7Introduction to sampling distribution:

Every statistical investigation aims at collecting information about some aggregate or collection of individuals or of their attributes, rather than the individuals themselves. In statistical language, such collection is called a population or universe.

e.g. 1) All the student in vit.2) Production of sugar bags in sugar factory. 3) Cooking rise in cooker.

A finite sub-set of a population is called sample and the process of selection of such sample is called sampling

The basic objective of theory of sampling is to draw inference about the population using the information of the sample.

Parameter: A statistical measures such mean, standard deviation calculated from the, whole universe is called a parameter.

Statistic: A statistical measures such mean, standard deviation calculated from the values of a sample is called a statistic.

Note: The value of a statistic will vary from one sample to another sample [as the values of the population members included in different samples may be different drawn from same population]. These difference in the values of statistic are said to be due to sampling fluctuations.

Sampling distribution:

If a number of samples each of same size n are drawn from same population and if for each sample the value of some statistic say mean is calculated, a set of values of the statistic will be obtained

If a number of samples is large. The value of statistic may be classified in the form of frequency table

The probability distribution of a statistic that would be obtained if the number samples each of same size, were infinitely large is called sampling distribution of the statistics.

The nature of the sampling distribution of a statistic can be obtained theoretically using the theory of probability provided the nature of the population distribution is known.

The **standard deviation** of the sampling distribution of a statistic is called the **standard error** of the statistics.

Estimation and Testing of hypothesis:

1) Some characteristic of the population in which we are interested may be completely unknown to us and we may like to make a guess about this characteristic entirely on the basis of a random sample drawn from the population. This type of problem is known as the problem of estimation.

2) Some information regarding the characteristic of the population may be available to us and we may like to know whether the information can be accepted on the basis of random sample drawn from the population. If it can be accepted, with what degree of confidence it can be accepted. This type of problem is known as the problem of testing of hypothesis.

Testing of hypothesis and Test of significance:

When we attempt to make a decision about the population on the basis of sample information, we have to make assumption or guess about population involved or about the value of some parameter of the population, such assumptions which may or may not be true are called statistical hypothesis

We set up a hypothesis which assumption that there is no significant difference between the sample statistic and the corresponding

population parameter or between two sample statistics. Such a hypothesis of no difference is called a null hypothesis and is denoted by **H_0**

A hypothesis that is different from the null hypothesis is called an alternative hypothesis and is

Denoted by **H_1**

A procedure for deciding whether to accept or to reject a null hypothesis is called testing of the hypothesis.

If **θ_0** is the population parameter and **θ** is the correspond sample statistic, usually there will be some difference between θ_0 & θ since θ is based on sample observation and is different for different samples such a difference which is due to sampling fluctuations is called insignificant difference.

The difference that arises either because the sampling procedure is not purely random or the sample has not been drawn from given population is known as significant difference. This procedure of testing whether the difference between θ_0 & θ is significant or not is called the test of significance.

Critical region and Level of significance:

If we are prepare to accept that the difference between a sample statistic and the corresponding population parameter is **significant.**

When sample statistic lies in a certain region or interval then that region is called critical region or region of rejection.

The region complementary to the critical region is called region of acceptance.

In the case of large samples, the sampling distribution of many statistics tends to become normal distribution.

If $\bar{x}$ is a statistic in a large samples, then $\bar{x}$ follows a normal distribution with mean μ =E(t), which is the corresponding population parameter and standard deviation equal to S.E.$(\bar{x})(\sigma/\sqrt{n},)$Hence

Z=$\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$ is a standard normal variate

i.e. Z follows a standard normal distribution with mean zero and standard deviation unity.

It is known from the study of normal distribution that the area under the standard normal curve between Z -1.96 to Z+1.96 is .95.

i.e. The area under the normal curve of t between [E (t)-1.96S.E. (t)] and [E (t) +1.96S.E.(t)] is .95

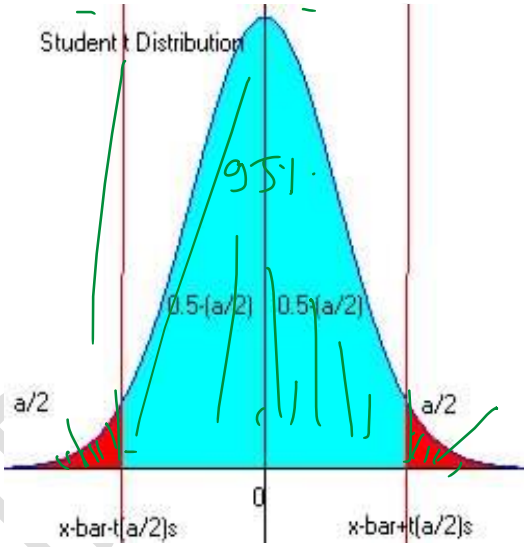


Figure no-01

Note: In figure take value of $a = 2.5$

i.e. 95% values of t will be between [E (t) ± 1.96 SE (t)] or

Only 5% values of t will lie outside this interval

If we are prepared to accept that the difference between t and E(t) is significant when t lies in either of the region[-∞,E(t)-1.96S.E.(t)] and [E(t)+1.96S.E.(t),+ ∞], then two region constitute the critical region of t

The probability that a random value (α) of the statistic lies in the critical region is called the level of significance (LOS) and is usually expressed in a percentage

The total area of a critical region is expressed as α% is the LOS

We know P [E (t)-1.96 S.E. (t) ≤ t ≤ E (t) +1.96 S.E. (t)] = .95

i.e. P [| $\frac{t-E(t)}{SE(t)}$ |≤ 1.96] = .95

Error in Testing of Hypothesis:

The error committed in rejecting **H_0** when it is really true is called **Type I error.** This is similar to a good product being rejected by the consumer and hence **Type I error** is also known as procedures risk. The error committed in accepting H_0. When it is false is called type **II error.** As this error is similar to that of accepting a product of inferior quality it is also known as consumers risk.

One tailed test and two tailed test:

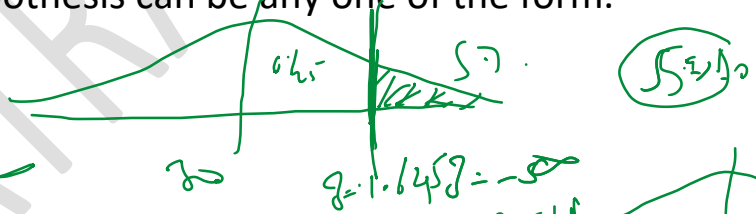
If θ_0 is the population parameter and θ is the corresponding sample statistic and if we set up the null hypothesis

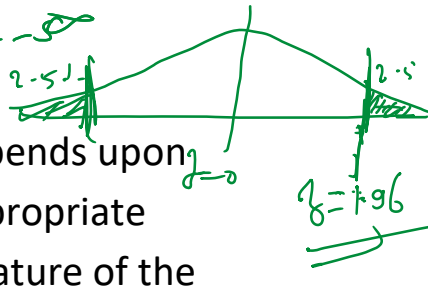
$H_0 : \theta = \theta_0$ then alternative hypothesis can be any one of the form.

1) $H_1 : \theta \neq \theta_0$ (two tailed test)

2) $H_1 : \theta > \theta_0$ (right tailed test)

3) $H_1 : \theta < \theta_0$ (left tailed test)





Note: Test no.2&3 are also called one tailed test.

Note: The application of one tailed test or two tail test depends upon the nature of the alternative hypothesis. The choice of appropriate alternative hypothesis depends on the situation and the nature of the problem concerned.

Critical value or significant value:

The value of the test statistic **Z** for which the critical region and acceptance region are separated is called Critical value or significant value of **Z** and is denoted by Z_α ,where **α** is the **LOS**

It is clear that the value of Z_α depends not only on **α** but also on the on the nature of the alternative hypothesis.

When $Z = \frac{t - E(t)}{S.E(t)}$

Then P (|Z|<1.96) =.95 & P (|Z|>1.96) =.05

Thus Z=±1.96 separate the critical and the acceptance region at 5% LOS for a two tailed test

i.e. Critical values of z in this case are ±1.96 See in the Figure no. 1

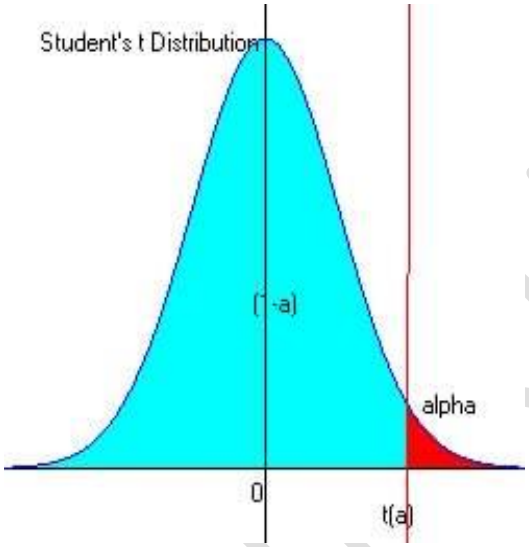
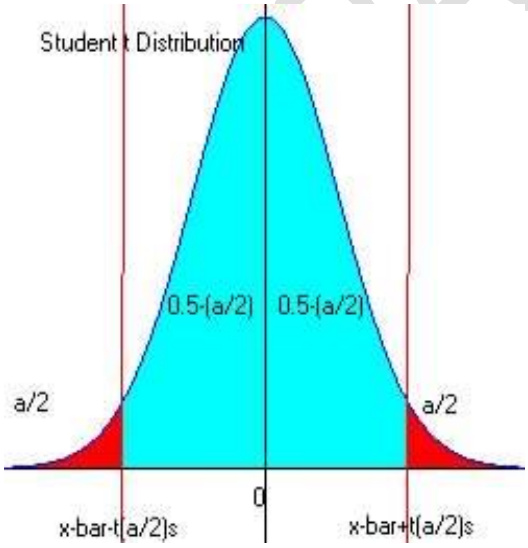


Figure no.2



Note: If Z_α is the critical value of z corresponding to the LOS **α** in the right tailed test then P (Z> Z_α)= α,and

Critical value of Z for α% LOS in two tailed test is same as (α/2)% LOS in one tailed test

by symmetry $P(Z<-Z_\alpha)=\alpha$

$P(|Z|>Z_\alpha)=P(Z>Z_\alpha, Z<-Z_\alpha)=P(Z>Z_\alpha)+P(Z<-Z_\alpha)=2\alpha$

Thus the value of Z for single tailed test at LOS α is the same as that for a two tailed test at LOS 2α

The critical values for some standard LOS are given below

Nature of test	**LOS**			
	1%(.01)	**2%(.02)**	**5%(.o5)**	**10%(.01)**
Two tail	$\|Z_\alpha\|=2.575$	$\|Z_\alpha\|=2.33$	$\|Z_\alpha\|=1.96$	$\|Z_\alpha\|=1.645$
Right tailed	$Z_\alpha=2.33$	$Z_\alpha=2.055$	$Z_\alpha=1.645$	$Z_\alpha=1.28$
Left tailed	$Z_\alpha=-2.33$	$Z_\alpha=-2.055$	$Z_\alpha=-1.645$	$Z_\alpha=-1.28$

Procedure for Testing of Hypothesis:

1) Null Hypothesis H_0 is defined

2) Alternative hypothesis H_1 is also defined after a carefully study of the problem and also the nature of the test

3) LOS α is fixed or taken from the problem if specified and Z_α is noted

4) Comparison is made between $|Z|$ and Z_α

i) If $|Z| < Z\alpha$, H_0 is accepted or H_1 is rejected.

i.e. it is concluded that the difference between t and E (t) is not significant at α % LOS

ii) If $|Z| > Z\alpha$, H_0 is rejected or H_1 is accepted

i.e. it is concluded that the difference between t and E (t) is significant at α % LOS

Interval estimation of population parameter:

$$\left|\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}\right| \le Z_\alpha$$

i.e. $$|\bar{x}-\mu| \le Z_\alpha \frac{\sigma}{\sqrt{n}}$$

i.e. $\bar{x}-\mu \le z_\alpha \frac{\sigma}{\sqrt{n}}$ and $-\bar{x}+\mu \le z_\alpha \frac{\sigma}{\sqrt{n}}$

i.e. $\bar{x} - z_\alpha \frac{\sigma}{\sqrt{n}} \le \mu$ and $\mu \le \bar{x} + z_\alpha \frac{\sigma}{\sqrt{n}}$

i.e. $\bar{x} - z_\alpha \frac{\sigma}{\sqrt{n}} \le \mu \le \bar{x} + z_\alpha \frac{\sigma}{\sqrt{n}}$

If z_α=1. 645 then $P[\bar{x} - z_\alpha \frac{\sigma}{\sqrt{n}} \le \mu \le \bar{x} + z_\alpha \frac{\sigma}{\sqrt{n}}]$=.90

If z_α=1.96 then $P[\bar{x} - z_\alpha \frac{\sigma}{\sqrt{n}} \le \mu \le \bar{x} + z_\alpha \frac{\sigma}{\sqrt{n}}]$=.95

If z_α=2.575 then $P[\bar{x} - z_\alpha \frac{\sigma}{\sqrt{n}} \le \mu \le \bar{x} + z_\alpha \frac{\sigma}{\sqrt{n}}]$=.99