UNIT
1 Simple Random Sampling

1. Define simple random sampling with replacement and without replacement from a finite population. Also Explain various procedures of drawing samples.

The simplest and common used method of sampling is simple random sampling.

Simple random sampling is a technique of drawing a sample in such a way that each and every unit of the population has an equal and independent chance of being included in the sample.

To draw a sample of size n from the population of size N in simple random sampling we have two methods.

1. Simple Random Sampling with replacement (SRSWR)
2. Simple Random Sampling without replacement (SRSWOR)

1. Simple Random sampling with replacement (SRSWR)

The technique of drawing each and every unit of Population has an equal and independent chance of being included into the sample in such a way that the selected unit should be replaced again into the population before going to select the next unit is called simple random sampling with replacement. Simply written as SRSWR.

Suppose we have to select a sample of size n from the population of size N in SRSWR we proceed as follows.

1. Probability of drawing a unit from the N population units at the first draw is $\frac{1}{N}$ and is same for all draws.

2. Number of possible samples in SRSWR are N^n

Probability of selecting a sample of size n from a population of size N in SRSWR is $\frac{1}{N^n}$

All the draws in SRSWR are independent and identical.

Simple Random sampling without Replacement (SRSWOR)

The technique of drawing each and every unit of the population has an equal and independent chance of being included into the sample in such a way that the selected unit should not be replaced again into the population before the next draw or any other subsequent draws is called simple random sampling without replacement. Simply written as SRSWOR.

Probability of drawing a unit from N units at the first draw is $\frac{1}{N}$

at the second draw is $\frac{1}{N-1}$ (since N-1 units remain)

at the third draw is $\frac{1}{N-2}$ and so on.

2. Number of possible samples in SRSWOR are N_{c_n}

3. Probability of selecting a sample of size n from a population of size N in SRSWOR is $\frac{1}{N_{c_n}}$

4. All draws in SRSWOR are independent but not identical.

. Explain selection of a simple Random sample.

Or

Procedures of selecting a random sample.

There are several methods to draw a sample of size n out of N units through simple Random sampling.

Selection of a simple random sample depends on the size and nature of the population.

A simple random sample can be obtained by the following two methods.

1. Lottery method.
2. Mechanical Randomization or Random numbers method.

1. Lottery method:- This is the simplest method of selecting a random sample.

These slips should be as homogeneous as possible in shape, size, colour etc to avoid human bias. Put these n slips in a bag and throughly.
shuffled and then n slips are drawn one by one
The n units (items) corresponding to numbers on the slips drawn will constitue a random sample. The only advantage in this method is its simplicity
If the population is large, the lottery method is time consuming and practically difficult.

Random Numbers method or mechanical Randomisation.

If the population is large, the lottery method is time consuming, expensive and very difficult to select a random sample.

The most inexpensive method of drawing a random sample is random numbers method.

In this method the random number tables have been identified from various records / blocks so that each of the digits 0, 1, 2, ... 9 appear with the same frequency and independent of each other.

If a random sample is to be selected from the population of size (N) ≤ 99 then the numbers can be combined by two digits from 00 to 99.

Similarly if N ≤ 999 then combine three digits from 000 to 999 and if N ≤ 9999 then combine four digits from 0000 to 9999 and so on.

obtaining a random sample from the random numb-ers method can be explained in the following steps:

Identify the total number of units in the population N with the numbers from 1 to N.

2. select any page of random number tables and pick up the numbers in any row or column or diagonal at random.

3. The selected numbers then constitute the requir-d random sample.

Some of the random number tables in common use for 1. Tippets random number tables
2. fisher & yates tables

There are some modified procedures of selecting random numbers.

The method of selecting a sample with the help of random number table is always advisable.

③ Explain SRSWOR Vs SRSWR

SRSWOR	SRSWR
1. If the unit selected in any draw is not replaced in the population before making the next draw then the sampling plan is called SRSWOR.	If the unit selected in any draw is replaced back before the making the next draw then the sampling plan is called SRSWR.
2. All draws are independent but not identical.	All draws are independent and identical.
3. Number of possible samples in SRSWOR are N_{C_n}	Number of possible samples in SRSWR are N^n
4. The probability of selecting a sample of size n from a population of size N is $\frac{1}{N_{C_n}}$	the probability of selecting a sample of size n from a population of size N is $\frac{1}{N^n}$
5. $E(\bar{x}) = \mu$	$E(\bar{x}) = \mu$
6. Sample mean square is an unbiased estimator of the population mean square ie $E(s^2) = S^2$	Sample mean square is an unbaised estimator of population varience i.e $E(s^2) = \sigma^2$
7. varience of the ~~squa~~ sample mean in SRSWOR is $\frac{N-n}{N} \frac{S^2}{N}$	varience of the sample mean in SRSWR is $\frac{N-1}{N} \frac{S^2}{n}$
since varience is less efficiency is more	The efficiency of the sample mean is the lesser than in SRSWOR
Probability of selecting a unit at first draw is $\frac{1}{N}$ at second draw is $\frac{1}{N-1}$ and so on	9. probability of selecting a unit in every draw is $\frac{1}{N}$ only.

SRSWR is suffered from the draw back having same unit two or more times in a sample.
In SRSWOR we have distinct elements in a sample.
In many cases SRSWOR is preferred to SRSWR.

4. Define merits and demerits of simple Random sampling merits (advantages).

1. Since simple Random sampling is probability sampling it eleminates personal bias, as such a simple random sample is more representative of the population as compared with the Judgement or Purposive sampling.
2. we can estimate the efficiency of the estimates through their standard error in simple random sampling method.
3. The calculations for estimating the population parameters are easy.
4. simple random sampling method specifies the three principles of a good sampling design.
5. simple Random sampling method is more popular and widely used sampling method which reduces the cost, time and labour.

Demerits (Limitations) or Disadvantages.

1. To select a simple random sample, it requires an up-to-date frame ie list or guide of the population. This is not possible every time.
A simple random sampling suffers from this draw-back severly.
2. The size of the sample in simple random sample should be more otherwise sample may not be a good representative of the population. Note that if the sample size increases, non-sampling errors also increases.
3. If the population is too large, simple random sampling method is not easy to apply.
4. If the population is heterogeneous the estimates by simple random sampling method are not reliable. Hence

cases we prefer stratified Random sampling.

5. simple Random sampling may give non-Random looking samples in some cases.

6. In simple random sampling with replacement we may get the same unit more than once. So the number of distinct observations is reduced. In such cases there may be increases in sampling Errors.

Notations and Terminology in simple Random sampli-ng.

Population

values of units $Y_1 Y_2 \cdots Y_N$

size $= N$

Population mean $= \bar{Y}_N = \frac{1}{N}\sum Y_i$

Sample

values of units $y_1 y_2 \cdots y_n$

size $= n$

Sample mean $= \bar{y}_n = \frac{1}{n}\sum_{i=1}^{n} y_i$

$= \frac{1}{n}\sum_{i=1}^{n} a_i y_i$

$a_i = \begin{cases} 1 \text{ if } i\text{th unit is included in the sample} \\ 0 \text{ if } \text{" " " not " " "} \end{cases}$

Varience $\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(Y_i - \bar{Y})^2$

$= \frac{1}{N}\left(\sum_{i=1}^{N} Y_i^2 - N\bar{Y}^2\right)$

$s^2 = \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2$

$= \frac{1}{n}\left(\sum_{i=1}^{n} y_i^2 - n\bar{y}^2\right)$

Population mean square

$S^2 = \frac{1}{N-1}\sum_{i=1}^{N}(Y_i - \bar{Y})^2$

$= \frac{1}{N-1}\left[\sum_{i=1}^{N} Y_i^2 - N\bar{Y}^2\right]$

Sample mean square

$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$

$= \frac{1}{n-1}\left(\sum_{i=1}^{n} y_i^2 - n\bar{y}^2\right)$

Note :- 1. Sample mean $= \bar{y}_n = \bar{y} = \frac{1}{n}\sum_{i=1}^{N} a_i y_i$

where $a_i = \begin{cases} 1 \text{ if } i^{th} \text{ unit is included in the sample} \\ 0 \text{ if } i^{th} \text{ unit is not included in the sample} \end{cases}$

$E(a_i) = 1 \cdot P(a_i = 1) + 0\, P(a_i = 0)$

$E(a_i) = 1 \cdot P(i^{th}$ unit is included in the sample of size $n)$

$+ 0 \cdot P(i^{th}$ unit is not included in the sample of size $n)$

$= P(a_i = 1) \cdot P(a_j = 1 / a_i = 1)$

$= \frac{n}{N} \cdot \frac{n-1}{N-1}$

$E(a_i a_j) = \frac{n(n-1)}{N(N-1)}$

4. $\left(\sum_{i=1}^{n} y_i\right)^2 = \sum_{i=1}^{n} y_i^2 + \sum_{i \neq j=1}^{n} \sum y_i y_j$

Theorem 1 :- The probability that a specified unit of the population being selected at any given draw (say r^{th} draw) is equal to the probability of it being selected at the first draw in Simple Random Sampling.

Proof :- Let A_i denotes the Event of selecting a specific unit at the i^{th} draw $i = 1, 2, \ldots r$. We are required to prove. $P(A_r) = P(A_1)$

In SRSWR, clearly $P(A_r) = P(A_1) = \frac{1}{N}$
Where N is the size of the population.

In SRSWOR, an item is selected in the r^{th} draw means that it is not selected in the Peri Previous $(r-1)$ draws

$P(A_r) = P(A_1' \cap A_2' \cap A_3' \cap \cdots \cap A_r)$

$= P(A_1') P(A_2' \backslash A_1') P(A_3' / A_1' \cap A_2') \cdots\cdots$

$\cdots P(A_r / A_1' \cap A_2' \cap A_3' \cdots \cap A_{r-1}')$

By the multiplication theorem of probability since draws are independent.

$P(A_r) = P(A_1') P(A_2') P(A_3') \cdots P(A_{r-1}') P(A_r)$

$P(A_r) = \left(1 - \frac{1}{N}\right)\left(1 - \frac{1}{N-1}\right)\left(1 - \frac{1}{N-2}\right) \cdots \left(1 - \frac{1}{N-r+2}\right) \frac{1}{N-r+1}$

$P(A_r) = \left(\frac{N-1}{N}\right)\left(\frac{N-2}{N-1}\right)\left(\frac{N-3}{N-2}\right) \cdots \frac{N-r+1}{N-r+2} \cdot \frac{1}{N-r+1}$

$P(A_r) = \frac{1}{N} = P(A_1)$

$\therefore P(A_r) = P(A_1) = \frac{1}{N}$

Note :- The probability that a specified ...

$E(\bar{y}_n) = \bar{Y}_N$

Proof :- suppose the population consists of N un $Y_1, Y_2 - - - Y_N$ from which the sample of size $y_1, y_2, ---, y_n$ are drawn without replacement

population mean $\bar{Y}_N = \frac{1}{N}\sum_{i=1}^{N} Y_i$

sample mean $\bar{y}_n = \frac{1}{n}\sum_{i=1}^{n} y_i$

$$\bar{y}_n = \frac{1}{n}\sum_{i=1}^{N} a_i y_i$$

Taking expectation on both sides

$$E(\bar{y}_n) = \frac{1}{n}\sum_{i=1}^{N} E(a_i) Y_i \quad \text{——— (1)}$$

we calculate $E(a_i)$, a_i takes the values

$$a_i = \begin{cases} 1, & i^{th} \text{ unit is included into the sample} \\ 0, & i^{th} \text{ unit is not included into the sample} \end{cases}$$

$$E(a_i) = \sum_{i=0}^{1} a_i P(a_i)$$

$$= 1\, P(a_i = 1) + 0\, P(a_i = 0)$$

$E(a_i) = 1\, P(i^{th}$ unit is included into the sample of size$)$
$+ 0\, P(i^{th}$ unit is not included into the sample of size$)$

$$E(a_i) = \frac{n}{N} \quad \text{——— (2)}$$

substitute equations (2) in (1) we get

$$E(\bar{y}_n) = \frac{1}{\cancel{n}}\sum_{i=1}^{N} \frac{\cancel{n}}{N} Y_i$$

$$= \frac{1}{N}\sum_{i=1}^{N} Y_i = \bar{Y}_N$$

In SRSWOR samplemean is an unbiased estimate of population mean.

3) Prove that in simple Random sampling with Replacement (SRSWR). Sample mean $(\bar{y})$ is an unbiased Estimate of the population mean $(\bar{Y})$

i.e $E(\bar{y}) = \bar{Y}$ & $E(\bar{y}_n) = \bar{Y}_N$

If $y_1, y_2 - - - y_n$ units are drawn from $Y_1, Y_2, ---$ $- Y_N$ units of population with the probability $\frac{1}{N}$ at every draw ($\because y_i$ gets from any one of

$\therefore E(y_i) = \overline{Y}_N \;\forall\; i$ ... indepe

$\therefore E(\bar{y}_n) = E\left[\frac{1}{n}\sum_{i=1}^{n} y_i\right)$

$= \frac{1}{n}\sum_{i=1}^{n} E(y_i)$

$= \frac{1}{n}\sum_{i=1}^{n} \overline{Y}_N$ [from eq^n (1)]

$= \frac{1}{n} n\,\overline{Y}_N = \overline{Y}_N$

In SRSWR sample mean is an unbiased estimator of population mean.

Theorem (3)^{R} :- show that in simple Random Sampling without Replacement (SRSWOR) the sample mean square (sample varience s^2) is an unbiased Estimator of the population mean square (S^2) ie $E(s^2) = S^2$

Proof :- Suppose the population consists of N units $Y_1, Y_2, \ldots Y_N$ from which the sample of size n $y_1, y_2, \ldots y_n$ are drawn without replacement

Sample mean $\bar{y}_n = \frac{\sum_{i=1}^{n} y_i}{n}$

We know that $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y}_n)^2$

$$s^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} y_i^2 - n\bar{y}_n^2\right]$$

$$= \frac{1}{n-1}\left[\sum_{i=1}^{n} y_i^2 - n\left[\frac{\sum_{i=1}^{n} y_i}{n}\right]^2\right]$$

$$= \frac{1}{n-1}\left[\sum_{i=1}^{n} y_i^2 - \frac{n}{n^2}\left[\sum_{i=1}^{n} y_i\right]^2\right]$$

$$= \frac{1}{n-1}\left[\sum_{i=1}^{n} y_i^2 - \frac{1}{n}\left[\sum_{i=1}^{n} y_i^2 + \sum\sum_{i \neq j=1}^{n} y_i y_j\right]\right]$$

$$\left[\because \left(\sum_{i=1}^{n} y_i\right)^2 = \sum_{i=1}^{n} y_i^2 + \sum\sum_{i \neq j=1}^{n} y_i y_j\right]$$

$$= \frac{1}{n-1}\sum_{i=1}^{n} y_i^2 - \frac{1}{n(n-1)}\sum_{i=1}^{n} y_i^2 - \frac{1}{n(n-1)}\sum\sum_{i \neq j=1}^{n} y_i y_j$$

$$= \left[\frac{1}{n-1} - \frac{1}{n(n-1)}\right]\sum_{i=1}^{n} y_i^2 - \frac{1}{n(n-1)}\sum\sum_{i \neq j=1}^{n} y_i y_j$$

$$= \frac{(n-1)}{n(n-1)}\sum_{i=1}^{n} y_i^2 - \frac{1}{n(n-1)}\sum\sum_{i \neq j=1}^{n} y_i y_j$$

$$= \frac{1}{n}\sum_{i=1} y_i^2 - \frac{1}{n(n-1)} \sum\sum_{i \neq j=1} y_i y_j \longrightarrow (1)$$

Taking Expectation on both sides

$$E(s^2) = \frac{1}{n} E\left[\sum_{i=1}^{n} y_i^2\right] - \frac{1}{n(n-1)} E\left[\sum\sum_{i \neq j=1}^{n} y_i y_j\right] \longrightarrow (2)$$

Consider $E\left[\sum_{i=1}^{n} y_i^2\right] = E\left[\sum_{i=1}^{N} a_i Y_i^2\right]$

where $a_i = \begin{cases} 1, \text{ if } i^{th} \text{ unit is included into the sample} \\ 0, \text{ if } i^{th} \text{ unit is not included into the sample} \end{cases}$

$$\therefore E\left[\sum_{i=1}^{n} y_i^2\right] = \sum_{i=1}^{N} E(a_i) Y_i^2$$

$$= \sum_{i=1}^{N} \frac{n}{N} Y_i^2$$

$$= \frac{n}{N} \sum_{i=1}^{N} Y_i^2 \text{ —— } (3)$$

and now consider

$$E\left[\sum\sum_{i \neq j=1}^{n} y_i y_j\right] = E\left[\sum\sum_{i \neq j=1}^{N} a_i a_j Y_i Y_j\right]$$

$$= \sum\sum_{i \neq j=1}^{N} E(a_i a_j) Y_i Y_j$$

Consider $E[a_i a_j] = 1\, P(a_i a_j = 1) + 0\, P(a_i a_j = 0)$

$$= 1\, P(a_i = 1 \cap a_j = 1)$$

$$= P(a_i = 1) \cdot P(a_j = 1 / a_i = 1)$$

$$E(a_i a_j) = \frac{n}{N} \frac{n-1}{N-1}$$

$$E\left(\sum\sum_{i \neq j=1}^{n} y_i y_j\right) = \sum\sum_{i \neq j=1}^{N} \frac{n(n-1)}{N(N-1)} Y_i Y_j$$

$$E\left(\sum\sum_{i \neq j=1}^{n} y_i y_j\right) = \frac{n(n-1)}{N(N-1)} \sum\sum_{i \neq j=1}^{N} Y_i Y_j \text{ —— } (4)$$

Substitute the values of (3) and (4) in Equation

we get

$$E(s^2) = \frac{1}{\cancel{n}} \frac{\cancel{n}}{N} \sum_{i=1}^{N} Y_i^2 - \frac{1}{n(\cancel{n-1})} \frac{n}{N} \frac{(\cancel{n-1})}{(N-1)} \sum\sum_{i \neq j=1}^{N} Y_i Y_j$$

$$E(s^2) = \frac{1}{N} \sum_{i=1}^{N} Y_i^2 - \frac{1}{N(N-1)} \sum\sum_{i \neq j=1}^{N} Y_i Y_j \text{ —— } (5)$$

(as compared with eqn (1) and eqn (5) for small letters, Just replacing Capital letters we get the same as above)

Compare eqn (1) and (5)

$E(s^2) = S^2$

$\therefore$ In SRSWOR the sample mean square is an unbiased estimate of population mean square (S^2)

Theorem ④ :- In SRSWOR varience of sample mean is given by $var(\bar{y}_n) = \frac{N-n}{N}\frac{S^2}{n}$ or $\left(1-\frac{n}{N}\right)\frac{S^2}{n} = \left(\frac{1}{n}-\frac{1}{N}\right)S^2$

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

where $f = \frac{n}{N}$ = sampling fraction

where S^2 = population mean square

Proof :- Sample mean = $\bar{y}_n = \frac{\Sigma y_i}{n}$

we have $var(\bar{y}_n) = E(\bar{y}_n^2) - (E(\bar{y}_n))^2$

$var(\bar{y}_n) = E(\bar{y}_n^2) - \bar{Y}_N^2 \longrightarrow ①$ $(\because E(\bar{y}_n) = \bar{Y}_N)$

$$\bar{y}_n = \frac{\sum_{i=1}^{n} y_i}{n}$$

Taking expectation and square on both sides we get

$$E(\bar{y}_n^2) = E\left(\frac{1}{n}\sum_{i=1}^{n} y_i\right)^2$$

$$= \frac{1}{n^2} E\left[\sum_{i=1}^{n} y_i\right]^2$$

$$= \frac{1}{n^2} E\left[\sum_{i=1}^{n} y_i^2 + \sum\sum_{i \neq j=1}^{n} y_i y_j\right]$$

$$\boxed{\because \left(\sum_{i=1}^{n} y_i\right)^2 = \sum_{i=1}^{n} y_i^2 + \sum\sum_{i\neq j=1}^{n} y_i y_j}$$

$$E(\bar{y}_n^2) = \frac{1}{n^2}\left[E\left(\sum_{i=1}^{n} y_i^2\right) + E\left[\sum\sum_{i\neq j=1}^{n} y_i y_j\right]\right] \longrightarrow ②$$

consider $E\left[\sum_{i=1}^{n} y_i^2\right] = E\left[\sum_{i=1}^{N} a_i Y_i^2\right] = \sum_{i=1}^{N} E(a_i) Y_i^2$

$$\boxed{E(a_i) = \frac{n}{N}}$$

$$E\left[\sum_{i=1}^{n} y_i^2\right] = \sum_{i=1}^{N} \frac{n}{N} Y_i^2 = \frac{n}{N}\sum_{i=1}^{N} Y_i^2 \longrightarrow ③$$

But $\sum_{i=1}^{N} (Y_i - \bar{Y}_N)^2 = \sum_{i=1}^{N} Y_i^2 - N\bar{Y}_N^2$

$$\sum_{i=1}^{N} Y_i^2 = \sum_{i=1}^{N} (Y_i - \bar{Y}_N)^2 + N\bar{Y}_N^2$$

$$\boxed{\because S^2 = \frac{1}{N-1}\sum_{i=1}^{N}(Y_i - \bar{Y}_N)^2}$$

$$\sum_{i=1}^{N} Y_i^2 = (N-1)S^2 + N\bar{Y}_N^2 \longrightarrow ④$$

substitute ④ and in ③ we get

$$E\left[\sum_{i=1}^{n} y_i^2\right] = \frac{n}{N}\left[(N-1)S^2 + N\bar{Y}_N^2\right]$$

$$E\left[\sum_{i=1}^{n} y_i^2\right] = n\left[\left(\frac{N-1}{N}\right)S^2 + \bar{Y}_N^2\right] \longrightarrow (5)$$

Consider $E\left[\sum\sum_{i\neq j=1}^{n} y_i y_j\right] = E\left[\sum\sum_{i\neq j=1}^{N} a_i a_j Y_i Y_j\right]$

$$E\left[\sum\sum_{i\neq j=1}^{n} y_i y_j\right] = \sum\sum_{i\neq j=1}^{N} E(a_i a_j) Y_i Y_j$$

$$E\left[\sum\sum_{i\neq j=1}^{n} y_i y_j\right] = \sum\sum_{i\neq j=1}^{N} \frac{n(n-1)}{N(N-1)} Y_i Y_j$$

$$E\left[\sum\sum_{i\neq j=1}^{n} y_i y_j\right] = \frac{n(n-1)}{N(N-1)} \sum\sum_{i\neq j=1}^{N} Y_i Y_j \longrightarrow (6)$$

$$\because \left(\sum_{i=1}^{N} Y_i\right)^2 = \sum_{i=1}^{N} Y_i^2 + \sum\sum_{i\neq j=1}^{N} Y_i Y_j$$

$$\sum\sum_{i\neq j=1}^{N} Y_i Y_j = \left(\sum_{i=1}^{N} Y_i\right)^2 - \sum_{i=1}^{N} Y_i^2$$

$$\sum\sum_{i\neq j=1}^{N} Y_i Y_j = (N\bar{Y}_N)^2 - (N-1)S^2 - N\bar{Y}_N^2 \quad [\text{from } (4)]$$

$$\sum\sum_{i\neq j=1}^{N} Y_i Y_j = N\bar{Y}_N^2[N-1] - (N-1)S^2$$

$$= (N-1)\left[N\bar{Y}_N^2 - S^2\right] \longrightarrow \cancel{(7)}$$

$$= N(N-1)\left(\bar{Y}_N^2 - \frac{S^2}{N}\right) \longrightarrow (7)$$

Substitute eq (7) in eqn (6)

$$E\left[\sum\sum_{i\neq j=1}^{n} y_i y_j\right] = \frac{n(n-1)}{N(N-1)}\left[N(N-1)\left(\bar{Y}_N^2 - \frac{S^2}{N}\right)\right]$$

$$E\left[\sum\sum_{i\neq j=1}^{n} y_i y_j\right] = n(n-1)\left[\bar{Y}_N^2 - \frac{S^2}{N}\right] \longrightarrow (8)$$

Substitute eqn (5) and eqn (8) in eqn (2)

$$E(\bar{y}_n^2) = \frac{1}{n^2}\left[E\left(\sum_{i=1}^{n} y_i^2\right) + E\left(\sum\sum_{i\neq j=1}^{n} y_i y_j\right)\right]$$

$$E(\bar{y}_n^2) = \frac{1}{n^2}\left[n\left(\frac{N-1}{N}S^2 + \bar{Y}_N^2\right) + n(n-1)\left(\bar{Y}_N^2 - \frac{S^2}{N}\right)\right]$$

$$E(\bar{y}_n^2) = \frac{1}{n}\left[\frac{N-1}{N}S^2 + \bar{Y}_N^2 + (n-1)\bar{Y}_N^2 - (n-1)\frac{S^2}{N}\right]$$

$$= \frac{1}{n}\left[\left(\frac{N-1-n+1}{N}\right)S^2 + (1+n-1)\bar{Y}_N^2\right]$$

$$= \frac{1}{n}\left(\frac{N-n}{N}S^2 + n\bar{Y}_N^2\right)$$

$$E(\bar{y}_n^2) = \frac{N-n}{Nn}S^2 + \bar{Y}_N^2 \longrightarrow (9)$$

Substitute the value of eqn (9) in eqn (1) we get

$$var(\bar{y}_n) = E(\bar{y}_n^2) - \bar{Y}_N^2$$

$$= \frac{N-n}{N}\frac{S^2}{n} + \bar{Y}_N^2 - \bar{Y}_N^2$$

$$var(\bar{y}_n) = \frac{N-n}{N}\frac{S^2}{n} = \left(1-\frac{n}{N}\right)S^2 = (1-f)S^2$$

$$var(\bar{y}_n) = \left(\frac{1}{n} - \frac{1}{N}\right)S^2$$

Hence the proof.

Standard Error of the sample mean is SRSWOR

$$SE(\bar{y}_n) = \sqrt{V(\bar{y}_n)} = \sqrt{\frac{N-n}{N}}\frac{S}{\sqrt{n}}$$

ESTIMATION OF SE OF SAMPLE MEAN IN SRSWOR

If S^2 is not known, then we estimate S^2 by the sample mean square s^2

since s^2 is an unbiased estimator of S^2.

Estimate of $SE(\bar{y}_n) = \sqrt{\frac{N-n}{N}}\frac{s}{\sqrt{n}}$

Sampling fraction

$\frac{n}{N}$ is called sampling fraction and is denoted by f

$\therefore f = \frac{n}{N}$

finite population correlation (f.p.c)

$(1-f)$ is called finite population correlation and is usually written as f.p.c

i.e f.p.c $= 1-f$

Note ① If population is very large, then

$f = \frac{n}{N} \to 0$ and the f.p.c $\to 1$

$$V(\bar{y}_n) = \frac{S^2}{n} \quad \left(\because V(\bar{y}_n) = (1-f)\frac{S^2}{n} = (1-0)\frac{S^2}{n} = \frac{S^2}{n}\right).$$

② since $\hat{Y} = N\bar{y}$

$$V(\hat{Y}) = V(N\bar{y})$$

$$= N^2V(\bar{y}) = N^2\frac{(N-n)}{N}\frac{S^2}{n}$$

$$= N(N-n)\frac{S^2}{n}$$

$$= \frac{N^2S^2}{n}$$

$$\therefore \frac{N-n}{N} = 1$$

$\therefore SE(\hat{Y}) = \frac{NS}{\sqrt{n}}$

Theorem ⑤ Prove that in SRSWR varience of the sample mean is given by $V(\bar{y}) = \frac{\sigma^2}{n}$ or $\frac{N-1}{N}\frac{S^2}{n}$ where σ^2, n are population varience and sample size.

Proof :- Let the sample observations $y_1, y_2, \ldots, y_n$ be independent and identically drawn from the population with the same varience σ^2

ie $var(y_i) = \sigma^2 \ \forall i$

$var(\bar{y}_n) = var\left[\frac{1}{n}\sum_{i=1}^{n} y_i\right]$

$= \frac{1}{n^2}\sum_{i=1}^{n} V(y_i)$ (covariance terms vanish since they are independent).

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

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

i.e $var(\bar{y}_n) = \frac{\sigma^2}{n} = \frac{N-1}{N}\frac{S^2}{n}$ ($\because (N-1)S^2 = N\sigma^2$)

If N is large $\frac{1}{N} \to 0$ $\left(\frac{N-1}{N} = 1 - \frac{1}{N} = 1 - 0 = 1\right)$

$\therefore V(\bar{y}_n) = \frac{S^2}{n}$

this is same as $V(\bar{y}_n)$ in SRSWOR

Note : ① Comparison of varience in SRSWOR and SRSWR varience of the sample mean in SRSWOR is less than that in SRSWR

we know $var(\bar{y}_n) = \frac{N-n}{N}\frac{S^2}{n}$ in SRSWOR

$var(\bar{y}_n) = \frac{N-1}{N}\frac{S^2}{n}$ in SRSWR

since sample size n is always greater than 1

$N-n < N-1$

$\therefore var(\bar{y}_n)$ SRSWOR $<$ $var(\bar{y}_n)$ SRSWR

Theorem ⑤ second method

Proof :- By defination varience of the sample mean $\bar{y}$

$$V(\bar{y}) = E(\bar{y} - E(\bar{y}))^2$$

$$= E(\bar{y} - \bar{Y})^2 \quad [\because E(\bar{y}) = \bar{Y}]$$

$$= E\left[\frac{\sum_{i=1}^{n} y_i}{n} - \bar{Y}\right]^2$$

$$= E\left[\frac{\sum_{i=1}^{n} y_i - n\bar{Y}}{n}\right]^2$$

$$= \frac{1}{n^2} E\left[\sum_{i=1}^{n} y_i - n\bar{Y}^2\right] = \frac{1}{n^2} E\left[\sum_{i=1}^{n}(y_i - \bar{Y})\right]^2$$

$$\Rightarrow V(\bar{y}) = \frac{1}{n^2} E\left[\sum_{i=1}^{n}(y_i - \bar{Y})^2 + \sum_{\substack{i=1 \\ i<j}}^{n}\sum_{j=1}^{n}(y_i - \bar{Y})(y_j - \bar{Y})\right]$$

$$\left(\because \left(\sum_{i=1}^{n} y_i\right)^2 = \sum_{i=1}^{n} y_i^2 + \sum_{i \neq j=1}\sum y_i y_j\right)$$

$$V(\bar{y}) = \frac{1}{n^2}\left(\sum_{i=1}^{n} E(y_i - \bar{Y})^2 + \sum_{i=1}^{n}\sum_{j=1}^{n} E(y_i - \bar{Y})(y_j - \bar{Y})\right] \rightarrow ①$$

$$\left.\begin{array}{l} E(y_i - \bar{y})^2 = V(y_i) = \sigma^2 \\ E(y_i - \bar{Y})(y_j - \bar{Y}) = \mathrm{cov}(y_i, y_j) = 0 \end{array}\right\} \text{—— ②}$$

In SRSWR the units are drawn independental

$\therefore$ Y_i and Y_j are Independent

Hence $\mathrm{cov}(Y_i, Y_j) = 0$

Substitute ② in ①

$$V(\bar{y}) = \frac{1}{n^2}\left[\Sigma\sigma^2 + 0\right] = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$$

$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(y_i - \bar{y})^2$ this is called the true variance

$$N\sigma^2 = \sum_{i=1}^{N}(y_i - \bar{Y})^2$$

But the Estimated varience $s^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \bar{Y})^2$

$$\Rightarrow (N-1)s^2 = \sum_{i=1}^{N}(y_i - \bar{Y})^2$$

$$(N-1)s^2 = N\sigma^2$$

$$\sigma^2 = \frac{N-1}{N}s^2$$

$$V(\bar{y}) = \frac{\sigma^2}{n} = \frac{N-1}{N}\frac{s^2}{n}$$

Theorem 6 :— In SRSWOR the sample mean square $\left(s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2\right)$ is an unbiased estimator of the population varience (σ^2) ie $E(s^2) = \sigma^2$

Proof :— As we know in SRSWOR

$$\left.\begin{array}{ll} E(y_i) = \bar{Y}_N & var(y_i) = \sigma^2 \\ E(\bar{y}_n) = \bar{Y}_N & var(\bar{y}_n) = \frac{\sigma^2}{n} \end{array}\right\} \forall i \quad \text{—} (1)$$

Now $E(s^2) = E\left[\frac{1}{n-1}\left[\sum_{i=1}^{n} y_i^2 - n\bar{y}_n^2\right]\right]$

$$= \frac{1}{n-1}\left[\sum_{i=1}^{n} E(y_i^2) - nE(\bar{y}_n^2)\right] \quad \text{—} (2)$$

we know that $var(y_i) = E(y_i^2) - (E(y_i))^2$ (from eqn (1))

$E(y_i^2) = var(y_i) + (E(y_i))^2 = \sigma^2 + \bar{Y}_N^2$ —(3)

and $E(\bar{y}_n^2) = var(\bar{y}_n) + (E(\bar{y}_n))^2 = \frac{\sigma^2}{n} + \bar{Y}_N^2 \rightarrow (4)$

Sub the values in eqn (3) and (4) in eqn (2) from (1) we get

$$E(s^2) = \frac{1}{n-1}\left[\sum_{i=1}^{n}(\sigma^2 + \bar{Y}_N^2) - n\left(\frac{\sigma^2}{n} + \bar{Y}_N^2\right)\right]$$

$$E(s^2) = \frac{1}{n-1}\left[n\sigma^2 + n\bar{Y}_N^2 - n\frac{\sigma^2}{n} - n\bar{Y}_N^2\right]$$

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

$E(s^2) = \sigma^2$ where $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$

Problem

1. Draw a sample of size 2 from the population consisting of 3, 5, 7 by SRSWR and SRSWOR methods. Compare the two methods through your results.

Sol Given population units are 3, 5, 7

Sample size $n = 2$

Population size $N = 3$

Population mean $\bar{Y} = \frac{\Sigma Y}{N} = \frac{3+5+7}{3}$

$= \frac{15}{3} = 5$

Population mean square

$$S^2 = \frac{1}{N-1}\Sigma(Y_i - \bar{Y})^2 = \frac{1}{3-1}\left[(3-5)^2 + (5-5)^2 + (7-5)^2\right]$$

$$S^2 = \frac{1}{2}\left[4+0+4\right] = \frac{8}{2} = 4$$

σ^2 = Population varience $= \frac{1}{N}\Sigma(Y_i - \bar{Y})^2$

$$\sigma^2 = \frac{1}{3}\left[(3-5)^2 + (5-5)^2 + (7-5)^2\right] = \frac{1}{3}\left[4+0+4\right] = \frac{8}{3}$$

no	sample values	Sample mean $\bar{y}_n$	$\bar{y}_n^2$	sample mean square deviation $s_i^2 = \frac{1}{n-1}\Sigma(y_i-\bar{y}_i)^2$
1	(3,3)	3	9	$s_1^2 = \frac{1}{2-1}[(3-3)^2+(3-3)^2] = 0$
2	(3,5)	4	16	$s_2^2 = \frac{1}{2-1}[(3-4)^2+(5-4)^2] = 2$
3	(3,7)	5	25	$s_3^2 = 8$
4	(5,3)	4	16	$s_4^2 = 2$
5	(5,5)	5	25	$s_5^2 = 0$
6	(5,7)	6	36	$s_6^2 = 2$
7	(7,3)	5	25	$s_7^2 = 8$
8	(7,5)	6	36	$s_8^2 = 2$
9	(7,7)	7	49	$s_9^2 = 0$
		45	237	24

$\Sigma\bar{y}_n = 45 \quad \Sigma\bar{y}_n^2 = 237 \quad \Sigma s_i^2 = 24$

$$V(\bar{y}_n) = \frac{\Sigma\bar{y}_{ni}^2}{N^n} - \left(\frac{\Sigma\bar{y}_{ni}}{N^n}\right)^2 = \frac{237}{9} - \left(\frac{45}{9}\right)^2 = 26.3333 - 25$$
$$= 1.3333$$

$$V(\bar{y}_n)_{SRSWR} = \frac{\sigma^2}{n} = \frac{2.6667}{2} = 1.3333$$ The formula is verified.

$$E(s^2) = \frac{1}{N^n}\sum_{i=1}^{N^n} s_i^2 = \frac{24}{9} = 2.6667 = \sigma^2$$

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

$$E(\bar{y}_n) = \frac{\Sigma\bar{y}_{ni}}{N^n} = \frac{45}{9} = 5 = \bar{Y}_N$$

ie $E(\bar{y}_n) = \bar{Y}_N$

SRSWOR possible samples $= N_{C_n} = {}^3C_2 = 3$

Sample no	sample values	$\bar{y}_n$	$\bar{y}_n^2$	s_i^2
1	(3,5)	4	16	2
2	(3,7)	5	25	8
3	(5,7)	6	36	2
		15	77	12

$\Sigma \bar{y}_n = 15 \quad \Sigma \bar{y}_n^2 = 77 \quad \Sigma s_i^2 = 12 \quad N_{C_n} = 3$

$$E(\bar{y}_n) = \frac{\Sigma \bar{y}_{n_i}}{N_{C_n}} = \frac{15}{3} = 5 = \bar{Y}_N$$

$$E(\bar{y}_n) = \bar{Y}_N$$

$$E(s^2) = \frac{\Sigma s_i^2}{N_{C_n}}$$

$$E(s^2) = \frac{12}{3} = 4$$

$$E(s^2) = S^2$$

$$V(\bar{y}_n) = \frac{\Sigma \bar{y}_{n_i}^2}{N_{C_n}} - \left(\frac{\Sigma \bar{y}_{n_i}}{N_{C_n}}\right)^2$$

$$= \frac{77}{3} - \left(\frac{15}{3}\right)^2$$

$$= 25.6667 - 25$$

$$= 0.6667$$

$$V(\bar{y}_n)_{SRSWOR} = \frac{N-n}{Nn} S^2 = \frac{3-2}{3\times 2} \times 4$$

$$= \frac{4}{6} = \frac{2}{3} = 0.6667$$

The formula is verified.

Estimation of Population mean, Population Total and varience of these estimates by SRSWR and SRSWOR (or) Estimation of Population parameters.

Estimation of population mean, Total and proportion is same in with and without replacement techniques whereas it will be different for Population variance.

① Estimation of Population mean

In simple Random sampling without replac... the ...

$\frac{N-1}{N}s^2$ is an unbiased estimator of the populat[ion] varience in SRSWOR

ie Population varience is Estimated by $\frac{N-1}{N}s^2$

ie $\hat{\sigma}^2 = \frac{N-1}{N}s^2$

Estimation of population varience in SRSWR.

In SRSWR we know that the sample mean square is an unbiased estimator of the population varience ie $E(s^2) = \sigma^2$

Therefore the population varience is estimated by sample mean square in SRSWR i.e $\hat{\sigma}^2 = s^2$

Question bank :-

1. Define simple random sampling. write its merits and demerits.
2. Distinguish between SRSWR and SRSWOR
3. Explain the selection procedure of simple Random sampling (what is simple random sampling ? mention the one of the method of drawing a random sample.) (or) Explain (a) Lottery method (b) Random numbers method in selecting SRS (Explain methods of selecting simple random sample).

- In SRSWOR show that $E(\bar{y}_n) = \bar{Y}_N$
- In SRSWOR show that $E(s^2) = S^2$
- In SRSWOR obtain $V(\bar{y}_n) = \frac{N-n}{N}\frac{S^2}{n}$

show that the sample mean is an unbiased estim-[at]or of population mean in ~~SRSWOR~~ and find its [va]rience in SRSWR $V(\bar{y}) = \frac{\sigma^2}{n}$

Define SRS show that sample mean is an unbia-[se]d estimator of population mean in SRSWOR

show that sample mean square is an unbiased [esti]mator of population mean square in SRSWOR.

UNIT-II

Stratified Random Sampling

① Explain the need for stratification of the population in sample surveys and the procedure of stratified sampling.

All the methods of sampling, the most commonly used procedure in surveys is stratified Random sampling when the population is hetrogeneous with respect to the variable or characteristic under study then the technique of stratified random sampling is used to obtain more efficient results.

If the population is heterogeneous, then we divide the entire population into relatively homogeneous sub-group called strata.

In every stratum (sub-group) we apply simple random sampling so that all the different groups should be represented proportionately equal.

Definition :- Stratification means division into layers. Auxiliary information ie past data or some other information related to the population characteristic under study may be used to divide the population into various groups such that

i) units within each group are as homogeneous as possible.

ii) The group means are as heterogeneous as possible.

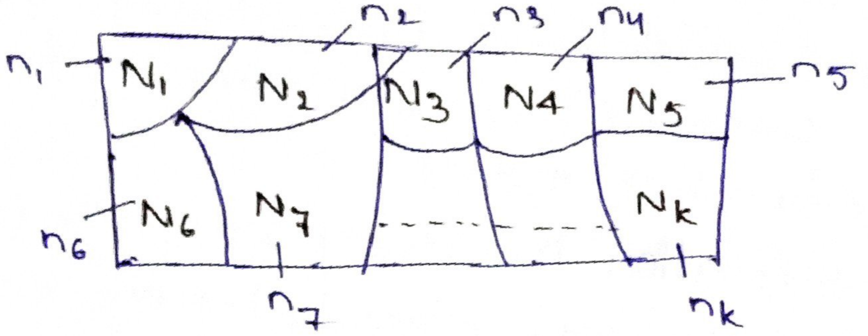
The population consist of N units is divided into k relatively homogeneous and mutually disjoint (means, non-overlapping) sub-groups, are known as strata of sizes $N_1, N_2, \ldots, N_k$ such that $\sum_{i=1}^{k} N_i = N$.

From each stratum, a sample is drawn uses simple random sampling without replacement (in general) of size n_i $(i = 1, 2, \ldots, k)$ respectively such that $\sum_{i=1}^{k} n_i = n$, the sample is known as stratified random sample of size n and the techinique of drawing such sample is called stratified random sampling.

The criterian used for the stratification of the universe into various strata is known as stratifying population into various strata and

ii) Selecting units at random from each stratum.

Diagramatic representation of stratified Random sampling



The stratified random sampling is to be conducted keeping in mind that the population is to be divided into strata properly and a suitable sample size is to be selected from each stratum. otherwise stratified random sample may not be trustworthy.

Example :- To study the cost of living of the people living in the state AP. In this case, the sampling data is to be collected from all income groups of people. since population units are heterogeneous, one cannot assure that all the income groups like high income group, middle income group, low income group etc. are equally represented into the sample if the sampling is done through the simple random sampling. this is the by the technique stratified random ".

Advantages of stratified Random sampling

• More Representative

usually some sampling techniques like simple random sampling, some groups may be over represented, some may under represented and some may not represented at all. stratified random sampling technique will be useful represent all the sub-groups (strata) in the population.

Hence stratified random sampling provide more representative cross-section of the population data.

Greater Accuracy :-

stratified random sampling provided more accurate increasing precision since it enables us to obtain the precision for each stratum.

3. Administrative Convenience

In stratified random sampling, the samples will be concentrated more geographically as compared to simple random sampling.

Hence the time and money involved in the collection of the data and interviewing the individuals may be reduced to a great extent.

The supervision of field work also may be done with greater case and convenience.

4. Sometimes the sampling Problems will be different in different strata.

for Example, the population may consist of literates and illiterates of people living in hostels, hospitals, Jails and those living in ordinary homes.

In such situations, Each of the above cases can be considered as a stratum.

This is a special advantages of stratified random sampling.

DISADVATAGES

1. Stratified Random sampling may give highly biased Estimates.

2. The selection of a stratified random sampling requires an up-to-date sampling frame. In practice up-to-dated sampling frame is not available then it is impossible to identify the sampling units. Hence Stratified Random sampling cannot be used.

Notations :-

Let N be the size of population

Let the population is divided into k strata of sizes $N_1, N_2 \dots N_k$ so that $N = \sum_{i=1}^{k} N_i$

Let us suppose that a simple random sample of size n_i (using SRSWOR) is drawn from the stratum N_i, $i = 1, 2, \dots k$ such that $n_i = \sum_{i=1}^{k} n_i$

N = Total number of population units

N_i = The number of population units in the i^{th} stratum

n_i = The number of units selected in SRSWOR from i^{th} stratum

$n = \sum_{i=1}^{k} n_i$: Total sample size drawn from all strata

Let Y_{ij} $(i=1,2,\ldots k,\ j=1,2,\ldots N_i)$ be the value of the j^{th} unit in the i^{th} stratum.

Population mean of i^{th} stratum $= \overline{Y}_{N_i} = \frac{1}{N_i}\sum_{j=1}^{N_i} Y_{ij}$

Population mean $= \overline{Y}_N = \frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{N_i} Y_{ij}$

$$= \frac{1}{N}\sum_{i=1}^{k} N_i \overline{Y}_{N_i} \qquad \left(\sum_{j=1}^{N_i} Y_{ij} = N_i \overline{Y}_{N_i}\right)$$

$$\overline{Y}_N = \sum_{i=1}^{k} P_i \overline{Y}_{N_i}$$

where $P_i = \frac{N_i}{N}$ is called the weight of i^{th} stratum.

Population mean square of the ith stratum

$$S_i^2 = \frac{1}{N_i - 1}\sum_{j=1}^{N_i}(Y_{ij} - \overline{Y}_{N_i})^2 \quad i=1,2,\ldots k$$

Population mean square $= S^2 = \frac{1}{N-1}\sum_{i=1}^{k}\sum_{j=1}^{N_i}(Y_{ij} - \overline{Y}_N)^2$

Let y_{ij} be the value of j^{th} sampled unit drawn from the stratum.

Mean of the sample selected from the i^{th} stratum

Mean of the sample selected from the i^{th} stratum

$$\overline{y}_{n_i} = \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij}, \quad i=1,2,\ldots k$$

Mean of the stratified sample

$$\overline{y}_n = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{N_i} y_{ij} \qquad \left(n_i \overline{y}_{n_i} = \sum_{j=1}^{n_i} y_{ij}\right)$$

or $\overline{y}_n = \frac{1}{n}\sum_{i=1}^{k} n_i \overline{y}_{n_i}$

Sample mean square of the i^{th} stratum

$$s_i^2 = \frac{1}{n_i - 1}\sum_{j=1}^{n_i}(y_{ij} - \overline{y}_{n_i})^2 \quad i=1,2\ldots k$$

Sample mean square

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \overline{y}_n)^2$$

We consider two estimates of the population mean $\overline{Y}_N$ which will be given below

$\overline{y}_n = \frac{1}{n}\sum_{i=1}^{k} n_i \overline{y}_{n_i}$, the stratified sample mean

If $\frac{n_i}{n} = \frac{N_i}{N}$, then stratified sample mean.

Considered as $\bar{y}_{st}$ and is given below

$$\bar{y}_{st} = \frac{1}{N}\sum_{i=1}^{k} N_i \bar{y}_{n_i}$$

$$= \sum_{i=1}^{k} P_i \bar{y}_{n_i}$$

where P_i is weight of the i^{th} stratum

Note :- $\bar{y}_n = \frac{1}{n}\sum_{i=1}^{k} n_i \bar{y}_{n_i}$

$$\bar{y}_{st} = \frac{1}{N}\sum_{i=1}^{k} N_i \bar{y}_{n_i}$$

these two estimates of the population mean are identical if $\frac{n_i}{n} = \frac{N_i}{N}$

$$\frac{n_i}{N_i} = \frac{n}{N} \text{ (constant)}$$

$$n_i = \frac{n}{N} N_i$$

$$n_i = C N_i$$

$$n_i \propto N_i$$

This is known as proportional allocation.

in stratified Random sampling

** Theorem 1 : show that in stratified random sample $\bar{y}_{st}$ is an the mean of the stratified random sample $\bar{y}_{st}$ is an unbiased estimator of the population mean $\bar{Y}_N$ ie

$$E(\bar{y}_{st}) = \bar{Y}_N$$

Proof : We know that in SRSWOR $E(\bar{y}_n) = \bar{Y}_N$

since the sample of size n_i $(i=1,2 \dots k)$ is drawn using SRSWOR from each of the stratum,

We know $E(\bar{y}_{n_i}) = \bar{Y}_{N_i}$ ——(1) $i=1,2,\dots k$

consider the stratified random sample mean

$$\bar{y}_{st} = \frac{1}{N}\sum_{i=1}^{k} N_i \bar{y}_{n_i}$$

where $\bar{y}_{n_i}$ is the mean of the sample drawn from the i^{th} stratum.

Now consider

$$E(\bar{y}_{st}) = E\left[\frac{1}{N}\sum_{i=1}^{k} N_i \bar{y}_{n_i}\right]$$

$$= \frac{1}{N}\sum_{i=1}^{k} N_i E(\bar{y}_{n_i})$$

$$= \frac{1}{N}\sum_{i=1}^{k} N_i \bar{Y}_{N_i} \quad (\because \text{from}(1))$$

$$= \bar{Y}_N \quad \left(\because \bar{Y}_N = \frac{1}{N}\sum_{i=1}^{k} N_i \bar{Y}_{N_i}\right)$$

$\because E(\bar{y}_{st}) = \bar{Y}_N$ in stratified Random Sampling

***

Theorem② show that stratified Random sampling variance of the sample mean (estimate of population mean) is

$$V(\bar{y}_{st}) = \frac{1}{N^2}\sum_{i=1}^{k} N_i(N_i - n_i)\frac{S_i^2}{n_i} \quad \text{where } S_i^2 = \frac{1}{N_i-1}\sum_{i=1}^{N_i}(y_{ij}-\bar{Y}_i)$$

$$= \sum_{i=1}^{k}\left(\frac{1}{n_i} - \frac{1}{N_i}\right)(P_i S_i)^2$$

Proof :- We know that in SRSWOR

$$V(\bar{y}_n) = \frac{N-n}{N}\frac{S^2}{n}$$

since the samples have drawn from each strata by using SRSWOR we have

$$V(\bar{y}_{n_i}) = \frac{N_i - n_i}{N_i}\frac{S_i^2}{n_i} \quad i = 1, 2, \dots k \quad \text{—①}$$

where N_i = size of i^{th} stratum

n_i = the sample is drawn from the i^{th} stratum by using SRSWOR

S_i^2 = population mean square of i^{th} stratum

Now consider

$$V(\bar{y}_{st}) = V\left(\frac{1}{N}\sum_{i=1}^{k} N_i \bar{y}_{n_i}\right)$$

$$= \frac{1}{N^2}\sum_{i=1}^{k} N_i^2 V(\bar{y}_{n_i})$$

$$= \frac{1}{N^2}\sum_{i=1}^{k} N_i^{\not 2}\frac{N_i - n_i}{N_i}\frac{S_i^2}{n_i} \quad [\text{from eqn ①}]$$

$$V(\bar{y}_{st}) = \frac{1}{N^2}\sum_{i=1}^{k} N_i(N_i - n_i)\frac{S_i^2}{n_i}$$

or $$V(\bar{y}_{st}) = \frac{1}{N^2}\sum_{i=1}^{k} N_i^2\left(\frac{1}{n_i} - \frac{1}{N_i}\right)S_i^2$$

$$= \sum_{i=1}^{k}\left(\frac{1}{n_i} - \frac{1}{N_i}\right)P_i^2 S_i^2 \quad \text{where } P_i = \frac{N_i}{N}$$

$$V(\bar{y}_{st}) = \sum_{i=1}^{k}\left(\frac{1}{n_i} - \frac{1}{N_i}\right)(P_i S_i)^2$$

Note ① $$V(\bar{y}_{st}) = \sum_{i=1}^{k}\frac{N_i^2}{N^2}\left(\frac{1}{n_i} - \frac{1}{N_i}\right)S_i^2$$

$$V(\bar{y}_{st}) = \sum_{i=1}^{k} P_i^2 S_i^2\left(\frac{1}{n_i} - \frac{1}{N_i}\right)$$

$$V(\bar{y}_{st}) = \sum_{i=1}^{k}\frac{P_i^2 S_i^2}{n_i} - \sum_{i=1}^{k}\frac{P_i^2 S_i^2}{N_i}$$

② $$V(\hat{y}_{st}) = V(N\bar{y}_{st}) = N^2 V(\bar{y}_{st})$$

$$= N^2\sum_{i=1}^{k}\frac{N_i}{N^2}(N_i - n_i)\frac{S_i^2}{n_i}$$

$$V(\hat{y}_{st}) = \sum_{i=1}^{k} N_i(N_i - n_i)\frac{S_i^2}{n_i}$$

We know that ... $\sum_{i=1} \frac{N_i^2}{N^2}\left(1-\frac{n_i}{N_i}\right)\frac{S_i^2}{n_i}$

$$= \sum_{i=1}^{k} P_i^2(1-f_i)\frac{S_i^2}{n_i} - \sum_{i=1}^{k} \frac{P_i^2 S_i^2}{n_i}$$

Explain the methods used for fixing the number of units to be selected from each stratum

(or)

Allocation of sample size of different strata

In stratified sampling, the allocation of the sample size to different strata is done by the consideration of three factors.

i. stratum size

ii. The variability with in the stratum

iii. The cost in talking sampling unit in the stratum.

A good allocation method is one which maximises the precision of the estimate with minimum resources.

To draw a sample size from the population, first of all we have to decide the size of the sub samples to be drawn from each strata.

In sample survey there are several types of allocation methods will be adopted, Among them most commonly used methods are.

1. Proportional allocation

2. Neymann allocation or optimum allocation.

1. Proportional allocation (Bowley allocation)

Proportional allocation is proposed by Bowley (1926)

This method of allocation is very common in practice, because of its simplicity. when no other information except stratum sizes is available we use this method. This procedure of allocation is called proportional allocation of n_i's to various strata is called proportional allocation.

Allocation of n_i's to various strata is called Proportional if the sample fraction is constant for each stratum.

ie $\frac{n_1}{N_1} = \frac{n_2}{N_2} = \cdots = \frac{n_i}{N_i} = \cdots = \frac{n_k}{N_k} = \frac{\sum_{i=1}^{k} n_i}{\sum_{i=1}^{k} N_i} = \frac{n}{N} = C$

ie $\frac{n_i}{N_i} = \frac{n}{N} = C$

$n_i = \frac{n}{N} N_i$

$n_i = C N_i$

from the various ... of optimization.
ie obtaining best results at minimum possible cost.
The rules given by Neyman for an optimum allocation n_i's $(i=1,2,\dots k)$ are determined so that
i) Minimize the varience for fixed sample size n
ie minimize $var(\bar{y}_{st})$ for fixed n.
ii) minimize the varience (ie, maximizing the precision) for fixed total cost c
ie $var(\bar{y}_{st})$ is minimum for fixed total cost c.
iii) minimise the total cost for fixed value of $V(\bar{y}_{st})$ (desired precision).
under the optimum allocation

$$n_i = n \frac{N_i S_i}{\sum_{i=1}^{k} N_i S_i}$$

$$n_i \propto N_i S_i \qquad \left(C = \frac{n}{\sum N_i S_i} = \text{Constant} \right)$$

Here n is the total sample size
n_i is the i^{th} sub sample size
N_i is the i^{th} stratum size
S_i^2 is the i^{th} stratum mean square deviation
and $$S_i^2 = \frac{\sum_{j=1}^{N_i} (y_{ij} - \bar{Y}_i)^2}{N_i - 1}$$

Theorem 4 :- With the usual notation, ...

$var(\bar{y}_{st})_{opt} \le var(\bar{y}_{st})_{prop} \le var(\bar{y}_n)_{ran}$

or $V_{opt} \le V_{prop} \le V_{ran}$

Comparision of stratified Random sampling with simple Random sampling

ie $var(\bar{y}_n)_{Ran} \ge var(\bar{y}_{st})_{prop} > var(\bar{y}_{st})_{opt}$

Proof :- we have variance of the Estimate of Population mean in different methods as

$$V(\bar{y})_{Ran} = \frac{N-n}{N}\frac{S^2}{n} = \left(\frac{1}{n}-\frac{1}{N}\right)S^2$$

$$V(\bar{y}_{st})_{prop} = \frac{N-n}{N}\frac{1}{n}\sum_{i=1}^{k} P_i S_i^2$$

$$= \left(\frac{1}{n}-\frac{1}{N}\right)\sum_{i=1}^{k} P_i S_i^2$$

$$V(\bar{y}_{st})_{opt} = \frac{1}{n}\left(\sum_{i=1}^{k} P_i S_i\right)^2 - \frac{1}{N}\sum_{i=1}^{k} P_i S_i^2$$

Consider $V_{prop} - V_{opt} = \left(\frac{1}{n}-\frac{1}{N}\right)\sum_{i=1}^{k} P_i S_i^2 - \frac{1}{n}\left(\sum_{i=1}^{k} P_i S_i\right)^2 + \frac{1}{N}\sum_{i=1}^{k} P_i S_i^2$

$$= \frac{1}{n}\sum_{i=1}^{k} P_i S_i^2 - \frac{1}{N}\sum_{i=1}^{k} P_i S_i^2 - \frac{1}{n}\left(\sum_{i=1}^{k} P_i S_i\right)^2 + \frac{1}{N}\sum_{i=1}^{k} P_i S_i^2$$

$$= \frac{1}{n}\left[\sum_{i=1}^{k} P_i S_i^2 - \left(\sum_{i=1}^{k} P_i S_i\right)^2\right]$$

$$= \frac{1}{n}\left[\sum_{i=1}^{k} P_i S_i^2 - (\bar{S})^2\right]$$ where $\bar{S} = \sum_{i=1}^{k} P_i S_i$ is a weighted average

$$= \frac{1}{n}\sum_{i=1}^{k} P_i (S_i - \bar{S})^2 \ge 0$$

$\therefore\ V_{prop} - V_{opt} \geq 0$

$V_{prop} \geq V_{opt} \Rightarrow V_{opt} \leq V_{prop} \longrightarrow (1)$

We know that $V_{Ran} = \left(\frac{1}{n} - \frac{1}{N}\right) s^2 \longrightarrow (2)$

we shall first express s^2 in terms of s_i^2

we have $s^2 = \frac{1}{N-1} \sum_{i=1}^{k} \sum_{j=1}^{N_i} (y_{ij} - \bar{Y}_N)^2$

$$(N-1)s^2 = \sum_{i=1}^{k} \sum_{j=1}^{N_i} (y_{ij} - \bar{Y}_{N_i} + \bar{Y}_{N_i} - \bar{Y}_N)^2$$

$$(N-1)s^2 = \sum_{i=1}^{k} \sum_{j=1}^{N_i} \left[(y_{ij} - \bar{Y}_{N_i}) + (\bar{Y}_{N_i} - \bar{Y}_N)\right]^2$$

$$(N-1)s^2 = \sum_{i=1}^{k} \sum_{j=1}^{N_i} (y_{ij} - \bar{Y}_{N_i})^2 + \sum_{i=1}^{k} \sum_{j=1}^{N_i} (\bar{Y}_{N_i} - \bar{Y}_N)^2 + 2 \sum_{i=1}^{k} \sum_{j=1}^{N_i} (y_{ij} - \bar{Y}_{N_i})(\bar{Y}_{N_i} - \bar{Y}_N)$$

$$= \sum_{i=1}^{k} (N_i - 1) s_i^2 + \sum_{i=1}^{k} N_i (\bar{Y}_{N_i} - \bar{Y}_N)^2 + 0$$

since $\sum_{j=1}^{N_i} (y_{ij} - \bar{Y}_{N_i}) = 0$ (algebraic sum of deviation take from its mean is always zero)

$$(N-1)s^2 = \sum_{i=1}^{k} (N_i - 1) s_i^2 + \sum_{i=1}^{k} N_i (\bar{Y}_{N_i} - \bar{Y}_N)^2$$

$\left(\because s_i^2 = \frac{1}{N_i - 1} \sum_{j=1}^{N_i} (y_{ij} - \bar{Y}_{N_i})^2\right)$

If we consider N_i and N are sufficiently large,

hence $N_i - 1 \simeq N_i$

$N - 1 \simeq N$

we get $NS^2 = \sum_{i=1}^{k} N_i s_i^2 + \sum_{i=1}^{k} N_i (\bar{Y}_{N_i} - \bar{Y}_N)^2$

$$s^2 = \sum_{i=1}^{k} \frac{N_i}{N} s_i^2 + \sum_{i=1}^{k} \frac{N_i}{N} (\bar{Y}_{N_i} - \bar{Y}_N)^2$$

$$s^2 = \sum_{i=1}^{k} P_i s_i^2 + \sum_{i=1}^{k} P_i (\bar{Y}_{N_i} - \bar{Y}_N)^2 \quad \left[P_i = \frac{N_i}{N}\right]$$

Substitute this value of s^2 in eqn (2), we get

$$V(\bar{y}_n)_{Ran} = \left(\frac{1}{n} - \frac{1}{N}\right)\left(\sum_{i=1}^{k} P_i s_i^2 + \sum_{i=1}^{k} P_i (\bar{Y}_{N_i} - \bar{Y}_N)^2\right)$$

$$V(\bar{y}_n)_{Ran} = \left(\frac{1}{n} - \frac{1}{N}\right) \sum_{i=1}^{k} P_i s_i^2 + \left(\frac{1}{n} - \frac{1}{N}\right) \sum_{i=1}^{k} P_i (\bar{Y}_{N_i} - \bar{Y}_N)^2$$

$$V(\bar{y}_n)_{Ran} = V_{prop} + \left(\frac{1}{n} - \frac{1}{N}\right) \sum_{i=1}^{k} P_i (\bar{Y}_{N_i} - \bar{Y}_N)^2$$

since $\left(\frac{1}{n} - \frac{1}{N}\right) \sum_{i=1}^{k} P_i (\bar{Y}_{N_i} - \bar{Y}_N)^2 \geq 0$

$V(\bar{y}_n)_{Ran} - V_{prop} \geq 0 \longrightarrow (3)$

... in stratified random sampling, the cost of obtaining information of a sample in one stratum will be usually different from other strata.

For example, the cost of collecting the data from rural areas will be more because of travelling expenses than from urban areas. Let c_i be the cost per unit in the i^{th} stratum and let 'a' be the overhead cost, then the cost function C in the stratified random sampling is $\boxed{C = a + \sum_{i=1}^{k} c_i n_i}$

Theorem 5 :- $var(\bar{y}_{st})$ is minimum for fixed total sample size n if $n_i \propto N_i S_i$

Proof :- We have to minimize

$$V(\bar{y}_{st}) = \frac{1}{N^2} \sum_{i=1}^{k} N_i(N_i - n_i)\frac{S_i^2}{n_i}$$

subject to the condition that $\sum_{i=1}^{k} n_i = n$

This is equivalent of minimising the Lagrangian function ϕ for the variations in n_i as given below.

$\phi = var(\bar{y}_{st}) + \lambda(\sum_{i=1}^{k} n_i - n)$ where λ is Lagrange multiple

$$\therefore \phi = \frac{1}{N^2}\sum_{i=1}^{k} N_i(N_i - n_i)\frac{S_i^2}{n_i} + \lambda\left(\sum_{i=1}^{k} n_i - n\right)$$

$$\phi = \frac{1}{N^2}\sum_{i=1}^{k} N_i\left(\frac{N_i}{n_i} - 1\right)S_i^2 + \lambda\left(\sum_{i=1}^{k} n_i - n\right)$$

minimise the function ϕ differentiating ϕ with respect ... and equate it to zero, we get

$$\frac{\partial \phi}{\partial n_i} = 0 \Rightarrow \frac{1}{N^2} N_i^2\left(\frac{-1}{n_i^2}\right)S_i^2 + \lambda(1) = 0 \longrightarrow (1)$$

$$\frac{-N_i^2 S_i^2}{N^2 n_i^2} + \lambda = 0 \quad \lambda = \frac{N_i^2 S_i^2}{N^2 n_i^2} \quad n_i^2 = \frac{N_i^2 S_i^2}{N^2 \lambda}$$

$$n_i = \frac{N_i S_i}{N\sqrt{\lambda}} \longrightarrow (2)$$

$$\frac{\partial^2 \phi}{\partial n_i^2} = \frac{N_i^2 S_i^2}{N^2}\left(\frac{2}{n_i^3}\right) > 0 \text{ from } (1)$$

value of n_i in eqn (2) provides minimum for this summing over i from $1,2,\dots k$ we get

$$\sum_{i=1}^{k} n_i = \frac{\sum_{i=1}^{k} N_i S_i}{N\sqrt{\lambda}}$$

$$n = \frac{\sum_{i=1}^{k} N_i S_i}{N\sqrt{\lambda}}$$

$$\sqrt{\lambda} = \frac{\sum_{i=1}^{k} N_i S_i}{Nn}$$

Substitute the value of $\sqrt{\lambda}$ in eqn (2)

we get $n_i = \dfrac{N_i S_i}{N \dfrac{\sum_{i=1}^{k} N_i S_i}{Nn}}$

$$n_i = \frac{n \cdot N_i S_i}{\sum_{i=1}^{k} N_i S_i} \quad \text{—— (3)}$$

this is the value of n_i in the optimum allocation for fixed total sample size n $\left(\frac{n}{\sum N_i S_i} = \text{constant} = c\right)$

$n_i \propto N_i S_i \quad i = 1,2 \dots k$

The value of n_i in eqn (3) is known as Neyman's formula for optimum allocation.

To obtain more precise estimates of the population mean we have to minimise the varience of the sample mean in optimum allocation.

For this, Neyman's optimum allocation suggests that the size of sample selected from the stratum to be more for the value of $N_i S_i$ is large

ie N_i is large and S_i is large.

Theorem 6 :- In stratified random sampling for a specified cost function, $var(\bar{y}_{st})$ is minimum if $n_i \propto \dfrac{N_i S_i}{\sqrt{C_i}}$

Proof :- We have to minimise $var(\bar{y}_{st}) = \frac{1}{N^2}\sum_{i=1}^{k} N_i(N_i - n_i)\frac{S_i^2}{n_i}$ → (1)

Subject to the condition that

$\dots t\ \sum^{k} C_i n_i$ —— (2) $\dots$ king the Lagrangian functi$\dots$ below where λ being Lagrange's multiplier

$$\therefore \phi = \frac{1}{N^2}\sum_{i=1}^{k} N_i(N_i-n_i)\frac{s_i^2}{n_i} + \lambda\left(\sum_{i=1}^{k} c_i n_i - C + a\right)$$

$$\phi = \frac{1}{N^2}\sum_{i=1}^{k} N_i\left(\frac{N_i}{n_i}-1\right)s_i^2 + \lambda\left(\sum_{i=1}^{k} c_i n_i - C + a\right) \quad \text{--- (3)}$$

To minimise the function ϕ differentiating ϕ with respect to n_i and equate it to zero, we get

$$\frac{d\phi}{dn_i} = 0 \Rightarrow \frac{N_i^2}{N^2}\left(\frac{-1}{n_i^2}\right)s_i^2 + \lambda c_i = 0$$

$$\frac{N_i^2 s_i^2}{N^2 n_i^2} = \lambda c_i$$

$$n_i^2 = \frac{N_i^2 s_i^2}{N^2 \lambda c_i}$$

$$n_i = \frac{N_i s_i}{N\sqrt{\lambda}\sqrt{c_i}} \quad \text{--- (4)}$$

$$\text{Now } \frac{d^2\phi}{dn_i^2} = \frac{N_i^2 s_i^2}{N^2}\left(\frac{2}{n^3}\right) > 0$$

$\therefore$ The value of n_i in Eqn (4) provides minimum for ϕ

$$\therefore n_i = \frac{N_i s_i}{N\sqrt{\lambda}\sqrt{c_i}}$$

We have to determine the value of λ for this summing over i from $1, 2, \ldots k$ we get

$$\sum_{i=1}^{k} n_i = \sum_{i=1}^{k}\frac{N_i s_i}{N\sqrt{\lambda}\sqrt{c_i}}$$

$$n = \frac{\sum_{i=1}^{k} N_i s_i/\sqrt{c_i}}{N\sqrt{\lambda}}$$

$$\sqrt{\lambda} = \frac{\sum_{i=1}^{k} N_i s_i/\sqrt{c_i}}{Nn}$$

Substitute the value of $\sqrt{\lambda}$ in Eqn (4) we get

$$n_i = \frac{N_i s_i/\sqrt{c_i}}{\frac{\sum_{i=1}^{k} N_i s_i/\sqrt{c_i}}{Nn}}$$

$$n_i = n\,\frac{N_i s_i/\sqrt{c_i}}{\sum_{i=1}^{k} N_i s_i/\sqrt{c_i}}$$

$$n_i \propto \frac{N_i s_i}{\sqrt{c_i}}$$

Systematic Random Sampling

** 1. Explain systematic sampling, merits and demerits of systematic sampling.

A. Systematic sampling is a very simple technique. It is generally used if the complete and up-to-date sampling frame (units) is available.

This sampling technique has a nice feature of selecting the whole sample with just one random strata.

A sampling technique in which only the first unit is selected with the help of random numbers or lottery method at random and the rest being automatically selecting according to some pre-determined pattern involving regular spacing of units is known as systematic random sampling or systematic sampling.

Suppose, there is a population with N units. If a sample of size n is to be selected, then divide the population into n groups such that each group consists of k units.

ie the population is divided into n groups in such a way that $N = nk$ or $k = \frac{N}{n}$ where k is an integer k is called the sampling interval.

Systamatic sampling consists in drawing a random number say i $(i \leq k)$ and selecting the unit corresponding to this number and every k^{th} unit subsequently.

Thus the systematic sample of size n consist of the units.

$i, i+k, i+2k, \dots, i+(n-1)k$

The random number i is called the random strata and its value determines the whole sample.

random Sampling.

Example :- If $N=200$ $n=10$ and $k=\frac{N}{n}=\frac{200}{10}=20$

suppose unit number 15 is selected at random from the form the first 20 units, then the remaining units at the sample are 35, 55, 75, 95, 115, 135, 155, 175 and 195.

Merits (Advantages)

1. Systematic sampling is a very simple technique operationally it is more convenient that simple random sampling & stratified random sampling.

2. Time, work and money required for selecting a systematic sample is relatively much less.

3. Systematic sampling may be more efficient than simple random sampling provided the sampling frame is arranged completely at random.

4. The most commonly used to achieve randomness is alphabetical order of the population units. for Example, Telephone directory.

Demerits (disadvantages)

• The main Disadvantages of systematic sampling is that systematic samples are not generally random samples, only the first unit is selected at random.

To achieve at randomness, it is very difficult to get randomly arranged sampling frame.

If the population size N is not a multiple of the sample size n, then the actual sample size is different from the required sample size.

In that case the sample mean will not be an unbiased estimate of the population mean.

... sampling may give highly biased

$$V(\bar{y}_{sys}) = \frac{1}{k}\sum_{i=1}^{k}(\bar{y}_{i.} - \bar{y})^2$$

It is not possible to obtain an unbiased estimate of this variance (population varience).

NOTATIONS AND TERMINOLOGY

Let y be the population characteristic

N = Total number of population units

n = Total number of sample units

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

y_{ij} = value of j^{th} unit of the i^{th} sample

$(i = 1, 2, \ldots k;\ j = 1, 2, \ldots, n)$

$\bar{y}_{sys} = \bar{y}_{i.}$ = Mean of i^{th} systematic sample (or) mean of the i^{th} random start in the sample.

$$= \frac{1}{n}\sum_{j=1}^{n} y_{ij}, \quad i = 1, 2, \ldots k.$$

$\bar{y}_{..}$ = population mean

$$= \frac{1}{nk}\sum_{i=1}^{k}\sum_{j=1}^{n} y_{ij}$$

$$(\text{or}) = \frac{1}{k}\sum_{i=1}^{k}\bar{y}_{i.}$$

S^2 = population mean square

$$= \frac{1}{N-1}\sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{..})^2 = \frac{1}{nk-1}\sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{..})^2$$

The k possible systematic sample together with means and probability are given below in the following table.

Random start	Sample Composition	Probability	Mean
1	$1, 1+k, \ldots 1+jk, \ldots 1+(n-1)k$	$1/k$	$\bar{y}_{1.}$
2	$2, 2+k, \ldots 2+jk, \ldots 2+(n-1)k$	$1/k$	$\bar{y}_{2.}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
i	$i, i+k, \ldots i+jk, \ldots i+(n-1)k$	$1/k$	$\bar{y}_{i.}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
k	$k, 2k, \ldots (j+1)k, \ldots nk$	$1/k$	$\bar{y}_{k.}$

In this table, the units of the ... of the ... also
-atic samples. The columns ... very clear from the ...
refered as n strata. It is ... N units occur once in only one
that each of the ... has an equal chance of being
the k samples and thus has an equal chance of being selected
included into the sample since the probability of selecting
the i^{th} systematic sample is $\frac{1}{k}$, $i = 1, 2, \ldots, k$.

Theorem 1:— Sample mean is an unbiased estimator of the population mean in systematic sampling.

Proof :— We know,

Population mean $= \bar{y}_{..} = \frac{1}{N} \sum_{i=1}^{k} \sum_{j=1}^{n} y_{ij}$ (or) $= \frac{1}{nk} \sum_{i=1}^{k} \sum_{j=1}^{n} y_{ij}$

$= \frac{1}{k} \sum_{i=1}^{k} \left(\frac{1}{n} \sum_{j=1}^{n} y_{ij} \right)$

(or) $= \frac{1}{k} \sum_{i=1}^{k} \bar{y}_{i.}$

Consider

$$E(\bar{y}_{sys}) = E(\bar{y}_{i.}) = \bar{y}_{1.}\left(\frac{1}{k}\right) + \bar{y}_{2.}\left(\frac{1}{k}\right) + \cdots + \bar{y}_{k.}\left(\frac{1}{k}\right)$$

$$= \frac{1}{k} \sum_{i=1}^{n} \bar{y}_{i} = \bar{y}_{..}$$

∴ The sample mean is an ~~ut~~ unbiased estimator of the population mean.

i.e., $E(\bar{y}_{i.}) = E(\bar{y}_{sys}) = \bar{y}_{..}$

****

Theorem 2 :— Varience of the systematic sample mean is given by $var(\bar{y}_{sys}) = \frac{N-1}{N} \cdot S^2 - \frac{k(n-1)}{N} \cdot S^2_{wsy}$

Where S^2_{wsy} = population mean square among the units which i.e., within the same systematic sample.

$$= \frac{1}{k(n-1)} \sum_{i=1}^{k} \sum_{j=1}^{n} (y_{ij} - \bar{y}_{i.})^2$$

Proof :— Varience of the systematic sample mean is $var(\bar{y}_{sys}) = var(\bar{y}_{i.})$

$$= E\left[\bar{y}_{i.} - E(\bar{y}_{i.})\right]^2$$

$$= E\left[\bar{y}_{i.} - \bar{y}_{..}\right)^2 \text{ (from theorem)}$$

$$= \frac{1}{k} \sum_{i=1}^{k} (\bar{y}_{i.} - \bar{y}_{..})^2$$

As we know, the population mean square in the systematic sampling, $S^2 = \frac{1}{N-1} \sum_{i=1}^{k} \sum_{j=1}^{n} (y_{ij} - \bar{y}_{..})^2$

$$(N-1)S^2 = \sum_{i=1}^{k} \sum_{j=1}^{n} (y_{ij} - \bar{y}_{i.} + \bar{y}_{i.} - \bar{y}_{..})^2$$

$$= \sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij}-\bar{y}_{i.})^2 + \sum_{i=1}^{k}\sum_{j=1}^{n}(\bar{y}_{i.}-\bar{y}_{..})^2 + 2\sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij}-\bar{y}_{i.})(\bar{y}_{i.}-\bar{y}_{..})$$

Covariance term vanish

$$\sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij}-\bar{y}_{i.})(\bar{y}_{i.}-\bar{y}_{..}) = \sum_{i=1}^{k}(\bar{y}_{i.}-\bar{y}_{..})\sum_{j=1}^{n}(y_{ij}-\bar{y}_{i.})$$

$$= \sum_{i=1}^{k}(y_{i.}-\bar{y}_{..})\left(\sum_{j=1}^{n}y_{ij}-n\bar{y}_{i.}\right)$$

$$= 0 \qquad \left(\because \sum_{j=1}^{n}y_{ij}=n\bar{y}_{i.}\right)$$

$$\therefore (N-1)s^2 = \sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij}-\bar{y}_{i.})^2 + \sum_{i=1}^{k}\sum_{j=1}^{n}(\bar{y}_{i.}-\bar{y}_{..})^2$$

$$(N-1)s^2 = \sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij}-\bar{y}_{i.})^2 + n\sum_{i=1}^{k}(\bar{y}_{i.}-\bar{y}_{..})^2$$

$$(N-1)s^2 = k(n-1)s^2_{wsy} + nk\,v(\bar{y}_{sys}) \quad (\text{from } 1)$$

Divide with $N=nk$, we get

$$\frac{N-1}{N}s^2 = \frac{k(n-1)}{N}s^2_{wsy} + \frac{\cancel{nk}\,v(\bar{y}_{sys})}{\cancel{N}}$$

$$v(\bar{y}_{sys}) = \frac{N-1}{N}s^2 - \frac{k(n-1)}{N}s^2_{wsy} \qquad (\because N=nk)$$

$$v(\bar{y}_{sys}) = \frac{N-1}{N}s^2 - \frac{k(n-1)}{nk}s^2_{wsy}$$

Theorem 3 :- varience of the systematic sample mean

$$v(\bar{y}_{sys}) = \frac{nk-1}{nk}\frac{s^2}{n}\left[1+(n-1)\rho\right]$$

where ρ is the intraclass correlation coefficient between the units of the same systematic sample and is given by

$$\rho = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij}-\bar{y}_{..})(y'_{ij}-\bar{y}_{..})}{(n-1)(nk-1)s^2}$$

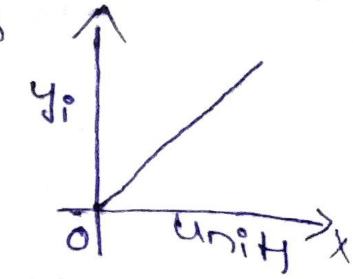
Proof :- We know

$$v(\bar{y}_{sys}) = \frac{1}{\ }\sum^{k}(\bar{y}_{i.}-\bar{y})^2$$

... If the population ... linear trend, then prove that

$$var(\bar{y}_{st}) \le var(\bar{y}_{sys}) \le var(\bar{y}_n)_{ran}$$

Proof :- Let us consider the population has a linear trend with the population units $Y_1, Y_2, \ldots, Y_N$ takes the values $1, 2, \ldots, N$ ie $Y_i = i$ $i = 1, 2, \ldots, N$



Then population Total

$Y_i = ai + b$

$a = 1$ $b = 0$

$Y_i = i$

$$\sum_{i=1}^{N} y_i = \sum_{i=1}^{N} i = 1 + 2 + \cdots + N$$

$$= \frac{N(N+1)}{2}$$

Population mean $= \bar{Y}_N = \frac{1}{N}\sum_{i=1}^{N} Y_i \Rightarrow \frac{1}{N}(1 + 2 + \cdots + N)$

$$\bar{Y}_N = \frac{\cancel{N}(N+1)}{2\cancel{N}} = \frac{N+1}{2}$$

$$\sum_{i=1}^{N} Y_i^2 = \sum_{i=1}^{N} i^2 = 1^2 + 2^2 + \cdots + N^2 = \frac{N(N+1)(2N+1)}{6}$$

Population mean square $= S^2 = \frac{1}{N-1}\left[\sum_{i=1}^{N} Y_i^2 - N\bar{Y}_N^2\right]$

$$S^2 = \frac{1}{N-1}\left[\frac{N(N+1)(2N+1)}{6} - N\left(\frac{N+1}{2}\right)^2\right]$$

$$= \frac{1}{N-1}\left[\frac{N(N+1)(2N+1)}{6} - \frac{N(N+1)^2}{4}\right]$$

$$= \frac{N(N+1)}{2(N-1)}\left[\frac{2N+1}{3} - \frac{N+1}{2}\right]$$

$$S^2 = \frac{N(N+1)}{2(N-1)}\left[\frac{4N+2-3N-3}{6}\right]$$

$$S^2 = \frac{N(N+1)}{2(\cancel{N-1})}\left[\frac{\cancel{N-1}}{6}\right]$$

$$S^2 = \frac{N(N+1)}{12}$$

$\therefore$ variance of the sample mean in SRSWOR

$$V(\bar{y}_n)_{Ran} = \frac{N-n}{Nn} S^2$$

$$= \frac{N-n}{\cancel{N}n}\,\frac{\cancel{N}(N+1)}{12} = \frac{nk-n}{n}\,\frac{(nk+1)}{12}$$

$$V(\bar{y}_n)_{Ran} = \frac{\cancel{n}(k-1)}{\cancel{n}}\,\frac{(nk+1)}{12} = \frac{(k-1)(nk+1)}{12} \quad \text{—— (1)}$$

In Stratified Random Sampling variance of the Estimate of Population mean

$$V(\bar{y}_{st}) = \frac{1}{N^2}\sum_{i=1}^{k} N_i(N_i-n_i)\frac{S_i^2}{n_i}$$

Let us assume that the varience in Each stratum of the population is Equal and it is S_i^2 and $n_i = \frac{n}{N}N_i$

$$V(\bar{y}_{st}) = \frac{1}{N^2}\sum_{i=1}^{k}\cancel{N_i}\left(N_i - \frac{n}{N}N_i\right)\frac{S_i^2}{\frac{n}{N}\cancel{N_i}}$$

~~Let us~~

$$= \frac{1}{N^2}\sum_{i=1}^{k} N_i\left(1-\frac{n}{N}\right)\frac{S_i^2 N}{n}$$

$$= \frac{S_i^2}{N^2}\sum_{i=1}^{k} N_i\left(1-\frac{n}{N}\right)$$

$$= \frac{S_i^2}{N^2}\,\frac{N-n}{N}\sum_{i=1}^{k} N_i$$

$$= \frac{S_i^2}{N^2}\left(\frac{N-n}{N}\right)\frac{N}{n}N$$

$$= \frac{S_i^2}{N^2}\left(\frac{N-n}{\cancel{N}}\right)\frac{\cancel{N}}{n}N$$

$$V(\bar{y}_{si}) = \frac{S_i^2(N-n)}{Nn}$$

We have $S^2 = \frac{N(N+1)}{12}$ for population as N units.

But we have k units in each stratum

$$= \frac{k(k+1)(nk-n)}{12nkn} \qquad (N=nk)$$

$$= \frac{k(k+1)n(k-1)}{12nkn} = \frac{k^2-1}{12n} \quad \text{—} (2)$$

Population units which have the linear trend can be divided as systematic samples

$$\bar{y}_{i.} = \frac{1}{n}\sum_{j=1}^{n} y_{ij}$$

$$= \frac{1}{n}\left[i+(i+k)+(i+2k)+ \cdots + i+(n-1)k\right]$$

$$= \frac{1}{n}\left[ni+[1+2+ \cdots +(n-1)]k\right]$$

$$= \frac{1}{n}\left(ni+\frac{(n-1)n}{2}k\right) = i+\frac{(n-1)}{2}k$$

and $\bar{y}_{..} = \bar{Y}_N = \frac{N+1}{2} = \frac{nk+1}{2}$

$$\therefore \bar{y}_{i.}-\bar{y}_{..} = i+\frac{(n-1)k}{2} - \frac{nk+1}{2} = i-\frac{k+1}{2}$$

Now $V(\bar{y}_{sys}) = \frac{1}{k}\sum_{i=1}^{k}(\bar{y}_{i.}-\bar{y}_{..})^2$

$$= \frac{1}{k}\sum_{i=1}^{k}\left(i-\frac{k+1}{2}\right)^2$$

$$= \frac{1}{k}\left[\sum_{i=1}^{k} i^2 + \sum_{i=1}^{k}\left(\frac{k+1}{2}\right)^2 - 2\frac{(k+1)}{2}\sum_{i=1}^{k} i\right]$$

$$= \frac{1}{k}\left[\frac{k(k+1)(2k+1)}{6} + \frac{k(k+1)^2}{4} - \frac{(k+1)k(k+1)}{2}\right]$$

$$= \cancel{k}\,\frac{k+1}{2}\left[\frac{2k+1}{3}+\frac{k+1}{2}-(k+1)\right]$$

$$= \frac{k+1}{2}\left[4k+2+3k+3-6k-6\right]$$

$$= \frac{(k+1)}{2}\cdot\frac{(k-1)}{6}$$

$$V(\bar{y}_{sys}) = \frac{k^2-1}{12} \quad \text{—} (3)$$

From equations (1), (2) and (3) we get

$V(\bar{y}_{st}) : V(\bar{y}_{sys}) : V(\bar{y}_n)_{Ran} : :$

$$\frac{k^2-1}{12n} : \frac{k^2-1}{12} : (k-1)(nk+1)$$

$\frac{k+1}{n} : k+1 : nk+1$

which is approximately equal to $\frac{1}{n} : 1 : n$.

from ①, ② & ③ and putting them in a ratio we get

$$V(\bar{y}_{st}) : V(\bar{y}_{sys}) : V(\bar{y}_{Ran}) :: \frac{k^2-1}{12n} : \frac{k^2-1}{12} : \frac{(k-1)(nk-1)}{12}$$

Dividing by $\frac{k-1}{12}$ we get the ratio as $\frac{1}{n} : 1 : \frac{nk+1}{k+1}$

But $\frac{nk+1}{k+1} = \frac{(nk+1)/k}{(k+1)/k}$ (dividing Nr and Dr by k)

$= \frac{n+\frac{1}{k}}{1+1/k} = n$ $\because$ for large k, $\frac{1}{k} \to 0$

The variance are in the ratio

$\frac{1}{n} : 1 : n$ (approx)

$\Rightarrow \frac{1}{n} \leq 1 \leq n$

Hence $V(\bar{y}_{st}) \leq V(\bar{y}_{sys}) \leq V(\bar{y}_{Ran})$

The equality sign holds good if $n=1$

In that case each varience $= (k+1)$

***

Theorem 5 :- With usual notations prove that mean of a systematic sample is more precise (more efficient) than the mean of a simple random sample if and only if $S^2_{wsy} > S^2$

In other works $V(\bar{y}_{sys}) < V(\bar{y}_{Ran})$ if $S^2_{wsy} > S^2$

Proof :- In simple Random sampling without replacement variance of the sample mean is given by

$$V(\bar{y}) = \frac{N-n}{N} \frac{S^2}{n}$$

Similarly in systematic sampling

$$V(\bar{y}_{sys}) = \frac{N-1}{N} S^2 - \frac{k(n-1)}{N} S^2_{wsy}$$

$V(\bar{y}_{sys})$ should be less than $V(\bar{y})$

$$\frac{N-1}{N} S^2 - \frac{k(n-1)}{N} S^2_{wsy} < \frac{N-n}{N} \frac{S^2}{n}$$

$$\left[(N-1) - \frac{(N-n)}{n}\right] S^2 < k(n-1) S^2_{wsy}$$

$\frac{N(n-1)}{N} S^2 < k(n-1) S^2_{wsy}$

$\frac{nk(n-1)}{n} S^2 < k(n-1) S^2_{wsy}$

$k(n-1) S^2 < k(n-1) S^2_{wsy}$

$S^2 < S^2_{wsy} \Rightarrow S^2_{wsy} > S^2$

$V(\bar{y}_{sys}) < V(\bar{y}_{Ran})$

Thus $V(\bar{y}_{sys}) < V(\bar{y}_{Ran})$ if $S^2_{wsy} > S^2$

Hence mean of a systematic sample.

UNIT III Analysis of variance (ANOVA)

1. Explain the concept of analysis of variance (ANOVA) and also assumption, uses.

Sol :- The analysis of varience is a powerful statical tool for the test of significance (hypothesis)

1. The test of significance based on t-distribution.
2. The significant of difference between two sample means using t-distribution.
3. But if we want test the significance difference of three ore more sample means.
4. Hence t-test cann't be used.
5. We need an alternative procedure is required for testing the homogenity of several means.

Eg :- five fertilizers are applied to four parts we may be intersted to find out if the effect of there fertilizers. The yield is significantly different.

* The answer to these problem is provided by the technique of analysis of varience to test the homogenity of several means.
* The alternative procedure is ANOVA.
* The term analysis of varience was introduced by Prof. R.A fisher in 1920. He said varience is inherent in nature.
* ANOVA is based on F-distribution.
* The analysis of varience is a powerful statical tool which support the controlled and uncontrolled variences in the data.

Defination :- In the words of R.A. fisher ANOVA is the separation of varience ascribable to one group of causes from the varience ascribable to the other group.

Causes of variation :-

1. The variation in any experiment is inherient in nature.
2. The total variation in any numerical data is due to number of factors or causes. They as classified as

3. The variation due to assignable causes are called controlled and measured. Hence there are variation.

4. The variation due to chance causes is beyond the control of human hand and they can't be controlled. This variation is also known as error variation.

5. This variation is also known as ANOVA.

6. It place an important role in ANOVA.

Uses :-

The technique of ANOVA is useful in many fields like agriculture, economics, biology, Education, Psycology, Business and so on.

Assumptions :-

The technique of ANOVA is based on f-test. for the validity of f-test, the following assumptions are made in analysis of varience.

1. The sample observations are independent.
2. The parent population from which the sample are taken is normal.
3. various effects are additive in nature.
4. E_{ij} is a random error.

   $E_{ij} \sim N(0, \sigma^2_e)$

5. The analysis of varience may be clasified as
   (i) one-way classification
   (ii) Two-way classification

Eg :-

Suppose four types of fertilizers are used to test their effects on the yeild of paddy is significantly different or they have the same effect.

② Discuss one-way classification of data and write down the analysis of variance table.

⇒ ANOVA is the simplest technique.

⇒ In one way ANOVA observations are classified into groups (or) classes (or) simples on the basic of single ceriterion i.e; one assignable cause

for example 'n' students of a class are arranged

According to their marks

⇒ Let us suppose that N observations y_{ij} $\begin{bmatrix} i=1,2,\dots k \\ j=1,2,\dots n_i \end{bmatrix}$ of sizes $n_1, n_2, \dots, n_k$

⇒ The sample data can be arrange into k classes as shown below.

classes (Treatment)	1 2 ... j ... n_i	total	Mean
1	y_{11} y_{12} ... y_{1j} ... y_{1n_1}	$T_{1.}$	$\bar{y}_{1.}$
2	y_{21} y_{22} ... y_{2j} ... y_{2n_2}	$T_{2.}$	$\bar{y}_{2.}$
⋮	⋮	⋮	⋮
i	y_{i1} y_{i2} ... y_{ij} ... y_{in_i}	$T_{i.}$	$\bar{y}_{i.}$
⋮	⋮	⋮	⋮
k	y_{k1} y_{k2} ... y_{kj} ... y_{kn_k}	$T_{k.}$	$\bar{y}_{k.}$
		G	$\bar{y}_{..}$

$$\bar{y} = \frac{\Sigma \bar{y}_i}{k}$$

G = Grand total

$\bar{y}$ = over all mean

If $n_1 = n_2 = \dots = n_k = n$ (say) then the data is called one way classified data with equal no. of observation. otherwise the data is known as one-way classified data with unequal no. of observations.

The total variations in the observations y_{ij} can be divided into the following two components.

i) The variation between the classes i.e., Treatment Effect; this is due to assignable causes which can be detected and control by human hand.

ii) The variation within the classes. is know as Error Effect. i.e., inherent variation is due to chance causes which cannot be control of human hand.

This is due to Chance causes which cannot be control of human hand.

⇒ The sources of variation in the data are

i) Effect of the treatment of d_i $i=1,2,\dots k$.

ii) Error due to chance causes [Random causes]

$\epsilon_{ij} \sim N(0, \sigma_e^2)$

The main objective of ANOVA technique is to Examine if the variation due to different classes is significant.

Assumptions :-

is normal.

⇒ various effects are additive in nature

⇒ E_{ij} is a random error.

$E_{ij} \sim N(0, \sigma^2)$

Statistical Analysis of one-way classification :-

T_i = Total yield of i^{th} treatment.

$$T_{i.} = \sum_{j=1}^{n_i} y_{ij}$$

$$G = \sum_{i=1}^{k} \sum_{j=1}^{n_i} y_{ij}$$

[G = Grand total
N = Total No. of observations]

$$N = \sum_{i=1}^{k} n_i = nk$$

$$\bar{y}_{..} = \frac{G}{N} \text{ [over all mean]}$$

$$\bar{y}_{i.} = \frac{\sum_{j=1}^{n_i} y_{ij}}{n_i} = \text{Mean of } i^{th} \text{ class.}$$

NULL Hypothesis :-

NH : H_0 : All the treatment are homogeneous

NH : H_0 : Population means are equal

i.e, $H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$

Mathematical Model of one-Way Classification :-

(linear additive model)

$$y_{ij} = \mu + \alpha_i + E_{ij}$$

$$y_{ij} = \mu + \alpha_i + \beta_j + E_{ij}$$

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_k + E_{ij}$$

Incase of one-way classification, the linear mathematical model is given by $\boxed{y_{ij} = \mu + \alpha_i + E_{ij}}$ $i = 1, 2, \cdots k$

$j = 1, 2, \cdots n_i$

where y_{ij} = The yield from the j^{th} unit by applying i^{th} treatment.

α_i = The i^{th} treatment effect.

E_{ij} = Error effect due to random.

μ = General mean effect

Estimation of Parameters :-

$$y_{ij} = \mu + \alpha_i + \epsilon_{ij}$$

$$\epsilon_{ij} = y_{ij} - \mu - \alpha_i$$

The parameters μ, α_i are estimated by using the principle of least squares an minimizing the error sum of squares

$$E = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \epsilon_{ij}^2$$

$$E = \sum_{i=1}^{k} \sum_{j=1}^{n} (y_{ij} - \mu - \alpha_i)^2$$

If minimize E w.r.to μ

$$\frac{dE}{d\mu} = 0$$

we get $\mu = \bar{y}_{..}$

If minimize E w.r.t α_i

$$\frac{dE}{d\alpha_i} = 0$$

we get $\alpha_i = \bar{y}_{i.} - \bar{y}_{..}$

Various sum of squares :-

In one way classification the total variation can be split into two parts.

i) frequent sum of squares

ii) Error sum of squares

The total sum of squares for corrected mean is

$$\sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{..})^2$$

$$\sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{..})^2 = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i.} + \bar{y}_{i.} - \bar{y}_{..})^2$$

$$\sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{..})^2 = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left[(y_{ij} - \bar{y}_{..}) + (\bar{y}_{i.} - \bar{y}_{..})\right]^2$$

$$\sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{..})^2 = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i.})^2 + \sum_{i=1}^{k} \sum_{j=1}^{n_i} (\bar{y}_{i.} - \bar{y}_{..})^2 + 2 \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i.})(\bar{y}_{i.} - \bar{y}_{..})$$

$$\sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{..})^2 = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i.})^2 + \sum_{i=1}^{k} n_i (\bar{y}_i - \bar{y}_{..})^2 + 0$$

since cross products vanish.

$$\sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{..})^2 = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i.})^2 + n_i \sum_{i=1}^{k} (\bar{y}_i - \bar{y}_{..})^2$$

Total sum of squares = Error of squares + Treatment sum of squares.

$$S_T^2 = S_E^2 + S_t^2$$

In numerical data for simple Arithmatic Calculations we make the use of following formulae.

i) Correction factor (CF) = $\frac{G^2}{N}$

Where G = Grand total

N = Total No. of observations.

ii) Total sum of squares $S_T^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij}^2 - CF$

iii) Frequent sum of squares $S_t^2 = \frac{\sum T_i^2}{n} - CF$ [equal size]

[or]

$S_t^2 = \sum_{i=1}^{k}\left(\frac{T_{i.}^2}{n_i}\right) - CF$ [unequal size]

iv) Error sum of squares $S_E^2 = S_T^2 - S_t^2$.

Degrees of freedom :-

i) The dof for total sum of squares = N-1

ii) The dof for treatment sum of squares = k-1

iii) The dof for Error sum of squares = N-1-(k-1)

= N-k

Mean sum of squares :- (MSS)

The sum of squares divided by its dof gives the Mean sum of squares. It is also called varience,

i) Mean sum of squares due to Treatments = $s_t^2 = \frac{S_t^2}{k-1}$

(or)

$Q_1 = \frac{S_t^2}{k-1}$

ii) Mean sum of squares due to Error = $s_E^2 = \frac{S_E^2}{N-k}$

[or]

$Q_2 = \frac{S_E^2}{N-k}$

Test statistic :-

Hence the test statistic for H_0 is

$F_{cal} = \frac{\text{MSS for treatment}}{\text{MSS for Error}}$ $\left(\begin{matrix} k-1, N-k \\ \text{at } 5\% \text{ or} \\ 1\% \text{ LOS}\end{matrix}\right)$

$= \frac{s_t^2}{s_E^2} = \frac{Q_1}{Q_2}$ ~ F

ANOVA Table :-

... SS	MSS	F-ratio F_{cal}	Ftable
		$F_{cal} = \frac{s_t^2}{s_E^2}$	$F_U(k-1, N-k)$ at 5% or 1%

Conclusion :-

If $f_{cal} \leq f_{table}$ for $(k-1, N-k)$ dof at required Probability level (5% or 1%) then we accept our H_0 otherwise we reject H_0.

③ Explain Estimation of parameters in one way classification.

A:- * The parameters in the model μ, α_i are estimated by using the principle of least squares. on minimizing the Error sum of squares.

* In one-way classification the linear mathematical model is

$$y_{ij} = \mu + \alpha_i + E_{ij} \qquad \begin{array}{l} i = 1, 2, \ldots k \\ j = 1, 2, \ldots n_i \end{array}$$

* Random Error $= E_{ij} = y_{ij} - \mu - \alpha_i$

* Error sum of squares $= E = \sum_{i=1}^{k} \sum_{j=1}^{n_i} E_{ij}^2$

$$= \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \mu - \alpha_i)^2$$

We minimize E w.r.to μ

$$\frac{\partial E}{\partial \mu} = 0$$

$$\frac{\partial \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \mu - \alpha_i)^2}{\partial \mu} = 0$$

$$2 \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \mu - \alpha_i)(-1) = 0$$

$$\sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \mu - \alpha_i) = 0$$

$$\sum_{i=1}^{k} \sum_{j=1}^{n_i} y_{ij} - \sum_{i=1}^{k} \sum_{j=1}^{n_i} \mu - \sum_{i=1}^{k} \sum_{j=1}^{n_i} \alpha_i = 0$$

$$\sum\sum y_{ij} - kn_i\mu - n_i \sum_{i=1}^{k} \alpha_i = 0$$

Since $\sum_{i=1}^{k} \alpha_i = 0$ $\qquad \begin{bmatrix} nk = N \\ n_i k = N \end{bmatrix}$

$$\sum\sum y_{ij} - N\mu = 0$$

$$N\mu = \sum\sum y_{ij}$$

$$\mu = \frac{\sum\sum y_{ij}}{N} = \bar{y}_{ij}$$

$$\boxed{\mu = \bar{y}_{..}}$$

We minimize E w.r.to α_i

$$\frac{\partial E}{\partial \alpha_i} = 0$$

$$\frac{\partial}{\partial \alpha_i} \sum_{j=1}^{n_i} (y_{ij} - \mu - \alpha_i)(-1) = 0$$

$$2\sum_{j=1}^{n_i}(y_{ij}-\mu-\alpha_i)(-1)=0$$

$$\sum_{j=1}^{n_i}(y_{ij}-\mu-\alpha_i)=0$$

$$\sum_{j=1}^{n_i} y_{ij}-\sum_{j=1}^{n_i}\mu-\sum_{j=1}^{n_i}\alpha_i=0$$

$$\sum_{j=1}^{n_i} y_{ij}-n_i\mu-n_i\alpha_i=0$$

$$n_i\alpha_i=\sum_{j=1}^{n_i} y_{ij}-n_i\mu$$

$$\alpha_i=\frac{\sum_{j=1}^{n_i} y_{ij}}{n_j}-\frac{\cancel{n_i}\mu}{\cancel{n_i}}$$

$$\alpha_i=\bar{y}_{i.}-\mu$$

$$\boxed{\alpha_i=\bar{y}_{i.}-\bar{y}_{..}}$$

DISCUSS TWO-WAY classification of data and also write down ANOVA table.

4. Explain statical analysis of two-way classification

A. Two-way Classification :-

In some cases the resulting observation are affected by two factors and we need to Examine the effect of this factors on the observations. Then stastical data is divided into on the basis of two factors. This type of classification is known as Two-way classification.

Ex:- 1) A sample of n individuals can be classified according to their height and weight & according to age and weight.

2) Let us take the yield of Paddy may be affected by differences in seeds and differences in different levels of fertilizers.

Let us known N observations are divided with respect to two Characteristics say A, B.

First we divide the observations in to k classes w.r.to characteristic A as rows and divide the observations into classes w.r.to characteristic B as columns.

$$\boxed{\text{i.e., } N=hk}$$

Here y_{ij} denote the yield from experimental unit of i^{th} and j^{th} column. In this case of observations are classified as follows.

Rows ↓ \ Columns →	Characteristic B II classification 1 2 ... j ... h	Total	Mean
Characteristic A I classification 1	y_{11} y_{12} ... y_{1j} ... y_{1h}	T_1	$\bar{y}_1$
2	y_{21} y_{22} ... y_{2j} ... y_{2h}	T_2	$\bar{y}_2$
i	y_{i1} y_{i2} ... y_{ij} ... y_{ih}	T_i	$\bar{y}_i$
k	y_{k1} y_{k2} ... y_{kj} ... y_{kh}	T_k	$\bar{y}_k$
Total	T_1 $T_{\cdot 2}$ ... $T_{\cdot j}$ ... $T_{\cdot h}$	G	—
Mean	$\bar{y}_{\cdot 1}$ $\bar{y}_{\cdot 2}$ ... $\bar{y}_{\cdot j}$ ... $\bar{y}_{\cdot h}$	—	—

Here the total variation is splitted into three parts.

* Total variation = variation b/w the rows + variation b/w the columns + Random variation.

→ The variation b/w the rows commonly known as Rows Effect.

→ The variation b/w the columns commonly known as Column Effect.

→ The variation with in classes commonly known as Error effect.

The main objective of ANOVA to test the homogenity of several sample means.

Assumptions :-

* All the observations y_{ij} are independent.
* The parent population from which the sample are taken is Normal.
* Varience Effects are additive in nature.
* E_{ij} is a random error

$E_{ij} \sim N(\mu, \sigma_e^2)$

Statistical Analysis of two-way classification :-

y_{ij} is the yield of i^{th} row in j^{th} column, $i=1,2,\ldots k$

$$\bar{y}_{i.} = \frac{\sum_{j=1} \quad \forall\ i = 1,2,}{h}$$

$$\bar{y}_{.j} = \frac{\sum_{i=1}^{k} y_{ij}}{k} \quad \forall\ j = 1,2,\ldots h$$

$$\bar{y}_{..} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{h} y_{ij}}{N} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{h} y_{ij}}{hk} = \frac{\sum_{i=1}^{k} \bar{y}_{i.}}{k} = \frac{\sum_{j=1}^{h} \bar{y}_{.j}}{h}$$

$N = hk$
k = No. of Rows
h = No. of colums.

Null Hypothesis :-

H_{01} : All the rows are homogeneous.
i.e., $H_{01} : R_1 = R_2 = \cdots = R_k$ $\left(\sum_{i=1}^{k} \alpha_i = 0\right)$

H_{02} : All the columns are homogeneous
i.e., $H_{02} : C_1 = C_2 = \cdots = C_h$ $\left(\sum_{j=1}^{h} \beta_j = 0\right)$

Mathematical model of two-way classification :-

The linear mathematical model of the observations for two-way classification is

$$y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij} \qquad i = 1,2,\ldots k \quad j = 1,2,\ldots h$$

y_{ij} = yield from the i^{th} row and j^{th} column.
μ = General mean effect.
α_i = i^{th} row Effect
β_j = j^{th} column effect.

ϵ_{ij} = Random effect. (Error effect due to random)

This equation is the basis equation for the calculation of VSS.

Estimation of ...

Sum of squares (ESS)

$y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij}$

$\epsilon_{ij} = y_{ij} - \mu - \alpha_i - \beta_j$

Error sum of squares $= E = \sum_{i=1}^{k} \sum_{j=1}^{h} \epsilon_{ij}^2$

we minimize E w.r.to μ

$$\frac{dE}{d\mu} = 0$$

we get, $\mu = \bar{y}_{..}$

we minimize E w.r.t α_i

$$\frac{dE}{d\alpha_i} = 0$$

we get $\alpha_i = \bar{y}_{i.} - \bar{y}_{..}$

we minimize E w.r.to β_j

$$\frac{dE}{d\beta_j} = 0$$

we get $\beta_j = \bar{y}_{.j} - \bar{y}_{..}$

Various sum of squares :- (VSS)

In two-way classification the total sum of squares can be split into three components

i) variation due to Rows

ii) variation due to columns

iii) variation due to Error

$\left[(a+b+c)^2 = a^2 + b^2 + c^2 + \text{cross product}\right]$

The total sum of squares for corrected mean.

$$\sum_{i=1}^{k} \sum_{j=1}^{h} (y_{ij} - \bar{y}_{..})^2 = \sum_{i=1}^{k} \sum_{j=1}^{h} \left[(y_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{..}) + (\bar{y}_{i.} - \bar{y}_{..}) + (\bar{y}_{.j} - \bar{y})\right]^2$$

$$\sum_{i=1}^{k} \sum_{j=1}^{h} (y_{ij} - \bar{y}_{..})^2 = \sum_{i=1}^{k} \sum_{j=1}^{h} (y_{ij} - \bar{y}_{.j} + \bar{y}_{..})^2 + \sum_{i=1}^{k} \sum_{j=1}^{h} (\bar{y}_{i.} - \bar{y}_{..})^2 + \sum_{i=1}^{k} \sum_{j=1}^{h} (\bar{y}_{ij} - \bar{y}_{..})^2 + \text{cross product.}$$

Since the cross product vanishes.

The algebraic sum of the deviation of observation

UNIT IV Design of Experiments

1. Describe randomised block design (RBD & CBD). Give its layout, advantages and Disadvantages.

RBD :- A method dividing the heterogeneous Experimental material into relatively homogeneous subgroups or blocks and the treatments are applied randomly to relatively homogeneous Experimental units with in Each block and replicated over all the blocks is known as Randomised Block Design (R.B.D)

for Exmaple. In agricultural Experimentation, the Experimental area.

ie field is not homogeneous and the fertility gradient is only one direction, then a simple method of controlling the variability of the Experimental material consists dividing the hetrogeneous field into homogeneous subgroups or blocks & replicates perpendicular to the direction of the fertility gradient.

Now if the treatments applied at random to homogeneous units within Each block and replicated over all the blocks, the design is called Randomised block design. In a CRD we the Experimental site and do not resort to the grouping of the Experimental units. allocate the treatments at random to the

But in RBD treatments are allocated at random within the units of Each block. i.e Randomization restricted.

Also variation among blocks is removed from variation due to Error.

Layout of RBD :-

In agricultural Experiment, if we consider 5 treatments A, B, C, D, E Each replicated four times. Then we divide the whole Experimental area is divided into four homogeneous blocks and Each block into five units.

Treatements (fertilizers) are then allocated at random the units of a block. fresh randomization being done for each block. The layout of RBD as follows.   Blocks   Treatments

To allocate treatments randomly, we any one of the methods, lottery method & random number tables method.

for randomization, we may use Tippetts random number tables.

Advantages (merits) :-

1. In this design three principles are used so it may be considered as a good design of Experiment.
2. The statistical analysis of RBD data is simple and Easy to understand.
3. RBD provides more accurate results that CRD since the experimental material is divided into blocks this results in decreasing Error variance. Therefore experimental Error is considerably reduced.
4. In RBD, no restrictions are placed on the number of treatments and on the number of replications. so the design is flexible. But atleast two replications are required to test the significance to treatments of Blocks.
5. It is more accurate (efficient) design than C.R.D
6. RBD is a very popular ~~Experiments~~ and Extensively used design in almost all the Scientific Experiments.

Disadvantages :-

1. RBD can not be used with unequal replications.
2. The randomization in RBD is restricted with in Each block.
3. It is suitable only if their is one favour of hetero-genity. If there is more than one factor of heterogenity, It is not suitable
4. RBD is not suitable for a large no. of treatments.
5. the analysis of RBD should be difficult in case of missing observation.

Statistical analysis of RBD :-

The statistical analysis of RBD is similar to the ANOVA for two way classified data.

Suppose a RBD Experiment is carried out with k

The ... can be arranged ...

Treatments	Blocks 1	2	...	j	...	h	Total	mean
1	y_{11}	y_{12}	...	y_{1j}	...	y_{1h}	T_1	$\bar{y}_1$
2	y_{21}	y_{22}	...	y_{2j}	...	y_{2h}	T_2	$\bar{y}_2$
⋮	⋮						⋮	⋮
i	y_{i1}	y_{i2}	...	y_{ij}	...	y_{ih}	T_i	$\bar{y}_i$
⋮	⋮						⋮	⋮
k	y_{k1}	y_{k2}	...	y_{kj}	...	y_{kh}	T_k	$\bar{y}_k$
Total	T_1	T_2	...	T_j	...	T_h	G	–
Mean	$\bar{y}_{.1}$	$\bar{y}_{.2}$	...	$\bar{y}_{.j}$	...	$\bar{y}_{.h}$	–	–

Total variation is splitted into three parts

$\therefore$ Total variation = variation in between Rows (Treatments effect) + variation in between columns (Block effect) + Error effect.

$N = hk$ = is the total number of experimental units

k = no. of treatments.

h = no. of blocks

$G = \sum_{i=1}^{k}\sum_{j=1}^{h} y_{ij}$ is grand Total $\quad T_{i.} = \sum_{j=1}^{h} y_{ij} \quad T_{.j} = \sum_{i=1}^{k} y_{ij}$

$\bar{y}_{i.} = \frac{1}{h}\sum_{j=1}^{h} y_{ij} \quad \bar{y}_{.j} = \frac{1}{k}\sum_{i=1}^{k} y_{ij} \quad \bar{y}_{..} = \frac{1}{hk}\sum_{i=1}^{k}\sum_{j=1}^{h} y_{ij}$

Null hypothesis :—

H_{01} : All the treatments are homogeneous

i.e H_{01} : $\alpha_1 = \alpha_2 = \cdots = \alpha_k = 0$

H_{02} : All the blocks are homogeneous

i.e H_{02} : $\beta_1 = \beta_2 = \cdots = \beta_h = 0$

Mathematical Model :— the linear mathematical model

becomes $y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij} \quad i = 1, 2, \cdots k \quad j = 1, 2, \cdots h$

where y_{ij} is the yield from j^{th} block by receiving i^{th} treatment. μ is general mean effect.

α_i is effect due to i^{th} treatment

β_j is effect due to j^{th} block

ϵ_{ij} is error effect due to random and $\epsilon_{ij} \sim iid\ N(0, \sigma_e^2)$

Estimator of Parameters :—

The parameters in the model μ, α_i, β_j are estimated by using the principle of least squares on minimizing the error sum of squares.

$\mu = \bar{y}_{..}$ $\alpha_i = \bar{y}_{i.} - \bar{y}_{..}$ $\beta_j = \bar{y}_{.j} - \bar{y}_{..}$

Various sum of squares :— The total sum of squares (Total variation) for the corrected mean.

$$\sum_{i=1}^{k}\sum_{j=1}^{h}(y_{ij}-\bar{y}_{..})^2 = \sum_{i=1}^{k}\sum_{j=1}^{h}\left((y_{ij}-\bar{y}_{i.}-\bar{y}_{.j}+\bar{y}_{..})+(\bar{y}_{i.}-\bar{y}_{..})+(\bar{y}_{ij}-\bar{y}_{..})\right)^2$$

$$\sum_{i=1}^{k}\sum_{j=1}^{h}(y_{ij}-\bar{y}_{..})^2 = \sum_{i=1}^{k}\sum_{j=1}^{h}(y_{ij}-\bar{y}_{i.}-\bar{y}_{.j}+\bar{y}_{..})^2 + \sum_{i=1}^{k}\sum_{j=1}^{h}(\bar{y}_{i.}-\bar{y}_{..})^2 + \sum_{i=1}^{k}\sum_{j=1}^{h}(\bar{y}_{.j}-\bar{y}_{..})^2 + \text{cross product}$$

The cross product in the above eqn vanish

$$\sum_{i=1}^{k}\sum_{j=1}^{h}(y_{ij}-\bar{y}_{..})^2 = \sum_{i=1}^{k}\sum_{j=1}^{h}(y_{ij}-\bar{y}_{i.}-\bar{y}_{ij}+\bar{y}_{..})^2 + h\sum_{i=1}^{k}(\bar{y}_{i.}-\bar{y}_{..})^2 + k\sum_{j=1}^{h}(\bar{y}_{.j}-\bar{y}_{..})^2 + 0$$

$TSS = ESS + \text{Treatment SS} + \text{Block SS}$

$S_T^2 = S_E^2 + S_t^2 + S_B^2$

In numerical data for simple arithmetic calculations, we make the use of following formulae.

$TSS = S_T^2 = \sum\sum y_{ij}^2 - CF$ $\quad CF = G^2/N$

$\text{Treatment SS} = S_t^2 = \dfrac{\sum T_i^2}{h} - CF$

$\text{Block SS} = S_B^2 = \dfrac{\sum T_{.j}^2}{k} - CF$

$\text{Error SS} = S_E^2 = S_T^2 - S_t^2 - S_B^2$

A NOVA TABLE FOR RBD

SV	Dof	SS	mss	f ratio variance ratio
Treatments	$k-1$	S_t^2	$s_t^2 = \dfrac{S_t^2}{k-1}$	$f_1 = \dfrac{s_t^2}{s_E^2} \sim F(k,(k-1)(h-1))$
Blocks	$h-1$	S_B^2	$s_B^2 = \dfrac{S_B^2}{h-1}$	$f_2 = \dfrac{s_B^2}{s_E^2} \sim F(h-1,(k-1)(h-1))$ at 5% & 1% LOS
Error	$(k-1)(h-1)$	S_E^2	$s_E^2 = \dfrac{S_E^2}{(k-1)(h-1)}$	—
Total	$hk-1 = N-1$	S_T^2	—	—

Condusion :— for treatments, Blocks.

If $f_{cal} \leq f_{table}$ value we accept our hypothesis otherwise we reject our hypothesis at required probability level (5% & 1%)

Problem: Analyse the RBD layout

sol:— NH: H_{01}: All the treatments (A, B, C, D, E) are homogene-ous H_{02}: All the blocks (I, II, III, IV) are homogenuous

LOS: $\alpha = 5\%$.

Treatments	Blocks I	II	III	IV	$T_{i.}$	$T_{i.}^2$	Σy_{ij}^2
A	12.5	12.4	14.2	11.6	50.7	2570.49	646.21
B	10.5	13.6	12.6	15.2	51.9	2693.61	685.01
C	11.2	12.5	13.7	12.3	49.7	2470.09	620.67
D	13.3	10.3	11.6	14.3	49.5	2450.25	622.03
E	11.1	14.1	10.4	12.3	47.9	2294.41	581.47
T_{ij}	58.6	62.9	62.5	65.7	249.7	12478.85	3155.39
$T_{.j}^2$	3433.96	3956.41	3906.25	4316.49 →	15613.11		—

Here k = no. of Treatments $= 5$ $\quad h$ = no. of blocks $= 4$

$N = hk = 5 \times 4 = 20 \quad G = 249.7 \quad \Sigma\Sigma y_{ij}^2 = 3155.39$

$\Sigma T_{i.}^2 = 12478.85 \quad \Sigma T_{.j}^2 = 15613.11$

$C.F = \frac{G^2}{N} = \left(\frac{(249.7)^2}{20}\right) = 3117.5045$

$TSS = S_T^2 = \Sigma\Sigma y_{ij}^2 - CF = 3155.39 - 3117.5045$
$= 37.8855$

Treatment SS $= S_t^2 = \frac{\Sigma T_{i.}^2}{h} - CF = \frac{12478.85}{4} - 3117.5045 = 2.208$

Block SS $= S_B^2 = \frac{\Sigma T_{.j}^2}{k} - CF = \frac{15613.11}{5} - 3117.5045 = 5.1175$

Error SS $= S_E^2 = S_T^2 - S_t^2 - S_B^2 = 37.8855 - 2.208 - 5.1175 = 30.5$

ANOVA Table :—

SV	dof	S.S	M.S.S	f-ratio fcal	ftable

Conclusion :—

Treatments : $f_1$cal $= 0.2167$ $\quad f_1$table $= 3.26$

$f_1$cal < $f_1$table value so, we accept our H_{01} at 5% LOS

All the treatments (A, B, C, D, E) are homogeneous.

Blocks : $f_2$cal $= 0.6698$ $\quad f_2$table value $= 3.49$

$f_2$cal < f_2 table value so, we accept our H_{02} at 5% LOS.

We conclude that All the blocks are homogeneou

I B(120) F(135) A(111) C(121) D(127)

I C(100) A(121) D(128) E(136) B(123)

II C(121) A(137) E(101) B(99) D(138)

$G = 1818 \quad \Sigma\Sigma y_{ij}^2 = 222822 \quad \Sigma T_{i.}^2 = 662922 \quad \Sigma T_{.j}^2 = 1101876$

$k = 5 \quad h = 3 \quad N = 15 \quad CF = 220341.6 \quad S_T^2 = 2480.4$

$S_t^2 = 632.4 \quad S_B^2 = 33.6 \quad S_E^2 = 1814.4 \quad F_{1cal} = 0.6990\ (3.84)$

$F_{2cal} = 0.07404\ (4.46)$

(8) In an agricultural field experimentation, three fertilizers are applied in four randomised blocks and the yield of wheat in kilos are given below.

Blocks

I	II	III	IV
A 8	C 10	A 6	B 10
C 12	B 8	B 9	A 8
B 10	A 8	C 10	C 9

1. Analyse the data using RBD and state conclusion

$F_1 = 7.8\,(5.14) \quad F_2 = 1.61\,(4.76)$

2. Find efficiency of RBD over CRD

$$E = \frac{h(k-1)S_E^2 + (h-1)S_E^2}{(hk-1)S_E^2} = 1.17$$

Latin Square Design (LSD)

(3) Describe LSD, Layout, mention its advantages and disadvantages.

Ans:- Latin Square Design (LSD)

The RBD attempts to control the variability in experimental material in one direction only. But there may be situations that hetrogenity may be in two perpendicular directions.

ie horizontally (⟺) as well as vertically (‖)

For example in agricultural experiments soil variation may be in two perpendicular directions. In such situations RBD may be of little use.

A layout known as LSD is of great use to control the variation due to two factors.

The entire heterogeneous experimental material is ... relatively homogeneous block with respect ... rows and columns, treatmen...

rows and columns in such a way that every treatment occurs once and only once in each row and in each column, such a design is called Latin square design. This helps in elimination of rows and columns effects from the experimental error and tries to make the experiment more sensitive. LSD is extensively used in agricultural experiments. Also it is used in industrial and animal husband experiments.

Layout of LSD :-

In LSD, the number of treatements and number of rows is equal to number of columns. If we consider m treatments, then there will be $m \times m = m^2$ experimental units. The whole experimental material is divided into m^2 experimental units arranged in a square so that each row and each column consists m experimental units. Then m treatments are allocated at random to these rows and columns in such a way that each and every treatment occurs once and only once in each row and in each column. This layout is called $m \times m$ LSD.

for example, if there LSD are five treatments, then 5X5 LSD layout can be explained in the following table.

A	B	C	D	E
B	C	D	E	A
C	D	E	A	B
D	E	A	B	C
E	A	B	C	D

Advantages of LSD :-

1) In field experiments, if the fertility gradient is in two directions then LSD is more efficient than RBD.
2) It is based on three principles of experimentation.
3) In LSD more number of factors are compared using less number of experimental units.
4) The statistical analysis of L.S.D is slightly complicated than that of RBD, but it is more efficient than RBD.
5) Local control is done with two types of grouping. Therefore experimental error will be much less in LSD compared to CRD and RBD. 6) LSD is a 3 way lay out.

DISADVANTAGES :-

... number of replication is equal to the number of treatments ... number ... therefore, if the ... due to large number of replication.

2) LSD is suitable for the number of treatments between 5 and 10. The design is not suitable and impracticable for more than 10 treatments.

3) If several units are missing in LSD, the statistical analysis is more difficult.

4) Randomization is also restricted with in each row and each column in this design

5) The fundamental assumption that there is no interaction between different factors may not be true in general.

4) Explain the statistical analysis of LSD data.

Suppose we have to compare m treatments for their effects, using a L.S.D. The m^2 experimental units needed for the LSD experiment are divided into m rows and m columns. The m treatments are allocated to the units of rows and columns in such a way that every treatment occurs only once in row & column.

From m^2 units we get m^2 yield.

NUll hypothesis :-

H_{01} : All the Rows and homogeneous ie

$H_{01} : L_1 = L_2 = \cdots = L_m = 0$

H_{02} : All the columns are homogeneous

i.e $H_{02} : \beta_1 = \beta_2 = \cdots \beta_m = 0$

H_{03} : All the treatments are homogeneous

i.e $H_{03} : V_1 = V_2 = \cdots V_m = 0$.

Mathematical Model :-

Let us suppose that y_{ijk} $(i,j,k = 1,2,\ldots m)$ be the yield from the experimental unit in the i^{th} row, j^{th} column and by receiving the k^{th} treatment.

The triplet (i,j,k) assumes m^2 experimental units.

The triplet (i,j,k) assumes m^3 values.

n LSD of the possible m^3 values by usually let us denote the set of m^2 values

we write $(i,j,k) \in S$

The linear mathematical model of LSD becomes

$$y_{ijk} = \mu + \alpha_i + \beta_j + V_k + \epsilon_{ijk} - (i,j,k) \in S$$

where μ = general mean effect

V_K — ... to the k^{th} treatment

ϵ_{ijk} = Error effect due to chance and

$\epsilon_{ijk} \sim$ iid $N(0, \sigma_e^2)$

consider $G = y_{...}$ = Total of all the m^2 observations

$R_i = y_{i..}$ = Total of m observations in the i^{th} row

$C_j = y_{.j.}$ = Total of m observations in the j^{th} column

$T_k = y_{..k}$ = Total of m observation from the k^{th} treatment.

Estimation of parameters :—

The parameters $\mu, \alpha_i, \beta_j, V_K$ are estimated by the principle of least squares. The least squares estimates of the parameters are $\mu = \frac{G}{N} = \frac{y_{...}}{m^2} = \bar{y}_{...}$

$\alpha_i = \bar{y}_{i..} - \bar{y}_{...} \quad i = 1, 2, --- m$

$\beta_j = \bar{y}_{.j.} - \bar{y}_{...} \quad j = 1, 2, --- m$

$V_K = \bar{y}_{..k} - \bar{y}_{...} \quad k = 1, 2, --- m$

$\bar{y}_{i..} = \frac{1}{m} \sum_{(j,k)} y_{ijk} \quad \bar{y}_{.j.} = \frac{1}{m} \sum_{(i,k)} y_{ijk} \quad \bar{y}_{..k} = \frac{1}{m} \sum_{(i,j)} y_{ijk}$

$\bar{y}_{...} = \frac{1}{m^2} \sum_{(i,j,k) \in S} y_{ijk}$

Various sum of squares :—

The total variation can be split into 4 parts total variation = variation b/w the rows + variation b/w the columns + variation b/w the treatments + Error variation.

Total sum of squares for corrected mean

$$\sum_{(i,j,k) \in S} (y_{ijk} - \bar{y}_{...})^2 = \sum_{(i,j,k) \in S} [(y_{ijk} - \bar{y}_{i..} - \bar{y}_{.j.} - \bar{y}_{..k} + 2\bar{y}_{...}) + (\bar{y}_{i..} - \bar{y}_{...}) + (\bar{y}_{.j.} - \bar{y}_{...}) + (\bar{y}_{..k} - \bar{y}_{...})]^2$$

$$\sum_{(i,j,k) \in S} (y_{ijk} - \bar{y}_{...})^2 = \sum_{(i,j,k) \in S} (y_{ijk} - \bar{y}_{i..} - \bar{y}_{.j.} - \bar{y}_{..k} + 2\bar{y}_{...})^2 + m \sum (\bar{y}_{i..} - \bar{y}_{...})^2 + m \sum_j (\bar{y}_{.j.} - \bar{y}_{...})^2 + m \sum_k (\bar{y}_{..k} - \bar{y}_{...})^2 + \text{Cross Product}$$

The product terms vanish since the algebraic sum of deviations from the mean is zero.

$T_{SS} = E_{SS} + R_{SS} + C_{SS}$ + Treatments

$S_T^2 = S_E^2 + S_R^2 + S_C^2 + S_t^2$

for numerical calculations of various sum of squares the following formula can be used

$$mv_k = \sum_{(i,j)} y_{ijk} - m\mu$$

$$v_k = \sum_{(i,j)} \frac{y_{ijk}}{m} - \frac{m\mu}{m}$$

$$v_k = \bar{y}_{..k} - \mu$$

$$\boxed{\hat{v}_k = \bar{y}_{..k} - \bar{y}_{...}}$$

Problems ①: – Analyse the LSD layout

A(12) B(14) C(11) D(10)
B(12) C(14) D(13) A(11)
C(9) D(15) A(13) B(10)
D(11) A(16) B(14) C(10)

Sol NH: H_{01}: All the rows and homogeneous.

H_{02}: All the columns are homogeneous

H_{03}: All the treatments are homogeneous

LOS: $\alpha = 5\%$

Rows	Column 1	2	3	4	R_i	R_i^2	$\sum y_{ij}^2$
1	12	14	11	10	47	2209	561
2	12	14	13	11	50	2500	630
3	9	15	13	10	47	2209	575
4	11	16	14	10	51	2601	673
					195	9519	2439
C_j	44	59	51	41		9699	—
C_j^2	1936	3481	2601	1681	–		

$m = 4 \quad N = m^2 = 4 \times 4 = 16 \quad G = 195 \quad \sum y_{ij}^2 = 2439$

$\sum R_i^2 = 9519 \quad \sum C_j^2 = 9699.$

Treatments	$\sum A$	$\sum B$	$\sum C$	$\sum D$	Total
Total T_k	52	50	44	49	
T_k^2	2704	2500	1936	2401	9541

$\sum T_k^2 = 9541$

$$CF = \frac{G^2}{N} = \frac{(195)^2}{16} = 2376.5625$$

$$S_T^2 = \sum y_{ij}^2 - CF = 2439 - 2376.5625 = 62.4375$$

$$S_R^2 = \frac{\sum R_i^2}{m} - CF = \frac{9519}{4} - 2376.5625 = 3.1875$$

$$S_C^2 = \frac{\sum C_j^2}{m} - CF = \frac{9699}{4} - 2376.5625 = 48.1875$$

$$S_t^2 = \frac{\sum T_k^2}{m} - CF = \frac{9541}{4} - 2376.5625 = 8.6875$$

$$S_E^2 = S_T^2 - S_R^2 - S_C^2 - S_t^2 = 62.4375 - 3.1875 - 48.1875 - 8.6875$$

ANOVA TABLE :-

SV	dof	SS	MSS	f-ratio fcal	ftable
Rows	4-1=3	$S_R^2 = 3.1875$	$s_R^2 = \frac{3.1875}{3} = 1.0625$	$f_1 = \frac{s_R^2}{s_E^2} = 2.6844$	$f_1(3,6)$ at 5% $= 4.76$
Columns	4-1=3	$S_C^2 = 48.1875$	$s_C^2 = 16.0625$	$f_2 = \frac{s_C^2}{s_E^2} = 40.5823$	"
Treat	4-1=3	$S_t^2 = 8.6875$	$s_t^2 = 2.8958$	$f_3 = \frac{s_t^2}{s_E^2} = 7.3163$	"
Error	(4-1)(4-2) = 6	$S_E^2 = 2.375$	$s_E^2 = 0.3958$	—	—
Total	16-1=15	$S_T^2 = 62.4375$	—	—	—

Infarence :- Rows $f_1$cal $= 2.6844$ ftable $= 4.76$

$f_1$cal < ftable value so we accept our H_0

All the rows are homogeneous

columns $f_2$cal $= 40.5823$ $f_2$table $= 4.76$

$f_2$cal > $f_2$table value we reject H_{02} at 5% LOS

All the columns are not homogeneous.

Treatments :-

f_3 cal $= 7.3163$ f_3 table value $= 4.76$

f_3 cal > f_3 table value we reject H_{03} at 5% LOS

All the treatments are not homogeneous.

efficiency of LSD over RBD

(i) when Rows are taken as blocks $E = \frac{s_C^2 + (m-1)s_E^2}{m s_E^2}$

(ii) when Columns are taken as blocks $E = \frac{16.0625 + (4-1)0.3958}{4 \times 0.3958}$

$E = \frac{s_R^2 + (m-1)s_E^2}{m s_E^2}$ $= 1.7063$

Efficiency of LSD over CRD

$$E = \frac{s_R^2 + s_C^2 + (m-1)s_E^2}{(m+1)s_E^2}$$

Problem no (2) The table below gives the yield of rice in kilos observed in a field experiment carried out in 4 x 4

UNIT V

Factorial Experiments.

1) Explain factorial Experiments, merits and Demerits.

In the Experiment of CRD & RBD & LSD we were mainly congerned with the comparison of single set of treatments. Such experiments dealing with one factor only are called simple Experiments.

In factorial Experiments, the effects of several factors of variation are investigated. In these Experiments the treatments be all the combinations of different factors under study.

In factorial Experiments the interaction effects shall be studied and Estimated.

Here we study experiments involving several factors & Experiments involving combinations of different levels of different factors. These factorial Experiments were considered as Complex Experiments in olden days.

Def :- An Experiment in which the treatments are different combinations of different levels of several factors is called a factorial Experiments.

Example :- We have two factors N and P each with two levels (0,1) say. Then we have factorial Combi-nations as follows.

$$N \to n_0,\ n_1 \qquad P \to p_0,\ p_1$$

$N_iP_i \rightarrow$ factorial combination

i.e., treatments in factorial Experiment.

Symmetrical factorial Experiment :-

In General f factors, l-levels for each factor the factorial experiments is l^{th} and this l^f factorial experiments is called Symmetrical factorial

$f=n$ $l=2$ $\Rightarrow 2^n$ factorial experiments
$f=2$ $l=3$ $\Rightarrow 3^2$ ''
$f=3$ $l=3$ $\Rightarrow 3^3$ ''
$f=n$ $l=3$ $\Rightarrow 3^n$ ''

$2^2, 2^3 \ldots 2^n$ and $3^3, 3^3 \ldots 3^n$ are symmetric factorial experiments.

Merits and demerits of factorial experiments.

Merits (advantages) :-

(1) When compared with single factor experiments factorial experiments have more flexipility regarding the study of different levels of several factors. In the factorial experiments each factor is to be studied over a wide set of experiments.

(2) Interaction between the effects of various factor can be studied only in factorial experiments.

(3) Factorial experiments have the advantage of economising on experimental resources.

When experiments are conducted factor by factor much more, resources are required for the same precision than when they are tried in factorial experiments.

(4) In factorial experiments on allocating all the combinations once we will get one complete replication. Besides such of replications there is hidden replication.

(5) Factorial experiments are most servisable than standard designs in biological & clinical experiments.

. A single factorial is easy to conduct than elaborate (entent) many single factorial experiment at a time.

Demerits :-

(1) If n is large the experiment becomes too complex. for example if n=10 is 2^n factorial design will have 1024 tr...

different levels of factors.

Explain 2^2 design.

Here we have two factors each at two levels so that there are $2 \times 2 = 4$ treatment combinations in all.

Let the capital letters A and B indicates the names of two factors under study. Also let the small letters a and b denote the levels of corresponding factors. The first level of A and B is generally expressed in the absence of corresponding level in the treatment combinations.

The four treatment combinations can be enumerated as follows.

$a_0 b_0$ or 1 factors A and B both at first level

$a_1 b_0$ or a A at second level and B at first level

$a_0 b_1$ or b A at first level and B at second level.

$a_1 b_1$ or ab A and B both at second level.

These four treatment combinations can be compared laying out (putting)

i) CRD with 4 treatments

ii) RBD with r replicates (say) each replicate (block) containing 4 units

iii) LSD as a 4×4 layout.

ANOVA can be carried out accordingly.

In the above cases there are 3 d.f associated with the treatment effects.

In factorial experiments our main objective is to carry out separate tests for the main effects and the interaction AB.

We split the 3 d.f of treatment sum of squares into 3 orthogonal components each with

main effects and interaction effects of 2^2 design

... the factorial experiment with $2^2=4$

Let [1] [a] [b] and [ab] denote the ... of the ... in units.
[1] [a] [b] [ab] receiving the treatments
1, a, b and ab respectively.
Let the corresponding mean values obtained on dividing these totals by r be denoted by (1) (a) (b) and (ab) respectively. The letters A, B and AB when they refer to numbers will represent the main effects due to the factors A and B and their interaction AB respectively.
The effect of A can be represented by the difference between mean yields obtained at each level.
Thus the effect of the factor 'A' at the first level b_0 of B = $(a_1 b_0) - (a_0 b_0) = (a) - (1) \rightarrow (1)$
Similarly the effect of A at the second level b_1 of B = $(a_1 b_1) - (a_0 b_1) = (ab) - (b) \rightarrow (2)$
These two effects (1) and (2) are termed as simple effects of the factor A.
The average effect of A over the two levels of B is called the main effect of A over the two levels of B is called the main effect due to A and is defined by
The average observed effect of A over the two levels of B is called the main effect due to A and is defined by $A = \frac{1}{2}\left[(ab) - (b) + (a) - (1)\right]$

$$A = \frac{1}{2}(a-1)(b+1) \rightarrow I$$

where the right hand side to be expanded algebrically and then the treatment combination are to be replaced by treatment mean.
Similarly the effect of factor B at first level a_0 of A = $(a_0 b_1) - (a_0 b_0) = (b) - (1) \rightarrow (1)$
The effect of B at second level a_1 of A
$= (a_1 b_1) - (a_1 b_0) = (ab) - (a) \rightarrow (2)$
The average of (1) and (2) is known as the

main Effect of B = $\frac{1}{2}((b)-(1)+(ab)-(a))$
$= \frac{1}{2}(a+1)(b-1)$ — II

The right hand side to be Expanded algerbaically and the treatment combinations are to be replaced by their means.

Interaction Effect AB

If the two factors act independently of one another if means ① and ② are the estimates of the same thing.

If the two factors are not act independently the two expressions ① and ② will not be same.

Then the difference of these two numbers is the generation between the factors A and B.

the generaction effect between the factors A and B defined as

Interaction effect $AB = \frac{1}{2}[(ab)-(b)-(a)+(1)]$
$= \frac{1}{2}(a-1)(b-1)$ — III

Where the RHS is to be Expanded algebrically and then the treatment combination are to be replaced by the corresponding treatments.

from III we notice that the Interaction between B and A is as

Interaction effect $BA = \frac{1}{2}(b-1)(a-1)$ — III

which are same as the Expressions III and III a

It means the interaction does not depend on the order of the factors.

④ Explain statistical Analysis of 2^2 factorial design.

Factorial experiments are conducted either in CRD or RBD & LSD and thus they can be analysed in the usual manner except that in this case the treatment sum of squares is split into three orthogonal components each with 1 d.f.

Suppose a 2^2 factorial experiment is conducted by a Randomised block design (RBD) layout with r blocks each contains 4 plots.

The data from the layout may be tabulated as follows

Yields

Blocks	Treatments a_0b_0	a_1b_0	a_0b_1	a_1b_1
1	y_{11}	y_{21}	y_{31}	y_{41}
2	y_{12}	y_{22}	y_{32}	y_{42}
⋮	⋮	⋮	⋮	⋮
j	y_{1j}	y_{2j}	y_{3j}	y_{4j}
⋮	⋮	⋮	⋮	⋮
r	y_{1r}	y_{2r}	y_{3r}	y_{4r}.

The yield totals under the 4 treatments can be written as $[a_0b_0]$ $[a_1b_0]$ $[a_0b_1]$ $[a_1b_1]$ where [ ] denotes the sum of yields from r plots.

Null hypothesis :-

H_{01} : All the blocks (replicates) are homogeneous

H_{02} : The main effect A is not significant

(or)

H_{02} : there is no significant effect of main effect A on increasing the yield.

(or)

H_{02} : There is no significant difference due to ME A.

H_{03} : There is no significant difference due to ME B

H_{04} : the Interaction effect AB is not significant

In testing these hypothesis the analysis is Just like as in RBD except in the case of treatment Sum of squares.

D.o.f for total $= 4r-1$

sum of squares due to blocks $= S^2_{Block} = \frac{\Sigma B_J^2}{4} - CF$

Here B_J = J^{th} block total

D.o.f for blocks $= r-1$

sum of squares due to treatments is divided into three components sum of squares due to main effect A, sum of squares due to main effect B, sum of squares due to interaction Effect AB.

The dof for treatments are $4-1=3$ is distributed to the three components equally. To find sum of squares due to different effects we may use traditional method.

sum of squares due to main Effect $A = S_A^2 = \frac{[A]^2}{4r}$

$[A] = [ab] + [a] - [b] - [1]$

sum of square due to main Effect $B = S_B^2 = \frac{[B]^2}{4r}$

$[B] = [ab] - [a] + [b] - [1]$

Sum of squares due to Interaction Effect AB

$= S_{AB}^2 = \frac{[AB]^2}{4r}$ $[AB] = [ab] - [a] - [b] + [1]$

Here $[A]$ $[B]$ $[AB]$ are the different Effects totals

$\therefore$ sum of squares due to treatments

$= S_A^2 + S_B^2 + S_{AB}^2$

$\therefore$ Sum of squares due to Error $= TSS - SSB - S_A^2 - S_B^2 - S_{AB}^2$

Error dof $= 4r - 1 - r + 1 - 3 = 3(r-1)$

To test the above null hypothesis we construct the following ANOVA Table using the above calculations.

ANOVA table :— for 2^2 factorial experiment ...

Conclusion :-

for blocks, Main effect A, and Interaction effect AB.

If Fcal ≤ Ftable value we accept our hypothesis otherwise we reject our hypothesis at 5% & 1% LOS.

⑤ Yates method of computing factorial effect Totals Totals.

for the calculations of various effect Totals for 2^n factorial experiments yates developed a special rule which enables us to avoid specific alegebraic formulae.

this method is very useful if the number of factors is more. Yates method for 2^2 factorial experiment consists in the following steps..

1) In the first column we write the treatment combinations in the standard order. for Example 2^2 factorial experiment with 2 factor A and B. The order of treatment combination will be 1 a b ab for 2^3 experiment with 3 factors A, B, c the order is 1 a b ab c ac bc abc.

2) In the second column we write the corresponding treatment totals

[1] [a] [b] [ab]

3) the third column can be splitted into two halves. the first half is obtained by writing the pairewise sum of the values in column ② in the given order and second half is obtained by writing in the same order the pairwise differences of the values in the column ②. The difference should be obtained by subtracting the first value from the second value in the pair.

... column is constructed by adopting the ... the column ④ gives ... total ... effect of A, total effect of B, Total interaction Effect of AB.

Table for a 2^2 factorial experiments

① Treatment combination	② Total yield from all replicates	③	④ effect totals
1	[1]	[ab]+[1]	[ab]+[b]+[a]+[1]=G
a	[a]	[ab]+[b]	[ab]-[b]+[a]-[1]=[A]
b	[b]	[a]-[1]	[ab]+[b]-[a]-[1]=[B]
ab	[ab]	[ab]-[b]	[ab]-[b]-[a]+[1]=[AB]

Main effect of $A = \frac{[A]}{2r}$ $[A] = [ab]+[a]-[b]-[1]$

Main effect of $B = \frac{[B]}{2r}$ $[B] = [ab]+[b]-[a]-[1]$

Interaction effect of $AB = \frac{[AB]}{2r}$ $[AB] = [ab]-[a]-[b]+[1]$

Problems: In a 2^2 factorial experiment, the factors are A&B. The yield from 3 replications of the four factors are given below.

Replication I	1 25	b 27	ab 32	a 26
Replication II	a 25	ab 31	1 24	b 27
Replication III	b 28	1 23	a 26	ab 32

find the various effects of treatments using traditional method and Yates method.

Analyse the design.

Traditional method.

Replication	1	a	b	ab
I	25	26	27	32
II	24	25	27	31
III	23	26	28	32
	72	77	82	95

$[1] = 72$ $[a] = 77$ $[b] = 82$ $[ab] = 95$

Main effect due to $A = \frac{(a-1)(b+1)}{2r} = \frac{[ab]+[a]-[b]-[1]}{2r}$

$$= \frac{95+77-82-72}{2\times 3} = 3$$

Main effect due to B $= \frac{(a+1)(b-1)}{2r} = \frac{[ab]-[a]-[b]-[1]}{2r}$

$= 4.6667$

Interaction Effect AB $= \frac{(a-1)(b-1)}{2r} = \frac{[ab]-[a]-[b]+[1]}{2r}$

$= 1.3333$

Yates method

①	②	③	④	
1	72	149	326	= G
a	77	177	18	= [A]
b	82	5	28	= [B]
ab	95	13	8	= [AB]

Main effect due to A

$= \frac{[A]}{2r} = \frac{18}{6} = 3$

M E B $= \frac{[B]}{2r} = 4.6667$

I E AB $= \frac{[AB]}{2r} = \frac{8}{6} = 1.3333.$

ab, ac, bc & abc.

Here we have 7 different treatment effects namely main effect $\left.\begin{matrix}A\\B\\C\end{matrix}\right\} 3c$,

Two factor Interaction Effects $\left.\begin{matrix}AB\\AC\\BC\end{matrix}\right\} 3c_2$

Three factor interaction Effects ABC $3c_3$

Total $= 2^3 - 1 = 7$

2^3 factorial experiment can be performed as a CRD with 8 treatments (or) RBD with r replicates (say) each replicate containing 8 treatments or LSD with m=8 and data can be analysed according.

In 2^3 experiment ~~can be performed as a CRD with 8 treatments (or) RBD with r replicates (say) each replicate containing 8 treatmen~~ we split up the treatment S.S with 7 d.f into 7 orthogonal components corresponding to the three main effects A, B and C three first order (or two factor interaction (or three factor) ABC each charging d.f

Here A, B, C, AB, AC, BC, ABC etc when they refer to numbers will represent the corresponding factorial effects.

7) Explain main effects and Interaction effects in 2^3 factorial Experiments

Suppose the factorial experiment with $2^3 = 8$ treatment is conducted in r blocks (replicates) Let [1][a][b][c][ab][ac][bc][abc] denotes the total yields of the runits (plots) receiving

... the ... they refer to the ... A, B, C and ... due to factors A, B, C and ... BC and ABC respectively.

The simple effect of A are calculated ... manner.

level of B	level of C	Simple effects of A
b_0	c_0	$(a_1b_0c_0) - (a_0b_0c_0) = (a) - (1)$
b_0	c_1	$(a_1b_0c_1) - (a_0b_0c_1) = (ac) - (c)$
b_1	c_0	$(a_1b_1c_0) - (a_0b_1c_0) = (ab) - (b)$
b_1	c_1	$(a_1b_1c_1) - (a_0b_1c_1) = (abc) - (bc)$

The main effect of A is the average of the above 4 simple effects.

Main effects of A $= \frac{1}{4}[(a)-(1)+(ac)-(c)+(ab)-(b)+(abc)-(bc)]$

$= \frac{1}{4}[(abc)+(ac)+(ab)-(bc)+(a)-(b)-(c)-(1)]$

$= \frac{1}{4}\{(a-1)(b+1)(c+1)\} \rightarrow (1)$

ME of A $= \frac{1}{4}\{(a-1)(b+1)(c+1)\} \rightarrow (2)$

llly ME of B $= \frac{1}{4}\{(a+1)(b-1)(c+1)\} \rightarrow (3)$

ME of C $= \frac{1}{4}\{(a+1)(b+1)(c-1)\}$

By expanding the expressions (1), (2) and (3) algebrically and the treatment combinations are replaced by the corresponding treatment average yield then we get the main effect of A, Main effect of B & Main effect of C.

First order interaction effects :-

The average effect of A (at one level of C) at the level b_0 of B is $\frac{1}{2}\{(ac)-(c)+(a)-(1)\}$

The average effect of A (at one level of C) at the level b_1 of B is $\frac{1}{2}\{(abc)-(bc)+(ab)-(b)\}$

∴ The interaction effect of AB is given by half the difference b/w the average effect of A at the second & first level of B is given by.

Interaction effect of AB $= \frac{1}{4}\{(abc)-(bc)+(ab)-(b)-(ac)+(c)-(a)+(1)\}$

Interaction effect of AB $= \frac{(a-1)(b-1)(c+1)}{4} \rightarrow (4)$

Similarly we can obtain expression for the interaction BC and AC.

Interaction effect of BC $= \frac{(a+1)(b-1)(c-1)}{4} \rightarrow (5)$

Interaction effect of AC $= \frac{(a-1)(b+1)(c-1)}{4} \rightarrow (6)$

Second order interaction effect :-

We obtained the expression ...

... at the level c_0 at c is
$\{(ab) - (b) - (a) - 1(1)\}$
of AB at the level c_1 of c is
$\frac{1}{2}\{(a,b,c) - (bc) - (ac) + (1)\}$
effect of AB with , is
) Effect of $ABC = \frac{1}{4}\{(abc) - (bc) - (ac) + (c) - (ab) + (b) + (a) - (1)\}$

$c = \frac{1}{4}((a,b,c) - (ab) - (ac) - (bc) + (a) + (b) + (c))$,
(or)
$= \frac{(a-1)(b-1)(c-1)}{4} \rightarrow$ (7)

nding (4) (5) (6) & (7) algebrically and substitu. atment combinations by the correspond. mean yields then we get the interaction c BC & ABC.

alysis of 2^3 design
(or)
^{3} factorial experiment.

ial experiments are conducted either in .D. they can be analysed in the usual the treatment sum of squares is components each with 1 d.f
ctorial experiment be performed in RBD have 8 treatments i.e 1, a, b, ab, c, ac, plicated r times.
yield are tabulated as follows

ocks	----	j	----	r	total
y_{1}	----	y_{1j}	---	y_{1r}	[1]
y_{2}	---	y_{2j}	---	y_{2r}	[a]
y_{32}	---	y_{3j}	---	y_{3r}	[b]
y_{42}	---	y_{4j}	---	y_{4r}	[ab]
y_{52}	---	y_{5j}	---	y_{5r}	[c]
y_{62}	---	y_{6j}	---	y_{6r}	[ac]
y_{72}	---	y_{7j}	---	y_{7r}	[bc]
y_{82}	---	y_{8j}	---	y_{8r}	[abc]
B_2	---	B_j	---	B_r	G

... yield from j^{th} block receiving all

y_{ij} $i = 1, 2, \dots 8$
$j = 1, 2, \dots r$

$= \frac{G^2}{8r}$ $i = 1, 4, \dots 8$

Σy_{ij} $j = 1, 4, \dots r$

$= S_T^2 = \Sigma\Sigma y_{ij}^2 - CF$

to blocks $= S_B^2 = \frac{\Sigma B_j^2}{8} - CF$

th block total $j = 1, 2, \dots, r$

ares is divided into 7 parts and

es follows.

ce to main effect $A = \frac{[A]^2}{8r}$

$bc] + [a] - [b] - [c] - [1]$

ue to main effect $B = \frac{[B]^2}{8r}$

$+(bc) - (a) + (b) + (c) - (1)$

ue to main effect $C = \frac{[C]^2}{8r}$

$+(bc) - (a) - (b) + (c) - (1)$

e to interation effect $AB = \frac{[AB]^2}{8r}$

$bc) - (a) - (b) + (c) + (1)$

to Interaction effect $AC = \frac{[AC]^2}{8r}$

$bc) + (a) - (b) - (c) + (1)$

to interaction effect of $ABC = \frac{(ABC)^2}{8r}$

$-(bc) + (a) + (b) + (c) - (1)$.

of squares are also calculated easily

ent sum of squares is the sum of

$S_{AB}^2 + S_{AC}^2 + S_{BC}^2 + S_{ABC}^2$

e to Error

S_t^2

S_t^2

total $= 8r - 1$

Source of variation	d.f	SS	MSS	F cal	F table
Blocks	$r-1$	S^2_{Block}	$s^2_{Block} = \frac{S^2_{Block}}{r-1}$	$F_1 = \frac{s^2_{Block}}{s^2_E}$	$F_{[(r-1), 7(r-1)]}$
Treatments					
ME A	1	S^2_A	$s^2_A = S^2_A$	$F_2 = \frac{s^2_A}{s^2_E}$	$F_{(1, 7(r-1))}$
ME B	1	S^2_B	$s^2_B = S^2_B$	$F_3 = \frac{s^2_B}{s^2_E}$	"
ME C	1	S^2_C	$s^2_C = S^2_C$	$F_4 = \frac{s^2_C}{s^2_E}$	"
IE AB	1	S^2_{AB}	$s^2_{AB} = S^2_{AB}$	$F_5 = s^2_{AB}/s^2_E$	"
IE AC	1	S^2_{AC}	$s^2_{AC} = S^2_{AB}$	$F_6 = s^2_{AC}/s^2_E$	"
IE BC	1	S^2_{BC}	$s^2_{BC} = S^2_{BC}$	$F_7 = s^2_{AB}/s^2_E$	"
IE ABC	1	S^2_{ABC}	$s^2_{ABC} = S^2_{ABC}$	$F_8 = s^2_{ABC}/s^2_E$	"
Error	$7(r-1)$	S^2_E	$s^2_E = \frac{S^2_E}{7(r-1)}$	—	—
Total	$8r-1$	S^2_T	—	—	

Conclusion :- We compare the F-calculated with the corresponding F-table values at specified LOS and draw the conclusion according.

Yates method for a 2^3 factorial experiment.

For the calculation of various effect totals Yates developed a specific algebraic formula. This is very useful if the no. of factors is more we construct the yates as discribed below.

1) In the first column we write the treatment combination in the standard order i.e., 1, a, b, ab, c, ac, bc, abc.

2) In the second column we write the corresponding treatment totals from all the replications. i.e [1] [a] [b] [ab], [c] [ac], [bc], [abc]

3) The entries in the third column is splitted into two halves. The first half is obtained by writing the pairwise sums of column 2 in the given order. The second half is obtained by subtracting the first value from the second value of the pair.

4) The fourth column is constructed by adopting the same procedure on column 3

5) The fifth column is also constructed by adopting the same procedure as explained in step 3 on column 4. The columns gives the Grand total G Total effect of A i.e [A] total. Total effect interaction effect of Effect of B ie [B]. Total effect of C i.e (C) total interaction effect ie [AB] Total effect of ... Effect of BC ie [BC] and

(1)	[(1)]	[a]+[1] = x_1	... $x_2+x_1=y_1$	[a][1] = G ... (B)
a	[a]	[ab]+[b] = x_2	(abc)+(bc)+(ac)+(c) $x_4+x_3=y_2$	y_2+y_1 ... G (abc)+(bc)+(ac)+(c)+(ab)-(b)+(a)-(1)
b	[b]	[ac]+[c] = x_3	(abc)-(b)+(a)+(1) $x_6+x_5=y_3$	$y_4+y_3=(A)$ +(abc) $y_6+y_5=[B]$
ab	[ab]	(abc)+(bc) = x_4	(abc)-(bc)+(ac)-(c) $x_8+x_7=y_4$	$y_8+y_7=[AB]$
c	[c]	(a)-(1) = x_5	(ab)+(b)-(a)-(1) $x_2-x_1=y_5$	$y_2-y_1=(C)$
ac	[ac]	(ab)-(b) = x_6	(abc)+(bc)-(ac)-(c) $x_4-x_3=y_6$	$y_4-y_3=(AC)$
bc	[bc]	(ac)-(c) = x_7	(ab)-(b)-(a)+(1) $x_6-x_5=y_7$	$y_6-y_5=(BC)$
(abc)	(abc)	(abc)-(bc) = x_8	(abc)-(bc)-(ac)+(c) $x_8-x_7=y_8$	$y_8-y_7=(ABC)$

Main effects and interaction effect can be obtained by dividing the effect totals by 4r.

Critical difference :-

If there is a significant difference in between the treatments then we would be interested to find out which pair of treatments differ significant. For the instead of calculating students 't' for different pairs treatment means, we calculate the test significant difference of the given level of significance. The least difference is known as the critical difference (CD).

CD at α level of significance is given by

$$CD = SE(\bar{x}_1 - \bar{x}_2) \times t_{\alpha}\% \text{ for error d.f}$$

For Example :- $H_0 : \mu_1 = \mu_2$ i.e, two treatment means do not differ significantly. If H_0 is rejected. then H_0 is accepted i.e $H_1 : \mu_1 \neq \mu_2$ may be accepted. Now we have to determine which pair of treatment differs significantly. to obtain this we have to calculate a student. 't' statistic for every pair. So instead of this we compute C.D as follows

$$CD(\bar{x}_1 - \bar{x}_2) = CD = SE(\bar{x}_1 - \bar{x}_2) \times t_{\alpha}\% \text{ for error dof}$$

$$= \sigma_e \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \times \text{ for } \cancel{\text{example}} \text{ dof}$$

$E(s_e^2) = \sigma_e^2$ and if each treatment replicated n times.

... $n_i = n$, $i = 1, 2, \dots, k$ then $CD = s_e \sqrt{\frac{2}{n}} \times t_{\alpha}$ for error d.o.f