# SAMPLING

Prof. Nandini Rai

KJSCE

# POPULATION

- Population is the set or collection or totality of objects, animate or inanimate, actual or hypothetical, under study.

- Thus mainly population consists of sets of numbers, measurements or observations which are of interest.

# SIZE

- Size of the population N is the number of objects or observations in the population.
- Population is said to be finite or infinite depending on the size N being finite or infinite.

# SAMPLING

- Since it is impracticable or uneconomical or time consuming to study the entire population, a finite subset of the population known as ***Sample*** is studied.
- Size of the sample is denoted by *n*.
- *Sampling* is the process of drawing samples from a given population.

# EXAMPLES

| POPULATION | SAMPLE |
|---|---|
| Population of India | Population of Maharashtra |
| Engineering colleges recognized by AICTE | Engineering colleges affiliated to MU |
| Cars produced in India | Maruti cars |
| Healthcare expenditure by central government | Healthcare expenditure by State government |

# LARGE & SMALL SAMPLING

- If $n \geq 30$ , the sampling is said to be large sampling.
- If $n < 30$, the sampling is said to be small sampling.

# STATISTICAL INFERENCE

Statistical Inference or inductive statistics deals with

the methods of drawing (arriving at) valid or logical generalizations and predictions

about the population

using the information contained in the sample alone,

with an indication of the accuracy of such inferences.

# PARAMETERS AND STATISTICS

- Statistical measures or constants obtained from the population are known as population parameters or simply *parameters.*
- Eg. Population mean, population variance etc.
- Statistical quantities computed from sample observations are known as sample statistics or simply *statistics.*
- Eg. Sample mean , sample variance etc.

# NOTATIONS

| | POPULATION | SAMPLE |
|---|---|---|
| MEAN | μ | $\bar{X}$ |
| STANDARD DEVIATION | $\sigma$ | $s$ |
| PROPORTION | $p$ | P |

# SAMPLING DISTRIBUTION

- Draw all possible samples of size $n$, from a given population of size $N$.
- Then the total number of such samples

$$= {}^{N}C_{n} = \frac{N!}{n!\,(N-n)!} = k$$

- Compute a statistic $S$ (such as the mean, standard deviation, median, mode etc.)for each of these sample using the sample data.

# SAMPLING DISTRIBUTION ……….

- Sampling distribution of the statistic is the set of values $\{S_1, S_2, \dots \dots.\}$ of the statistic $S$ obtained one for each sample.
- Thus sampling distribution describes how a statistic $S$ will vary from one sample to the other of the same size.
- If the statistic $S$ is mean, then the corresponding distribution of the statistics is known as sampling distribution of means.

# SAMPLING DISTRIBUTION ……….

- Mean of the sampling distribution of $S$

$$= \bar{S} = \frac{1}{k}\sum_{i=1}^{k} S_i$$

- Thus we can have mean of the sampling distribution of means, mean of the sampling distribution of variance etc.
- Sampling distribution of statistic helps to learn information about the corresponding population parameters.

# STANDARD ERROR

- Standard Error is the standard deviation of the sampling distribution of a statistic $S$.
- It gives an index of the precision of the estimate of the parameters.
- As the sample size $n$ increases, standard error decreases.
- Standard error plays an important role in large sample theory and forms the basis in test of hypothesis.

# CENTRAL LIMIT THEOREM

- ***When large samples usually greater than thirty are taken into consideration then the distribution of sample arithmetic mean approaches the normal distribution irrespective of the fact that random variables were originally distributed normally or not.***

# SAMPLING DISTRIBUTION OF MEAN

If $\bar{x}$ is the mean of the sample of size $n$ drawn from the population with mean $\mu$ and standard deviation $\sigma$

then $\bar{x}$ is normally distributed with mean $\mu$ and standard deviation $\sigma/\sqrt{n}$ i.e.

$$Z = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} \text{ is a S.N.V. as } n \to \infty$$

# TEST OF SIGNIFICANCE

An important aspect of the sampling theory is to study the test of significance which will enable us to decide, on the basis of the result of the sample, whether

**(i)** the deviation between the observed sample statistic and hypothetical parameter value or

(ii) the deviation between two sample statistics

is significant or might be attribute due to change or the fluctuation of the sampling

# TEST OF SIGNIFICANCE………

- For applying the tests of significance, we first set up a hypothesis which is a definite statement about the population parameter called **Null hypothesis** denoted by $H_0$
- Any hypothesis which is complementary to the null hypothesis $(H_0)$ is called an **Alternative hypothesis** denoted by ${H_0}'\ or\ H_a$

# TEST OF SIGNIFICANCE………

**For example** if we want to test the null hypothesis that the population has a specified mean $\mu_0$, then we have $H_0: \mu = \mu_0$

Alternative hypothesis will be

**(i)** $H_a: \mu \neq \mu_0 (\mu > \mu_0 \; or \; \mu < \mu_0)$ (two tailed alternative hypothesis)

**(ii)** $H_a: \mu > \mu_0$ (right tailed alternative hypothesis or single tailed)

**(iii)** $H_a: \mu < \mu_0$ (left tailed alternative hypothesis or single tailed)

Hence alternative hypothesis helps to know whether the test is two tailed test or one tailed test

# CRITICAL REGION

- A region( corresponding to a statistic , in the sample space S) which amounts to rejection of the null hypothesis $H_0$ is called critical region of rejection.
- The region (of the sample space S) which amounts to the acceptance of $H_0$ is called acceptance region.

# LEVEL OF SIGNIFICANCE

- The probability of the value of the variate falling in the critical region is known as level of significance
- The probability $\alpha$ that a random value of the statistic  belongs to the critical region is known as the level of significance.
- $P(t \in \omega | H_0) = \alpha$

# CRITICAL VALUE OR SIGNIFICANT VALUE

The value of the test statistic which separates the critical region and acceptance region is called the critical values or significant value. This value is dependent on

**(i)** the level of significance used and

**(ii)** the alternative hypothesis, whether it is one tailed or two tailed.

- For larger samples corresponding to the statistic $t$, the variable $z = \frac{t-E(t)}{S.E.(t)}$ is normally distributed with mean 0 and variance 1.
- The value of $z$ given above under the null hypothesis is known as test statistic.

The critical value of $z_\alpha$ of the test statistic at level of significance $\alpha$ for a two tailed test given by $p(|z| > z_\alpha) = \alpha$

i.e. $z_\alpha$ is the value $z$ so that the total area of the critical region on the both tails is α.

Since the normal curve is symmetrical, we get $p(z > z_\alpha) + p(z < -z_\alpha) = \alpha;$

$i.e. 2p(z > z_\alpha) = \alpha;\ \ p(z > z_\alpha) = \alpha/2$

i.e. the area of each tail is $\alpha/2$

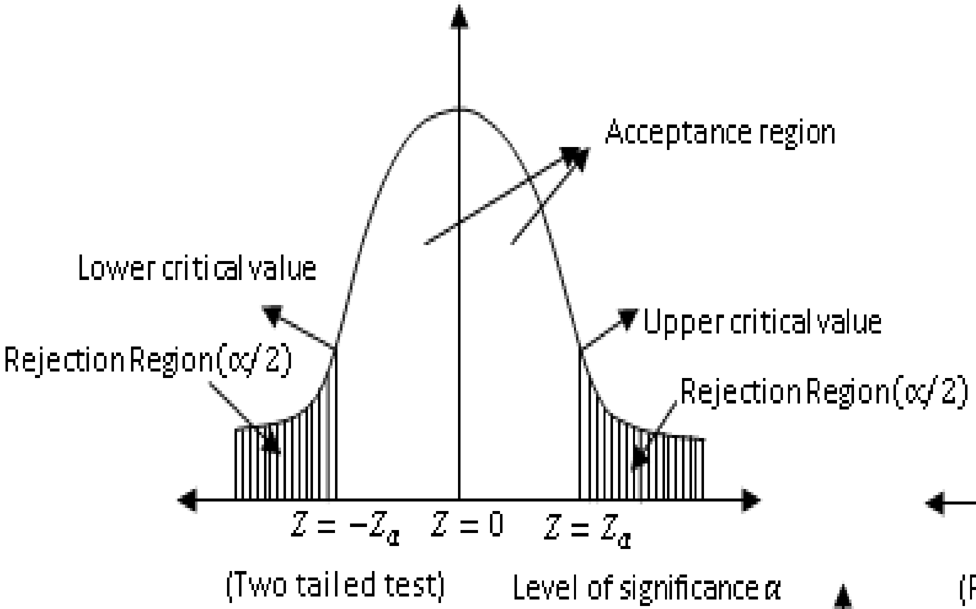
The critical value $z_\alpha$ is that value such that the area to the right of $z_\alpha$ is $\alpha/2$ and the area to the left of $-z_\alpha$ is $\alpha/2$

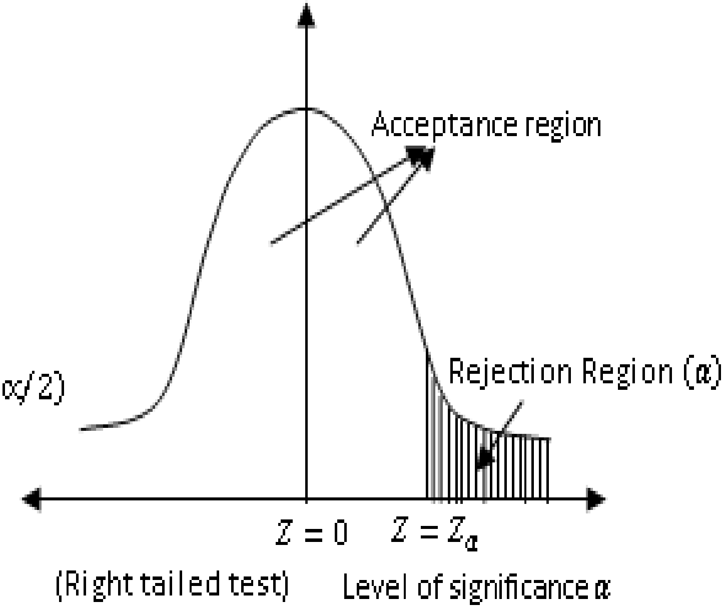
In the case of one tailed test

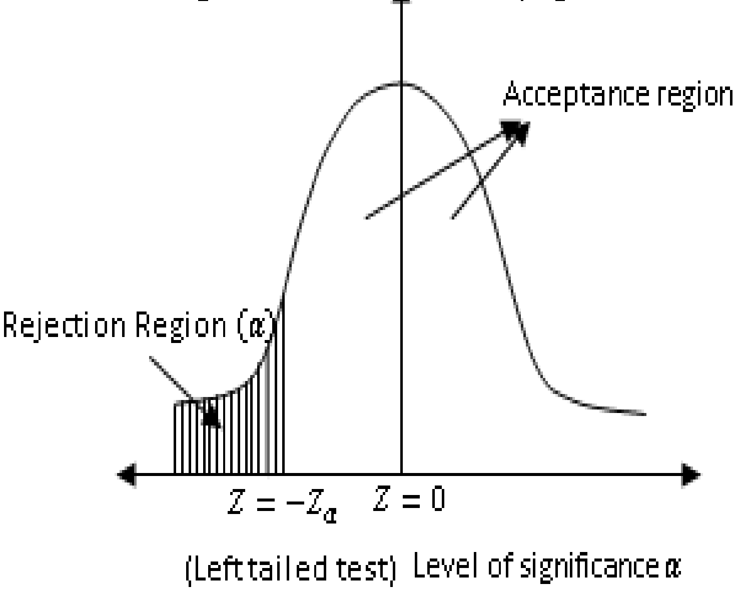
$p(z > z_\alpha) = \alpha$ if it is right tailed:

$p(z < -z_\alpha) = \alpha$ if it is left tailed

The critical value of $z$ for a single- tailed test (right or left) or level of significance $\alpha$ is same as the critical value of $z$ for two- tailed test at the level of significance $2\alpha$


Acceptance region
Lower critical value
Upper critical value
Rejection Region(α/2)
Rejection Region(α/2)
$Z = -Z_\alpha$
$Z = 0$
$Z = Z_\alpha$
(Two tailed test)
Level of significance α


Acceptance region
Rejection Region (α)
$Z = 0$
$Z = Z_\alpha$
(Right tailed test)
Level of significance α


Acceptance region
Rejection Region (α)
$Z = -Z_\alpha$
$Z = 0$
(Left tailed test)
Level of significance α

Using the equation, also using the normal tables, the critical value of $z$ at different level of significance $(\alpha)$ for both single tailed and two tailed test are calculated and listed below. The equations are

- $p(|z| > z_\alpha) = \alpha$;(two tailed)
- $p(z > z_\alpha) = \alpha$;(right tailed)
- $p(z < -z_\alpha) = \alpha$ (left tailed)

| Level of significance | | | |
|---|---|---|---|
| | 1% (0.01) | 5% (0.05) | 10% (0.1) |
| Two tailed test | $\lvert z_\alpha \rvert = 2.58$ | $\lvert z_\alpha \rvert = 1.966$ | $\lvert z_\alpha \rvert = 1.645$ |
| Right tailed | $z_\alpha = 2.33$ | $z_\alpha = 1.645$ | $z_\alpha = 1.28$ |
| Left tailed | $z_\alpha = -2.33$ | $z_\alpha = -1.645$ | $z_\alpha = -1.28$ |

# TESTING OF HYPOTHESIS

**Step 1:Null hypothesis.** Set up $H_0$ in clear term

**Step2:Alternative hypothesis.** Set up $H_a$, so that we could decide whether we should use one-tailed test or two tailed test

**Step 3:Level of significance.** Select the appropriate level of significance in advance depending on the reliability of the estimates

# TESTING OF HYPOTHESIS

**Step 4: Test statistic.** Compute the test statistics $z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$ under the null hypothesis

**Step 5: Conclusion.** Compare the compute value of $z$ with the critical value $z_\alpha$ at the level of significance $(\alpha)$.

If $|z| > z_\alpha$, we **reject** $\boldsymbol{H_0}$ and conclude that there is significant difference.

If $|z| < z_\alpha$, we **accept** $\boldsymbol{H_0}$ and conclude that there is no significant difference.

# ESTIMATION

How do we know the mean or proportion of the population?

The procedure of finding this is called estimation.

There are two ways of estimating the mean or proportion of population.

We may estimate the value or the interval in which the value may lie.

The first is called **point estimation** and the second is called the **interval estimation**.

# POINT ESTIMATION

It can be proved that the best estimate of population mean is the sample mean.

In other words, If the sample mean is $\bar{X}$ we estimate that the population mean $\mu$ is also $\bar{X}$.

For example: If average salary of a sample of 100 workers in a factory employing 1000 workers is found to be Rs. 500, we estimate that the average salary of all the 1000 workers is also Rs. 500.

# INTERVAL ESTIMATION

However, more often we estimate the interval in which the population mean would lie.

The sample mean is normally distributed with
mean = the population mean $\mu$
standard deviation = $\sigma/\sqrt{n}$.

We know that at 5% level of significance $|Z| < 1.96$

i.e $\left|\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}\right| < 1.96$

$$i.e\ |\bar{X} - \mu| < 1.96\frac{\sigma}{\sqrt{n}}$$

# INTERVAL ESTIMATION

$$\bar{X} - \mu < 1.96\frac{\sigma}{\sqrt{n}} \quad \boldsymbol{or} \quad \mu - \bar{X} < 1.96\frac{\sigma}{\sqrt{n}}$$

$$\bar{X} - 1.96\ \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}$$

This gives the interval in which we estimate $\mu$ to lie with 95% confidence. They are called **confidence limits.**

Other limits at other confidence levels can be similarly found.

# CONFIDENCE LEVEL

The Probability which we associate with an interval estimate is called confidence level.

The higher the probability, the more our confidence. The most commonly used confidence levels are 90%, 95%, 98% and 99%.

From the table of areas under normal curve we find that 90% area lies between $\mu - 1.64\,\sigma$ and $\mu + 1.64\sigma$ where $\mu$ is the mean of the population and $\sigma$ is the standard deviation of the population.

# CONFIDENCE LEVEL…….

Similarly 95% area lies between $\mu - 1.96\,\sigma$ and $\mu + 1.96\sigma.$

98% area lies between $\mu - 2.33\sigma$ and $\mu + 2.33\sigma$

99% area lies between $\mu - 2.58\sigma$ and $\mu + 2.58\sigma.$

The values $1.64, 1.98, 2.33, 2.56\ of\ S.N.V\ z$ corresponding to confidence levels

$\alpha = 90\%, 95\%, 98\%$ and $99\%$ are called the critical values and are denoted by $z_\alpha.$

# ERRORS IN TESTING OF HYPOTHESIS

When a statistical hypothesis is tested there are only two results either we accept it or we reject it. We never know whether the hypothesis is true or false. Hence there arise four possibilities.

# ERRORS IN TESTING OF HYPOTHESIS

**(i)** A true hypothesis is rejected

**(ii)** A true hypothesis is accepted

(iii) A false hypothesis is rejected

(iv) A false hypothesis is accepted

If outcome of the test leads to the possibilities

(i) rejecting true hypothesis------ **type I error**

(iv) accepting a false one ------ **type II error**.

# TYPE -I ERROR

Type I error arises when a true hypothesis is rejected

when the difference between the sample value and hypothetical value exceeds the confidence limits.

The error can be minimized by increasing the confidence limits.

But then because of this the error of type II i.e of accepting a false hypothesis is increased,

because we do not know whether the hypothesis is true or false in reality.

# TYPE -II ERROR

Type II Error arises when a false hypothesis is accepted

When the difference between the sample value and the hypothetical value lies within the limits.

The error, can be minimized by decreasing the confidence limits.

But then the possibility of error of type I i.e of rejecting a true hypothesis is increased,

because we again do not know whether the hypothesis is true or false in reality.

The four situations arising in the process of decision making can be described in the form of a table as

| | $H_0$ is accepted | $H_0$ is rejected |
|---|---|---|
| $H_0$ is true | Correct decision | Type I error |
| $H_0$ is false | Type II error | Correct decision |