# MODULE :3

# SAMPLING THEORY

# Sampling: Introduction

- The group of individuals under study is called population or universe. It may be finite or infinite.
- A part selected from the population is called a sample. The process of selection of a sample is called sampling. A Random sample is one in which each member of population has an equal chance of being included.

# Symbols for Population and Sample

| Characteristic | Population | Sample |
|---|---|---|
| | Parameter | Statistic |
| Symbols | population size = $N$<br>population mean = $\mu$<br>population standard deviation = $\sigma$ | sample size = $n$<br>sample mean = $\bar{x}$<br>sample standard deviation = $s$ |

# Aims of Sample

The population parameters are not known generally. Then the sample characteristics are utilised to approximately determine or estimate of the population. Thus, static is an estimate of the parameter. To what extent can we depend on the sample estimates?

The estimate of mean and standard deviation of the population is a primary purpose of all scientific experimentation. The logic of the sampling theory is the logic of *induction*. In induction, we pass from a particular (sample) to general (population). This type of generalization here is known as *statistical inference*. The conclusion in the sampling studies are based not on certainties but on probabilities.

## Sample Size (n)

- If $n \geq 30$, sample is known as Large Sample
- If $n < 30$, sample is known as Small Sample

For large sample size, any data set that is randomly sampled from a population, regardless of distribution, will be approximately normal

Study of statistics is mainly divided into two categories :

1.Estimation

2. Testing of hypothesis

## Estimation

- Guessing the value of unknown population parameter, using sample observations is called estimation.
- Results of estimation can be expressed as a single value, known as a point estimate, or a range of values, known as a confidence interval.

## Testing of hypothesis

- Any statistical statement or assumption about the population or form of the population or about the parameters of the population is called statistical hypothesis.
- It is always a statement about the population and not about the sample.
- Null hypothesis: An assumption about the population which is believed to be true is called null hypothesis (H0). It is a hypothesis of no difference.
- Alternative hypothesis: Any other logical alternative statement to null hypothesis is called alternative hypothesis (H1).
- Testing of hypothesis: It is a two actions decision problem, two actions being either reject null hypothesis or accept null hypothesis based on sample observations.

## Critical region and Acceptance region and Types of Errors

- To take a decision about acceptance or rejection of null hypothesis, divide the range of sample statistics into two complementary regions say C and C’.
- If the observed value of sample statistics belongs to region C, then our decision is to reject null hypothesis. Region C is called critical region.
- If the observed value of sample statistics belongs to region C’, then our decision is to accept null hypothesis. Region C’ is called acceptance region.
- Since decision about acceptance or rejection of null hypothesis is taken on the basis of sample observations only, therefore decision taken need not be always correct. If the decision taken is not correct, it is called as a mistake or error.
- When null hypothesis is true and it is rejected, the mistake is called type I error. Probability of type I error is denoted by $\alpha$.
- When null hypothesis is false and it is accepted, the mistake is called type II error. Probability of type II error is denoted by $\beta$.

# Level of Significance

$Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ is a standard normal variate. For SNV, 95% area under the curve lies between -1.96 and +1.96, 99% between -2.58 and +2.58. That is only 5% area under the curve lies beyond $\mp 1.96$ and 1% beyond $\mp 2.58$ for the cases defined.

The probability level below which we reject the hypothesis is known as the *level of significance*. The region in which a sample value falling is rejected, is known as the *critical region*. We generally take two critical regions which cover 5% and 1% areas of the normal curve. The shaded portion in the figure corresponds to 5% level of significance. Thus the *probability of the value of the variate falling in the critical region is the level of significance.*

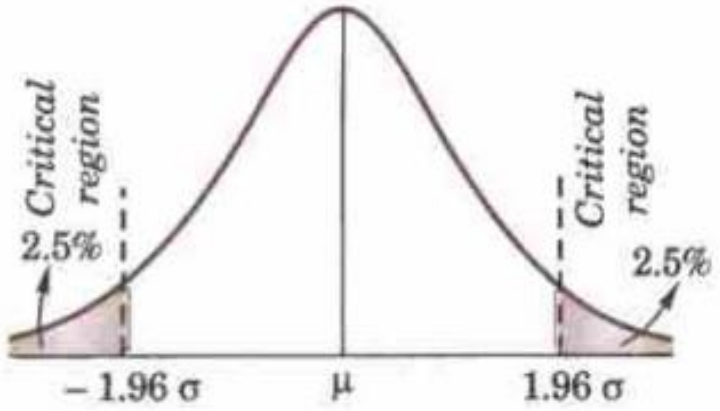


Fig. 27.1

Depending on the nature of the problem, we use a *single-tail test* or *double-tail test* to estimate the significance of a result. In a double-tail test, the areas of both the tails of the curve representing the sampling distribution are taken into account whereas in the single tail test, only the area on the right of an ordinate is taken into consideration.

## Standard Critical Values

The Critical Values of $|z_\alpha|$ for some standard level of significance $\alpha\%$ for both two-tailed and one-tailed tests are given as follows

| Nature of test | LOS | |
|---|---|---|
| | 5%(.05) | 1%(.01) |
| 2-tailed | 1.96 | 2.58 |
| 1-tailed | 1.64 | 2.33 |

# Problems on Interval Estimation

Let the population from which a random sample of size n is drawn, have mean μ and standard deviation ó. If μ is not known, there will be a range of values of μ for which observed mean $\bar{x}$ of the sample is not significant at any assigned level of probability. For 5% LOS, we have $|Z| = |\frac{\bar{x}-\mu}{ó/\sqrt{n}}|<1.96$.

$$|(\bar{x}-\mu)\sqrt{n}/\sigma| < 1.96 \quad i.e. \quad \bar{x} - 1.96\sigma/\sqrt{n} < \mu < \bar{x} + 1.96\sigma/\sqrt{n}$$

*Thus 95% confidence or fiducial limits for the mean of the population corresponding to given sample are* $\bar{x} \pm 1.96\sigma/\sqrt{n}$.

Similarly *99% confidence limits for* μ *are* $\bar{x} \pm 2.58\sigma/\sqrt{n}$.

Ex1-A random sample of 100 items from a normal population of unknown mean has mean 10 and standard deviation 1.5. What are 95% and 99% fiducial limits for the population mean?

- For 5% LOS or 95% confidence level, we have $|Z|=\left|\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}\right| <1.96$
- n=100, $\bar{x}$=10, $\sigma$=1.5
- Confidence Interval is $(\bar{x}\text{-}1.96\frac{\sigma}{\sqrt{n}}\ , \bar{x}+1.96\frac{\sigma}{\sqrt{n}})$

  =(10-1.96*1.5/10 , 10+1.96*1.5/10)

  =(9.706, 10.294)

Ex2- Two samples are drawn from two different population gave the following results . Find 95% and 99% confidence limits for the difference between the population means.

| | size | mean | s.d. |
|---|---|---|---|
| Sample 1 | 400 | 124 | 14 |
| sample2 | 250 | 120 | 12 |

- For 5% LOS or 95% confidence level, we have $|Z| = \left|\frac{(\bar{x}_1-\bar{x}_2)-\mu}{\sqrt{\frac{{s_1}^2}{n_1}+\frac{{s_2}^2}{n_2}}}\right| < 1.96$
- $n_1 = 400, n_2 = 250,\ \bar{x}_1 = 124,\ \bar{x}_2 = 120$, $s_1$=14, $s_2$=12
- $\bar{x}_1-\bar{x}_2 = 4$
- Confidence Interval is $((\bar{x}_1 - \bar{x}_2)$-1.96$\left(\sqrt{\frac{{s_1}^2}{n_1} + \frac{{s_2}^2}{n_2}}\right)$ , $(\bar{x}_1 - \bar{x}_2)$ +1.96$\left(\sqrt{\frac{{s_1}^2}{n_1} + \frac{{s_2}^2}{n_2}}\right)$)

  =(4-1.96*1.03 , 10+1.96*1.03)

  =(1.98, 6.02)

## Exercise

1. A sample of 900 numbers has a mean 3.4 cms and s.d. 2.61 cms. If the population is normal, find the 95% and 98% fiducial limits of the true mean.

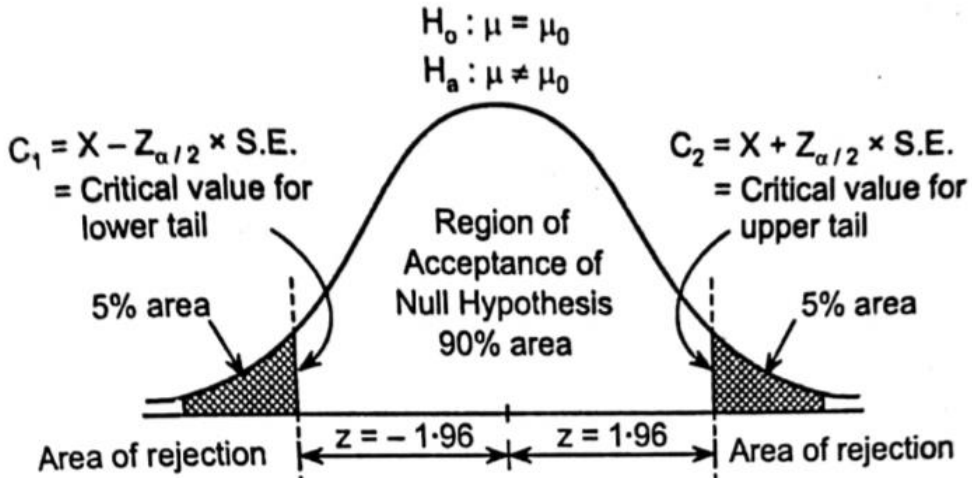
   ans (3.5705, 3.2295) , (3.6027, 3.1973)

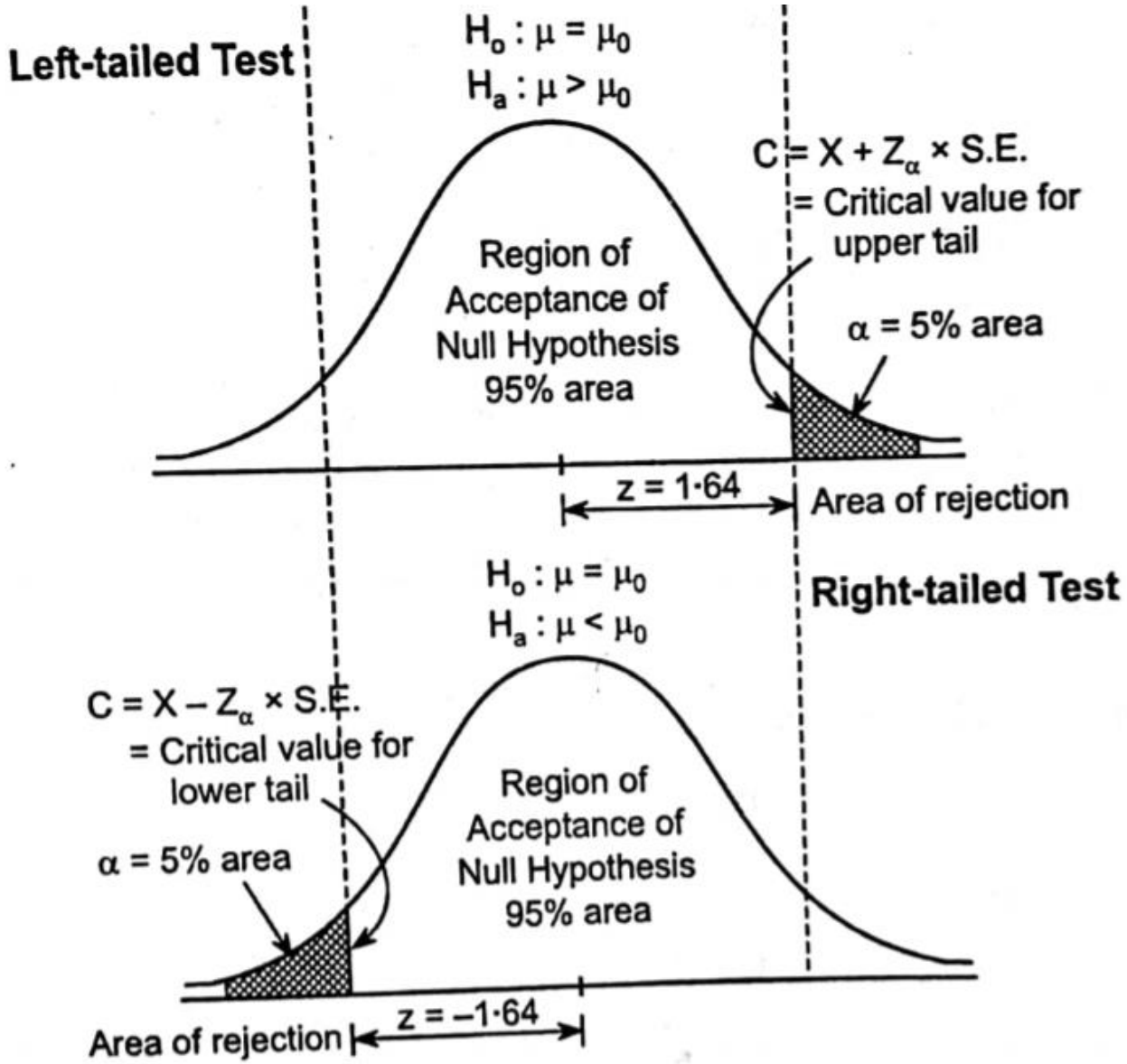
# Problems on Large Sample: Testing Hypothesis--method

1. Null hypothesis H0 is defined
2. Alternative hypothesis H1 is defined after deciding whether the test is one-tailed or two-tailed
3. LOS α is fixed and critical value $z_\alpha$ is noted
4. Test statistic $z_{cal}$ is calculated
5. If $|z_{cal}| < z_\alpha$ , null hypothesis is accepted
6. Conclusion: the difference between population parameter and sample statistic is not significant $|z_\alpha|$

Note: If $|z_{cal}| > z_\alpha$ , null hypothesis is rejected and H1 is accepted and the conclusion is that the difference between population parameter and sample statistic is significant $|z_\alpha|$.

# LOS α=10%


$H_o : \mu = \mu_0$
$H_a : \mu \neq \mu_0$
$C_1 = X - Z_{\alpha/2} \times S.E.$
= Critical value for lower tail
$C_2 = X + Z_{\alpha/2} \times S.E.$
= Critical value for upper tail
Region of Acceptance of Null Hypothesis 90% area
5% area
5% area
Area of rejection
z = – 1·96
z = 1·96
Area of rejection


Left-tailed Test
$H_o : \mu = \mu_0$
$H_a : \mu > \mu_0$
$C = X + Z_\alpha \times S.E.$
= Critical value for upper tail
Region of Acceptance of Null Hypothesis 95% area
α = 5% area
z = 1·64
Area of rejection
Right-tailed Test
$H_o : \mu = \mu_0$
$H_a : \mu < \mu_0$
$C = X - Z_\alpha \times S.E.$
= Critical value for lower tail
α = 5% area
Region of Acceptance of Null Hypothesis 95% area
Area of rejection
z = –1·64

## 1.Test for Significance between sample mean and population mean

$$z_{cal} = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$$

Ex1- A random sample of 400 observations has mean 4.45 cm. Can it be a random sample from a population whose mean is 5 cm and variance is 4 cm ?

Given: n=400, $\bar{x}$=4.45, μ=5, $\sigma = \sqrt{4}$=2

1. H0: $\bar{x}$=μ
2. H1: $\bar{x} \neq \mu$ (as nature of the test is two tailed)
3. Let LOS α be 5% so critical value $z_\alpha$=1.96
4. $z_{cal} = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} = \frac{4.45 - 5}{2/20} = 5.5$
5. $|z_{cal}| = 5.5 > z_\alpha$ , null hypothesis is rejected
6. Conclusion: The random sample can not be regarded as the sample taken from the population.

Ex2-The mean height of a random sample of 100 individuals from a population is 165 cm. The sd of the sample is 20. Would it be reasonable to suppose that the mean height of population is 165 cm?

Given: n=100, $\bar{x}$=160, μ=165, $s = 20$

1. H0: $\bar{x}$=μ
2. H1: $\bar{x} \neq \mu$ (as nature of the test is two tailed)
3. Let LOS α be 5% so critical value $z_\alpha$=1.96

   If Let LOS α be 1% so critical value $z_\alpha$=2.58

4. $z_{cal} = \frac{\bar{x}-\mu}{s/\sqrt{n}} = \frac{160-165}{20/10} = -2.50$
5. $|z_{cal}| = 2.5 > 1.96 = z_\alpha$ for LOS α be 5% , null hypothesis is rejected
6. Conclusion: The random sample can not be regarded as the sample taken from the population under 5% LOS
7. $|z_{cal}| = 2.5 < 2.58 = z_\alpha$ for LOS α be 1% , null hypothesis is accepted hence random sample can be regarded as the sample taken from the population under 1% LOS.

Note: here instead of population sd, sample sd is given so replace $\sigma$ by s

Ex3- Can it be concluded that the average life span of an Indian is more than 70 years if a random sample of 100 Indians has an average life span of 71.8 years with standard deviation of 7.8 years. Use 5% LOS.

Given: n=100, $\bar{x}$=71.8, μ=70, s = 7.8

1. H0: $\bar{x}$=μ
2. H1: $\bar{x} >$ μ (as nature of the test is left tailed)
3. Let LOS α be 5% so critical value $z_\alpha$=1.645
4. $z_{cal} = \frac{\bar{x}-\mu}{s/\sqrt{n}} = \frac{71.8-70}{7.8/10} = 2.02$
5. $|z_{cal}| = 2.02 > 1.64 = z_\alpha$ for LOS α be 5% , null hypothesis is rejected
6. Conclusion: the average life span of an Indian is more than 70 years under 5% LOS

# Exercise

1

**Example** . *A sample of 900 members has a mean 3·4 cms,. and s.d. 2·61 cms. Is the sample from a large population of mean 3·25 cms. and s.d. 2·61 cms. ?*

## 2.Test for Significance between two sample means

(i) If samples are taken from different population

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

(ii) If samples are taken from same population

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

Ex1- A sample of 200 fish of a particular kind taken at random from one end of the lake had mean weight 20 pounds and standard deviation 2 pounds. At the other end of the lake a sample of 80 fish oh the same kind had mean weight 20.5 pounds and standard deviation 2 pounds. Is the difference between the two mean weights significant?

Given: $n_1 = 200, n_2 = 80,\ \bar{x}_1 = 20,\ \bar{x}_2 = 20.5$, $s_1$=2, $s_2$=2

1. H0: $\bar{x}_1 = \bar{x}_2$
2. H1: $\bar{x}_1 \neq \bar{x}_2$(as nature of the test is two tailed)
3. Let LOS α be 5% so critical value $z_\alpha$=1.96
4. Formula used $z_{cal} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$ (as the samples are taken from same population)

$$z_{cal} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{20 - 20.5}{\sqrt{\frac{4}{200} + \frac{4}{80}}} = -1.89$$

5. $|z_{cal}| = 1.89 < 1.96 = z_\alpha$ for LOS α be 5% , null hypothesis is accepted
6. Conclusion: the difference between the two mean weights is not significant under 5% LOS.

Ex2-Average height of a sample of 6400 persons from one population was found to be 67.85 inches with standard deviation 2.56 inches. Average height of a sample of 1600 persons from another population was found to be 68 inches with standard deviation 2.52 inches. Is the difference between the two mean heights of the two samples significant? Use 5% and 1% LOS

Given: $n_1 = 6400, n_2 = 1600,\ \bar{x}_1 = 67.85,\ \bar{x}_2 = 68,\ s_1 = 2.56,\ s_2 = 2.52$

1. H0: $\bar{x}_1 = \bar{x}_2$
2. H1: $\bar{x}_1 \neq \bar{x}_2$ (as nature of the test is two tailed)
3. If LOS α be 5% so critical value $z_\alpha$=1.96

   If Let LOS α be 1% so critical value $z_\alpha$=2.58
4. Formula used $z_{cal} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$ (as the samples are taken from different population)

$$z_{cal} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{67.85 - 68}{\sqrt{\frac{2.56^2}{6400} + \frac{2.52^2}{1600}}} = -2.12$$

$= -2.12$

5. $|z_{cal}| = 2.12 > 1.96 = z_\alpha$ for LOS α be 5% , null hypothesis is rejected
6. Conclusion: the difference between the two mean heights is significant under 5% LOS.
7. $|z_{cal}| = 2.12 < 2.58 = z_\alpha$ for LOS α be 1% , null hypothesis is accepted and the difference between the two mean heights is not significant under 1% LOS.

# Exercise

1

**Example** . *The means of two single large samples of 1000 and 2000 members are 67·5 inches and 68·0 inches respectively. Can the samples be regarded as drawn from the same population of standard deviation 2·5 inches ? (Test at 5% level of significance).*

2. .

**Example** . *In a certain factory there are two independent processes manufacturing the same item. The average weight in a sample of 250 items produced from one process is found to be 120 ozs. with a standard deviation of 12 ozs. while the corresponding figures in a sample of 400 items from the other process are 124 and 14. Obtain the standard error of difference between the two sample means. Is this difference significant ? Also find the 99% confidence limits for the difference in the average weights of items produced by the two processes respectively.*

3. A tyre company claims that the lives of the tyres have mean of 42000 kms with S.D. of 4000 kms. A change in production process is believed to result in a better product. A test sample of 81 new tyres has a mean life of 42500 kms. Test at 5% LOS that the new product is significantly better than the current one.