UNIT II

SAMPLING DISTRIBUTION AND ESTIMATION

Population:

A population consists of collection of individual units, which may be person's or experimental

outcomes, whose characteristics are to be studied

Sample:

A sample is proportion of the population that is studied to learn about the characteristics of the population.

Random sample:

A random sample is one in which each item of a population has an equal chance of being selected.

Sampling:

The process of drawing a sample from a population is called sampling

Sample size:

The number of items selected in a sample is called the sample size and it is denoted by 'n'. If $n \geq 30 \leq$,the sample is called large sample and if $n \leq 30$ it is called small sample

Sampling distribution:

Consider all possible samples of size' n' drawn from a given population at random. We calculate mean values of these samples.

If we group these different means according to their frequencies, the frequency distribution so formed is called sampling distribution.

The statistic is itself a random variate. Its probability distribution is often called sampling distribution.

All possible samples of given size are taken from the population and for each sample, the statistic is calculated. The values of the statistic form its sampling distribution.

Standard error:

The standard deviation of the sampling distribution is called the standard error.

Notation:

Population Size $= N$	Sample size $= n$
Population mean $\mu \quad P\, p$	Sample mean $=\bar{x}$
Population standard deviation $= \sigma$	Sample standard deviation$= \ s$
Population proportion $= P$	Sample population $= p$

Null Hypothesis (H_0 **)**

The hypothesis tested for possible rejection under the assumption that it is true is usually called null hypothesis. The null hypothesis is a hypothesis which reflects no change or no difference. It is usually denoted by H_0

Alternative Hypothesis (H_1 **)**

The Alternative hypothesis is the statement which reflects the situation anticipated to be correct if the null hypothesis is wrong. It is usually denoted by H_1

For example:

If $H_0: \mu_1 = \mu_2$ (There is no difference between the means) then the formulated alternative hypothesis is $H_1: \mu_1 \neq \mu_2$

$$ie., either\ H_1: \mu_1 < \mu_2\ \ (or) H_1: \mu_1 > \mu_2$$

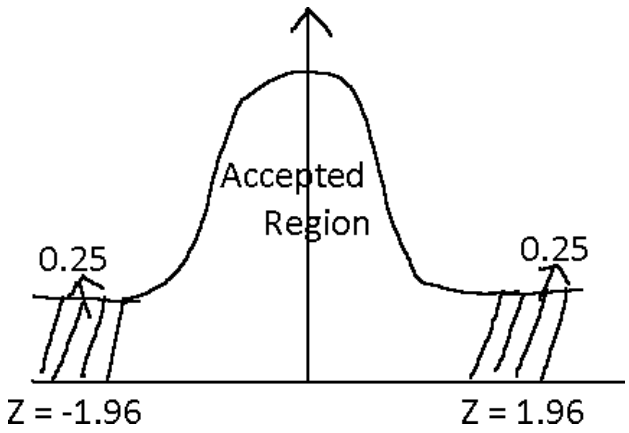
Level of significance

It is the probability level below which the null hypothesis is rejected. Generally, 5% and 1% level of significance are used.

Critical Region (or) Region of Rejection

The critical region of a test of statistical hypothesis is that region of the normal curve which corresponds to the rejection of null hypothesis.

The shaded portion in the following figure is the critical region which corresponds to 5% LOS



Critical values (or) significant values

The sample values of the statistic beyond which the null hypothesis will be rejected are called critical values or significant values

Level of significance

Types of test	1%	5%	10%
Two tailed test	2.58	1.96	1.645
One tailed test	2.33	1.645	1.28

Two tailed test and one-tailed tests:

When two tails of the sampling distribution of the normal curve are used, the relevant test is called two tailed test.

The alternative hypothesis $H_1: \mu_1 \neq \mu_2$ is taken in two tailed test $H_0: \mu_1 = \mu_2$

When only one tail of the sampling distribution of the normal curve is used, the test is described as one tail test $H_1: \mu_1 < \mu_2 \ (or) H_1: \mu_1 > \mu_2$

$$\left.\begin{matrix} H_0: \mu_1 = \mu_2 \\ H_1: \mu_1 \neq \mu_2 \end{matrix}\right\} \ two\ tailed\ test.$$

Type I and Type II Error

Type I Error: Rejection of null hypothesis when it is correct

Type II Error: Acceptance of null hypothesis when it is wrong

Procedure for testing Hypothesis:

1. Formulate H_0 and H_1
2. Choose the level of significance $\propto$
3. Compute the test statistic Z, using the data available in the problem
4. Pick out the critical value at $\propto$ % level say Z_α
5. conclusion: If $|Z| < Z_\alpha$, accept H_0 at $\propto$ % level. Otherwise reject H_0

Test of Hypothesis (Large Sample Tests)

Large sample tests (Test based in Normal distribution.)

Type - I: (Test of significance of single mean)

To test whether the difference between Population mean μ and sample mean x is significant or not and this sample comes from the normal population whose mean is μ or not.

H_0: μ = a specified value

H_1: $\mu \neq$ a specified value

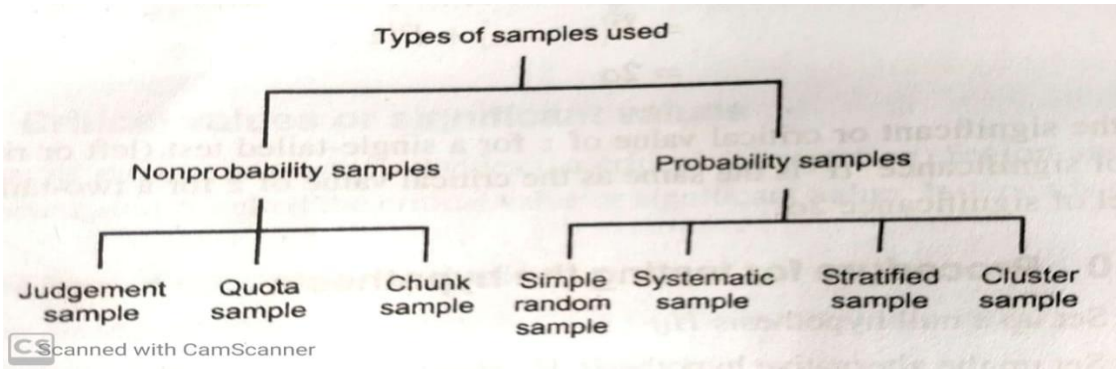
we choose $\propto = 0.05\ (5\%)\ (or) 0.01\ (1\%)$ as the Level of significance.

The test statistics

$$Z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

Types of Sampling Methods

There are two methods of selection samples from populations; the sample and the probability sample.



A nonprobability sample is one in which the items are chosen without regard to their probability of occurrence.

A probability sample is one in which the subjects of the sample are chosen on the basis of known population.

Probability sampling should be used whenever possible because it is the only method by which correct statistical inferences can be made from a sample. The four types of

probability samples most commonly used are simple random, systematic, stratified, and cluster. These sampling methods vary from one another in their cost, accuracy, and complexity.

Simple random sampling

A simple random sample is one in which every item from a population has the same chance of selection as every other item. In addition, every sample of a fixed size has the same chance of selection as every other sample of that size. Simple random sampling is the most elementary random sampling technique.

With simple random sampling, n is used to represent the sample size and N is used to represent the population size. Every item in the population is numbered from 1 to N. The chance that any particular member of the frame is selected on the first draw is $\frac{1}{N}$.

There are two basic methods by which sample are selected with replacement and without replacement.

Sampling with replacement means that after an item is selected, it is returned to the frame, where it has the same probability of being selected again. Sampling without replacement means that an item, once selected, is not returned to the frame, and, therefore, cannot be selected again.

Systematic sampling

In a systematic sample, the N items in the population are partitioned into k groups by dividing the size of the population by the desired sample size n. That is

$$k = \frac{N}{n}$$

Where k is rounded to the nearest integer. To obtain a systematic sample, the first item to be selected is chosen at random from the k items in the first partitioned group in the population, and the rest of the sample is obtained by selecting every k^{th} item thereafter from the entire population.

If you wanted to take a systematic sample of 40 from the population of N = 600 employees, the population of 600 would be partitioned into 600/40=15 groups. A random number would be selected from the first 15 items, and every fifteenth item after the first selection would be included in the sample For example, if the first number selected was 005, the subrequest selections would be 020, 035, 050, 065...

It will be seen that it is extremely convenient to select a sample in this way. The main point to note is that once the first unit in the sample is selected, the selection of subsequent units in the sample becomes obvious.

Stratified sampling

In a stratified sample, then N items in the population are first subdivided into separate subpopulations, or strata, according to some common characteristic. A simple random sample is conducted within each of the strata, and the results from the separate simple random samples are then combined. Such sampling methods are more efficient than either simple random sampling or systematic sampling because they ensure of items across the entire population, which ensures a greater precision in the estimates of underlying population parameters. The homogeneity of items within each stratum, when combined across strata, provides precision.

Imagine that you are working as a Marketing manager in a consumer product company. Suppose you are studying the customer attitudes towards your product in order to improve your sales. Suppose there are three typical cities that will influence your sales. Suppose the customers within each city are similar and between cities are vastly different. Selection of the customers for the study has to be a random sample of customers chosen from each city so that meaningful and reliable inferences can be drawn, which in turn will enable the marketing manager to develop suitable strategies. This is an example of stratified random sampling.

Cluster sampling

In a cluster sample, the N items in the population are divided into several clusters so that each cluster is representative of the entire population. A random sampling of cluster is then taken, and all items in each selected cluster are then studied Clusters can be naturally occurring designations such as countries, election districts, city blocks, apartment buildings, or families.

Cluster sampling methods can be more cost-effective than simple

random sampling methods, particularly if the population is spread over a wide geographic region. However, cluster sampling methods tend to be less efficient than either simple random sampling methods or stratified sampling methods and often re quire a larger overall sample size to obtain results. The steps involved in cluster sampling are:

1. Divide the population into a number of clusters based on geographic boundaries.

2. Select a random sample of clusters from this population of clusters.

3. Either measure all the units within the randomly chosen clusters or do further random sampling in each cluster.

Judgement sampling and Quota sampling

Under judgement sampling there is a deliberate selection based on the judgement of the person entrusted with the job. It is also called "Purposive Sampling". The worth of this method depends on the sampling design. The purpose of represes tativeness can also be realized if the selection is objective and proper judgement is exercised by an expert in the field who knows the limitations of such selection In quota sampling method, quotas are fixed according to the basic parameters of the population determined earlier and each field investigator is assigned with quotas of number of elementary units to be interviewed.

Sampling and Non-Sampling Errors

The errors involved in the collection, processing and analysis of data may be broadly classified under sampling and Non-sampling errors.

1. Sampling Errors

Sampling errors have their origin in sampling and arise due to the fact that only a part of the population (i.e. sample) has been used to estimate

populations parameters and draw inferences about them. Errors in sampling are primarily due to the following reasons.

(a)Faulty selection of the sample
(b) Substitution
(c) Faulty demarcation of sampling units
(d) Variability of the population

2. Non-Sampling Errors

As distinct from sampling errors which are due to the inductive process of drawing inference about the population on the basis of sample, the non sampling errors primarily arise at the stages of observation, approximation and processing of the data and are thus present in both the complete enumeration and the sample survey. Non-sampling errors can occur at every stage of the planning or execution of census or sample survey. However, a careful examination of the major phases of a survey indicate that some of the more important non-sampling errors arise from the following.

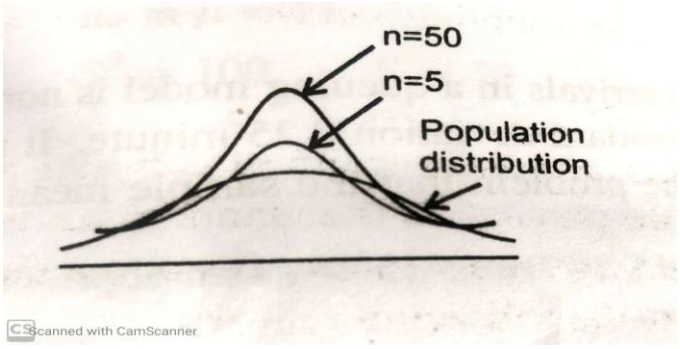
(a) Faulty planning
(b) Errors in response
(c) Non-response bias
(d) Errors in design of the survey
(e) Error in compilation
(f) Publication errors.

Now, we shall discuss about the various sampling distributions which are provide the necessary basis for drawing statistical Inferences and making decisions about the population.

Sampling Distribution of the Mean

The sampling distribution of $\bar{x}$ is the probability distribution of all possibles values of the sample mean $\bar{x}$.

If a population is normal, the sampling distribution of the mean ($\bar{x}$) is also normal for samples of all sizes, which can be seen from the following diagram.



1. The time between two arrivals in a queuing model is normally distributed with a mean 2 minutes and standard deviation 0.25 minute. It a random sample of size 36 is drawn, what is the problem that the sample mean will be greater than 2.1 minutes?

   Solution:

   Since the population is normally distributed, the sampling distribution of the sample mean will also follow a normal distribution with mean

   $$\mu_{\bar{x}} = \mu = 2$$

   And $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = 0.042$

   ∴the probability that the sample mean will be greater than 2.1 minutes given by $p[\bar{x} \geq 2.1]$

   this normal variate $\bar{x}$ must be converted into a standard normal variate given by, $p\left[\frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}} \geq \frac{2.1-\mu}{\frac{\sigma}{\sqrt{n}}}\right]$

   or , $p\left[Z \geq \frac{2.1-2}{0.042}\right] = p[Z \geq 2.38] = 0.5 - 0.4913 = 0.0087$

   only 0.87% of all possible sample of size n = 36, the sample mean will greater than 2.1 minutes.

2. A random sample of size 9 is obtained from a normal population with $\mu = 25$. If the sample variance is equal to 100, find the probability that the sample mean exceeds 31.2

Solution

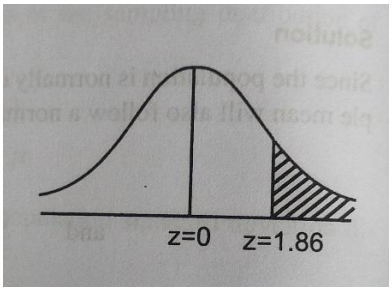
$$\mu_{\bar{x}} = \mu = 25$$

$S^2 = 100, S = 10$

$$\sigma_{\bar{x}} = \frac{S}{\sqrt{n}} = \frac{10}{\sqrt{9}} = 3.33$$

the probability that the sample mean will exceed 31.2 is

$p[\bar{x} > 31.2]$



$$p\left[\frac{\bar{x}-\mu}{\frac{S}{\sqrt{n}}} > \frac{31.2-25}{3.33}\right]$$

$$p[Z > 1.862] = 0.5 - 0.4686 = 0.0314.$$

3. The mean strength of a certain coming tool is 41.5 hrs with a standard deviation of 2.3 hrs. What is the probability that a random sample of size 50 drawn from this population will have a mean between 40.5 hrs and 42 hrs.

   Solution

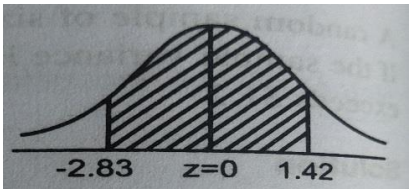
$$\mu_{\bar{x}} = \mu = 41.5$$

$$\sigma_{\bar{x}} = \frac{S}{\sqrt{n}} = \frac{2.5}{7.07} = 0.353$$

probability that the sample mean will be between 40.5 and 42 is

$$p[40.5 < \bar{x} < 42] = p\left[\frac{40.5 - 41.5}{0.353} < \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} < \frac{42 - 41.5}{0.353}\right]$$

$$= p[-2.83 < Z < 1.42]$$



$$= 0.4222 + 0.4977 = 0.9199$$

4. The diameter of component produced on a semi-automatic machine is known to be distributed normally with a mean of 10mm and a standard deviation of 0.1mm. If we pick up a random sample of size 5, what is the probability that the sample mean will be between 9.95mm and 10.05mm?

   Solution:

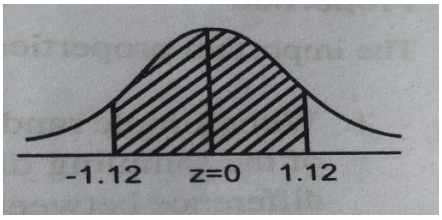
$$\mu_{\bar{x}} = \mu = 10$$

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{0.1}{\sqrt{5}}$$

probability that the sample mean will be between 9.95 and 10.05 is

$$p[9.95 < \bar{x} < 10.05] = p\left[\frac{9.95 - 10}{\frac{0.1}{\sqrt{5}}} < \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} < \frac{10.05 - 10}{\frac{0.1}{\sqrt{5}}}\right]$$

$$= p[-1.12 < Z < 1.12]$$



$$= 2p[0 < Z < 1.12]$$

$$= 2 * 0.3686 = 0.7372$$

5. For a particular brand of T.V. picture tube, it is known that the mean operating life of the tubes is 1000 hrs with a standard deviation of 250 hrs, what is the probability that the mean for a random sample of size 25 will be between 950 and 1050 hours?

   Solution:

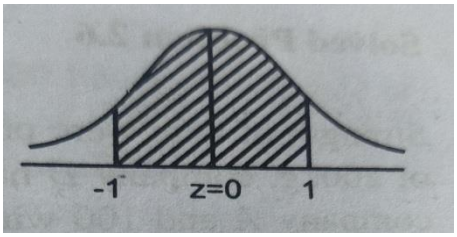
$$\mu_{\bar{x}} = \mu = 1000$$

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{250}{\sqrt{25}}$$

Required probability is

$$p[950 < \bar{x} < 1050] = p\left[\frac{950 - 1000}{\frac{250}{\sqrt{25}}} < \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} < \frac{1050 - 1000}{\frac{250}{\sqrt{25}}}\right]$$

$$= p[-1 < Z < 1]$$



$$= 2\,p[0 < Z < 1] = 2 * 0.3413 = 0.6826$$