Unit - I

Inferential Statistics - I

Inferential Statistics :-

* Allows to make predictions from that data.

* Can take data from samples and make generalizations about population.

In other words we can say,

* It is used to draw conclusions about a population by examining the sample.

Example :- 50% of people dislike shopping malls.

Populations :-

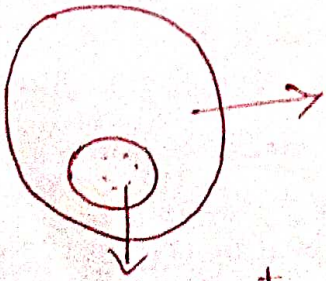
*A population is the set of all items or individuals of interest.

example:- All parts produced today.

Sample :-

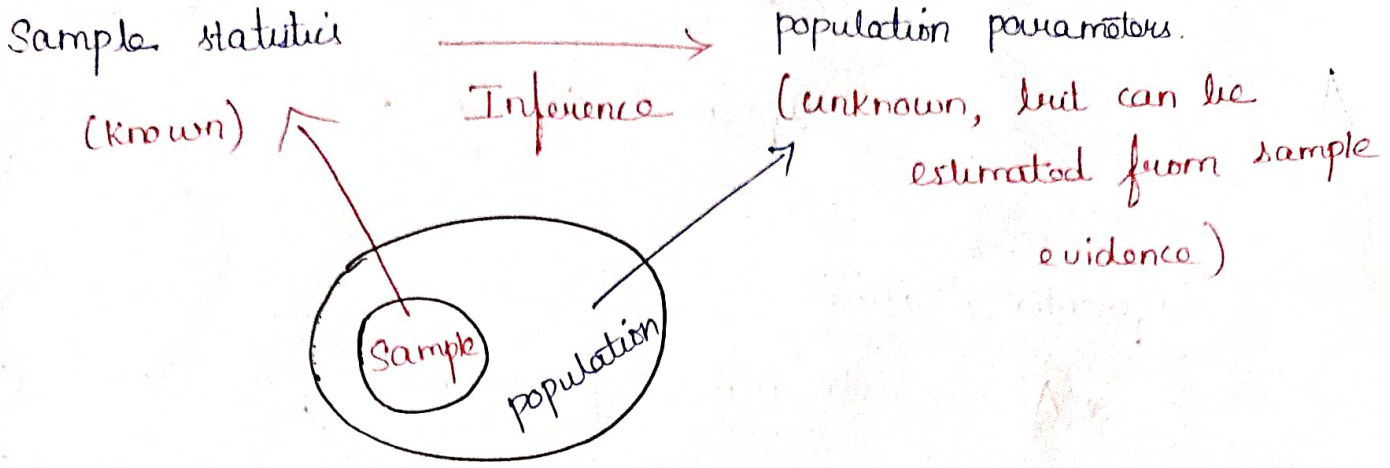
* A sample is a subset of the population.

example:- A few parts selected for destructive testing.



all parts produced today (population)

all parts are produced today (Sample).



Inferential statistics includes;

* estimation of parameters.
* Hypothesis testing.

Estimation of parameter:-

* Taking a statistic from your/the sample data (Sample mean) and using it to say something about a population parameter (ie., population mean).

Hypothesis testing:-

* Hypothesis testing in statistics is a way for to test the results of a survey or experiment to see if if (data) have meaningful results.

Random Sampling:-

* Every unit of the population has the same probability of being included in the sample.
* A chance mechanism is used in the selection

process.

*eliminates bias in the selection process.

(or)

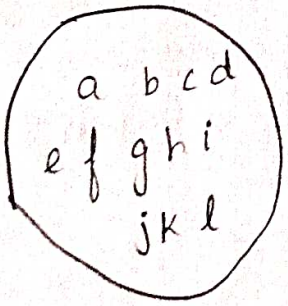
In other words,

*it is the part of the sampling technique in which each sample has an equal probability of being chosen.

* Random Sampling is also known as probability sampling.

* If for some reasons, the sample does not represent the population, the variation is called a Sampling error.

• Sampling error → a statistical error that occurs when an analyst does not select a sample that represents the entire population of data.



population

randomly selected (or) biased Sample

→ Sampling error occurs

Random Sampling Techniques :-

4 Techniques :-

1) Simple Random Sampling

2) Stratified Random Sampling.

3) Cluster Random Sampling.

4) Systematic Random Sampling.

1) Simple Random Sampling :-

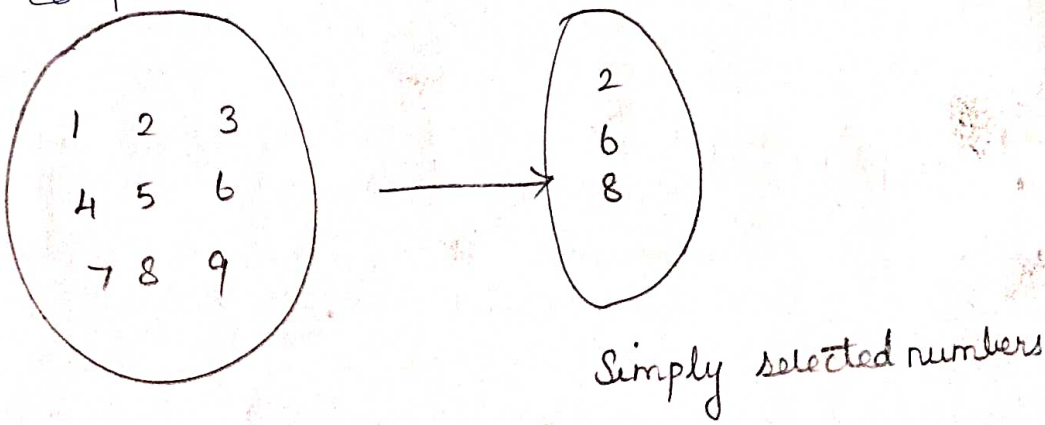
* every object in the population has an equal chance of being selected.

* Objects are selected independently.

* Samples can be obtained from a table of random numbers or computer random number generators.

* It is the ideal against which other sample methods are compared.

Ex:-



Simple Random Numbers

2) Stratified Random Sampling :-

* It starts off by dividing a populations into groups with Similar attributes.

* Then a random sample is taken from each group.

* It has the potential for reducing the sampling error.

Stratified random Sampling includes,

- proportionate
- disproportionate.

proportionate :-

→ the percentage of these samples taken from each stratum is proportionate to the percentage that each stratum is within the population.

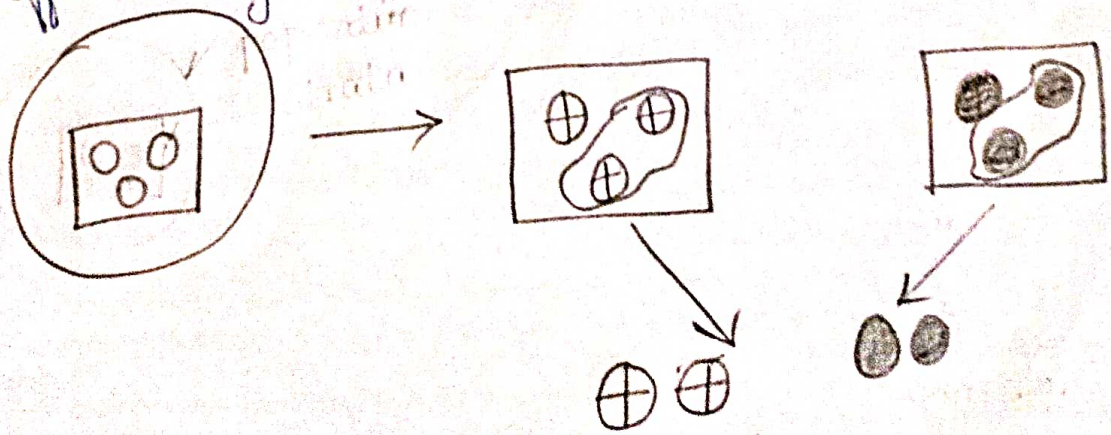
disproportionate :-

→ the proportions of the strata within the sample are different than the proportions of the strata within the population.

[Strata → population is divided into non-overlapping subpopulations]

[stratum → a random sample is selected from each stratum]

* This stratified random sampling method is used to ensure that different segments in a population are equally represented

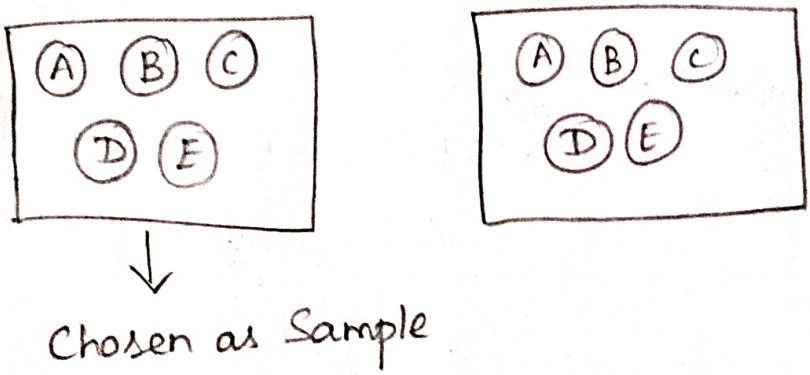


3) Cluster random Sampling :-

* Starts by dividing a population into groups or clusters.

* each cluster is a miniature of the population.

* a subset of cluster is selected randomly for the sample.



Example :-

If an elementary school had five different Grade. Cluster random sampling might be used and only one class would be chosen as sample.

Merits :-

* most economical form of Sampling.
* larger sample for a similar fixed cost.
* less time for listing and implementation.
* reduce travel and other administrative costs.

Demerits :-

* May not reflect the diversity of the community.
* Standard errors are of the estimates are high,

*Comparing to other sampling designs with same sample size.

4) Systematic Random Sampling:-

*It is a very common technique.

*In this, the population elements are an ordered sequence (at least, conceptually).

*The first sample element is selected randomly from the first 'k' population elements.

*Thereafter the sample elements are selected at a constant interval, k from the ordered sequence frame.

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

n = Sample Size

N = population Size.

k = size of selection interval.

Example:-

id	Name
1	A
2	B
3	C
4	D
5	E
6	F

N = 6

n = 2

$K = \frac{6}{2}$

K = 3

Merits:-

*Simple and convenient

*less time consuming.

Demerits:-

*population with hidden periodicities.

Merits for Simple random sampling:-

* No personal bias
* Sample more representative of population.
* accuracy can be assessed as sampling errors follow principals of chance.

Demerits for Simple random sampling:-

* requires completely catalogued universe.
* Cases too widely dispersed - more time and cost.

Merits for Stratified random Sampling:-

* More representative.
* Greater accuracy.
* Greater geographical concentration.

Demerits for stratified random Sampling:-

* atmost care in dividing strata.
* skilled Sampling supervisors.
* Cost per observation may be high.

Probability and statistics Sampling distribution :-

A sampling distribution is a probability distribution of a statistic obtained from a larger number of samples drawn from a specific population.

The sampling distribution of a given population is the distribution of frequencies of a range of different outcomes that could possibly occur for a statistic of a population.

Mean :-

$$\bar{x} = \frac{(x_1 + x_2 + \dots + x_n)}{n}$$

Median :-

Total number of value odd,

$$\text{Median} = \left(\frac{n+1}{2}\right)^{th} \text{term.}$$

Total number of value even,

$$\text{Median} = \left(\frac{\left(\frac{n}{2}\right)^{th} \text{term} + \left(\frac{n}{2}+1\right)^{th} \text{term}}{2}\right)^{th} \text{term}$$

Mode :- Value appears most. (mode → 2)

2, 4, 6, 8, 2, 1

Functions of sampling distribution :-

* Mean
* range

* mean absolute value of the deviation from the mean.
* SD of the sample.
* Unbiased estimate of the sample.
* Variance of the sample.

Probability	Statistics
* Theoretical ("pure Math")	* Applicable ("Applied Math")
* deduction (Rules → Data)	* induction (data → rule)
* ideal	* real
* Certainty	* estimation
* predict	* Summarize
* future	* past.

Creating a sampling distribution :-

Example :- Assume there is a population. size $N = 4$

random variable x is the age of individuals,

Values of $x = 18, 20, 22, 24$ (years)

$$\mu = \frac{\sum x_i}{N}$$

$$\mu = \frac{18 + 20 + 22 + 24}{4}$$

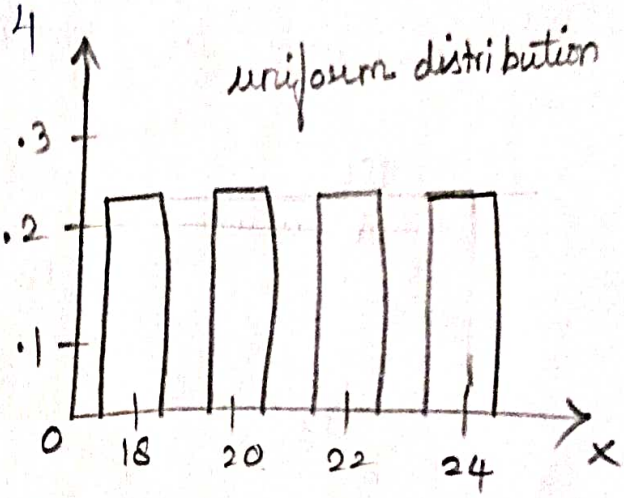
$$\mu = 21$$

$$\sigma = \sqrt{\frac{\Sigma(x_i - \mu)^2}{N}}$$

$$\sigma = \sqrt{\Sigma(18-21)^2 + (20-21)^2 + (22-21)^2 + (24-21)^2}$$

$$\sigma = \sqrt{\frac{20}{14}} \Rightarrow \sqrt{5}$$

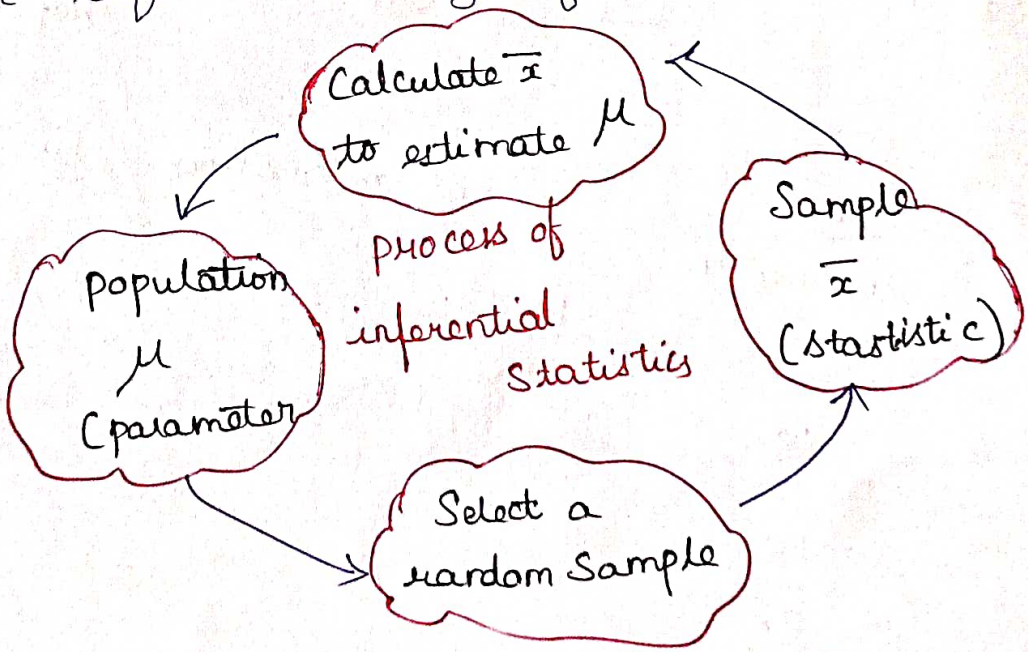
$$\sigma = 2.236$$



Sampling distribution of all sample means:-

sampling distribution of $\bar{x}$

proper analysis and interpretation of a sample statistic requires knowledge of its distribution.



[Mean of all sample means $\mu_{\bar{x}}$]

Mean of the sampling distribution of the mean always equals the mean of the population.

[μ of Sampling distribution = μ of population

$(\mu_{\bar{x}} = \mu)$

Sample means distribution:-

Consider all possible samples of size $n=2$.

1st observation	2nd observation			
	18	20	22	24
18	(18,18)	(18,20)	(18,22)	(18,24)
20	(20,18)	(20,20)	(20,22)	(20,24)
22	(22,18)	(22,20)	(22,22)	(22,24)
24	(24,18)	(24,20)	(24,22)	(24,24)

→ 16 possible samples (samples with replacement).

1st observation	2nd observation:-			
	18	20	22	24
18	18	19	20	21
20	19	20	21	22
22	20	21	22	23
24	21	22	23	24

→ 16 sample means

Mean $\mu_{\bar{x}} = \frac{\sum \bar{x}_i}{N}$

$$\mu_{\bar{x}} = \frac{18+19+19+20+\dots+24}{16}$$

$$\mu_{\bar{x}} = \frac{336}{16}$$

$$\boxed{\mu_{\bar{x}} = 21}$$

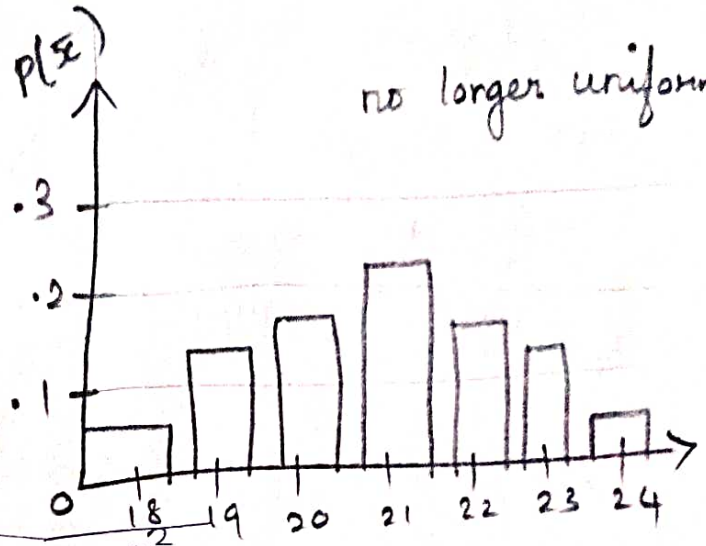
$$\sigma_{\bar{x}} = \sqrt{\frac{\sum(\bar{x}_i - \mu_{\bar{x}})^2}{N}}$$

$$= \sqrt{\frac{(18-21)^2+(19-21)^2+\dots+(24-21)}{16}}$$

$$= \sqrt{\frac{40}{16}}$$

$$= \sqrt{2.5}$$

$$\boxed{\sigma_{\bar{x}} = 1.58}$$



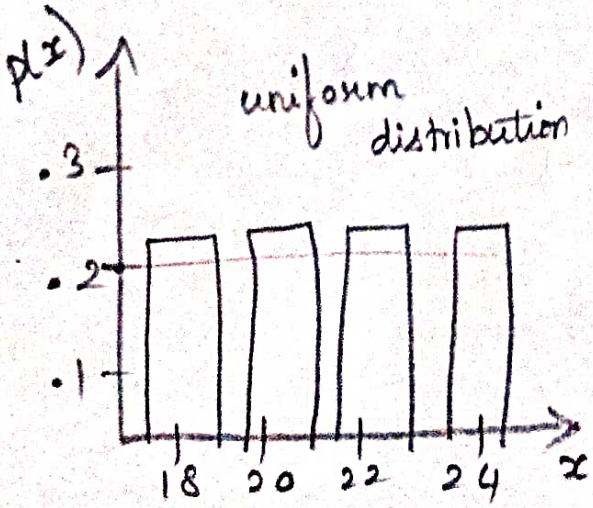
no longer uniform

Comparing the population distribution to / with the sample means distribution :-

population

$N = 4$

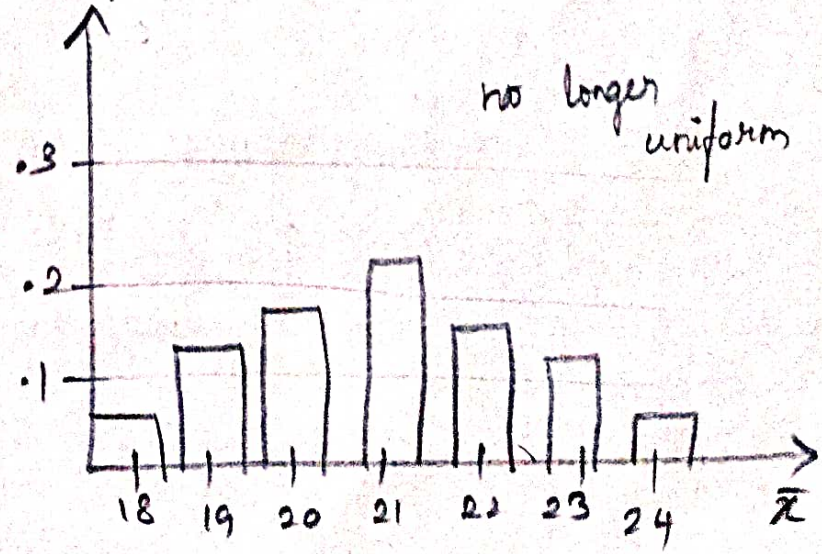
$\mu = 21$; $\sigma = 2.236$



Sample means distribution

$n = 2$

$\mu_{\bar{x}} = 21$; $\sigma_{\bar{x}} = 1.58$



Standard error of the mean:-

Different samples of the same size from the same population will yield different sample means.

A measure of the variability in the mean from sample to sample is given by the standard error of the mean:

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

Standard error of the mean decreases as the sample size increases.

If the population is Normal:-

$$\mu_{\bar{x}} = \mu \quad \text{and} \quad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

If the sample size n is not large relative to the population size N, then

$$\mu_{\bar{x}} = \mu \quad \text{and} \quad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$$

Z-value for the sampling distribution of the mean:-

$$Z = \frac{(\bar{x} - \mu)}{\sigma_{\bar{x}}}$$

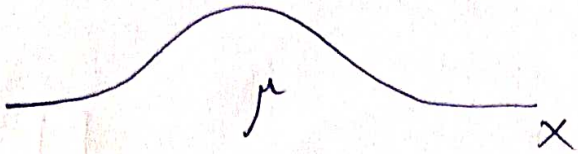
$\bar{x}$ → Sample mean

μ → population Mean

$\sigma_{\bar{x}}$ → Standard error of the mean.

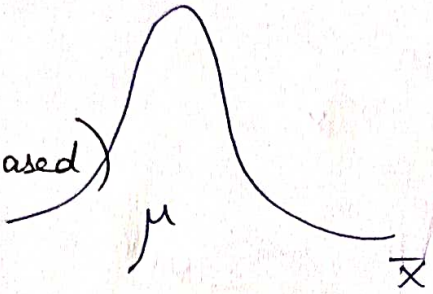
Sampling distribution properties :-

normal population distribution



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

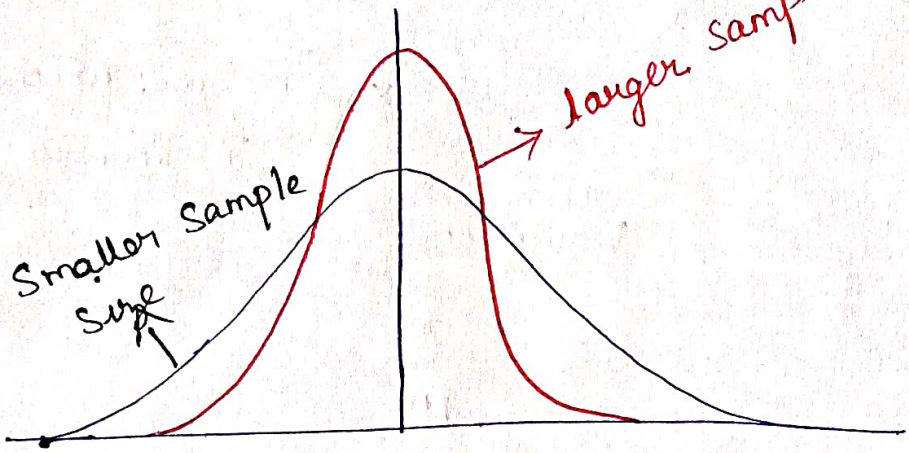
(i.e., $\bar{x}$ is unbiased)



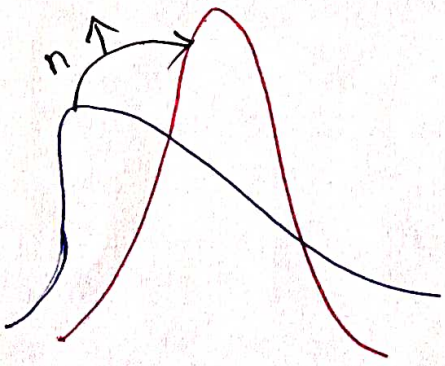
normal sampling distribution (has the same mean)

For sampling with replacement :-

As n increases, $\sigma_{\bar{x}}$ decreases



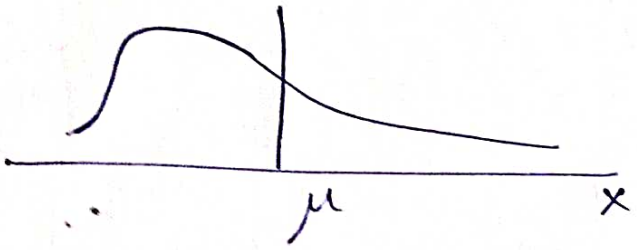
If the population is not normal apply central limit Theorem :-



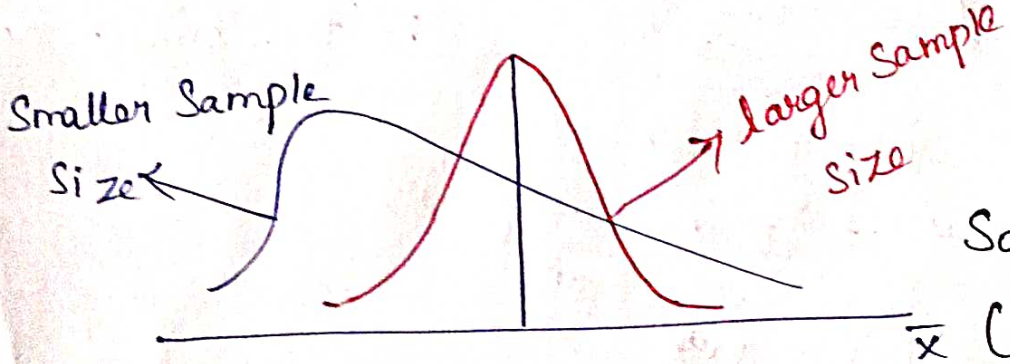
As the sample size gets larger enought, the sampling distribution becomes almost normal regardless of shape of population.

If the population is not normal :-

Sampling distribution properties :-



population distribution



Sampling distribution (becomes normal as 'n' increases)

Central Tendency

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

Variation

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

Hypothesis testing :-

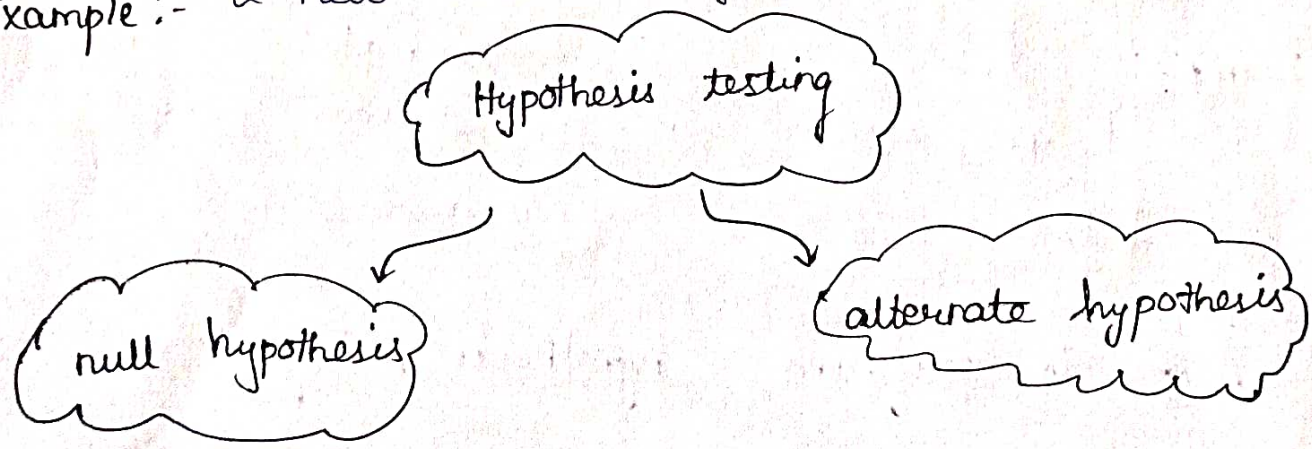
hypothesis :-

Claim or statement about a property of a population (in our case, about the mean or a proportion of the population).

Hypothesis test is a standard procedure for testing a claim or statement about a property of a population.

Hypothesis testing can be used to determine whether a statement about the value of a population parameter should or should not be rejected.

Example :- a new medicine you think might work.



Null hypothesis :-

predicts no effect or no relationship between variables.

The Sample observations results purely from Chance denoted by Ho.

Alternate hypothesis:-

States research prediction of an effect or relationship.

Sample variables are influenced by some non-random cause.

denoted by H_A.

Difference between Ho and H_A:-

Ho (null hypothesis)	H_A (alternate hypothesis)
Statement that you try to find evidence against. Statement of no difference Statement of no effect Example:- I am going to win $ 1,000	statement that you try to find evidence for Statement of no difference Statement of no effect Example:- I am going to win $1,000 or more.

Steps to perform hypothesis testing:-

1) State the hypothesis (Ho, Hi) about the population.

2) Set the Significance level, criteria for decision.

level of Significance → it mean the degree of Significance in which we accept or reject the null hypothesis.

Since in most of the experiments 100%, accuracy is not possible for accepting or rejecting a hypothesis, so we use the level of Significance.

denoted by α (alpha)

3) Compute the test statistics.

4) Make a decision.

Just for reference :-

Hypothesis testing :-

2 procedures

1) Z-test (unit I, II)

2) t-test (unit III)

1) Z-test

steps to perform

1-tail

2-tail

α

calculations

H_0, H_A, etc,.

2) t-test

steps to perform

1-tail

2-tail

α

calculations

H_0, H_A etc.,-

Z-test :-

Any statistical test for which the distribution of the test statistics under the null hypothesis can be approximated by a normal

distribution.

It is used to determine whether two population means are different when the variances are known and the sample size is large.

Example:-

The Sample Size is greater than or equal to 30.

Steps to be performed in Z-test :-

1) State the H_0 and H_A.

2) Choose an α level.

3) find the Critical Value of z in a z-table.

4) Calculate the z-test statistic.

5) Compare the test statistic to the critical z value and decide if you should support or reject the null hypothesis.

Types of Z -test :-

* one - sample (mean)
  - left tail
  - right tail
  - 2-tail
* Two - sample (mean)
* One - Sample (proportion)
* Two - Sample (proportion)

Conditions for left and right tail:

Left -tail:

$H_0 : \mu = \mu_0$

$H_1 : \mu < \mu_0$

$Z < -1.645$; H_0 is rejected

$Z > -1.645$, H_0 is accepted

Right tail :-

$H_0 : \mu = \mu_0$

$H_1 : \mu > \mu_0$

$Z < 1.645$, H_0 is accepted

$Z > 1.645$, H_0 is rejected.

Two-tail :-

$H_0 : \mu = \mu_0$

$H_1 = \mu \neq \mu_0$

Critical Values	Level of Significance, α		
	1%	5%	10%
Two-tail	$\|Z\| = 2.58$	$\|Z\| = 1.96$	$\|Z\| = 1.645$
right-tail	$Z = 2.33$	$Z = 1.645$	$Z = 1.28$
left-tail	$Z = -2.33$	$Z = -1.645$	$Z = -1.28$

Left tail Example :-

1) A simple Sample of heights of 6,400 englishmen has a mean of 170 inches and a SD of 6.4 inches, while a simple sample of heights of 1600 australians has a mean of 172 inches and a SD of 6.3 inches. Do the data indicate that australians are on the average taller than englishmen? level of Significance 5%.

Given:-

$n_1 = 6400$ $\quad \bar{x}_1 = 170$ $\quad S_1 = 6.4$

$n_2 = 1600$ $\quad \bar{x}_2 = 172$ $\quad S_2 = 6.3$

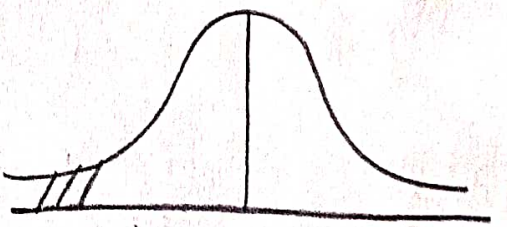
Sol:-

$H_0 : \mu_1 = \mu_2$

$H_1 = \mu_1 < \mu_2$

$\alpha = 5\%$

$-Z_\alpha = -1.645$



$\alpha = 5\%$

$-Z_\alpha = -1.645$

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

$$Z = \frac{170 - 172}{\sqrt{\frac{(6.4)^2}{6400} + \frac{(6.3)^2}{1600}}}$$

$$Z = -11.32$$

$-1.645 \neq 11.32$

H_0 is rejected.

Right tail example:-

A principal claims that the students in his school are above average intelligence. A random sample of 30 IQ scores have a mean score of 112.5. Is there sufficient evidence to support the principal's claim? The mean population IQ is 100 with SD of 15? level of significance 5%?

Given:-

$\bar{x} = 112.5 \quad ; \sigma = 15 \quad ; n = 30$

$\alpha = 5\%$

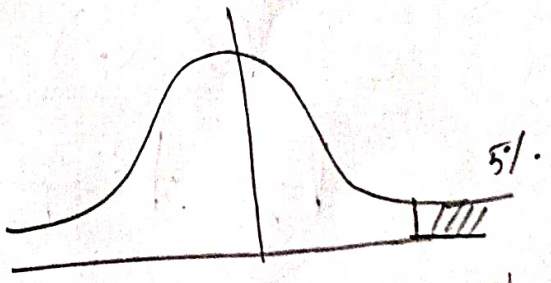
Sol:-

$H_0 : \mu = \mu_0$

$H_1 \quad \mu > \mu_0$

$H_0 = \mu = 100$

$H_1 = \mu > 100$



$\alpha = 5\%$

$Z\alpha = 1.645$

$\alpha = 5\%$

$Z\alpha = 1.645$

$$Z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

$$Z = \frac{112.5 - 100}{15/\sqrt{30}}$$

$Z = 4.56$

$Z = 4.56 > 1.65$

Ho is rejected.

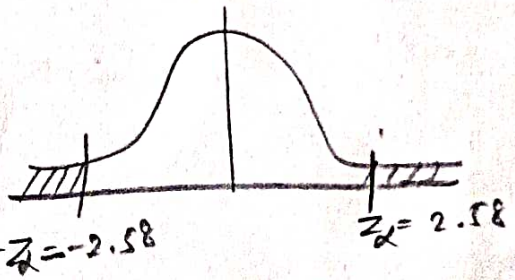
Two-tailed test example:-

In a random sample of size 500, the mean is found to be 20. In another independent sample of size 400, the mean is 15. Could the samples have been drawn from sample population with S.D 4? $\alpha = 1\%$.

Given:- $n_1 = 500$, $\bar{x}_1 = 20$

$n_2 = 400$, $\bar{x}_2 = 15$

$\sigma_1 = \sigma_2 = 4$



Sol:-

Ho : $\mu_1 = \mu_2$

H1 : $\mu_1 \neq \mu_2$

$\alpha = 1\%$. $-Z\alpha = -2.58$ and $Z_\alpha = 2.58$

$$Z = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{20 - 15}{\sqrt{\frac{(4)^2}{500} + \frac{(4)^2}{400}}}$$

$Z = 18.66$

$-Z_\alpha < Z < Z_\alpha$

$-2.58 < 18.66 \not< 2.58$

We reject Ho.