Sample Theory

Population :- A population refers to the entire set of individuals, items, or data points that someone is interested in studying or analyzing.
It includes every possible observation or data point relevant to the research.

Ex :- (a) If you're studying the average height of adults in a country, the population would be "all the adults in that country."

(b) If a company wants to know how satisfied its customers are, the population could be "all customers who have ever bought a product from them.

Sample :- A finite subset of the population, selected for examinations or study of the population

Example :- (a) In (a) part; you might take a sample of 1000 adults from different regions of the country.

(b) In (b) part :- For the company, measuring customers satisfaction; a sample might include 500 recent customers who completed a survey.

Parameter and Statistics :-

Parameter :- A statistical measure which describes the characteristics of a population:
Example of parameters include the population mean 'μ', population standard deviation 'σ',

population proportion, etc.

Statistics:- A statistical measure describing the characteristics of a sample.

Example of statistics include the sample mean ($\bar{x}$), sample standard deviation (s), sample proportion, etc.

Sampling Error:- variation among sample statistics due to change.

Definition:- sampling error is the discrepancy between the sample statistics and the true population parameter.

- It occurs because a sample is only a subset of the population, and therefore might not reflect all the characteristics of the population.
- It occurs due to the inherent limitations of using a sample rather than the entire population.

Non-sampling Errors:- These are not related to the sample size or method, but instead come from other factors like data collection errors, measurement errors, or biases.

Standard Error:- The standard deviation of the sampling distribution of a statistics. and denoted as:

$$SE = \frac{\sigma}{\sqrt{n}},$$

- σ : standard deviation of the population
- n : sample size.

Note :- Sampling Error is a difference amount between a particular sample and the population.

where as;

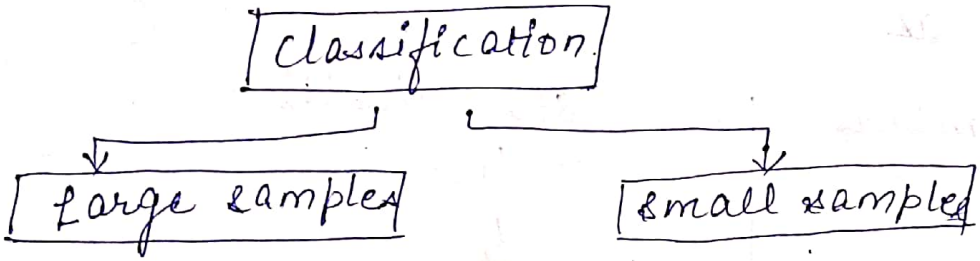
Standard Error is a measure of the variability or spread of a sample statistics if we were to take many samples from the population.

Classification of Samples

Number of possible samples :→ Taking a population of size N, drawing a sample of size 'n', then the total number of possible samples is

$$NC_n = \frac{(N)!}{n!\,(N-n)!} = K$$

where k is a constant.



→ Large samples :- If the size of the sample of $n \geq 30$, then it is said to be large sample.

→ Small samples :- If the size of the sample $n < 30$, then it is said to be small sample or exact sample.

Sampling distribution

A sampling distribution is a probability distribution of a statistics that is obtained through repeated sampling of a specific population.

→ sampling distribution describes a range of possible outcomes for a statistic, such as the mean, mode, etc.

If considering a 'N' size population and drawing the sample of size n, then the sampling distribution of any statistic, say 't', (like mean or maybe variance), and 't' can be defined as a function of sample observations.

i.e. $t = t(x_1, x_2 \text{——} x_n)$,

with 'K' possible samples of equal size 'n', can be defined as.

Sample Numbers	Statistics (t)
1	t_1
2	t_2
3	t_3
4	t_4
|	|
K	t_K

The values $t_1, t_2, \text{——} t_K$ determine the sampling distribution of a statistics 't'.

- If we want to compute the statistical constant for this statistic(t), then mean can be obtained by

$$\bar{t} = \frac{t_1 + t_2 + \text{——} + t_K}{K} = \frac{1}{K} \sum_{i=1}^{K} t_i.$$

The variance of a statistic 't' is

$$Var(t) = (t_1 - \bar{t})^2 + \cdots (t_k - \bar{t})^2$$

$$= \frac{1}{k} \sum_{i=1}^{k} (t_i - \bar{t})^2$$

Drawing of samples from a population (N size)

(a) ordered sampling with replacement $= N^n$

(b) Unordered sampling with replacement $= {}^{N+n-1}C_{N+1}$

(c) Ordered sampling without replacement $= n!\ {}^{N}C_n$

(d) Unordered sampling without replacement $= {}^{N}C_n$

Example :- Population $= \{A, B, C, D\}$

Sample size $= n = 2$.

Formation of sample $= (\_, \_)$.

(a) $N^n = 4^2 = \underline{16}$

(b) ${}^{N+n-1}C_{N-1} = {}^{4+2-1}C_{4-1} = {}^{5}C_3 = \frac{5 \cdot 4 \cdot 3!}{3!\, 2} = \underline{10}$

(c) $n!\ {}^{N}C_n = \frac{4 \cdot 3 \cdot 2!}{2!\, 2!} = \underline{12}$

(d) ${}^{N}C_n = {}^{4}C_2 = \frac{4 \cdot 3 \cdot 2!}{2 \cdot 2!} = \underline{6}$.

Central Limit Theorem

Theorem :- If $\bar{x}$ is the mean of a random sample of size n taken from a population with mean μ and variance σ^2, then the sampling distribution of the sample mean tends to be normal distribution with mean μ and variance $\frac{\sigma^2}{n}$ as the sample size tends to be large ($n \geq 30$), regardless the form of the parent population,

i.e $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$.

NOTE : ① In term of ~~standaar~~ standard normal variate, we can write it as

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1).$$

② In case of $n < 30$, the approximation is good only if the population is not too different from a normal distribution.

③ If the population is to be normal, then the sampling distribution of $\bar{x}$ is normal for any size of n.

④ CLT is really useful because it characterizes large samples from any distribution.

Examples

Ex 1 :- Let $X_1, X_2, \dots X_n, \dots$ be a independently and identically distribution Poisson variates with parameter λ. Use CLT to estimate $P(120 \le S_n \le 160)$. where

$$S_n = X_1 + X_2 + \dots + X_n \quad ; \quad \lambda = 2, \text{ \& } \quad n = 75.$$

Sol^n :- Since X_i's are i.i.d $P(\lambda)$

$$E(X_i) = \lambda$$
$$V(X_i) = \lambda \quad ; \quad i = 1, 2, \dots n.$$

Expected value of the random variable S_n is

$$E(S_n) = E(X_1 + X_2 + \dots + X_n)$$
$$= E(X_1) + E(X_2) + \dots + E(X_n)$$
$$= \underbrace{\lambda + \lambda + \dots + \lambda}_{n\text{-times}}$$
$$= n\lambda$$

similarly for variance;

$$\text{Var}(S_n) = \text{var}[X_1 + \dots + X_n]$$
$$= \text{var}(X_1) + \dots + \text{var}(X_n)$$
$$= n\lambda$$

which implies: By CLT

$$S_n \sim N(n\lambda, n\lambda)$$

(or exactly in terms of sample mean)

$$\bar{X} = \frac{S_n}{n} \sim N\left(\lambda, \frac{\lambda}{n}\right)$$

$S_n \sim N(75\times\lambda, 75*\lambda)$

$N(\mu=150, \sigma^2=150)$

$$P(120 \le S_n \le 160) = P\left(\frac{120-150}{\sqrt{150}} \le Z \le \frac{160-150}{\sqrt{150}}\right)$$

$$= P(-2.45 \le Z \le 0.82)$$

$\because Z \sim N(0,1)$

$$= P(-2.45 \le Z \le 0) + P(0 < Z \le 0.82)$$

$$= P(0 \le Z \le 2.45) + P(0 < Z \le 0.82)$$

$$= 0.4929 + 0.2939 \quad (\text{By } z\text{-table})$$

$$= 0.7868$$

Example 2:- The mean and standard deviation of the tax value of all vehicles registered in a certain state are $\mu = \$13,525$, and $\sigma = \$4,180$. Suppose random samples of size 100 are drawn from the population of vehicles. What are the mean $\mu_{\bar{x}}$ and standard deviation $\sigma_{\bar{x}}$ of the sample mean of $\bar{x}$?

Sol^n:- $n = 100$,

$\mu_{\bar{x}} = \mu$ (equal to population mean)

$= \$13,525$

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{\$4.180}{\sqrt{100}} = \$418$$

Ex 3. An automobile battery manufacturer claims that its midgrade battery has a mean of life of 50 months with a standard deviation of 6 months. Suppose the distribution of battery lives of this particular brand is approximately normal.

(a) On the assumption that the manufacturer's claim are true, find the probability that a randomly selected battery of this type will last less than 48 months.

(b) On the same assumption, find the probability that the mean of a random sample of 36 such batteries will be less than 48 months.

Sol^n :- (a) Population X ~ normal

$$P(X < 48) = P\left(Z < \frac{48-50}{6}\right)$$

$$= P(Z < -0.33)$$

$$= 0.3707$$

(b) $\mu_{\bar{x}} = \mu = 50$, $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{6}{\sqrt{36}} = 1.$

$$P(\bar{X} < 48) = P\left(Z < \frac{48 - \mu_{\bar{x}}}{\sigma_{\bar{x}}}\right)$$

$$= P(Z < 48-50)$$

$$= P(Z < -2) = 0.0228$$

Sampling Distribution of Sample Mean for a normal population

Situations :- • For example;

→ An investigator may interested to estimate average income of the population of particular area.

→ To estimate the average life of electric bulbs manufactured by a company

→ A researcher may want to estimate the mean time required to complete a certain analysis, and many more.

In the above cases, an estimate of population mean is required and one may estimate this on the basis of a sample taken from that population. For this, sampling distribution of sample mean is required.

"The probability distribution of all possible values of sample mean that would be obtained by drawing all possible samples of the same size from all the population is called **sampling distribution of sample mean.**" (say SDSM for further study).

Shape of sampling distribution of sample mean:

For different popopulation with varying sample size, we have the following conclusion:

N = normal dist^n

D = any distribution, other than normall

n = sample size

X = population

If $X \sim N(\mu, \sigma^2)$ n = large then $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$	If $X \sim N(\mu, \sigma^2)$ n = small then $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$
If $X \sim D(\mu, \sigma^2)$ n = large then $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$	If $X \sim D$ n = small then Dist^n of $\bar{X}$ = having no specific distrib-ution

The mean and variance of SDSM :-

By selecting one sample from the population, we will draw the inferences about the population.

Let $X_1, X_2, \ldots X_n$ be a random sample of size n taken from a ~~natu~~ normal population with mean μ and variance σ^2 then it has also established that sampling distribution of sample mean $\bar{X}$ is also normal. The mean and the variance of sampling distribution of $\bar{X}$ can be obtained as

mean of $(\bar{X}) = E(\bar{X})$ where $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$

$= E\left(\frac{X_1 + \ldots + X_n}{n}\right)$ ~~and var~~

$= \frac{1}{n}[E(X_1) + \ldots E(X_n)]$

$\because E(X_i) = \mu \quad \forall i \in \{1, 2, \ldots n\}$

$= \frac{1}{n} \times n\mu$

$= \mu$. (= equal to population mean)

variance of $\bar{X} = \text{Var}(\bar{X})$

$= \text{Var}\left(\frac{X_1 + \ldots + X_n}{n}\right)$

$= \frac{1}{n^2}[\text{Var}(X_1) + \ldots + \text{Var}(X_n)]$

$\because \text{Var}(X_i) = \sigma^2 \; \forall i$

$= \frac{1}{n^2} \times n\sigma^2$

$= \frac{\sigma^2}{n}$.

Hence, we conclude that if the samples are drawn from normal population with μ and σ^2 as mean and variance resp., then the

SDSM is also normal i.e

if $X_i \sim N(\mu, \sigma^2)$

then $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$.

The standard Error of sample mean is the standard deviation (by definition)

$$SE(\bar{X}) = SD(\bar{X}) = \frac{\sigma}{\sqrt{n}}.$$

NOTE :- We assume that the random sample is selected from the infinite popul-ation or from a finite population with replacement.

Examples

Ex4. Diameter of a steel ball bearing produced on a semi-automatic machine is known to be distributed normally with mean 12 cm and standard deviation 0.1 cm. If we take a random sample of size 10. then find,

(i) mean and ~~deviation~~ variance of sampling distribution of mean.

(ii) The probability that the sample mean lies between 11.95 cm and 12.05 cm.

Sol^n:- Population mean $= \mu = 12$

$\sigma = 0.1$

$n = 10.$

(i) $\bar{X} \sim N(\mu = 12, \sigma^2 = 0.01)$

$E(\bar{X}) = \mu = 12$

$\text{var}(\bar{X}) = \frac{\sigma^2}{n} = \frac{0.01}{n} = 0.001.$

(ii) $P[11.95 \leq \bar{X} < 12.05]$

$= P\left[\frac{\bar{X} - 12}{0.001} \leq \right.$

$= P\left(\frac{11.95 - 12}{\sqrt{0.001}} \leq \frac{\bar{X} - 12}{\sqrt{0.001}} \leq \frac{12.05 - 12}{\sqrt{0.001}} \right)$

$= P(-1.67 < Z < 1.67)$

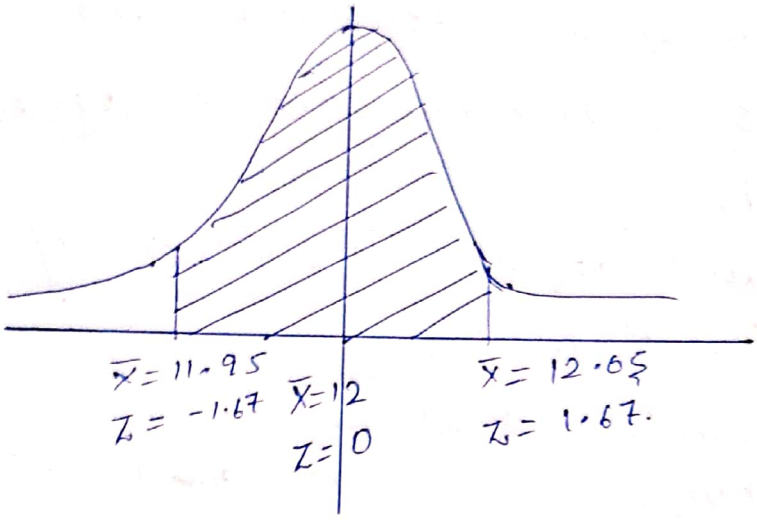
$= P[-1.67 < Z < 0] + P[0 < Z < 1.67)$

$= P[0 < Z < 1.67) + P[0 < Z < 1.67)$

$= 2P[0 < Z < 1.67]$

$= 2 \times 0.4525$ ∵ By Z-table

$= 0.9050.$



NOTE :- In above discussion, the variance is known.

Ex 5 :- The weight of certain type of a truck tyre is known to be distributed normally with mean 200 pounds and standard deviation 4 pounds. A random sample of 10 tyre is selected.

(i) What is the sampling distributed of sample mean ? also obtained the mean and variance of this dist?

(ii) Find the probability that the mean of this sample is greater than or equal to 202 pounds.

Sol:- $\mu = 200, \quad \sigma = 4, \quad n = 10.$

(i) parent population $X \sim N(200, 4^2)$

$$\mu = E(\bar{X}) = 200$$

$$\text{var}(\bar{X}) = \frac{\sigma^2}{n} = \frac{16}{10} = 1.6$$

(ii) Now, the probability that the sample mean $\bar{X}$ is greater than or equal to 202 pounds;

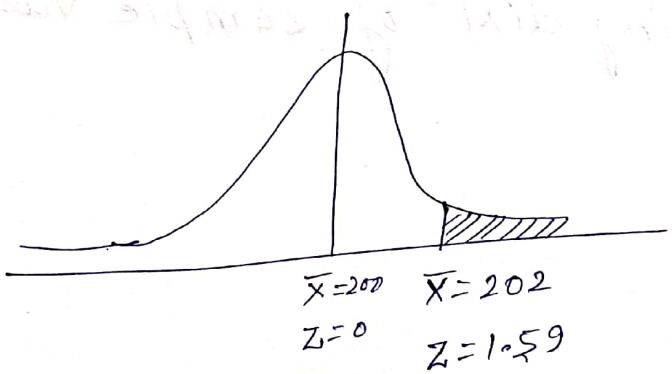
$$P[\bar{X} \geq 202] = P\left(\frac{\bar{X} - 200}{\sqrt{1.6}} \geq \frac{202 - 200}{\sqrt{1.6}}\right)$$

$$= P(Z \geq 1.59)$$

$$= 0.5 - P[0 \leq Z \leq 1.59]$$

$$= 0.5 - 0.4441$$

$$= 0.0559$$



Sampling distribution of sample variance for a normal population

In many practical situations, one may want the variability.

For example :-
- A life insurance company may be interested in the variation of the no. of polices in different years.
- A manufacturing of steel ball bearings may want to know about the variation of diameter of steel ball bearing.

In such cases, we need the information about the sampling dist^n of sample variance.

"The probability distribution of all values of the sample variance would be obtained by drawing the all possible samples of same size from the parent population is called the sampling dist^n of the sample variance."

In general, the sample variance of a sample of size n, which is constituted by $x_1, x_2 \ldots x_n$ random variables from a population with mean μ and variance σ^2 is defined as;

$$s^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2.$$

Now, we are interested in finding expected value of the sample variance s^2;

$$E(s^2) = E\left(\frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2 \right)$$

$$= \frac{1}{n} E\left(\sum_{i=1}^{n} ((x_i - \mu) - (\bar{x} - \mu))^2 \right)$$

$$= \frac{1}{n} E\left(\sum_{i=1}^{n} [(x_i - \mu)^2 + (\bar{x} - \mu)^2 - 2(\bar{x} - \mu)(x_i - \mu)] \right)$$

$$= \frac{1}{n} \left[\sum_{i=1}^{n} E(x_i - \mu)^2 + E(\bar{x} - \mu)^2 \sum_{i=1}^{n} 1 - 2E\left((\bar{x} - \mu) \left(\sum_{i=1}^{n} \frac{x_i}{n} - \mu \right) \times n \right) \right]$$

$$= \frac{1}{n} \left[\sum_{i=1}^{n} \text{var}(x_i) + \text{var}(\bar{x}) \cdot n - 2E\left((\bar{x} - \mu)^2 \cdot n \right) \right]$$

$$= \frac{1}{n} \left[\sigma^2 \cdot n + \frac{\sigma^2}{n} \cdot n - 2\,\text{var}(\bar{x}) \cdot n \right]$$

$$= \frac{1}{n} \left[\sigma^2 \cdot n + \sigma^2 - 2\frac{\sigma^2}{n} \cdot n \right]$$

$$= \frac{1}{n} \sigma^2 (n - 1)$$

$$E(s^2) = \frac{n-1}{n} \sigma^2.$$

Hence $E(s^2) \neq \sigma^2$, sample variance 's^2' is not an unbiased estimate of population variance.

To make it unbiased we have to modify it as

$$E(s^2) = \frac{n-1}{n}\sigma^2.$$

$$\frac{n}{n-1} E(s^2) = \sigma^2$$

$$E\left(\frac{n}{n-1}\sum_{i=1}^{n}\frac{(x_i-\bar{x})^2}{n}\right) = \sigma^2$$

$$E\left(\frac{1}{(n-1)}\sum_{i=1}^{n}(x_i-\bar{x})^2\right) = \sigma^2$$

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

Thus, S^2 is an unbiased estimate of population variance.

Note :- (1) In small sampling theory, whenever σ^2 is not known, we will use S^2 for practical problems.

(2) As $s^2 = \left(1-\frac{1}{n}\right)S^2$

Hence for large n we obtained

$$s^2 = S^2 \quad \text{as } n \to \infty.$$

Thus, for large ~~small~~ samples ($n \to \infty$) we can take $\sigma^2 = s^2$

(3) For further studies, sample variance means S^2.

Sampling distribution of sample variance s^2, cannot be predicted alike the sample mean. There is not any statement like CLT.

Here, it is to be noted that s^2 always be positive, which implies s^2 does not follow normal distribution as normal takes neg. and posi. values together.

In order to work with sample variance, we need to make a transformation to s^2 by multiplying $\left(\frac{n-1}{\sigma^2}\right)$, which gives a new variable i.e.

$$\chi^2 = \frac{s^2(n-1)}{\sigma^2} \rightsquigarrow \chi^2_{n-1},$$

here the new variable is χ^2 which follows the chi-square distribution with $(n-1)$ degree of freedom.

Note:- For numerical examples, in order to solve the problems related to sampling distribution of sample variance, we convert s^2 into χ^2-variate by above transform. For particular values of χ^2 we have χ^2-table alike z-table.

Ex.6 :- A manufacturer of steel ball bearings has found the variance of diameter of ball bearings 0.0018 inches^2. What is the probability that a random sample of 25 ball bearings will result in a sample variance at least 0.002 inches^2?

Sol^n :- $\sigma^2 = 0.0018$, $n = 25$

The probability that the sample variance is at least 0.002 inches^2 is given;

$$P[S^2 \geq 0.002] = ?$$

$$P(S^2 \geq 0.002) = P\left(\frac{(n-1)S^2}{\sigma^2} \geq \frac{24 \times 0.002}{0.0018}\right)$$

$$= P\left(\chi^2_{24} \geq 26.67\right)$$

Using chi-square table; check the value corresponding to (25-1) degree of freedom 'row' and α level of column, if the exact value is not matching then pick a range for given value $x = 26.67$, this value lies between 15.659 and 33.196 of '0.9' and '0.1' significance level (α) in 24 df row.

$P(\chi^2_{24} \geq 15.659) = 0.9$ and $P(\chi^2_{24} \geq 33.20) = 0.10$.

$\Rightarrow$ $0.10 < P(\chi^2_{24} \geq 26.67) < 0.90$

$\Rightarrow$ $0.10 < P(\chi^2_{24} S^2 \geq 0.2) < 0.90$

Degree of freedom

The maximum number of the independent values, which are free to vary is called degree of freedom; can be determined as

$df = N - p$, where N is sample size and p is given condition / parameter or relationship.

For example :- ① $x_1 + x_2 + x_3 + x_4 = 0$

In this example, we have total 4 variables, and given condition or relationship between these unknown variable is one.

$df = 4 - 1 = 3$,

i.e. we are free to choose any '3' variable and then the fourth one will automatically be dependent on other '3' variables at chosen first.

② In a symmetric matrix 'A' of $n \times n$ order

$$A = \begin{bmatrix} a_{11} & \square & \square \\ a_{21} & a_{22} & \\ | & \diagdown & | \\ a_{n1} & & a_{nn} \end{bmatrix}$$

, $df = \frac{n(n+1)}{2}$, here are free to choose Lower triang-ular / upper triangular matrix, having total no. of elements are $\frac{n(n+1)}{2}$.

Note:- ① degree of freedom is very useful in chi-square test, t-test, F-test.

Chi-Square Distribution

Let X_i, $(i = 1, 2, \ldots n)$ are n independent normal variates with means μ_i and variance σ_i^2, $(i = 1, 2 \ldots n)$, then

$$\chi^2 = \sum_{i=1}^{n} \left(\frac{X_i - \mu_i}{\sigma_i} \right)^2$$ is a chi-square variate with 'n' degree of freedom.

NOTE → (1) If a random variable X has a chi-square dist^n with 'n' d.f., i.e. $X \sim \chi^2_n$ then its p.d.f is

$$f(x) = \frac{1}{2^{n/2} \lceil n/2} e^{-x/2} x^{n/2 - 1} \; ; \; 0 \le x < \infty.$$

(2) Let $X \sim N(\mu, \sigma^2)$, then $Z = \frac{x - \mu}{\sigma} \sim N(0, 1)$ and $Z^2 = \left(\frac{x - \mu}{\sigma} \right)^2 \sim \chi^2_1$, i.e is chi-square with '1' d.f.

(3) For large samples ($n \ge 30$), the χ^2- ~~variate~~ distribution tends to normal distribution.
It can be easily proved by moment generating function.
Therefore, in practice for $n \ge 30$, we can use the normal distribution tables for testing the significance of the value of χ^2.
That's why in χ^2- table, the significant values of χ^2 have been given till $n = 30$ only.

(4) The mean of the chi-square dist^n is

$$E(\chi^2_n) = n$$

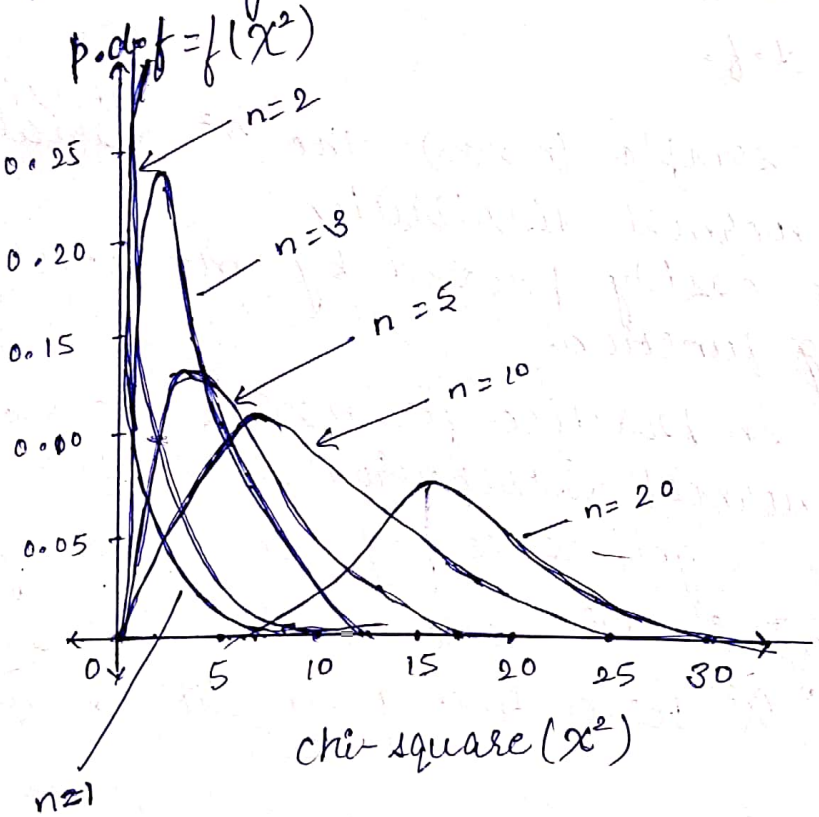
(5) The variance is

$$Var(\chi^2_n) = 2n$$

(6) Since the chi-square dist^n is the sum of std. normal random variables so it cannot be negative, i.e. it begin at zero and continues to infinity.

(7) It has only one parameter 'n' i.e. degree of freedom. The shape of the distribution depends on it, therefore, there is a different chi-square distribution for each value of 'n'.

(8) The chi-square distribution is positively skewed and asymmetrical.



Example

Ex. 7 :- If X is chi-square variate with n d.o.f., then prove that for large n, $\sqrt{2X} \sim N(\sqrt{2n}, 1)$.

Sol^n :- Since $X \sim \chi^2_n$

$$E(X) = n, \quad V(X) = 2n$$

$$Z = \frac{X - E(X)}{\sqrt{V(X)}} = \frac{X-n}{\sqrt{2n}} \sim N(0,1) \text{ for large } n \quad \text{---(1)}$$

$$P\left(\frac{X-n}{\sqrt{2n}} \le Z\right) = P(X \le Z\sqrt{2n} + n)$$

$$= P\left(\sqrt{2X} \le \sqrt{(2Z\sqrt{2n} + 2n)}\right)$$

$$= P\left(\sqrt{2X} \le \sqrt{2n}\left(\frac{\sqrt{2}}{\sqrt{n}}Z + 1\right)^{1/2}\right)$$

using fractional Binomial formula.

i.e.

$$(1+x)^{1/2} = 1 + \frac{1}{2}x + \frac{\left(\frac{1}{2}\right)\left(-\frac{1}{2}\right)}{2!}x^2 + \text{---} + \left(\frac{1}{2}\right)\left(-\frac{1}{2}\right)\cdots$$

$$\frac{\cdots\left(-\frac{2k-3}{2}\right)}{k!}x^k + \text{- - - - -}$$

valid for $|x| < 1$.

$$= P\left[\sqrt{2X} \le \sqrt{2n}\left(1 + \frac{Z\sqrt{\frac{2}{n}}}{2!} + \frac{\left(\frac{1}{2}\right)\left(-\frac{1}{2}\right)}{2!}\left(Z\sqrt{\frac{2}{n}}\right)^2 + \text{- - -}\right)\right]$$

$$= P\left[\sqrt{2X} \le \left(\sqrt{2n} + \sqrt{2n}\frac{Z}{\sqrt{2n}} - \sqrt{2n}\frac{Z^2}{4n} + \text{---}\right)\right]$$

$$= P\left[\sqrt{2X} \le \sqrt{2n} + Z\right] \text{ for large } n \text{ i.e } n \to \infty.$$

$$P\left(\frac{\chi - n}{\sqrt{2n}} \le z\right) = P\left[\sqrt{2\chi} - \sqrt{2n} \le z\right]$$ for large n.

We have a condition for large n in Eq (1)

$$z = \frac{\chi - n}{\sqrt{2n}} \sim N(0,1) \text{ for large } n.$$

$$\sqrt{2\chi} - \sqrt{2n} \sim N(0,1) \text{ for large } n.$$

i.e. $$\frac{\sqrt{2\chi} - \text{mean}}{\sqrt{\text{variance}}} \sim N(\text{mean, variance})$$

$$\sqrt{2\chi} \sim N(\sqrt{2n}, 1)$$

Ex. 8: → The weight of certain type of truck tyres has a variance of 11 pounds^2. A random sample of 20 tyres is selected. What is the probability that the variance of this sample is greater than or equal to 16 pounds^2?

Soln:-

$\sigma^2 = 11, \; n = 20$

The probability that the given sample is greater than 16 pounds^2 is given by $P[s^2 \le 16]$

→ Convert s^2 into χ^2, Here d.o.f $= 20 - 1 = 19$

$$\chi^2_{19} = \frac{(n-1)s^2}{\sigma^2} = \frac{19 s^2}{11}$$

$$P\left(\frac{19S^2}{11} \geq \frac{16 \cdot 19}{11}\right) = P\left(\chi^2_{19} \geq 27.64\right)$$

from chi-square table,

The value, say $x_c = 27.64$, $\longrightarrow = x_a$

x_c lies between 27.20 and

$x_b = 30.14$ corresponding to signific - ance level (α) (in 19 d.f. row) 0.10 and 0.05 respectively column.

For $x_a = 27.20$ and $x_b = 30.14$

$$P(\chi^2_{19} \geq x_a) = 0.10 \text{ and}$$

$$P(\chi^2_{19} \geq x_b) = 0.05$$

which implies

$P(\chi^2_{19} \geq x_c)$ lies between 0.10 & 0.05

i.e $P(\chi^2_{19} \geq 27.64)$ lies between 0.10 & 0.05.

Hence, ~~0.P0~~

$$0.05 < P(S^2 \leq 16) < 0.10.$$

Ex. 9:- The p.d.f of a chi-square dist^n is given as follows:

$$f(\chi^2) = \frac{1}{2} e^{-\chi^2/2} \quad ; \quad 0 < \chi^2 < \infty.$$

obtain the degree of freedom of the chi-square. Also, find its mean and variance.

Sol^{n} :- p.d.f of chi-square is

$$f(\chi^2) = \frac{1}{2^{n/2}\sqrt{\frac{n}{2}}} e^{-\chi^2/2} (\chi^2)^{n/2-1} \quad 0 < \chi^2 < \infty. \quad \text{— (1)}$$

comparing it with the given p.d.f; the given one can be written as,

$$f(\chi^2) = \frac{1}{2^{2/2}\sqrt{\frac{2}{2}}} e^{-\chi^2/2} (\chi^2)^{\frac{2}{2}-1} \quad \text{— (2)}$$

By comparing (1) and (2), we get

$n = 2$, i.e d.f. $= 2$.

mean = d.f $= 2$

variance = $2 \times$ d.f.

$= 2 \times 2 = 4$.

Student's t-distribution

When the population is normal and the population standard deviation is not known, then we may use the sample standard deviation 's^2' in place of 'σ'. However, due to the error between s and σ, when the sample deviation 's' is calculated from a very small sample, then the distribution of the statistic does not follow the standard normal distribution, i.e.

the z-variate $z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$, now became

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$ (because σ is unknown and s is unbiased estimator of 'σ')

and applied the normal tests for small samples, and it has been found that

'the distribution of 't' is far from the normal dist" for small samples, then

(Although the dist" of 't' for large ~~sam~~ samples, is asymptotically normal.)

the t-variate is named as new statistic, and say that t-variate follows t-distribution.

Let $X_i, (i = 1, 2, \ldots n)$ be a random sample of size n from a normal population with mean μ and variance σ^2. Then student's t is defined by the statistic:

$$t = \frac{\bar{X} - \mu}{S/\sqrt{n}}, \quad \text{where } \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \text{ is the}$$

sample mean and the $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$, is an unbiased estimator of population variance σ^2, and it follows t-dist^n with $(n-1)$ degree of freedom. The probability density function of t-dist^n with 'n' d.f. is defined as

$$f(t) = \frac{1}{\sqrt{n}\, B\left(\frac{1}{2}, \frac{n}{2}\right)} \cdot \frac{1}{\left(1 + \frac{t^2}{n}\right)^{\frac{(n+1)}{2}}}, \quad -\infty < t < \infty.$$

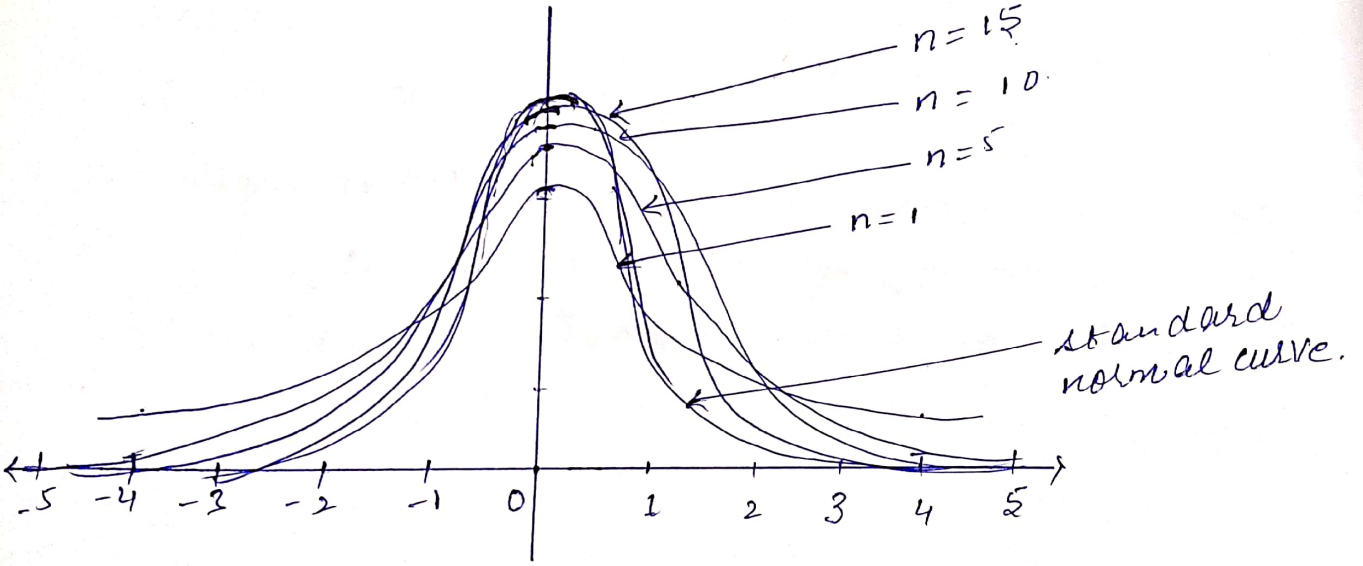
and abbreviated as '$t \sim t_n$', $B(a,b)$ is a beta function can be given as

$$B(a,b) = \int_0^1 x^{a-1}(1-x)^{b-1}\, dx$$

$$= \frac{\Gamma a \, \Gamma b}{\Gamma a+b}, \quad \because \Gamma = \text{gamma function.}$$

Properties :-

- The t-dist^n has only one parameter i.e. d.f.
- Shape of the dist^n is totally dependent on d.f.
- For large 'n', t-dist^n tends to standard normal dist^n with mean 0 and variance 1.
- The mean of t-dist^n is '0'.
- The variance of t-dist^n is dependent on d.f., say 'n', $= \frac{n}{n-2}$, $n > 2$.



- If z-variate follows standard normal dist^n and χ^2 follows chi-square with 'n' d.f., then

$\frac{z}{\sqrt{\chi^2/n}}$ will follow t-dist^n with n d.f.

i.e. $t = \frac{z}{\sqrt{\chi^2/n}} \sim t_n$ if $z \sim N(0,1)$ and $\chi^2 \sim \chi^2_n$.

Examples

Ex. 10 :- A mobile company claim that on average the people of India change their mobile phones after 2 years. To test the claim of the company, a student of the a programme collects such information about 16 mobile users randomly and observe that the people change their mobile phone after 2.02 years with 0.5 years standard deviation. What would be the t-statistic represented by this test and the d.f. of this test?

Sol^n :- Population mean (μ) = 2 years (μ).

Sample mean = 2.02 years ($\bar{x}$)

Standard deviation = 0.5 years (s)

Sample size = 16 (n)

t-statistic $\Rightarrow$

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{2.02 - 2}{0.5/\sqrt{16}} = 1.6$$

$$d.f = 16 - 1 = \underline{\underline{15}}.$$

F - distribution

Definition:-

If $\chi^2_{(n_1-1)}$ and $\chi^2_{(n_2-1)}$ are two independent chi-square variates with (n_1-1) and (n_2-1) degree of freedom respectively, then F-dist^n is defined by

$$F = \frac{\chi^2_{(n_1-1)}/(n_1-1)}{\chi^2_{(n_2-1)}/(n_2-1)} \quad \text{---} \quad (1)$$

In other words, F is defined as the ratio of two independent chi-square variates divided by their respective d.f. and denoted as

$$F \sim F(n_1-1, n_2-1).$$

Further, as we known $\chi^2_{n_1-1} = \frac{(n_1-1)S_1^2}{\sigma^2}$

therefore, eq (1) became

$$F = \frac{\frac{(n_1-1)S_1^2}{(n_1-1)\sigma_1^2}}{\frac{(n_2-1)S_2^2}{(n_2-1)\sigma_2^2}} = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F(n_1-1, n_2-1)$$

Here, F follows F-dist^n with (n_1-1, n_2-1) d.f.

NOTE :- • The shape of the dist^n depends on the degree of freedom (two parameters n_1 & n_2).

• The p.d.f is given by

$$f(F) = \frac{\left(\frac{n_1}{n_2}\right)^{\frac{n_1}{2}}}{B\left(\frac{n_1}{2}, \frac{n_2}{2}\right)} \cdot \frac{F^{\frac{n_1}{2}-1}}{\left(1+\frac{n_1}{n_2}F\right)^{\frac{n_1+n_2}{2}}}, \quad 0 < F < \infty.$$

where $B(a,b)$ is beta function, and F has been considered as variable.

• mean $= \frac{n_2}{n_2-2}$ for $n_2 > 2$, when d.o.f is (n_1, n_2).

• variance $= \frac{2n_2^2(n_1+n_2-2)}{n_1(n_2-2)^2(n_2-4)}$ for $n_2 > 4$, when d.o.f is (n_1, n_2).

Descriptive statistics

Definition :- Descriptive statistics involves describing, summarizing and organizing the data so it can be easily understood.

Frequency :- The frequency of any value occurring in a data set, is the number of times the value occurs in the set.

Once we ~~can~~ know the frequency with which the values occur in a data set, we can construct a table showing the frequency of each value against it, this table is frequency distribution

Frequency distribution :- A table structing the data into classes of suitable intervals. showing number of observations falling into a certain class interval.

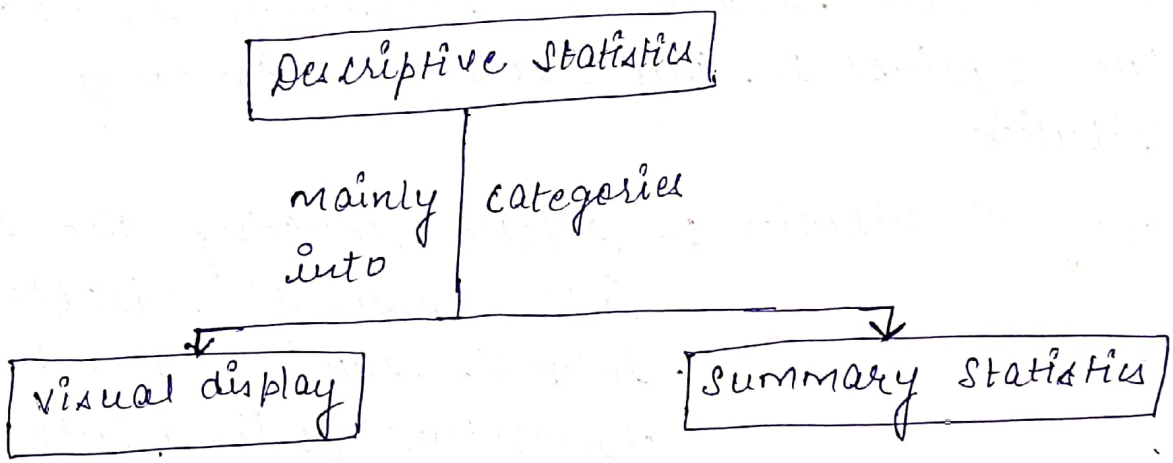
Frequency curve :- A frequency polygon modification by smoothing ~~curve~~ classes and data points for a data set.

Frequency polygon :- A line graph connecting the midpoints of each class in a data set at the frequency height of a class.

Relative frequency :- Relative frequency of a value occurring in a data set is the frequency of a value as a fraction or a percentage of the total number of observations.

Relative frequency distribution :-

The presentation of data showing fraction or percentage of the total data under a particular class.



- Bar diagram
- Histogram
- Pie chart
- Frequency polygon

- Measures of Central Tendency / Location
- Measures of dispersion/ variability

Visual Display :- (Graphical Representation)

Q: The following table gives the raw data relating to the marks, out of 10, of 100 students in a statistics examination.

2	5	0	5	7	6	6	7	4	8
4	6	7	3	6	6	5	6	2	6
6	4	5	7	4	4	7	4	6	4
3	4	8	1	5	8	7	5	7	7
7	6	5	7	4	5	5	3	6	6
5	8	6	6	7	7	3	4	3	5
9	4	8	5	3	5	9	5	5	7
1	9	3	5	5	7	6	8	8	2
5	4	4	4	6	3	5	6	4	4
8	2	8	5	5	6	7	3	6	9

→ Marks in the test is variable.

→ Marks are in whole numbers between 0 and 10.

→ Variable has '11' values, which are isolated i-e is discrete variable. (possible values = 11).

→ This is ungrouped data.

→ It is like a mesh, difficult to extract the characteristics of given data.

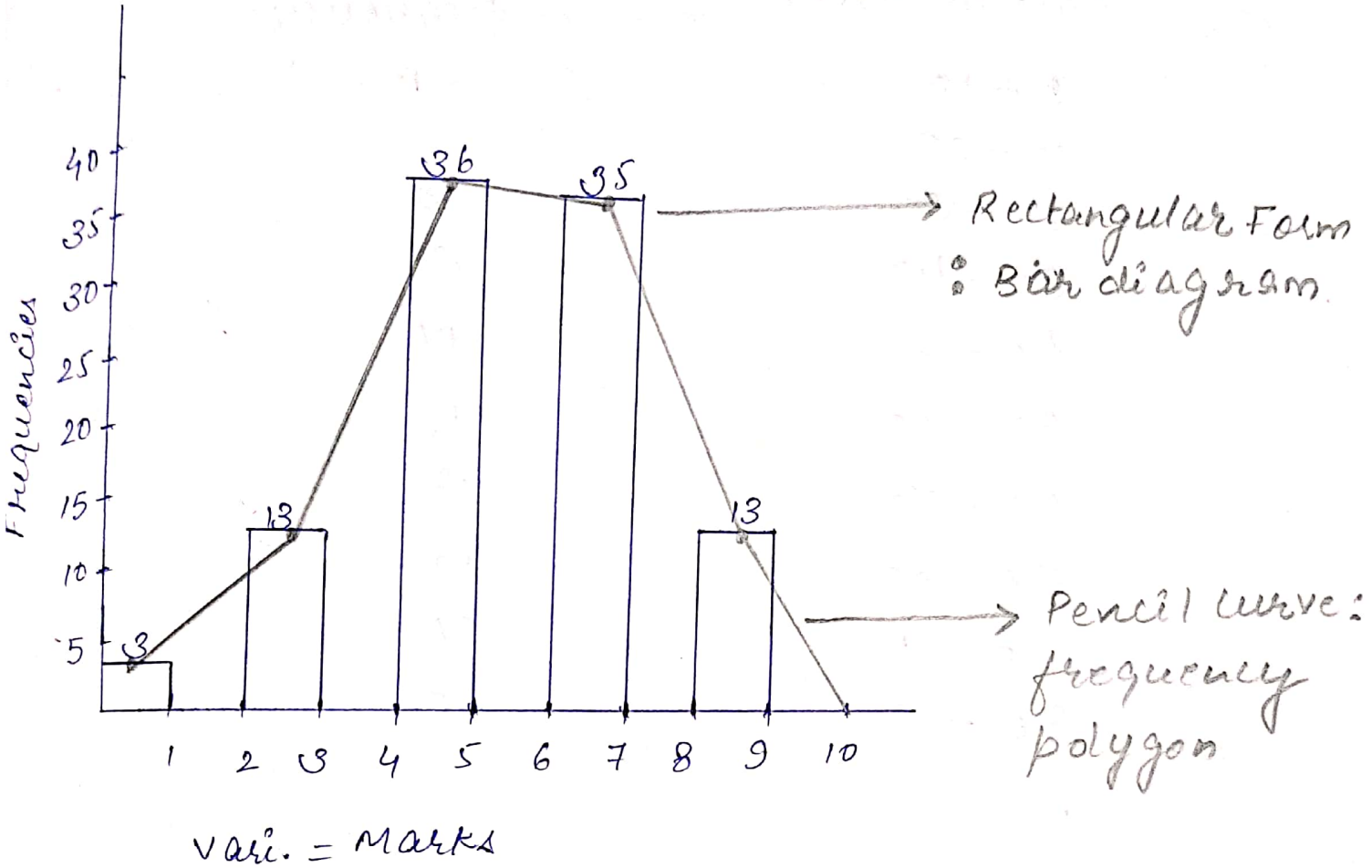
→ ~~Bar diagram~~ :-

→ Organised form :-

Marks	Frequency	Relative frequency
0	1	0.01
1	2	0.02
2	4	0.04
3	9	0.09
4	15	0.15
5	21	0.21
6	20	0.20
7	15	0.15
8	9	0.09
9	4	0.04
Total	100	1.00

Table: Frequency distribution of marks of 100 students.

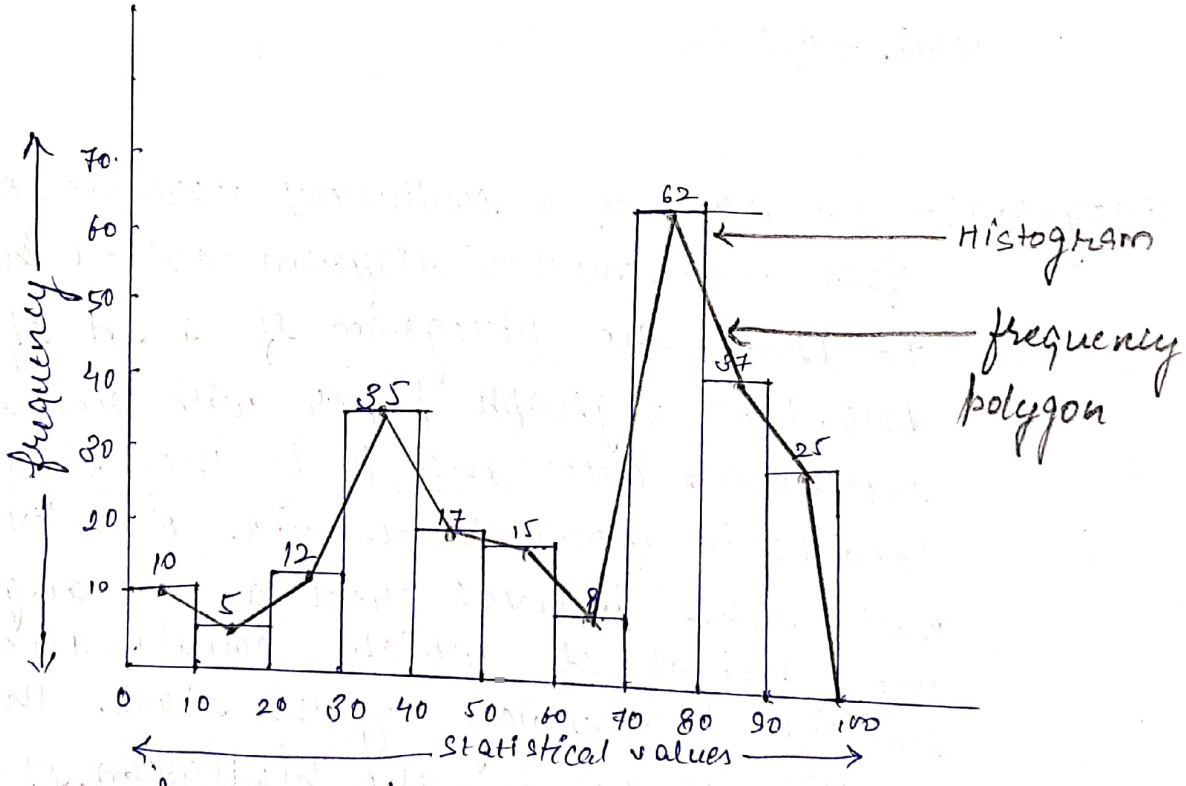
→ Bar diagram :-



Histogram :- In case of a continuous variable, one often uses another diagram called histogram To draw the histogram of a set of given data, take a graph paper with rectangular coordinates axes and plot the class boundaries along the x-axis. Now over each class interval erect a rectangle the height of which equals the relative frequency of this class. The resulting figure is the histogram of given data.

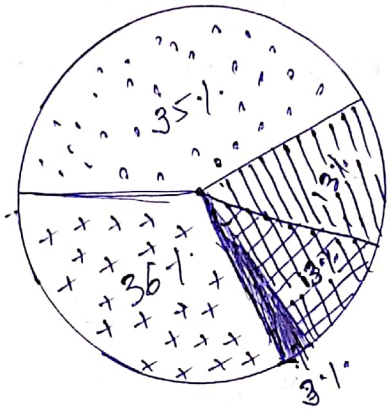
Ex :- Draw the histogram for the following frequency distribution.

Statistic value	frequency
0 – 10	10
10 – 20	5
20 – 30	12
30 – 40	35
40 – 50	17
50 – 60	15
60 – 70	8
70 – 80	62
80 – 90	37
90 – 100	25



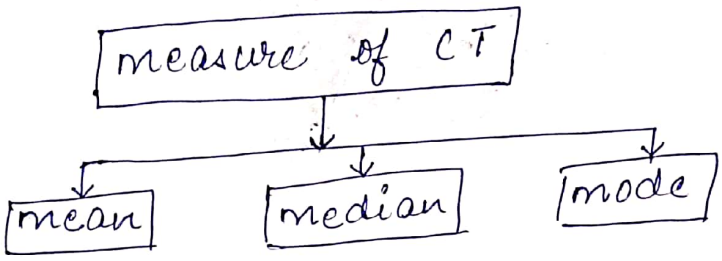
∴ Histogram of given data.

→ **Pie Chart** :- The different segments of a circle show percentage contribution of various constituents to its total picture. This sub-divided circle is called an angular or pie-diagram.



Measure of Central Tendency

In any distribution, majority of the observations pile up or cluster around in a particular region or concentration of the values in the central part of the distribution, is the measure of central tendency / measure of location (position).



→ **Mean:-** Mean / Arithmetic mean of a set of observations is their sum divided by the number of observations, is given as:

If $\{x_1, x_2, \text{---} x_n\}$ is a set of observations, then the mean $\bar{x}$ is

$$\bar{x} = \frac{1}{n}(x_1 + x_2 + \text{---} + x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Ex:- Find the arithmetic mean of the following frequency distribution:

Marks:	0-10	10-20	20-30	30-40	40-50	50-60
No. of Students:	12	18	27	20	17	6

Sol:-

Marks	(f) No. of St.	Mid pt. (x)	fx
0-10	12	5	60
10-20	18	15	270
20-30	27	25	675
30-40	20	35	700
40-50	17	45	765
50-60	6	55	330
Total	100		2,800

$$\bar{x} = \frac{1}{n}\sum fx$$

$$= \frac{1}{100} \times 2,800 = \underline{\underline{28}}.$$

→ Median :- Median of a distribution is the value of variable which divides it into two equal parts.

← Discrete →

Ex:- obtain the median for the following frequency distribution:

x :	1	2	3	4	5	6	7	8	9	
f :	8	10	11	16	20	25	15	9	6	$= \sum_i f_i = N = 120$

Sol:-

cf :	8	18	29	45	65	90	105	114	120.

In case of discrete frequency distributions, cumulative frequencies will be considered to find the median value.

Steps:-

① Find $\frac{\sum_i f_i}{2} = \frac{N}{2}$ (say)

② See the c.f. just greater than $\frac{N}{2}$.

③ The corresponding value of the variable 'x' is median.

In this example;

$N = 120$

$\frac{N}{2} = 60$

The c.f. just greater than $\frac{N}{2}$ is 65, the corresponding value of x to 65 is 5.

The median is '5', for discrete case.

Continuous Case

Median Formula:

$$\text{Median} = l + \frac{h}{f}\left(\frac{N}{2} - c\right)$$

l : is the lower limit of median class
f : is the frequency of the median class
h : is the magnitude of the median class
'c' : is the c.f. of the class preceding the median class.

and $N = \sum_i f_i$,

median class : the class corresponding

to the cummulative frequency just greater than $\frac{N}{2}$.

Ex:- Find the median wages of the following distribution:

Wages (in Rs.) :	2000-3000	3000-4000	4000-5000	5000-6000	6000-7000
No. of Worker :	3	5	20	10	5

Sol^{n}:-

Wages	No. of Worker (f)	c.f
2000-3000	3	3
3000-4000	5	8
4000-5000	20	28
5000-6000	10	38
6000-7000	5	43
	N = 43	

Median class : (just greater than $\frac{N}{2}$)

$\frac{N}{2} = \frac{43}{2} = 21.50$

c.f. just greater than 21.50 = 28

corresponding ~~class of~~ to 28 = 4000-5000

l = lower value of median class
= 4000

h = magnitude = 1000

f = frequency = 20

c = c.f. of the class preceding to median class
= 8

$$\text{Median} = 4{,}000 + \frac{1{,}000}{26}(21.50 - 8)$$

$$= 4{,}000 + 50 \times 13.50$$

$$= 4{,}000 + 675$$

$$= 4{,}675 \text{ wages.}$$

→ **Mode :-** Mode is the ~~most~~ value which occurs most frequently in a set of observations and around which the other items of the set cluster densely.

Ex :- Frequency distribution! mode?

x :	1	2	3	4	5	6	7	8
f :	4	9	16	25	22	15	7	3

Sol^n The value of x to the maximum frequency

$\Rightarrow x = 4$ is the mode with maximum frequency '25'.

—: For continuous case :→

Formula :- Modal class : class interval to maximum frequency.

$$\text{mode} = l + \frac{h(f_1 - f_0)}{(f_1 - f_0) - (f_2 - f_1)}$$

l : lower limit of modal class (M.C.)

h : magnitude of model class

f_1 : frequency of modal class.

f_0 : frequencies of preceding class to modal class.

f_2 : frequency of succeeding class to M.C.

Ex:- Find the mode for the following dist:

Class interval	0-10	10-20	20-30	30-40	40-50	50-60	60-70	70-80
Frequency	5	8	7	12	28	20	10	10

Sol:- Maximum frequency = 28

Modal class = 40-50

$l = 40$, $h = 10$, $f_1 = 28$, $f_0 = 12$, $f_2 = 20$.

$$\text{mode} = 40 + \frac{10(28-12)}{(28-12)-(20-28)}$$

$$= 40 + 6.666$$

$$= 46.67 \text{ Approx.}$$

Statistical Inference :- It is a technique of drawing conclusions about the population from sample data.

→ It is based on the assumption that the sample should be drawn randomly.

→ In other words; statistical inference uses the statistic to make predictions.

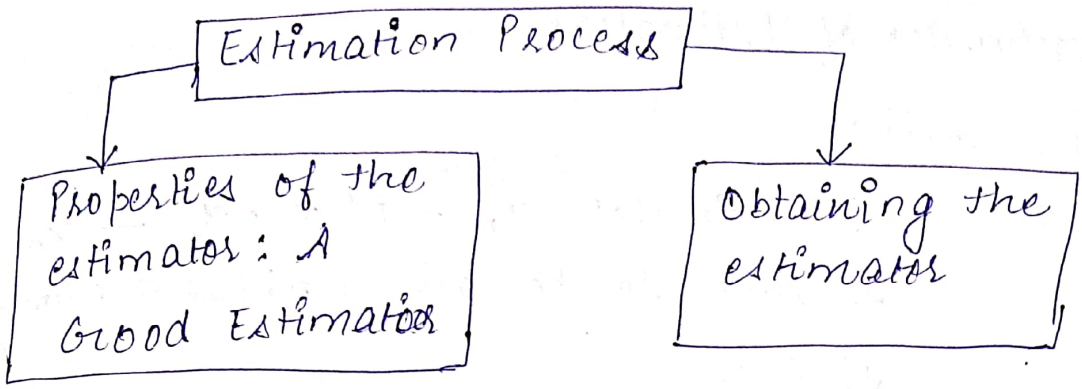
Estimation :- The technique of finding an estimator to produce an estimate (approx. value) of unknown parameter of the population on the basis of a sample is called estimation.

Estimator :- Any statistic used to estimate an unknown population parameter is known as estimator.

Estimate :- A specific value of the estimator.

Point Estimate :- A single value of the estimate of the population parameter.

Interval Estimate :- A range of ~~values~~ ~~estimate~~ values to estimate a population parameter.



Criterion of a good estimator

1. Unbiased
2. Consistent
3. Efficient
4. Sufficient

Method of Estimation

→ The method of moment

→ the method of maximum likelihood estimation.

Characteristics of Estimators :-

→ **Unbiasedness :-**

Definition : An estimator $T_n = T(x_1, x_2, \dots x_n)$ is said to be an unbiased estimator of θ, if

$$E(T_n) = \theta.$$

Ex :- The sample mean $\bar{x}$ is an unbiased estimator of population mean μ; as

$$E(\bar{x}) = \mu \quad \text{(already done)}$$

Ex :- The sample variance with the value

$$s^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2$$ is not an unbias-

-ed estimator, but

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$$ is an ~~unb~~ unbiased

estimator as; $E(s^2) = \sigma^2$.

Note :- If $E(T_n) \neq \theta$, then T_n is biased estim-ator with biased value;

$$\text{biased value} = E(T_n) - \theta.$$

Ex :- As sample variance s^2 is biased estimator of σ^2 with biased value

$$E(T_n) - \theta = BV.$$

$$E(s^2) - \sigma^2 = BV$$

$$\left(1 - \frac{1}{n}\right)\sigma^2 - \sigma^2 = BV$$

$$BV = -\sigma^2/n.$$

→ **Consistent Estimator:-** An estimator $T_n = T(x_1, x_2 \ldots x_n)$, is said to be consistent estimator of population parameter θ if it converges in probability to θ, i.e.

$$\lim_{n \to \infty} p(T_n \to \theta) = 1.$$

Sufficient condition for consistency:- Let T_n be an estimator for the population parameter θ, then it is said to be consistent ~~parameter~~ estimator of θ if

(i) it is unbiased estimator of θ as $n \to \infty$.

(ii) the variance of estimator T_n decreases with increasing sample size.

In other words, the variance of the estimator approaches zero as $n \to \infty$, i.e.

$$Var(T_n) \to 0 \quad \text{as} \quad n \to \infty.$$

Ex:- The sample mean is always a consistent estimator of the population mean (μ), provided that the population has finite variance.

Sol^n:- $E(\bar{x}) = \mu$ and $Var(\bar{x}) = \frac{\sigma^2}{n}$ are given,

Now check the sufficient conditions

(i) Expected value of $\bar{x}$ is μ.

(ii) $Var(\bar{x}) = \frac{\sigma^2}{n}$

which implies $Var(\bar{x}) \to 0$ as $n \to \infty$.

i.e. $\lim_{n \to \infty} Var(\bar{x}) = \lim_{n \to \infty} \frac{\sigma^2}{n} = 0.$

both the conditions are satisfied.
Hence sample mean is a consistent estimator of the population mean.

Ex:- s^2 is the a consistent estimator of population variance, where $s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$

Sol^n :- To check:-

(i) $E(s^2) \to \sigma^2$ as $n \to \infty$

(ii) $\text{var}(s^2) \to 0$ as $n \to \infty$.

We know that

$$E(s^2) = \left(1 - \frac{1}{n}\right)\sigma^2$$

$$\lim_{n\to\infty} E(s^2) = \lim_{n\to\infty}\left(1 - \frac{1}{n}\right)\sigma^2 = \sigma^2.$$

i.e $E(s^2) \to \sigma^2$ as $n \to \infty$

$$\text{Var}(s^2) = \text{var}\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right)$$

$$= \frac{1}{n^2}\,\text{var}\left(\sum_{i=1}^{n}(x_i - \bar{x})^2\right)$$

$$= \frac{1}{n^2}\,\text{var}\left(\frac{n-1}{\sigma^2}\,\sigma^2 \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}\right)$$

$$= \frac{1}{n^2}\,\text{var}\left(\sigma^2\,\frac{n-1}{\sigma^2}\,s^2\right)$$

$$= \frac{1}{n^2}\,\text{var}\left(\sigma^2\,\chi^2_{n-1}\right)$$

$$= \frac{\sigma^4}{n^2}\,\text{var}\left(\chi^2_{n-1}\right)$$

$$= \frac{\sigma^4}{n^2}\,2(n-1)$$

$$\lim_{n\to\infty}\text{var}(s^2) = \lim_{n\to\infty}\frac{2\sigma^4}{n^2}(n-1) = 2\sigma^4 \lim_{n\to\infty}\left(\frac{1}{n} - \frac{1}{n^2}\right)$$

$= 2\sigma^4 \times 0$

$= 0.$

i.e $Var(s^2) \longrightarrow 0$ as $n \longrightarrow \infty$.

Hence (i) and (ii) both are proved.

Note :- (i) A consistent estimator may or may not be unbiased.

(ii) Consistent estimator is related to large sample size i.e. $n \longrightarrow \infty$.

Method of Estimation

→ Method of Maximum likelihood Estimation:-

In this method, we will use maximum likelihood function and try to maximize it.

Likelihood function:-

Let $\{x_1, x_2, \dots x_n\}$ be a random sample of size 'n' from a population with density function $f(x, \theta)$. Then the likelihood function of the sample values $(x_1, x_2, \dots x_n)$ is their joint density / Mass function, given by

$$L = L(\theta) = L(\theta / x_1, \dots x_n)$$

$$= f(x_1, x_2 \dots x_n / \theta) = f(x_1, \dots x_n ; \theta)$$

i.e

$$L(\theta) = f(x_1; \theta) \cdot f(x_2; \theta) \cdot \dots \cdot f(x_n; \theta)$$

$$= \prod_{i=1}^{n} f(x_i / \theta)$$

for continuous dist^n.

Note:- (i) For discrete case:-

$$L(\theta) = \prod_{i=1}^{n} P(x_i = x_i / \theta)$$

(ii) The joint density function and likelihood function both are different in interpretation.

(iii) As j.d.f. is interpreted as a function of values of random variable $(x_1, x_2, \dots x_n)$, 'given' values of the parameters θ.

whereas likelihood function is interpreted as a function of values of the parameters θ, given values of the random variable $(x_1, x_2 \text{—} x_n)$.

Maximum likelihood Estimation :-

"The maximum likelihood estimation is a method is a method that determines parameters values in such a way that they maximize the likelihood function."

Let $x_1, x_2 \text{—} x_n$ be a random sample from a distribution with a parameter θ, then a maximum likelihood estimate of θ, say $\hat{\theta}$ is a value of θ that maximizes the likelihood function

$$L(\theta / x_i).$$

→ **Likelihood Equation :-** By the principle of maxima and minima, maximum likelihood Estimators is the solution of the following conditions:

$$\frac{\partial L}{\partial \theta} = 0, \quad \frac{\partial^2 L}{\partial \theta^2} < 0.$$

Ex:- Let $(x_1, x_2, \ldots x_n)$ is a random sample of size n for a normal population $N(\mu, \sigma^2)$ with mean μ and variance σ^2. then find the MLE for μ & σ^2.

Sol^n:- P.d.f

$$f(x, \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} \quad ; -\infty < x < \infty, \; -\infty < \mu < \infty, \; \sigma > 0,$$

$$L = L(\theta) = f(x_1/\theta) \cdot f(x_2/\theta) \cdot \ldots \cdot f(x_n/\theta).$$

$$= \prod_{i=1}^{n} f(x_i/\theta)$$

$$L(\mu) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x_i-\mu)^2}$$

$$= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \cdot e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2}$$

(∵ L and log L attains their extrem at same pt.)

taking log on both sides

$$\log L(\mu) = \log e^{-\frac{1}{2\sigma^2}\cdot\sum_{i=1}^{n}(x_i-\mu)^2} - \log(\sqrt{2\pi\sigma^2})^n$$

$$= -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2 - \frac{n}{2}\log(2\pi\sigma^2)$$

first partial derivative w.r.t 'θ' is

$$\frac{\partial(\log L)}{\partial\mu} = -\frac{\partial}{\partial\mu}(\text{const}^{tt}) - \left(\frac{1}{2\sigma^2}\right)\sum_{i=1}^{n}\frac{\partial}{\partial\mu}(x_i-\mu)^2$$

$$= -\frac{1}{2\sigma^2}\sum_{i=1}^{n} 2(x_i-\mu)(-1)$$

$$= \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu)$$

Equating this last result to zero and solve for μ yields

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu)=0.$$

$$\sum_{i=1}^{n}(x_i-\mu)=0$$

$$\sum_{i=1}^{n}x_i-n\mu=0$$

$$\hat{\mu}=\frac{1}{n}\sum_{i=1}^{n}x_i$$

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

$$\rightarrow \frac{\partial^2(\log L)}{\partial \mu^2}=\frac{\partial}{\partial \mu}\left(\frac{1}{\sigma^2}\sum(x_i-\mu)\right)$$

$$=\frac{1}{\sigma^2}\sum_{i=1}^{n}(-1)$$

$$=\frac{1}{\sigma^2}-n$$

$$=\frac{-n}{\sigma^2}.$$

$$<0.$$

Hence $L(\theta)$ is maximum at $\theta=\bar{x}$ where $\theta=\mu$.

$\Rightarrow$ Maximum Likelihood estimator for μ is the sample mean $\bar{x}$.

(ii) $$\log L(\sigma^2)=-\frac{n}{2}\log(2\pi\sigma^2)-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2$$

first partial derivative w.r.t. 'σ^2'.

$$\frac{\partial}{\partial \sigma^2}\log L=-\frac{n}{2}\frac{2\pi}{2\pi\sigma^2}+\frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i-\mu)^2$$

$$=\frac{-n}{2\sigma^2}+\frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i-\mu)^2$$

Equate to zero, i.e.

$$\frac{\partial}{\partial \sigma^2} \log L = 0.$$

$$\Rightarrow \hat{\sigma}^2 = \frac{\sum_{i=1}^{n} (x_i - \hat{\mu})^2}{n} = s^2, \quad \text{as } \hat{\mu} = \bar{x},$$

second partial derivate is

$$\frac{\partial^2}{\partial(\sigma^2)^2} \log L = -\frac{n}{2}\left(-\frac{1}{\sigma^4}\right) + \frac{1}{2}\frac{(-2)}{(\sigma^2)^3} \sum_{i=1}^{n} (x_i - \hat{\mu})^2$$

$$= \frac{n}{2\sigma^4} - \frac{n}{\sigma^6} \sum_{i=1}^{n} \frac{(x_i - \hat{\mu})^2}{n}$$

$$= \frac{n}{2\sigma^4} - \frac{n}{\sigma^6} s^2$$

$$= \frac{n}{2\sigma^6} [\sigma^2 - 2s^2]$$

at $\sigma^2 = \hat{\sigma}^2$ point;

$$\frac{\partial^2}{\partial(\sigma^2)^2} \log L = \frac{n}{2\sigma^6} (\hat{\sigma}^2 - 2s^2)$$

$$= \frac{n}{2\sigma^6} (-s^2) < 0,$$

This shows that $\sigma^2 = \hat{\sigma}^2$ is the maxima pt. of σ^2

Hence $L(\sigma^2)$ is maximum at $\hat{\sigma}^2 = s^2$,

$\Rightarrow \hat{\sigma}^2 = s^2$ is the maximum likelihood estimator of population variance.

→ **Method of moment:-**

Let $X_1, X_2, \ldots X_n$ be a random sample from the p.dist^n a population whose probability density (Mass) function is $f(x; \theta_1, \theta_2, \ldots \theta_k)$ with k unknown parameters $(\theta_1 \ldots \theta_k)$. Then the r^{th} sample moment about origin is defined as

$$M'_r = \frac{1}{n} \sum_{i=1}^{n} X_i^r \quad \text{and about mean is}$$

$$M_r = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})^r,$$

while the r^{th} population moment about origin is $\mu'_r = E(X)^r$ and about mean

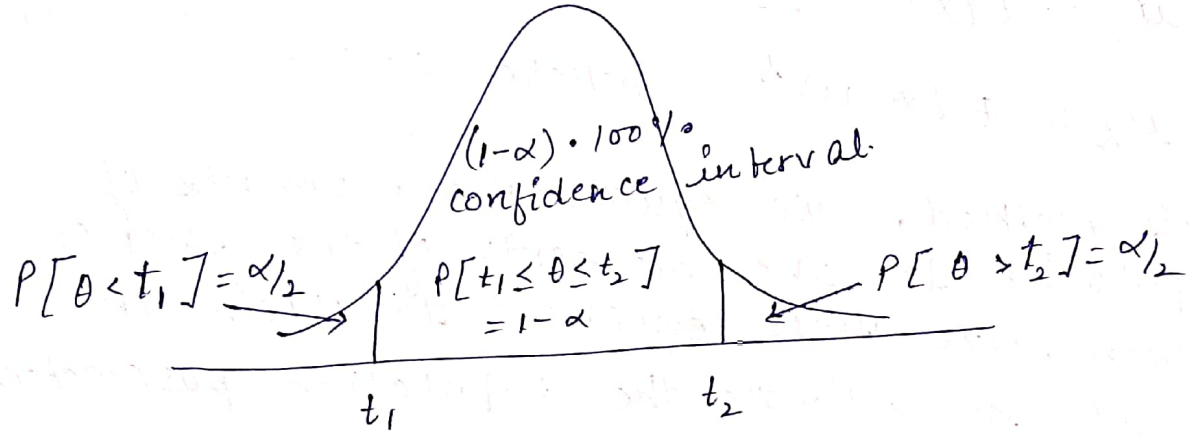
$$\mu_r = E(X-\mu)^r, \quad \text{where } r = 1, 2, \ldots$$

In this method, we equate the moments of the population to the moments of the samples and then solve these equation to obtain the value of the estimate the population parameters.

Confidence interval for parameters

Let $x_1, x_2 \text{---} x_n$ be a random sample of size 'n' from a population whose p.d/m.f is $f(x, \theta)$. Let t_1 and t_2 $(t_1 \leq t_2)$ be two statistics, such that the probability that the random interval $[t_1, t_2]$ includes the true value of population parameter θ is $(1-\alpha)$, i.e.

$$P[t_1 \leq \theta \leq t_2] = 1 - \alpha,$$ where α is independent of θ.



→ $[t_1, t_2]$ is the confidence interval.

→ the factor $(1-\alpha)$ is the confidence coefficient or confidence level

→ If $\alpha = 0.05$ then confidence limit is 95%

as $(1-\alpha) \times 100\% = 0.95 \times 100\%$

$= 95\%$

→ If $\alpha = 0.01$ then confidence limit is 99%.

→ Higher the confidence limit, the more strongly we believe that the value of the parameter being estimated lies within the interval.

→ Confidence interval for mean (one-sample)

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

for ~~large~~ known variance.

For unknown population variance; we can use sample variance,

$$P\left[\bar{X} - (t_{n-1})_{\alpha/2}\frac{S}{\sqrt{n}} \leq \mu \leq \bar{X} + (t_{n-1})_{\alpha/2}\frac{S}{\sqrt{n}}\right]$$

$$= 1-\alpha.$$

→ For two samples which are paired (confid. interval for the difference of two means (unknown variance))

$$P\left[\bar{D} - (t_{n-1})_{\alpha/2}\frac{S_D}{\sqrt{n}} \leq \mu_D \leq \bar{D} + (t_{n-1})_{\alpha/2}\frac{S_D}{\sqrt{n}}\right]$$

$$= 1-\alpha$$

where, $\bar{D} = \frac{1}{n}\sum_{i=1}^{n} D_i$, $S_D^2 = \frac{1}{n-1}\sum_{i=1}^{n}(D_i - \bar{D})^2$.

$D_i = X_i - Y_i \quad \forall\ i = 1, 2, \ldots n$ with pairwise

$(X_1, Y_1), \ldots (X_n, Y_n)$ random variable.

Confidence interval for population variance

→ Known 'μ' = population mean

$$P\left[\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{(\chi^2_n)_{\alpha/2}} \le \sigma^2 \le \frac{\sum_{i=1}^{n}(x_i-\mu)^2}{(\chi^2_n)_{1-\alpha/2}}\right]$$

$$= 1-\alpha.$$

→ Unknown mean

$$P\left[\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{(\chi^2_{n-1})_{\alpha/2}} \le \sigma^2 \le \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{(\chi^2_{n-1})_{1-\alpha/2}}\right]$$

$$= 1-\alpha$$

Confidence interval for population proportion

→

$$P\left[\hat{p} - Z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leq P \leq \hat{p} + Z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right]$$

where sampling dist^{n} of sample proportion with mean P & variance $\frac{P(1-P)}{n}$.

and $\hat{p}$ is the variable, here P is unknown and sample is large.