sigma/\sqrt{n}} = \frac{2.04-2}{0.4/\sqrt{50}} = \frac{0.04}{0.05656} = 0.71 < 1.96$

$\therefore z_{cal} = 0.71$ and at 5% level of significance, we know that $z_\alpha = 1.96$

Since $z_{cal} < z_\alpha$, we conclude that the null hypothesis is accepted.

Thus, the performance of the machine producing the copper plates is normal.

Example 3:

Uniliver Company manufacture water filters and claim that their water filters have a life of atleast 18 months. Test their claim if a sample 100 water filters taken at random had an average life of 16 months with standard deviation 6 months.

Ans: Given: $\mu = 18$, $X = 16$, $\sigma = 6$ and $n = 100$

This is an example of one tailed test. Here the null hypothesis is that the average life is atleast 18 months.

Let H$_0$: $\mu \geq 18$ and H$_1$: $\mu < 18$

Now, $z = \frac{X - \mu}{\sigma/\sqrt{n}} = \frac{16-18}{6/\sqrt{100}} = \frac{-2}{0.6} = -3.33$

$\therefore z_{cal} = -3.33$ and at 5% level of significance, we know that $z_\alpha = -1.96$

Since $z_{cal} < z_\alpha$, we conclude that the null hypothesis should be accepted.

Thus, the claim of the company is proved correct.

Example 4:

A pay commission is appointed to study the wages of government employees. It was provided with the information that the average salaries of the employees are Rs. 8,400 with standard deviation Rs. 3000. But the commission selected 100 employees at random and found that their average salary is Rs. 8,800. Test at 5% level of significance, whether the sample chosen is a representative of the population?

Ans: Given: $\mu = 8400$, $X = 8800$, $\sigma = 3000$ and $n = 100$

Let H$_0$: $\mu = X$ and H$_1$: $\mu \neq X$

Now, $z = \frac{X - \mu}{\sigma/\sqrt{n}} = \frac{8800-8400}{3000/\sqrt{100}} = \frac{400}{300} = 1.33 < 1.96$

$\therefore z_{cal} = 1.33$ and at 5% level of significance, we know that $z_\alpha = 1.96$

Since $z_{cal} < z_\alpha$, we conclude that the null hypothesis is accepted.

The sample chosen by the commission represents the population of employees.

Difference between means

If two samples of size n_1 and n_2 are drawn from a population with means μ_1, μ_2 and standard deviations σ_1, σ_2 respectively then the sampling distribution of the difference between the sample means X_1 and X_2 follows a normal distribution with mean $\mu_1 - \mu_2$, standard error SE = $\sqrt{\frac{\sigma_1^{\,2}}{n_1} + \frac{\sigma_2^{\,2}}{n_2}}$ and $z = \frac{X_1 - X_2}{SE}$

In problems of this kind we first find the standard error and then the test statistic z.

Example 5:

The average income of 100 men in a city is Rs. 15,000 with standard deviation Rs. 8,500 and the average income of 100 women is Rs. 12,000 and standard deviation Rs. 9000. Can it be said at 5% level of confidence that there is a significant difference between the average income of men and women?

Ans: $H_0: \mu_1 = \mu_2$ and $H_1: \mu_1 \neq \mu_2$

Given: For men: $n_1 = 100, X_1 = 15000, \sigma = 8500$

For women: $n_2 = 100, X_2 = 12000, \sigma = 9000$

$$\text{SE} = \sqrt{\frac{{\sigma_1}^2}{n_1} + \frac{{\sigma_2}^2}{n_2}} = \sqrt{\frac{(8500)^2}{100} + \frac{(9000)^2}{100}} = \sqrt{722500 + 810000}$$

$\therefore$ SE = 1237.94

$$\therefore z = \frac{X_1 - X_2}{SE} = \frac{15000 - 12000}{1237.94} = 2.42 > 1.96$$

At 5% level of significance the value of $z_\alpha = 1.96$

It is observed that $z_{cal} > z_\alpha$. Thus, the null hypothesis is rejected.

Hence there is a significant difference between the salaries of men and women.

Large Sample Tests for Proportion

In situations when the population and sample is expressed in percentages or proportions, the method of testing the hypothesis is as follows:

If the population proportion is π with standard deviation σ and the sample proportion is p with standard error SE = $\sqrt{pq/n}$, then the test statistic is $z = \frac{\pi - p}{SE}$. In problem for testing the hypothesis fro proportion we first find the standard error and then the test statistic.

Example 6:

A manufacturer claims that 10% of his product is defective. A sample of 300 items selected at random had 32 defective items. Test his claim at 1% level of significance.

Ans: H_0: $\pi = 10\% = 0.1$ and H_1: $\pi \neq 0.1$

Given: $\pi = 10\% = 0.1, p = \frac{32}{300} = 0.11$ $\Rightarrow q = 1 - p = 0.89$

$$\therefore \text{SE} = \sqrt{pq/n} = \sqrt{\frac{0.11 \text{ x } 0.89}{400}} = 0.016$$

$$\therefore z = \frac{\pi - p}{SE} = \frac{0.1 - 0.11}{0.016} = -\,0.625$$

At 1% level of confidence the value of $| z_\alpha | = 2.58$

$\therefore |z_{cal}| = 0.626 < 2.58$

Thus, the null hypothesis is accepted. Hence the manufacturer's claim is accepted.

Example 7:

A die is thrown 5000 times and a throw of 2 or 6 is observed 1520 times. Test whether the die is biased?

Ans: Let H_0: The die is unbiased and H_1: The die is biased

Given: $\pi = \frac{1}{6} + \frac{1}{6} = \frac{1}{3} = 0.33$ (since the probability of getting a 2 or 6 is 1/6)

$$p = \frac{1520}{5000} = 0.304 \Rightarrow q = 1 - p = 0.696$$

$$\therefore \text{SE} = \sqrt{pq/n} = \sqrt{\frac{0.304 \text{ x } 0.696}{5000}} = 0.006$$

$$\therefore z = \frac{\pi - p}{SE} = \frac{0.33 - 0.304}{0.006} = 4.33 > 1.96$$

At 5% level of confidence the value of $z_\alpha = 1.96$

$\therefore z_{cal} > z_\alpha$, we reject the null hypothesis and conclude that the die is biased.

For Difference between proportions

Let two random samples of size n_1 and n_2 with proportions p_1 and p_2 respectively have the standard error SE = $\sqrt{pq\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$, where $p = \frac{n_1p_1+n_2p_2}{n_1+n_2}$.

The test statistic is then given by $z = \frac{p_1-p_2}{SE}$. We first find the combined proportion p and compute the SE. Then the test statistic is calculated.

Example 8:

An old machine produced 10 defective bolts in a batch of 300. After the servicing was done the same machine was found to produce 6 defective bolts in a batch of 200. Help the manufacturer to conclude whether the machine has improved after the servicing?

Ans: Let H$_0$: $p_1 = p_2$ and H$_1$: $p_1 \neq p_2$

Given: $p_1 = \frac{10}{300} = 0.033$ and $p_2 = \frac{6}{200} = 0.03$

$n_1 = 300$ and $n_2 = 200$

$$\therefore p = \frac{n_1p_1+n_2p_2}{n_1+n_2} = \frac{10+6}{500} = 0.032 \quad \Rightarrow q = 1-p = 0.968$$

$$\therefore \text{SE} = \sqrt{pq\left(\frac{1}{n_1}+\frac{1}{n_2}\right)} = \sqrt{0.032 \text{ x } 0.968 \text{ x } \left(\frac{1}{300}+\frac{1}{200}\right)} = \sqrt{0.0309 \text{ x } \left(\frac{500}{60000}\right)}$$

$\therefore$ SE = 0.016

$$\therefore z = \frac{p_1-p_2}{SE} = \frac{0.033-0.03}{0.016} = 0.1875 < 1.96$$

At 5% level of confidence, we observe that $z_{cal} < z_\alpha$.

Thus, the null hypothesis is accepted.

The machine has not improved.

Interval Estimation

We know from the discussion till now the different level of significance and their confidence limits. For 5% level of significance the limits for *z* are ± 1.96. This can be written as $-1.96 \le \frac{X-\mu}{SE} \le 1.96$ ⇒ $-$ 1.96 SE ≤ $X - \mu$ ≤ 1.96 SE

⇒ μ – 1.96 SE ≤ X ≤ μ + 1.96 SE

⇒ the confidence interval for sample at 5% level of confidence is (μ –1.96SE , μ + 1.96SE)

Similarly we can derive the *confidence intervals* for different levels of confidence.

The following table demonstrates these formulae:

	Level of Significance			
	1%	5%	10%	
Sample mean	μ ± 2.58SE	μ ± 1.96 SE	μ ± 1.645SE	Here, SE = $\frac{\sigma}{\sqrt{n}}$
Population Mean	X ± 2.58SE	X ± 1.96 SE	X ± 1.645SE	
Sample proportion	π ± 2.58SE	π ± 1.96 SE	π ± 1.645SE	SE = $\frac{\sqrt{pq}}{n}$
Population proportion	p ± 2.58SE	p ± 1.96 SE	p ± 1.645SE	SE= $\frac{\sqrt{\pi(1-\pi)}}{n}$

Example 9:

A coin was tossed 200 times and heads was observed 105 times. Compute the confidence intervals at 5% level of significance.

Ans: The confidence interval for sample proportion at 5% level of significance is $\pi \pm$ 1.96 SE, where SE = $\frac{\sqrt{pq}}{n}$

Now, $\pi = 0.5$ (the probability of getting heads is 1/2)

$p = \frac{105}{200} = 0.525 \Rightarrow q = 1 - p = 0.475$ and $n = 200$

SE = $\sqrt{\frac{0.525 \text{ x } 0.475}{200}} = 0.035$

$\therefore$ the confidence interval for sample proportion is $0.5 \pm (1.96 \text{ x } 0.035) = 0.5 \pm 0.069$

Thus, the confidence interval is (0.431, 0.569).

Determination of Sample Size

If the confidence level for a testing of hypothesis is known and the maximum error allowed ($E = X - \mu$) is given along with the standard deviation (σ) of the population then the size of the sample (n) can be determined by the formula: $n = \left(\frac{\sigma z}{E}\right)^2$

If the confidence level for testing a hypothesis is known, the maximum error allowed ($E = \pi - p$) along with the population proportion is given then the sample size (n) is calculated by the formula: $n = \frac{pq\, z^2}{E^2}$

Example 10:

The students of College of Engineering, Nagpur have designed a robot. The time this robot takes to react after a command is given has a standard deviation of 0.8sec. How large a sample of measuring the time should be taken by the students to be 95% confident of not exceeding an error of 0.1 sec?

Ans: Given: σ = 0.8, E = 0.1 and z = 1.96 (for 95% confidence level)

$$\therefore n = \left(\frac{\sigma z}{E}\right)^2 = \left(\frac{0.8 \text{ x } 1.96}{0.1}\right)^2 = 245.86 \approx 246$$

Thus, the sample size that the students should take for measurements is 246.