Unit-4

Random Sample ; Sampling Distribution :

Concept of Population & Sample :

Population is a distinct group of individuals whether that group comprises a nation or a group of people with common characteristics.

Instead of examining the entire group, called population, we may examine only a small part of this population, which is called a **sample.**

Parameters ~~of Popn~~ and Statistics

- Population parameter : μ, σ, σ^2

  mean

  Standard deviation

  Variance.

- Sample parameter : $\bar{x}$, s, s^2

  Mean

  Standard deviation

  Variance.

Notation → Population (N)

Sample (n).

Sampling and Non sampling Error:

1> The difference between parameters and statistic is known as the sampling error.

2> This can't be controlled by human.

3> This errors come into existence when various sample ~~we~~ are taken.

(b) i> There are various systematic errors and random error that are not due to sampling.

2> These can control by human.

3> Non-sampling errors are systematic error.

Sampling Techeniques:

There are two types of sampling techeniques —

i> Probability sampling (unplaned)

(a) ~~Simple random sampling~~

(b)

ii> Non-probability sampling (Planned.

(a) ~~Snow fall sampling~~

(b)

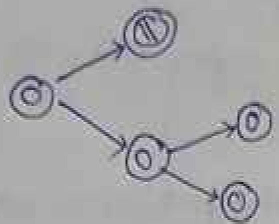
* Small sample.
* Biased (depend on research)
* qualitative research
* ~~at~~ known as non-random sampling

⊙ Types of non-probability sampling.

1) Snowball sampling –

↓

moving work chainwise or choosing sampling chainwise.



Ex: How many people of W.B comes from outside ~~india~~ W.B with covid 19.

One person can not find the total data.

One can distribute the work district wise, district can distribute this zila wise and zila also can distribute ~~using~~ area wise or block wise and so-on.

2. Convenience Sampling:

using some question-answer method to prepare sample.

3. Quota Sampling:

Restrict the data to collect sample.

Ex: behaviour of human of age above 60.

[Here the quota is the age]

4. Purposive sampling:

To find the people who take ~~loan~~ loan and not return in the time we need to take data from the manager of bank informing our purpose.

☑ Probability sampling.

ⓐ For large data.

<u>use</u> i) No biased

ii) Quantitative data

iii) Sample size large.

1) <u>Simple random sampling</u>:

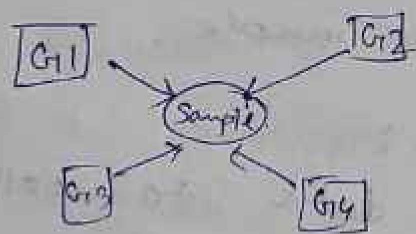
ⓐ each item has equal chance to be selected.

2) <u>Stratified Sampling</u>:

ⓐ Samples are drawn at random.

ⓑ the researcher divides the population into separate groups called strata.

ⓒ Then the probability sample is drawn from each group.



3) Systematic Sampling.

ⓐ this sample member from a larger population are selected according to a random starting point but with a fixed periodic interval.

ⓑ This interval is calculated by dividing the population size by desired sample size

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

4> Cluster Sampling:

Divide population into differen cluster and take sample from some cluster not all

5> Mutistage Sampling:

Taking sample by making populating small step by step.

<u>Note</u>: Parameter : Any term of population like mean, S.D. correlation coefficient

Statistics : Any term of sample is called statistics.

☑ Sampling Distribution:

Sample is a collection of data taken from the population.

Sampling is dividing a population into some sample of same size and choose a statistics from each sample. This is called sampling distribution.

⊛ <u>Sample distribution</u>: involves a singular sample from a population, and interpreting the data.

⊛ <u>Sampling distribution</u>: is a distribution of a statistics made from multiple simple Random samples drawn from a specific population.

<u>Note</u>: Sampling distribution of ~~a a~~ the sample mean is usually normal distribution.

Sampling Distribution of a statistics :

If we draw a sample of size n from a given finite population of size N, then the total number of possible sample is :

$$^{N}C_{n} = \frac{N!}{n!(N-n)!} = k \text{ (say)} .$$

For each of these k samples we can compute some statistic $t = t(x_1, x_2, \dots x_n)$ in particular the mean $\bar{x}$, the variance s^2, etc., as given below,

Sample Number	Statistics		
	t	$\bar{x}$	s^2
1	t_1	$\bar{x}_1$	s_1^2
2	t_2	$\bar{x}_2$	s_2^2
3	t_3	$\bar{x}_3$	s_3^2
⋮	⋮	⋮	⋮
k	t_k	$\bar{x}_k$	s_k^2

The mean and variance of the sampling distribution of the statistic t are given by

$$\bar{t} = \frac{1}{k}(t_1 + t_2 + \cdots + t_k) = \frac{1}{k}\sum_{i=1}^{k} t_i$$

and $$Var(t) = \frac{1}{k}\left[(t_1 - \bar{t})^2 + (t_2 - \bar{t})^2 + \cdots + (t_k - \bar{t})^2\right]$$

Standard Error:

The standard deviation of the sampling distribution of a statistics is known as its Standard Error. The standard errors of some of the well-known statistics, for large samples, are given below, where n the sample size, σ^2 the population variance, P the population proportion, and $Q = 1 - P$, n_1 and n_2 represent the sizes of two independent random samples respectively drawn from the given population(s):

Statistics		Standard Error
1. Sample mean : $\bar{x}$	$\longrightarrow$	$\sigma/\sqrt{n}$
2. Observed sample proportion 'p' :	$\longrightarrow$	$\sqrt{PQ/n}$
3. Sample s.d : s	$\longrightarrow$	$\sqrt{\sigma^2/2n}$
4. Sample variance : s^2	$\longrightarrow$	$\sigma^2\sqrt{2/n}$
5. Sample moment : μ_3	$\longrightarrow$	$\sigma^3\sqrt{96/n}$
6. Sample moment : μ_4	$\longrightarrow$	$\sigma^4\sqrt{96/n}$

Utility of S.E :

S.E plays a very important role in the large sample theory and forms the basis of the testing of hypothesis. If t is any statistic, then for large sample,

$$z = \frac{t - E(t)}{\sqrt{V(t)}} \sim N(0,1)$$

$$\Rightarrow Z = \frac{t - E(t)}{S.E(t)} \sim N(0,1) \text{, for large sample.}$$

Thus, if the discrepancy between the observed and expected value of statistics is greater than

i) The magnitude of the standard error gives an index of the precision of the estimate of the parameter. The reciprocal of the S.E is taken as the measure of reliability of the statistics.

ii) S.E enables us to determine the probable limits within which the population parameter may be expected to lie.

Test of significance :

~~Level of Significance →~~

Test of significance enable us to decide wheather our hypothesis is strong (or significant) or not.

Test of significance enable us to decide on the basis of the sample results, if

i) the deviation between the observed

sample value statistics and the hypothetical parameter value ~~or~~

or

ii) the deviation between two independent sample statistics, is significant or might be attributed to chance or fluctuation of sampling.

For large n almost all the distributions are very closely approximated by normal distribution, so we use normal test of significance for large sample.

Some other types of test of significance are. t-test, F-test, χ^2-test. Z-test.

Level of significance:

5% level of significance means and accuracy 95%. error in the data is 5%.

T-test (Student t-test):

Sample size < 30

Basic difference between the tests:

T-test	Z-test	χ^2-test.
* Sample size < 30	sample size ~~30<~~50 > 30	sample size ≥ 50
* One sample t-test → S.mean P.mean	Mainly used in tea driker	Mainly used in Genetic problem.
Two-sample t-test ↓ Sample mean		
* Paired t-test → D.S unpaired t-test → I.S.		

Difference Between Parametric and non-parametric test of significance:

Parametric	Non-parametric
i) Normal distribution [mean = median = mode]	i) Skewed distribution
ii) Quantitative	ii) Qualitative.
iii) Compear between SD / mean as statistice.	iii) use percentance or other statistics to compear.
Ex: t-test	Ex: x^2- test.

Null and alternative hypothesis:

Hypothesis is a definite statement about a population parameter.

A hypothesis which is usually a hypotesis of no difference is called null hypotesis and denoted by Ho.

According to Prof. R.A. Fisher, null hypothesis is a hypothesis which is tested for possible rejection under the assumption that it is true.

For example,

in case of single statistics :

H_0 : Sample statistics does not differ significantly from the hypothetical parameter value.

in case of two statistics :

H_0 : Sample statistics do not differ significantly.

If at some level of significance the deviation comes out to be significant, then the null hypothesis rejected and if the deviation is not significant then the null hypothesis is retained.

⊛ Alternative hypothesis: Any hypothesis which is complementary to the null hypothesis is called alternative hypothesis. and usually denoted by H_1

(The population has a specified mean μ_1)

If $H_0 : \mu = \mu_1$ then

i) $H_1 : \mu \neq \mu_1$ → two tailed

ii) $H_1 : \mu > \mu_1$ → Right tailed

iii) $H_1 : \mu < \mu_1$ → Left tailed.

i), ii), iii) all are alternative hypothesis.

* Type of error :

Type I : Reject H_0 when it is true.

Type II : Accept H_0 when it is wrong.

$P\{\text{Reject } H_0 \mid H_0\}$ = Probability of type I error $= \alpha$

$P\{\text{Accept } H_0 \mid H_1\}$ = Probability of type II error $= \beta$

α is known as producer's risk
β is known as customer's risk.

☑ critical region :

A region in the sample space S which amounts to rejection of H_0 is termed as critical region.

If w is the critical region and t is the value of the statistics on a random sample of size n, then,

$P\{t \in w \mid H_0\} = \alpha$ and $P\{t \in \bar{w} \mid H_1\} = \beta$.

$\bar{w}$ is alled acceptance region.

$w \cup \bar{w} = S$ (sample space)

$w \cap \bar{w} = \emptyset$.

⊛ Level of significance means the size of type I error, i.e α.

The level of significance is always fixed in advance before collecting a sample information.

14·4·5. Critical Values or Significant Values. The value of test statistic which separates the critical (or rejection) region and the acceptance region is called the *critical value* or *significant value*. It depends upon :

(*i*) The level of significance used, and

(*ii*) The alternative hypothesis, whether it is two-tailed or single-tailed.

As has been pointed out earlier, for large samples, the standardised variable corresponding to the statistic t, *viz.*,

$$Z = \frac{t - E(t)}{S.E(t)} \sim N(0, 1), \qquad \ldots (*)$$

asymptotically as $n \to \infty$. The value of Z given by (*) under the null hypothesis is known as *test statistic*. The critical value of the test statistic at level of significance α for a two-tailed test is given by z_α, where z_α is determined by the equation :

$$P(|Z| > z_\alpha) = \alpha \qquad \ldots (14\cdot 2c)$$

i.e., z_α is the value so that the total area of the critical region on both tails is α. Since normal probability curve is a symmetrical curve, from (14·2*c*), we get

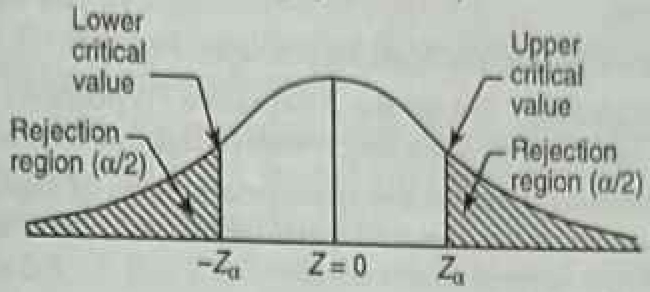
$$P(Z > z_\alpha) + P(Z < -z_\alpha) = \alpha \;\Rightarrow\; P(Z > z_\alpha) + P(Z > z_\alpha) = \alpha \qquad \text{[By symmetry]}$$

$$\Rightarrow \quad 2P(Z > z_\alpha) = \alpha, \;\Rightarrow\; P(Z > z_\alpha) = \alpha/2$$

In other words, the area of each tail is $\alpha/2$. Thus z_α is the value such that area to the right of z_α is $\alpha/2$ and to the left of $(-z_\alpha)$ is $\alpha/2$, as shown in the following diagram :

TWO-TAILED TEST

(*Level of Significance* 'α')



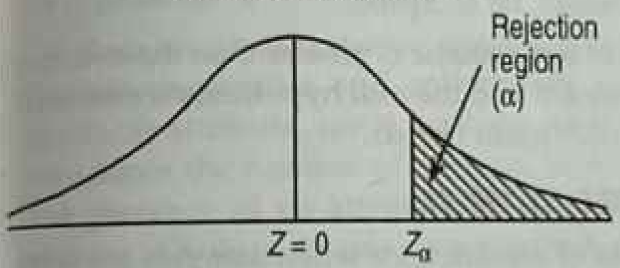
In case of single-tail alternative, the critical value z_α is determined so that total area to the right of it (for right-tailed test) is α and for left-tailed test the total area to the left of $(-z_\alpha)$ is α (See diagrams below), *i.e.*,

For Right-tailed Test : $P(Z > z_\alpha) = \alpha$ ... (14·2*d*)

For Left-tailed Test : $P(Z < -z_\alpha) = \alpha$ ... (14·2*e*)

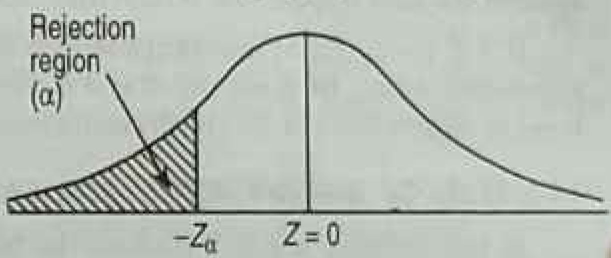
RIGHT-TAILED TEST

(*Level of Significance* 'α')



LEFT-TAILED TEST

(*Level of Significance* 'α')



Thus *the significant* or *critical value of Z for a single-tailed test (left or right) at level of significance 'α' is same as the critical value of Z for a two-tailed test at level of significance '2α'.*

We give below, the critical values of Z at commonly used levels of significance for both two-tailed and single-tailed tests. These values have been obtained from equations (14·2*c*), (14·2*d*), (14·2*e*), on using the Normal Probability Tables as explained in § 14·6.

Critical value (z_α)	*Level of significance* (α) 1%	5%	10%
Two-tailed test	$\lvert Z_\alpha \rvert = 2·58$	$\lvert Z_\alpha \rvert = 1·96$	$\lvert Z_\alpha \rvert = 1·645$
Right-tailed test	$Z_\alpha = 2·33$	$Z_\alpha = 1·645$	$Z_\alpha = 1·28$
Left-tailed test	$Z_\alpha = -2·33$	$Z_\alpha = -1·645$	$Z_\alpha = -1·28$

Remark. If n is small, then the sampling distribution of the test statistic Z will not be normal and in the that case we can't use the above significant values which have been obtained from normal probability curves. In this case, *viz.*, n small (usually less than 30), we use the