L=I

12-10-2020. "Probability theory" Monday.

Sampling theory:-

⇒ Sampling a specific thing used in every fields. Impossible to get complete information.

⇒ less resource, less time, stock of information gathered from population.

"In order to collect information from units on the basis of this information."

⇒ Sampling is statistical technique.

⇒ Sampling is a process used in statistical analysis in which a predetermined number of observations are taken from a larger population.

⇒ Sampling theory → field of statistics that is involved with the collection, analysis & interpretation of data gathered from random samples of a population under study.

Scanned with CamScanner

13-10-2020 L=2 Tuesday-

Primary data:- → (collecting direct information)

Examples:- (questioning & answering by mail or by messaging)

Secondary data:-

Examples:-

① Official publication of Federal trade & local govt.

② Reports of committees & commission

③ Data released by magazines, journals & newspapers.

④ Publications of different internation organizations like united nation organization (UNO), world bank, international monitory fund, international labor organizations, food & agriculture organizations.

⇒ (How inference is made)

↳ (We gather information from the population)

(Get exact information)

(We just make an idea about anything)

For example, media collects information about the votes being casted by the people.

Scanned with CamScanner

They collect this information from the people & then make an inference that which party will win & upto how much margin.

Element:-

→ (registered voters)

Population:- (or universe)

→ (complete list of voters)

↙ (statistical population)

→ **Types:-** (finite or infinite)

Sample:-

→ (Subset of population)

For example we have a list of houses

House → Sampling unit. List of houses → Sampling frame

Members in house → elements

When we go from individual to generalize
→ (element to population)

From generalize to individual
→ (population to element)

Scanned with CamScanner

14-10-2020.

L=3

Wednesday

Census → (پوری آبادی)

n → denotes samples

N → population size

$\frac{n}{N}$ → sampling fraction

Probability:- → (dealing with uncertain situations)

→ (how likely an event to be occur)

→ "Probability is a branch of mathematics concerning numerical description of how likely an event to be occur or how likely it is that a proposition is true."

⇒ '0' indicates impossible events.

⇒ '1' indicates certain events.

Probability sampling:- → (also known as Random sampling)

⇒ Write the names of probability sampling procedures-

① Simple random sample.
② Stratified random sample. → (each unit has probability but not equal probability)
③ Cluster sample.
④ Systematic sample.
⑤ Multistage sample.
⑥ Multiphase sample.
⑦ Sequential sampling

(Q:- On which type of situation you use probability sampling or non-probability sampling)

Scanned with CamScanner

(on the basis of personal opinion)

Non-probability sampling: → (alternative of probability sampling)

↳ Convenient sampling → (We take a spoon from our dish & judge its taste)

→ Purposive Sampling → (judgement sampling)

↳ (having special skills)

19-10-2020 L = 4 Monday

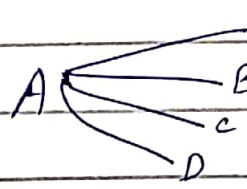
"With & without replacement sampling."

With replacement samplings-

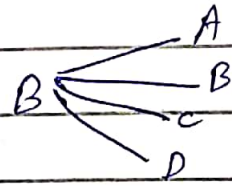
(N^n)

For example,

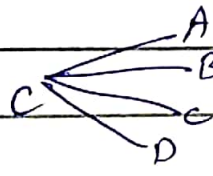
A, B, C, D → $4^2 = 16$



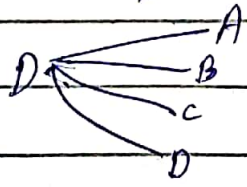
✓A,A
✓A,B
✓A,C
✓A,D



✓B,A
✓B,B
✓B,C
✓B,D

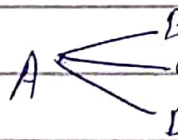


✓C,A
✓C,B
✓C,C
✓C,D

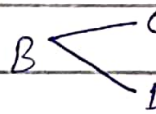


✓D,A
✓D,B
✓D,C
✓D,D

Without replacement samplings-



✓A,B
✓A,C
✓A,D



✓B,C
✓B,D

C — D ✓C,D

Scanned with CamScanner

Types of Probability sampling:-

(1) Simple Random sampling:-

For example, our teacher has the list of our class. The can call anyone of the students from the list.

(2) Stratified Random sampling:- → (also called random quota sampling)

We divide population on the basis of category in different groups.
One individual is involved in one group only so that the group will not overlap but represent the entire population.
We must have a variable in stratified random sampling. It may be variable of sex, age, status, income.
Further, we use simple random sampling in stratified random sampling.
When we have relevant information about an individual then we can move towards stratified random sampling.
When we have complete population available & we also know a characteristic related to population, so we use stratified sampling).

(3) Random cluster sampling:-

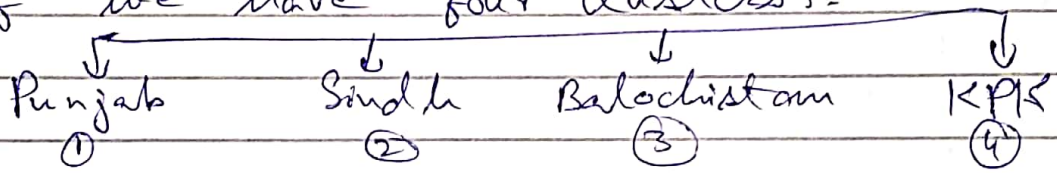
For Examples- I make a group of one block.
Departments in this block

↓ Mathematics ↓ Statistics ↓ Geographic ↓ English

Scanned with CamScanner

When we have to go far apart where population is scattered & we have less cost, then we are restricted to add all geographically related people in one group (either male or female).
Another examples.

Of we have four clusters:-



And we have cost to study only one cluster out of these four. So we draw one sample from these four samples.
(When we do not have enough cost & the list of population is not completely available then we use cluster sampling).

(4) Systematic sampling:-

We move according to a system.
We select one unit randomly & then we select a number. For example, our class strength is 34. The teacher chooses roll no. 12 then she will move towards the next 12th roll no. which is 24. (every 12th)
When we have population in serial numbers, then we use systematic sampling.
The chances of 1st person to be chosen is random.

Scanned with CamScanner

And the next ones are selected according to a system.

(5) Multiphase sampling:-

I have a population & I am bound to use stratified random sampling but I have no information about population then I'll use multi-phase sampling.

If I have to get any information but I am not doing sampling. My information which I want is that I have to do grouping. I may be able to get some information that how I divide population into groups then I'll use multi-phase sampling.

In multi-phase sampling, at first we have ~~double~~ two-phase sampling. We take information about two groups.

We make different groups..

If we get information for two groups, then it is two-phase sampling & if we get information for more than two groups, then it is multi-phase sampling.

(6) Multistage sampling:-

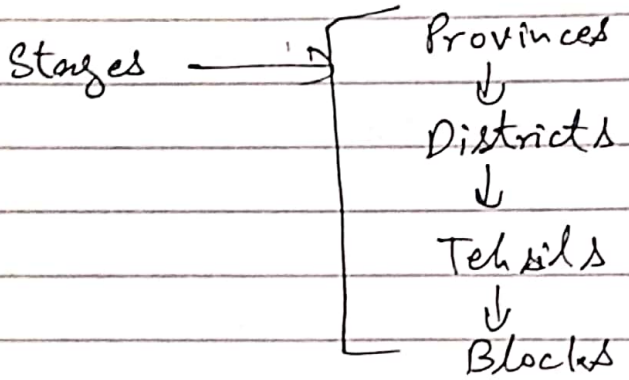
We select two geographical clusters.

We choose two groups.

KPK & Punjab

Scanned with CamScanner

Then we do further sampling rather than complete study of all individuals in sample.
(In sample of Punjab & in sample of KPK)
Then we make different stages for sampling
If we didn't need complete information but a part of information, then we'll use multistage sampling.



Auxiliary information → Rather than a sample, I need such an information which will help me in collecting sampling or it is helpful to me in any stage then we call it auxiliary information or extra information. In this case, we'll use multi-phase sampling.

(If we want accurate information, we use probability sampling.)

(Sampling bias means our sample is true representative of population.)

Scanned with CamScanner

20-10-2020. L=5 Tuesday.

"Non-Probability Sampling."

→ Subjective manner in non-probability sampling.
→ Objective manner in probability sampling.
(In medical, we use probability sampling because this matter is about life or death.)

Non-probability sampling is based on the subjective judgement of the researcher. The researcher should be efficient. The selection of candidates entirely depends on the researchers.

Pilot surveys:-

When we need a piece of information earlier than census, then we conduct a small survey called pilot survey.

Types of Non-probability sampling:-

(1) Convenience Samplings-

We use this sampling when we have low cost or no cost available. When accuracy is not required we use convenience sampling.

(2) Consecutive samplings → (involves steps)

Researcher selects a small number of candidates.

Scanned with CamScanner

(3) Quota sampling:-

We make groups.
For examples- Male & Female
Further, we can use convenience or consecutive sampling.

(4) Judgemental or Purposive samplings-

We need accurate or exact information so we choose people who can provide us exact information.

(5) Snowball samplings → (Reference sampling)

For example, when a snowball starts falling it is a small piece of snow but when it continues falling down & then stops at a point, its size becomes larger & its weight becomes heavier. → (Example to explain the meaning of snowball)
We select our favourite person & then he refers to the others who are equally efficient or more efficient than our selected person.
We move forward with reference.

In-depth analysis → (when we need this, we move towards consecutive sampling)

Scanned with CamScanner

Advantages of non-probability samplings-

→ less time is consumed.
→ less cost required.

Sample selection:-

The group size may be different. So we have to do sampling according to the size of group.

$L = \bar{b}$

21-10-2020 Wednesday

"Sampling Distribution"

⇒ It is used in case of probability sampling. We move towards probability sampling when we have complete list of population (population is available), accuracy & time are required, enough cost is available to conduct a random sample & population and characteristics under study are also available. In that case, we use sampling. ~~distribution~~.

Proportion → Mean → Average value → statistical measure
↳ is a tool or formula to measure proportions.

⇒ Whenever we describe our objective, we have to clearly tell that at the end we are going to use

Scanned with CamScanner

such kind of statistics to calculate our results. We continue our process of sampling & get an average value. Then we want to check our average value. Is it true or not? Is it truly representative of population or not? To for this purpose, we'll again take a sample & conduct an average value & we continue this process & repeat our sample 10 times & get 10 average values then we have to see how much these average values are close to each other. This procedure is known as sampling distribution.

⇒ In simple words, we say that sampling distribution is a procedure in which we conduct samples again & again and against every sample we get a statistical value. We draw this statistical value then we see that either this average is truly representative of population or not. If in case, it is not truly representative of population then in this situation which type of average we need to use which is truly representative of population.

⇒ When we talk about probability then statistics (probability distribution) does not apply on a single value. So when we have to check accuracy of our statistical value then we conduct different samples.

outcomes → statistical values

Scanned with CamScanner

⇒ We have different statistical values. We note out the frequencies of these statistical values. We draw these frequencies & then make a curve. Then we come to know that how much our statistical value is accurate. Either it is truly representative or it is deviating towards small values or its deviation is towards large values.

Understanding sampling distributions

⇒ Sampling mean comes through sampling distribution.

⇒ When we draw different samples & get different statistical values then we take mean of these statistical values. This is called sampling mean. Sampling mean is called mean of sample means.

⇒ Sample & sampling are two different things.

⇒ Sampling distribution is conducted from sample averages rather than sample units.

⇒ In averages, we include mean, median, mode, quartiles, quintiles, harmonic mean, geometric mean etc.

⇒ Standard deviation, range & variance measure the variability of the sampling distribution.

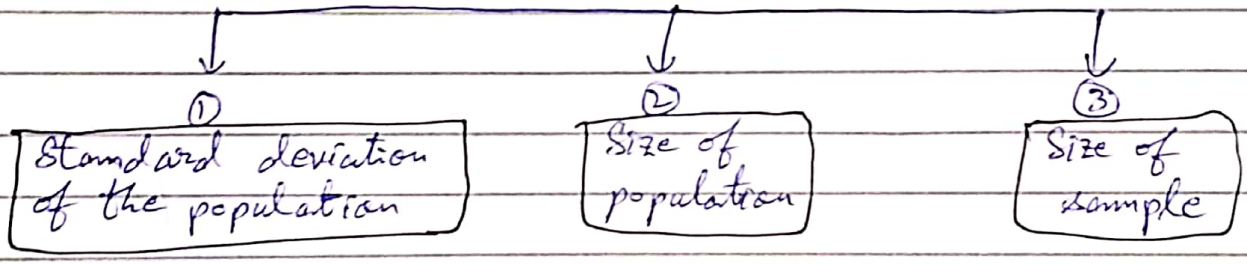
Scanned with CamScanner

(The mean we get through one sample is called sample mean.)

⇒ Standard deviation of a sampling distribution is called standard error.

⇒ When we talk about variation of units then we call it as variance or standard deviation. When we talk about variability of different statistical values, then we call it as standard error.

⇒ Standard error depends on:-



Special Considerations:-

⇒ Normal distributions-

↳ (symmetric or same-same behaviour)

↳ We show this as a bell-curved shape.

Scanned with CamScanner

26-10-2020. L=7 Monday.

(14.4) "Sampling Distribution."

(14.4.1) Sampling Distribution of means:-

⇒ Sum of all probabilities is always equal to one.

Types of Sampling Distribution:-

Most frequently used types in statistical inferences.
Binomial, Normal, t-distribution, chi-square distribution & F distribution.

⇒ Sampling distribution is different from sample distribution.

Sampling mean → $\mu_{\bar{x}}$ (Proofs → No need to do)
Population parameter → μ
Standard deviation → $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}}$
(without replacement)
(σ = population s.d)

(With replacement): → $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$

FPC → Finite population correction → $\frac{N-n}{N-1}$

27-10-2020. L=8 Tuesday

(Example (14.8) & (14.9) from book) Variance = σ^2
(Example 1.48.1) → (from pdf book)

Scanned with CamScanner

28-10-2020. L=9 Wednesday.

Population parameter:-

$$\mu = \text{mean} = \frac{\sum_{i=1}^{N} X_i}{N}$$

$$OR\ \sigma^2 = \text{variance} = \frac{\sum_{i=1}^{N} (X_i - \mu)^2}{N}$$

$$\sigma = \text{standard deviation} = \sqrt{\sigma^2}$$

Sample statistic:-

$$\text{mean} = \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

$$\text{variance} = S^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}$$

$$\text{Standard deviation} = S = \sqrt{S^2}$$

'0' represents failure.
'1' represents success.

Sample proportions → (how much 'yes' & 'no')

$$\mu_p = np \quad \text{where } p = \frac{X}{n}$$

$$\sigma_p = \sqrt{\frac{pq}{n}}$$

Sampling distribution of 'p':-

Objectives:-

Compute the mean & standard deviation of the sampling distribution of 'p'.

Scanned with CamScanner

Assume that in an election race between Candidate A and Candidate B, 0.60 of the voters prefer Candidate A. If a random sample of 10 voters were polled, it is unlikely that exactly 60% of them (6) would prefer Candidate A. By chance the proportion in the sample preferring Candidate A could easily be a little lower than 0.60 or a little higher than 0.60. The sampling distribution of p is the distribution that would result if you repeatedly sampled 10 voters & determined the proportion (p) that favored Candidate A.

The sampling distribution of p is a special case of the sampling distribution of the mean. Table 1 shows a hypothetical random sample of 10 voters. Those who prefer Candidate A are given scores of '1' & those who prefer Candidate B are given scores of '0'. Note that seven of the voters prefer Candidate A so the sample proportion (p) is:-

$$p = \frac{X}{n} = \frac{7}{10} = 0.70$$

As you can see, p is the mean of the 10 preference scores.

Table 1. Sample of voters

Voter	Preference
1	
2	0
3	1
4	1
5	1
6	0
7	1
8	0
9	1
10	1

n=10

X=7

Scanned with CamScanner

The distribution of p is closely related to the binomial distribution.

(Binomial distribution:-

When we answer according to categories & we have two categories.)

The binomial distribution is the distribution of the total number of successes (favoring Candidate A, for example) whereas the distribution of p is the distribution of the mean number of successes. The mean, of course, is the total divided by the sample size, N. Therefore, the sampling distribution of p & the binomial distribution differ in that p is the mean of the scores (0-70) & the binomial distribution is dealing with the total number of successes(7).

The binomial distribution has a mean of

$$\mu = Np$$

The standard deviation of the binomial distribution is:-

$$\sqrt{Npq} \qquad \boxed{\because q = 1-p}$$

$$\Rightarrow \sqrt{Np(1-p)}$$

$$= \sqrt{10 \times 0.7 (1-0.7)} = \sqrt{10 \times 0.7 (0.3)} = \sqrt{10 \times 0.21} = \sqrt{2.1}$$

$$\boxed{1.45}$$

$$\sigma_p = \sqrt{\frac{pq}{n}} = \sqrt{\frac{0.7 \times 0.3}{10}} = \sqrt{\frac{0.21}{10}} = \sqrt{0.021}$$

$$\Rightarrow \boxed{\sigma_p = 0.145}$$

Scanned with CamScanner

02-11-2020 L=10 Monday.

"Statistical Inference: Estimation."

Parameters:-

A parameter is a descriptive measure computed from an entire population of data.

OR

A value usually a numerical value that describes a population.

Statistic:-

A statistic is a descriptive measure computed from a sample of data.

OR

A value usually a numerical value that describes a sample.

Population mean $= \mu = \frac{\sum_{i=1}^{N} X_i}{N}$

Population standard deviation $= \sigma = \sqrt{\frac{\sum_{i=1}^{N} (X-\mu)^2}{N}}$

Population variance $= \sigma^2 = \frac{\sum (X_i - \mu)^2}{N}$

Population size $= N$

Sample size $= n$

→ Population parameter

Scanned with CamScanner

Statistic

- Sample mean $= \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$
- Sample variance $= S^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}$
- Sample standard deviation $= S = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n}}$
- Population $= \mu = \frac{X}{n}$

value → estimate
formula → estimator

$$s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{(n-1)}}$$

(Example (15.1) → On notes)

Q:- A sample of size $n=10$ yields values 8, 4, 10, 5, 5, 4, 9, 4, 3, 7. Estimate the mean & variance of population. Also find standard error of the sample mean.

Sol:- The sample mean is

$$\bar{X} = \frac{\sum X_i}{n} = \frac{8+4+10+5+5+4+9+4+3+7}{10} = \frac{59}{10}$$

$$\Rightarrow \boxed{\bar{X} = 5.9} \text{ Ans.}$$

Now, Standard deviation $= S = \sqrt{\frac{1}{n} \sum (X_i - \bar{X})^2}$

$$= \sqrt{\frac{1}{10}\left[(8-5.9)^2+(4-5.9)^2+(10-5.9)^2+(5-5.9)^2+(5-5.9)^2+(4-5.9)^2+(9-5.9)^2+(4-5.9)^2+(3-5.9)^2+(7-5.9)^2\right]}$$

Scanned with CamScanner

$$S = \sqrt{\frac{1}{10}\left[(2.1)^2+(-1.9)^2+(4.1)^2+(-0.9)^2+(-0.9)^2+(-1.9)^2+(3.1)^2+(-1.9)^2+(-2.9)^2+(1.1)^2\right]}$$

$$= \sqrt{\frac{1}{10}\left(4.41+3.61+16.81+0.81+0.81+3.61+9.61+3.61+8.41+1.21\right)}$$

$$= \sqrt{\frac{1}{10}(52.9)}$$

$$S = \sqrt{52.9}$$

$$\Rightarrow \boxed{S = 7.27}$$

Variance = $\boxed{S^2 = 52.9}$ Ans.

Standard error = $S_{\bar{x}} = \frac{S}{\sqrt{n}} = \frac{7.27}{\sqrt{10}} = \frac{7.27}{3.16}$

$$\Rightarrow \boxed{S_{\bar{x}} = 2.3}$$ Ans.

(15.3.1) Criteria for Good Point Estimators:-

(i) Unbiasedness:-

An estimator is defined to be unbiased if the statistic used as an estimator has its expected value equal to the true value of population parameter being estimated.

Expected of $\hat{\theta} = \theta \rightarrow \boxed{E(\hat{\theta}) = \theta}$

(ii) Consistency:-

An estimator is said to be consistent if the statistic to be used as estimator becomes closer & closer to the population parameter being estimated as the sample size 'n' increases.

Scanned with CamScanner

03-11-2020 L = II Tuesday.

Properties of a good point estimator:-

(i) Unbiasedness:-

An estimator is defined to be unbiased of the statistic used as an estimator has its expected value equal to the true value of population parameter being estimated. In other words let $\hat{\theta}$ be an estimator of a parameter θ, then $\hat{\theta}$ will be called an unbiased estimator of θ i.e. $E(\hat{\theta}) = \theta$.

In similar ways

$$E(\bar{X}) = \mu$$

$$E(s^2) = \sigma^2, \text{ where } s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$

$$\& \quad \sigma^2 = \frac{\sum_{i=1}^{N} (X - \mu)^2}{N}$$

$$E(\hat{p}) = P \text{ where } \hat{p} = \frac{X}{n}.$$

(ii) Consistency:-

An estimator is said to be consistent if the statistic used as an estimator becomes closer & closer to the population parameter being estimated as the sample size n increases.

Let an estimator $\hat{\theta}$ is called a consistent estimator of θ if the probability that $\hat{\theta}$ becomes closer & closer to θ approaches unity with the increasing sample size

i.e. $$\lim_{n \to \infty} P\{|\hat{\theta} - \theta| \leq e\} = 1$$

To check consistency, we find variance of that

Scanned with CamScanner

estimator. If variance of estimator approaches zero, we shall say that estimator is consistent.

i.e. $E(\bar{X}) = \mu$

$$Var(\bar{X}) = \frac{\sigma^2}{n} \; ; \; \lim_{n\to\infty} Var(\bar{X}) = \frac{\sigma^2}{n} = 0$$

so $\bar{X}$ is a consistent estimator.

(Probability = 1 when we are certain)

(iii) **Efficiency:-**

An unbiased estimator is defined to be efficient if the variance of its sampling distribution is smaller than that of the sampling distribution of any other unbiased estimator.

Suppose there are two unbiased estimators T_1 & T_2 of same parameter θ, then T_1 will be said to be more efficient estimator than T_2, if $Var(T_1) < Var(T_2)$.

The relative efficiency of T_1 compared to T_2 is given by the ratio $E_f = \frac{Var(T_2)}{Var(T_1)} > 1$

Let $Var(\bar{X}) = \frac{\sigma^2}{n}$, $Var(\text{median}) = \frac{\pi\sigma^2}{2n}$.

$$\text{then } \frac{Var(\text{median})}{Var(\bar{X})} = \frac{\pi\sigma^2/2n}{\sigma^2/n} = \frac{\pi\sigma^2}{2n} \times \frac{n}{\sigma^2} = \frac{\pi}{2} > 1$$

Hence the sample mean is more efficient than sample median. ($\bar{X}$ → represents/uses complete information)

(iv) **Sufficiency:-**

An estimator is said to be sufficient if the statistic used as an estimator uses all information that is contained in a sample. (Example 15.3)

Scanned with CamScanner

9.5 Normal Distribution:-

The normal probability distribution, which is considered the cornerstone of the modern statistical theory. was discovered by Abraham de Moivre (1667 - 1754) as the limiting form of the binomial distribution by increasing n, the number of trials, to a very large number for a fixed value of p.
The normal distribution is also called the Gaussian distribution in honour of the great German mathematician Carl F. Gauss (1777-1855), who also derived its mathematically as the probability distribution of the errors of measurements. It was karl Pearson who in 1893 called it the normal distribution and is best known by this name today.

Scanned with CamScanner

A normal distribution is defined by the P.d.f.

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\{(x-\mu)^2/2\sigma^2\}}.$$

for $-\infty < x < \infty$, and $\sigma > 0$

where 'μ' is the mean, σ is the standard deviation and π (= 3.1416) and e (= 2.7183) are constants.

Obviously a normal distribution is characterized by two Parameters 'μ' and 'σ', its mean and standard deviation.

Since $\int_{-\infty}^{\infty} f(x)\,dx = 1$, the function $f(x)$ is a proper Probability density function.

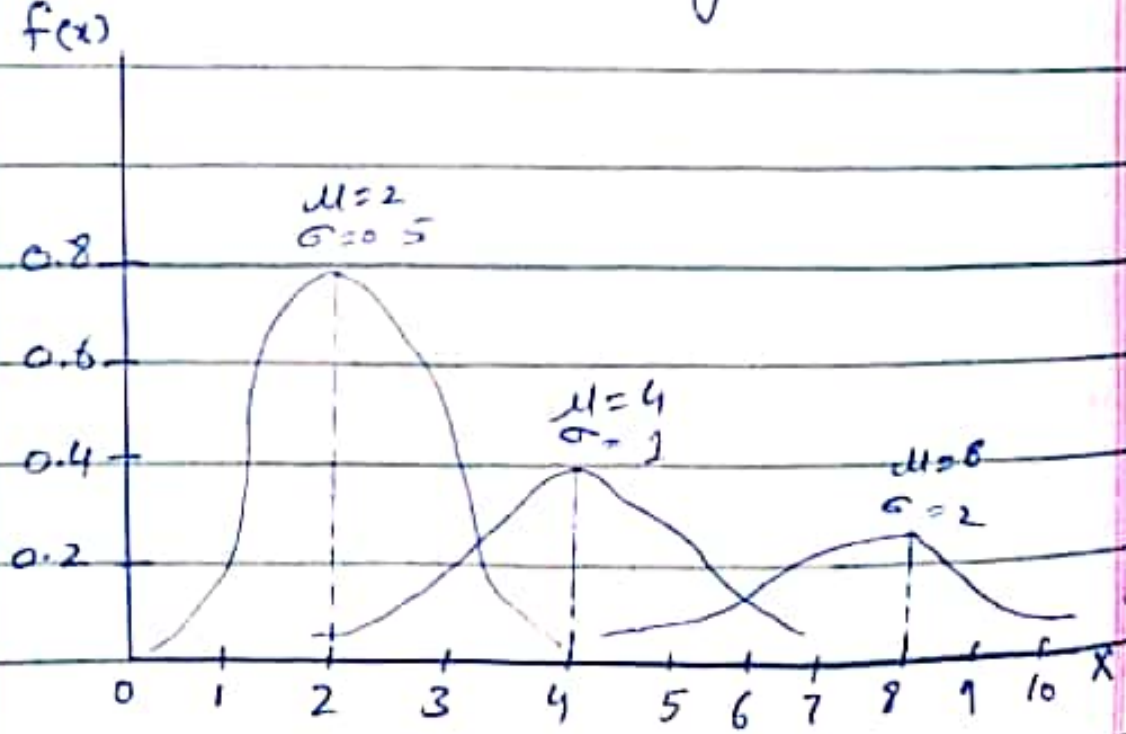
The normal distribution having mean 'μ' and variance 'σ^2' is usually denoted by $N(\mu, \sigma^2)$. Thus 'X' is $N(\mu, \sigma^2)$ means that a r.v is normally distributed with mean 'μ' and variance σ^2.

Scanned with CamScanner

The graph of the normal distribution, which is symmetrical bell-shaped curve, is called the normal curve.

The location and shape of the normal curve are determined by 'μ' and 'σ'. To put it differently, 'μ' changes the position of the normal curve along horizontal axis while 'σ' determines the horizontal spread.

A sketch for different values of 'μ' and 'σ' is given below:-



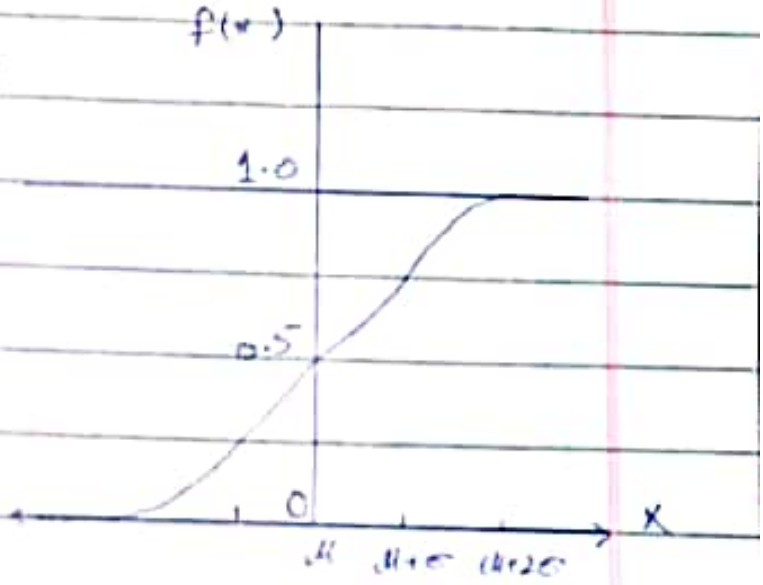
Scanned with CamScanner

The distribution function of the normal Probability distribution is given by:-

$$F(x) = P(X < x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-[(t-\mu)^2/2\sigma^2]} dt$$

which is sketched below:-

This curve is the ogive of the normal curve.



9.5.1 Standardized Normal Distribution:-

A normal Probability depends on the values of the parameters 'μ' and 'σ2' and the various possible values for these two parameters will result in an unlimited number of different normal distributions. The r.v. $Z = \frac{X-\mu}{\sigma}$, as we have seen, has zero mean and unit variance.

Scanned with CamScanner

Every normally distributed r.v. X with mean = μ and variance = σ^2 is therefore Conveniently transformed into a new normal r.v. Z with zero mean and unit variance by using the following expression:-

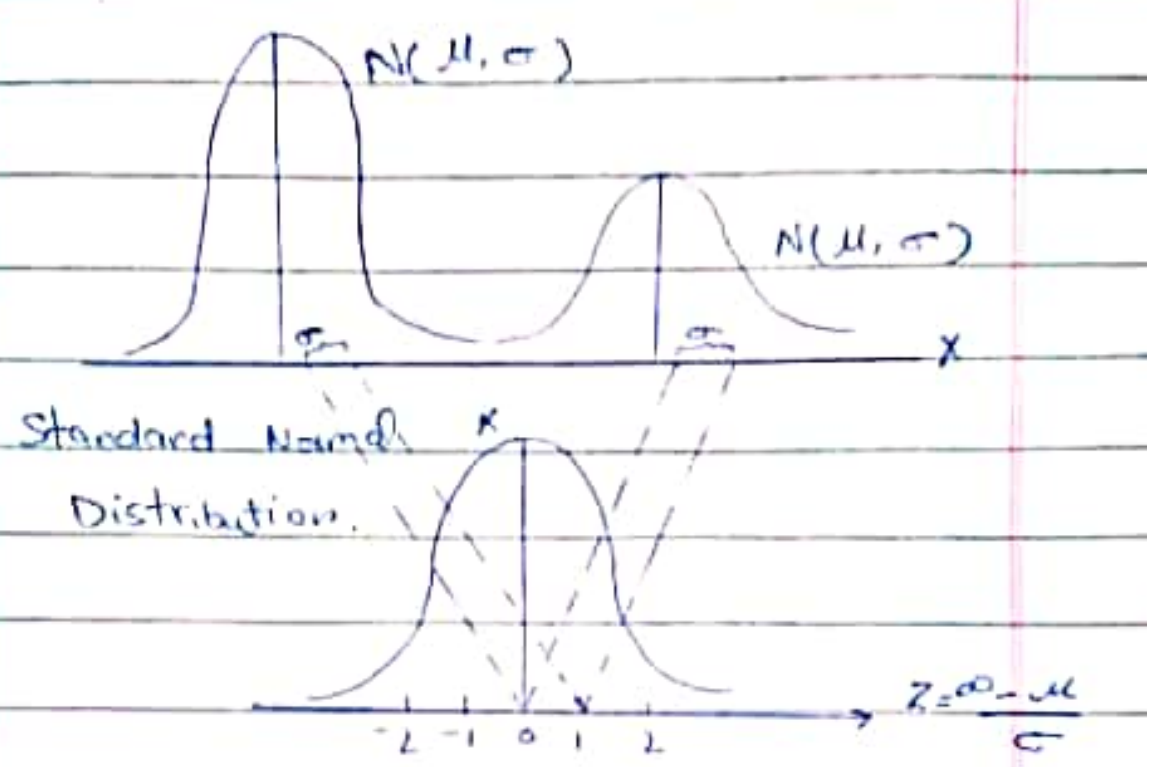
$$Z = \frac{X - \mu}{\sigma}$$

Then the P.d.f of Z, denoted by $\phi(z)$ ('ϕ' is Pronounced phi) becomes:

$$\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-t^2/2}, \text{ for } -\infty < z < \infty.$$

The transformation of the original normal distribution into the standard normal distribution.

Scanned with CamScanner



The normal Probability distribution of 'z' which has zero mean and unit variance, is called the standardized normal distribution or unit normal distribution and is denoted by $N(0,1)$. The distribution function of the standard normal distribution, usually denoted by $\Phi(z)$ (Φ is capital ϕ), is:

$$\Phi(z) = P(Z \leq z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-t^2/2}\, dt,$$

which has been tabulated for positive values of z. The values of

Scanned with CamScanner

$\Phi(z)$ for negative values of 'z' are obtained from the identity $\Phi(-z) = 1 - \Phi(z)$.

It should be noted that $F(x) = \Phi\left[\frac{x-\mu}{\sigma}\right]$ and for any 'a' and 'b' (positive or negative) $P(a < z < b) = \Phi(b) - \Phi(a)$.

Properties of Normal Distribution:-

(i):- The function $f(x)$ defining the normal distribution is a proper p.d.f i.e $f(x) \geq 0$ and the total area under the normal curve is unity.

(ii) The mean and variance of the normal distribution are 'μ' and 'σ^2' respectively.

(iii) The median and the mode of the normal distribution are each equal to 'μ', the mean of the distribution.

(iv) The mean deviation of the normal distribution is approximately '$\frac{4}{5}$' of its standard deviation.

(v) The normal curve has points of inflection which are equidistant

Scanned with CamScanner

from the mean.

The point of inflexion by which we mean a point at which the concavity changes.

The two points of inflection of normal curve are $\left[\mu-\sigma, \frac{1}{\sigma\sqrt{2\pi e}}\right]$ and $\left[\mu+\sigma, \frac{1}{\sigma\sqrt{2\pi e}}\right]$.

In other words, the points of inflection occur on the right and on the left of the mean at a distance equal to standard deviation and thus the graph of the normal curve is bell-shaped.

(vi) For the normal distribution, the odd order moments about the mean are all zero and the even order moments are given by:-

$$\mu_{2n} = (2n-1)(2n-3)\ldots 5\cdot 3\cdot 1\,\sigma^{2n}.$$

(vii) If 'X' is $N(\mu, \sigma^2)$ and if $Y = a+bX$, then Y is $N(a+b\mu, b^2\sigma^2)$.

Scanned with CamScanner

imp

(iii) The sum of independent normal variables is a normal variable. Stated differently, if 'X_1' is $N(\mu_1, \sigma_1^2)$ and X_2 is $N(\mu_2, \sigma_2^2)$, then for independent 'X_1' and X_2, $X_1 + X_2$ is $N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$.

(iv) No matter what the values of 'μ' and 'σ' are, areas under normal curve remain in certain fixed proportions within a specified no. of S.D's on either side of μ.

e.g the interval :-

(i) $\mu \pm \sigma$ will always contain 68.26%.

(ii) $\mu \pm 2\sigma$ will always contain 95.44%.

(iii) $\mu \pm 3\sigma$ will always contain 99.73%.

practically all of the area is b/w '$\mu - 3\sigma$' and $\mu + 3\sigma$, the range of the distribution is therefore approximately 6 S. deviations (theoretically curve goes from $-\infty$ to $+\infty$), and we usually terminate the graph at these points.

Scanned with CamScanner

This is a very important property of the normal distribution as most of the tests of significance for large samples are based on it.

(x) The Quartile Deviation Q_1 is found as:-

$$\frac{1}{\sigma\sqrt{2\pi}} \int_{\mu-Q}^{\mu+Q} e^{-(x-\mu)^2/2\sigma} dx = \frac{1}{2}$$

or

$$\frac{1}{\sqrt{2\pi}} \int_{0}^{Q/\sigma} e^{-z^2/2} dz = \frac{1}{4}, \text{ where } z = \frac{x-\mu}{\sigma}$$

we find from area table, that $\frac{Q}{\sigma} = 0.6745$ or $Q = 0.6745\sigma$ which is also called the Probable Error, a term not used nowadays.

This also gives the values of the Quartiles which are :-

$$Q_1 = \mu - 0.6745\sigma \text{ and } Q_3 = \mu + 0.6745\sigma$$

(xi) The normal curve approaches, but never fully touches, the horizontal axis on either side of the mean

Scanned with CamScanner

towards plus and minus infinity. That is the curve is asymptotic to the horizontal axis as $x \to \pm \infty$.

Example# 9.4:-

Let $X \sim N(0, 1)$ mean that 'X' has a normal distribution with zero mean and unit variance. what will be the distribution of $2X-3$; $\frac{7}{2}X+5$; and $4X$?

Sol:-

(i) $2X-3$

Given that $E(X)=0$ and $Var(X)=1$

Let $Y = 2X-3$. Then

$E(Y) = E(2X-3)$

$E(Y) = 2E(X) - 3$

$E(Y) = 2(0) - 3$

$\boxed{E(Y) = -3}$ and;

$Var(Y) = Var(2X-3)$

$= 2^2 Var(X)$

$= 4(1)$

$\boxed{Var(Y) = 4}$

Hence the distribution of '$2X-3$' is $N(-3, 4)$.

(ii) $\frac{7}{2}X+5$.

Let $Y = \frac{7}{2}X+5$. Then $E(Y) = E(\frac{7}{2}X+5)$

$E(Y) = \frac{7}{2}E(X) + 5 = 0+5$

$\boxed{E(Y) = 5}$

$Var(Y) = Var(\frac{7}{2}X+5)$

$Var(Y) = (\frac{7}{2})^2 Var(X)$

$\boxed{Var(Y) = \frac{49}{64}}$

$N(5, 49/64)$

(iii) Let $Y = 4X$. Then $E(Y) = E(4X)$

$E(Y) = 4E(X)$.

$\boxed{E(Y) = 0}$

$Var(Y) = Var(4X)$

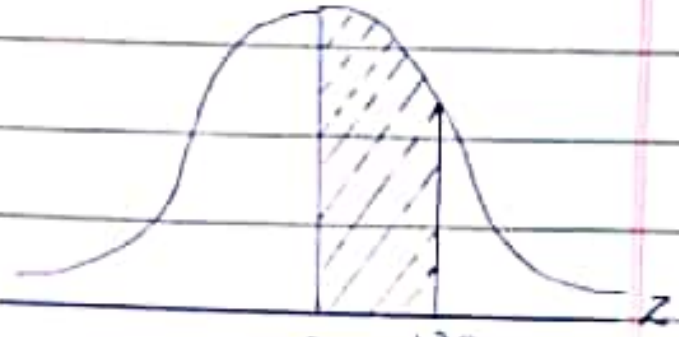
$= 4^2 Var(X)$.

$\boxed{Var(Y) = 16}$

Scanned with CamScanner

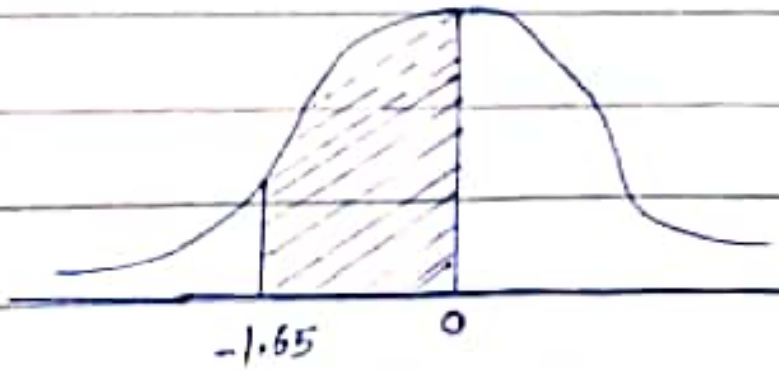
Example#7.6 Let the r.v. Z have the standard distribution. Find

(i) P= $P(0 \leq Z \leq 1.20)$



To find $P(0 \leq Z \leq 1.20)$ in Table 9.2 Page (365) we move downward the column marked 'Z' until 1.2 is reached, and then move across that row to the column headed 0.00 to find entry 0.3849. Therefore $P(0 \leq Z \leq 1.20) = 0.3849$.

(ii) $P(-1.65 \leq Z \leq 0)$.



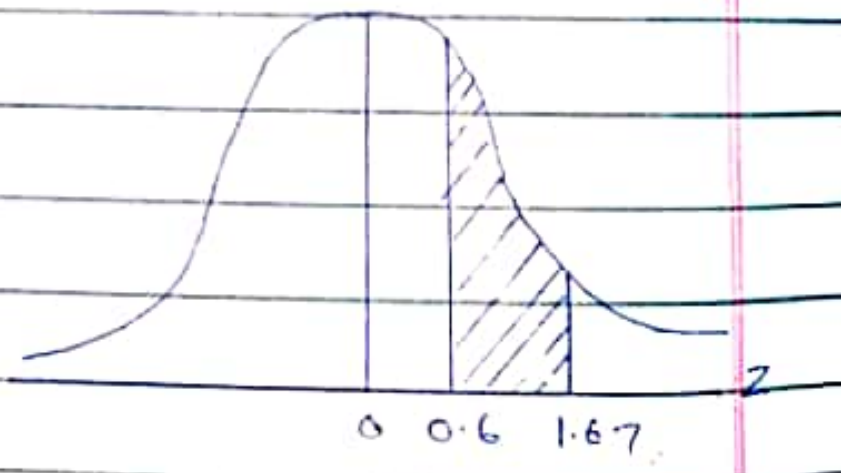
Scanned with CamScanner

Since the normal curve is symmetrical about the mean, therefore area b/w $z=0$ and positive value of 'z' is equal to the area b/w $z=0$ and a negative value of 'z' of the same magnitude.

Hence, using Table 9.2 Page S65, we have $P(-1.65 \leq z \leq 0) = P(0 \leq z \leq 1.65)$

$= 0.4505.$

(iii) $P(0.6 \leq z \leq 1.67)$.



$P(0.6 \leq z \leq 1.67)$

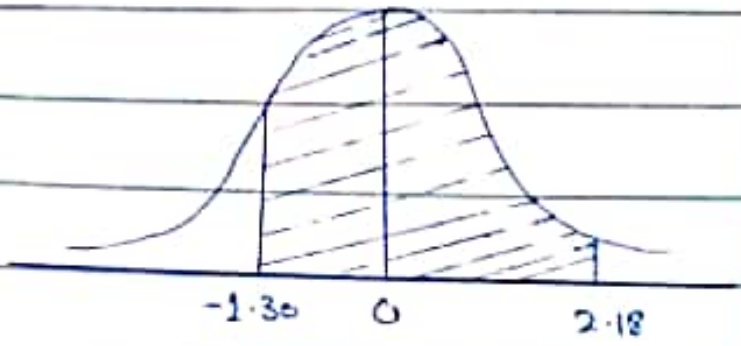
$= P(0 \leq z \leq 1.67) - P(0 \leq z \leq 0.6)$

$= 0.4525 - 0.2257$

$= 0.2268$ (From area tables)

Scanned with CamScanner

(iv) $P(-1.30 \leq Z \leq 2.18)$.



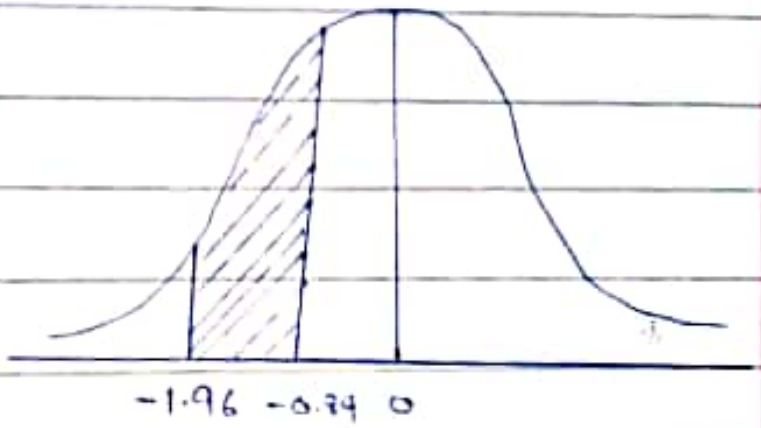
$P(-1.30 \leq Z \leq 2.18)$

$= P(-1.30 \leq Z \leq 0) + P(0 \leq Z \leq 2.18)$

$= 0.4032 + 0.4854$

$= 0.8886$ (From area tables).

(v) $P(-1.96 \leq Z \leq -0.84)$



$= P(-1.96 \leq Z \leq 0) - P(-0.84 \leq Z \leq 0)$

$= 0.4750 - 0.2995$

$= 0.1755$ (From area tables)

Scanned with CamScanner

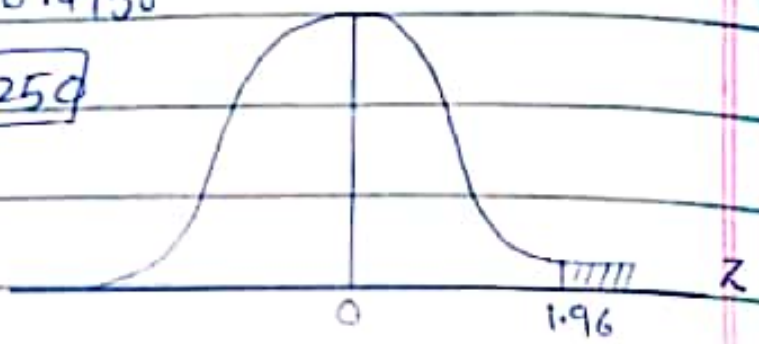
(vi) $P(Z \geq 1.96)$, and

Sol:-

$P(Z \geq 1.96) = 0.5 - P(0 \leq Z \leq 1.96)$

$= 0.5 - 0.4750$

$= 0.0250$



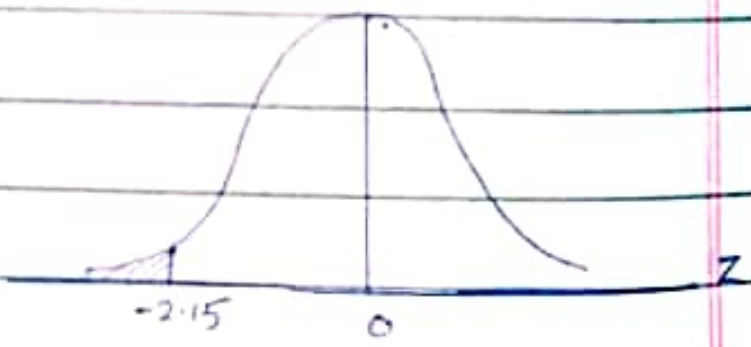
(vii) $P(Z \leq -2.15)$.

Sol:-

$P(Z \leq -2.15) = 0.5 - P(-2.15 \leq Z \leq 0)$

$= 0.5 - 0.4842$

$= 0.0158$



Example 9.7:-

A random variable 'X' is normally distributed with $\mu = 50$ and $\sigma^2 = 25$. Find the probability

Scanned with CamScanner

(a) that it will fall b/w (i) 0 and 40 (ii) 55 and 100; (b) that it will be (i) large than 54, (ii) smaller than 57.

Sol:-

We draw the normal curve Sketch showing 'x' and 'z' values, and the desired area for each part.

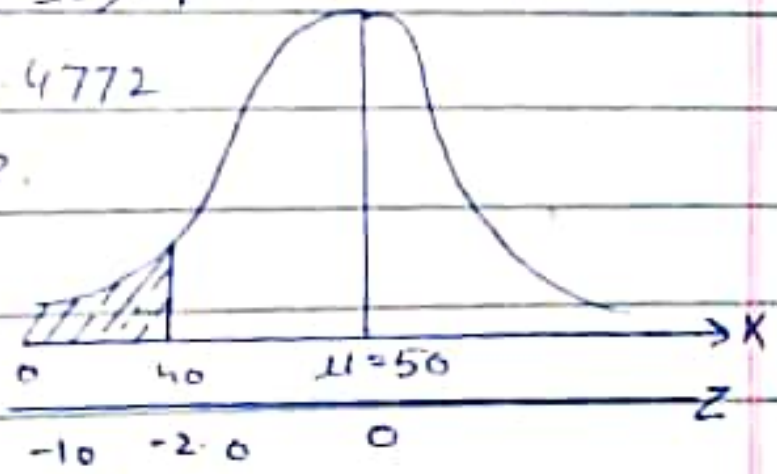
with $\mu = 50$ and $\sigma = 5$, we have:-

$$z = \frac{x - 50}{5}$$

(a) (i) At $x = 0$, we find $z = \frac{0 - 50}{5} = -10$, and at $x = 40$, we find $z = \frac{40 - 50}{5} = -2.0$

Hence, using Table 9.2 Page (365), we have:-

$$P(0 \leq X \leq 40)$$
$$= P(-10 \leq Z \leq -2)$$
$$= P(-10 \leq Z \leq 0) - P(-2 \leq Z \leq 0)$$
$$= 0.5 - 0.4772$$
$$= 0.0228.$$



Scanned with CamScanner

(ii) we have for $x = 55$,

$$z = \frac{55 - 50}{2} = +1.0$$

For $x = 100$, $z = \frac{100 - 50}{5}$

$$= 10.0$$

The 'x' values and the corresponding z' values are shown in the figure.

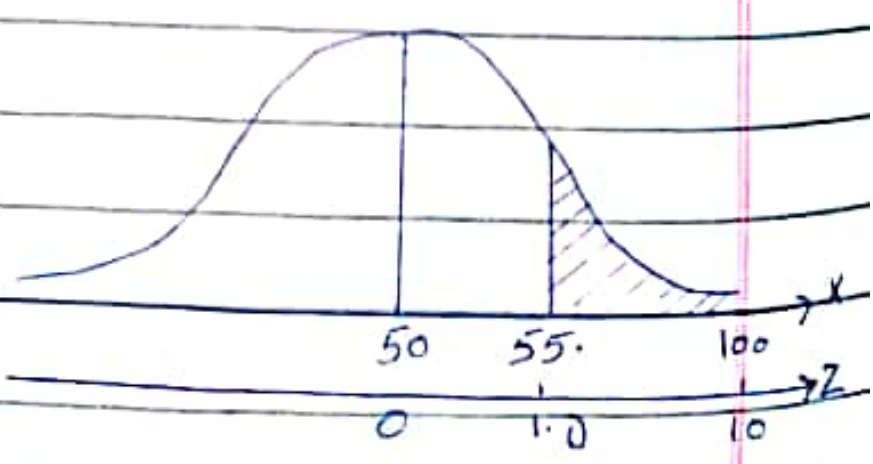
Therefore using Table 9.2, we have:-

$$P(55 \leq X \leq 100) = P(1.0 \leq Z \leq 10.0)$$

$$= P(0 \leq Z \leq 100) - P(0 \leq Z \leq 1.0)$$

$$= 0.5 - 0.3413$$

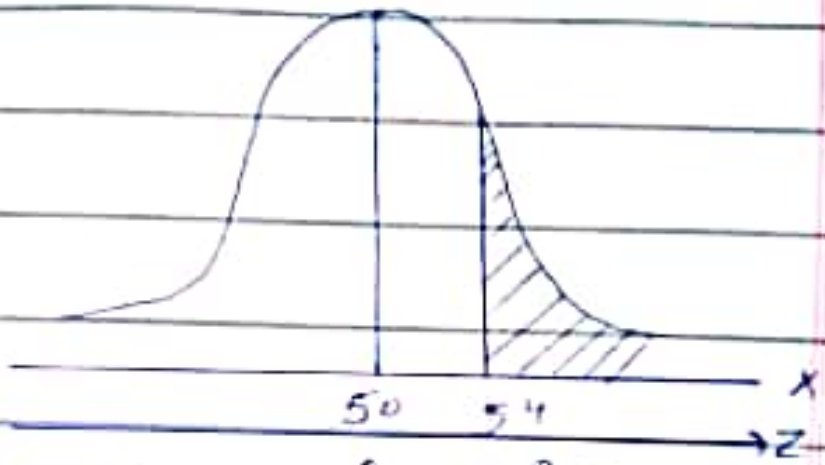
$$= 0.1587$$



Scanned with CamScanner

(b) (i) With $\mu = 50$ and $\sigma = 5$, we have:-

for $x = 54$, $z = \frac{54-50}{5} = 0.80$



Hence using Table 9.2, we have:-

$P(X \geq 54) = P(Z \geq 0.8)$

$= 0.5 - P(0 \leq Z \leq 0.8)$

$= 0.5 - 0.2881 = 0.2119$

(ii) At $x = 57$, $z = \frac{57-50}{5} = 1.40$

Therefore using Table 9.2, we have:-

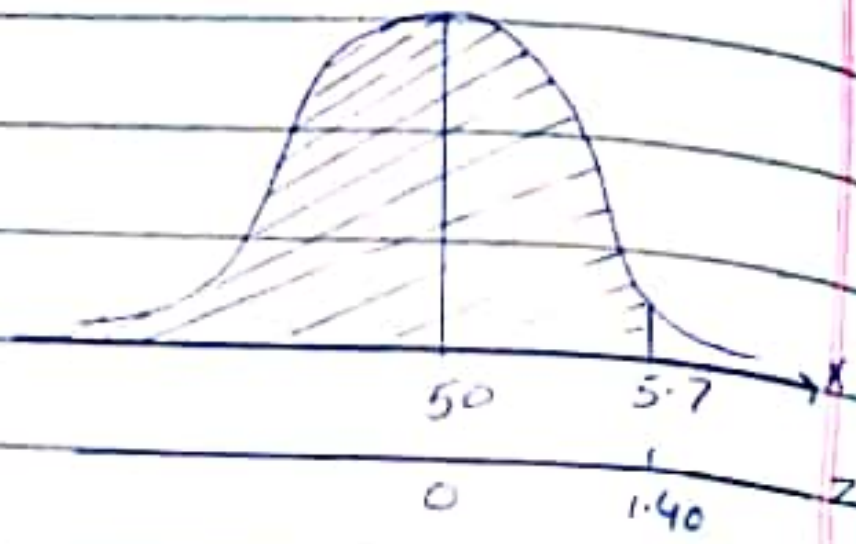
$P(X < 57) = P(Z < 1.40)$

$= 0.5 + P(0 \leq Z \leq 1.40)$

$= 0.5 + 0.4192$

$= 0.9192$

Scanned with CamScanner



Example #9.8

The length of life for an automatic dishwasher is approximately normally distribution with a mean of 3.5 years and a S.D of 1.0 years. If this type of dishwasher is guaranteed by 12 months, what fraction of the sales will require replacement?

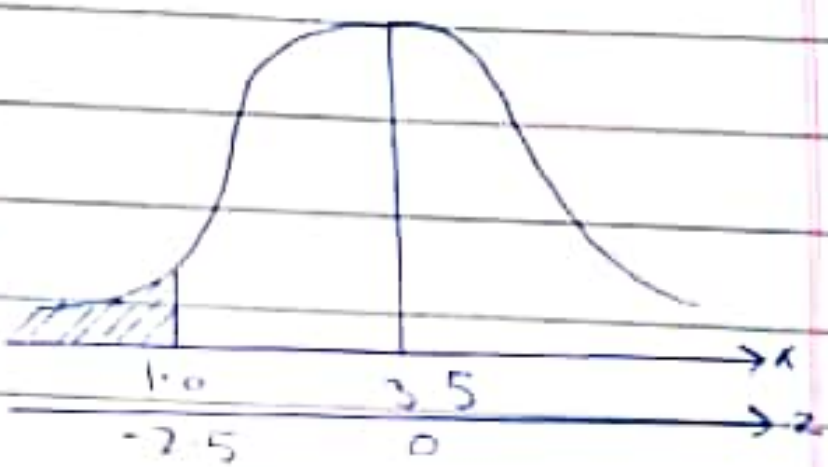
Sol:-

The fraction of sales requiring replacement is equal to the area under the normal curve for $x \leq 1$ year, the guaranteed Period.

Scanned with CamScanner

Thus we find $Z = \frac{1.0 - 3.5}{1.0}$

$$\boxed{Z = -2.5}$$



The 'x' value and the corresponding 'z' values are indicated in the figure. From Table 9.2 we find:-

$$P(X \leq 1.0) = P(Z \leq -2.5)$$

$$= 0.5 - P(-2.5 \leq Z \leq 0)$$

$$= 0.5 - 0.4938$$

$$= 0.0062$$

Hence 0.62 % of sales need replacement before 12 months.

Scanned with CamScanner

Example 9.9 :-

The mean height of Soldiers is 68.22 inches with a variance of $10.8\ (in.)^2$. And the distribution of heights to be normal, how many Soldiers in a regiment of 1000 would you expect over 6 feet tall?

Sol:-

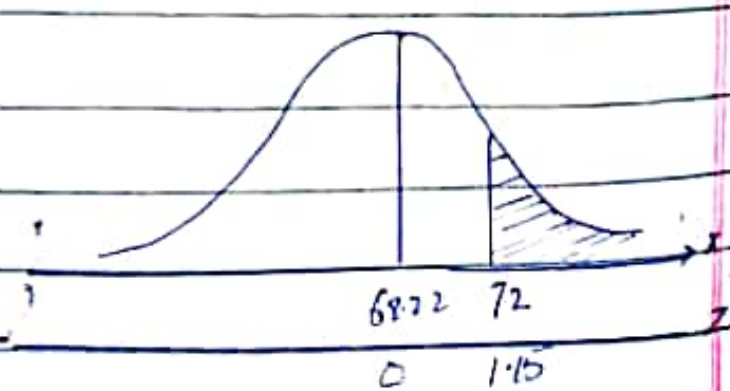
With $\mu = 68.22$ and $\sigma^2 = 10.8\ (in.)^2$, we first compute the 'z' value.

At $x = 72$ (6ft), we have :-

$$z = \frac{72 - 68.22}{\sqrt{10.8}} = \frac{3.78}{3.29}$$

$$\boxed{Z = 1.15}$$

The 'x' values and the corresponding 'Z' values are shown in normal curve sketch and we need the right tail area which is shaded.



Scanned with CamScanner

Therefore using Table 9.2 Page (365), we find:-

$$P(X \geq 72) = P(Z \geq 1.15)$$
$$= 0.5 - P(0 \leq Z \leq 1.15)$$
$$= 0.5 - 0.3749$$
$$\boxed{= 0.1251}$$

If there are 1000 Soldiers in the regiment, then number to be over 6 feet (or 72 inches) is $1000 \times 0.1251 = 125$.

Example 9.10:- If the moment generating function of 'X' is $M(t) = e^{166t + 200t^2}$, find (i) $P(170 < X < 200)$

(ii) $P(148 \leq X \leq 172)$

Sol:-

Comparing $M(t) = e^{166t + 200t^2}$ with the m.g.f of $N(\mu, \sigma^2)$, we find $\mu = 166$ and $\sigma^2 = 400$.

To find the desired probabilities, we transform 'x' values to 'z' values.

using $Z = \frac{X - 166}{20}$ Therefore:-

Scanned with CamScanner

(ii) At $x = 170$, we get:-

$$z = \frac{170 - 166}{20} = 0.2 \text{ and}$$

at $x = 200$ we find

$$z = \frac{200 - 166}{20} = 1.7$$

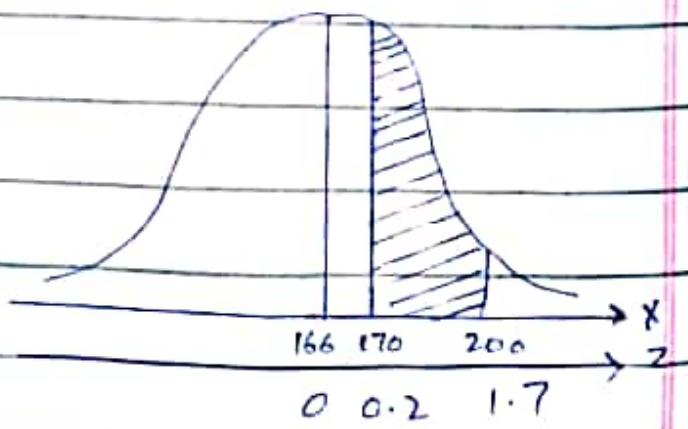
Hence, using Table 9.2 Page (365), we find:-

$$P(170 < X < 200) = P(0.2 < Z < 1.7)$$

$$= P(0 < Z < 1.7) - P(0 < Z < 0.2)$$

$$= 0.4554 - 0.0793$$

$$= 0.3761$$



Scanned with CamScanner

(ii)

At $x = 148$, we compute:-

$$z = \frac{148 - 166}{20} = -0.9, \text{ and}$$

for $x = 172$, we find:-

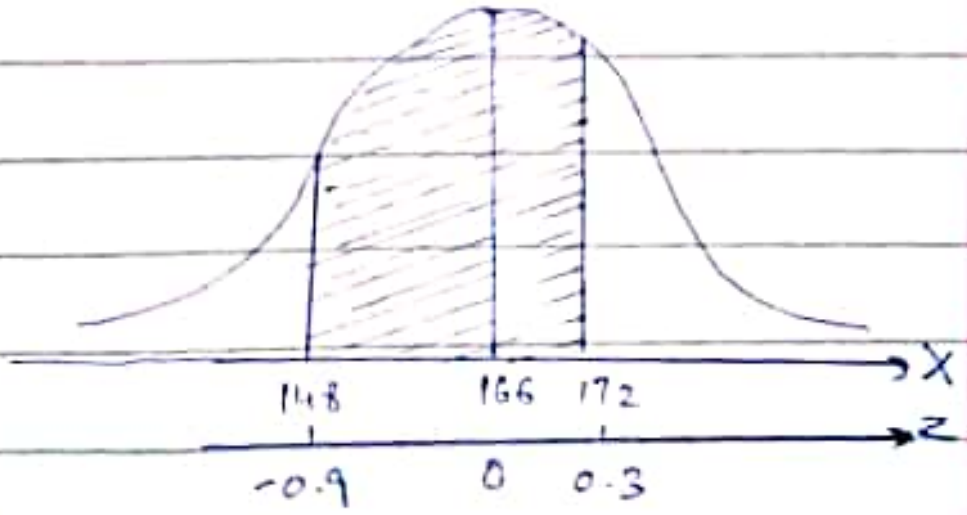
$$z = \frac{172 - 166}{20} = +0.3$$

Hence, using Table 9.2 Page (365), we get:-

$$P(148 \leq X \leq 170) = P(-0.9 \leq Z \leq 0) + P(0 \leq Z \leq 0.3)$$

$$= 0.3159 + 0.1179$$

$$= \boxed{0.4338}$$



Scanned with CamScanner

16-11-2020 L#14 Probability Theory.

Example 14.13:- Given the Population 1,1,1,3,4,5,6,6,6 and 7.

Sol:-

The mean and S.D of the Population are:-

$$\mu = \frac{\Sigma X}{N} = \frac{40}{10} = 4, \text{ and}$$

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

$$\sigma = \sqrt{\frac{210}{10} - \left(\frac{40}{10}\right)^2}$$

$$\sigma = \sqrt{\frac{210}{10} - 16}$$

$$\sigma = \sqrt{\frac{210 - 160}{10}}$$

$$\sigma = \sqrt{\frac{50}{10}} \quad 5 = \sqrt{5}$$

$$\boxed{\sigma = 2.236}$$

To calculate mean and standard deviation, we may describe the Population by the following Probability distribution:-

x	1	3	4	5	6	7
P(X=x)	3/10	1/10	1/10	1/10	3/10	1/10

Scanned with CamScanner

(a) Find the Probability that a random variable of Size '36' selected with replacement will yield a Sample mean b/w 3.26 and 4.74. (78)

Sol:-

As the sampling is performed with replacement, therefore a Sample of any Size can be selected. A sample of Size $n=36$ is large enough for the Central limit theorem to apply. The Sampling distribution of '$\bar{x}$' is therefore approximately normal with mean $\mu_{\bar{x}} = \mu = 4$ and $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{2.236}{\sqrt{36}}$

$$\sigma_{\bar{x}} = \frac{2.236}{6} = \boxed{0.373}$$

That is,

$$Z = \frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}} = \frac{\bar{x} - 4}{0.373} \text{ is}$$

approximately $N(0, 1)$

Scanned with CamScanner

To find the Probability that the mean of a random sample of size $n = 36$ will fall b/w 3.26 and 4.74, we transform 3.26 and 4.74 to z values.

Thus at $\bar{x} = 3.26$, we find:-

$$z = \frac{3.26 - 4}{0.373}$$

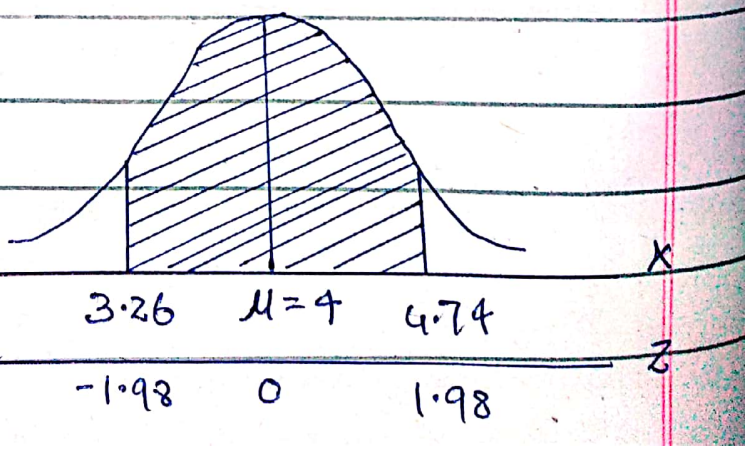
$$\boxed{Z = -1.98}$$

and at $\bar{x} = 4.74$, we find:-

$$z = \frac{4.74 - 4}{0.373}$$

$$Z = \frac{0.74}{0.373}$$

$$\boxed{Z = 1.98}$$



Scanned with CamScanner

Hence using Table of areas under normal Curve, we find:-

$$P(3.26 \leq X \leq 4.74) = P(-1.98 \leq Z \leq 1.98)$$

$$= P(-1.98 \leq Z \leq 0) + P(0 \leq Z \leq 1.98)$$

$$= 0.4762 + 0.476$$

$$\boxed{= 0.9524}$$

(b) Find the mean and S.D for the Sampling distribution of means for a Sample of size '4' selected at random without replacement. Between what two values would you expect at least $\frac{3}{4}$ of the Sample means to fall?

Sol:-

As the sample is without replacement and Sample size $n = 4$ is greater than 5% of the Population size $N = 10$, therefore the mean and S.D of the Sampling distribution of $\bar{x}$, are:-

$\mu_{\bar{x}} = \mu = 4$ and

Scanned with CamScanner

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}}$$

$$\sigma_{\bar{x}} = \frac{2.236}{\sqrt{4}} \cdot \sqrt{\frac{10-4}{10-1}}$$

$$\sigma_{\bar{x}} = \frac{2.236}{2} \cdot \sqrt{\frac{6}{9}}$$

$$\sigma_{\bar{x}} = (1.118)(0.816)$$

$$\boxed{\sigma_{\bar{x}} = 0.912}$$

The chebyshev's inequality says "at least $\left[1 - \frac{1}{k^2}\right]$ fraction of the data lies in the interval mean $\pm k(s.d)$" and the Problem says "at least $\frac{3}{4}$ of the sample means should fall in the same interval." So $\frac{3}{4}$ is $1 - \frac{1}{k^2}$ that is,

$$1 - \frac{1}{k^2} = \frac{3}{4} \quad \text{or} \quad -\frac{1}{k^2} = \frac{3}{4} - 1$$

$$-\frac{1}{k^2} = -\frac{1}{4}$$

$$\sqrt{k^2} = \sqrt{4} \Rightarrow \boxed{k = 2}$$

Scanned with CamScanner

Hence we would expect atleast $\frac{3}{4}$ of the sample means to fall in the interval $\mu_{\bar{x}} = \pm 2\sigma_{\bar{x}}$, that is between $4 - 2(0.912)$ and $4 + 2(0.912)$ or b/w 2.2 and 5.8.

Example 14.14:- A random sample of size '25' is selected from a Poisson distribution with $\mu = 3$. Find, using the central limit theorem, the Probability that the sample mean will be greater than 4.

Sol:-

Let 'X' denote the Poisson distribution distribution with $\mu = 3$.

Then $var(X) = 3$.

By the central limit theorem, '$\bar{X}$' is approximately $N(3, \frac{3}{25})$

we require $P(\bar{X} > 4)$

Thus

$$P(\bar{X} > 4) = P\left(\frac{\bar{X} - \mu}{\sigma_{\bar{x}}} = \frac{4-3}{\sqrt{3/25}}\right)$$

$$= P(Z > 2.89)$$

Scanned with CamScanner

$= 0.0019$ ANS

14.4.3 Sampling Distribution of Difference b/w Means:-

The Sampling distribution of the difference $\bar{X}_1 - \bar{X}_2$ has the following Properties:-

(i) The mean of the Sampling distribution of $\bar{X}_1 - \bar{X}_2$, denoted by $\mu_{\bar{X}_1 - \bar{X}_2}$, is equal to the difference between Population means, that is:-

$$\mu_{\bar{X}_1 - \bar{X}_2} = \mu_1 - \mu_2$$

$$[\because E(\bar{X}_1 - \bar{X}_2) = E(\bar{X}_1) - E(\bar{X}_2) = \mu_1 - \mu_2].$$

(ii) The Standard deviation of the Sampling distribution of $\bar{X}_1 - \bar{X}_2$ (standard error of $\bar{X}_1 - \bar{X}_2$), denoted by $\sigma_{\bar{X}_1 - \bar{X}_2}$, is given by:-

$$\sigma_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

$$[\because Var(\bar{X}_1 - \bar{X}_2) = Var(\bar{X}_1) + Var(\bar{X}_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}]$$

(samples are independent).

Scanned with CamScanner

(•) This expression for the S.E. of $\bar{x}_1 - \bar{x}_2$ also holds for finite Populations when sampling is performed with replacement. When Population S.D are equal or both the samples come from the same population, the expression for the S.E becomes:-

$$\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

$$= \sqrt{\frac{2\sigma^2}{n}}, \text{ when } n_1 = n_2 = n.$$

(•) If the values of 'σ_1' and 'σ_2' are not known and if both Sample sizes are large, they are replaced by 's_1' and s_2, the S.D of the respective samples. The S.E becomes:-

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

(•) If, on the other hand, the Populations are finite, sampling is done without replacement and the Sample size

Scanned with CamScanner

are large than 5% of the Population sizes, then the S.E. is:-

$$\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} \cdot \frac{N_1 - n_1}{N_1 - 1} + \frac{\sigma_2^2}{n_2} \cdot \frac{N_2 - n_2}{N_2 - 1}}$$

(iii) Shape of the distribution:-

If the Populations are normally distributed, the Sampling distribution of $\bar{x}_1 - \bar{x}_2$, regardless of Sample sizes, will be normal with mean '$\mu_1 - \mu_2$' and variance $\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$. In other words, the variable,

$$Z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

is normally distributed with zero mean and unit variance. If the Populations are non-normal and if both Sample sizes are large, (≥ 80), then the Sampling distribution of differences b/w means is approximately a normal by the

Scanned with CamScanner

Central limit theorem.

Example 14.15:- Draw all Possible random Samples of size $n_1 = 2$ with replacement from a finite Population consisting of 4, 6, 8. Similarly draw all Possible random Samples of size $n = 2$ with replacement from another finite Population consisting of 1, 2, 3.

Sol:-

There are $(3)^2 = 9$ Possible Samples which can be drawn with replacement from each Population. These are two sets of Samples and their means are given below:-

From Population-1			From Population-2		
Sample No.	Sample values	$\bar{x}_1$	Sample No.	Sample values	$\bar{x}_2$
1	4, 4	4	1	1, 1	1
2	4, 6	5	2	1, 2	1.5
3	4, 8	6	3	1, 3	2.0
4	6, 4	5	4	2, 1	1.5

Scanned with CamScanner

From Population-1			From Population-2		
Sample No.	Sample Values	$\bar{x}_1$	Sample	Sample values	$\bar{x}_2$
5	6, 6	6	5	2, 2	2.0
6	6, 8	7	6	2, 3	2.5
7	8, 4	6	7	3, 1	4
8	8, 6	7	8	3, 2	2.5
9	8, 8	8	9	3, 3	3.0

(a) Find the Possible differences b/w the sample means of the two Populations.

Sol:-

(a) The '81' Possible differences $\bar{x}_1 - \bar{x}_2$ are Presented in the following table.

$\bar{x}_2$	$\bar{x}_1$								
	4	5	6	5	6	7	6	7	8
1.0	3.0	4.0	5.0	4.0	5.0	6.0	5.0	6.0	7.0
1.5	2.5	3.5	4.5	3.5	4.5	5.5	4.5	5.5	6.5
2.0	2.0	3.0	4.0	3.0	4.0	5.0	4.0	5.0	6.0
1.5	2.5	3.5	4.5	3.5	4.5	5.5	4.5	5.5	6.5
2.0	2.0	3.0	4.0	3.0	4.0	5.0	4.0	5.0	6.0
2.5	1.5	2.5	3.5	2.5	3.5	4.5	3.5	4.5	5.5
2.0	2.0	3.0	4.0	3.0	4.0	5.0	4.0	5.0	6.0
2.5	1.5	2.5	3.5	2.5	3.5	4.5	3.5	4.5	5.5
3.0	1.0	2.0	3.0	2.0	3.0	4.0	3.0	4.0	5.0

Scanned with CamScanner

(b) Construct the Sampling distribution of $\bar{X}_1 - \bar{X}_2$ and compute its mean and variance.

Sol:-

The Sampling distribution of $\bar{X}_1 - \bar{X}_2$ (i.e, the relative frequency distribution of the possible differences $\bar{x}_1 - \bar{x}_2$) is constructed below and the mean and variance of this distribution are also computed below.

$(\bar{x}_1 - \bar{x}_2 = d)$	Tally	f	Probability $f(\bar{x}_1 - \bar{x}_2)$	$df(d)$	$d^2f(d)$
1.0	I	1	1/81	1/81	1.0/81
1.5	II	2	2/81	3/81	4.5/81
2.0	~~IIII~~	5	5/81	10/81	20.0/81
2.5	IIII I	6	6/81	15/81	37.5/81
3.0	IIII IIII	10	10/81	30/81	90.0/81
3.5	IIII IIII	10	10/81	35/81	122.5/81
4.0	IIII IIII III	13	13/81	52/81	208.0/81
4.5	IIII IIII	10	10/81	45/81	202.5/81
5.0	IIII IIII	10	10/81	50/81	250.0/81
5.5	IIII I	6	6/81	33/81	181.5/81
6.0	IIII	5	5/81	30/81	180.0/81
6.5	II	2	2/81	13/81	84.5/81
7.0	I	1	1/81	7/81	49.0/81
52	81	81	81/81 = 1	324/81	1431/81

Scanned with CamScanner

Thus the mean and the variance are:-

$$\mu_{\bar{x}_1 - \bar{x}_2} = \Sigma(\bar{x}_1 - \bar{x}_2) f(\bar{x}_1 - \bar{x}_2)$$

$$= \Sigma d f(d) = \frac{324}{81} = 4$$

$$\boxed{\mu_{\bar{x}_1 - \bar{x}_2} = 4}$$ and,

$$\sigma^2_{\bar{x}_1 - \bar{x}_2} = \Sigma (d - \mu_{\bar{x}_1 - \bar{x}_2})^2 f(d) =$$

$$\Sigma d^2 f(d) - [\Sigma d f(d)]^2$$

$$\sigma^2_{\bar{x}_1 - \bar{x}_2} = \frac{1431}{81} - \left(\frac{324}{81}\right)^2$$

$$\sigma^2_{\bar{x}_1 - \bar{x}_2} = \frac{1431}{81} - 16$$

$$\sigma^2_{\bar{x}_1 - \bar{x}_2} = \frac{1431 - 1296}{81}$$

$$\sigma^2_{\bar{x}_1 - \bar{x}_2} = \frac{135}{81} = \frac{15}{9} = \frac{5}{3}$$

$$\boxed{\sigma^2_{\bar{x}_1 - \bar{x}_2} = 5/3}$$ ANS.

(c) verify that $\mu_{\bar{x}_1 - \bar{x}_2} = \mu_1 - \mu_2$ and $\sigma^2_{\bar{x}_1 - \bar{x}_2} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_1}$.

Sol:-

The mean and variance of the first Population are:-

Scanned with CamScanner

$$\mu_1 = \frac{4+6+8}{3} = \frac{18}{3} = 6$$

$$\boxed{\mu_1 = 6} \text{ and,}$$

$$\sigma_1^2 = \frac{(4-6)^2 + (6-6)^2 + (8-6)^2}{3}$$

$$\sigma_1^2 = \frac{(-2)^2 + (0)^2 + (2)^2}{3}$$

$$\sigma_1^2 = \frac{4+4}{3} = \frac{8}{3}$$

$$\boxed{\sigma_1^2 = \frac{8}{3}}$$

The mean and variance of the second Population are :-

$$\mu_2 = \frac{1+2+3}{3} = \frac{6}{3} = 2$$

$$\boxed{\mu_2 = 2} \text{ and}$$

$$\sigma_2^2 = \frac{(1-2)^2 + (2-2)^2 + (3-2)^2}{3}$$

$$\sigma_2^2 = \frac{(-1)^2 + (0)^2 + (1)^2}{3}$$

$$= \frac{1+1}{3} = \frac{2}{3}$$

$$\boxed{\sigma_2^2 = \frac{2}{3}}$$

Scanned with CamScanner

Now $\mu_{\bar{x}_1 - \bar{x}_2} = 4 = 6 - 2 = \mu_1 - \mu_2$

and

$$\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} = \frac{4}{3}\cdot\frac{1}{2} + \frac{2}{3}\cdot\frac{1}{2}$$

$$\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} = \frac{4}{3} + \frac{1}{3}$$

$$\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} = \frac{4+1}{3} = 5/3$$

$$\boxed{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} = \sigma^2_{\bar{x}_1 - \bar{x}_2}}$$

14.4.4. Sampling Distribution of Sample Proportion:-

A sample proportion 'p' may be identified with the population mean, where the mean is obtained from the units whose possible values are either 0's and 1's. In other words, let:-

$Y_i = 1$, if the ith unit possesses the characteristic of interest,

$Y_i = 0$, if the ith unit does not possess the

Scanned with CamScanner

characteristic of interest. Then the mean is:-

$$\mu = \frac{1}{N} \sum_{i=1}^{N} Y_i$$

= Number of units having the characteristic of interest.

Total number of units in the Population.

= $\frac{X}{N}$, where 'X' represents the number of units having the characteristics of interest.

Thus the mean is simply the Proportion of 1's in the Population and we write 'p' for μ, meaning Proportion (usually called the Proportion of Success).

Similarly, the Sample Proportion '$\hat{p}$' is defined as:-

$$\hat{p} = \bar{y} = \frac{1}{n} \sum_{i=1}^{n} Y_i = \frac{X}{n}$$

Note that '$\sum Y_i = X$' is a binomial random variable and the binomial Parameter 'p' is being called a Proportion of Success here.

Scanned with CamScanner

The Sample Proportion $\hat{p}$ has different values in different samples. It is obviously a random variable and has a Probability distribution. This Probability distribution of the Proportions of all Possible random Samples of size n, is called the Sampling distribution of $\hat{p}$.

The Sampling distribution of '$\hat{p}$' has the following important Properties:-

(i) The mean of the Sampling distribution of Proportions, denoted by '$\mu_{\hat{p}}$' is equal to the Population Proportion p, that is

$$\mu_{\hat{p}} = p$$

(ii) The S.D of the Sampling distribution of Proportions, called the Standard error of '$\hat{p}$' and denoted by $\sigma_{\hat{p}}$, is given as

$$\sigma_{\hat{p}} = \sqrt{\frac{pq}{n}}$$, when the Sampling

Scanned with CamScanner

is Performed with replacement

or

$$\sigma_{\hat{p}} = \sqrt{\frac{pq}{n} \cdot \frac{N-n}{N-1}}$$, when Sampling is done without replacement from a finite Population is to be used when the sample Size n is 5% or more than 5% of the Population Size N.

(c) when the Population Proportion p' is not known and both the Population and the sample Sizes are large, then the sample Proportion $\hat{p}$ obtained from Sample data is used in place of 'p' in the expression for the S.E of P, getting

$$S_{\hat{p}} = \sqrt{\frac{\hat{p}\hat{q}}{n}}$$, where $\hat{q} = 1 - \hat{p}$.

when the sample is selected without replacement from a finite Population of size N, the S.E becomes:-

$$S_{\hat{p}} = \sqrt{\frac{\hat{p}\hat{q}}{n} \cdot \frac{N-n}{N-1}}$$

Scanned with CamScanner

Example 14.17:- A Population Consists of N=6 values 1, 3, 6, 8, 9 and 12. Draw all Possible Samples of size n=3 without replacement from the population and find the Proportion of even numbers in the samples. Construct the Sampling distribution of Sample Proportions and verify that:-

(i) $\mu_{\hat{p}} = p$, (ii) $\text{var}(\hat{p}) = \frac{pq}{n} \cdot \frac{N-n}{N-1}$

Sol:-

Where $q = 1-p$; '$\hat{p}$' and 'p' are Sample and Population Proportions respectively.

The number of Possible Samples of Size $n=3$ that could be selected without replacement is $\binom{6}{3} = 20$. Let '$\hat{p}$' represent the Proportion of even numbers in the sample, then the 20 Possible Samples and the Population of even numbers are given as

Scanned with CamScanner

Sample No.	Sample Data	Sample Proportion ($\hat{p}$)
1	1, 3, 6	1/3
2	1, 3, 8	1/3
3	1, 3, 9	0
4	1, 3, 12	1/3
5	1, 6, 8	2/3
6	1, 6, 9	1/3
7	1, 6, 12	2/3
8	1, 8, 9	1/3
9	1, 8, 12	2/3
10	1, 9, 12	1/3
11	3, 6, 8	2/3
12	3, 6, 9	1/3
13	3, 6, 12	2/3
14	3, 8, 9	1/3
15	3, 8, 12	2/3
16	3, 9, 12	1/3
17	6, 8, 9	2/3
18	6, 8, 12	3/3 = 1
19	6, 9, 12	1/3
20	8, 9, 12	2/3

Scanned with CamScanner

The Sampling distribution of Sample Proportion, is given below:-

$\hat{p}$	No. of samples	Probability $f(\hat{p})$	$\hat{p}f(\hat{p})$	$\hat{p}^2 f(\hat{p})$
0	1	1/20	0	0
1/3	9	9/20	3/20	1/20
2/3	9	9/20	6/20	4/20
1	1	1/20	1/20	1/20
$\Sigma = 2$	20	$\Sigma = 1$	10/20	6/20

Now :-

$$\mu_{\hat{p}} = \Sigma \hat{p} f(\hat{p}) = \frac{10}{20} = 0.5, \text{ and}$$

$$\sigma^2_{\hat{p}} = \Sigma \hat{p}^2 f(\hat{p}) - [(\Sigma \hat{p} f(\hat{p})]^2$$

$$\sigma^2_{\hat{p}} = \frac{6}{20} - \left[\frac{10}{20}\right]^2$$

$$\sigma_{\hat{p}}^2 = \frac{6}{20} - \left(\frac{1}{2}\right)^2$$

$$\sigma_{\hat{p}}^2 = \frac{6}{20} - \frac{1}{4}$$

$$\sigma_{\hat{p}}^2 = \frac{6-5}{20} = \frac{1}{20}$$

$$\boxed{\sigma_{\hat{p}}^2 = 0.05}$$

To verify the given result, we first calculate the Population

Scanned with CamScanner

Proportion 'p' and the Population variance Pq. Thus:-

$p = \frac{X}{N}$, where 'X' represent the number of even numbers.

$p = \frac{3}{6} = \frac{1}{2} = 0.5$, and

$\sigma^2 = pq = (0.5)(0.5) = 0.25$

Therefore $\mu_{\hat{p}} = 0.5 = p$, and:-

$$\frac{pq}{n} \cdot \frac{N-n}{N-1} = \frac{0.25}{3} \cdot \frac{6-3}{6-1}$$

$$= \frac{0.25}{3} \cdot \frac{3}{5}$$

$$= \frac{0.25}{5} = 0.05 = \text{Var}(\hat{p})$$

Example# 14.18:- Ten percent of the 1-kilogram boxes of sugar in a large warehouse are underweight. Suppose a retailer buys a random sample of 144 of these boxes. What is the probability that at least '5' per cent of the sample boxes will be underweight?

Sol:-

Scanned with CamScanner

Here the statistic is the Sample Proportion $\hat{P}$.

The Sample Size ($n=144$) is large enough to assume that the Sample Proportion '$\hat{P}$' is approximately normally distributed with mean:-

$\mu_{\hat{P}} = P = 0.10$, and standard error

$$\sigma_{\hat{P}} = \sqrt{\frac{pq}{n}} = \sqrt{\frac{(0.10)(0.9)}{144}}$$

$$\sigma_{\hat{P}} = \frac{0.3}{12} = 0.025$$

$\Rightarrow q = 1 - P$
$= 1 - 0.1$
$= 0.9$

$$\boxed{\sigma_{\hat{P}} = 0.025}$$

Therefore, $Z = \dfrac{\hat{P} - \mu_{\hat{P}}}{\sigma_{\hat{P}}} = \dfrac{\hat{P} - P}{\sqrt{pq/n}}$

$Z = \dfrac{\hat{P} - P}{0.025}$ is approximately $N(0,1)$.

we are asked to find the Probability that the Sample Proportion of the underweight is equal to or greater than 5%.

ie, we require $P(\hat{P} \geq 0.05)$

Scanned with CamScanner

Thus $P(\hat{p} \geq 0.05) \Rightarrow P(\hat{p} \geq 0.05 - \frac{1}{(2)(144)})$

[Continuity Correction].

$$= P\left[\frac{\hat{p} - 0.10}{0.025} \geq \frac{(0.05 - 1/288) - 0.10}{0.025}\right]$$

$$= P(Z \geq -2.14)$$

$$= P(-2.14 \leq Z \leq 0) + P(0 \leq Z \leq \infty)$$

$$= 0.4838 + 0.5$$

$$= 0.9838 \quad \text{ANS.}$$

14.4.5:- Sampling Distribution of Differences between Proportions:-

Suppose there are two binomial Populations with Proportions of Successes 'P_1' and 'P_2' respectively. Let independent random Samples of Size 'n_1' and 'n_2' be drawn from the respective Populations, and the differences $\hat{P}_1 - \hat{P}_2$ between the Proportions of all Possible Pairs of Samples be computed. Then Probability distribution of the differences $\hat{P}_1 - \hat{P}_2$ can be obtained.

Scanned with CamScanner

Such a Probability distribution is called Sampling distribution of the differences between the Proportions $\hat{P}_1 - \hat{P}_2$, which has the following important Properties:-

(i) The mean of the sampling distribution of $\hat{P}_1 - \hat{P}_2$, denoted by $\mu_{\hat{P}_1 - \hat{P}_2}$, is equal to the difference b/w the Population Proportions, that is $\mu_{\hat{P}_1 - \hat{P}_2} = P_1 - P_2$.

(ii) The Standard deviation of the Sampling distribution of $\hat{P}_1 - \hat{P}_2$, (i.e the S.Error of $\hat{P}_1 - \hat{P}_2$) denoted by $\sigma_{\hat{P}_1 - \hat{P}_2}$ is given by

$$\sigma_{\hat{P}_1 - \hat{P}_2} = \sqrt{\frac{P_1 q_1}{n_1} + \frac{P_2 q_2}{n_2}}$$, where $q_1 = 1 - P_1$

(iii) If both Populations have the Same Proportions of Successes, i.e $P_1 = P_2 = P$ or if both the Samples have been drawn from a Common binomial Population, then-

$$\sigma_{\hat{P}_1 - \hat{P}_2} = \sqrt{Pq\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

Scanned with CamScanner

(ii) Whenever the value of the Common Proportion 'P' is not known, then for sufficiently large sample sizes, it is replaced with its estimate $\hat{P}_c$, which is computed by taking a weighted mean of the two observed sample proportions $\hat{P}_1$ and $\hat{P}_2$ as follows:-

$$\hat{P}_c = \frac{n_1\hat{P}_1 + n_2\hat{P}_2}{n_1+n_2} = \frac{\text{Sum of successes in the two samples}}{\text{Total sample size}}$$

The standard error of $\hat{P}_1 - \hat{P}_2$ then becomes:-

$$S_{\hat{P}_1-\hat{P}_2} = \sqrt{\hat{P}_c\hat{q}_c\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}, \text{ where } \hat{q}_c = 1-\hat{P}_c$$

(iii) Shape of the distribution

The sampling distribution of $\hat{P}_1 - \hat{P}_2$ is approximately normal for sufficiently large sample sizes.

Scanned with CamScanner

Example 14.19:- Two random samples of sizes $n_1 = 40$ and $n_2 = 45$ are drawn from a binomial population with $P = 0.60$. what is the Probability that $-0.15 < \hat{P}_1 - \hat{P}_2 < +0.15$?

Sol:-

Both sample sizes ($n_1 = 40$ and $n_2 = 45$) are large enough to assume that the sampling distribution of '$\hat{P}_1 - \hat{P}_2$' is approximately a normal with mean:-

$$\mu_{\hat{P}_1 - \hat{P}_2} = P_1 - P_2 = 0, \text{ and}$$

Standard deviation

$$\sigma_{\hat{P}_1 - \hat{P}_2} = \sqrt{PQ\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

$$= \sqrt{(0.60)(0.40)\left(\frac{1}{40} + \frac{1}{45}\right)}$$

$$= \sqrt{(0.24)(0.0472)}$$

$$\boxed{= 0.106}$$

Thus the variable:-

$$Z = \frac{(\hat{P}_1 - \hat{P}_2) - 0}{\sqrt{PQ\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

$$= \frac{\hat{P}_1 - \hat{P}_2}{0.106} \text{ is approximately } N(0, 1).$$

Scanned with CamScanner

Now, at $\hat{P}_1 - \hat{P}_2 = -0.15$, we find that

$$z = \frac{-0.15}{0.106} = -1.42,$$

Hence using Table of areas under normal curve, we find:-

$$P(-0.15 < \hat{P}_1 - \hat{P}_2 < 0.15) = P(-1.42 < Z < 1.42)$$

$$= P(-1.42 < Z < 0) + P(0 < Z < 1.42)$$

$$= 0.4222 + 0.4222$$

$$\boxed{= 0.8444}$$

The desired Probability is therefore 0.8444.

Scanned with CamScanner

Probability Theory L#15 17-11-2020

Estimation of Confidence level:-

A Confidence interval estimate of the unknown Parameter 'θ' is an interval computed from a random sample of 'n' values $x_1, x_2, \ldots, x_n$ with a statement of how Confident (e.g. 90 Percent, 95 percent or 99 percent) we are that the interval Contains the unknown Parameter θ. A Confidence interval estimate is in the from $(L < \theta < U)$, where 'L' and 'U' depend upon the value of the statistic 'θ̂' of a random sample selected from the Population and the Sampling distribution of the statistic. To make an assertion that 'θ' lies in the interval (L, U), we may determine from the Sampling distribution of 'θ̂' two values 'L' and 'U' such that $P(L < \theta < U)$ is equal to any Specified Probability.

Scanned with CamScanner

conventionally denoted by '1-α'. If, of instance, L and U are two statistics such that for all θ

$P(L<\theta<U) = 1-\alpha$ for $0<\alpha<1$.

then the probability of the interval (L,U) containing the population parameter 'θ' is '1-α'. The interval (L,U) is called a 100(1-α) percent confidence interval for the unknown parameter, the probability (1-α) associated with interval estimate is called the confidence coefficient or the confidence level, and 'α' is the probability that the parameter 'θ' will lie outside the interval (L,U). Thus an interval has a specified probability (1-α) of containing the true value of the parameter. For example, if α=0.05, then the probability that the interval (L,U) contains θ, is 0.95

Scanned with CamScanner

The endpoints that bound the Confidence interval, are called the lower and upper confidence limits for θ. These limits are random variables as they can be different for different samples. The width of the Confidence, i.e the difference $U-L$, is called the Precision of the estimate. The Precision may be increased either by increasing the sample size or by decreasing the confidence level. The concept of a Confidence interval was introduced in 1937 by the Polish-English-American statistician Jerzy Neyman (1894-1981).

Some of the most commonly used Confidence intervals for Population Parameters are discussed in the sections that follow,

15.5.1:- Confidence Interval Estimate of a Population Mean.

Scanned with CamScanner

To Compute a confidence interval for the Population mean μ, we have to see whether or not the Population is normal, whether or not the Population standard deviation is known, and whether the Small size is large or small. We discuss these different cases below.

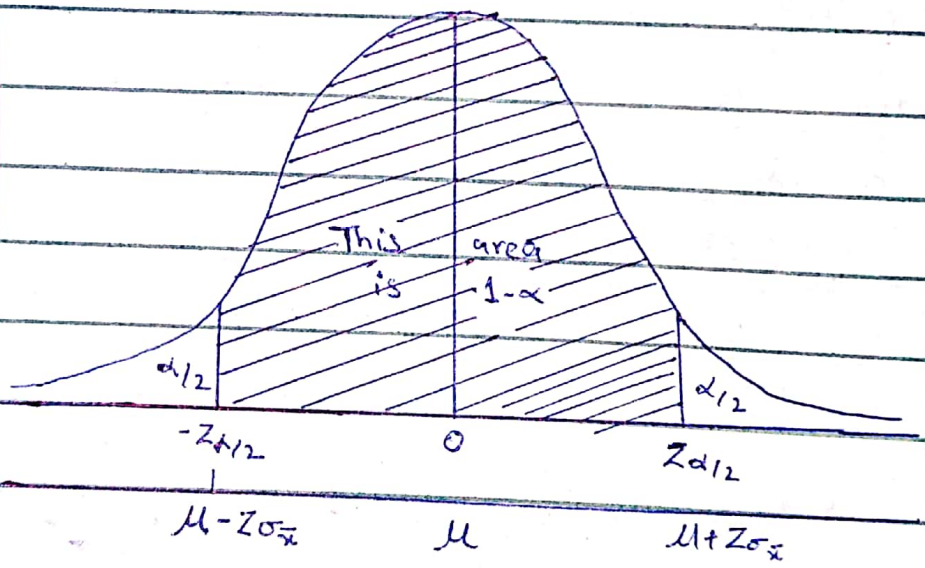
(i) Normal Population with σ known:-

Let a random sample $x_1, x_2, \ldots x_n$ of size 'n' be drawn from a normal Population with an unknown mean 'μ' and a known standard deviation σ. Then the Sampling distribution of the mean '$\bar{x}$' will be normal with a mean 'μ' and a standard deviation. $\frac{\sigma}{\sqrt{n}}$; and the variable $Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$ will be exactly standard normal, no matter how Small the size will be.

Scanned with CamScanner

The normal distribution tells us that the Probability that a value of 'Z' will fall in the interval from $-Z_{\alpha/2}$ to $Z_{\alpha/2}$ is equal to '$1-\alpha$', where '$Z_{\alpha/2}$' is equal to '$\alpha/2$'. That is we can make the following Probability statement:-

$$P\left[-Z_{\alpha/2} < \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} < Z_{\alpha/2}\right] = 1-\alpha.$$



To put inside inequalities within the brackets, we Proceed as below:-

Scanned with CamScanner

(a) we multiply all terms inside the brackets by '$\frac{\sigma}{\sqrt{n}}$' and get:-

$$-Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \bar{X} - \mu < Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

(b) we subtract '$\bar{X}$' from each term and have:-

$$-\bar{X} - Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < -\mu < -\bar{X} + Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

(c) we multiply all terms by -1 (remember, we inverse the direction of the inequality sign when we multiply both sides of the inequality by a negative number) and obtain:-

$$\bar{X} + Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} > \mu > \bar{X} - Z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}$$

which is equivalent to:-

$$\bar{X} - Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

(d) we substitute this result in the Bobability statement and get:-

Scanned with CamScanner

$$P\left[\bar{X} - Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \bar{X} < Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right] = 1-\alpha.$$

Hence, for a particular sample of size n, a $100(1-\alpha)$ percent Confidence interval for 'μ' is given by:-

$$\left[\bar{x} - Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{x} + Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$$

which may be expressed more Compactly as:-

$$\bar{x} \pm Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}.$$

If, for instance, we desire a 95% Confidence interval, i.e $(1-\alpha)100 = 95\%$ then from the table of areas under the normal curve, we find that the value $Z_{0.025}$ is 1.96 and the 95% confidence interval will be from $\bar{x} - 1.96\frac{\sigma}{\sqrt{n}}$ to $\bar{x} + 1.96\frac{\sigma}{\sqrt{n}}$. This means that about 95% of the intervals found in this way will Contain the Parameter 'μ'.

Scanned with CamScanner

Example #15.18:- The standard deviation of the amounts poured into bottles by an automatic filling machine is 1.8ml (millimeter). The amounts of fill in a random sample of bottles, in ml, were 451, 479, 482, 480, 477, 478, 481 and 482. Suppose the Population of amounts of fill is normal. Construct a 90% confidence interval for the mean interval for the mean amount in all bottles filled by the machine.

Solution:-

The 90% Confidence interval for the mean amount in all bottles, μ is given by:-

$$\bar{x} \pm Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

Here,

$$\bar{x} = \frac{\Sigma X}{n} = \frac{451+479+482+480+477+478+481+482}{8}$$

$$\bar{x} = \frac{3840}{8} = 480$$

$$\sigma = 1.8 \quad , n = 8$$

Scanned with CamScanner

$$\alpha = 0.10$$

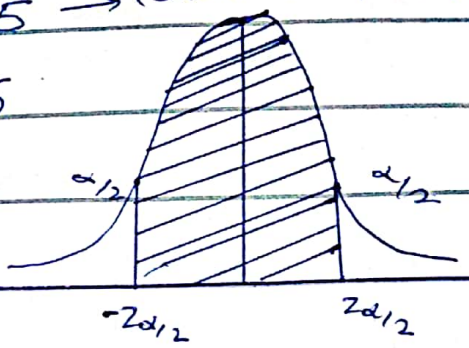
$$\frac{\alpha}{2} = \frac{0.10}{2} = 0.05$$

$$Z_{\alpha/2} = Z_{0.05}$$

$$Z_{\alpha/2} = 0.5 - 0.05$$

$$Z_{\alpha/2} = 0.495 \rightarrow \text{(check in table)}$$

$$Z_{0.05} = 1.645$$



Substituting these values, we get:-

$$480 \pm 1.645\left(\frac{1.8}{\sqrt{8}}\right).$$

or $480 \pm (1.645)(0.636)$

or 480 ± 1.05

or 478.95 to 481.05.

Hence the 90.1. confidence interval for 'u' calculated from the given sample is (478.95, 481.05).

Example 15.19:-

A confidence interval is constructed from a sample of size 25, for the mean of a normal Population which has $\sigma = 5$.

Scanned with CamScanner

The limits for the interval are 110.2 and 135.8. what confidence coefficient was used?

Solution:-

The $100(1-\alpha)\%$ confidence limits are given by:-

$\bar{X} - Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ and $\bar{X} + Z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}$.

Substituting the given values, we get:-

$$\bar{X} - Z_{\frac{\alpha}{2}}\frac{50}{\sqrt{25}} = 110.2$$

$$\bar{X} + Z_{\frac{\alpha}{2}}\frac{50}{\sqrt{25}} = 135.8$$

$$-2Z_{\alpha/2}(10) = 110.2 - 135.8$$

$$+2Z_{\alpha/2}(10) = +25.6$$

$$Z_{\alpha/2} = \frac{25.6}{20}$$

$$\boxed{Z_{\alpha/2} = 1.28}$$

We know that '$Z_{\alpha/2}$' denotes that the area to the right of $Z_{\alpha/2}$ is $\alpha/2$. From area tables, we find that the area to the right

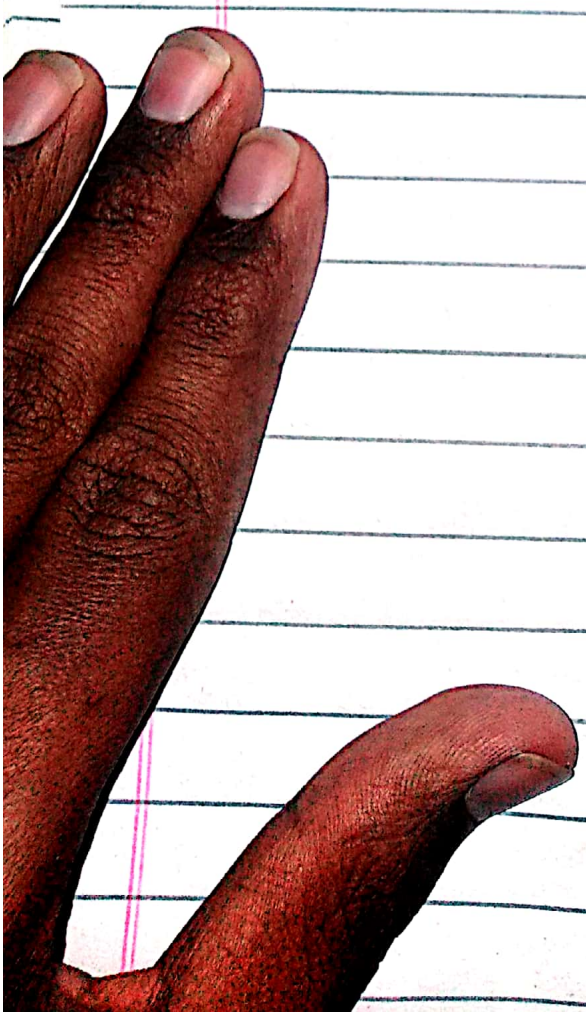
Scanned with CamScanner

of the value = 0.5 - 0.4 = 0.1,

implying that $\alpha/2 = 0.1$ or $\alpha = 0.1 \times 2 = 0.2$

Thus $1 - \alpha = 1 - 0.2 = 0.8$

$$\boxed{1 - \alpha = 0.8}$$



Scanned with CamScanner

(iii) Normal Population with σ unknown:-

when a random sample $X_1, X_2, \ldots, X_n$ of size 'n' is drawn from a normal population with σ unknown, we estimate 'σ' by the sample S.D, which is then used in place of σ. If the sample size is sufficiently large ($n \geq 30$), then the Central limit Theorem allows us to assume that the sampling distribution of '$\bar{x}$' is approximately normal with a mean of 'μ' and a standard deviation of $\frac{S}{\sqrt{n}}$, where 'S' is the sample standard deviation.

The Probability expression for estimating 'μ' then becomes:-

$$P\left[\bar{x} - z_{\alpha/2}\frac{S}{\sqrt{n}} < \mu < \bar{x} + z_{\alpha/2}\frac{S}{\sqrt{n}}\right] = 1-\alpha.$$

Thus a $100(1-\alpha)$ percent confidence interval for 'μ' is given by:-

$$\bar{x} \pm z_{\alpha/2}\frac{S}{\sqrt{n}}$$

when 'σ' is unknown and Sample size is small ($n < 30$), the sampling distribution of '$\bar{x}$' will not be normally distributed.

Scanned with CamScanner

The Sampling distribution of $\bar{x}$ then follows a distribution, known distribution.

Example #15.20:- The Punjab Highway Department is studing traffic Pattern on the G.T. Road near Lahore. As Part of the study, the department needs to estimate the average number of vehicles that Pass the Ravi boidge each day. A random sample of 64 days gives $\bar{x} = 5410$ and $S = 680$. Find the 90 Percent confidence interval estimate for μ, the average number of vehicles Per day.

Sol:-

The 90% confidence interval for 'μ' is:-

$$\bar{x} \pm Z_{\alpha/2} \frac{S}{\sqrt{n}}$$

where $\bar{x} = 5410$, $S = 680$, $n = 64$ and $Z_{0.05} = 1.645$.

Substituting these values, we get:-

$$5410 \pm (1.645)\left(\frac{680}{\sqrt{64}}\right)$$

Scanned with CamScanner

or $5410 + (1.645)(85)$

or $5410 + 139.8$

or 5270.2 to 5549.8

Thus the 90% confidence interval estimate for μ is $(5270, 5550)$.

(iii) Non-Normal Population with known or unknown σ :- (large samples).

The central limit theorem tells us that for large sample sizes, the sampling distribution of the mean $\bar{x}$ is approximately a normal, even though the population sampled is non-normal. That is, the random variable $z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$ is approximately standard normal and consequently:-

$$P\left[-z_{\alpha/2} < \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} < z_{\alpha/2}\right] = 1 - \alpha.$$

Therefore an approximate $100(1-\alpha)$ percent confidence interval for μ, the mean of a non-normal population with 'σ' known is given by:-

$$\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}.$$

In case 'σ' is unknown and is estimated by the sample S.D 's',

Scanned with CamScanner

the confidence interval estimate for 'μ' becames:-

$$\bar{x} \pm z_{\alpha/2} \frac{S}{\sqrt{n}}$$

If we sample without replacement from a finite Population of size 'N' and Sample size 'n' is greater than 5% of Population size, then the confidence interval estimate for 'μ' is given by:-

$$\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$$

Example# 15.21:- A sample of '100' observations from a Population known to be non-normal yielded the sample values $\bar{x} = 182$ and $s^2 = 299$. Find an approximation 99% Confidence interval for μ.

Solution:-

The Sample size ($n=100$) is large enough to allow us to assume that the Sampling distribution of '$\bar{x}$' is approximately

normal with mean = μ and S.D $= \frac{S}{\sqrt{n}}$. Therefore the approximate $100(1-\alpha)$ percent confidence interval for 'u' is :-

$$\bar{x} \pm Z_{\alpha/2} \frac{S}{\sqrt{n}}.$$

Now $\bar{x} = 182$, $S = \sqrt{299} = 17.29$, $n = 100$ and $Z_{0.005} = 2.58$ as the confidence coefficient interval for 'u' is:-

$$182 \pm (2.58)\left[\frac{17.29}{\sqrt{100}}\right].$$

or $182 \pm (2.58)(1.729)$

or 182 ± 4.46 or 177.54 to 186.46.

Hence an approximate 99% confidence interval for 'u' is $(177.54, 186.46)$.

Example # 15.22 :- A random sample of size $n = 200$, selected without replacement from a population of size $N = 1000$, we therefore use the following expression to calculate the desired 95% confidence interval for u:-

Sol:-

$$\bar{x} + Z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$$

Scanned with CamScanner

Substituting the values, we get:-

$$69.2 \pm (1.96)\frac{1.08}{\sqrt{200}}\sqrt{\frac{1000-200}{1000-1}}$$

or $69.2 \pm (1.96)\left(\frac{1.08}{14.141}\right)(0.8949)$

or $69.2 \pm (1.96)(0.068)$

or 69.2 ± 0.13 or 69.07 to 69.33

Hence the 95% confidence interval for 'μ' is $(69.07, 69.33)$.

Example #15.23:- A sample of readings from a normal population with unknown mean 'μ' and unknown variance 'σ^2' gave the following data:-

x	17.4	17.5	17.6	17.7	17.8
f	12	16	19	23	10

Sol:-

A second sample of readings taken from the same population gave $n_2 = 72$, $\Sigma x = 1267.2$, $\Sigma x^2 = 22536$. Combine the two samples to give estimates of 'μ' and σ^2, and give the approximate 90% confidence interval for μ.

Scanned with CamScanner

First we Calculate the sample means and the Sample variances.

For sample 1: $\bar{x}_1 = \frac{\Sigma fx}{n_1} = \frac{1408.3}{80} = 17.60$

$$S_1^2 = \frac{\Sigma fx^2}{n_1} - \left[\frac{\Sigma fx}{n_1}\right]^2 = \frac{24792.63}{80} - \left\{\frac{1408.3}{80}\right\}^2$$

$$S_1^2 = 309.9079 - 309.8920$$

$$\boxed{S^2 = 0.0159}$$

For sample 2: $\bar{x}_2 = \frac{\Sigma x}{n_2} = \frac{1267.2}{72} = 17.60$

$$S_2^2 = \frac{\Sigma x^2}{n_2} - \left[\frac{\Sigma x}{n_2}\right]^2 = \frac{22536}{72} - \left\{\frac{1267.2}{72}\right\}^2$$

$$= 313 - 309.76 = 3.24$$

$$\boxed{S_2^2 = 3.24}$$

Next we calculate the Pooled estimates $\bar{x}_c$ and S_p^2 as below:-

$$\bar{x}_c = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2} = \frac{80(17.60) + 72(17.60)}{80+72}$$

$$\boxed{\bar{x}_c = 17.60}$$

$$S_p^2 = \frac{n_1 S_1^2 + n_2 S_2^2}{n_1 + n_2 - 2} = \frac{80(0.0159) + 72(3.24)}{80+72-2}$$

$$S_p^2 = \frac{234.552}{150} = 1.5637$$

$$\boxed{S_p^2 = 1.5637}$$

Scanned with CamScanner

A 90% Confidence interval for μ, based on the combined samples, is:-

$$\bar{x}_c \pm 1.645 \frac{s_p}{\sqrt{n}}$$

or $$17.60 \pm 1.645 \frac{\sqrt{1.5637}}{\sqrt{152}}$$

or $17.60 \pm 1.645 (0.1014)$

or 17.60 ± 0.17 or 17.43 to 17.77

Hence, the 90% Confidence interval for μ, on the basis of two samples, is $(17.43, 17.77)$

15.5.2 :- Interpretation of a Confidence Interval :-

A $100(1-\alpha)$ Percent confidence interval for 'μ' is:-

$$P\left[\bar{x} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}\right] = 1-\alpha$$

It is to be emphasized that in this expression, 'μ' is constant and it is the endpoints (i.e. limits) of the interval which are random variables. Therefore after computing the confidence interval for a particular

Scanned with CamScanner

Sample, it is erroneous to say that:-

$$P\left[\bar{x} - z\alpha_{/2} \frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z\alpha_{/2} \frac{\sigma}{\sqrt{n}}\right] = 1-\alpha,$$

b/c in this expression, no random variable appears, whereas a Probability statement is made about random variables. This means that the Probability measure cannot be attached to the stated interval. If the statement is correct, in the sense that it include μ, the Probability is 1; and if it is incorrect, the Probability is zero. In neither case, the Probability is $(1-\alpha)$. However, of all possible intervals, $100(1-\alpha)$ Percent will include 'μ' and 'α' Percent of the intervals will not include 'μ' in the long run.

To be specific, suppose an actual Sample of size $n=16$ is selected from a normal Population with an unknown mean 'μ' and a known S.D

$\sigma = 2$. Let the sample mean $\bar{x} = 6.2$ and the confidence coefficient $(1-\alpha) = 0.95$. Substituting these values, we get:-

$$P\left[6.2 - 1.96\left(\frac{2}{\sqrt{16}}\right) < \mu < 6.2 + 1.96\left(\frac{2}{\sqrt{16}}\right)\right] = 0.95$$

Or $P(5.22 < \mu < 7.18) = 0.95$

This probability statement is erroneous b/c 'μ' is not a random variable. The parameter 'μ' is either in the interval or it is not. If μ lies in the interval $(5.22, 7.18)$, then $P(5.22 < \mu < 7.18) = 1$ and if it does not lie in the interval $(5.22, 7.18)$, then $P(5.22 < \mu < 7.18) = 0$. We can say that:-

$$P\left[\bar{X} - 1.96\left(\frac{2}{\sqrt{16}}\right) < \mu < \bar{X} + 1.96\left(\frac{2}{\sqrt{16}}\right)\right] = 0.95$$

or $P(\bar{X} - 0.98 < \mu < \bar{X} + 0.98) = 0.95$.

The interval $\bar{X} \pm 0.98$ is a random variable b/c '$\bar{X}$' does not have any particular numerical value but takes different values in different samples. It is therefore correct to say that:-

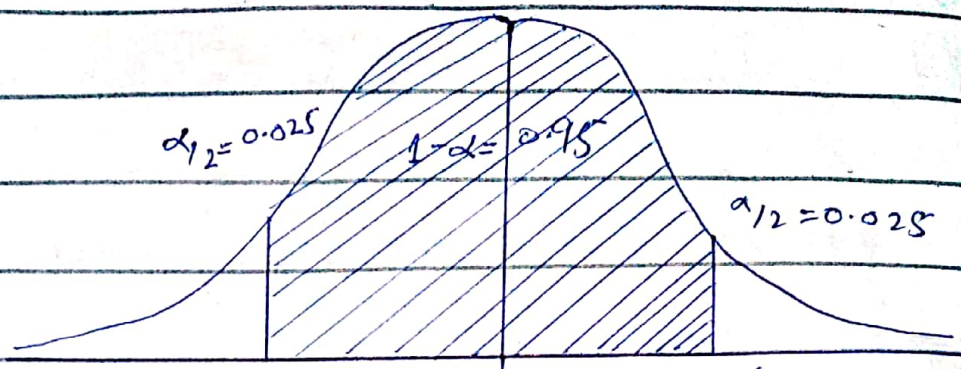
Scanned with CamScanner

$P(\bar{x} - 0.98 < \mu < \bar{x} + 0.98) = 0.95$.

meaning thereby that "the probability that the random interval $(\bar{x} - 0.98, \bar{x} + 0.98)$ covers the true value of 'μ' is 0.95". In other words, in repeated samples of size '16' from a normal population with S.D 2, the interval $(\bar{x} - 0.98, \bar{x} + 0.98)$ will contain the true unknown value of μ about 95 Percent of time.

To illustrate, let us draw '100' samples of '16' observations each calculate $\bar{x}$' for each sample and hence find the interval $(\bar{x}_i - 0.98, \bar{x}_i + 0.98)$ for each sample. These interval estimates based on '100' possible values of the random variable '$\bar{x}$' are shown in the figure on next page:

Scanned with CamScanner



$\bar{x}$-scale

-1.96 | 0 | $+1.96$

$\mu - 1.96\frac{\sigma}{\sqrt{n}}$ | μ | $\mu + 1.96\frac{\sigma}{\sqrt{n}}$

$\bar{x}_1 - 0.98$ | $\bar{x}_1$ | $\bar{x}_1 + 0.98$

$\bar{x}_2 - .98$ | $\bar{x}_2$ | $\bar{x}_2 + 0.98$

$\bar{x}_3 - 0.98$ | $\bar{x}_3$ | $\bar{x}_3 + 0.98$

(This interval does not cover μ)

$\bar{x}_{100} - .98$ | $\bar{x}_{100}$ | $\bar{x}_{100} + .98$

on the average, about 95 of these 100 intervals will contain the true value of μ. Thus we see that having taken our sample and found $\bar{x} = 6.2$, we cannot say that:-

$$P[6.2 - 0.98 \leq \mu \leq 6.2 + 0.98] = 0.95$$

Scanned with CamScanner

Rather, we say that we are 95 percent Confident that the true Population mean 'μ' will be in the interval (6.2-0.98, 6.2+0.98).

15.5.3 Confidence Interval for Difference of Means:-

To Construct the confidence interval for the difference btw two means, $\mu - \mu_2$, the following three cases are to be considered:-

(i) Both the Populations are normal with known standard deviations.

(b) Both the Populations are normal with unknown standard deviations.

(c) Both the Populations are non-normal, in which case, both sample sizes are necessarily large.

(i) Normal Populations with known Standard Deviations:-

Suppose we have two normal populations. Population '1' has an unknown mean 'μ' and a known S.D 'σ' and Population '2' has an unknown

mean 'μ_2' and a known S.D σ_2.
Independent samples of size 'n_1' and 'n_2' are taken from the Populations and Sample means calculated.
Let the Sample means be denoted by $\bar{X}_1$ and $\bar{X}_2$. Then the sampling distribution of the difference, $\bar{X}_1 - \bar{X}_2$ is normally distributed with a mean '$\mu_1 - \mu_2$', and a S.D $\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$.
In other words, the variable:-

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

is exactly standard normal, no matter how small the sample size are.
We can therefore make the following probability statement:-

$$P\left[-Z_{\alpha/2} < \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} < Z_{\alpha/2}\right] = 1 - \alpha.$$

Multiplying each term inside the bracket by $\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$, subtracting $(\bar{X}_1 - \bar{X}_2)$ and then multiplying by '-1' (inequality signs reversed), we get:-

Scanned with CamScanner

$$P[(\bar{X}_1 - \bar{X}_2) - Z_{\alpha/2}\sqrt{\frac{\sigma_1^2+\sigma_2^2}{n_1+n_2}} < (\mu_1-\mu_2) < (\bar{X}_1-\bar{X}_2)+Z_{\alpha/2}\sqrt{\frac{\sigma_1^2+\sigma_2^2}{n_1+n_2}}] = 1-\alpha$$

Hence the $100(1-\alpha)\%$ Confidence interval for Particular Samples obtained, for '$\mu_1 - \mu_2$' is:-

$$(\bar{x}_1 - \bar{x}_2) \pm Z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Example # 15.24:- Two independent samples of 100 mechanists and '100' carpenters are taken to estimate the difference b/w the weekly wages of the two categories of workers. The relevant data are given below:-

	Sample mean wage	Population variance
Mechanists	345	196
Carpenters	340	204

Determine the 95 and 99% confidence limits for the true difference b/w the average wages for mechanists and carpenters.

Sol:-

The 95% confidence limits for '$\mu_1 - \mu_2$' are given by:-

Scanned with CamScanner

$$(\bar{x}_1 - \bar{x}_2) + 1.96 \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Substituting the values, we get:-

$$(345 - 340) \pm 1.96 \sqrt{\frac{196}{100} + \frac{204}{100}}$$

or $5 \pm (1.96)(2)$

or 5 ± 3.92 or 1.08 to 8.92

Hence the 95% confidence limits for the true difference b/w the average weekly wages for mechanists and carpenters are 1.08 to 8.92.

Similarly, the 99% confidence limits for $\mu_1 - \mu_2$ are:-

$$(\bar{x}_1 - \bar{x}_2) \pm 2.58 \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

or $(345 - 340) \pm 2.58 \sqrt{\frac{196}{100} + \frac{204}{100}}$

or $5 \pm 2.58 (2)$

5 ± 5.16 or -0.16 to 10.16.

Thus the 99% confidence limits for the true difference between the average weekly wages are -0.16 to 10.16.

(ii) Normal Populations with unknown standard Deviations:-

when the independent samples of sizes 'n_1' and 'n_2' are

Scanned with CamScanner

drawn from normal populations with unknown standard deviations, we estimate them by the respective sample S.D. If sample sizes are sufficiently large, then we can assume that the sampling distribution of the difference $\bar{X}_1 - \bar{X}_2$, is approximately normal with mean $\mu_1 - \mu_2$ and standard deviation $\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$. Hence the $100(1-\alpha)$ percent confidence interval estimate for $\mu_1 - \mu_2$ goo particular samples obtained, is

$$(\bar{x}_1 - \bar{x}_2) \pm Z_{\alpha/2} \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$$

If, on the other hand, sample sizes are small and the populations have unknown equal standard deviations, then we use student's t-distribution to construct the confidence interval. The t-distribution shall be discussed in ch. 18.

Example 15.25:- A test in statistics was given to '50' girls and 75 boys. The girls made an average grade of '76' with a standard deviation

Scanned with CamScanner

of 6, while the boys made an average grade of '82' with a standard deviation of 8. Find a 96% confidence interval for the difference $'\mu_1 - \mu_2'$, where $'\mu_1'$ is the mean score of all boys and $'\mu_2'$ is the mean score of all girls who might take this test.

Sol:-

As the sample sizes are sufficiently large $(n_1, n_2 > 30)$, we can therefore use the sample standard deviations $'s_1'$ and $'s_2'$ in place of Population standard deviations $'\sigma_1'$ and $'\sigma_2'$. Assuming the Populations to be normally distributed, the 96% confidence interval for $'\mu_1 - \mu_2'$ is:-

$$(\bar{x}_1 - \bar{x}_2) \pm 2.504 \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

where $z_{\alpha/2}$ i.e $z_{0.02} = 2.054$.

Substituting the given values, we get:-

Scanned with CamScanner

$$(82-76) \pm 2.054\sqrt{\frac{(8)^2}{75}+\frac{(6)^2}{50}}$$

or $6 \pm 2.054\sqrt{\frac{64}{75}+\frac{36}{50}}$.

or $6 \pm (2.054)(1.254)$

or $6 \pm (2.58)$ or 3.42 to 8.58.

Hence the desired 96% Confidence interval for $\mu_1 - \mu_2$ calculated from the given values, is (3.42, 8.58).

(iii) Non-normal Populations:-

If the sample sizes are sufficiently large, then the Central limit theorem tells us that the sampling distribution of the distance $\bar{X}_1 - \bar{X}_2$ will be approximately normal even though the Populations are non-normal. An approximate $100(1-\alpha)$ Percent Confidence intervals for $\mu_1 - \mu_2$ when the Population Standard deviations are known, would be:-

$$(\bar{x}_1 - \bar{x}_2) \pm Z\alpha_{/2}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}$$

where $\bar{x}_1$ and $\bar{x}_2$ are the means of the two Particular

Scanned with CamScanner

Samples.

If the Population S.D's are unknown, then they are estimated by the Sample S.Ds. The approximate $100(1-\alpha)$ Percent Confidence interval for '$\mu_1 - \mu_2$' is then given by:-

$$(\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_1}}.$$

where 's_1' and 's_2' are the Sample s.Ds.

(Further on Photocopy).

Scanned with CamScanner

Statistical Inference: Hypothesis Testing.

16.1:- Introduction:-

Hypothesis testing is a very important phase of statistical inference. It is a procedure which enables us to decide on the basis of information obtained from sample data whether to accept or reject a statement or an assumption about the value of a Population Parameter. Such a statement or assumption which may or may not be true, is called a statistical hypothesis. we accept the hypothesis as being true, when it is supported by the sample data. we reject the hypothesis when the sample data fail to support it.

It is important to understand what we mean by the terms reject and accept in hypothesis testing. The rejection of a hypothesis is to

Scanned with CamScanner

declare it false. The acceptance of a hypothesis is to conclude that there is not sufficient evidence to reject it. Acceptance does not necessarily mean that the hypothesis is true.

16.1.1:- Null and Alternative Hypothesis:-

A null hypothesis, generally denoted by the symbol H_0, is any hypothesis which is to be tested for possible rejection under the assumption that it is true. Today the term is used for any hypothesis that is being tested. The word null in the term null hypothesis implies that usually 'H_0' is the hypothesis of no effect. A null hypothesis should always be precise such as "the given is unbiased" or "a drug is ineffective in curing a particular disease" or the difference b/w the two teaching methods is null or zero". The hypothesis is usually assigned a numerical value. For example, Suppose we

Scanned with CamScanner

think that the average height of students in all colleges is 62". This statement is taken as a hypothesis and is written symbolically as $H_0 : \mu = 62"$. In other words, we hypothesize that $\mu = 62"$.

An alternative hypothesis is any other hypothesis which we accept when the null hypothesis 'H_0' is rejected. It is customarily denoted by H_1 or H_A. A null hypothesis 'H_0' is thus tested against an alternative hypothesis H_1. For example, if our null hypothesis is $H_0 : \mu = 62"$, then our alternative hypothesis may be $H_1 : \mu \neq 62"$ or $H_1 : \mu > 62"$ or $H_1 : \mu < 62"$.

16.1.2:- Simple and Composite Hypotheses:-

A simple hypothesis is one in which all parameters of the distribution are specified. For example, if the heights of college

Scanned with CamScanner

students are normally distributed with $\sigma^2 = 4$, the hypothesis that its mean ill' is, say 62", that is $H: \mu = 62$, we have stated a simple hypothesis, as the mean and variance together specify a normal distribution completely. A simple hypothesis, in general, states that $\theta = \theta_0$ where 'θ_0' is the specified value of a parameter θ, (θ may represent $\mu, p, \mu_1 - \mu_2$ etc).

A hypothesis which is not simple (i.e in which not all of the parameters are specified) is called a composite hypothesis. For instance, if we hypothesize that $H: \mu > 62$ (and $\sigma^2 = 4$) or $H: \mu = 62$ and $\sigma^2 < 4$, the hypothesis becomes a composite hypothesis b/c we cannot know the exact distribution of the population in either case. Obviously, the parameters $\mu > 62$" and $\sigma^2 < 4$ have

Scanned with CamScanner

more than one value and no specified values are being assigned. In general form of a composite hypothesis is $\theta \leq \theta_0$ or $\theta \geq \theta_0$, that is the parameter 'θ' does not exceed or does not fall short of a specified value θ_0. The concept of simple and composite hypotheses applies to both null hypothesis and alternative hypothesis.

Hypotheses may also be classified as exact and inexact. A hypothesis is said to be an exact hypothesis if it selects a unique value for the parameter such as $H: \mu = 62$ or $p = 0.5$. A hypothesis is called an inexact hypothesis when it indicates more than one possible values for the parameter such as $H: \mu \neq 62$ or $H: p > 0.5$. A simple hypothesis must be an exact one while an exact hypothesis is not necessarily a simple hypothesis. An

Scanned with CamScanner

inexact hypothesis is a composite hypothesis.

16.1.3 Test- Statistic:-

A sample statistic which provides a basis for testing a null hypothesis, is called a test-statistic. Every test-statistic has a Probability (sampling) distribution which gives the Probability of obtaining a specified value of the test-statistic when the null hypothesis is true. It is important to remember that a test-statistic does not Prove the hypothesis to be correct but if furnishes an evidence against the hypothesis. The sampling distributions of the most commonly used test statistics are normal, t, chi-square or F.

16.1.4:- Acceptance and Rejection Regions:-

All Possible values which a test-statistic may assume can be divided into two mutally exclusive

groups:- One group consisting of values which appear to be consistent with the null hypothesis, and the other having values which are unlikely to occur if 'Ho' is true. The first group is called the acceptance region and the second set of values is known as the rejection region for a test. The rejection region is also called the critical region. The value(s) that separates the critical region from the acceptance region, is called the critical value(s). The critical value which can be in the same units as the parameter or in the standardized units, is to be decided by the experimenter keeping in view the degree of confidence he is willing to have in the null hypothesis.

16.1.5:- Type I and Type II Errors:-

when a peoform a hypothesis test, we drive the

evidence from the sample in the form of a test-statistic. There is a possibility that the sample evidence may lead us to make a wrong decision. we may reject a null hypothesis H_0, when it is, in fact, true or we may accept a null hypothesis H_0, when it is actually false. The form type is called an error of the first kind or a Type I-error, while the latter, an error of the second kind or a Type II error. The decision and the corresponding two types of error may be displayed in a tabular form as below:-

True situation	→ Decision	
	Accept H_0	Reject H_0 (or accept H_1)
H_0 is true	Correct decision (No error)	wrong decision (Type-I error)
H_0 is false	wrong decision (Type-II error)	Correct decision (No error).

Scanned with CamScanner

A legal analogy will help in understanding the difference between Type I and Type II errors. In a Court trial, the supposition of Law is that the accused (the defendant) is innocent. This supposition of innocence may be regarded as a kind of null hypothesis 'Ho' that is to be rejected or accepted. After having heard the evidence presented during the trial, the Judge arrives at a decision. Suppose the accused is, in fact, innocent (i.e Ho is true), but the finding of the Judge is guilty. The Judge has rejected a true null hypothesis and in so doing, has made a Type I error. If, on the other hand, the accused is, in fact, guilty (i.e Ho is false) and the finding of the Judge is innocent, the Judge has accepted a false null hypothesis and by accepting a false hypothesis,

Scanned with CamScanner

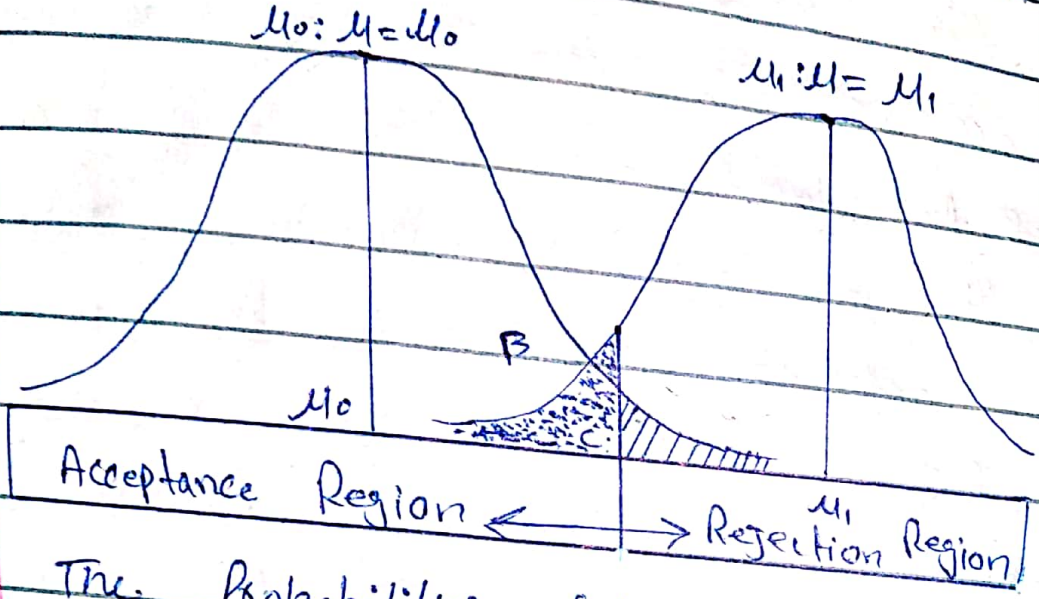
he has Committed a Type II error. The Probability of making a Type I error is Conventionally denoted by α (alpha) and that of Committing a Type II error is indicated by 'β' (beta). Thus 'α' is the Probability of rejecting 'H_0' when 'H_0' is true and 'β' is the Probability of accepting 'H_0' when 'H_0' is false (i.e. H_1 is true). In symbols, we may write:-

$$\alpha = P(\text{Type I error}) = P(\text{reject } H_0 / H_0 \text{ is true})$$

$$\beta = P(\text{Type II error}) = P(\text{accept } H_0 / H_0 \text{ is false})$$

Let us Consider two distributions:- one under the null hypothesis $H_0: \mu = \mu_0$ (i.e. distribution assuming 'H_0' is true) and the other under alternative hypothesis $H_1: \mu = \mu_1$ (i.e. distributing assuming 'H_1' is true).

Scanned with CamScanner



The Probabilities of 'α' and 'β' are the shaded and dotted areas respectively of the distributions under the null hypothesis and under the alternative hypothesis. When our null hypothesis 'H_0' is true, then any value greater than or equal to C (the critical point) constitutes the rejection region equal to α (one-sided). That is 'α' is associated with extreme values of the μ_0-distribution. The commonly used values of 'α' are 0.05 and 0.01. on the other hand, 'β' is associated with the area

under the H_1 distribution in the acceptance region established from H_0-distribution. The probability of accepting 'H_0' when 'H_1' is true. i.e. β, thus depends both on the null hypothesis 'H_0' and on the alternative hypothesis H_1. In order to determine β (the probability of Type II error) we require α (the probability of Type I error) and the values of both 'H_0' and 'H_1'. When 'α' becomes smaller, β tends to become larger and when 'α' becomes larger, β tends to become smaller. Thus there is an inverse relationship between 'α' and 'β'. We can decrease both 'α' and 'β' by increasing the sample size.

Example#16.1 :- The proportion of adults living in a small town from are matriculates is estimated to be $p=0.8$. To test this hypothesis a random sample of 15 adults is selected. If the

Scanned with CamScanner

number of matriculates in our sample is anywhere from '2' to 7, we shall accept the null hypothesis that $p=0.3$; otherwise we shall conclude that $p \neq 0.3$.

Evaluate 'α' assuming $p=0.3$.

Evaluate 'β' for the alternatives $p=0.2$ and $p=0.4$.

Sol:-

The null and alternative hypothesis are given as:-

$H_0: p = 0.3$ and $H_1: p \neq 0.3$

Let 'x' denote the number of adults who are matriculates. Then the test-statistic has the binomial distribution with $p=0.3$ and $n=15$. The acceptance region, as given, consist of all values from $x=2$ to $x=7$. Then the critical region is composed of two parts:- all values less than 2 and all values greater than 7. Thus the Probability of making Type I

error ie α Consists of $P(X<2)$ and $P(X>7)$.

Hence $\alpha = P(X<2 \text{ when } p=0.3) + P(X>7 \text{ when } p=0.3)$

$$= \sum_{x=0}^{1} b(x; 15, 0.3) + \sum_{x=8}^{15} b(x; 15, 0.3)$$

$$= \sum_{x=0}^{1} b(x; 15, 0.3) + [1 - \sum_{x=0}^{7} b(x; 18, 0.3$$

$= 0.0353 + [1 - 0.9500]$ (From Binomial Probability tables)

$= 0.0853$

To compute β, the Probability of Type II error, we need a specific alternative hypothesis. Now, we are given $H_0: p = 0.3$; and $H_1: p = 0.2$. A Type II error results when a false null hypothesis is accepted. That is a Type II error if any error occurs if any value of the distribution under $H_1: p = 0.2$ falls in the region $X = 2$ to $X = 7$, the acceptance region of the

distribution under null hypothesis

$H_0: p = 0.3$.

Hence $\beta = P(2 \leq X \leq 7 \text{ when } H_1: p = 0.2)$

$$= \sum_{x=2}^{7} b(x; 15, 0.2)$$

$$= \sum_{x=0}^{7} b(x; 15, 0.2) - \sum_{x=0}^{1} b(x; 15, 0.2)$$

$= 0.9958 - 0.1671$ (From binomial probability tables)

$= 0.8287$

Similarly, when $H_1: p = 0.4$, we have:-

$$\beta = P(2 \leq X \leq 7, \text{ when } p = 0.4)$$

$$= \sum_{x=2}^{7} b(x; 15, 0.4)$$

$$= \sum_{x=0}^{7} b(x; 15, 0.4) - \sum_{x=0}^{1} b(x; 15, 0.4)$$

$= 0.7869 - 0.0052 = 0.7817$

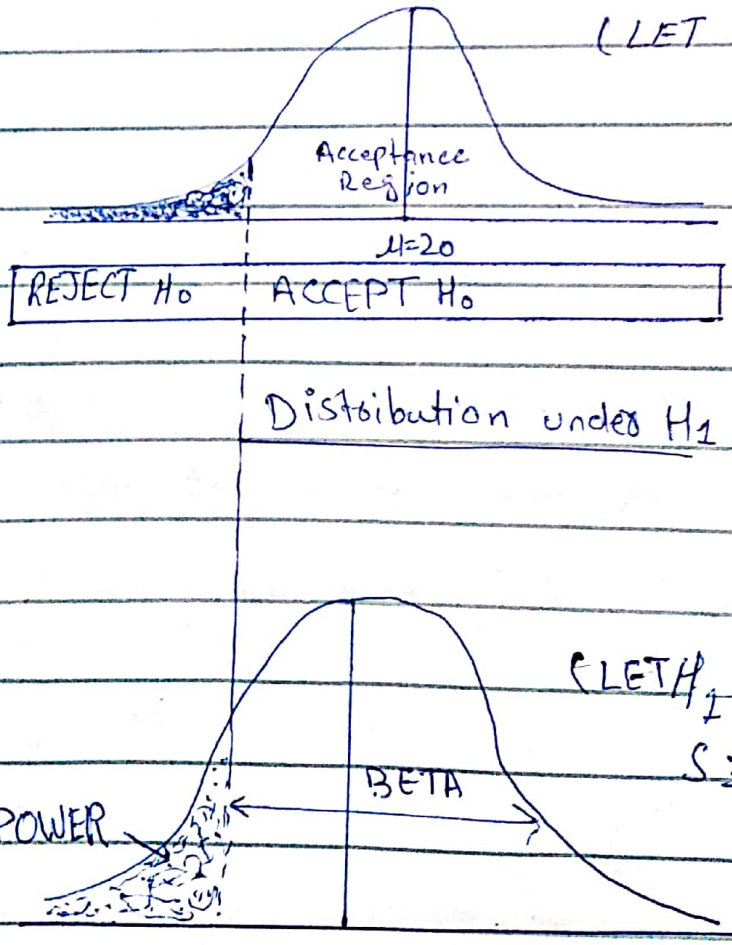
16.1.6. The Power of a Test with respect to a specified alternative hypothesis, is the probability of rejecting a null hypothesis when it is actually false. The power is the complement of β, the probability

Scanned with CamScanner

Committing a Type II error. It is therefore numerically equivalent to minus β. Symbolically,

Power = P(reject H_0 / H_0 is false)
= 1 − β.

To represent α, β and Power of a test graphically, we show the distribution of the test-statistic under both hypotheses 'H_0' and H_1 as below:-



Scanned with CamScanner

The shaded area in the lower diagram represents power. This Probability corresponds to the rejection region of the distribution under H_0. The Power generally increases with an increases in the Sample size. A test for which 'β' is small is defined to be a Powerful test. A Curve giving the Probabilities of making Type II errors for various Parametric values under alternative hypothesis, is called an operating characteristic curve or simply the OC curve. The power curve which may be regarded as the complement of the OC curve, shows the Probabilities of rejecting the null hypothesis 'H_0' for various values of the Parameter θ.

(Size of critical region)

16.1.7:- The Significance level:-

The Significance level of a test is the Probability

used as a standard for rejecting a null hypothesis 'H₀' when 'H₀' is assumed to be true. This probability is equal to some pre-assigned value, conventionally denoted by α. The value 'α' is also known as the size of the critical region. It is note-worthy that the significance level and the probability of Type I error are equivalent. The most frequency used values of α, the significance level, are 0.05 and 0.01, i.e 5 percent and 1 percent but occasionally 0.10 or 0.001 is used. By α=5%, we mean that there are about '5' chances in '100' of incorrectly rejecting a true null hypothesis. To put it in another way, we say that we are 95% confident in making the correct decision.

Scanned with CamScanner

16.1.8 Test of Signifance:-

A test of signifance is a rule or Procedure by which sample results are used to decide whether to accept or reject a null hypothesis. Such a Procedure is usually based on a test-statistic and the sampling distribution of such a statistic under H_0. A value of the statistic is said to be statistically significant when the Probability of its occurence under 'H_0' is equal to or less than the significance level α, that is the value falls in the rejection region, 'H_0' in this case is rejected. If, on the other hand, the value falls in the acceptance region, it is said to be statistically insignificant. In this case, 'H_0' may be accepted. These are two

Scanned with CamScanner

desirable qualities for a test of Significance. First, when the null hypothesis is actually true, it must have a low Probability of rejecting H_0, and secondly, when 'H_0' is actually false, it must have a high Probability of rejecting H_0. It is to be noted that the word Significant is used in a special sense.

Scanned with CamScanner