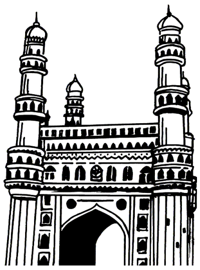
Rahul's ✔
Topper's Voice

AS PER CBCS SYLLABUS



B.Sc.

III Year VI Sem

Latest 2023 Edition

STATISTICS

PAPER-VI-A: APPLIED STATISTICS-II

- **Study Manual**
- **FAQ's & Important Questions**
- **Short Question & Answers**
- **Choose the Correct Answers**
- **Fill in the blanks**
- **Solved Model Papers**
- **Solved Previous Question Papers**

` .199/-

- by -

WELL EXPERIENCED LECTURER

Rahul Publications™

Hyderabad. Cell : 9391018098, 9505799122

B.Sc.

III Year VI Sem

STATISTICS

PAPER-VI-A: APPLIED STATISTICS-II

Price ` 199-00

Sole Distributors : **Cell : 9391018098, 9505799122**

VASU BOOK CENTRE

Shop No. 2, Beside Gokul Chat, Koti, Hyderabad.

Maternity Hospital Opp. Lane, Narayan Naik Complex, Koti, Hyderabad.

Near Andhra Bank, Subway, Sultan Bazar, Koti, Hyderabad -195.

STATISTICS

PAPER-VI-A: APPLIED STATISTICS-II

CONTENTS

STUDY MANUAL

SOLVED MODEL PAPERS

SOLVED PREVIOUS QUESTION PAPERS

SYLLABUS

UNIT - I

Analysis of Variance and Design of Experiments : Concept of Gauss-Markoff linear model with examples, statement of Cochran's theorem, ANOVA – one-way, two-way classifications with one observation per cell Expectation of various sums of squares, Statistical analysis, Importance and applications of design of experiments.

UNIT - II

Principles of experimentation: Analysis of Completely randomized Design (C.R.D), Randomized Block Design (R.B.D) and Latin Square Design (L.S.D) including one missing observation, expectation of various sum of squares. Comparison of the efficiencies of above designs.

UNIT - III

Vital statistics : Introduction, definition and uses of vital statistics. Sources of vital statistics, registration method and census method. Rates and ratios, Crude death rates, age specific death rate, standardized death rates, crude birth rate, age specific fertility rate, general fertility rate, total fertility rate. Measurement of population growth, crude rate of natural increase- Pearl's vital index. Gross reproductive rate sand Net reproductive rate, Life tables, construction and uses of life tables and Abridged life tables.

UNIT - IV

Indian Official Statistics: Functions and organization of CSO and NSSO. Agricultural Statistics, area and yield statistics. National Income and its computation, utility and difficulties in estimation of national income.

Index Numbers : Concept, construction, uses and limitations of simple and weighted index numbers. Laspeyer's, Paasche's and Fisher's index numbers, criterion of a good index numbers, problems involved in the construction of index numbers. Fisher's index as an ideal index number. Fixed and chain base index numbers. Cost of living index numbers and wholesale price index numbers. Base shifting, splicing and deflation of index numbers.

Contents

UNIT - IV

Frequently Asked & Important Questions

UNIT - I

1. Discuss the concept of Gauss-Markoff Linear Model with Examples.

Ans : **(July-22)**

Refer Unit-I, Q.No. 1

---

2. State and prove Cochran's Theorem.

Ans : **(Dec.-21,Oct-20)**

Refer Unit-I, Q.No. 2

---

3. What is an ANOVA? State the assumptions and applications of an ANOVA.

Ans : **(Imp.)**

Refer Unit-I, Q.No. 3

---

4. Give complete statistical analysis of ANOVA One-way classification.

Ans : **(July-22, July-21)**

Refer Unit-I, Q.No. 5

---

5. Explain the analysis of variance of two-way classification.

Ans : **(July-22)**

Refer Unit-I, Q.No. 6

UNIT - II

1. Explain in detail about Completely randomized Design (C.R.D).

Ans : **(July-21)**

Refer Unit-II, Q.No. 1

---

2. Discuss in detail about statistical analysis of CRD.

Ans : **(July-22)**

Refer Unit-II, Q.No. 2

---

3. Discuss about Randomized Block Design (R.B.D).

Ans : **(July-21, June-19)**

Refer Unit-II, Q.No. 3

4. Discuss in detail the statistical analysis of RBD.

Ans : **(June-19)**

Refer Unit-II, Q.No. 4

5. How can you estimate the missing observations in RBD?

Ans : **(Dec.-21, Oct.-20, June-19)**

Refer Unit-II, Q.No. 6

6. Define Latin Square Design (L.S.D). State its merits and demerits.

Ans : **(July-21, June-19)**

Refer Unit-II, Q.No. 7

7. Explain the Statistical Analysis of LSD.

Ans : **(June-19)**

Refer Unit-II, Q.No. 8

8. How do you estimate the missing observations in LSD.

Ans : **(July-22, July-21, Oct.-20)**

Refer Unit-II, Q.No. 9

9. Compare the efficiency of RBD relative to CRD.

Ans : **(July-22, Dec.-21, Oct.-20)**

Refer Unit-II, Q.No. 12

UNIT - III

1. Define vital statistics. Explain the uses of vital statistics.

Ans : **(July-22)**

Refer Unit-III, Q.No. 1

2. Explain various measures of mortality rate.

Ans : **(July-21)**

Refer Unit-III, Q.No. 4

3. Explain different types of fertility rates.

Ans : **(July-22, Dec.-21, Oct-20)**

Refer Unit-III, Q.No. 5

4. Write a detail note on population growth and how it can be measured.

Ans : **(July-21, Oct.-20)**

Refer Unit-III, Q.No. 6

5. Define Life table. What are the assumptions of Life table.

Ans : **(July-22, Dec.-21)**

Refer Unit-III, Q.No. 7

6. Explain various components of life tables.

Ans : **(July-22)**

Refer Unit-III, Q.No. 8

7. Write in detail about Abridged life tables.

Ans : **(July-22)**

Refer Unit-III, Q.No. 9

8. Explain Reed Merrell Method of Abridged life tables.

Ans : **(Imp.)**

Refer Unit-III, Q.No. 10

UNIT - IV

1. Explain Functions and organization of CSO.

Ans : **(Dec.-21, Oct.-20, June-19)**

Refer Unit-IV, Q.No. 2

2. Explain briefly about National Sample Survey Organisation (NSSO).

Ans : **(July-21)**

Refer Unit-IV, Q.No. 3

3. How are the agricultural statistics of area and yield collected in India.

Ans : **(Imp.)**

Refer Unit-IV, Q.No. 6

4. Define National Income. Explain the uses of National Income.

Ans : **(Dec.-21)**

Refer Unit-IV, Q.No. 7

5. Explain the importance of index numbers.

Ans : **(Imp.)**

Refer Unit-IV, Q.No. 11

6. What are the various methods of Constructing Index Numbers?

Ans : **(Imp.)**

Refer Unit-IV, Q.No. 14

7. What are the problems involved in construction of index numbers ? Explain.

Ans : **(Imp.)**

Refer Unit-IV, Q.No. 21

8. Define Cost of Living Index Numbers. Explain various methods of construction of Cost of Living Index Numbers.

Ans : **(July-22)**

Refer Unit-IV, Q.No. 24

9. Distinguish between fixed base and chain base index numbers. From the fixed base index numbers given below, construct chain base index numbers.

Year	2003	2004	2005	2006	2007	2008
Fixed Base index	94	98	102	95	98	100

Sol : **(July-22)**

Refer Unit-IV, Prob. 9

UNIT I

Analysis of Variance and Design of Experiments : Concept of Gauss-Markoff linear model with examples, statement of Cochran's theorem, ANOVA – one-way, two-way classifications with one observation per cell Expectation of various sums of squares, Statistical analysis, Importance and applications of design of experiments.

1.1 Concept of Gauss-Markoff Linear Model with Examples

Q1. Discuss the concept of Gauss-Markoff Linear Model with Examples.

(OR)

Explain the concept of Gauss-Markoff Linear Model.

Ans : **(July-22)**

Marks consider a set of n independent random variables $y_1, y_2 .....y_i .....y_n$ whose expectations given has linear function of 'm' unknown parameters $\beta_1, \beta_2 ....\beta_j ....\beta_m \,(m \le n)$ with known coefficient a and whose variance are a constant σ^2 does

$$E(y_i) = a_{i1}, \beta_1 + a_{i2}\beta_2 + ... + a_{ij}\beta_j + ...a_{im}\beta_{mj} \quad ... (1)$$

$$v(y_i) = \sigma^2 \operatorname{cov}(y_i, y_j) = 0 \quad (i \ne j)$$

Also $y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}_{n\times 1} \quad \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_m \end{bmatrix}_{m\times 1}$

$$A = \begin{bmatrix} a_{11} & a_{12}........a_{1m} \\ a_{12} & a_{22}.......a_{2m} \\ \vdots & \\ a_{i1} & a_{i2}.......a_{im} \\ \vdots & \\ a_{n1} & a_{n2}.......a_{nm} \end{bmatrix}_{n \times m}$$

from (1) $E(y) = A\beta$ ... (2)

$D(y) = \sigma^2 I$

Here D(y) denotes the dispersion matrix of y and I is the Identity matrix of order n in matrix from the general linear model may be written as

$$y_{n\times 1} = A_{n\times m}\beta_{m\times 1} + \Sigma_{n\times m} \quad ... (3)$$

E(E) = 0 (null vector)

$D(y) = \sigma^2 I$

The model $\rightarrow$(1) (or) $\rightarrow$ (2) (or) $\rightarrow$ (3) is known as Gauss mark off linear model (or) G.M model.

1.2 Statement of Cochran's Theorem

Q2. State and prove Cochran's Theorem.

Ans : **(Dec.-21,Oct-20)**

Let $x_1, x_2 .... x_n$ denote a random sample from normal population N (0, σ^2). Let the sum of the squares of these values be written in the form.

$$\sum_{i=1}^{n} x_i^2 = Q_1 + Q_2 + ... + Q_k$$

Where Q_j is a quadratic form in $x_1, x_2 ......x_n$, with rank (degree of freedom) r_j, j = 1, 2....k. Then the random variables $Q_1, Q_2 .....Q_k$ are mutually independent and Q_j / σ^2 is x^2 – variate with r_j degrees of freedom if and only if $\sum_{j=1}^{k} r_j = n$.

Proof :

If $\sum_{j=1}^{k} r_i = n$, then there exists an orthogonal matrix A such that from transformation X = Ay we get,

$$Q_1 = y_1^2 + y_2^2 + ... + y_{r_1}^2$$

$$Q_2 = y_{r_1+1}^2 + y_{r_1+2}^2 + ... + y_{r_1+r_2}^2$$

$$\vdots$$

$$Q_k = y_{n-r_k+1}^2 + y_{n-r_k+2}^2 + ... + y_n^2$$

Here, y_1, y_2, y_3,......y_n are independent and so Q_1, Q_2......Q_k are also independent. Similarly, when X = Ay, then $x \in N(0,\sigma^2)$ and since $\sum_{j=1}^{k} r_j = n$, Q/σ^2 is χ^2 - variate with rank r_c

1.3 ANOVA – One-way, Two-way Classifications with one Observation per Cell Expectation of Various Sums of Squares, Statistical Analysis

Q3. What is an ANOVA? State the assumptions and applications of an ANOVA.

Ans : **(Imp.)**

Meaning

The variance test is also known as ANOVA. ANOVA is the acronym for Analysis of variance. Analysis of variance is a statistical technique specially designed to test whether the means of more than two quantiative population are equal i.e., to make inferences about whether those samples are drawn from the populations having the same mean.

The test is called 'F' test as it was developed by R.A Fisher in 1920's. The test is conducted in situations where we have three (or) more to consider at a time an alternative procedure (to t-test) needed for testing the hypothesis that all samples could likely be drawn from the same population.

Example :

Five fertilizers are applied to four plots of wheat and yield of wheat and yield of wheat on these plots is given. We are interested in finding out whether the samples have come from the same population. ANOVA answers this question.

Assumptions

Analysis of variance test is based on the test statistic F (or variance ratio).

It is based on the following assumptions.

(i) Observation are independent.

(ii) Each sample is drawn randomly from a normal population as the sample statistics reflect the characteristic of the population.

(iii) Variance and means are identical for those population from which samples have been drawn.

Applications

The applications of ANOVA are as follows,

1. Anova is used in education, industry, business, psychology field mainly in their experiment design.
2. Anova helps to save time and money as several population means can be compared simultaneously.
3. Anova is used to test the linearity of the fitted regression line and correlation ratio, significance test statistic of anova = F(r – 1, n – r).

Q4. Explain briefly about One-Way ANOVA.

Ans : **(Jan.-20, June-19)**

The following are the various steps under One way Anova :

1. Set the null hypothesis H_0 and alternate hypothesis H_1.
2. Computation of Test statistic. The test statistic is obtained by the following steps :

Steps I :

Find the sum of the values of all the items of all the samples and denote it by T, i.e., obtain

$T = \sum_i \sum_j x_{ij}$.

Step II :

Compute the correction factor C, where

$C = \frac{T^2}{N}$. N is the total number of all the observations of all the samples and T is obtained in STEP 1.

Step III :

Find the sum of the squares of all items of all the samples i.e., find $\sum_i \sum_j x_{ij}^2$.

Step IV :

Find the total sum of the squares SST, where

$SST = \sum_i \sum_j x_{ij}^2 - C$, or $SST = \sum_i \sum_j x_{ij}^2 - \frac{T^2}{N}$.

where C is the correction factor obtained in STEP 2.

Step V :

Find the sum of the squares between the samples, i.e., SSC by the following steps

(i) Square the totals and divide it by the number of items in each sample.

$$P = \frac{(\Sigma X_i)^2}{n_i} + \frac{(\Sigma X_2)^2}{n_2} + \frac{(\Sigma X_3)^2}{n_3}$$

(ii) Subtract the correction factor $C = \frac{T^2}{N}$ from P, then the resulting figure would be the sum of the squares between the samples SSC, i.e.,

$$SSC = P - \frac{T^2}{N}$$

Step VI :

Find the sum of the squares within the samples SSE from the result

SSE = SST – SSC.

[From Step 4 and Step 5]

Step VII :

Set up the ANOVA table and calculate F, which is the test statistic.

3) Level of Significance. Let the level of significance be $\propto = 0.05$.

4) Decision

(a) If the computed value of F is less than the tabled value of F at $\alpha = 0.05$, then accept the null hypothesis H_0.

(b) If the computed value of F is greater than the tabled value of Fat $\alpha = 0.05$, then reject the null hypothesis H_0 and accept the alternate hypothesis H_1.

Q5. Give complete statistical analysis of ANOVA One-way classification.

(OR)

Find the expectation of treatment and error sum of square in one-way classification.

Ans : **(July-22, July-21)**

Null Hypothesis (H_0)

There is no significance difference between the effect of different treatments.

i.e., $\alpha_1 = \alpha_2 = ... = \alpha_k$

Alternative Hypothesis (H_1)

There is a significant difference between the effect of different treatments.

Notations

$\bar{x}_{i.}$ = Mean of the i[th] treatment

$$\bar{x}_{i.} = \frac{1}{n_i} \sum_{j=1}^{n_i} x_{ij}$$

$\bar{x}_{..}$ = Overall mean

$$\bar{x}_{..} = \frac{1}{N} \sum_{i=1}^{k} \sum_{j=1}^{n_i} x_{ij}$$

We have to estimate μ, α_i of equation (1) using principle of least squares such that the error sum of squares is minimum.

i.e., $\in = \sum_{i=1}^{k}\sum_{j=1}^{n_i} \in_{ij}^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij} - \mu - \alpha_i)^2$

is minimum

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

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

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

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

$$\left[\because \sum_{j=1}^{n_i} n_i = N\right]$$

$$N\mu = \sum_{i=1}^{k}\sum_{j=1}^{n_i} x_{ij} + 0$$

$$\left[\because \sum_{i=1}^{k} n_i\alpha_i = 0\right]$$

$$N\mu = \sum_{i=1}^{k}\sum_{j=1}^{n_i} x_{ij}$$

$$\mu = \frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{n_i} x_{ij}$$

$$\mu = \bar{x}_{...} \quad ... (2)$$

$$\frac{\partial E}{\partial \alpha_i} = 0 \Rightarrow \sum_{j=1}^{n_i} 2(x_{ij} - \mu - \alpha_i)^{2-1}(-1) = 0$$

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

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

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

$$\sum_{j=1}^{n_i} x_{ij} = n_i(\mu + \alpha_i)$$

$$\mu + \alpha_i = \frac{1}{n_i}\sum_{j=1}^{n_i} x_{ij}$$

$$\mu + \alpha_i = \bar{x}_{i.}$$

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

$$\alpha_i = \bar{x}_{i.} - \bar{x}_{..} \quad ... (3) [\because \text{ from } (2)]$$

Substitute μ, α_i in equation (1) we get,

$$x_{ij} = \bar{x}_{..} + \bar{x}_{i.} - \bar{x}_{..} + \in_{ij}$$

$$(x_{ij} - \bar{x}_{..}) = (\bar{x}_{i.} - \bar{x}_{..}) + (x_{ij} - \bar{x}_{i.})$$

$[\because \in_{ij} = x_{ij} - \bar{x}_{i.}]$

Squaring and taking summation on both sides.

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_{..})^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i} [(\bar{x}_{i.} - \bar{x}_{..}) + (x_{ij} - \bar{x}_{i.})^2]$$

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

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_{..})^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(\bar{x}_{i.} - \bar{x}_{..})^2 + \sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij.} - \bar{x}_{i.})^2 + 0 \quad ... (4)$$

[$\because$ the product term is zero]

$S_T^2 = \sum_i\sum_j (x_{ij} - \bar{x}_{..})^2$ is known as total sum of squares (T.S.S)

$S_E^2 = \sum_i\sum_j (x_{ij} - \bar{x}_{i.})^2$ is called with in sum of squares (or) error sum of squares (E.S.S) and

$S_t^2 = \sum_i n_i(\bar{x}_{i.} - \bar{x}_{..})^2$ is called S.S due to treatments (t.s.s) then T.S.S = E.S.S + t.S.S

Degree of Freedom

For total sum of squares (T.S.S) = N = 1

For treatment sum of squares (t.S.S) = k – 1

For error sum of squares (E.S.S) = N – k

Mean sum of squares (M.S.S)

The sum of squares divided by its degrees of freedom gives the corresponding variance or the mean sum of squares (M.S.S). Thus,

$\frac{St}{k-1} = \frac{T.S.S}{(k-1)} = S_t^2$ (say) is the M.S.S due to treatments and $\frac{S_E^2}{(N-k)} = \frac{E.S.S}{N-k} = S_E^2$ (say) is the M.S.S due to error.

Expectations of Various sum of squares of one-way classification

For obtaining appropriate test statistics for testing $H_0 : \mu_1 = \mu_2 ... \mu_k$ or its equivalent hypothesis $H_0 : \alpha_1 = \alpha_2 = ... = \alpha_k = 0$, we need the expectation of the S.S due to each of the independent factors.

Starting with the linear model. $x_{ij} = \mu + \alpha_i + \in_{ij}$ summing over j and dividing by n_i, we get,

$$\frac{1}{n_i}\sum_j x_{ij} = \frac{1}{n_i}(n_i \mu + n_i \alpha_i + \sum_j \in_{ij})$$

$$\bar{x}_{i.} = \mu + \alpha_i + \bar{\in}_i \quad .... (5)$$

Summing (1) over i and j and dividing by $N = \sum_i n_i$, we get,

$$\frac{1}{N}\sum_i\sum_j x_{ij} = \frac{1}{N}(N\mu + \sum_i n_i \alpha_i + \sum_i \sum_j^i \in_{ij})$$

$$\bar{x}_{..} = \mu + \bar{\in}_{..} \quad ... (6)$$

[$\because$ on using $\sum_{i=1}^{k} n_i \alpha_i = 0$]

Expectation of treatment sum of squares (T.S.S)

$$E(T.S.S) = E(S_t^2) = E. \sum_{i=1}^{k} n_i(\bar{x}_{i.} - \bar{x}_{..})^2$$

$$= E\{\sum_{i=1}^{k} n_i[(\mu + \alpha_i + \bar{\in}_{i.}) - (\mu + \bar{\in}_{..})^2\} \; [\because \text{ using eq(1) and (6)}]$$

$$= E[\sum_i n_i(\alpha_i + \bar{\in}_{i.} - \bar{\in}_{..})^2]$$

$$= E\{\sum_i n_i [\alpha_i^2 + (\bar{\in}_{i.} - \bar{\in}_{..})^2 + 2\alpha_i(\bar{\in}_{i.} - \bar{\in}_{..})]\}$$

$$= \sum_i E[n_i \alpha_i^2 + n_i(\bar{\epsilon}_{i.} - \bar{\epsilon}_{..})^2 + 2\sum_i (n_i \alpha_i E(\bar{\epsilon}_{i.} - \bar{\epsilon}_{..})]$$

$$= \sum_i E[n_i \alpha_i^2 + n_i(\bar{\epsilon}_{i.} - \bar{\epsilon}_{..})^2] + 2\sum_i [n_i \alpha_i E(\bar{\epsilon}_{i.} - \bar{\epsilon}_{..})]$$

$$= \sum_i n_i \alpha_i^2 + E[\Sigma n_i(\bar{\epsilon}_{i.} - \bar{\epsilon}_{..})^2]$$

$$[\because E(\bar{\epsilon}_{i.} - \bar{\epsilon}_{..}) = 0, \text{ as } \epsilon_{ij} \text{ are i.i.d } N(0, \sigma_e^2)]$$

$$= \sum_i n_i \alpha_i^2 + E[\sum_i n_i(\bar{\epsilon}_{i.} - N\bar{\epsilon}_{..}^2)]$$

$$= \sum_i n_i \alpha_i^2 + \sum_i n_i E(\bar{\epsilon}_{i.}^2) - NE(\bar{\epsilon}_{..}^2) \quad \ldots (*)$$

Since ϵ_{ij} are i.i.d $N(0, \sigma_e^2)$, $E(\bar{\epsilon}_{i.}) = 0$ and $E(\bar{\epsilon}_{..}) = 0$

$\therefore$ $\text{Var}(\bar{\epsilon}_{i.}) = E(\bar{\epsilon}_{i.}^2) - [E(\bar{\epsilon}_{i.})]^2 = E[\bar{\epsilon}_{i.}^2]$

$E[\bar{\epsilon}_{i.}^2] = \text{var}(\bar{\epsilon}_{i.}) = \sigma_e^2/n_i$

[Since the variance of the mean of a random sample of size n from population with variance σ^2 is σ^2/n]

Similarly, we have

$E(\bar{\epsilon}_{..}^2) = \text{var}(\bar{\epsilon}_{..}) = \sigma_e^2/N$

Substitution in $\rightarrow$(*), we get,

$$E(\text{t.S.S}) = \sum_i n_i \alpha_i^2 + \sum_i n_i \left(\frac{\sigma_e^2}{n_i}\right) - N.\frac{\sigma_e^2}{N}$$

$$= \sum_i n_i \alpha_i^2 + \sum_i \sigma_e^2 - \sigma_e^2$$

$$= \Sigma n_i \alpha_i^2 + K\sigma_e^2 - \sigma_e^2$$

$$= \Sigma n_i \alpha_i^2 + (K-1)\sigma_e^2$$

$$E\left[\frac{\text{T.S.S}}{k-1}\right] = \sigma_e^2 + \frac{1}{(k-1)} \sum_{i=1}^{k} n_i \alpha_i^2 \quad \ldots (7)$$

Hence under $H_0 : \sigma_1 = \sigma_2 = \ldots = \sigma_k = 0$ M.S.S due to treatments provides an unbiased estimate of σ_e^2.

Expectation of Error sum of Squares (E.S.S)

$$E(\text{E.S.S}) = E[\sum_i \sum_j (x_{ij} - \bar{\epsilon}_{i.})^2]$$

$$= E\,[\sum_i \sum_j \{(\mu + \alpha_i + \in_{ij}) - (\mu + \alpha_i + \overline{\in}_{i.})\}^2]$$

[$\because$ using (1) and (6)]

$$= E[\sum_i \sum_j (\in_{ij} - \overline{\in}_{i.})^2\}$$

$$= E[\sum_i (\sum_j (\in_{ij} - \overline{\in}_{i.})^2)]$$

$$= E[\sum_i (\sum_j (\in_{ij}{}^2 - n_i \overline{\in}_{i.}{}^2)]$$

$$= E[\sum_i \sum_j \in_{ij}{}^2 - \Sigma n_i\, \overline{\in}_{i.}{}^2]$$

$$= \sum_i \sum_j E(\in_{ij}{}^2) - \sum_i n_i\, E(\overline{\in}_{i.}{}^2)$$

$$= \sum_i \sum_j \sigma_e{}^2 - \sum_{i=1}^{k} n_i \left(\frac{\sigma_e{}^2}{n_i}\right)$$

$$= N\,\sigma_e{}^2 - \sum_{i=1}^{k} \sigma_e{}^2$$

$$= N\,\sigma_e{}^2 - k\sigma_e{}^2$$

$$= (N - k)\sigma_e{}^2$$

$$E\left[\frac{E.S.S}{N-k}\right] = \sigma_e{}^2 \qquad \ldots (8)$$

i.e., the error mean sum of squares always gives an unbiased estimate of $\sigma_e{}^2$.

Thus we have

$$E\left[s_t^2 = \frac{t.S.S}{k-1}\right] = E\left[s_E^2 = \frac{E.S.S}{N-k}\right] = \sigma_e{}^2 \qquad \ldots (9)$$

Under H_0, otherwise

$$E(s_t{}^2) > E(s_E{}^2) \qquad \ldots (10)$$

Hence the test statistics for H_0 is provided by the variance ratio.

$$F = \frac{s_t^2}{s_E^2} \qquad \ldots (11)$$

If H_0 is true then F should take the value 1, otherwise it should be greater than unity. In order to find out if an observed value of F is significantly greater than unity, we have to obtain the sampling distribution of the statistic F.

ANOVA table for One-Way classified Data

Source	Sum of squares	d.f	Mean sum of squares	Variance ratio
Treatment	S_t^2	K – 1	$s_t^2 = \frac{s_t^2}{(k-1)}$	
Error	S_E^2	N – k	$s_E^2 = \frac{s_E^2}{(N-k)}$	$\frac{s_t^2}{s_E^2} = F_{(k-1, N-k)}$
Total	S_T^2	N – 1		

Conclusion

If an observed value of F obtained from is (11) is greater than the tabulated value of F for [(k – 1, N – k)] d.f. at specified level of significance, (usually 5% or 1%) then H_0 is refuted at that level otherwise it may be retained.

(or)

If the calculated value of F is less than or equal to the critical value of F then accept H_0 i.e., there is no significant difference between the effect of different treatments otherwise reject H_0.

PROBLEMS

1. A farmer applied three types of fertilizers on 4 separate plots. The figure on yield per acre are tabulated below:

Fertilizers Plots	Yield				
	A	B	C	D	Total
Nitrogen	6	4	8	6	24
Potash	7	6	6	9	28
Phosphates	8	5	10	9	32

Find out if the plots are materials different in fertility, as also, if the three fertilizers make materials difference in yields.

Sol :

Correction Factor : C $= \frac{T^2}{N} = \frac{(84)^2}{12} = 588.$

SST = Total sum of square $= \Sigma\Sigma x_{ij}^2 - C$

$= [(6^2 + 7^2 + 8^2) + (4^2 + 6^2 + 5^2) + (8^2 + 6^2 + 10^2) + (6^2 + 9^2 + 9^2)] - 588$

$(36 + 49 + 64) + (16 + 36 + 25) + (64 + 36 + 100) + (36 + 81) + (81) - 588$

$= 624 - 588 = 36$

SSC = sum of square between plots = $\dfrac{\Sigma\left(\sum_j x_i\right)^2}{r} - C$

$= \left[\dfrac{(21)^2}{3} + \dfrac{(15)^2}{3} + \dfrac{(24)^2}{3} + \dfrac{(24)^2}{3}\right] - 588 = \dfrac{1818}{3} - 588$

$= \dfrac{441}{3} + \dfrac{225}{3} + \dfrac{576}{3} + \dfrac{576}{3} - 558$

$= 147 + 75 + 192 + 192 - 568$

$= 606 - 558 = l8$

Sum of squares between Fertilizers = $\left(\sum_{i=1}^{r} T_I^2 / c\right) - C$

$= \dfrac{(24)^2}{4} + \dfrac{(28)^2}{4} + \dfrac{(32)^2}{4} - 588$

$= \dfrac{576}{4} + \dfrac{784}{4} + \dfrac{1024}{4} - 588$

$= \dfrac{2384}{4} - 588 = 596 - 588 = 8$

Error sum of squares = 36 – 18 – 8 = 10

Null Hypothesis : (i) Plots are equally fertile

(ii) Fertilizer are equally effective

ANOVA Table

Prices of variation	**d.o.f.**	**S.S.**	**MM = $\dfrac{S.S.}{d.o.f}$**	**FC**	**F_{Tab}(5%)**
Between Plots	3	18	MSC = $\dfrac{18}{3}$ = 6	$\dfrac{MSC}{MSE}$ = 3.6	4.76
Between Fertilizer	2	8	MSR = $\dfrac{8}{2}$ = 4	$\dfrac{MSR}{MSE}$ = 2.4	5.14
Error	6	10	MSE = $\dfrac{10}{6}$ = 1.667		
Total	**C_{r-1} = 11**	**36**			

Conclusion

We accept the hypothesis at (i) and (ii), what is, the plots are equally effective and the fertilizer have the same effect.

2. Three different methods of teaching statistics are used on three groups of students, Random samples of size 5 are taken from each group and the results are shown below. The grades are on a 10 point scale :

Group A	Group B	Group C
7	3	4
6	6	7
7	5	7
7	4	4
8	7	8

Determine on the basis of the above data whether there is a difference in the teaching methods.

Sol : **(Nov.-21)**

(i) Null Hypothesis (H_0) : There is no difference between in the teaching methods.

(ii) Alternative Hypothesis (H_1) : There is a difference between in the teaching methods

(iii) Level of Significance (α) : 0.05 (Assume)

(iv) Calculations

Group 'A'	Group 'B'	Group 'C'
7	3	4
6	6	7
7	5	7
7	4	4
8	7	8
$T_1 \rightarrow 35$	$T_2 \rightarrow 25$	$T_3 \rightarrow 30$

Grand Total = 35 + 25 + 30 = 90

(i) Correction factor (c.f) $= \frac{(GT)^2}{N} = \frac{(90)^2}{15} = \frac{8100}{15} = 540$

(ii) Total sum of square (TSS) $= \Sigma(X_{ij}^2) - C.F = 576 - 540$

$$TSS = 36$$

(iii) Sum of square between samples

$$= \frac{\Sigma T_j^2}{n_j} - C.F$$

$$= \left[\frac{(35)^2}{5} + \frac{(25)^2}{5} + \frac{(30)^2}{5}\right] - 540$$

$$= \frac{1225}{5} + \frac{625}{5} + \frac{900}{5} - 540$$

$= 245 + 125 + 180 - 540$

$= 550 - 540$

$= \boxed{SSB = 10}$

(iv) Sum of the squares of errors

$SSE = TSS - SSR$

$36 - 10$

$\boxed{SSE = 26}$

ANOVA One Way Table

Source of Variation	**Degree of Freedom**	**Sum of Square**	**Mean Sum of Square**	**F – Ratio & $F_{calculation}$**
SSB	$3 - 1 = 2$	10	$10/2 = 5$	
SSE	$15 - 3 = 12$	26	$26/12 = 2.166$	$F_{cal} = \frac{5}{2.166} = 2.308$
TSS	$15 - 1 = 14$	36	–	

Conclusion :

F_{cal} & $F_{Ratio} = 2.166$

F_{tab} at (2, 12) degree of freedom $F_{tab} = 3.74$

$F_{tab} < F_{tab}$

$\therefore$ H_0 is accepted

We conclude that there is no difference in the teaching method.

3. The table below shows the lives of batches of electric lamps in hours.

Batches	**Life of Bulbs Calculated in Hours (y_{ij})**							
1	**1600**	**1610**	**1650**	**1680**	**1700**	**1720**	**1800**	**–**
2	**1580**	**1640**	**1640**	**1700**	**1750**	**–**	**–**	
3	**1460**	**1550**	**1600**	**1620**	**1640**	**1660**	**1740**	**1820**
4	**1510**	**1520**	**1530**	**1570**	**1600**	**1680**	**–**	**–**

Calculate analysis of variance on this data and show that significance test don't reject their homogeneity.

Sol:

Assume j^{th} observation in i^{th} batch as y_{ij} where $i = 1, 2, 3, 4$.

Setting up the Hypotheses

Null Hypothesis

$H_0 = \mu_1 = \mu_2 = \mu_3 = \mu_4$

Assume that the above four batches of electric lamps are homogeneous in nature.

H_1 : Here two means should be different now shift the origin to 1640 and then divide the integer by 10. In simple terms it can be given as, $u_{ij} = \frac{(y_{ij} - 1640)}{10}$. The table below shows the necessary calculations where y_{ij} = 1600, 1610, 1650.

Batches	Decoded Data : $u_{ij} = (y_{ij} - 1640)/10$								$T_i = \sum_j u_{ij}$	$\sum_j u_{ij}^2$
1	– 4	– 3	1	4	6	8	16		– 4 – 3 + 1 + 4 + 6 + 8 + 16 = 28	$(-4)^2 + (-3)^2 + (1)^2 + (4)^2 + (6)^2 + (8)^2 + (16)^2 = 398$
2	– 6	0	0	6	11				– 6 + 6 + 11 =11	$(-6)^2 + (6)^2 + (11)^2 = 193$
3	–18	–9	– 4	–2	0	2	10	18	–18 – 9 – 4 – 2 + 0 + 2 + 10 + 18 = – 3	$(-18)^2 + (-9)^2 + (-4)^2 + (-2)^2 + (0)^2 + (2)^2 + (10)^2 + (18)^2 = 853$
4	–13	–12	–11	–7	–4	4			–13 –12 – 11 – 7 – 4 + 4 = – 43	$(-13)^2 + (-12)^2 + (-11)^2 + (-7)^2 + (-4)^2 + (4)^2 = 515$
	ToTal								$G = \sum_i \sum_j u_{ij} = -7$	$\sum_i \sum_j u_{ij}^2 = 1959$

$$\therefore \quad \text{R.S.S} = \sum_i \sum_j u_{ij}^2 = 1959$$

$$G = \sum_i \sum_j u_{ij} = -7$$

Calculating the Correction Factor (C.F)

$$\text{C.F} = \frac{G^2}{N}$$

N = Number of bulbs

$$= \frac{(-7)^2}{26} = \frac{49}{26} = 1.884615$$

Total S.S is calculated by taking difference of R.S.S and C.F

Total S.S = R.S.S – C. F

= 1,959 – 1.884615

= 1957.115385

Calculating sum of squares between batches = $\sum_{i=1}^{4} \left[\frac{T_i^2}{n_i}\right] - \text{C.F}$

Where, T_i = 28, 11, –3, – 43

n_i = Number of bulbs = 7, 5, 8, 6

$$= \left[\frac{(28)^2}{7}+\frac{(11)^2}{5}+\frac{(-3)^2}{8}+\frac{(-43)^2}{6}\right]-C.F$$

$$= \left[\frac{784}{7}+\frac{121}{5}+\frac{9}{8}+\frac{1849}{6}\right]-1.884615$$

= [112 + 24.2 + 1.125 + 308.16] – 1.884616

= [445.485] – [1.884615] = 443.560385

Now the difference between total S.S and between S.S is calculated in order to generate sum of square.

∴ Sum of squares with in (batches) = Total S.S – Between S.S

= 1957.115385 – 443.60385

= 1513.611535

Calculating the Degrees of Freedom

Total S.S = N – 1

$$= \sum_{i=1}^{4} n_i - 1 = (7 + 5 + 8 + 6) - 1 = 26 - 1 = 26$$

Between the batches S.S

= k – 1

= 4 – 1 = 3

Now calculate the error = 25 – 3 = 22

Tabulating the value in ANOVA table.

Sources of Variation	d.f	Sum of squares	Mean S.S = $\frac{\text{Sum of squares}}{\text{d.f}}$	V.R (F)
Between batches	3	443.600385	$\frac{443.600385}{3} = 147.86795$	$\frac{147.86795}{68.795978} = 2.1493$
With batches (error)	22	1513.511535	$\frac{1513.511535}{22} = 68.795978$	
Total	**25**	**1957.11539**		

∴ The calculated $F_{0.05}$ corresponding to 3 and 22 d.f is 3.05 while for $F_{0.01}$ corresponding to 3 and 22 d.f is 4.82.

At 5% and 1% levels, the generated value of F is not effective and does not reject H_0. Thus, the four batches of electric lamps is considered as homogeneous.

4. The average number of days survived by mice after getting effected by 5 strains of typhoid organisms with their SD and total number of mice participated in experiments. Calculate the strains of typhoid organisms based on the data in the table below.

Strains of Typhoid			
	No. of Mice (n_i)	**Average ($\overline{y}_1$)**	**Standard Deviation (S_i)**
A	**10**	**10.9**	**12.72**
B	**6**	**13.5**	**5.96**
C	**8**	**11.5**	**3.29**
D	**11**	**11.2**	**5.65**
E	**5**	**15.4**	**3.64**

Sol:

Null Hypothesis, H_0

Assume various strains of typhoid organisms to be homogeneous.

Now making $\mu_A = \mu_B = \mu_C = \mu_D = \mu_E$

H_i = Here, two means should be different

Consider T_i as the total for i^{th} strain of typhoid

G = The grand total

$$= \sum_i T_i$$

Then $\overline{y}_i = \frac{T_i}{n_i} = T_i = n_i \overline{y}_i$

Subequently, $S_i^2 = \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij}^2 = n_i(S_i^2 + \overline{y}_i^2)$

This generates S.S of observations for i^{th} strain of typhoid.

Calculations for different sum of squares is given in the below table,

Strains of Typhoid						
	n_i	$\overline{y}_i$	$T_i = n_i\overline{y}_i$	$\overline{y}_i^2$	S_i^2	$n_i(S_i^2 - \overline{y}_i^2)$
A	10	10.9	10 × 10.9 = 109	10.9 × 10.9 = 118.81	12.72 × 12.72 = 161.79	2806
B	6	13.5	6 × 13.5 = 81	135 × 13.5 = 182.25	5.96 × 5.96 = 35.52	1306.62
C	8	11.5	8 × 11.5 = 92	11.5 × 11.5 = 135.25	3.25 × 3.25 = 10.49	1141.92
D	11	11.2	11 × 11.2 = 123.2	11.2 × 11.2 = 125.44	5.65 × 5.65 = 31.92	1730.96
E	5	15.4	5 × 15.4 = 77	15.4 × 154 = 237.16	3.64 × 3.64 = 13.24	1252
Total	**40**		**G = 482.2**			**R.S.S = 8237.5**

Now, calculate C.F by dividing G^2 and N.

$$C.F = \frac{G^2}{N} = \frac{(482.2)^2}{40} = \frac{232516.84}{40} = 5812.92$$

$$R.S.S = \sum_i \sum_j y_{ij}^2$$

$$= \sum_i (S_i^2 + y_i^2)$$

$$= 8237.5$$

Calculate T.S.S by taking difference of R.S.S and C.F = R.S.S – C.F

$= 8237.5 – 5812.92$

$= 2424.58$

Estimate S.S due to strain of typhoid

$$= \sum_i \frac{T_i^2}{n_i} - C.F$$

$$= \left[\frac{(109)^2}{10} + \frac{(81)^2}{6} + \frac{(92)}{8} + \frac{(123.2)^2}{11} + \frac{(77)^2}{5}\right] - 5812.92$$

$$= \left[\frac{11881}{10} + \frac{6561}{6} + \frac{8464}{8} + \frac{15178.24}{11} + \frac{5929}{5}\right] - 5812.92$$

$= [11881 + 1093.5 + 1058 + 1379.84 + 1185.8] – 5812.92$

$= 5905.24 – 5812.92$

$= 92.32$

Estimating the error = T.S.S – S.S due to strain

$= 2425.58 – 92.32$

$= 2332.26$

Tabulate the given values in ANOVA Table.

Sources of variation	**Sum of squares**	**d.f**	**Mean S.S = $\frac{\text{Sum of Squares}}{\text{d.f}}$**	**V.R (F)**
Between Strains of Typhoid	4	92.32	$\frac{92.32}{4} = 23.08$	$\frac{66.636}{23.08} = 2.88$
Error	35	2332.26	$\frac{2332.26}{35} = 66.636$	
Total	39	2424.58		

$\therefore$ The tabulated data $F_{0.05}$ with respect to 4 and 35 d.f is 5.735 as the generated value pertaining to F is less than the tabulated value as is not considered at 5% level of significance while the null hypothesis H_0 is considered.

Q6. Explain briefly about two way ANOVA.

(OR)

Explain the analysis of variance of two-way classification.

Ans : **(July-22)**

Two way classification/two factor ANOVA is defined where two independent factors have an effect on the response variable of interest.

Example : Yield of crop affected by type of seed as well as type of fertilizer.

Procedure

(a) Calculate the variance between columns,

$$SSC = \sum_{j=1}^{c} \frac{T_j^2}{n_j} - \frac{T^2}{N}$$

(b) Calculate the variance between rows,

$$SSR = \sum_{i=1}^{r} \frac{T_i^2}{n_i} - \frac{T^2}{N}$$

(c) Compute the total variance,

$$SST = \sum X_{ij}^2 - \frac{T^2}{N}$$

(d) Calculate the variance of residual or error,

SSE = TSS – (SSC + SSR)

(e) Divide the variances of between columns, between rows and residue by their respective degrees of freedom to get the mean squares.

(f) Compute F ratio as follows,

F-ratio concerning variation between columns,

$$= \frac{\text{Mean square between columns}}{\text{Mean squares of residual}}$$

F-ratio concerning variation between rows,

$$= \frac{\text{Mean square between rows}}{\text{Mean squares of residual}}$$

(g) Compare F-ratio calculated with F-ratio from table,

If F-ratio (calculated) < F-ratio (table), H_1 accepted,

If F-ratio (calculated) $\geq$ F-ratio (table), H_0 rejected,

H_1 accepted $\Rightarrow$ no significant differences

H_0 rejected $\Rightarrow$ significant differences

Two-Way ANOVA with Interation

Under two-way ANOVA with interaction, the total sum of squares SST is divided into four components, which are as follows,

1. The Sum of Squares for factor 'A' (SSA)
2. The Sum of Squares for factor 'B' (SSB)
3. The Sum of Squares for the interaction between two factors 'SSAB'.
4. The Error of Sum of Squares (SSE).

These factors can be represented as,

$$SST = SSA + SSB + SSAB + SSE$$

The main purpose of using two-way ANOVA with interaction is to understand the relationship between factors 'A' and 'B'. Such relationship will help to find out the impact, effect or influence of factor 'A' on factor 'B' and factor 'B' on factor 'A'.

Two-Way ANOVA without Interaction

Under two-way ANOVA without interaction, the total variability of data is divided into three components, which are as follows,

1. Treatment i.e., factor 'A'
2. Block i.e., factor 'B'
3. Chance.

However, the term 'block' refers to a matched group of observations from each population. When units of each block are assigned randomly to each treatment then the design of such experiment is referred as randomized block design.

Note :

Two factors are said to interact if the difference between levels (treatments) of one factor depends on the level of the other factor. Factors that do not interact are called additive.

A combination of a treatment from one factor with a treatment from another factor results in an interaction.

An interaction between two factors exists when for atleast one combination of treatments, the effect of combination is not additive.

ANOVA Table for Two-way Classified Data with m-Observation Per Cell

Sources of Variation	Degree of Freedom	S.S	M.S.S	Variance Ratio F
Factor A	p – 1	S_A^2	$S_A^2 = \frac{S_A^2}{p-1}$	$F_A = \frac{S_A^2}{S_E^2}$
Factor B	q – 1	S_B^2	$S_B^2 = \frac{S_B^2}{q-1}$	$F_B = \frac{S_B^2}{S_E^2}$

Interaction AB	(p–1) (q–1)	S^2_{AB}	$F_{AB} = \frac{S^2_{AB}}{S^2_E}$	
Factor AB	pq (m – 1)	S^2_E	$S^2_E = \frac{S^2_E}{pq(m-1)}$	
Total	pqm – 1			

Remark

The calculation of various sum of squares is facilitated to a great extent by the use of following formulae,

$$C.F = \frac{G^2}{pqm} = \frac{T^2}{pqm}$$

$$TSS = \sum_i \sum_j \sum_k x_{ij}k^2 - C.F = RSS - CF = \sum_i T_i^2$$

$$S^2_A = \frac{i}{M} CF$$

$$S^2_B = \frac{j}{M} CF$$

$$S^2_{AB} = \sum \sum T_{ij}^2$$

SS due to Means (SSM)

$$= \frac{ij}{mp} CF$$

$$S^2_{AB} = SSM - S^2_A - S^2_B$$

$$S^2_E = TSS - S^2_A - S^2_B - S^2_{AB}$$

Hypothesis Tests in Two-way ANOVA

- **Factor A Test**

  Hypothesis is designed to determine whether there are any factor A main effects. Null Hypothesis true if and only if there are no differences in means due to different treatments (population) of factor A.

- **Factor B Test**

  Hypothesis is test designed to detect factor B main effects. Null hypothesis is true if and only if there are no differences in means due to different treatments (populations) of factor B.

- **Test for AB Interactions**

  Test for existence of interactions between levels of the two factors Null hypothesis is true if and only if there are no two way interactions between levels of factor A and levels of factor B, means factor effects are additive for two way ANOVA.

Q7. Describe the statistical analysis of two-way ANOVA.

Ans :

Null Hypothesis

We set up the null hypothesis that the treatment as well as varieties are homogeneous. In other words, the null hypothesis for treatments and varieties are respectively.

$H_t : \mu_{1.} = \mu_{2.} = \cdots = \mu_{k} = \mu$

$H_v : \mu_{.1} = \mu_{.2} = \cdots = \mu_{.l} = \mu$

(or)

$H_t : \alpha_1 = \alpha_2 = \cdots = \alpha_k = 0$

$H_v : \beta_1 = \beta_2 = \cdots = \beta_l = 0$

Alternative Hypothesis

There is a significance difference between different treatments and between different varieties.

Notations

$\bar{x}_{i.}$ = Mean of the i^{th} treatment

$$= \frac{1}{l}\sum_{j=1}^{l} x_{ij}$$

$$\bar{x}_{.j} = \frac{1}{k}\sum_{i=1}^{k} x_{ij} = \text{Mean of } j^{th} \text{ variety}$$

$\bar{x}_{..}$ = Over all mean

$$= \frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{l} x_{ij}$$

where $N = kl$

We have to estimate μ, α_i, β_j in equation (1) of mathematical model using the principle of least squares such that the error S.S. is minimum.

$$\text{i.e., } E = \sum_{i=1}^{k}\sum_{j=1}^{l} \in_{ij}^{2} = \sum_{i=1}^{k}\sum_{j=1}^{l} (x_{ij} - \mu - \alpha_i - \beta_j)^2 \text{ is minimum}$$

$$\frac{\partial E}{\partial \mu} = 0 \Rightarrow \sum_{i=1}^{k}\sum_{j=1}^{l} 2(x_{ij} - \mu - \alpha_i - \beta_j)(-1) = 0$$

$$\sum_{i=1}^{k}\sum_{j=1}^{l} X_{ij}(x_{ij} - \mu - \alpha_i - \beta_j) = 0$$

$$\sum_{i=1}^{k}\sum_{j=1}^{l} x_{ij} - \sum_{i=1}^{k}\sum_{j=1}^{l}(\mu) - \sum_{i=1}^{k}\sum_{j=1}^{l}\alpha_i - \sum_{i=1}^{k}\sum_{j=1}^{l}\beta_j = 0$$

$$\sum_{i=1}^{k}\sum_{j=1}^{l} x_{ij} - \mu(N) - l\sum_{i=1}^{k}\alpha_i - k\sum_{j=1}^{l}\beta_j = 0$$

$$\sum_{i=1}^{k}\sum_{j=1}^{l} x_{ij} = N\mu \qquad \left[\because \sum_{i=1}^{k}\alpha_i = 0, \sum_{j=1}^{l}\beta_j = 0\right]$$

$$\mu = \frac{1}{N}\sum_{i=1}^{K}\sum_{j=1}^{l} x_{ij}$$

$$\mu = \bar{x}_{..} \qquad \ldots (2)$$

$$\frac{\partial E}{\partial \alpha_i} = 0 \Rightarrow \sum_{j=1}^{l} 2(x_{ij} - \mu - \alpha_i - \beta_j)^{2-1}(-1) = 0$$

$$\sum_{j=1}^{l}(x_{ij} - \mu - \alpha_i - \beta_j) = 0$$

$$\sum_{j=1}^{l} x_{ij} - \sum_{j=1}^{l}\mu - \sum_{j=1}^{l}\alpha_i - \sum_{j=1}^{l}\beta_j = 0$$

$$\sum_{j=1}^{l} x_{ij} - l(\mu + \alpha_i) - 0 = 0 \qquad \left[\because \sum_{j=1}^{l}\beta_j = 0\right]$$

$$\sum_{j=1}^{l}(x_{ij}) = l(\mu + \alpha_i)$$

$$\alpha_i = \frac{1}{l}\sum_{j=1}^{l} x_{ij} - \mu$$

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

$$\alpha_i = \bar{x}_{i.} - \bar{x}_{..} \qquad \ldots (3)$$

$$\frac{\partial E}{\partial \beta j} = 0 \Rightarrow \sum_{i=1}^{k} 2(x_{ij} - \mu - \alpha_i - \beta_j)^{2-1}(-1) = 0$$

$$\sum_{j=1}^{K}(x_{ij} - \mu - \alpha_i - \beta_j) = 0$$

$$\sum_{j=1}^{K} x_{ij} - k(\mu + \beta_j) = 0$$

$$\beta_j = \frac{1}{k}\sum_{i=1}^{K} x_{ij} - \mu$$

$$\beta_j = \bar{x}_{.j} - \bar{x}_{..} \qquad \ldots (4)$$

Substitute (2), (3), (4) in (1) we get,

$$x_{ij} = \bar{x}_{..} + (\bar{x}_{i.} - \bar{x}_{..}) + (\bar{x}_{.j} - \bar{x}_{..}) + \in_{ij}$$

$$(x_{ij} - \bar{x}_{..}) = (\bar{x}_{i.} - \bar{x}_{..}) + (\bar{x}_{.j} - \bar{x}_{..}) + \in_{ij}$$

$$(x_{ij} - \bar{x}_{..}) = (\bar{x}_{i.} - \bar{x}_{..}) + (\bar{x}_{.j} - \bar{x}_{..}) + (x_{ij} - \bar{x}_{i.} - \bar{x}_{.j} + \bar{x}_{..})$$

Squaring and taking summation on both sides.

$$\sum_{i=1}^{K}\sum_{j=1}^{l}(x_{ij} - \bar{x}_{..})^2 = \sum_{i=1}^{K}\sum_{j=1}^{l}[(\bar{x}_{i.} - \bar{x}_{..}) + (\bar{x}_{.j} - \bar{x}_{..}) + (x_{ij} - \bar{x}_{i.} - \bar{x}_{.j} + \bar{x}_{..})]$$

$$\sum_{i=1}^{K}\sum_{j=1}^{l}(x_{ij} - \bar{x}_{..})^2 = \sum_{i=1}^{K}\sum_{j=1}^{l}(\bar{x}_{i.} - \bar{x}_{..})^2 + \sum_{i=1}^{K}\sum_{j=1}^{l}(\bar{x}_{.j} - \bar{x}_{..})^2$$

$$+ \sum_{i=1}^{K}\sum_{j=1}^{l}(x_{ij} - \bar{x}_{i.} - \bar{x}_{.j} + \bar{x}_{..})^2 \quad [\because \text{ the product terms are zero}]$$

$$\sum_{i=1}^{K}\sum_{j=1}^{l}(x_{ij} - \bar{x}_{..})^2 = l\sum_{i=1}^{K}(\bar{x}_{i.} - \bar{x}_{..})^2 + k\sum_{j=1}^{l}(\bar{x}_{.j} - \bar{x}_{..})^2 +$$

$$\sum_{i=1}^{K}\sum_{j=1}^{l}(x_{ij} - \bar{x}_{i.} - \bar{x}_{.j} + \bar{x}_{..})^2$$

Total S.S = Treatment S.S + Variety S.S + Error S.S

(or) $S_T^2 = S_t^2 + S_v^2 + S_E^2$

Where

$S_T^2 = \sum_i \sum_j (x_{ij} - \bar{x})^2$ is the total S.S

$S_t^2 = l\sum_i (\bar{x}_{i.} - \bar{x}_{..})^2$ is the S.S due to treatments

$S_v^2 = k\sum_j (\bar{x}_{.j} - \bar{x}_{..})^2$ is the S.S due to varieties

$S_E^2 = \sum_i \sum_j (\bar{x}_{ij} - \bar{x}_{i.} - \bar{x}_{.j} + \bar{x}_{..})^2$ is the error (or) residual S.S

Degrees of Freedom For Various S.S

The total S.S., S_T^2 being computed from $N = lk$ quantities $(x_{ij} - \bar{x}_{..})$ which are subject to one linear constraint $\sum_i \sum_j (x_{ij} - \bar{x}_{..}) = 0$, will carry (N – 1) d.f

Similarly S_t^2 will be based on (k – 1) d.f., since

$\sum_i (\bar{x}_{i.} - \bar{x}_{..}) = 0$ and S_v^2 will have $(l - 1)$ d.f., since

$\sum_j(\bar{x}_{.j} - \bar{x}_{..}) = 0$ and S_E^2 will carry (N – 1) – (K – 1) – (*l* – 1) = (*l* – 1) (k – 1) d.f

($\because$ N = *l*k)

Thus the partitioning of d.f is as follows.

(*l*k – 1) = (k – 1) + (*l* – 1) + (*l* – 1) (k – 1) which implies that the d.f. are additive.

Test Statistic

In order to obtain appropriate test statistic to test the hypothesis H_t and H_v we need the expectations of the various mean S.S due to each of independent factors.

Using the same notations for the mean S.S as in the case of one-way classified data, we get

Mean S.S due to treatments = $\dfrac{S_t^2}{k-1} = s_t^2$, (say)

Mean S.S due to varieties = $\dfrac{S_v^2}{l-1} = s_v^2$, (say)

Error mean S.S = $\dfrac{S_E^2}{(l-1)(k-1)} = S_E^2$, (say)

Summing (1) over j from 1 to *l* and dividing by h, we get,

$$\frac{1}{l}\sum_j x_{ij} = \frac{1}{l}[l\mu + l\alpha_i + \sum_j \beta_j + \sum_j \in_{ij}]$$

$$\bar{x}_{i.} = \mu + \alpha_i + \bar{\in}_{i.} \quad \text{... (5)} \quad [\because \sum_j \beta_j = 0]$$

Similarly summing (1) over i from 1 to k and dividing by k and using $\sum_j \alpha_i = 0$, we shall get,

$$\bar{x}_{.j} = \mu + \beta_j + \bar{\in}_{i.} \quad \text{... (6)}$$

Summing (1) over i and j both and dividing by hk and using (6) we shall get.

$$\bar{x}_{..} = \mu + \bar{\in}_{..} \quad \text{... (7)}$$

Expectations of Various sum of Squares

$$E(s_t^2) = E[l \sum_{i=1}^{k}(\bar{x}_{i.} - \bar{x}_{..})^2]$$

$$= E[h\sum_i(\mu + \alpha_i + \bar{\in}_{i.} - \mu - \bar{\in}_{..})^2][\text{from (5) \& (7)}]$$

$$= E[l \sum_i\{\alpha_i + (\bar{\in}_{i.} - \bar{\in}_{..})\}^2]$$

$$= E[l\sum_i\alpha_i^2 + l\sum_i(\bar{\in}_{i.} - \bar{\in}_{..})^2 + 2l\sum_i\alpha_i(\bar{\in}_{i.} - \bar{\in}_{..})]$$

$$= l\sum_i\alpha_i^2 + lE[\Sigma(\bar{\in}_i - \bar{\in}_{..})^2 + 2l\sum_i\alpha_i E(\bar{\in}_i - \bar{\in}_{..})$$

$$= l\sum_i\alpha_i^2 + l\,E[\sum_i \bar{\in}_{i.}^2 - k\,\bar{\in}_{..}^2) \quad \because E(\bar{\in}_{i.} - \bar{\in}_{..}) = 0 \text{ as } \in_{ij} \text{ are i.d } N(0, \sigma_e^2)$$

$$= l\sum_i\alpha_i^2 + l[\sum_i E(\bar{\in}_{i.}^2)] - kE(\bar{\in}_{..}^2)]$$

$$= l\sum_i \alpha_i^2 + l\left[\sum_i\left(\frac{\sigma_e}{h}\right)^2 - k.\frac{\sigma_e^2}{kl}\right]$$ [$\because$ $\in_{ij}$ are i.i.d N(0, σ_e^2)]

$$= l\sum_i \alpha_i^2 + l\left[\frac{k\sigma_e^2}{l} - \frac{\sigma_e^2}{l}\right]$$

$$E(s_t^2) = l\sum_{i=1}^{k} \alpha_i^2 + (k-1)\,\sigma_e^2 \qquad \dots (8)$$

$$E\left[\frac{s_t^2}{k-1}\right] = \frac{l}{k-1}\sum_{i=1}^{k}\alpha_i^2 + \sigma_e^2$$

$$E[s_t^2] = \sigma_e^2 + \frac{l}{(k-1)}\sum_{i=1}^{k}\alpha_i^2 \qquad \dots (9)$$

Similarly we can prove that

$$E(S_v^2) = K\sum_{j=1}^{l}\beta_j + (l-1)\sigma_e^2 \qquad \dots (10)$$

$$E\left[\frac{s_v^2}{l-1}\right] = \frac{k}{l-1}\sum_j\beta_j + \sigma_e^2$$

$$E(s_v^2) = \sigma_e^2 + \frac{k}{(l-1)}\sum_{j=1}^{l}\beta_j^2 \qquad \dots (11)$$

Expectation of error sum of square is given by

$$E(s_E^2) = E[\sum_i\sum_j(x_{ij} - \bar{x}_{i.} - \bar{x}_{.j} + \bar{x}_{..})^2]$$

On substituting from (1) and (5) to (7) and simplifying we shall get,

$$E(s_E^2) = E[\sum_i\sum_j(\in_{ij} - \bar{\in}_{i.} - \bar{\in}_{.j} - \bar{\in}_{j.} + \bar{\in}_{..})^2]$$

$$E[\sum_i\sum_j(\in_{ij}^2 + \bar{\in}_{i.}^2 + \bar{\in}_{.j}^2 + \bar{\in}_{..}^2 - 2\in_{ij}\bar{\in}_{i.} - 2\in_{ij}\bar{\in}_{.j} + 2\bar{\in}_{ij}\bar{\in}_{..} + 2\bar{\in}_{i.}\bar{\in}_{.j}$$

$$- 2\bar{\in}_{i.}\bar{\in}_{.j} - 2\bar{\in}_{.j}\bar{\in}_{..})]$$

$$= E[\sum_i\sum_j\in_{ij}^2 + l\sum_i\bar{\in}_{i.}^2 + k\sum_j\bar{\in}_{ij}^2 - lk\,\bar{\in}_{..}^2 - 2l\sum_i\bar{\in}_{i.}^2 -$$

$$2k\sum_j \bar{\in}_{.j}^2 + 2lk\, \bar{\in}_{i.}^2 + 2lk\, \bar{\in}_{..}^2 - 2lk\, \bar{\in}_{..}^2 - 2lk\, \bar{\in}_{..}^2]$$

$$= \sum_i \sum_j E(\in_{ij}^2) + l\sum_{i=1}^{k} E(\bar{\in}_{i.}^2) + k\sum_{j=1}^{l} E(\bar{\in}_{.j}^2) + lk\, E(\bar{\in}_{..}^2) - 2l\sum_{i=1}^{k} E(\bar{\in}_{i.}^2) - 2k\sum_{j=1}^{l} E(\bar{\in}_{.j}^2)$$

$$= \sum_i \sum_j \sigma_e^2 + l\sum_{i=1}^{k}\left(\frac{\sigma_e^2}{l}\right) + k\sum_{j=1}^{l}\left(\frac{\sigma_e^2}{k}\right) + lk\,\frac{\sigma_e^2}{lk} - 2l\sum_{i=1}^{k}\left(\frac{\sigma_e^2}{l}\right) - 2k\sum_{i=1}^{l}\left(\frac{\sigma_e^2}{k}\right)$$

$[\because\ \in_{ij}$ are i.i.d $N(0, \sigma_e^2)]$

$$= lk\,\sigma_e^2 + k\sigma_e^2 + l\sigma_e^2 + \sigma_e^2 - 2k\,\sigma_e^2 - 2l\,\sigma_e^2$$

$$= (lk - k - l + 1)\sigma_e^2$$

$$E(s_E^2) = (l-1)(k-1)\sigma_e^2 \qquad \text{... (12)}$$

$$E\left[\frac{s_E^2}{(l-1)(k-1)}\right] = \sigma_e^2$$

$$E(s_E^2) = \sigma_e^2 \qquad \text{... (13)}$$

Thus we see that M.S.S always provides unbiased estimate of σ_e^2 whereas s_t^2 and s_v^2 are unbiased estimates of σ_e^2 under the null hypothesis H_t and H_v as given in (6) respectively.

Since $\sum_i \alpha_i^2 \geq 0$ and $\sum_j \beta_i^2 \geq 0$, we get

$$E(S_t^2) = E(S_E^2),\ \text{under } H_t \qquad \text{... (14)}$$

Otherwise $E(S_t^2) > E(S_E^2)$

and $E(S_v^2) = E(S_E^2)$, under H_v ... (15)

Otherwise $E(S_v^2) > E(S_E^2)$

Since various S.S as well as their respective d.f are additive and since under the null hypothesis H_t and H_v, each of S_T^2, S_t^2, S_v^2 and S_E^2 provides an unbiased estimate of normality of parent population, we get by cochran's theorem.

$$\frac{S_t^2}{\sigma_e^2},\ \frac{S_v^2}{\sigma_e^2} \text{ and } \frac{S_E^2}{\sigma_e^2}$$

are mutually independent x^2 - variates with (k - 1), $(l-1)$ and $(l-1)(k-1)$ d.f. respectively. Hence under H_t and H_v respectively, we get.

$$F_t = \frac{S_t^2}{\sigma_e^2(k-1)} \div \frac{S_E^2}{\sigma_e^2(l-1)(k-1)}$$

$$= \frac{S_t^2}{S_E^2},\ \text{conforms to } F_{(k-1),(l-1)(k-1)}$$

and $F_v = \frac{S_v^2}{\sigma_e^2(l-1)} \div \frac{S_E^2}{\sigma_e^2(l-1)(k-1)}$

$= \frac{S_v^2}{S_E^2}$, conforms to $F_{(l-1),\ (l-1)(k-1)}$

ANOVA table for two-way classified data

Sources of variation	Sum of squares (S.S)	d.f	M.S.S	Variance ratio
Treatments	$S_T^2 = \sum_i l(\bar{x}_{i.} - \bar{x}_{..})^2$	k-1	$S_t^2 = \frac{s_T^2}{(k-1)}$	$F_t = \frac{S_t^2}{S_t^2} \sim F_{(k-1),(l-1)(k-1)}$
Varieties	$S_v^2 = \sum_i k(\bar{x}_{.j} - \bar{x}_{..})^2$	l-1	$S_v^2 = \frac{s_v^2}{(l-1)}$	$F_v = \frac{s_v^2}{s_E^2} \sim F_{(l-1)(l-1)(k-1)}$
Error	$S_E^2 = \sum_i \sum_j (x_{ij} - \bar{x}_{i.} - \bar{x}_{.j} + \bar{x}_{..})^2$	(l-1) x(k-1)	$S_E^2 = \frac{S_E^2}{(l-1)(k-1)}$	
Total	$\sum_i \sum_j (x_{ij} - \bar{x}_{..})^2$	$lk - 1$		

Conclusion

Calculated value of F is less than tabulated value of F for respective d.f then we accept H_0, the null hypothesis of the homogeneity of various treatments and various varieties may be rejected or accepted.

Q8. Find the expectation of error sum of squares in two-way classification.

Ans : **(July-22)**

Now, computing expectation of error sum of squares.

$$E(S_E^2) = E\left[k\sum_i \sum_j \bar{y}_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{..})^2\right]$$

$$= E\left[\sum_i \sum_j (\mu + \alpha_i + \beta_j + \varepsilon_{ij} - \mu - \alpha_i + \bar{\varepsilon}_i - \mu - \beta_j - \bar{\varepsilon}_{.j} + \mu + \bar{\varepsilon}_{..})^2\right] \quad \therefore\ y_{ij} = \mu + \alpha_1 + \beta_j + \varepsilon_{ij}$$

$\bar{y}_{i.} = \mu + \alpha_1 + \bar{\varepsilon}_{i.}$

$$= E\left[\sum_i \sum_j (\varepsilon_{ij} - \bar{\varepsilon}_{i.} - \bar{\varepsilon}_{.j} + \bar{\varepsilon}_{..})^2\right]$$

$\bar{y}_{j.} = \mu + \beta_1 + \bar{\varepsilon}_{j.}$

$\bar{y}_{..} = \mu + \bar{\varepsilon}_{..}$

$$= E\left[\sum_i \sum_j (\varepsilon_{ij} - \bar{\varepsilon}_{i.} - \bar{\varepsilon}_{.j} + \bar{\varepsilon}_{..})^2\right]$$

$$= E\left[\sum_i\sum_j(\varepsilon_{ij}^2 + \overline{\varepsilon}_i.^2 - \overline{\varepsilon}._j^2 + \overline{\varepsilon}..^2 - (2\varepsilon_{ij}\overline{\varepsilon}_i.) + (2\varepsilon_{ij}\overline{\varepsilon}._j) + 2\varepsilon_{ij}\overline{\varepsilon}..2\overline{\varepsilon}_i.\overline{\varepsilon}._j - 2\overline{\varepsilon}_j.\overline{\varepsilon}.. - 2\overline{\varepsilon}._j\,\overline{\varepsilon}.)\right]$$

$$= E\left[\sum_i\sum_j\varepsilon_{ij}^2 h\sum_i\overline{\varepsilon}_i.^2 k\sum_j\overline{\varepsilon}._j^2 + hk\overline{\varepsilon}..^2 - 2h\sum_i\overline{\varepsilon}..^2 - 2k\sum_i\overline{\varepsilon}._j^2 + 2hk\overline{\varepsilon}..^2 + 2hk\overline{\varepsilon}..^2 - 2hk\overline{\varepsilon}..^2 - 2hk\overline{\varepsilon}..^2\right]$$

$$= \sum_i\sum_j E(\varepsilon_{ij}^2) + h\sum_{i=1}^{k}E(\overline{\varepsilon}_i.^2) + k\sum_{j=1}^{h}E(\overline{\varepsilon}._j^2) + khE(\overline{\varepsilon}..^2) - 2h\sum_{i=1}^{k}E(\overline{\varepsilon}_i.^2) - 2k\sum_{j=1}^{h}E(\overline{\varepsilon}._j^2)$$

$$= \sum_i\sum_j\sigma_e^2 + h\sum_{i=1}^{k}\left(\frac{\sigma_e^2}{h}\right) + k\sum_{j=1}^{h}\left(\frac{\sigma_e^2}{k}\right) + hk\left(\frac{\sigma_e^2}{hk}\right) - 2h\sum_{i=1}^{k}\left(\frac{\sigma_e^2}{h}\right) - 2k\sum_{j=1}^{h}\left(\frac{\sigma_e^2}{k}\right)$$

$$\left[\because E(\varepsilon_{ij}^2) = \sigma_e^2, E(\overline{\varepsilon}_i.^2) = \frac{\sigma_e^2}{h}, E(\overline{\varepsilon}._j^2) = \frac{\sigma_e^2}{k}, E(\overline{\varepsilon}..^2) = \frac{\sigma_e^2}{hk}\right]$$

$$= hk\sigma_e^2 + k\sigma_e^2 + h\sigma_e^2 + \sigma_e^2 - 2k\sigma_e^2 - 2h\sigma_e^2$$

$$= (hk - k - 1)\sigma_e^2$$

$$= (h - 1)(k - 1)\sigma_e^2$$

$$E(S_E^2) = (h - 1)(k - 1)\sigma_e^2$$

$$\Rightarrow \quad E\left[\frac{S_E^2}{(h-1)(k-1)}\right] = \sigma_e^2$$

$$\Rightarrow \quad E(S_E^2) = \sigma_e^2$$

Thus, error M.S.S provides unbiased estimate of and σ_e^2 S_t^2, S_V^2 are unbiased estimates of σ_e^2 under the null hypotheses.

<u>**PROBLEMS**</u>

5. Four different drugs have been developed for a certain disease. These drugs are used under three different environments. It is assumed that the environment might affect efficacy of drugs. The number of cases of recovery from the disease per 100 people who have taken the drugs is tabulated as follows :

Environment	Drug A1	Drug A2	Drug A3	Drug A4
I	19	8	23	8
II	10	9	12	6
III	11	10	13	16

Test whether the drugs differ in their efficacy to treat the disease, also whether there is any effect of environment on the efficacy of disease.

Sol :

Null Hypothesis

H_o = There is no significant difference in the efficacy of drugs to treat the disease.

H_o = There is no significant effect of environment on the efficacy of disease.

	Drugs				
Environment	**A_1**	**A_2**	**A_3**	**A_4**	**Total**
I	19	8	23	8	58
II	10	9	12	6	37
III	11	10	13	16	50
Total	**40**	**27**	**48**	**30**	**GT=145**

$$\text{Correction Factor, (CF)} = \frac{(\text{Grand Total})^2}{N}$$

$$= \frac{(145)^2}{12} = 1752.08$$

Total Sum of Squares (TSS)

$$(\text{TSS}) = \sum_i \sum_j X_{ij}^2 - \text{C.F}$$

$$= [(19)^2 + (8)^2 + (23)^2 + (8)^2 + (10)^2 + (9)^2 + (12)^2 + (6)^2 + (11)^2 + (10)^2 + (13)^2 + (16)^2] - \text{C.F}$$

$$= 361 + 64 + 529 + 64 + 100 + 81 + 144 + 36 + 121 + 100 + 169 + 256 - 1752.08$$

$$= 2025 - 1752.08$$

$\therefore$ TSS = 272.92

Sum of Squares Between Drugs (Column)

$$\text{SSC} = \sum_j \frac{T_j^2}{n_j} - \frac{(GT)^2}{N}$$

$$= \left[\frac{(40)^2}{3} + \frac{(27)^2}{3} + \frac{(48)^2}{3} + \frac{(30)^2}{3}\right] - 1752.08$$

$$= (533.33 + 243 + 768 + 300) - 1752.08$$

$$= 1844.33 - 1752.08$$

$\therefore$ SSC = 92.25

Degree of freedom (r) = (C – 1)

= (4 – 1) = 3

Sum of Squares Between Environment (Rows)

$$\text{SSR} = \sum_i \frac{T_i^2}{n_i} = \frac{(GT)^2}{N}$$

$$= \left[\frac{(58)^2}{4} + \frac{(37)^2}{4} + \frac{(50)^2}{4}\right] - \text{C.F}$$

$$= (841 + 342.25 + 625) - 1752.08$$

$$= 1808.25 - 1752.08$$

$\therefore$ SSR = 56.17

Degree of freedom,

$V_m = (r - 1) - (3 - 1) - 2$

Residual = Total sum of squares – (Sum of squares between columns + Sum of squares between rows)

= TSS – (SSC + SSR) = 272.92 – (92.25 + 56.17)

= 272.92 – 148.42 = 124.5

ANOVA TABLE

Sources of variation	Sum of squares	Degrees of Freedom	Means squares	Variance Ratio (F)
Between Drugs	92.25	(C–1) = (4–1) = 3	$\frac{92.25}{3} = 30.75$	$F_{(3.6)} = \frac{30.75}{20.75} = 1.48$
Between Environment	56.17	(r–1) = (3–1) = 2	$\frac{56.17}{2} = 28.08$	$F_{(2,6)} = \frac{28.09}{20.75} = 1.353$
Residual or Error (e)	124.5	(C–1) (r–1) = 3 × 2 = 6	$\frac{124.5}{6} = 20.75$	
Total	**272.92**	**(12 – 1) = 11**		

[**Note:** As level of significance is not given in the problem, assume 5% level of significance]

Critical value of $F_{0.05}$	Computed value of F
Drugs at V_0 (3,6) = 4.76	1.48
Environment at V_m(2,6) = 5.14	1.354

Table values are calculated as per 5% level of significance.

Decision

1. Drugs

Since the calculated value of F(1.48) is less than the table value (4.76), null hypothesis is accepted. Hence, there is no significant difference in the efficacy drugs.

2. Environment

Since the calculated value of F(1.354) is less than the table Value (5.14), null hypothesis is accepted. Hence, there is no affect of environment on the efficacy of disease.

6. Suppose that we are interested in establishing the yield producing ability of four types of soya beans A, B, C and D. We have three blocks of land X,Y and Z which may be different in fertility. Each block of land is divided into four plots and the different types of soya beans are assigned to the plots in each block by a random procedure. The following results are obtained:

	Soya Bean			
Block	**Type A**	**Type B**	**Type C**	**Type D**
X	5	9	11	10
Y	4	7	8	10
Z	3	5	8	9

Test whether A,B,C and D are significantly different.

Sol :

Null Hypothesis

H_o : There is no significant difference between A,B,C and D.

	Soya bean				
Block	**Type A**	**Type B**	**Type C**	**Type D**	**Total**
X	5	9	11	10	35
Y	4	7	8	10	29
Z	3	5	8	9	25
Total	12	21	27	29	GT = 89

$$\text{Correction Factor (CF)} = \frac{(\text{Grand Total})^2}{N} = \frac{(89)^2}{12}$$

$$= 660.08$$

Total Sum of Squares (TSS)

$$= \sum_i \sum_j X_{ij}^2 - \frac{(GT)^2}{N}$$

$$= [(5)^2 + (9)^2 + (11)^2 + (10)^2 + (4)^2 + (7)^2 + (8)^2 + (10)^2 + (3)^2 + (5)^2 + (8)^2 + (9)^2 1] - 660.08$$

$$= [25 + 81 + 121 + 100 + 16 + 49 + 64 + 100 + 9 + 25 + 64 + 81] - 660.8$$

$$= 735 - 660.08$$

$$\therefore \text{TSS} = 74.92$$

Sum of Squares Between Soya Bean (Columns)

$$\text{SSB} = \sum_j \frac{T_j^2}{n_j} - \frac{(GT)^2}{N}$$

$$= \frac{(12)^2}{3} + \frac{(21)^2}{3} + \frac{(27)^2}{3} + \frac{(29)^2}{3} - 660.08$$

$= [48 + 147 + 243 + 280.33] - 660.08$

$= 718.33 - 660.08$

$\therefore$ SSB $= 58.25$

Degree of freedom (r) $= (K - 1)$

$= (4 - 1)$

$= 3$

Mean sum of squares between the soya beans $= \dfrac{58.25}{3} = 19.41$

Sum of Squares within Blocks (SSW)

SSW $=$ TSS $-$ SSB

$= 74.92 - 58.25$

$= 16.67$

Mean sum of squares within the blocks

$$= \frac{16.67}{12-4} = \frac{16.67}{8} = 2.08$$

ANOVA TABLE

Sources of variation	Sum of squares	Degrees of Freedom	Means squares
Between soya bean type	58.25	(k – 1) = (4 – 1) = 3	19.42
Within blocks	16.67	(n – k) = (12 – 4) = 8	2.08
Total		(n – 1) = (12 – 1) = 11	

$$\text{F-Ratio} = \frac{\text{Mean square between soya bean type}}{\text{Mean square within blocks}}$$

$$= \frac{19.42}{2.08} = 9.33$$

[Note: Assuming level of significance as 5%]

F-Ratio$_{(3,\ 8)}$, calculated $= 9.34$

F-Ratio from table for $V_1 = 3$ and $V_2 = 8$ at 5% level of significance $= 4.07$

Decision

The calculated value of F is more than the table value. Therefore we reject null hypothesis (H_0) which means that there is a significant difference between types of soya beans.

7. Given the Observations,

	Blocks				
Treatment 1	**14**	**6**	**11**	**0**	**9**
Treatment 2	**14**	**10**	**16**	**9**	**16**
Treatment 3	**12**	**7**	**10**	**9**	**12**
Treatment 4	**12**	**8**	**11**	**6**	**7**

Construct the analysis of variance table and test for differences among the treatments using 0.05 test.

Sol:

1. Means of each observation is calculated.

	Blocks				
Treatment 1	14	6	11	0	9
Treatment 2	14	10	16	9	16
Treatment 3	12	7	10	9	12
Treatment 4	12	8	11	6	7
Total	52	31	48	24	44
Mean ($\bar{y}$)	$\bar{y}_1$ = 52/4 = 13	$\bar{y}_2$ = 31/4 = 7.75	$\bar{y}_3$ = 48/4 = 12	$\bar{y}_4 = 24/4 = 6$	$\bar{y}_5$ = 44/4 = 11

2. Combined mean of each observation is calculated.

$$\bar{Y} = \frac{13+7.75+12+6+11}{5} = \frac{49.75}{5} = 9.95$$

3. $S^2_{between} = P_1(\bar{y}_1 - \bar{Y})^2 + P_2(\bar{y}_2 - \bar{Y})^2 + P_3(\bar{y}_3 - \bar{Y})^2 + P_4(\bar{y}_4 - \bar{Y})^2 + P_5(\bar{y}_5 - \bar{Y})^2$

$= 4(13 - 9.95)^2 + 4(7.75 - 9.95)^2 + 4(12 - 9.95)^2 + 4(6 - 9.95)^2 + 4(11 - 9.95)^2$

$= 4 \times 9.303 + 4 \times 4.84 + 4 \times 4.203 + 4 \times 15.603 + 4 \times 1.103$

$= 37.212 + 19.360 + 16.812 + 62.412 + 4.412$

$= 140.208$

4. $M^2_{between} = \frac{S^2_{between}}{df_{between}}$

$df_{between}$ = Number of columns – 1 = 5 – 1 = 4

$\therefore \; M^2_{between} = \frac{140.208}{4} = 35.052$

5. $S^2_{within} = \Sigma(y_1 - \bar{y}_1)^2 + \Sigma(y_2 - \bar{y}_2)^2 + \Sigma(y_3 - \bar{y}_3)^2 + \Sigma(y_4 - \bar{y}_4)^2 + \Sigma(y_5 - \bar{y}_5)^2$

I Sample

(y_1)	$(y_1 - \bar{y}_1)$; $\bar{y}_1 = 13$	$(y_1 - \bar{y}_1)^2$
14	14 – 13 = 1	1
14	14 – 13 = 1	1
12	12 – 13 = – 1	1
12	12 – 13 = – 1	1
		$\Sigma(y_1 - \bar{y}_1)^2 = 4$

II Sample

(y_2)	$(y_2 - \bar{y}_2)$; $\bar{y}_2 = 7.75$	$(y_2 - \bar{y}_2)^2$
6	6 – 7.75 = – 1.75	3.0625
10	10 – 7.75 = 2.25	5.0625
7	7 – 7.75 = – 0.75	0.5625
8	8 – 7.75 = 0.25	0.0625
		$\Sigma(y_2 - \bar{y}_2)^2 = 8.75$

III Sample

(y_3)	$(y_3 - \bar{y}_3)$; $\bar{y}_3 = 12$	$(y_3 - \bar{y}_3)^2$
11	11 – 12 = – 1	1
16	16 – 12 = 4	16
10	10 – 12 = – 2	4
11	11 – 12 = – 1	1
		$\Sigma(y_3 - \bar{y}_3)^2 = 22$

IV Sample

(y_4)	$(y_4 - \bar{y}_4)$; $\bar{y}_4 = 6$	$(y_4 - \bar{y}_4)^2$
0	0 – 6 = – 6	36
9	9 – 6 = 3	9
9	9 – 6 = 3	9
6	6 – 6 = 0	0
		$\Sigma(y_4 - \bar{y}_4)^2 = 54$

V Sample

(y_5)	$(y_5 - \bar{y}_5)$; $\bar{y}_5 = 11$	$(y_5 - \bar{y}_5)^2$
9	9 – 11 = – 2	4
16	16 – 11 = 5	25
12	12 – 11 = 1	1
7	7 – 11 = – 4	16
		$\Sigma(y_5 - \bar{y}_5)^2 = 46$

$\therefore\ S^2_{within} = 4 + 8.75 + 22 + 54 + 46 = 134.75$

6. $M^2_{within} = \dfrac{S^2_{within}}{df_{within}}$

df_{within} = Number of items within all the observations – Number of columns

= 20 – 5 = 15

$\therefore\ M^2_{within} = \dfrac{134.75}{15} = 8.983$

7. ANOVA table

Source of variance	S^2	df	M^2
Variance between group	140.208	4	35.052
Variance within group or residual variance	134.75	15	8.983
Total	**274.958**	**19**	**44.035**

8. F - value/F - statistic

$$\text{F- value} = \frac{\text{Variance between group}}{\text{Variance within group (or) residual variance}}$$

$$= \frac{M^2_{between}}{M^2_{within}} = \frac{35.052}{8.983} = 3.902$$

Therefore, the calculated F-value is 3.902. Then for tabulated value is read by taking 15 df as m_2 and 4 df as m_1. Hence the tabulated value of F is 3.06

Conclusion

As the calculated F - value is greater than the tabulated F - value, there exists significant difference among the treatments.

1.4 Importance and Applications of Design of Experiments

Q9. Discuss the important definitions used in experimental design.

Ans :

The following are the important definitions are used in experimental design.

1. Experiment

An experiment is a device of getting an answer to the problem under consideration. The experiment is of two types.

1. Absolute experiment
2. Comparative experiment

2. Absolute Experiment

It consists of determining the absolute value of some characteristic like.

- Obtaining the average intelligence of group of people.
- Finding the correlation between two variables.

3. Comparative Experiment

It is designed to compare the effect of two (or) more objects on some population characteristics,

e.g.,

1. Comparison of different fertilizers
2. Comparison of different medicines for a disease.

4. Treatments

The various objects of comparison in a comparative experiment are known as treatments.

Example

1. Different fertilizers
2. Different varieties of crop (or) different methods of cultivation are the treatments.

5. Experimental Material

The whole experimental area which is used for experimentation, is known as experimental material.

6. Experimental Unit or Plot

The smallest division of experimental material to which we apply the treatments and on which we make the observation is known as experimental unit e.g., in field experiments the plots of "land" is the experimental unit. In other experiments unit may be patient in a hospital, a lump of dough, a group of pigs in a pen or a batch of seeds.

7. Blocks

In agricultural experiments, most of the times we divide the whole experimental unit into relatively homogeneous sub groups or strata. These strata, which are more uniform amongst themselves than the field as a whole, are known as blocks.

8. Yield

The measurement of the variable under study on different experimental units e.g., plots, in field experiments are termed as yields.

9. Experimental Error

Let us suppose that a large homogeneous field is divided into different plots (of equal shape and size) and different treatments are applied to these plots. If the fields from some of the treatments are more than those of the others, the experimenter is faced with the problem of deciding if the observed differences are really due to treatment effect or they are due to chance factors. In field experimentation, it is a common experience that the fertility gradient of the soil does not follow any systematic pattern but behaves in a erratic fusion. Experience tells us that even if the same treatment is used on all the plots, the yields would still very due to the difference in soil fertility such variation from plot to plot, which is due to random factors beyond human control, is spoken of as experimental error. It may be pointed out that the term 'error' used here is not synonymous with 'mistake' but is a technical term which includes all types of extraneous variations due to

- The inherent variability in the experimental material to which treatments are applied.
- The lack of uniformity in the methodology of conducting the experiment or in otherwords failure to standardise the experimental technique, and.
- Lack of representativeness of the sample to the population under study.

10. Efficiency of a Design

Consider the designs D_1 and D_2 with error variances per unit σ_1^2 and σ_2^2 and replications r_1 and r_2 respectively. Then the variance of the difference between two treatment means is given by

$\frac{2\sigma_1^2}{r_1}$ and $\frac{2\sigma_2^2}{r_2}$ for D_1 and D_2 respectively

then the ratio.

$$E = \frac{2\sigma_2^2}{r_2} \cdot \frac{r_1}{2\sigma_1^2} = \frac{r_1}{\sigma_1^2} \div \frac{r_2}{\sigma_2^2}$$

is termed as efficiency of design D_1 w.r.t. D_2. In other words, efficiency of D_1 w.r.t D_2 may be defined as the "ratio of the precisions of D_1 and D_2".

If E = 1, then both the designs D_1 and D_2 are said to be equally efficient.

If E > 0 (E < 1) then D_1 is said to be more (less) efficient than D_2.

11. Uniformity Trial

The fertility of the soil may be distributed randomly over the entire field but does not in general increase (or) decrease uniformly in any direction uniformity trials help us to divide the field into zones of equal soil fertility (or) otherwise, the whole field may be made as homogeneous blocks with same size and shape of equal fertility and so, the errors get reduced uniformity trials also give us a graphic picture of the variation of soil fertility.

Q10. What is an experimental design? Explain the principles of experimental design.

(OR)

Write in detail about principles of Experimentation.

Ans : **(Oct.-20, June-19)**

Meaning

"The systematic design of an experiment which may define the degree of uncertainty in an experiment" is experimental design. The collection of data and analysis should be well planned and organized so as to obtain a productive design with least sampling error.

Principles

Prof R.A. fisher suggested the following basic principles of experimentation to conduct valid experiments in the presence of many naturally fluctuating conditions.

1. Replication
2. Randomisation
3. Local control (or) Error control

1. Replication

The process of repeating the same treatment on different experimental units under similar conditions is known as replication. It is concerned with no. of experimental units under each treatment.

Suppose that a treatment is allotted to r-experimental units, then it is said to be replicated r-no. of times and hence, the standard error of the mean is $\sigma/\sqrt{r}$.

Advantages

1. Replications are essential to get valid estimate of the experimental error.
2. The larger no. of replications reduces the standard error of treatments and hence, the precision of the experiment increases.
3. Replications give more precise estimates of treatment effects.

Disadvantages

1. Replications cannot be increased infinitely as it increases the cost of experimentation.
2. A large no. of replications cannot be taken as it causes the scarcity of the resources.

3. Replication of treatments may some times subject to bias.

2. Randomisation

A process in which the treatments are allocated to the experimental units in such a way that each and every experimental unit has got equal chance receiving any of the treatments is called randomisation.

Randomisation is efficiently implemented with the help of random number tables (or) tossing of a coin (or) dice. Randomization is a matter of allocating the treatments randomly to the experimental units.

Advantages

1. It eliminates human biases.
2. Randomisation leads to an unbiassed estimate of error variance and unbiased estimate of treatment differences.
3. Randomisation makes the experiment free from any systematic effects of environment.
4. Randomisation introduces the independence among the experimental errors, hence, F-test.

Disadvantages

1. Well trained persons are necessary to perform the technique of randomisation hence, it leads to more expenditure.
2. If there exists a large no. of experimental units, then performing randomisation becomes difficult.

Local Control (or) Error Control

In experimental designs it is noticed that the error can enter at any stage of the experimentation, even so the replication and randomisation are greatly considered. Perhaps, this could be due to heterogeneous experimental material. Here, in agricultural field the whole heterogeneous experimental material is divided into relatively homogeneous sub groups or strata, on which the treatments are applied randomly with required no. of replications. Eventually, the error can also be controlled by means of local control.

"The process of reducing the experimental error by dividing the relatively heterogenous experimental area (field) into homogeneous blocks" is known as local control.

Advantages

1. Local control is meant to make the design more efficient.
2. Local control reduces the experimental error.
3. It makes any test of significance more sensitive and powerful.
4. A reduction in the experimental error consequently helps the investigator to detect the small real difference between the treatments.

Disadvantages

1. Soil fertility, an important factor which distributed randomly over the entire field, hence, making the blocks becomes difficult.
2. Some times the available experimental material itself homogenous, now making them into blocks is wasting the time and money.

Q11. Explain the Importance of Design of Experiments

Ans :

1. The design of experiments is meant for reducing the effect of errors due to expected sources and due to other sources.
2. The design of experiments is a must to improve the precision of the experiment.
3. Appropriate design of experiments is vital in researches where we need to accurately ascertain the statistical significance of various factors.
4. Design of experiments is imperative in business settings, where we need to determine an unbiased data for forecasting and for continuous improvement.
5. To organise the experiment properly the design of experiments is very required.
6. To Design of experiments is the backbone of any product design as well as any process/ product improvement efforts.
7. Design of experiments is important as a formal way of maximizing information gained while resources required.

Q12. Explain the Applications of Design of Experiments

Ans :

The following are the Applications of Design of Experiments

1. Design of experiments and its applications are tremendously used in the field of agriculture, green house studies, laboratories, marketing etc.,
2. DOE is applied in the field of pharmaceutical statistics to analysis of drug trials and to issues of commercialization of a medicine.
3. The application of experimental methods is applied to experimental economics to study the various economic questions.
4. In clinical trials, DOE is applied for safety collection of efficacy data of new drugs (or) devices.
5. DOE is a tool that has been used by many industries for the purpose of optimizing processes.
6. DOE has variety of applications in the field of software packages and in military also in industry.

PROBLEMS

8. Calculate the minimum number of replications so that an observed difference of 10% of the mean will be taken as significant at 5% level, the C.V. of the plot values being 12%

Sol :

Given that,

Difference of means = 10% of mean

t is significant at 5% level

C.V. of the plot values = 12%

We know that,

Difference of means = $\bar{x}_1 - \bar{x}_2$

= 10% μ

= 10% μ

$$C.V. = 100\frac{\sigma}{\mu}$$

$$\Rightarrow 12 = 100 \times \frac{\sigma}{\mu}$$

$$\Rightarrow 12\mu = 100\sigma$$

$$\Rightarrow \sigma = \frac{12\mu}{100} = 0.12\mu$$

When mean difference of two treatments at certain level of significance is given, the minimum number of replications can be calculated by the formula,

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sigma\sqrt{\frac{2}{r}}}$$

$$= \frac{0.10\mu}{0.12\mu\sqrt{\frac{2}{r}}}$$

$$t = \frac{10}{12} \times \frac{1}{\sqrt{\frac{2}{r}}}$$

Given that, t is significant at 5% level. $\therefore |t| > 1.96$

$$\Rightarrow \left|\frac{10}{12}\frac{1}{\sqrt{\frac{2}{r}}}\right| > 1.96$$

$$\Rightarrow \frac{10}{12}\frac{1}{\sqrt{\frac{2}{r}}} > 1.96$$

$$\Rightarrow \frac{10}{12} \times \frac{1}{\frac{\sqrt{2}}{\sqrt{r}}} > 1.96$$

$$\Rightarrow \frac{10}{12} \times \frac{\sqrt{r}}{\sqrt{2}} > 1.96$$

$$\Rightarrow \sqrt{r} > 1.96 \times \frac{12}{10} \times \sqrt{2}$$

$$\Rightarrow \sqrt{r} > 1.96 \times 1.2 \times \sqrt{2}$$

$$\Rightarrow (\sqrt{r})^2 > (1.96 \times 1.2 \times \sqrt{2})^2$$

$$\Rightarrow r > (1.96 \times 1.2)^2 \times (\sqrt{2})^2$$

$$\Rightarrow r > (2.35)^2 \times 2$$

$$\Rightarrow r > 5.53 \times 2$$

$$\Rightarrow r > 11.1$$

$\therefore$ r $\simeq$ 12 (approximate value or next integer greater than 11.1)

Therefore, the number of replications (r) is 12.

Short Question and Answers

1. Explain the concept of Gauss-Markoff Linear Model.

Ans :

Marks consider a set of n independent random variables $y_1, y_2 .....y_i .....y_n$ whose expectations given has linear function of ‘m’ unknown parameters $\beta_1, \beta_2 ....\beta_j....\beta_m (m \le n)$ with known coefficient a and whose variance are a constant σ^2 does

$$E(y_i) = a_{i1},\beta_1 + a_{i2}\beta_2 + ... + a_{ij}\beta_j + ...a_{im}\beta_{mj} \quad ... (1)$$

$$v(y_i) = \sigma^2 \text{cov}(y_i, y_j) = 0 \quad (i \ne j)$$

Also $y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}_{n\times 1}$ $\beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_m \end{bmatrix}_{m\times 1}$

$$A = \begin{bmatrix} a_{11} & a_{12}........a_{1m} \\ a_{12} & a_{22}.......a_{2m} \\ \vdots & \\ a_{i1} & a_{i2}.......a_{im} \\ \vdots & \\ a_{n1} & a_{n2}.......a_{nm} \end{bmatrix}_{n\times m}$$

from (1) $E(y) = A\beta$ ... (2)

$D(y) = \sigma^2 I$

Here D(y) denotes the dispersion matrix of y and I is the Identity matrix of order n in matrix from the general linear model may be written as

$$y_{n\times 1} = A_{n\times m}\beta_{m\times 1} + \Sigma_{n\times m} \quad ... (3)$$

E(E) = 0 (null vector)

$D(y) = \sigma^2 I$

The model $\rightarrow$(1) (or) $\rightarrow$ (2) (or) $\rightarrow$ (3) is known as Gauss mark off linear model (or) G.M model.

2. Find the expectation of error sum of squares in two-way classification.

Ans :

Now, computing expectation of error sum of squares.

$$E(S_E^2) = E\left[k\sum_i\sum_j \bar{y}_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{..})^2\right]$$

$$= E\left[\sum_i\sum_j(\mu + \alpha_i + \beta_j + \varepsilon_{ij} - \mu - \alpha_i + \bar{\varepsilon}_i - \mu - \beta_j - \bar{\varepsilon}_{.j} + \mu + \bar{\varepsilon}_{..})^2\right] \quad \therefore \quad y_{ij} = \mu + \alpha_1 + \beta_j + \varepsilon_{ij}$$

$$\bar{y}_i. = \mu + \alpha_1 + \bar{\varepsilon}_i.$$

$$= E\left[\sum_i\sum_j(\varepsilon_{ij} - \bar{\varepsilon}_i. - \bar{\varepsilon}_{.j} + \bar{\varepsilon}_{..})^2\right] \qquad \bar{y}_j. = \mu + \beta_1 + \bar{\varepsilon}_j.$$

$$\bar{y}_{..} = \mu + \bar{\varepsilon}_{..}$$

$$= E\left[\sum_i\sum_j(\varepsilon_{ij} - \bar{\varepsilon}_i. - \bar{\varepsilon}_{.j} + \bar{\varepsilon}_{..})^2\right]$$

$$= E\left[\sum_i\sum_j(\varepsilon_{ij}^2 + \bar{\varepsilon}_i.^2 - \bar{\varepsilon}_{.j}^2 + \bar{\varepsilon}_{..}^2 - (2\varepsilon_{ij}\bar{\varepsilon}_i.) + (2\varepsilon_{ij}\bar{\varepsilon}_{.j}) + 2\varepsilon_{ij}\bar{\varepsilon}_{..}2\bar{\varepsilon}_i.\bar{\varepsilon}_{.j} - 2\bar{\varepsilon}_j.\bar{\varepsilon}_{..} - 2\bar{\varepsilon}_{.j}\bar{\varepsilon}.)\right]$$

$$= E\left[\sum_i\sum_j\varepsilon_{ij}^2 h\sum_i\bar{\varepsilon}_i.^2 k\sum_j\bar{\varepsilon}_{.j}^2 + hk\bar{\varepsilon}_{..}^2 - 2h\sum_i\bar{\varepsilon}_{..}^2 - 2k\sum_i\bar{\varepsilon}_{.j}^2 + 2hk\bar{\varepsilon}_{..}^2 + 2hk\bar{\varepsilon}_{..}^2 - 2hk\bar{\varepsilon}_{..}^2 - 2hk\bar{\varepsilon}_{..}^2\right]$$

$$= \sum_i\sum_j E(\varepsilon_{ij}^2) + h\sum_{i=1}^{k}E(\bar{\varepsilon}_i.^2) + k\sum_{j=1}^{h}E(\bar{\varepsilon}_{.j}^2) + khE(\bar{\varepsilon}_{..}^2) - 2h\sum_{i=1}^{k}E(\bar{\varepsilon}_i.^2) - 2k\sum_{j=1}^{h}E(\bar{\varepsilon}_{.j}^2)$$

$$= \sum_i\sum_j\sigma_e^2 + h\sum_{i=1}^{k}\left(\frac{\sigma_e^2}{h}\right) + k\sum_{j=1}^{h}\left(\frac{\sigma_e^2}{k}\right) + hk\left(\frac{\sigma_e^2}{hk}\right) - 2h\sum_{i=1}^{k}\left(\frac{\sigma_e^2}{h}\right) - 2k\sum_{j=1}^{h}\left(\frac{\sigma_e^2}{k}\right)$$

$$\left[\because E(\varepsilon_{ij}^2) = \sigma_e^2, E(\bar{\varepsilon}_i.^2) = \frac{\sigma_e^2}{h}, E(\bar{\varepsilon}_{.j}^2) = \frac{\sigma_e^2}{k}, E(\bar{\varepsilon}_{..}^2) = \frac{\sigma_e^2}{hk}\right]$$

$$= hk\sigma_e^2 + k\sigma_e^2 + h\sigma_e^2 + \sigma_e^2 - 2k\sigma_e^2 - 2h\sigma_e^2$$

$$= (hk - k - 1)\sigma_e^2$$

$$= (h-1)(k-1)\sigma_e^2$$

$$E(S_E^2) = (h-1)(k-1)\sigma_e^2$$

$$\Rightarrow \quad E\left[\frac{S_E^2}{(h-1)(k-1)}\right] = \sigma_e^2$$

$$\Rightarrow \quad E(S_E^2) = \sigma_e^2$$

Thus, error M.S.S provides unbiased estimate of and σ_e^2 S_t^2, S_V^2 are unbiased estimates of σ_e^2 under the null hypotheses.

3. Replication.

Ans :

The process of repeating the same treatment on different experimental units under similar conditions is known as replication. It is concerned with no. of experimental units under each treatment.

Suppose that a treatment is allotted to r-experimental units, then it is said to be replicated r-no. of times and hence, the standard error of the mean is $\sigma/\sqrt{r}$.

Advantages

1. Replications are essential to get valid estimate of the experimental error.
2. The larger no. of replications reduces the standard error of treatments and hence, the precision of the experiment increases.
3. Replications give more precise estimates of treatment effects.

Disadvantages

1. Replications cannot be increased infinitely as it increases the cost of experimentation.
2. A large no. of replications cannot be taken as it causes the scarcity of the resources.
3. Replication of treatments may some times subject to bias.

4. Local Control.

Ans :

In experimental designs it is noticed that the error can enter at any stage of the experimentation, even so the replication and randomisation are greatly considered. Perhaps, this could be due to heterogeneous experimental material. Here, in agricultural field the whole heterogeneous experimental material is divided into relatively homogeneous sub groups or strata, on which the treatments are applied randomly with required no. of replications. Eventually, the error can also be controlled by means of local control.

"The process of reducing the experimental error by dividing the relatively heterogenous experimental area (field) into homogeneous blocks" is known as local control.

Advantages

1. Local control is meant to make the design more efficient.
2. Local control reduces the experimental error.
3. It makes any test of significance more sensitive and powerful.
4. A reduction in the experimental error consequently helps the investigator to detect the small real difference between the treatments.

Disadvantages

1. Soil fertility, an important factor which distributed randomly over the entire field, hence, making the blocks becomes difficult.
2. Some times the available experimental material itself homogenous, now making them into blocks is wasting the time and money.

5. Randomisation.

Ans :

A process in which the treatments are allocated to the experimental units in such a way that each and every experimental unit has got equal chance receiving any of the treatments is called randomisation.

Randomisation is efficiently implemented with the help of random number tables (or) tossing of a coin (or) dice. Randomization is a matter of allocating the treatments randomly to the experimental units.

Advantages

1. It eliminates human biases.
2. Randomisation leads to an unbiassed estimate of error variance and unbiased estimate of treatment differences.
3. Randomisation makes the experiment free from any systematic effects of environment.
4. Randomisation introduces the independence among the experimental errors, hence, F-test.

Disadvantages

1. Well trained persons are necessary to perform the technique of randomisation hence, it leads to more expenditure.
2. If there exists a large no. of experimental units, then performing randomisation becomes difficult.

6. Analysis of variance of two-way classification.

Ans :

Two way classification/two factor ANOVA is defined where two independent factors have an effect on the response variable of interest.

Example : Yield of crop affected by type of seed as well as type of fertilizer.

Procedure

(a) Calculate the variance between columns,

$$SSC = \sum_{j=1}^{c} \frac{T_j^2}{n_j} - \frac{T^2}{N}$$

(b) Calculate the variance between rows,

$$SSR = \sum_{i=1}^{r} \frac{T_i^2}{n_i} - \frac{T^2}{N}$$

(c) Compute the total variance,

$$SST = \sum X_{ij}^2 - \frac{T^2}{N}$$

(d) Calculate the variance of residual or error,

$$SSE = TSS - (SSC + SSR)$$

(e) Divide the variances of between columns, between rows and residue by their respective degrees of freedom to get the mean squares.

(f) Compute F ratio as follows,

F-ratio concerning variation between columns,

$$= \frac{\text{Mean square between columns}}{\text{Mean squares of residual}}$$

F-ratio concerning variation between rows,

$$= \frac{\text{Mean square between rows}}{\text{Mean squares of residual}}$$

(g) Compare F-ratio calculated with F-ratio from table,

If F-ratio (calculated) $<$ F-ratio (table), H_1 accepted,

If F-ratio (calculated) $\geq$ F-ratio (table), H_0 rejected,

H_1 accepted $\Rightarrow$ no significant differences

H_0 rejected $\Rightarrow$ significant differences

7. Assumptions and Applications of ANOVA.

Ans :

Analysis of variance test is based on the test statistic F (or variance ratio).

It is based on the following assumptions.

(i) Observation are independent.

(ii) Each sample is drawn randomly from a normal population as the sample statistics reflect the characteristic of the population.

(iii) Variance and means are identical for those population from which samples have been drawn.

Applications

The applications of ANOVA are as follows,

1. Anova is used in education, industry, business, psychology field mainly in their experiment design.
2. Anova helps to save time and money as several population means can be compared simultaneously.
3. Anova is used to test the linearity of the fitted regression line and correlation ratio, significance test statistic of anova = F(r – 1, n – r).

8. Importance of Design of Experiments.

Ans :

1. The design of experiments is meant for reducing the effect of errors due to expected sources and due to other sources.
2. The design of experiments is a must to improve the precision of the experiment.
3. Appropriate design of experiments is vital in researches where we need to accurately ascertain the statistical significance of various factors.
4. Design of experiments is imperative in business settings, where we need to determine an unbiased data for forecasting and for continuous improvement.
5. To organise the experiment properly the design of experiments is very required.

6. To Design of experiments is the backbone of any product design as well as any process/product improvement efforts.

9. Applications of Design of Experiments.

Ans :

1. Design of experiments and its applications are tremendously used in the field of agriculture, green house studies, laboratories, marketing etc.,
2. DOE is applied in the field of pharmaceutical statistics to analysis of drug trials and to issues of commercialization of a medicine.
3. The application of experimental methods is applied to experimental economics to study the various economic questions.
4. In clinical trials, DOE is applied for safety collection of efficacy data of new drugs (or) devices.
5. DOE is a tool that has been used by many industries for the purpose of optimizing processes.
6. DOE has variety of applications in the field of software packages and in military also in industry.

10. Efficiency of a Design.

Ans :

Consider the designs D_1 and D_2 with error variances per unit σ_1^2 and σ_2^2 and replications r_1 and r_2 respectively. Then the variance of the difference between two treatment means is given by

$\frac{2\sigma_1^2}{r_1}$ and $\frac{2\sigma_2^2}{r_2}$ for D_1 and D_2 respectively then the ratio.

$$E = \frac{2\sigma_2^2}{r_2} \cdot \frac{r_1}{2\sigma_1^2} = \frac{r_1}{\sigma_1^2} \div \frac{r_2}{\sigma_2^2}$$

is termed as efficiency of design D_1 w.r.t. D_2. In other words, efficiency of D_1 w.r.t D_2 may be defined as the "ratio of the precisions of D_1 and D_2".

If $E = 1$, then both the designs D_1 and D_2 are said to be equally efficient.

If $E > 0$ ($E < 1$) then D_1 is said to be more (less) efficient than D_2.

Choose the Correct Answer

1. The term 'Analysis of variance' was introduced by R.A fisher in the year ______. [b]
   (a) 1918's (b) 1920's
   (c) 1930's (d) 1950's

2. ______ is defined as the square of standard deviation (σ), root mean square deviation, and mean error. [d]
   (a) ANOVA (b) Degrees of freedom
   (c) Mean (d) Variance

3. Total sum of squares is calculated by ______. [a]
   (a) SI – CF (b) CF – SI
   (c) $\frac{T^2}{N}$ (d) $\frac{T}{N}$

4. The basic purpose of ANOVA is to test the ______ of several means [c]
   (a) Efficiency (b) Consistency
   3) Homogeneity (d) None of the above

5. S.S.A + S.S.E = ______. [a]
   (a) T.S.S (b) M.S.A
   (c) M.S.E (d) S.S.T

6. Mixed effect model is dependent on ______. [c]
   (a) Fixed effect model (b) Random effects model
   (c) Both (a) & (b) (d) None of the above

7. Correlation factor CF = ______. [d]
   (a) $\sum x$ (b) $\sum \frac{x}{n}$
   (c) $\frac{\sum x^2}{n^2}$ (d) $\frac{\left(\sum x\right)^2}{n}$

8. The following can be achieved by practising the principles of design of experiments. [d]
   (a) Replication (b) Randomization
   (c) Local control (d) All the above

9. In a Gauss-Markov linear model $E(\varepsilon_i)$ = ______ [a]
   (a) 0 (b) 1
   (c) ∞ (d) None of the above

10. In ANOVA ______ types of models are used. [c]
   (a) Three (b) Four
   (c) Two (d) Five

Fill in the blanks

1. The variance test is also known as _____.
2. The test is called 'F' test as it was developed by _____ .
3. An _____ is a device of getting an answer to the problem under consideration.
4. The various objects of comparison in a comparative experiment are known as _____.
5. The whole experimental area which is used for experimentation, is known as _____.
6. The process of repeating the same treatment on different experimental units under similar conditions is known as _____.
7. _____ are essential to get valid estimate of the experimental error.
8. _____ leads to an unbiassed estimate of error variance and unbiased estimate of treatment differences.
9. _____ is meant to make the design more efficient.
10. A reduction in the _____ error consequently helps the investigator to detect the small real difference between the treatments.

Answers

1. ANOVA
2. R.A Fisher
3. Experiment
4. Treatments
5. Experimental material.
6. Replication
7. Replications
8. Randomisation
9. Local control
10. Experimental

UNIT II	**Principles of experimentation:** Analysis of Completely randomized Design (C.R.D), Randomized Block Design (R.B.D) and Latin Square Design (L.S.D) including one missing observation, expectation of various sum of squares. Comparison of the efficiencies of above designs.

2.1 Principles of Experimentation

2.1.1 Analysis of Completely randomized Design (C.R.D) and Randomized Block Design (R.B.D)

Q1. Explain in detail about Completely randomized Design (C.R.D).

Ans : **(July-21)**

The completely randomised design is the simplest of all the designs, based on principles of randomisation and replication. In this design treaunents are allocated at random to the experimental units over the entire experimental material. Let us suppose that we have v treatments, the ith treatment being replicated r, times, i =1, 2, ..., v. Then the whole experimental material is divided into $n=\Sigma r_i$ experimental units and the treatments are distributed completely at random over the units subject to the condition that the ith treatment occurs r< times. Randomisation assures that extraneous factors do not continually influence one treatment. In particular case if

$$r_t = r \ \forall \ i = I,2,..., v$$

i.e., if each treatment is repeated an equal number of times r, then n=rv and randomisation gives every group of r units an equal chance of receiving the treatments. In general, equal number of replications for each treatment should be made except in particular cases when some treatments are of greater interest than others or when practical limitations dictate otherwise.

Advantages

1. C.R.D. results in the maximum use of the experimental units since all the experimental material can be used.
2. The design is very flexible. Any number of treatments can be used and different treatments can be used unequal number of times with-out unduly complicating the statistical analysis in most of the cases.
3. The statistical analysis remains simple if some or all the observations for any treatment are rejected or lost or missing for some purely random accidental reasons. We merely carry out the standard analysis on the available data. Moreover the loss of information due to missing data is smaller in comparison with any other design.
4. It provides the maximum number of degrees of freedom for the estimation of the error variance, which increases the sensitivity or the precision of the experiment for small experiments, i.e., for experiments with small number of treatments.

Disadvantages

1. CRD is used only on homogeneous blocks but in the field experimentation the available experimented material is always heterogeneous.
2. CRD is suitable when small no. of treatments are taken but it is not practicable.
3. In CRD the local control principle is not used, hence, it leads to more the experimental error.

Q2. Discuss in detail about statistical analysis of CRD.

Ans : **(July-22)**

Statistical analysis of a C.R.D. is analogous to the ANOVA for a one way classified data, the linear model (assuming various effects to be additive) becomes.

$$y_{ij} = \mu + \tau_i + \varepsilon_{ij} \begin{pmatrix} i = 1, 2, ..., v \\ j = 1, 2, ..., r_i \end{pmatrix}$$

where $y_{(i}$ is the yield or response from the jth unit receiving the with treatment, μ. is the general mean effect, τ_i- is the effect due to the with treatment, and ε_{ij} is error effect due to chance such that ε_{ij} a dentically and independently distributed (i. i. d.) N(0, σ_e^{2}). Then $n = 1 \sum_{i=1}^{v} r$, is the total number of experimental units. If we write

$$\sum_i \sum_j y_{ij} = y.,= G = \text{Grand total of all the n observations.}$$

$$\sum_{j=1}^{ri} y_{ij} = y_i = T_i = \text{Total resported the units incase of ith treatment. Then, as in ANOVA}$$

$$\text{Correction Factor, CF} = \frac{(G)^2}{n}$$

$$G = \text{Grand Total} \left(\sum_i \sum_j y_{ij} \right)$$

$$n = \text{Number of treatments} \left(\sum_{i=1}^{n} r_i \right)$$

$$\Rightarrow \quad T.S.S = \sum_i \sum_j (y_{ij} - \bar{y}..)^2 = \sum_i \sum_j y_{ij}^2 - \frac{G^2}{n}$$

$$\therefore \quad T.S.S = \sum_i \sum_j y_{ij}^2 - C.F$$

$$\Rightarrow \quad S.S.T = \sum_i r_i (\bar{y}_i. - \bar{y}..)^2$$

$$\therefore \quad S.S.T = \sum_i \left(\frac{T_i^2}{r_i} \right) - C.F$$

$$\Rightarrow \quad S.S.T = T.S.S - S.S.T$$

$$= \sum_i \sum_j y_{ij}^2 - C.F - \left(\sum_i \left(\frac{T_i^2}{r_i} \right) - C.F \right)$$

$$\therefore \quad S.S.E = \sum_i \sum_j y_{ij}^2 - \sum_i \left(\frac{T_i^2}{r_i} \right)$$

ANOVA Table for C.R.D

Source of variation	d.f	S.S.	M.S.S	Variance ratio
Treatment	v – 1	S_T^2	$s_T^2 = \frac{S_T^2}{(v-1)}$	
Error	n – v	S_E^2	$s_E^2 = \frac{S_E^2}{(n-v)}$	$F_T = \frac{s_T^2}{s_E^2}$
Total	n - 1	$S_T^2 + S_E^2$		

Under the null hypothesis $H_0 : \tau_1 = \tau_2 = ... = \tau_v$ against the alternative that all τ's are not equal, the statistic,

$$F_T = \frac{S_T{}^2}{S_E{}^2} \sim F(v-1, n-v)$$

i.e., F_T follows F (central) distribution with (v – 1, n – v) d.f.

If $F_T > F_{\alpha,(v-1, n-v)}$, then H_0 is refuted at α% level of significance and we conclude that treatments differ significantly. If $F_T < F_{\alpha,(v-1, n-v)}$, H_0 may be accepted, i.e., the data do not provide any evidence to prefer one treatment to the other and as such all of them can be considered alike.

Remrak. The following formulae for the calculation of various S.S. are maeh convenient to use from practical point of view :

$$\text{T.S.S.} = \sum_i \sum_j \left(y_{ij} - \overline{y..}\right)^2 = \sum_i \sum_j y_{ij}{}^2 - n\bar{y}..^2$$

$$= \sum_i \sum_j y_{ij}^2 - \frac{G^2}{n}, \qquad n = \sum_{i=1}^{v} r_i$$

$\therefore$ T.S.S. = Raw S.S. – Correction Factor = R.S.S. – C.F.,

$$\text{S.S.T.} = \Sigma r_i(\bar{y}_i, - \bar{y}..)^2$$

$$= \Sigma r_i(\bar{y}_i{}^2, - n\bar{y}..^2$$

$$= \sum_{i=1}^{v} \left(\frac{T_i^2}{r_i}\right) - \text{C.F.}$$

$\therefore$ S.S.E. = T.S.S. – S.S.T.

$$= \sum_i \sum_j y_{ij}^2 - \sum_i \left(\frac{T_i^2}{r_i}\right)$$

$$= \Sigma_i \Sigma_j \, y_{ij}^2 - \Sigma_i r_{\backslash i} \bar{y}_i^2$$

PROBLEMS

1. A set of data involving Jour "tropical feed stuffs A, B, C, D" tried on 20 chicks is given below. All the twenty chicks are treated alike in all respects except the feeding treatments and each feeding treatment ts given to 5 chicks. Analyse the data.

Weight gain of baby chicks fed on different feeding materials composed of tropical feed stuffs :

Feed	Gain in weight					Total T_i
A	55	49	42	21	52	219
B	61	112	30	89	63	355
C	42	97	81	95	92	407
D	169	137	169	85	154	714
					Grand Total	G = 1,695

Sol :

$H_0 : \tau_A = \tau_B = \tau_C = \tau_D$

i.e., the treatment effects are same. In other words, all the treatments {A, B, C, D) are alike as regards their effect on increase in weight.

Calculations. Raw S.S.(R.S.S.) = $\sum_i \sum_j y_{ij}^2 = 55^2 + 49^2 + ... + 85^2 + 154^2$

$= 1,81,445$

Correction factor (C.F.) = $\frac{G^2}{N} = \frac{(1695)^2}{20} = 1,43,651.25$

$\therefore$ Total S.S (T.S.S.) = R.S.S. – C.F.

= 181445 – 143651.25 = 37,793.75

$$\text{Treatment S.S.} = \frac{T_1^2 + T_2^2 + T_3^3 + T_4^4}{5} - \text{C.F.}$$

$$= \frac{47961 + 126025 + 165649 + 509796}{5} - 1,43,651.25$$

= 26,234.95

Error S.S. = Total S.S. – Treatment S.S.

= 37793.75 – 26234.95 = 11,558.80

AVONA Table

Source of variation	**S.S**	**d.f.**	**M.S.S.** = $\frac{\textbf{S.S.}}{\textbf{d.f.}}$	**Variance ratio 'F'**
Treatments	26234.95	3	8744.98	$F_T = \frac{8744.98}{722.42} = 12.105$
Error	11558.80	16	722.42	
Total	37793.75	19		

$F_T \sim F(3, 16)$

Tabulated $F_{0.05}$ (3, 16) = 3.06

Hence Fj is highly significant and we refute H_0 at 5% level of significance and conclude that the treatments A, B, C and D differ significantly.

Since H_0 is rejected in this case, we proceed further to find out which of the treatment means differ significantly. For this we find out the 'critical difference' (C.D.), i.e., the least difference between any two means to be significant.

S.E. of the difference between any two treatment means is

$$S_E\sqrt{\frac{2}{r}} = \sqrt{\frac{2\times 722.42}{5}} = 16.99$$

$$\therefore \quad C.D. = \sqrt{\frac{2S_E^2}{r}} \times t_{0.05} \text{ for error d.f.} = 16.99 \times 2.12 = 36.018$$

The treatment mean effects, arranged in descending order of magnitude, are given below :

Treatment	Mean gain in weight	Difference	
D	$\frac{714}{5} = 142.8$		
C	$\frac{407}{5} = 81.4$	61.4*	
B	$\frac{315}{5} = 71.0$	10.4	37.6*
A	$\frac{219}{5} = 43.8$	27.2	

Comparing these differences with the C.D., we find that

(i) Treatment D differs significantly from each of the treatments A, B and C.

(ii) The treatments A and C also differ significantly and

(iii) All the remaining differences are not significant.

Conclusions

(i) Treatments A, B, C and D are not alike.

(ii) The highest treatment mean effect is 142·8 due to the feedstuff D.

(iii) Hence if a choice is to be made among the four treatments A, 3, C and D, treatment D is the best and most effective. Moreover, if a choice is to be made between A and C (which differ significantly), the treatment C is to be preferred since the average gain in weight due to treatment C is more than due to the treatment A. All other possible combinations of treatment pairs are alike.

Q3. Discuss about Randomized Block Design (R.B.D).

Ans : **(July-21, June-19)**

If the whole experimental material is not homogeneous then divide the experimental area into relatively homogeneous blocks and then allocate the treatments in each block randomly. The designs so obtained is called randomised block design.

Let us suppose that the experimental area is divided into *l* homogeneous blocks and there be k treatments.

Therefore divide each block into k subdivisions and allocate the k treatments randomly to each block. This design is based on all the three principles. i.e., the principle replication randomization and local control.

Layout

Let us suppose that there be four treatments A, B, C, D and the whole experimental material is divided into three relatively homogeneous blocks then one of the possible layout of RBD is

Block I	Block II	Block III
A	D	D
B	B	C
C	C	A
D	A	B

Advantages

1. This design is completely flexible to have any no. of treatments and blocks.
2. It provides more accurate results than CRD due to grouping.
3. The statistical analysis is simple and less time consuming.
4. Relatively easy statistical analysis even with missing data.
5. The number of replicates provides enough degrees of freedom to the error sum of squares.

Disadvantages

1. The design is not suitable for large no. of treatments because blocks become too large.
2. It is not suitable when complete block contains considerable variability.
3. In many situations the criterion of blocking is not easily selectable.
4. RBD is not suitable to control a two way variability.
5. The error increases if there exists interaction between block and treatment.

Applications

RBD is one of the most widely used designs in field experimentation, green house studies, laboratories etc.,

Q4. Discuss in detail the statistical analysis of RBD.

Ans : **(June-19)**

The statistical analysis of R.B.D is performed similar to the ANOVA for a two way classified data (with one observation per cell) for fixed effect model. Then, the linear model in case of additive effects will be,

$$y_{ij} = \mu + \tau_i + b_i + \varepsilon_{ij} \begin{cases} i = 1, 2, 3, ...t \\ j = 1, 2, 3, ...r \end{cases}$$

Where,

y_{ij} = Response or yield of the experimental unit from/"'WeclLin case of j^{th} treatment.

τ_i = Effect in case of i^{th} treatment

μ = General mean effect

b_j = Effect in case of replicate or j^{th} block

ε_{ij} = Error effect which is identically independently distributed N(0, σ_e^2)

Also, μ, τ_i's and b_j's are constants. Therefore,

$$\sum_{i=1}^{t} \tau_i = 0, \ \sum_{i=1}^{r} b_j = 0$$

Let us consider the following,

General total of the entire t x r observation i.e.,

$$G = \sum_i \sum_j y_{ij} = y..$$

Total for treatment i.e., $T_i = \sum_j y_{ij} = y_i.$

Total for j^{th} block i.e., $B_j = \sum_i y_{ij} = y_{ij}$

Then, the total sum of squares is given by,

$$\sum_i \sum_j (y_{ij} - \bar{y}...)^2 = \sum_i \sum_j \left[(\bar{y}_i. - \bar{y}..) + (\bar{y}._j - \bar{y}...) + (\bar{y}_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}..)\right]^2$$

$$= r\sum_i (\bar{y}_{i.} - \bar{y}..)^2 + t\sum_j (\bar{y}_{.j} - \bar{y}..)^2 + \sum_i \sum_j (y_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}..)^2$$

Here, the product terms get eliminated as the algebraic sum of deviations from mean is zero.

Where,

$\sum_i \sum_j (y_{ij} - \bar{y}..)^2$ = Total Sum of Squares (T.S.S)

$r\sum_i (\bar{y}_{i.} - \bar{y}..)^2$ = Sum of Squares due to Treatments (S.S.T)

$t\sum_j (\bar{y}_{.j} - \bar{y}..)^2$ = Sum of Squares due to Blocks (S.S.B)

$\sum_i \sum_j (y_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}..)^2$ = Sum of Squares due to Error (S.S.E) = T.S.S – (S.S.T + S.S.B)

Analysis of Variance or Variance Analysis

Consider the example of 3×4 randomized block design containing four treatments with three replications.

Blocks/Replications		Plots / units / Treatments				
	Treatment	**1**	**2**	**3**	**4**	**Total**
	1	a_1	b_1	c_1	d_1	B_1
	2	d_2	a_2	b_2	c_2	B_2
	3	c_3	d_3	a_3	b_3	B_3
	Total	A_1	A_2	A_3	A_4	A

The three sources of variations which may arise in RBD are either due to the block or due the treatment or due to the experimental error. These variations can be computed from the above table.

1. Correction Factor (CF)

$$CF = \frac{(\text{final total})}{m \times n} = \frac{(A)^2}{3 \times 4}$$

Where,

m – Number of treatments (columns)

n – Number of blocks (rows)

2. Total Sum of Square (T.S.S)

$$T.S.S = (a_1^2 + b_1^2 + c_1^2 + d_1^2 + a_2^2 + b_2^2 + c_2^2 + d_2^2 + a_3^2 + b_3^2 + c_3^2 + d_3^2) - CF$$

3. Sum of Squares Due to Treatments (S.S.T)

$$S.S.T = \frac{A_1^2 + A_2^2 + A_3^2 + A_4^2}{\text{Number of treatment (m)}} - CF$$

4. Sum of Squares Due to Block (S.S.B)

$$S.S.B = \frac{B_1^2 + B_2^2 + B_3^2}{\text{Number of blocks (n)}} - CF$$

5. Sum of Squares Due to Errors (S.S.E)

S.S.E = T.S.S – (S.S.T + S.S.B)

Source of variation	**S.S**	**d.f.**	**M.S.S.** $= \frac{\text{S.S.}}{\text{d.f.}}$	**Variance Ratio** $F_{calculated}$
Block (replication)	r – 1	S_B^2	$s_B^2 = \frac{S_B^2}{r-1}$	$F_B = \frac{s_B^2}{s_E^2}$
Treatment (Column)	t – 1	S_T^2	$s_T^2 = \frac{S_B^2}{t-1}$	$F_T = \frac{s_T^2}{s_E^2}$
Error	(r – 1) (t – 1)	S_E^2	$s_E^2 = \frac{S_E^2}{(t-1)(r-1)}$	
Total	**rt – 1**			

Table : ANOVA for carrying F-test

Suppose,

$H_{0t} : \tau_1 = \tau_2 = ... = \tau_t$ and

$H_{1t} : \tau_1 \neq \tau_2 \neq ..... \neq \tau_t$ and

Then, the test statistic for treatments

$$F_T = \frac{s_T^2}{s_E^2} \sim F[(t-1), (t-1)(r-1)]$$

If $F_T > F_{[(0-1), (t-1)(r-1)]-1)]}$ then, the null hypothesis H_{0t} is rejected otherwise, H_{0t} is accepted.

Similarly,

$H_{0b} : b_1 = b_2 = ... = b_r$ and

$H_{1t} : b_1 \neq b_2 \neq .... \neq b_r$

Then, the test statistic for blocks

$$F_B = \frac{s_B^2}{s_E^2} \sim F[(r-1), (r-1)(t-1)]$$

If $F_B > F_{[(r-1)(r-1)(r-1)]}$ then H_{0b} is rejected. Otherwise, H_{0b} is accepted.

Q5. Discuss in detail about R.B.D with random effect model.

Ans :

The R.B.D with random effect model allows the selection of treatment from the given large amount of treatments. It is defined as,

$$y_{ij} = \mu + \tau_i + b_j + \varepsilon_{ij} \; ; \begin{cases} i = 1, 2, 3, ...t \\ j = 1, 2, 3, ...r \end{cases}$$

Where,

μ = General mean effect

y_{ij} = Yield of the experimental units in j[th] block in case of i[th] treatment

τ_i = Effect in case of i[th] treatment

b_j = Effect incase of replica or fh block

ε_{ij} = Error effect which is identically independently distributed τ_i i.i.d N(0, σ_e^2) and ε_{ij} i.i.d N(0, σ_e^2) and b_j i.i.d N(0, σ_e^2) are not dependent on each other.

Statistical Analysis of R.B.D With Random Effect Model

Let A. B be two sources of variation where A has 'p' levels and B has 'q' levels i.e., A_i, i = 1,2,....t and B_j, j = 1,2,...r respectively. These levels of A, B are selected randomly from large population.

The linear mathematical model for Random Effect is as follows,

$y_{ij} = \mu + \tau_i + b_j + c_{ij} + e_{ij}$, where y.. represents the observation under i[th], j[th] level of A, B respectively

Here,

t_1 i.i.d rt(0, σ_t^2)

b_j i.i.d rt(0, σ_b^2)

c_{ij} i.i.d rt(0, σ_c^2)

ε_{ij} i.i.d rt(0, σ_e^2)

Where,

i = 1, 2, 3. ....t

j = 1, 2, 3, ... r

μ = Denotes general mean effect

τ_i, b_j = Denotes the additional effect due to i^{th} level of A and j^{th} level of B respectively

ε_{ij} = Denotes random error effect

c_{ij} = Denotes interaction effect.

Null Hypothesis

Null hypothesis is considered as the appropriate hypothesis which can be used to test the equality of all the class means. The equality is checked due to the different levels of treatments and blocks. The hypothesis is as follows,

(i) $H_{04} : \sigma_\tau^2 = 0$ against $H_{14} : \sigma_\tau^2 > 0$

(ii) $H_{1B} : \sigma_b^2 = 0$ against $H_{1B} : \sigma_b^2 > 0$

Q6. How can you estimate the missing observations in RBD?

Ans : **(Dec.-21, Oct.-20, June-19)**

Consider the table of R.B.D containing or missing observation in case of i^{1h} treatment and j^{th} block. Let the missing observation be $a = y_{ij}$

← Blocks →

Treatments	1	2	…	j	…	r	Total
1	y_{11}	y_{12}	…	y_{1j}	…	y_{1r}	$y_{1}.$
2	y_{21}	y_{22}	…	y_{2j}	…	y_{2r}	$y_{2}.$
⋮	⋮	⋮		⋮		⋮	⋮
i	y_{i1}	y_{i2}	…	a	…	y_{ir}	$y_i' + x$
⋮	⋮	⋮		⋮		⋮	⋮
t	y_{t1}	y_{t2}	…	y_{tj}	…	y_{tr}	$y_{t}.$
Total	$y._1$	$y._2$	…	$y._{j'+x}$	…		$y'.. + x$

Table: R.B.D with One Missing Observation

Where,

$y'_i.$ denotes the total of all known observations in case of i^{th} treatment.

$y.'_j$ denotes the total of all known observations in case of j^{th} block.

$y..'$ denotes the total of all known observations.

Then,

Sum of Squares due to Treatments (S.S.T) $= \frac{1}{r}[(y'_{i.} + a)^2 + \text{constant with respect to a}] - C.F$

Sum of Squares due to Blocks (S.S.B) $= \frac{1}{t}[(y.'_j + a)^2 + \text{constant with respect to a}] - C.F$

Total Sum of Squares (T.S.S) $= \sum\sum y_{ij}^2 - C.F$

$= a^2 + \text{constant with respect to a} - C.F$

Where C.F $= \frac{(y'.. + a)^2}{rt}$

E = Residual Sum of Square,

Residual SS = T.S.S – S.S.B – S.S.T

$$E = a^2 - \frac{1}{t}(y.'_j + a)^2 - \frac{1}{r}(y'_i. + a)^2 + \frac{(y'.. + a^2)}{rt} + \text{constant terms independent of 'a'.}$$

The value of 'a' should be chosen in such away that E has minimum value which can be achieved when E is differentiated with respect to 'a'

i.e., $\frac{\partial E}{\partial a} = 0$

$$\therefore \frac{\partial\left(a^2 - \frac{1}{t}(y.'_j + a)^2 - \frac{1}{r}(y'_{i.} + a)^2 + \frac{(y'.. + a)^2}{rt} + \text{constant terms}\right)}{\partial a} = 0$$

$$\Rightarrow 2a - \frac{2}{t}(y.'_j + a) - \frac{2}{r}(y'_{i.} + a) + \frac{2}{rt}(y'.. + a) + 0 = 0$$

$$\Rightarrow 2a - \frac{2y.'_j}{t} - \frac{2a}{t} - \frac{2y'_i.}{r} - \frac{2a}{r} + \frac{2y'..}{rt} + \frac{2a}{rt} = 0$$

$$\Rightarrow 2a - \frac{2a}{t} - \frac{2a}{r} + \frac{2a}{rt} = \frac{2y'_{.j}}{t} + \frac{2j'_i.}{r} - \frac{2y'..}{rt}$$

$$\Rightarrow 2a\left(1 - \frac{1}{t} - \frac{1}{r} + \frac{1}{rt}\right) = 2\left(\frac{y'_{.j}}{t} + \frac{y'_i.}{r} - \frac{y'..}{rt}\right)$$

$$\Rightarrow a - \left(1 - \frac{1}{t} - \frac{1}{r} + \frac{1}{rt}\right) = \frac{y'._j}{t} + \frac{y'_i.}{r} - \frac{y'..}{rt}$$

$$\Rightarrow \quad a\left(\frac{rt - r - t + 1}{rt}\right) = \frac{ry'_{.j} + ty'_{i.} - y'..}{rt}$$

$$\Rightarrow \quad a\ (r(t - 1) = t + 1) = ry'_{.j} + ty'_{i.} - y..$$

$$\Rightarrow \quad a(r(t - 1) - (t - 1)) = ry'_{.j} + ty'_{i.} - y.'.$$

$$\Rightarrow \quad a\ ((r - 1)\ (t - 1)) = ry'_{.j} + ty'_{i.} - y'..$$

$$\boxed{\therefore a = \frac{ry'_{.j} + ty'_{i.} - y'..}{(r-1)(t-1)}}$$

Statistical Analysis

In general, ANOVA is performed only after the substitution of the estimated values of the missing observations. For each and every missing observation one degree of freedom (d.f) is subtracted from total and consequently from error d.f. The adjusted treatment s.s. is obtained by performing the subtraction of so called adjustment factor

$$\frac{\left[y'_{.j} + ty'_{i.} - y'..\right]^2}{\left[t(t-1)(r-1)^2\right]}$$ from the treatment s.s.

2.2 Latin Square Design (L.S.D) Including one Missing Observation

Q7. Define Latin Square Design (L.S.D). State its merits and demerits.

Ans : **(July-21, June-19)**

In RBD whole of the experimental area is divided into relatively homeogeneous groups (blocks) and treatments are allocated at random to units within each block, i.e., randomisation was subjected to one restriction, i.e., within blocks. But in field experimentation, it may happen that experimental area (field) exhibits fertility in strips, eg., cultivation might result in alternative strips of high or low fertility. R.B.D will be effective if the blocks happen to be parallel to these strips and would be extremely inefficient if the blocks are across the strips.

Initially, fertility gradient is seldom known a useful method of eliminating fertility variations consists in an experimental layout which will control variation in two perpendicular directions such a layout in a Latin square design (L.S.D.).

Layout of Design

In field plot experiments, the Latin square is usually laid out in the conventional square with the rows and columns corresponding to possible fertility trends in two directions across the field. In other types of experiments, the rows and columns may be made to correspond to different sources of error as in animal feeding experiment where the column groups may correspond with initial weight and the row group with age.

In this design the number of treatments is equal to the number of replications. Thus in case of m treatments, there have to be $m \times m = m^2$ experimental units. The whole of experimental area is divided into m^2 experimental units (plots) arranged in a square, so that each row as well as each column contains m units (plots).

The m treatments are then allocated at random to these rows and columns in such a way that every treatments occurs once and only once in each row and in each column. Such a layout is known as $m \times m$ latin square design (L.S.D) and is extensively used in agricultural experiments.

For examples if we are interested in studying the effects of m types of fertilizers on the yield of a certain variety of wheat, it is customary to conduct the experiments on a square field with m^2 - plots of equal area and to associate treatments with different fertilizers and row and column effects with variations in fertility of soil.

Obviously, there can be many arrangements for an $m \times m$ L.S.D. and a particular layout in an experiment must be determined randomly.

For 2×2 and 3×3 latin squares, only one standard square exists.

A	B
B	A

A	B	C
B	C	A
C	A	B

For a 4×4 latin square design, 4 standard squares are possible, one of the design is

A	B	C	D
B	C	D	A
C	D	A	B
D	A	B	C

Advantages

- LSD controls more of the variation than CRD and RBD.
- The statistical analysis is simple even though slightly complicated than RBD.

Disadvantages

- In LSD the number of treatments are restricted to the number of replications therefore it is not a flexible design.
- In case of missing plots when several units are missing than the statistical analysis becomes complicate.
- In field layout RBD is much easy to manage the LSD.

Q8. Explain the Statistical Analysis of LSD.

Ans : **(June-19)**

Let Y_{ijk} (i,j, k = I, 2,..., m) denote the response from the unit (plot, in field experimentation) in the ith row, jth column and receiving the kth treatment. The triplet (i, j, k) assumes only m^2 of the possible m^3 values of an L.S. selected by the experiment. If S represents the set of m^2 values then symbolically (i, j, k)$\in$S. If a single observation is made per experimental unit then the linear additive model is

$$y_{ijk} = \mu + \alpha_i + \beta_j + \tau_k + \varepsilon_{ijk}, \text{ (i. j. k)} \in S$$

where μ is the constant mean effect;

α, β, and τ_k are the effects due to the fth row.

yth column and kth treatment respectively and

ε_{ijk} is error effect due to random component assumed to be normally distributed with mean zero and variance σ_e^2, i.e, $\varepsilon_{ijk} \sim N(0, \sigma_e^2)$. If we write

$G = y... =$ Total of all the m^2 observations

$R_i = y_i.. =$ Total of the m observations in the ith row

$C_j = y_{.j.} =$ Total of the m observations in the jth column

$T_k = y._k =$ Total of the m observations from fcth treatment, then heuristically, we have

$$\sum_{i,j,k \in S} (y_{ijk} - y....)^2 = \sum_{i,j,k \in S} [(\bar{y}_i.. - \bar{y}...) + (\bar{y}_{.j.} - \bar{y}...) + \bar{y}_{..k} - \bar{y}...) + (y_{ijk} - \bar{y}_i..\bar{y}_{..k} + 2y...)]^2$$

$$= m\sum_i (\bar{y}_i.. - \bar{y}...)^2 + m\sum_j (\bar{y}_{.i.} - \bar{y}...)^2$$

$$+ m\sum_k (\bar{y}_{..k} - \bar{y})^2 + \sum_{i,j,k \in S} (y_{ijk} - \bar{y}_i... - \bar{y}_{.i.} - \bar{y}_{..k} + 2\bar{y}...)^2$$

the product terms vanish, since the algebraic sum of deviations from mean is zero.

$\therefore$ T.S.S. = S.S.R. + S.S.C. + S.S.T. + S.S.E.

where T.S.S. is the total sum of squares and S.S.R., S.S.C., S.S.T. and S.S.E. represent sum of squares due to rows, columns, treatments and error -respectively, given by

T.S.S. = $\sum_{i,j,k \in S} (y_{ijk} - \bar{y}...)^2$; S.S.R. = $S_R^2 = m\sum_i (\bar{y}_i... - \bar{y}...)^2$

S.S.C. = $S_C^2 = m\sum_j (\bar{y}_{.j.} - \bar{y}...)^2$; S.S.T. = $S_T^2 = m\sum_k (\bar{y}_{..k} - \bar{y})^2$, and

S.S.E. = S_E^2 = T.S.S. – S.S.R. – S.S.C. – S.S.T.

ANOVA TABLE FOR m × m L.S.D.

Source of Variation	d. f.	S.S.	M.S.S.	Variance Ratio 'F'
Rows	m –1	S_R^2	$^SR^2 = S_R^2/(m-1)$	$F_R = {^SR^{2/S}E^2}$
Columns	m – 1	S_C^2	$^SC^2 = S_C^2/(m-1)$	$F_C = {^SC^{2/S}E^2}$
Treatments	m – 1	S_T^2	$^ST^2 = S_T^2/(m-1)$	$F_T = {^ST^{2/S}E^2}$
Error	(m – l)(m – 2)	S_E^2	$^SE^2 = S_E^2/(m-1)(m-2)$	
Total	$m^2 - 1$			

Let us set up the null hypotheses

for row effects, $H_\alpha : \alpha_1 = \alpha_2 = ... = \alpha_m = 0$,

for column effects, $H_\alpha : \beta_1 = \beta_2 = ... = \beta_m = 0$ and

for treatment effects, $H_\tau : \tau_1 = \tau_2 = ... = \tau_m = 0$,

The variance ratios F_R, F_C and F_T follow (central) F distribution with (m –1), (m –1) (m – 2) d.f. under the null hypotheses H_α, H_β and H_τ respectively.

Let $F_\alpha = F_\alpha$ {(m –1), (m –1) (m – 2)} be the tabulated value of F for (m – 1) (m – 2) d.f. at the level of significance 'α'. Thus if $F_R > F_\alpha$ we reject H_α and if $F_R < F_\alpha$ we may accept H_α.

Similarly we can test for H_β and H_T .

Remarks 1.

For numerical computations of various sum of square the following formulae are much more convenient to use.

$$\text{T.S.S.} = \sum_{i,j,k} (y_{ijk} - \bar{y}...)^2 = \sum_{i,j,k} y_{ijk}^2 - m^2\ \bar{y}...^2$$

$$= \Sigma y_{ijk}^2 - m^2\left(\frac{y...}{m^2}\right)^2 = \Sigma y_{ijk}^2 = \frac{y...^2}{m^2}$$

$$= \Sigma y_{ijk}^2 - \frac{G^2}{N},\ N = m^2$$

i.e., T.S.S. = Raw S.S. – C.F.

$$\text{S.S.R.} = S_R^2 = m\sum_i (\bar{y}_i.. - \bar{y}... = m[\sum_i \bar{y}_i..^2 - m\bar{y}...^2]$$

$$= m\left[\sum_i\left(\frac{y_i..}{m}\right)^2 - m\left(\frac{y...}{m^2}\right)^2\right]$$

$$= \frac{1}{m}\sum_i R_i^2 - \text{C.F.} \qquad (R_i = y_i..)$$

Similarly, we shall get

$$\text{S.S.C.} = \sum_i\left(\frac{y_{.j.}^2}{m}\right) - \text{C.F.} = \frac{1}{m}\sum_i C^2 - \text{C.F,}$$

S.S.T. = T.S.S. – S.S.R. – S.S.C. – S.S.T.

2. Standard Error (S.E.) of the difference between any two treatment means is $\left[S_E^2\left(\frac{1}{m}+\frac{1}{m}\right)\right]^{\frac{1}{2}} = \left(2S_E^2 / m\right)^{\frac{1}{2}}$ and the critical difference (C.D.) for significance of difference between any two treatment means at r el of significance 'α' is

$$\text{C.D.} = t_\alpha \text{ (for error d.f.)} \times \text{S.E.}$$

$$= [t_\alpha \text{ for } (m - l)(m - 2)\text{d.f.}] \times \left(2S_E^2 / m\right)^{\frac{1}{2}}$$

Q9. How do you estimate the missing observations in LSD.

Ans : **(July-22, July-21, Oct.-20)**

Let us suppose that in mxm Latin Square, the observation occuring in the ith row, jth column and receiving the kth treatment is missing. Let us assume that its value is x, i.e., $y_{ijk} = x$.

R = Total of the known observations in the ith row, i.e., the row containing 'x'.

C = Total of known observations in the jth column, i.e:, the column containing 'x'.

T = Total of known observations receiving kth treatment, i.e., total of all known treatment values containing 'x'.

S = Total of known observations.

Then

$$T.S.S. = y^2 + \text{constants w.r.t. } x - \frac{(S+x)^2}{m^2}$$

$$S.S.R. = \frac{(R+x)^2}{m} + \text{constant w.r.t. } x - \frac{(S+x)^2}{m^2}$$

$$S.S.C. = \frac{(C+x)^2}{m} \text{ h constant w.r.t. } x - \frac{(S+x)^2}{m^2}$$

$$S.S.T. = \frac{(T+x)^2}{m} + \text{constant w.r.t. } x - \frac{(S+x)^2}{m^2}$$

$\therefore$ E = Residual Sum of Squares (S.S.E.)

= T.S.S. – S.S.R. – S.S.C. – S.S.T.

$$= x^2 - \frac{1}{m}[(R+x)^2 + (C+x)^2 + (T+x)^2] + 2\frac{(S+x)^2}{m^2}$$

We will choose x so as to minimise E.

$$\therefore \quad \frac{\partial E}{\partial x} = 0 = 2x - \frac{2}{m}[R + C + T + 3x] + \frac{4(S+x)}{m^2}$$

$$\Rightarrow \quad (m^2 - 3m + 2)x = m(R + C + T) - 2S$$

$$\Rightarrow \quad x = \frac{m(R+C+T)-2S}{(m-1)(m-2)}$$

Remark

The same procedure may be followed for estimating more than one, say k missing values and then missing values are obtained by solving k-equations simultaneously.

Statistical Analysis. After inserting the estimated value for missing observation, we perform the usual analysis of variance, substracting one , d.f. for total S.S. and consequently for Error S.S. Adjusted treatment S.S. is obtained by subtracting the quantity

$$[(m-1)T + R + C - S]^2 \; [(m-1)(m-2)]^2$$

from the treatment S.S.

Standard Error (S.E.) of the difference between two treatment means, none of which corresponds to missing values is given by $S_E\sqrt{2/m}$. The S.E. of difference of the means of two treatments, one of which corresponds to missing observation is given by

$$S_E\left[\frac{2}{m} + \frac{1}{(m-1)(m-2)}\right]^{\frac{1}{2}}$$

provided the treatments show significant effect.

PROBLEMS

2. In a Random Block Design there are only two blocks. Let k be the number of treatments x-and Xi and X_2 be the average yield of the two blocks respectively. Show that the between blocks sum of squares is given as $\frac{k}{2}\left(\bar{X}-\bar{X}_2\right)^2$

Sol :

Given that,

Number of blocks = 2

Number of treatments = 1, 2, 3,.... k

Average yield of block 1 = $\bar{X}_1$

Average yield of block 2 = $\bar{X}_2$

The table for the given data is given as shown below,

		Blocks		
		Block 1	Block 2	Total
Treatment	1	x_{11}	x_{12}	T_1
	2	x_{21}	x_{22}	T_2
	3	x_{31}	x_{32}	T_3
	⋮			
	k	x_{k1}	x_{k1}	T_k
Total		$T._1$	$T._2$	ΣT_{ij} = T.. = G

$$\frac{T_{.1}}{k} = \bar{X}_1 \Rightarrow T_{.1} = k\bar{X}_1$$

$$\frac{T_{.2}}{k} = \bar{X}_2 \Rightarrow T_{.2} = k\bar{X}_2$$

N = 2Xk = 2k

$$\sum T_{ij} = T.. = T_{.1} + T_{.2}$$

$$= k\bar{X}_1 + k\bar{X}_2 \qquad [\because \text{ From equations (1) and (2)}]$$

$$= k(\bar{X}_1 + \bar{X}_2)$$

$$\text{C.F} = \frac{G^2}{N} = \frac{T^2}{N} = \frac{T..^2}{2K}$$

Sum of squares = $T_{.1}^2 + T_{.2}^2$

Between block sum of squares $= \dfrac{\text{Sum of squares}}{k} - C.F$

$$= \frac{T_{.1}^2 + T_{.2}^2}{k} - \frac{T..^2}{2k}$$

$$= \frac{T_{.1}^2 + T_{.2}^2}{k} - \frac{\left[k\left(\bar{X}_1 + \bar{X}_2\right)\right]^2}{2k}$$

$$= \frac{\left(k\bar{X}_1\right)^2 + \left(k\bar{X}_2\right)^2}{k} - \frac{\left[k\left(\bar{X}_1 + \bar{X}_2\right)\right]^2}{2k}$$

$$= \frac{1}{k}\left[k^2\bar{X}_1^2 + k^2\bar{X}_2^2 - \frac{\left(\bar{X}_1 + \bar{X}_2\right)^2}{2}\right]$$

$$= \frac{k^2}{k}\left[\bar{X}_1^2 + \bar{X}_2^2 - \frac{\left(\bar{X}_1 + \bar{X}_2\right)^2}{2}\right]$$

$$= \frac{k\left[2\left(\bar{X}_1^2 + \bar{X}_2^2\right) - \left(\bar{X}_1 + \bar{X}_2\right)^2\right]}{2}$$

$$= k\frac{\left[2\left(\bar{X}_1^2 + \bar{X}_2^2\right) - \left(\bar{X}_1 + \bar{X}_2\right)^2\right]}{2}$$

$$= \frac{k}{2}\left[2\bar{X}_1^2 + 2\bar{X}_2^2 - \left(\bar{X}_1^2 + \bar{X}_2^2 + 2\bar{X}_1\bar{X}_2\right)\right]$$

$$= \frac{k}{2}\left[2\bar{X}_1^2 + 2\bar{X}_2^2 - \bar{X}_1^2 - \bar{X}_2^2 + 2\bar{X}_1\bar{X}_2\right]$$

$$= \frac{k}{2}\left[\bar{X}_1^2 + \bar{X}_2^2 - 2\bar{X}_1\bar{X}_2\right]$$

Between block sum of squares $= \dfrac{k}{2}\left[\bar{X}_1 - \bar{X}_2\right]^2$

3. In the below table are the yields of 6 varieties in a 4 replicate experiment for which one value is missing. Estimate, the missing value and analyze the data.

Block	1	2	3	4	5	6	Block Totals (B_1)
1	18.5	15 7	16.2	14.1	13.0	13.6	91.1
2	11.7	–	12.9	14.4	16.9	12.5	68.4
3	15.4	16.6	15.5	20.3	18.4	21.5	107.8
4	16.5	18.6	12.7	15.7	16.5	18.0	98.0
Treatment Totals (T_1)	62.1	50.9	57.3	64.5	64.8	65.7	365.3

Sol :

Estimating Missing Value

Given that,

Total of known observations in i[th] block $y'._j$ = 68.4

Total of known observations incase of i[th] treatment, $y_{i.}'$ = 50.9

Total of all known observations $y'._j$ = 365.3

Treatments (t) = 6

Blocks or replicates (r) = 4

The missing value 'a' can be calculated by the formula

$$a = \frac{ry._j' + ty_{i.}' - y..'}{(r-1)(t-1)}$$

$$= \frac{4\times 68.4 + 6\times 50.9 - 365.3}{(4-1)(6-1)} = \frac{273.6 + 3.5.4 - 365.3}{3\times 5}$$

$$= \frac{213.7}{15}$$

a = 14.25

Analyzing the Data

Null Hypothesis (H_0)

The treatments and replicates are same

$$\text{i.e., } H_0 : \begin{cases} H_{0t} : \tau_1 = \tau_2 = \tau_3 = ... = \tau_6 \\ H_{0b} : b_1 = b_2 = b_3 = ... = b_6 \end{cases}$$

Alternative Hypothesis

$$\text{i.e., } H_1 : \begin{cases} H_{1t} : \text{Minimum two } \tau_i\text{'s are not same} \\ H_{1b} : \text{Minimum two blocks are not same} \end{cases}$$

The value of missing observation ‘a’ when substituted in the table, we get

Treatment Totals	Block Totals
T_1 = l 32.1	B_1 = 91.1
T_2 = 65.15	B_2 = 82.65
T_3 = 57.3	B_3 = 107.8
T_4 = 64.5	B_4 = 98.0
T_5 = 64.8	
T_6 = 65.7	

Grand total, $G = \sum_i \sum_j y_{ij} = 379.55$

$$N = rt = 4 \times 6 = 24$$

$$C.F = \frac{G^2}{N} = \frac{(379.55)^2}{24} = \frac{144058.20}{24} = 6002.42$$

Raw Sum of squares, (RSS) $= \sum_{i=1}^{6} \sum_{j=1}^{4} y_{ij}^2 = 6151.94$

Block sum of squares (BSS) $= \frac{1}{r} \sum_{j=1}^{t} B_j^2 - C.F$

$$= \frac{1}{6} \sum_{j=1}^{t} B_j^2 - C.F$$

$$= \frac{1}{6} [(91.1)^2 + (82.65)^2 + (107.8)^2 + (98)^2] - 6002.42$$

$$= \frac{1}{6}(8299.21 + 6831.02 + 11620.84 + 9604) - 6002.42$$

$$= \frac{1}{6}(36355.07) - 6002.42$$

$$= 6059.18 - 6002.42$$

$$= 56.76$$

Treatment sum of squares

$$= \frac{1}{t} \sum_{i=1}^{r} T_i^2 - C.F$$

$$= \frac{1}{4} \sum_{i=1}^{6} T_i^2 - C.F = \frac{1}{4}(62.1^2 + 65.15^2 + 57.3^2 + 64.5^2 + 64.8^2 + 65.7^2) - 6002.42$$

$= \frac{1}{4}(3856.41 + 4244.52 + 3283.29 + 4160.25 + 4199.04 + 4316.49) - 6002.42$

$= \frac{1}{4}(24060) - 6002.42$

$= 6015 - 6002.42$

$= 12.58$

Total sum of squares = RSS – C.F

= 6151.94-6002.42 = 149.52

Error sum of squares, ESS = Total S.S – SS – BSS

= 149.52 – 12.58 – 56.76 = 80.18

Adjustment factor of treatment sum of squares is calculated by the formula

$$\text{Adjustment factor} = \frac{\left(y_{.j}' + ty_{i}.' - y..'\right)^2}{t(r-1)(t-1)^2}$$

$$= \frac{(68.4 + 50.9 - 365.3)^2}{6(4-1)(6-1)^2}$$

$$= \frac{(68.4 + 305.4 - 365.3)^3}{6(3)(5)^2}$$

$$= \frac{(8.5)^2}{6(3)(25)}$$

$$= \frac{72.25}{450}$$

$= 0.16$

Adjusted value of treatment S.S = Treatment SS - Adjustment factor

= 12.58 – 0.16

= 12.42

ANOVA table obtained from the above data is shown below,

Sum of Variation	**Degrees of Freedom d.f**	**Sum of Squares (S.S)**	**Mean Sum of Squares (M.S.S)**	**Variation Ratio**
Treatments (Adjusted)	5	12.42	$\frac{12.42}{5} = 2.48$	$F_1 = \frac{2.48}{5.73} = 0.43$
Blocks (Adjusted)	3	56.76	$\frac{56.76}{3} = 18.92$	$F_b = \frac{18.92}{5.73} = 3.30$
Error	14	80.18	$\frac{80.18}{14} = 5.73$	
Total	22*			

Tabulated values of F_t and F_b at 5% level of significance are,

$F_{0.05}(5, 14) = 2.96$

$F_{0.05}(3. 14) = 3.34$

$\therefore$ Both F_t and F_b are not significant since their calculated values (0.43, 3.30) are lesser than tabulated values (2.96, 3.34). Hence, H_0 is accepted that means treatments and blocks are same.

2.3 Expectation of Various Sum of Squares

Q10. Discuss in detail about the expectations of sum of squares in CRD.

Ans :

Expectation of Sum of Squares.

Here we shall obtain the shall obtain the expectation of various expectation of various sum of squares under the assumption that the treatment effects τ_i and the random effects ε_{ij} are i.i.d. random variables such that

$$\left.\begin{aligned} E(\tau_i) = E(\varepsilon_{ij}) = 0, (i = 1, 2, 3, ..., v) \\ (j = 1, 2, ..., r_i) \\ Var(\tau_i) = E(\tau_i^2) = \sigma_\tau^2 \\ Var(\varepsilon_{ij}) = E(\varepsilon_{ij}^2) = \sigma_e^2 \end{aligned}\right\}$$

The linear model gives

$$\begin{cases} y_{ij} = \mu + \tau_i + \varepsilon_{ij} \\ \bar{y}_i, = \mu + \tau_i + \bar{\varepsilon}_i, \text{summing } y_{ij} \text{ over j from 1 to } r_i \text{ and dividing by } r_i) \\ \bar{y}.., = \mu + \frac{1}{n}\Sigma r_i \tau_i + \bar{\varepsilon}. \text{summing } y_{ij} \text{ over i and j and dividing by } n = \Sigma r_i \end{cases}$$

$$E\left(y_{ij}^2\right) = E\,[\mu + \tau_i + \varepsilon_{ij}]^2 = \mu^2 + \sigma_\tau^2 + \sigma_e^2 \qquad ...(i)$$

$$E\left(\bar{y}_{ij}^2\right)^2 = E[\mu + t_i + \bar{\varepsilon}_{i,}]^2$$

$$= \mu^2 + \sigma_\tau^2 + \frac{\sigma_e^2}{r_i} \qquad ...(ii)$$

$$\therefore \quad E(\varepsilon_i.)^2 = E\left[\frac{1}{r_i}\sum_{j=1}^{r_i}\varepsilon_{ij}\right]^3 = \frac{1}{r_i^2}\sum_{j=1}^{r_i}E\left(\varepsilon_{ij}^2\right) = \frac{\sigma_e^2}{r_i}$$

and $E(\bar{y}..^2) = E\left[\mu + \frac{\Sigma r_i \tau_i}{n} + \bar{\varepsilon}..\right]^2$

$$= \mu^2 + \frac{1}{n^2}(\Sigma r_i^2)\,\sigma_\tau^2 + \frac{\sigma_e^2}{n} \qquad ...(iii)$$

Using the results in (i), (ii) and (iii), we get from

$$E(S.S.T) = E(\Sigma_i r_i \bar{y}_i^2 - n\,\bar{y}..^2]$$

$$= \Sigma_i r_i E(\bar{y}_{i.})^2 - nE(\bar{y}..^2)$$

$$= \sum_i r_i \left[\mu^2 + \sigma_\tau^2 + \frac{\sigma_e^2}{r_i}\right] - n\left[\mu^2 + \frac{\Sigma r_i^2}{n^2}\sigma_\tau^2 + \frac{\sigma_e^2}{n}\right]$$

$$= (\mu^2 + \sigma_\tau^2)\left(\sum_i r_i\right) + v\sigma_e^2 - n\left[\mu^2 + \frac{\Sigma r_i^2}{n^2}\sigma_\tau^2 + \frac{\sigma_e^2}{n}\right]$$

$$= (y - 1)\,\sigma_e^2 + \left[n - \frac{\Sigma r_i^2}{n}\right]\sigma_\tau^2$$

$$\Rightarrow \quad E\left(\frac{S.S.T.}{v-1}\right) = \sigma_e^2 + \frac{1}{(v-1)}\left[n - \frac{\Sigma r_i^2}{n}\right]\sigma_\tau^2$$

Now from we have

$$E(S.S.E.) = E[\Sigma_i \Sigma_j y_{ij}^2 - \Sigma_i r_i \bar{y}_{i.}^2] = \Sigma_i \Sigma_j E(y_{ij}^2) - \Sigma_i r_i E(\bar{y}_{i.}^2)$$

$$= \sum_i \sum_j (\mu^2 + \sigma_\tau^2 + \sigma_e^2) - \sum_i r_i \left[\mu^2 + \sigma_\tau^2 + \frac{\sigma_e^2}{r_i}\right]$$

$$= n(\mu^2 + \sigma_\tau^2 + \sigma_e^2) - n(\mu^2 + \sigma_\tau^2) - v\sigma_e^2$$

$$= (n - v)\,\sigma_e^2$$

$$\Rightarrow \quad E\left(\frac{S.S.T.}{n-v}\right) = \sigma_e^2$$

Hence error mean sum of squares provides an unbiased estimate of σ_e^2. Finally,

$$E(T.S.S) = E(S.S.T) + E(S.S.E.)$$

$$= (n - 1)\,\sigma_e^2 + \left[n - \frac{\Sigma r_i^2}{n}\right]$$

Remarks.

1. If we assume that each treatment is replicated r times, i.e., $r_i = \ \forall\, i = I, 2, ...v$ then $n = rv$ and the expressions for expectations of various mean squares become

$$E\left(\frac{S.S.T}{y-1}\right) = \sigma_e^2 + \frac{1}{y-1}\left[rv - \frac{vr^2}{rv}\right]\sigma_\tau^2$$

$$\Rightarrow \quad E\left(\frac{S.S.T}{v-1}\right) = \sigma_e^2 + r\sigma_\tau^2$$

and $\quad E(T.S.S) = (n - l)\ \sigma_e^2 + r(v - l)\sigma_\tau^2$

2. In deriving the expression for the expectations of various sum of squares above, after substituting the fitted values, we assume that $\sum_i t_i = 0$ while the assumption $\sum_i \tau_i = 0$, as made in the calculation of variance of the estimates is not necessary.

3. The above expressions for the expectations of various sum of squares have been obtained under the assumption that τ_i as well ε_{ij} in model are i.i.d random variables as defined in. Instead, if we assume that ε_{ij} are i.i.d N(0, σ_e^2) while τ_i are completely enumerated and fixed, which is usually the case in most of the experiments, then we have

$$E(\tau_i) = \tau_i \text{ and } E = (\tau_i^2) = \tau_i^2$$

In this case we also require that $\sum_{i=1}^{v} \tau_i = 0$

If we take the case $r_i = r\ \forall\ i = 1, 2,..., v$, we get from on using and

$$E(y_{ij}^2) = E\,[\mu + \tau_i + \varepsilon_{ij}]^2$$

$$= \mu^2 + \tau_i^2 + \sigma_e^2$$

$$E(\bar{y}_{i.}^2) = E[\mu + \tau_i + \varepsilon_i]^2$$

$$= \mu^2 + \tau_i^2 + \frac{\sigma_e^2}{r}$$

$$E(\bar{y}..^2) = E\left[\mu + \frac{1}{v}\Sigma\tau_i + \bar{\varepsilon}..\right]^2$$

$$= \mu^2 + \frac{\sigma_e^2}{n}, \qquad (\because \Sigma\tau_i = 0)$$

Proceeding exactly as before and using and, we get

$$E(S.S.T) = rE[\sum_i (\bar{y}_{i.} - \bar{y}..)^2]$$

$$= rE[\sum_i E(\bar{y}_{i.}^2) - vE(\bar{y}..^2)]$$

$$= r\left[\sum_i \left(\mu^2 + \tau_i^2 + \frac{\sigma_e^2}{r}\right) - v\left(\mu^2 + \frac{\sigma_e^2}{n}\right)\right]$$

$$= r\left[\Sigma\tau_i^2 + \left(\frac{v}{r} - \frac{v}{n}\right)\sigma_e^2\right]$$

$$= r\Sigma\tau_i^2 + (v-1)\ \sigma_e^2$$

$$\Rightarrow \qquad E\left[\frac{S.S.T}{v-1}\right] = \sigma_e^2 + \frac{r}{v-1}\sum_{i=1}^{v}\tau_i^2$$

Similarly we shall get, as before

$$E\left[\frac{S.S.E}{n-v}\right] = \sigma_e^2$$

Hence under the null hypothesis

$H_0 : \tau_1 = \tau_2 = ... = \tau_\upsilon = 0,$

$$E\left[\frac{S.S.T}{v-1}\right] = \sigma_e^2$$

i.e., under H_0 treatment mean sum of squares also gives an unbiased estimate of σ_e^2.

Q11. Discuss in detail about the expectations of sum of squares in RBD.

Ans :

Let, S.S due treatments $S_i^2 = r\sum_i(\bar{y}_{i.} - \bar{y}..)^2$

Now, computing the expectation of SS due treatments,

$$S(S_i^2) = E\left[r\sum_{i=1}^{t}(\bar{y}_i. - \bar{y}..)^2\right]$$

$$= E[r\sum_i(\mu + \tau_i + \bar{\varepsilon}_i. - \mu - \bar{\varepsilon}..)^2] \qquad \left[\begin{array}{l}\because \bar{y}_i. = \mu + \tau_i + \bar{\varepsilon}_i. \\ \bar{y}.. = \mu + \bar{\varepsilon}..\end{array}\right]$$

$$= E[r\sum_i\tau_i^2 + \sum_i(\bar{\varepsilon}_i. - \bar{\varepsilon}..)^2 + 2r\sum_i\tau_i(\bar{\varepsilon}_i. - \bar{\varepsilon}..)]$$

$$= r\sum_i\tau_i^2 + E[\sum_i(\bar{\varepsilon}_i.^2 - \bar{\varepsilon}..)^2] + 2r\sum_i\tau_i\ E(\bar{\varepsilon}_i. - \bar{\varepsilon}..)]$$

$$= r\sum_i\tau_i^2 + rE[\sum_i(\bar{\varepsilon}_i.^2 - \bar{\varepsilon}..^2 + 0) \qquad [\because E(\bar{\varepsilon}_i. - \bar{\varepsilon}..) = 0$$

$$= r\sum_i\tau_i^2 + r[\sum_i E(\bar{\varepsilon}_i.^2) - tE(\bar{\varepsilon}..^2)$$

$$= r\sum_i\tau_i^2 + r\left[\sum_i\left(\frac{\sigma_e^2}{r}\right) - t.\frac{\sigma_e^2}{rt}\right] \qquad \left[\because E(\varepsilon_i.^2 = \frac{\sigma_e^2}{r} \text{ and } E(\varepsilon..^2) = \frac{\sigma_e^2}{rt}\right]$$

$$= r\sum_{i}\tau_i^2 + \left(\frac{t\sigma_e^2}{r} - \frac{\sigma_e^2}{r}\right)$$

$$= r\sum_{i}\tau_i^2 + r(t-1)\,\sigma_e^2$$

$$E\left(\frac{S_1^2}{t-1}\right) = \frac{r}{t-1}\sum_{i=1}^{t}\tau_i^2 + \sigma_e^2$$

$$E\left(s_1^2\right) = \sigma_e^2 + \frac{r}{(t-1)}\sum_{i=1}^{t}\tau_i^2$$

$$\boxed{\therefore E(s_1^2) = \sigma_e^2 + \frac{r}{(t-1)}\sum_{i=1}^{t}\tau_i^2}$$

Now, computing expectation of S.S due to varieties

$$E\,(S_b^2) = E\left[t\sum_{j}(\bar{y}._j - \bar{j}..)^2\right]$$

$$= E\left[t\sum_{j}(\mu + b_j + \bar{\varepsilon}._j - \mu - \bar{\varepsilon}..)^2\right]$$

$$= E\left[t\sum_{j}(b_j + \bar{\varepsilon}._j - \bar{\varepsilon}..)^2\right]$$

$$= E\left[t\sum_{j}\{b_j + (\bar{\varepsilon}_j - \bar{\varepsilon}..)\}^2\right]$$

$$= E\left[t\sum_{j}b_j^2 + t\sum_{j}(\bar{\varepsilon}._j - \bar{\varepsilon}..)^2 + 2t\sum_{j}b_j(\bar{\varepsilon}._j - \bar{\varepsilon}..\right]$$

$$= t\sum_{j}b_j^2 + tE\left[\sum_{j}(\bar{\varepsilon}._j - \bar{\varepsilon}..)^2\right] + 2t\sum_{j}b_jE(\bar{\varepsilon}._j - \bar{\varepsilon}..)$$

$$= t\sum_{j}b_j^2 + tE\left(\sum_{j}(\bar{\varepsilon}._j^{\,2} - r\bar{\varepsilon}..)^2\right) \qquad \left[\because E(\bar{\varepsilon}._j - \bar{\varepsilon}) = 0\right]$$

$$= t\sum_{j}b_j^2 + t\left[\sum_{j}E(\bar{\varepsilon}._j^{\,2}) - rE(\bar{\varepsilon}..^2)\right]$$

$$= t\sum_{j}b_j^2 + t\left[\sum_{j}\left(\frac{\sigma_e^2}{t}\right) - r.\frac{\sigma_e^2}{rt}\right] \qquad \left[\because E(\varepsilon._j^{\,2})\frac{\sigma_e^2}{j} \text{ and } E(\varepsilon..^2) = \frac{\sigma_e^2}{hk}\right]$$

$$= t\sum_j b_j^2 + t\left(\frac{r\sigma_e^2}{t} - \frac{\sigma_e^2}{t}\right)$$

$$E\left(S_b^2\right) = t\sum_{j=1}^{r} b_j^2 + (r-1)\ \sigma_e^2$$

$$E\left(\frac{S_b^2}{r-1}\right) = \frac{t}{r-1}\sum_i b_j^2 + \sigma_e^2$$

$$\therefore E(s_b^2) = \sigma_e^2 + \frac{t}{r-1}\sum_{j=1}^{r} b_j^2$$

Now, computing expectation of error sum of squares,

$$E\left(S_E^2\right) = E\left[\sum_i\sum_j (\bar{y}_{ij} - \bar{y}_i. - \bar{y}._j + \bar{y}..)^2\right]$$

$$= E\left[\sum_i\sum_j (\mu + \tau_i + n_j + \varepsilon_{ij} - \mu - \tau_i + \varepsilon_i. - \mu - b_j - \bar{\varepsilon}._j + \mu + \bar{\varepsilon}..)^2\right]$$

$$\left[\because \begin{array}{l} y_{ij} = \mu + \tau_i + b_j + \varepsilon_{ij} \\ \bar{y}_i. = \mu + \tau_i + \bar{\varepsilon}_i. \\ \bar{y}._j = \mu + b_j + \bar{\varepsilon}._j \\ \bar{y}.. = \mu + \bar{\varepsilon}... \end{array}\right]$$

$$= E\left[\sum_i\sum_j (\varepsilon_{ij} - \bar{\varepsilon}_i. - \varepsilon._j + \bar{\varepsilon}..)^2\right]$$

$$= E\left[\sum_i\sum_j (\varepsilon_{ij} - \bar{\varepsilon}_i. - \bar{\varepsilon}._j + \bar{\varepsilon}..)^2\right]$$

$$= E\left[\sum_i\sum_j (\varepsilon_{ij}^2 - \bar{\varepsilon}_i.^2 - \bar{\varepsilon}._j^2 + \bar{\varepsilon}..^2 - (2\varepsilon_{ij}\bar{\varepsilon}_i.) + (2\varepsilon_{ij}\bar{\varepsilon}._j) + 2\varepsilon_{ij}\bar{\varepsilon}.. + 2\bar{\varepsilon}_i.\bar{\varepsilon}.. - 2\varepsilon._j\bar{\varepsilon}..)\right]$$

$$= E\left[\sum_i\sum_j \varepsilon_{ij}^2 + r\sum_i \bar{\varepsilon}(\bar{\varepsilon}_i.^2 + t\sum_j \bar{\varepsilon}._j^2 + rt\bar{\varepsilon}..^2 - 2r\sum_i \bar{\varepsilon}_i.^2 - 2t\sum_i \bar{\varepsilon}._j^2 + 2rt\bar{\varepsilon}..^2 - 2rt\bar{\varepsilon}..^2 - 22rt\bar{\varepsilon}..^2\right]$$

$$= \sum_i\sum_j E\left(\varepsilon_{ij}^2\right) + r\sum_{i=1}^{t} E\left(\bar{\varepsilon}_i.^2\right) + t\sum_{j=1}^{r} E\left(\bar{\varepsilon}._j^2\right) + rt\left(\bar{\varepsilon}..^2\right) - 2r\sum_{i=1}^{t} E\left(\bar{\varepsilon}_i.^2\right) - 2t\sum_{j=1}^{r} E(\bar{\varepsilon}._j^2)$$

$$= \sum_i\sum_j \sigma_e^2 + r\sum_{i=1}^{t}\left(\frac{\sigma_e^2}{r}\right) + t\sum_{j=1}^{r}\left(\frac{\sigma_e^2}{t}\right) + rt\left(\frac{\sigma_e^2}{rt}\right) - 2r\sum_{i=1}^{t}\left(\frac{\sigma_e^2}{r}\right) - 2t\sum_{j=1}^{r}\left(\frac{\sigma_e^2}{t}\right)$$

$$\left[\because E(\varepsilon_{ij}^2) = \sigma_e^2, E(\bar{\varepsilon}_i.^2) = \frac{\sigma_e^2}{r}, E(\bar{\varepsilon}._j^2) = \frac{\sigma_e^2}{t}, E(\bar{\varepsilon}..^2) = \frac{\sigma_e^2}{rt}\right]$$

$= rt\sigma_e^2 + t\sigma_e^2 + r\sigma_e^2 + \sigma_e^2 - 2t\sigma_e^2 - 2r\sigma_e^2$

$= (rt - t - r + 1)\sigma_e^2$

$= (r - 1)(t - 1)\sigma_e^2$

$E(S_E^2) = (r - 1)(t - 1)\sigma_e^2$

$\therefore E(s_E^2) - \sigma_e^2$

Thus, error M.S.S provide unbiased estimate of σ_e^2 and S_i^2, S_b^2 are unbiased estimates of σ_e^2 under the null hypotheses.

Since $\sum_i \tau_i^2 \geq 0$ and $\sum_j b_j^2 \geq 0$,

(i) $E(s_t^2) = E(s_E^2)$ for H_{0t} hypothesis otherwise $E(s^2) > E(s^2)$

(ii) $E(s_b^2) = E(s_E^2)$ for H_{0V} hypothesis otherwise $E(s_t^2) > E(s_E^2)$

As the various types of S.S and their d.f are additive and the null hypothesis H_{0t}, H_{0b} of S_b^2, S_t^2, S_E^1 provides unbiased estimates for σ_e^2.

2.4 Comparison of the Efficiencies of above Designs

Q12. Compare the efficiency of RBD relative to CRD.

Ans : **(July-22, Dec.-21, Oct.-20)**

The ANOVA table for R.B.D in case of random effects with treatments 't' and replicas 'r' is given by,

Source of Variation	Degrees of Freedom (d.f)	Mean of Squares (M.S.S)	Expectations of Mean of squares E(M.S.S)
Treatments (column)	t – 1	S_1^2	$\sigma_e^2 + r\sigma_\tau^2$
Blocks (replicas)	r – 1	S_B^2	$\sigma_e^2 + \sigma_b^2$
Error	(r – l) (t – l)	4	4
Total	(rt – 1)		

Table (a): ANOVA Table for R.B.D with Random Effects

The ANOVA table for C.R.D incase of random effects with treatments 'f and replicas V' is given by,

Source of variation	Degrees of freedom (d.f)	Mean sum of squares (M.S.S)	Expectations of mean of squares E(M.S.S)
Treatments	(t-D	4	$a^2 + res^2$
Error	rt – t = t(r – 1)	4	a^2
Total	rt – 1		

Table (b): ANOVA Table for C.R.D With Random Effects

Equalizing the total sum of squares of R.B.D and C.R.D as they will be same when the experiment is same

$$(f - l)(\sigma_e^2 + r\,\sigma_\tau^2) + (r - l)(\sigma_e^2 + t\,\sigma_b^2) + (r - l)(f - 1)\,\sigma_e^2 = (t - l)(\sigma_e'^2 + r\,\sigma_\tau^2) + t(r - 1)\,\sigma_e'^2$$

$$\Rightarrow \quad (f - l)\,\sigma_e^2 + r(t - 1)\,\sigma_\tau^2 + (r - l)\,\sigma_e^2 + t(r - l)\,\sigma_b^2 + (r - l)(t - 1)\,\sigma_e^2 = (t - 1)\,\sigma_e'^2 + r(t - 1)\,\sigma_\tau^2 + t(r - 1)\,\sigma_e'^2$$

$$\Rightarrow \quad \sigma_e^2\,[\{t - 1) + (r - 1) + (r - 1)(t - 1)] + t(r - l)\,\sigma_b^2 = (t - 1)\,\sigma_e'^2 + r(t - l)\,\sigma_\tau^2 + t(r - l)\,\sigma_e'^2 - r(t - 1)\,\sigma_\tau^2$$

$$\Rightarrow \quad \sigma_e^2\,[(t - 1) + (r - 1) + (r - l)(t - 1)] + t(r - l)\,\sigma_b^2 = (t - l)\,\sigma_e'^2 + t(r - 1)\,\sigma_e'^2$$

$$\Rightarrow \quad \sigma_e^2\,[(r - 1) + (r - 1) + (r - l)(t - 1)] + t(r - l)\,\sigma_b^2 = \sigma_e'^2\,[(t - 1) + t(r - 1)]$$

$$\Rightarrow \quad \sigma_e^2\,[(rt - l) + t(r - l)\,\sigma_b^2$$

$$\Rightarrow \quad \sigma_e'^2\,(rt - 1)$$

[$\because$ Total degrees of freedom in both C.R.D & R.B.D = rt – 1] ... (1)

The efficiency of R.B.D against C.R.D can be evaluated by the ratio of their amount of precision, where amount of precision is the inverse of the error variance.

Therefore,

$$\text{efficiency } E = \frac{\left(\dfrac{1}{\sigma_e^2}\right)}{\left(\dfrac{1}{\sigma_e'^2}\right)} = \frac{\sigma_e'^2}{\sigma_e^2}$$

$\dfrac{\sigma_e'^2}{\sigma_e^2}$ can be calculated from equation (1) as follows,

$$\sigma_e^2\,(rt - 1) + t(r - l)\,\sigma_b^2 = \sigma_e'^2\,(rt - 1)$$

Dividing both side by σ_e^2

$$\frac{\sigma_e^2(rt-1)}{\sigma_e^2} + \frac{t(r-1)\sigma_b^2}{\sigma_e^2} = \frac{\sigma_e'^2}{\sigma_e^2}\,(rt - 1)$$

$$\Rightarrow \quad (rt - 1) + \frac{t(r-1)\sigma_b^2}{\sigma_e^2} = \frac{\sigma_e'^2}{\sigma_e^2}\,(rt - 1)$$

$$\Rightarrow \quad \frac{1}{rt-1}\left((rt-1) + \frac{r(r-1)\sigma_b^2}{\sigma_e^2}\right) = \frac{\sigma_e'^2}{\sigma_e^2}$$

$$\Rightarrow \quad \frac{(rt-1)}{(rt-1)} + \frac{t(r-1)\sigma_b^2}{(rt-1)\sigma_e^2} = \frac{\sigma_e'^2}{\sigma_e^2}$$

$$\Rightarrow \quad 1+\frac{t(r-1)\sigma_b^2}{(rt-1)\sigma_e^2} = \frac{\sigma_e'^2}{\sigma_e^2} = E \qquad ...(2)$$

But from table (a).

$E(s_E^2) = \sigma_e^2$ and

$E(s_B^2) = \sigma_e^2 + t\,\sigma_b^2$

Then, σ_e^2 can be calculated as $\frac{s_B^2 - s_E^2}{t}$

$$= \frac{\sigma_e^2 + t\sigma_b^2 - \sigma_e^2}{t} = \frac{t\sigma_b^2}{t}$$

$$\therefore \quad E\left[\frac{s_B^2 - s_E^2}{t}\right] = \sigma_b^2$$

Therefore, the unbiased estimates of σ_e^2 and σ_b^2 will be,

$$\hat{\sigma}_e^2 = s_E^2 \text{ and} \qquad ...(3)$$

$$\hat{\sigma}_b^2 = \frac{s_B^2 - s_E^2}{t} \qquad ...(4)$$

Substituting equations (3) and (4) in equation (2), we get,

$$E = \frac{\sigma_e'^2}{\sigma_e^2} = 1 + \frac{t(r-1)}{(rt-1)}\frac{\left(\frac{s_B^2 - s_E^2}{t}\right)}{EMS}$$

$$= 1 + \frac{t(r-1)\left(s_B^2 - s_E^2\right)}{(rt-1)t} \times \frac{1}{s_E^2}$$

$$= \frac{(rt-1)ts_E^2 + t(r-1)(s_B^2 - s_E^2)}{(rt-1)ts_E^2}$$

$$= \frac{t[(rt-1)s_E^2 + (r-1)(s_B^2 - s_E^2)}{(rt-1)ts_E^2}$$

$$= \frac{(rt-1)s_E^2 + (r-1)(s_B^2 - s_E^2)}{(rt-1)s_E^2}$$

$$= \frac{(rt-1)s_E^2 + (r-1)s_B^2 - (r-1)s_E^2}{(rt-1)s_E^2}$$

$$= \frac{[(rt-1)-(r-1)]s_E^2 + (r-1)s_B^2}{(rt-1)s_E^2}$$

$$= \frac{(rt-1-r+1)s_E^2 + (r-1)s_B^2}{(rt-1)s_E^2}$$

$$= \frac{(rt-r)s_E^2 + (r-1)s_B^2}{(rt-1)s_E^2}$$

$$\therefore\ E = \frac{r(t-1)s_E^2 + (r-1)s_B^2}{(rt-1)s_E^2}$$

Short Question & Answers

1. Discuss in detail about statistical analysis of CRD.

Ans :

Statistical analysis of a C.R.D. is analogous to the ANOVA for a one way classified data, the linear model (assuming various effects to be additive) becomes.

$$y_{ij} = \mu + \tau_i + \varepsilon_{ij} \begin{pmatrix} i = 1, 2, ..., v \\ j = 1, 2, ..., r_i \end{pmatrix}$$

where $y_{(i}$ is the yield or response from the jth unit receiving the with treatment, μ. is the general mean effect, τ_i- is the effect due to the with treatment, and ε_{ij} is error effect due to chance such that ε_{ij} a dentically and independently distributed (i. i. d.) N(0, σ_e^2). Then $n = 1 \sum_{i=1}^{v} r$, is the total number of experimental units. If we write

$$\sum_i \sum_j y_{ij} = y.., = G = \text{Grand total of all the n observations.}$$

$$\sum_{j=1}^{ri} y_{ij} = y_i = T_i = \text{Total resported the units incase of ith treatment. Then, as in ANOVA}$$

$$\text{Correction Factor, CF} = \frac{(G)^2}{n}$$

$$G = \text{Grand Total} \left(\sum_i \sum_j y_{ij} \right)$$

$$n = \text{Number of treatments} \left(\sum_{i=1}^{n} r_i \right)$$

$$\Rightarrow \quad \text{T.S.S} = \sum_i \sum_j (y_{ij} - \bar{y}..)^2 = \sum_i \sum_j y_{ij}^2 - \frac{G^2}{n}$$

$$\therefore \quad \text{T.S.S} = \sum_i \sum_j y_{ij}^2 - \text{C.F}$$

$$\Rightarrow \quad \text{S.S.T} = \sum_i r_i (\bar{y}_i. - \bar{y}..)^2$$

$$\therefore \quad \text{S.S.T} = \sum_i \left(\frac{T_i^2}{r_i} \right) - \text{C.F}$$

$$\Rightarrow \quad \text{S.S.T} = \text{T.S.S} - \text{S.S.T}$$

$$= \sum_i \sum_j y_{ij}^2 - \text{C.F} - \left(\sum_i \left(\frac{T_i^2}{r_i} \right) - \text{C.F} \right)$$

$$\therefore \quad S.S.E = \sum_i \sum_j y_{ij}^2 - \sum_i \left(\frac{T_i^2}{r_i} \right)$$

ANOVA Table for C.R.D

Source of variation	d.f	S.S.	M.S.S	Variance ratio
Treatment	$v - 1$	S_T^2	$s_T^2 = \dfrac{S_T^2}{(v-1)}$	
Error	$n - v$	S_E^2	$s_E^2 = \dfrac{S_E^2}{(n-v)}$	$F_T = \dfrac{s_T^2}{s_E^2}$
Total	$n - 1$	$S_T^2 + S_E^2$		

Under the null hypothesis $H_0 : \tau_1 = \tau_2 = ... = \tau_v$ against the alternative that all τ's are not equal, the statistic,

$$F_T = \frac{S_T{}^2}{S_E{}^2} \sim F(v-1, n-v)$$

i.e., F_T follows F (central) distribution with $(v - 1, n - v)$ d.f.

If $F_T > F_{\alpha,(v-1, n-v)}$, then H_0 is refuted at α% level of significance and we conclude that treatments differ significantly. If $F_T < F_{\alpha,(v-1, n-v)}$, H_0 may be accepted, i.e., the data do not provide any evidence to prefer one treatment to the other and as such all of them can be considered alike.

Remrak. The following formulae for the calculation of various S.S. are maeh convenient to use from practical point of view :

$$T.S.S. = \sum_i \sum_j \left(y_{ij} - \bar{y}_{..} \right)^2 = \sum_i \sum_j y_{ij}{}^2 - n\bar{y}_{..}{}^2$$

$$= \sum_i \sum_j y_{ij}^2 - \frac{G^2}{n}, \qquad n = \sum_{i=1}^{v} r_i$$

$$\therefore \quad T.S.S. = \text{Raw S.S.} - \text{Correction Factor} = R.S.S. - C.F.,$$

$$S.S.T. = \Sigma r_i(\bar{y}_i, - \bar{y}_{..})^2$$

$$= \Sigma r_i(\bar{y}_i{}^2, - n\bar{y}_{..}{}^2$$

$$= \sum_{i=1}^{v} \left(\frac{T_i^2}{r_i} \right) - C.F.$$

$$\therefore \quad S.S.E. = T.S.S. - S.S.T.$$

$$= \sum_i \sum_j y_{ij}^2 - \sum_i \left(\frac{T_i^2}{r_i} \right)$$

$$= \sum_i \sum_j y_{ij}^2 - \sum_i r_{\backslash i} \bar{y}_i^2$$

2. ANOVA TABLE FOR m × m L.S.D.

Ans :

Source of Variation	d. f.	S.S.	M.S.S.	Variance Ratio 'F'
Rows	m –1	$S_R^{\ 2}$	${}^{S}R^2 = S_E^{\ 2}/(m-1)$	$F_R = {}^{S}R^{2/S}E^2$
Columns	m – 1	$S_C^{\ 2}$	${}^{S}C^2 = S_C^{\ 2}/(m-1)$	$F_C = {}^{S}C^{2/S}E^2$
Treatments	m – 1	$S_T^{\ 2}$	${}^{S}T^2 = S_T^{\ 2}/(m-1)$	$F_T = {}^{S}T^{2/S}E^2$
Error	(m – l)(m – 2)	$S_E^{\ 2}$	${}^{S}E^2 = S_E^{\ 2}/(m-1)(m-2)$	
Total	$m^2 - 1$			

Let us set up the null hypotheses

for row effects, $H_\alpha : \alpha_1 = \alpha_2 = ... = \alpha_m = 0$,

for column effects, $H_\alpha : \beta_1 = \beta_2 = ... = \beta_m = 0$ and

for treatment effects, $H_\tau : \tau_1 = \tau_2 = ... = \tau_m = 0$,

The variance ratios F_R, F_C and F_T follow (central) F distribution with (m –1), (m –1) (m – 2) d.f. under the null hypotheses H_α, H_β and H_τ respectively.

Let $F_\alpha = F_\alpha$ {(m –1), (m –1) (m – 2)} be the tabulated value of F for (m – 1) (m – 2) d.f. at the level of significance 'α'. Thus if $F_R > F_\alpha$ we reject H_α and if $F_R < F_\alpha$ we may accept H_α.

Similarly we can test for H_β and H_T .

3. Efficiency of a design.

Ans :

The efficiency of a design can be defined as a measure of accuracy of various estimates performed by two various designs with the equal number of experimental units and blocks.

Therefore,

$$\text{The efficiency of a design} = \frac{\text{Precision of Design (1)}}{\text{Precision of Design (2)}}$$

$$= \frac{r_1 / \sigma_1^2}{r_2 / \sigma_2^2}$$

$$\therefore \text{The efficiency of a design} = \frac{r_1\sigma_2^2}{r_2\sigma_1^2}$$

Where. r_1, r_2 are replications of Design(l) and Design(2) respectively.

σ_1, σ_2 are error variances per each unit of Design (1) and Design (2) respectively.

4. Advantages of LSD over CRD.

Ans :

The different advantages of LSD over CRD are as follows,

1. LSD is more efficient compared to CRD.
2. LSD uses the principle of local control to reduce the experimental error by dividing the heterogeneous experimental area into homogeneous groups. Whereas, CRD does not use the principle of local control.
3. LSD is applicable for the experiments which are homogeneous and heterogeneous groups. But, CRD is applicable only to the experiments that are homogeneous.
4. LSD is useful when the variation is bidirectional. But, CRD is applicable to only small number of experiments.

5. Advantages of RBD.

Ans :

1. This design is completely flexible to have any no. of treatments and blocks.
2. It provides more accurate results than CRD due to grouping.
3. The statistical analysis is simple and less time consuming.
4. Relatively easy statistical analysis even with missing data.
5. The number of replicates provides enough degrees of freedom to the error sum of squares.

6. How do you estimate the missing observations in LSD?

Ans :

Let us suppose that in mxm Latin Square, the observation occuring in the ith row, jth column and receiving the kth treatment is missing. Let us assume that its value is x, i.e., y_{ijk} = x.

R = Total of the known observations in the ith row, i.e., the row containing 'x'.

C = Total of known observations in the jth column, i.e:, the column containing 'x'.

T = Total of known observations receiving kth treatment, i.e., total of all known treatment values containing 'x'.

S = Total of known observations.

Then

$$\text{T.S.S.} = y^2 + \text{constants w.r.t. } x - \frac{(S+x)^2}{m^2}$$

$$\text{S.S.R.} = \frac{(R+x)^2}{m} + \text{constant w.r.t. } x - \frac{(S+x)^2}{m^2}$$

$$\text{S.S.C.} = \frac{(C+x)^2}{m} \text{ h constant w.r.t. } x - \frac{(S+x)^2}{m^2}$$

$$\text{S.S.T.} = \frac{(T+x)^2}{m} + \text{constant w.r.t. } x - \frac{(S+x)^2}{m^2}$$

$\therefore$ E = Residual Sum of Squares (S.S.E.)

= T.S.S. – S.S.R. – S.S.C. – S.S.T.

$$= x^2 - \frac{1}{m}[(R + x)^2 + (C + x)^2 + (T + x)^2] + 2\frac{(S+x)^2}{m^2}$$

We will choose x so as to minimise E.

$$\therefore \quad \frac{\partial E}{\partial x} = 0 = 2x - \frac{2}{m}[R + C + T + 3x] + \frac{4(S+x)}{m^2}$$

$$\Rightarrow \quad (m^2 - 3m + 2)x = m(R + C + T) - 2S$$

$$\Rightarrow \quad x = \frac{m(R+C+T)-2S}{(m-1)(m-2)}$$

7. Explain in detail about Completely randomized Design (C.R.D).

Ans :

The completely randomised design is the simplest of all the designs, based on principles of randomisation and replication. In this design treaunents are allocated at random to the experimental units over the entire experimental material. Let us suppose that we have v treatments, the ith treatment being replicated r, times, i = 1, 2, ..., v. Then the whole experimental material is divided into $n = \Sigma r_i$ experimental units and the treatments are distributed completely at random over the units subject to the condition that the ith treatment occurs r< times. Randomisation assures that extraneous factors do not continually influence one treatment. In particular case if

$$r_t = r \; \forall \; i = l,2,..., v$$

i.e., if each treatment is repeated an equal number of times r, then n = rv and randomisation gives every group of r units an equal chance of receiving the treatments. In general, equal number of replications for each treatment should be made except in particular cases when some treatments are of greater interest than others or when practical limitations dictate otherwise.

Advantages

1. C.R.D. results in the maximum use of the experimental units since all the experimental material can be used.
2. The design is very flexible. Any number of treatments can be used and different treatments can be used unequal number of times with-out unduly complicating the statistical analysis in most of the cases.
3. The statistical analysis remains simple if some or all the observations for any treatment are rejected or lost or missing for some purely random accidental reasons. We merely carry out the standard analysis on the available data. Moreover the loss of information due to missing data is smaller in comparison with any other design.
4. It provides the maximum number of degrees of freedom for the estimation of the error variance, which increases the sensitivity or the precision of the experiment for small experiments, i.e., for experiments with small number of treatments.

Disadvantages

1. CRD is used only on homogeneous blocks but in the field experimentation the available experimented material is always heterogenous.

2. CRD is suitable when small no. of treatments are taken but it is not practicable.
3. In CRD the local control principle is not used, hence, it leads to more the experimental error.

8. Define Latin Square Design (L.S.D).

Ans :

In RBD whole of the experimental area is divided into relatively homeogeneous groups (blocks) and treatments are allocated at random to units within each block, i.e., randomisation was subjected to one restriction, i.e., within blocks. But in field experimentation, it may happen that experimental area (field) exhibits fertility in strips, eg., cultivation might result in alternative strips of high or low fertility. R.B.D will be effective if the blocks happen to be parallel to these strips and would be extremely inefficient if the blocks are across the strips.

Initially, fertility gradient is seldom known a useful method of eliminating fertility variations consists in an experimental layout which will control variation in two perpendicular directions such a layout in a Latin square design (L.S.D.).

Layout of Design

In field plot experiments, the Latin square is usually laid out in the conventional square with the rows and columns corresponding to possible fertility trends in two directions across the field. In other types of experiments, the rows and columns may be made to correspond to different sources of error as in animal feeding experiment where the column groups may correspond with initial weight and the row group with age.

In this design the number of treatments is equal to the number of replications. Thus in case of m treatments, there have to be $m \times m = m^2$ experimental units. The whole of experimental area is divided into m^2 experimental units (plots) arranged in a square, so that each row as well as each column contains m units (plots).

The m treatments are then allocated at random to these rows and columns in such a way that every treatments occurs once and only once in each row and in each column. Such a layout is known as $m \times m$ latin square design (L.S.D) and is extensively used in agricultural experiments.

For examples if we are interested in studying the effects of m types of fertilizers on the yield of a certain variety of wheat, it is customary to conduct the experiments on a square field with m^2 - plots of equal area and to associate treatments with different fertilizers and row and column effects with variations in fertility of soil.

Obviously, there can be many arrangements for an $m \times m$ L.S.D. and a particular layout in an experiment must be determined randomly.

For 2×2 and 3×3 latin squares, only one standard square exists.

A	B
B	A

A	B	C
B	C	A
C	A	B

For a 4×4 latin square design, 4 standard squares are possible, one of the design is

A	B	C	D
B	C	D	A
C	D	A	B
D	A	B	C

9. Merits and Demerits.

Ans :

Merits

- LSD controls more of the variation than CRD and RBD.
- The statistical analysis is simple even though slightly complicated than RBD.

Demerits

- In LSD the number of treatments are restricted to the number of replications therefore it is not a flexible design.
- In case of missing plots when several units are missing than the statistical analysis becomes complicate.
- In field layout RBD is much easy to manage the LSD.

10. Randomized Block Design (R.B.D).

Ans :

If the whole experimental material is not homogeneous then divide the experimental area into relatively homogeneous blocks and then allocate the treatments in each block randomly. The designs so obtained is called randomised block design.

Let us suppose that the experimental area is divided into *l* homogeneous blocks and there be k treatments.

Therefore divide each block into k subdivisions and allocate the k treatments randomly to each block. This design is based on all the three principles. i.e., the principle replication randomization and local control.

11. Explain the concept of critical difference.

Ans :

The critical difference (CD) is a parameter used to compare the means of different treatments, that have an equal number of replications.

$$C.D = t(m-1)(n-1) \times \sqrt{\frac{2EMS}{m}}$$

Where,

m = Number of treatments in the columns

n = Number of treatments in the rows.

Choose the Correct Answer

1. In RCBD we may assume that the treatment are fixed and the blocks are random, such a model is called ______. [b]
   (a) Random effect model (b) Mixed effect model
   (c) Rare effect model (d) Fixed effect model
2. The simplest type of the basic designs. [a]
   (a) CRD (b) RCBD
   (c) ANOCOVA (d) BCR
3. A design in which the treatments are assigned to the experimental unit completely at random. [c]
   (a) ANOCOVA (b) RCBD
   (c) CRD (d) BCR
4. CRD gives accurate information if all the experimental units present in the experiment are______. [b]
   (a) Heterogeneous (b) Homogeneous
   (c) Not clear (d) Clear
5. CRD is very simple. [b]
   (a) But not easily laid out (b) And easily laid out
   (c) But gives biased result (d) And unbiased
6. If in the CRD some observations are missing then also the analysis is very simple, because the missing observations are discarded and carry out the experiment without losing the ______. [a]
   (a) Efficiency of the design (b) Degree of freedom
   (c) Confidentiality (d) Sufficiency of the design
7. CRD provides maximum number of degree of freedom for the ______. [b]
   (a) Sum of squares (b) Error sum of squares
   (c) Experiment (d) Calculations
8. In CRD due to the maximum number of degree of freedom, the experimental error is ______. [d]
   (a) Increased (b) Remained the same
   (c) Not remained the same (d) Reduced
9. Completely flexible design i.e. any number of treatments and any number of units per treatment may be used ______. [c]
   (a) GLSD (b) LSD
   (c) CRD (d) RCBD
10. Design is not useful when the experimental units are heterogeneous. [d]
    (a) GLSD (b) LSD
    (c) RCBD (d) CRD

Fill in the blanks

1. CRD Stands for ______.
2. ______ is the process where in the treatments are randomly allocated such that each treatment unit will have an equal chance of receiving any treatment.
3. In R.B.D, sum of squares due to blocks is ______.
4. In randomized block design, the expectation of squares due to treatment is ______.
5. A Latin square in which the first row and first column contains treatment in alphabetical order is called ______ Latin square.
6. The order of 2 × 2 and 3 × 3 have only ______ possible standard square.
7. The linear additive model for Latin square is ______.
8. RBD Stands for ______.
9. In Latin square design, the variance of estimate is ______.
10. In C.R.D, the total sum of squares is ______.

Answers

1. Completely randomized Design
2. Completely randomized design
3. $\frac{\sum_i y._j^2}{t} - CF$
4. $E\left(\frac{S.S.T}{t-1}\right) = E(M.S.T) = \sigma_e^2 + r\sigma_\tau^2$
5. Standard
6. One
7. $y_{ijk} = \mu + \alpha_i + \beta_j + \tau_i + \varepsilon_{ijk}$
8. Randomized Block Design
9. Var $(\hat{\alpha}_i) = \left(\frac{m-1}{m^2}\right)\sigma_e^2$
10. $\sum_i \sum_j (y_{ij} - \bar{y}..)^2$

UNIT III	**Vital statistics :** Introduction, definition and uses of vital statistics. Sources of vital statistics, registration method and census method. Rates and ratios, Crude death rates, age specific death rate, standardized death rates, crude birth rate, age specific fertility rate, general fertility rate, total fertility rate. Measurement of population growth, crude rate of natural increase- Pearl's vital index. Gross reproductive rate sand Net reproductive rate, Life tables, construction and uses of life tables and Abridged life tables.

3.1 Vital Statistics

3.1.1 Introduction, Definition and Uses of Vital Statistics

Q1. Define vital statistics. Explain the uses of vital statistics.

Ans : **(July-22)**

Introduction

Vital statistics is defined as that a branch of Biometry which deals with data and the laws of human mortality, morbidity and demography. The term vital statistics refers to the numerical data or the techniques used in the analysis of the data pertaining to vital events occurring in the given section of the population. By vital events we mean such events of human life as fertility and mortality (births and deaths), marriage, divorce, separation, adoptions, etc.

Definition

The term vital statistics refers to the Numerical data (or) the technique use in the numerical data (or) the technique use in the analysis of data. Vital statistics forms perhaps the most important branch of statistic as it deals with mankind in the aggregate.

Uses

Vital statistics are being Extensively used in almost all the spheres of human activity. Given below are some of the important applications of vital statistics.

1. **Study of Population Trend:** The study of births and deaths gives us an idea of the population trend of any region. Community (or) country.

   If 'Birth rate > Death rate ; there is an increasing trend . If Birth rate < Death rate, there is decreasing trend.

   The division of the population of different regions by birth and death rates enables is to form some idea about the population trend of the regions or countries and the general standard of living and virility of the races.

   In under developed countries, the birth rate is fairly high but at the same time it is accompanied by high infant mortality rate showing thereby the lack of medical facilities poor hygenic conditions, malnutrition and low standard of living.

2. **Use of Public Administration:** The study of population movement i.e., population estimation, population projections and other allied studies together with birth and death statistics according to age and sex distributions provides any administration with fundamental tools which are indispensable for the overall planing and evaluation of economic and social development programes.

3. **Mortality** and natality statistics also provide guide sports for use by the researchers in Medical and pharmaceutical profession.

4. **Use to Operating Agencies:** The facts relating to births, deaths and marringes are of Extreme importance to various official agencies for a variety of administrative purposes. Mortality statistics serve as a guide to the health authorities for sanitary improvements, improved medical facilities and public cleanliness.

5. **The whole of actuarial science,** including life insurance is based on the mortality or life tables. The vital records concerning all possible factors contributing to deaths in various ages are indispensable tool in numerous life insurance schemes.

3.2 Sources of Vital Statistics

3.2.1 Registration Method and Census Method

Q2. Explain briefly about various Sources of Vital Statistics.

Ans :

The vital statistics data are usually obtained by the following methods.

1. Registration Method

The most important source of obtaining vital statistics data is the registration method which consists in continuous and permanent recording of vital events pertaining to births, deaths, marriages, migration, *etc*. These, data in addition to their statistical utility also have their value as legal documents. Registration of births provide information on place of birth, sex, age and religion of the parents, legitimacy, number of previous issues and their sexes, father's occupation and birth place of parents. Similarly death registration furnishes information on place of death, sex, age, marital status, number of issues, birth place, occupation and cause of death. Similar information is also obtained with respect to marriages and migrants.

Many countries require compulsory registration of births and deaths under the law. For example, every new birth has to be reported to the authorities along with the information as given above. Similarly the death of a person is automatically recorded since the disposal of the body requires an appropriate death certificate from the authorities.

Shortcomings

- In India, in rural areas there is no legislation which makes the registration of vital events (births, deaths, marriages) and reporting of epidemics compulsory and the requisite information is collected by village.
- Pauls under the administrative orders of the government.
- Consequently, a number of births are likely to remain unregistered especially in scattered rural areas. Even in municipal areas where registration is compulsory, the laws in respect of registration vary from State to State.
- Thus in India, the statistics of be his suffer from the error of understimation as pointed out in Census of India paper 6, 1954, published by Registrar General of India. "The reg stration of births is non-existent in some parts of the country and incomplete in varying degrees in all parts of the country.
- Hence statistics of births suffer from errors of under estimation". Similarly in registration of deaths, the data regarding age at death, cause of the death, etc., are usually unreliable.
- Due to non-availability of qualified doctors in villages and interior rural areas, quite often the disease and consequently the cause of death remains undiagnosed.
- Moreover, people have a general tendency to withhold information regarding their diseases particularly in respect of infection/ or contagious diseases.

Moreover, in our country there are no proper records about the ages of mothers at the time to marriage, at the birth of first child and to subsequent children. The religious customs do not require the compul-sory registration of marriages in Hindus and Muslims. Hence we do not get any reliable data in respect of marriages for whole of the country.

In order to ensure a continuous permanent recording of vital events suitable legislation, uniform all over the country, should be introduced, making time registration of births, deaths, marriages, etc., compulsory. Such legislation should also provide sanctions for the enforcement of the obligation. Separate organisation should be set up to collect this data more completely and systematically.

2. Census Method

Almost in all the countries all over the world population census is conducted at regular intervals of time, usually ten years. Census consists of complete enumeration of the population of the particular area under study and collecting information from individuals regarding age, sex, marital status, occupation, religion and other economic and social characteristics.

The main drawback of the census method is that it provides vital statistics only for the census years and fails to give any information about the vital events in the intercensal period.

3.3 Rates and Ratios of Vital Statistics

Q3. Explain the ratios of Vital Statistics.

Ans :

In vital statistics, the vital rates are the rates of vital events like, birth, death etc. They are simply defined as the frequency of occurrence of vital events over a given interval of time.

Vital Ratios

In vital statistics, the vital ratios are similar to the vital rats of vital events like, birth, death etc. They are simply defined as the relationship of one group with another group of population.

Both vital rates and vital ratios are used interchangeably for the calculations and estimation of vital events in vital statistics. The basic formula used for calculating the rate of vital event is as follows,

$$\text{Rate of Vital Event} = \frac{\text{Number of cases of even considered}}{\text{Total population exposed to the risk of event}}$$

However, rate of vital events is mostly expressed based on thousand persons.

Types of Rates and Ratios of Vital Statistics

The rates and ratios of vital statistics are classified into three different measures. They are as follows,

1. Measures of Mortality or Death
2. Measure of Fertility or Birth
3. Measure of Population Growth

3.4 Crude Death Rates, Age Specific Death Rate, Standardized Death Rates

Q4. Explain various measurement of mortality.

(OR)

Explain various measures of mortality rate.

Ans : **(July-21)**

The following are the principle rates used in measurement of mortality.

1. Crude Death rate (C.D.R)
2. Age Specific Death rate (A.S.D.R)
3. Standard Death rate (St.D.R.)

1. Crude Death Rate (C.D.R)

This is the simplest of all the indices of mortality and is defined as the number of deaths per k persons in the population of any given region or community during a given period. Thus, in particular, the annual crude death rate (C.D.R) denoted by m for any given region (or) community is given by.

$$CDR = \frac{\text{Annual Deaths}}{\text{Annual Mean Population}} \times k$$

where k = 1,000

The crude death rate for any period gives the rate at which the populated is depleted through deaths over the course of the period.

Merits

- It is simple to understand and calculate.
- C.D.R is perhaps the most widely used of any vital statistics rates. As an index of mortality, it is used in numerous demographic and public health problems.
- Since the entire population of the region is Exposed to the risk of mortality.

Demerits

- The important drawback of crude death rate is "it completely ignores the age and sex distribution of the population".

Example

Children in the early ages of their life and the older generation are exposed to higher risk of mortality as compared to the middle age people.

- Crude death rate is not suitable for comparing the mortality in two places (or) at same place in two different periods unless the population of the places being compared have more (or) less the same age and sex distribution.

2. Age Specific Death Rate (A.S.D.R)

To formulate ideas mathematically, Let ${}_nD_x$ = Number of deaths in the age -group (x, x + n) i.e., the number occurring during a given period, t.

${}_nP_x$ = Total population of the age -group x to (x+n) then the age -specific death rate for the age -group x to x+n, usually denoted by ${}_nm_x$ is given by

$${}_{n}m_{x} = \frac{{}_{n}D_{x}}{{}_{n}P_{x}} \times 100$$

Taking $n = 1$, we get the annual A.S.D.R given by $m_x = \frac{D_x}{P_x} \times 1000$.

To be more specific, the A.S.D.R for males is given by

$${}_{n}^{m}m_{x} = \frac{{}_{n}^{m}D_{x}}{{}_{n}^{m}P_{x}} \times 1000 \qquad \text{... (1)}$$

Where ${}_{n}^{m}P_{x}$ is the number of males in the population in the age group x to $x + n$ and ${}_{n}^{m}D_{x}$ is the number of deaths amongst this population.

Similarly, the A.S.D.R for females is given by the formula.

$${}^{f}_{n}m_{x} = \frac{{}^{f}_{n}D_{x}}{{}^{f}_{n}P_{x}} \times 1000 \qquad \text{... (2)}$$

Equation (1) and (2) give the death rates specific to both age and sex.

Merits

The death rates specific to age and sex overcome the drawback of C.D.R, Since they are computed by taking into consideration the age and sex composition of the population. By eliminating the variation in the death rates due to age-sex distribution of the population S.D.R's provide more appropriate measures of the relative mortality situation in the regions.

- For general analytical purpose, the death rate specific for age and sex is one of the most important and widely applicable type of death rates.

Demerits

More over, in addition to age and sex distribution of the population social, occupational and topographical factors come into operation causing what is called differential mortality. S.D.R'S completely ignore these factors.

3. Standardized Death Rates

The crude death rates in terms of age specific death rates for two regions A and B are given respectively by

$$m^a = \frac{D^a}{P^a} \times 1000 = \frac{\sum_x m_x^a p_x^a}{\sum_x P_x^a} \qquad \text{... (1)}$$

and $$m^b = \frac{D^b}{P^b} \times 1,000 = \frac{\sum_x m_x^b P_x^b}{\sum_x P_x^b} \qquad \text{... (2)}$$

The Expressions in (1) & (2) are nothing but the weighted arithmetic means of the age S.D.R., the weights being the corresponding populations in the age groups we observe that even if A.S.D.R'S are some, i.e.,

$$m_x^a = m_x^b \ \forall x,\ m^a \neq m^b \text{ since in general, } \frac{P_x^a}{\sum_x P_x^a} \neq \frac{P_x{}^b}{\sum_x P_x^b}$$

i.e., Since the age distributions of the populations in the two regions A and B are not identical. This drawback is removed if the same set of weights is used in $\rightarrow$(1) & (2) for computing the weighted average of the A.S.D.R'S. This is what is done in standardized death rates or adjusted death rates, used with a prefix to identify the basis of adjustment as, for Example, age-adjusted death rates and so on. We discuss below the two methods of age adjustments which are in common use.

(i) Direct Method of Standardization

This method consists in weighting the age specific death rates not by the corresponding population of the area to which they refer (as is done in C.D.R). But by the population contribution of another region chosen as a standard Thus if P_x^S is the number of persons in the age - group x to $x+1$ in the standard population, then the standardised death rates for the regions A and B are given respectively by

$$(\text{STDR})_A = \frac{\sum_x m_x^a P_x^s}{\sum_x P_x^s} \text{ and } (\text{STDR})_B = \frac{\sum_x m_x^a P_x^s}{\sum_x P_x^s}$$

These age adjusted death rates for regions A and B respectively are nothing but the crude death rates that would be observed in the standard population if it were subject to the age S.D.R of the regions A and B.

(ii) Indirect Standardization

In compiling the standardised death rates by formula in direct method. It is necessary to know the number of persons and the age - specific death rates for different age -groups. Quite often we have a population classified by age but the age S.D.R's may not be known . However, the total number of deaths and hence C.D.R. may be known.

In such case, we use the indirect method of standardisation which consists in multiplying the crude death rate of the region A, say, by adjustment factor 'c' measuring the relative 'mortality'. Proneness of the population of the region such that the result is equal to the standardised death rate. Thus the problem is to find C such that.

ST. D.R = C.D.R × C

$$\frac{\sum_x m_x{}^s P_x{}^s}{\sum_x P_x{}^s} = C \times \frac{\sum_x m_x{}^s p_x{}^a}{\sum_x P_x{}^a}$$

(or)

$$C = \frac{\sum_x m_x{}^a P_x{}^s}{\sum_x P_x{}^s} \div \frac{\sum_x m_x{}^a P_x{}^a}{\sum P_x{}^a}$$

Since $m_x{}^a$ are not usually known, an approximate value of c is obtained on replacing m_x^a by m_x^s, the age - S.D.Rs for the standard population.

Thus giving

$$\hat{C} = \frac{\sum_x m_x{}^s P_x{}^s}{\sum_x P_x{}^s} \div \frac{\sum_x m_x{}^s P_x{}^a}{\sum_x P_x{}^a} = \text{(C.D.R. for standard Population)} \div \frac{\sum_x m_x{}^s P_x{}^a}{\sum_x P_x{}^a}$$

Finally, the indirect standardised death rate for region A is given by.

STDR for A = (C.D.R for A) × $\hat{C}$.

Problems

1. Compute the crude and standardised death rates of the two populations A and B, regarding A as standard population, from the given data.

	A		B	
Age-group (years)	**Population**	**Deaths**	**Population**	**Deaths**
Under 10	20,000	600	12,000	372
10 - 20	12,000	240	30,000	660
20 - 40	50,000	1250	62,000	1612
40 - 60	30,000	1050	15,000	325
Above 60	10,000	500	3,000	180

Sol :

	Population A			Population B			
Age group (years)	**Population** $p_x{}^a$	**Deaths** $D_x{}^a$	**Death rate per 1,000** $m_x{}^a$	**Population** $P_x{}^b$	**Deaths** $D_x{}^b$	**Death rate per 1000** $m_x{}^b$	$m_x{}^b p_x{}^a$
Under 10	20,000	600	30	12,000	372	31	6,20,000
10 - 20	12,000	240	20	30,000	660	22	2,64,000
20 - 40	50,000	1250	25	62,000	1612	26	13,00,000
40 - 60	30,000	1050	35	15,000	525	35	10,50,000
above 60	10,000	500	50	3,000	180	60	6,00,000
	1,22,000	3,640		1,22,000	3,349		38,34,000

Crude Death Rates

C.D.R for population

$$A = \frac{\sum_x D_x{}^a}{\sum_x P_x{}^a} \times 1{,}000 = \frac{3{,}640}{1{,}22{,}000} \times 1{,}000 = 29.8$$

C.D.R for population

$$B = \frac{\sum_x D_x{}^b}{\sum_x P_x{}^b} \times 1{,}000 = \frac{3349}{1{,}22{,}000} \times 1{,}000 = 27.4$$

Standard Death Rates

Since population A is taken as Standard Population,

STDR for A = C.D.R for A = 29.8

$$\textbf{STDR for B} = \frac{\sum_x m_x{}^b P_x{}^a}{\sum_x P_x{}^a} = \frac{38{,}34{,}000}{1{,}22{,}000} = 31.4$$

We, Thus conclude that death rate in population B is greater than in population A.

2. Estimate the standardised death rates for the two countries from the data given below.

Death Rate per 1000

Age group in years	Country A	Country B	Standardised Population (in lakhs)
0 - 4	20.00	5	100
5 - 14	1.00	0.50	200
15 - 24	1.40	1.00	190
25 - 34	2.00	1.00	180
35 - 44	3.30	2.00	120
45 - 54	7.00	5.00	100
55 - 64	15.00	12.00	70
65 - 74	40.00	35.00	30
75 and above	120.00	110.00	10

Sol :

Death rate per 1,000					
Age group in years	Country A m_x^a	Country B m_x^b	Standardised population	$m_x^a p_x^s$	$m_x^b p_x^s$
0 - 4	20.0	5.0	100	2,000	500
5-14	1.0	0.5	200	200	100
15-24	1.4	1.0	190	266	190
25-34	3.0	1.0	180	360	180
35-44	3.3	2.0	120	396	240
45-54	7.0	5.0	100	700	500
55-64	15.0	120.0	70	1050	840
65-74	40.0	35.0	30	1200	1050
75 & above	120.0	110.0	10	1200	1,100
Total			1000	7,372	4,700

Standardised Death rate for country A = $(STDR)_A$

$$= \frac{\sum m_x{}^a p_x{}^s}{\sum p_x{}^s}$$

$$= \frac{7372}{1000} = 7.372$$

Standardised Death rate for country B = $(STDR)_B$

$$= \frac{\sum m_x{}^b p_x{}^s}{\sum p_x{}^s} = \frac{4700}{1000} = 4.7$$

3. Find the standarised death rate by Direction and Indirect methods for the data given below.

Standard Population			Population A	
Age	Population	Specific Death rate	Population	Specific Death rate
0 - 5	8	50	12	48
5 - 15	10	15	13	14
15 - 50	27	10	15	9
50 & above	5	60	10	59

Sol :

Computation of SIDR by direct and Indirect method

	Standard Population			Population A				
Age	P_x^s	m_x^s	$p_x^s\ m_x^s$	P_x^a	m_x^a	$p_x^a m_x^a$	$m_x^a p_x^s$	$m_x^s p_x^a$
0 - 5	8000	50	4,00,000	12,000	48	5,76,000	3,84,000	6,00,000
5 - 15	10,000	15	1,50,000	13,000	14	1,82,000	1,40,000	1,95,000
15 - 50	27,000	10	2,70,000	15,000	9	1,35,000	2,43,000	1,50,000
50 & above	50,000	60	3,00,000	10,000	59	5,90,000	2,95,000	6,00,000
Total	50,000		11,20,000	50,000		14,83,000	10,62,000	15,45,000

Direct Method

The points $(\text{STDR})_A = \dfrac{\sum m_x{}^a p_x{}^s}{\sum p_x{}^s} = \dfrac{10,62,000}{50,000} = 21.24$

Indirect method

$$(\text{CDR})_A = \frac{\sum m_x{}^a p_x{}^a}{\sum p_x{}^a} = \frac{1483000}{50,000} = 29.66$$

Adjustment Factor

$$\hat{C} = \frac{\sum m_x{}^s p_x{}^s}{\sum p_x{}^s} \times \frac{\sum p_x{}^a}{\sum m_x{}^s p_x{}^a}$$

$$= \frac{11,20,000}{50,000} \times \frac{50,000}{15,45,000} = 0.7249$$

$$(\text{STDR})_A = \hat{C} \times (\text{CDR})_A$$

$$= 0.7249 \times 29.66 \ = 21.5005 \cdot$$

3.5 Measures of Fertility

3.5.1 Crude birth rate, age specific fertility rate, general fertility rate, total fertility rate

Q5. Explain different types of fertility rates.

Ans : **(July-22, Dec.-21, Oct-20)**

Meaning

In demography, the word fertility is used in relation to the actual production of children (or) occurrence of births, specially live births. Fertility must be distinguished from fecundity which refers to the capacity to bear children. In fact, fecundity provides on upper band for fertility. As a measure of the rate of growth of population various fertility rates are computed.

The following are the principle rates used in measurement of fertility.

(1) Crude birth rate (C.B.R)

(2) General Fertility rate (GFR)

(3) Specific Fertility Rate (S.F.R)

(4) Age Specific Fertility Rate (A.S.F.R)

(5) Total Fertility Rate (T.F.R).

1. Crude Birth Rate (C.B.R)

This is the simplest of all the measures of fertility and consists in relating the number of live births to the total population. This provides an index of the relative speed at which additions are being made through child birth. The fertility pattern of the above mentioned measure is given by crude birth rate (C.B.R) defined as follows.

$$\text{Crude Birth Rate} = \frac{B^t}{P^t} \times k$$

Where

B^t = Total number of live births in the given region or locality drawing a given period, say t.

P^t = Total population of the given region during the period t.

K = A constant, usually 1000.

Merits

It is simple, easy to calculate and readily comprehensible.

It is based only on the number of births (B^t) and the total size of the population (P^t) and does not necessitate the knowledge of these figures for different sections of the community or the population.

Demerits

- The crude birth rate, though simple, is only a crude measure of fertility and is unreliable since it completely ignores the age and sex distribution of the population.
- C.B.R is the not a probability ratio, since the whole population P^t cannot be regarded as Exposed to the risk of producing children . In fact, only the females and only those between the child bearing age group (usually 15 to 49 years) are Exposed to risk and as such whole of the male population and the female population outside the child-bearing age should be Excluded from P^t, the demoninator in ....(1). Moreover, even among the females who are Exposed to risk, the risk varies from one age group to another, a woman under 30 is certainly under greater risk as compared to a woman over 40.
- As a consequence of variation of climatic conditions in various countries, the child bearing age groups are not identical in all the countries. In tropical countries, the period starts at an apparent earlier date than in countries with cold weather. Accordingly, crude birth rate does not enable us to compare the fertility situation in different countries.
- Crude birth rate assumes that women in all the ages have the same fertility, an assumption which is not true since younger women have, in general higher fertility than elderly women. CBR thus gives us an estimate of a hetrogeneous figure and is unsuitable for comparative studies.

- The level of crude birth rate is determined by a number of factors such as age and sex distribution of the population, fertility of the population, sex, ratio, marriage rate, migration, family planning measures and so on.

2. General Fertility Rate (G.F.R)

This consists in relating the total number of live births to the number of females in the reproductive (or) child bearing ages and is given by the formula.

$$\text{General Fertility Rate (GFR)} = \frac{\text{Number of live births}}{\text{Number of Women of child bearing age}} \times 1000$$

Merits

- General fertility rate is a probability rate since the denominator in .... (2) consists of the entire female population which is Exposed to the risk of producing children.
- G.F.R reflects the Extent to which the female population in the reproductive ages increases the Existing Population through live births. Obviously G.F.R takes into account the sex distribution of the population and also the age structure to a certain extent.

Demerits

G.F.R gives a heterogeneous figure since it overlooks the age composition of the female population in the child bearing age.

Hence it suffers from the draw back of non- comparability in respect of time and country.

3. Specific Fertility Rate (S.F.R)

The concept of specific fertility rate originated from the fact fertility is affected by a number of factors such as age, marriage, migration, state or region etc. The Fertility rate computed with respect to any specific factor is called specific fertility rate (S.F.R) and is defined as

$$\text{S.F.R} = \frac{\text{No. of births to the female population of the specific section in a given period}}{\text{Total number of female population in the Specified section}} \times \text{K.} \qquad \text{... (3)}$$

Where K = 1000, usually.

4. Age Specific Marital Fertility Rate (ASMFR)

The Age Specific Marital Fertility Rate (ASMFR) is directly concerned with married women belongs to a age specific group. It is defined as the number of child bom alive in a given year per thousand married women of specific age group.

The formula used to calculate ASMFR is as follows,

$$\text{ASMFR} = \frac{\text{Number of live births to married women}}{\text{Number of married women of child bearing age}} \times 1000$$

5. Total Fertility Rate (TFR)

Under Total Fertility Rate (TFR) all the women who are in the reproductive age are considered. TFR is the mean number of children which a women around the age of 15 can expect to bear if she lives until the age of 50 and subject to the given fertility conditions over the whole of her child bearing period. Flowever, for calculating TFR, Age Specific Fertility Rate (ASFR) is required.

The formula used for calculating TFR is as follows,

TFR = Sum of ASFR ^{x} C

Where,

C = Interval or Magnitude of the age class.

(or)

$$TFR = \sum \frac{SFR}{1000}$$

PROBLEMS

4. Calculate the age specific fertility rate, general fertiltiy rate and total fertility rate from the following information:

Age Group (Years)	No. of Women ('.000)	No. of Live Births
05 - 09	60	200
10 - 14	80	800
15 - 19	30	1,000
20 - 24	20	600
25 - 29	60	400
30 - 34	20	600
35 - 39	10	700
40 - 44	80	1,400
45 - 49	40	1,300
Total	400	7,000

Sol:

Age Group (Years)	No. of Women ('.000)	No. of Live Births	Age specific Fertility Rate (ASFR)
05 - 09	60	200	3.33
10 - 14	80	800	10.00
15 - 19	30	1,000	33.33
20 - 24	20	600	30.00
25 - 29	60	400	6.66
30 - 34	20	600	30.00
35 - 39	10	700	70.00
40 - 44	80	1,400	17.50
45 - 49	40	1,300	32.50
Total	400	7,000	ΣASFR = 233.32

Working Notes

$$\text{ASFR} = \frac{\text{Number of live births of female in the age group x to (x + c)}}{\text{Average number of female in the age group x to (x + c)}} \times 100$$

$$05 - 09 = \frac{200}{60,000} \times 1,000 = 3.33$$

$$10 - 14 = \frac{800}{80,000} \times 1,000 = 10.00$$

$$15 - 19 = \frac{1,000}{30,000} \times 1,000 = 33.33$$

$$20 - 24 = \frac{600}{20,000} \times 1,000 = 30.00$$

$$25 - 29 = \frac{400}{60,000} \times 1,000 = 6.66$$

$$30 - 34 = \frac{600}{20,000} \times 1,000 = 30.00$$

$$35 - 39 = \frac{700}{10,000} \times 1,000 = 70.00$$

$$40 - 44 = \frac{1,400}{80,000} \times 1,000 = 17.50$$

$$45 - 49 = \frac{1,300}{40,000} \times 1,000 = 32.50$$

Calculation of General Fertility Rate (GFR)

$$\text{General Fertility Rate (GFR)} = \frac{\text{Number of live births}}{\text{Number of women of child bearing age}} \times 1,000$$

$$= \frac{7,000}{4,00,000} \times 1,000$$

General Fertility Rate (GFR) = 17.5

Calculation of Total Fertility Ratio

Total Fertility Ratio (TFR) = Sum of ASFR × C

= 233.32 × 5

∴ Total Fertitlity Ratio (TFR) = 1,166.60

3.6 Measurement of Population Growth

3.6.1 Crude rate of natural increase- Pearl's vital index. Gross reproductive rate sand Net reproductive rate

Q6. Write a detail note on population growth and how it can be measured.

Ans : **(July-21, Oct.-20)**

Fertility rates are inadequate to give as any idea about the rate of population growth since they ignore the sex of the newly born children and their mortality. Obviously the population increases though female births. Thus if a majority of births are those of girls, the population is bound to increase while it will have a downward trend if the majority of births are boys. Similarly if we ignore the Mortality of the newly born children we cannot form a correct idea of the rate of growth of the population, since it is possible that a number of female children may die before reaching the reproductive age. In the following sections we shall study some measures of the growth of population under the assumption that in future also it is subject to the current fertility and mortality rates.

There are four measures of population growth.

(1) Crude rate of Natural increase

(2) Pearls vital index

(3) Gross Reproductive rates (G.R.R)

(4) Net Reproductive rates (N.R.R)

1. Crude Rate of Natural Increase

The simplest measure of the population growth known as crude rate of Natural increase is defined as the difference between the crude birth rate (per thousand) and the crude death rate (per thousand) and is given by

Crude Rate of Natural increase = C.B.R – C. D. R

2. Pearl's Vital Index

Another indicator of population growth based on births and deaths taken together is provided by R. pearl's vital index. Defined as follows :

$$\text{Pearl's vital index} = \frac{\text{No. of births in the given period t}}{\text{No. of deaths in the given period t}}$$

$$= \frac{B^t}{D^t} \times 100 \qquad \text{... (1)}$$

Dividing numerator and denominator in ...... (1) by the population P^t in the given period t, we get

$$\text{Pearl's vital index} = \frac{\text{C.B.R}}{\text{C.D.R}} \times 100$$

3. Gross Reproductive Rate

The population growth mainly depends on the birth of female children who are the future mothers. In order to have a better idea about the rate of population growth, In addition to the age and sex composits of the population. We must take into account the sex of the new born children because it is ultimately the female births who are the potential future mother's and result in an increase in the population. One of such measure is the G.R.R. and it is defined as the sum of the annual female age specific fertility rate.

The composition of Gross Reproductive rate requires the classification of births and the sex of the newly born babies. In such cases an approximate value of G.R.R. may be obtained under the assumption that sex ratio of birth remain more (or) less constants at all the age of the human in the reproductive period and is given by

$$G.R.R = \frac{\text{Total no. of female births}}{\text{No. of births}} \times T.F.R$$

4. Net Reproductive rates (N.R.R)

The demographic year book of United Nations in 1954 define Net Reproductive Rate (NRR) as the rate may be interpreted as the average number of daughters that would be produced by women throughout their life time if they were exposed at each age to the fertility and mortality rates on which the calculation is based.

In simple words, NRR can be defined as the measure which helps to calculate the extent to which the female infants who may continue to survive their maximum reproductive age are able to reproduce the infants of their same sex.

The formula for NRR is as follows,

$$NRR = \frac{(\text{S Number of Female Births} \times \text{Survival Rate})}{1{,}000}$$

(or)

$$NRR = \Sigma(\text{Specific Fertility Rate x Survival Rate}) \text{x CI}$$

Where,

CI = Class Interval

(or)

$$NRR = \frac{\text{Number of Female Children bom and Survived to 1,000 Women}}{1{,}000}$$

PROBLEMS

Q5. Calculate the gross and net reproduction rates from the data given below: number of women in age groups and number of female children born in one year.

Aqe Group	Female Population ('000)	Femal Births ('000)	Survival Rate
05-09	1,500	30	0.900
10-14	1,200	20	0.960
15-19	1,400	10	0.850
20-24	1,800	60	0.888
25-29	1,000	80	0.720
30-34	2,000	70	0.999
35-39	1,500	30	0.820
40-44	1,700	60	0.860
45-49	900	40	0 890

Sol :

Aqe Group	Female Population ('000)	Femal Births ('000)	Age Specific Birth Rate per Women (ASBR)
05 - 09	1,500	30	0.020
10 - 14	1,200	20	0.017
15 - 19	1,400	10	0.007
20 - 24	1,800	60	0.033
25 - 29	1,000	80	0.080
30 - 34	2,000	70	0.035
35 - 39	1,500	30	0.020
40 - 44	1,700	60	0.035
45 - 49	900	40	0.044
Total	**13,000**	**400**	**0.291**

Working Notes

$$\text{Age Specific Fertility Rate (ASFR)} = \frac{\text{Female Births}}{\text{Female Population}} \times 1{,}000$$

$$05 - 09 = \frac{30{,}000}{15{,}00{,}000} = 0.020$$

$$10 - 14 = \frac{20{,}000}{12{,}00{,}000} = 0.017$$

$$15 - 19 = \frac{10{,}000}{14{,}00{,}000} = 0.007$$

$$20 - 24 = \frac{60{,}000}{18{,}00{,}000} = 0.033$$

$$25 - 29 = \frac{80{,}000}{10{,}00{,}000} = 0.080$$

$$30 - 34 = \frac{70{,}000}{20{,}00{,}000} = 0.035$$

$$35 - 39 = \frac{30{,}000}{15{,}00{,}000} = 0.020$$

$$40 - 44 = \frac{60{,}000}{17{,}00{,}000} = 0.035$$

$$45 - 49 = \frac{40{,}000}{9{,}00{,}000} = 0.044$$

Calculation of TFR

TFR = Sum of ASFR = C

= 0.291 × 5

∴ TFR = 1.455

Calculation of Grosss Reproduction Rate

$$\text{Gross Reproduction Rate (GRR)} = \text{TFR} \times \frac{\text{Number of Female Total Births}}{\text{Total Births}}$$

$$= 1.455 \times \frac{400}{13,000} = 1.455 \times 0.031$$

∴ GRR = 0.045

Calculation of Net Reproduction Rate

Age Group	Female Population ('000)	Female Births ('000)	Age Specific Fertility Rate Per Women (ASFR)	Survival Rate (S)	ASFR × S
05-09	1,500	30	0.020	0.900	0.018
10-14	1,200	20	0.017	0.960	0.016
15-19	1,400	10	0.007	0.850	0.006
20-24	1,800	60	0.033	0.888	0.029
25-29	1,000	80	0.080	0.720	0.058
30-34	2,000	70	0.035	0.999	0.035
35-39	1,500	30	0.020	0.820	0.016
40-44	1,700	60	0.035	0.860	0.030
45-49	900	40	0.044	0.890	0.039
Total	**13,000**	**400**	**0.291**	**7.887**	**Σ(ASFR × S) = 0.247**

NRR = S (Specific Fertility Rate × Survival Rate) × CI

= 0.257 × 5

∴ NRR = 1.235

6. Compute G.F.R, T.F.R for the data given below.

Age of women	No. of womens	A.S.F.R (per 1000)
15 - 19	212, 619	98.0
20 - 24	198,732	169.6
25 - 29	162, 800	158. 2
30 - 34	145,362	139.7
35 - 39	128,109	98.2
40 - 44	106, 211	42.8
45 - 49	86, 753	16.9

Sol :

No. of womes	A.S.F.R.	Total Births
212,619	98.0	$\frac{ASFR \times W_x}{1000} = \frac{98.0 \times 212,619}{1000} = 20836.662$
198,732	169.6	33704.9472
162,800	158.2	25754.96
145,362	139.7	20300.0864
128,109	98.2	12580.3038
106,211	42.8	4545.8308
86,753	16.9	1466.1257
1040586	723.4	119195.9009

$$\text{G.F.R} = \frac{B}{\sum f_{P_x}} \times 1000 = 114.5456$$

$$\text{T.F.R} = 5\,\Sigma(\text{A.S.F.R}) = 5 \times 723.4 = 3617$$

$$\therefore \text{T.F.R} = 3617$$

3.7 Life tables, construction and uses of Life Tables

Q7. Define Life table. What are the assumptions of Life table.

(OR)

What is life table.

Ans : **(July-22, Dec.-21)**

Meaning

The life table gives the life history of a hypothetical group or cohort as it is gradually diminished by deaths. It is a conventional method of Expressing the most fundamental and essential facts about the age distribution of mortality in a tabular form and is a powerful tool for measuring the probability of life and death of various age sectors. A life table provides answers to the following questions.

(i) How will a group of infants all born at the same time and experiencing unchanging mortality conditions throughout the life time, gradually die out ?

(ii) When in the course of time all these infants die, what would be the average longevity per person?

(iii) What is the probability that persons of specified age will survive a specified number of years.

(iv) How many persons, out of selected number of persons living at some initial age, survive on the average to each attained age.

The life table thus gives a summary of the mortality experience of any population group during a given period and is a very effective and comprehensive method for providing concise measures of the longevity of that population.

The data for constructing a life table are the census data and death registration data. Life tables are generally constructed for various sections of the people which, as experience shows, have sharply different patterns of mortality thus there are life tables constructed for different races, occupational groups and sex. Life tables are as well constructed on regional basis and other factors accounting differential mortality.

Definition

It is a conventional method of expressing the most fundamental and essential facts about the age distribution of mortality in a tabular form. It is a powerful tool for measuring the probability of life death of various age stress.

The life table gives the life history of hypothetical group as it is gradually decreased by deaths.

Assumptions

The assumptions which are used in the construction of life tables are

- The hypothetical group is closed for migration i.e., there is no change in the group expect the losses due to deaths.
- The individuals die at each age according to some predetermined schedule which is fixed and doesn't change.
- The deaths are distributed uniformly over the period of age 'x' to 'x+1'.

Q8. Explain various components of life tables.

Ans : **(July-22)**

The life table is constructed with the help of different columns and every column will have its own value or calculation. Usually, a life table begins with a population of 1,00,000 people but as per the conveniency life table can be constructed for any amount of population like 1,000, 10,000 etc. Such figure of table is known as 'Radix' of table.

The various columns and structure of life table are as follows,

Column 1

Column 1 indicates the age in years with a symbol 'x' (exact age in years till last birthday).

Column 2

Column 2 indicates persons living at age 'x' with a symbol of 'L_x'. It means number of survivors at the exact age 'x' out of the population group (i.e., 1,00,000, 10,000, 1000 etc).

Column 3

Column 3 indicates persons dying between age x and x + 1 = $d_x = L_x - L_{x+1}$, d_x means the number death in the interval x to (x + 1) in the initial population group (i.e., 1,00,000, 10,000, 1000 etc).

Column 4

Column 4 indicates the probability of persons dying between age x and x + 1 = $q_x = \frac{d_x}{L_x}$.

Column 5

Column 5 indicates the probability of persons surviving between age x and x + 1 = p_x = 1 – $q_,$ = $\frac{L_{x+1}}{L_x}$

Column 6

Column 6 indicates the persons living between the age x and x + 1 = $L_x = \frac{L_x + L_{x+1}}{2} = L_x - \frac{1}{2} \times d_x$

Column 7

Column 7 indicates the persons living above the age x = $T_x = L_x + L_{x+1} + ... = L_x + T_{x+1}$

$\therefore \quad T_{x+1} = T_x - L_x$.

Column 8

Column 8 indicates the expectation of life at age x = $e_x^{\alpha} = \frac{T_x}{L_x}$

The illustrative life table with imaginary information is constructed below,

x (i)	L_x (ii)	d_x (iii)	q_x (iv)	P_x (v)	L_x (vi)	T_x (vii)	e_x^{θ} (viii)
0	1,00,000	9,872	.09872	.90128	95,064	57,68,318	57.68
1	90,128	3,939	.04370	.95630	88,158	56,73,254	62.95
2	86,189	2,654	.03079	.96921	84,862	55,85,096	64.80
3	83,535	2,476	.02964	.97036	82,297	55,00,234	65.84
4	81,059	1,772	.02186	.97814	80,173	54,17,937	66.84
5	79,287	1,234	.01556	.98444	78,670	53,37,764	67.32
6	78,053	987	.01264	.98736	77,560	52,59,094	67.38
7	77,066	765	.00993	.99007	76,684	51,81,534	67.24
8	76,301	477	.00625	.99375	76,062	51,04,850	66.90
9	75,824	462	.00609	.99391	75,593	50,28,788	66.32
10	75,362	418	.00555	.99445	75,153	49,53,195	65.72
-	-	-	-	-	-	-	-
-	-	-	-	-	-	-	-
-	-	-	-	-	-	-	-
-	-	-	-	-	-	-	-
96	371	158	.42588	.57412	292	788	2.12
97	267	104	.38951	.61049	215	496	1.86
98	210	57	.27143	.72857	182	281	1.34
99	181	291	.16022	.83978	166	99	0.55

3.8 Abridged Life Tables

Q9. Write in detail about Abridged life tables.

Ans : **(July-22)**

Meaning

A abridged life table is similar to the complete life table under which most of the important ages are mentioned for a period of more than 1 year that could be 5 years or 10 years.

The abridged life table can be available in two different forms or types. Such as,

1. It could be available in the form where some integral values of x which are at some distance apart, usually 5 or 10 years.
2. It could be available in the form where the values of the age group x are stated for 5 or 10 years.

Construction of Abridged Life Table

The abridged life table is constructed with different columns, which are as follows,

Column I

In the first column of abridged life table, the age intervals x to (x + n), i.e., x_0, x_0 + n, x_0 + 2_n.... will be mentioned.

Column 2

The second column is L_x which indicates the number of persons out of a group of L_x persons who are living at the beginning of the interval x to (x + n).

Column 3

The third column is $_nq_x$ which indicates the probability of the person dying in the age interval x to (x + n). It is expressed as,

$$_nq_x = -\ _np_x = 1 - \frac{L_{x+n}}{L_x}$$

Column 4

Fourth column is $_nd_x$ which indicates the number of deaths occurred in the age interval x to (x + n). It is expressed as,

$$_nq_x = \frac{_nd_x}{L_x} = \ _nd_x = L_x \times \ _nq_x$$

Column 5

Fifth column is $_nL_x$ which indicates the number of members of the life table stationary population in the age group (x, x + n). It is expressed as,

$$_nL_x = \int_0^n L_{x+t}dt = \frac{n}{2}(L_x + L_{x+n})$$

Column 6

Sixth column is $T_x = \int_0^\infty L_{x+t}\, dt$ which indicates the number of persons lived after age x or the number of members of the life table stationary population of age x or above.

Column 7

Seventh column is $e_x^0 = \frac{T_x}{L_x}$ which indicates complete expectation of life at age *x.*

x(1)	L_x(2)	${}_nq_x$(3)	${}_nd_x$(4)	${}_nL_x$(5)	T_x(6)	e_x^0(7)

Fig.: Structure of Abridged Life Table

Q10. Explain Reed Merrell Method of Abridged life tables.

Ans : **(Imp.)**

This method due to L.J. Reed and M. Merrell is bad on the following fundamental result which we state in the form of a lemma.

Lemma. $${}_nq_x^z = \frac{2n({}_nm_x^z)}{2+n({}_nm_x^z)} \quad \ldots (1)$$

Where ${}_nP_x^z$ and ${}_nd_x^z$ are respectively the average, number of persons and the number of deaths between ages x to (x + n) in the calender year z and

$${}_nm_x^z = \frac{{}_nd_x^z}{{}_nP_x^z}$$

is the central rate of mortality in the calender year z, n being the length of the age group.

Proof

Let the life table population be $l_x = E_x^z$.

Assuming that deaths are uniformly distributed in the interval (x, x + n) or equivalently assuming the linearit of l_{x+t} for $t \in [0, n]$, we get

$${}_nP_x^z = \int_0^n l_{n+t}\, dt \simeq \frac{n}{2}(l_x + l_{n+x})$$

$$= \frac{n}{2}\,[l_x + (l_x - {}_nd_x^z)]$$

$$= nE_x^z - \frac{n}{2}({}_nd_x^z) \qquad (\because l_x = E_x^z)$$

$$\Rightarrow \qquad E_x^z = \frac{1}{n}({}_nP_x^z) + \frac{1}{2}({}_nd_x^z)$$

By definition, we have

$$ {}_nq_x^z = \frac{{}_nd_x^n}{E_x^z} = \frac{{}_nd_x^n}{\frac{1}{n}({}_nP_x^z) + \frac{1}{2}({}_nd_z^x)} $$

$$ = \frac{\dfrac{{}_nd_x^x}{{}_nP_x^z}}{\dfrac{1}{n} + \dfrac{1}{2}\left(\dfrac{{}_nd_x^z}{{}_nP_x^z}\right)} = \frac{{}_nm_x^z}{\dfrac{1}{n} + \dfrac{1}{2}({}_nm_x^z)} = \frac{2n({}_nm_x^z)}{2 + n({}_nm_x^z)} $$

Description of the Method

For each possible value of x, the values of ${}_nP_x^z$ and ${}_nd_x$ are known from the census and registration data respectively. Using these values we can find ${}_nm_x^z$ by the relation.

Thus starting with given radix l_x, we can find other values l_{x+ns} l_{s+2ns} and so on by the relation.

$l_{x+ns} = lx \,.\, {}_np_x$

$l_{x+ns} = ls \,.\, {}_np_x + n$

and so on where ${}_np_x = 1 - {}_nq_x$. Other functions of the life table can now be obtained in the usual way.

Q11. Explain King's Method of Abridged life tables.

Ans : **(Imp.)**

This method due to G. King is intended if the life table functions (l_x, q_x, L_x, ...) are to be obtained for the values of x at some distance apart, say 5 years or 10 years and is based on the valid application of King's interpolation formula.

In the usual notations let ${}_nP_x$ be the observed population and nDx be the number of deaths in age group [x, x + n]. Then we can write

$$ {}_nP_x = P\,x - \left[\frac{n-1}{2}\right]^{+P} x + 1 - \left[\frac{n-1}{2}\right]^{+...+P} x + \left[\frac{n-1}{2}\right] $$

and $$ {}_nD_x = D\,x - \left[\frac{n-1}{2}\right]^{+D} x + 1 - \left[\frac{n-1}{2}\right]^{+...+D} x + \left[\frac{n-1}{2}\right] $$

where P_x and D_x are respectively the population and the number of deaths for the age x and [(n–1)/2] stands for the larget positive integer contained in (n–1)/2.

The main purpose is to obtain an estimate of the population P_x^0 and the deaths D_x^0 for the centralage in the age group [x, x + n], from the given values of ${}_nP_x$ and ${}_nD_x$. Under the assumption that P_x^0 and D_x^0 can be approxiated by a second degree parabola, Kind obtained their estimates from the formulae.

$$P_x^0 = \frac{1}{n}({}_nP_x) - \frac{1}{24}\left(\frac{1}{n}\right)\left(1-\frac{1}{n^2}\right)\Delta^2(nP_x)$$

$$D_x^0 = \frac{1}{n}({}_nD_x) - \frac{1}{24}\left(\frac{1}{n}\right)\left(1-\frac{1}{n^2}\right)\Delta^2(nP_x)$$

Using the estimates of P_x^0 and D_x^0 obtained from, the central rate of mortality at age x is given by

$$m_x = \frac{D_x^0}{P_x^0}, \ (x = x_0, x_0 + n, x_0 + 2_n, ...)$$

and consequently the pivotal column of the life table, viz., q_x is obtained from the relation.

$$a_x = \frac{2m_x}{2+m_x}, \ (x = x_0, x_0 + n, x_0 + 2_n, ...)$$

assuming that death are uniformly distributed over the given interval. Next we obtain p_n, the probability of survival at age x by the relation.

$$p_x = 1 - q_x, \ (x = x_0, x + n, x_0 + 2_n, ...)$$

For the remaining columns of the life table, we obtain i_x for the ages $x = x_0, x_0 + n, x_0 + 2n$, etc., by the relation.

$$l_{x+n} = l_x \ ({}_np_x)$$

$$l_{x+2n} = l_{x+n} \ ({}_np_{x+n})$$

and so on, where ${}_ap_x$ is the probability that a person aged x survives next n years.

${}_np_x$ is obtained from the value of p_x computed in by using the formula.

$${}_np_x = p_{x.}p_{x+1} \ ..... \ p_{x+n-1} = \prod_{i=0}^{n-1} p_{x+i}$$

$$\Rightarrow \qquad \log({}_np^x) = \sum_{i=0}^{n-1} \log P_{x+i}$$

King obtained the values of ${}_np_x$ from the available values of px by using Everett's Central Difference formula, viz.,

$$U_{x+n} = [y\, U_{x+n} + \frac{y(y^2-1)}{3!}\Delta^2 U_x + ...] + [tU_x + \frac{t(t^2-1)}{3!} D^z U_{x-n} + ...]$$

where $0 \le h \le n$, $y = h/n$ and $t = 1 - y$

Taking $U_x = \log p_x$ and h = 1, 2, ..., (n –1) successively in (*), we get correct up to second order differences,

$$\log p_{x+1} = \frac{1}{n}\log p_{x+n} + \left(1-\frac{1}{n}\right)\log p_x + \frac{1}{3!}\ \frac{1}{n}\left(\frac{1}{n^2}-1\right)\Delta^2 \log p_x$$

$$+\ \frac{1}{3!}\left(1-\frac{1}{n}\right)\left[\left(1-\frac{1}{n}\right)^2-1\right]\Delta^2 \log P_{x-n}$$

$$\log P_{x+2} = \frac{2}{n}\log p_{x+n} + \left(1-\frac{2}{n}\right)\log p_x + \frac{1}{3!}\cdot\frac{2}{n}\left[\left(\frac{2}{n}\right)^2-1\right]\Delta^2 \log px$$

$$+\ \frac{1}{3!}\left(1-\frac{2}{n}\right)\left[\left(1-\frac{2}{n}\right)^2-1\right]\Delta^2 \log p_{x-n}$$

$$\vdots \qquad \vdots \qquad \vdots$$

$$\log p_{x+n-1} = \frac{n-1}{n}\log p_{x+n} + \left(1-\frac{n-1}{n}\right)\log p_x$$

$$+\ \frac{1}{3!}\left(\frac{n-1}{n}\right)\left[\left(\frac{n-1}{n}\right)^2-1\right]\Delta^2 \log p_x$$

$$+\ \frac{1}{3!}\left(1-\frac{n-1}{n}\right)\left[\left(1-\frac{n-1}{n}\right)^2-1\right]\Delta^2 \log p_{x-n}$$

Adding the above results, we get

$$\log ({}_np_x) = \sum_{i=0}^{n-i}\log\ p_{x+1} = \log p_x + \log p_{x+1} + \log p_{x+2} + \ldots + \log p_{x+n-1}$$

$$= \left(n-\frac{1+2+\ldots+(n-1)}{n}\right)\log p_x + \left(\frac{1+2+\ldots.+(n-1)}{n}\right)\log p_{x+n}$$

$$-\ \frac{n^2-1}{24n}[\Delta^2 \log p_x + \Delta^2 \log p_{x-n}] \text{ (on simplification)}$$

$$= \frac{n+1}{2}\log p_x + \frac{n-1}{2}\log p_{x+n}$$

$$= \frac{n^2-1}{24n}[\Delta^2 \log p_x + \Delta^2 \log p_{x-n}]$$

King obtained the values of ${}_np_x$ form and then using obtained the values of l_x f or $x = x_0, x_0 + n, x_0 + 2_n$ and so on.

Next column in the life table is to obtain $T_{x,n}$*, the number of years lived by the radix l_x during the age interval [x, n + n] and under the assumption that deaths are distributed uniformly over the interval [x, x+n] is approximately given by,

$$T^*_{x:n} = L_x + L_{x+1} + \ldots L_{x+n-1}$$

$$= \frac{1}{2}(l_x + l_{x+1}) + \frac{1}{2}(l_{x+1} + l_{x+2}) + \ldots + \frac{1}{2}(l_{x+n-1} + l_{x+n})$$

$$= (l_x + l_{x+1} + \ldots + l_{x+n-1}) - \frac{1}{2}(l_x - l_{x+n})$$

$$= \sum_{i=0}^{n-1} l_{x+n} - \frac{1}{2}\, l_x \left[1 - \frac{l_{x+n}}{l_x}\right]$$

On using with log p_x replaced by l_x, we get up to second order differences.

$$T_{x:n}^* = \frac{n+1}{2}\, l_x + \frac{n-1}{2}\, l_{x+n} - \frac{n^2-1}{24n}(\Delta^2 l_x + \Delta^2 l_{x-n}) - \frac{1}{2} l_x$$

$$\Rightarrow \quad T^*_{x:n} = N^*_{x:n} - (l_2\, l_x)$$

where $\quad N^*_{x:n} = \sum_{i=0}^{n-1} l_x + i$

is the total number of complete years lived by l_x persons having ages x to (x + n).

Finally, e_x^0, the expected number of years lived by the l_x persons before crossing the age interval (x, x+n) is obtained from the formula.

$$e_x^0 = \frac{T^*_{x:n}}{l_x}$$

$$= \frac{N^*_{x:n}}{l_x} = \frac{1}{2}\cdot {}_nq_x$$

Q12. Explain Greville's Method of Abridged life tables.

Ans :

For the estimation of ${}_nq_x$ from the observed age-specific death rate ${}_nm_x$, Greville used

$$_nq = \frac{2n\,({}_nm_x)}{1 + {}_nm_x\left[n + \frac{n^2}{6}\left({}_nm_x - \log_e C\right)\right]} \quad \ldots (1)$$

where c is estimated from the assumption that ${}_nm_x$ follows Gompertz (exponential) law.

$$_nm_x = BC^x \quad \ldots (2)$$

The formula (1) may be regarded as a refinement over the formula (2) after allowing for departure of linearity by Euler-Maclaurin's summation formula and enables us to obtain the values of ${}_nq_x$ for mode rate values of n.

As in the construction of complete life table, the values of ${}_nq_w$ for early ages (x = 0, 1, 2, ... etc.) are obtained by any of the method involving birth and death statistics. From these probabilities we can get a value l_x which will serve as a radix for the construction of the abridged life table. We have

$${}_nd_x = l_x \cdot {}_nq_x \qquad \text{... (3)}$$

and $${}_np_x = \frac{l_{x+n}}{l_x} = 1 - {}_nq_x$$

$$\Rightarrow \quad {}_nL_x = l_x\,(1 - {}_nq_n) = l_x - {}_ndx \qquad \text{... (4)}$$

Where ${}_nd_x$ is the total number of deaths in the life table stationary population in the age sector x to x + n.

Now starting with he radix l_x and computing the ${}_nq_x$ values from the relation, we can obtain the values of l_{x+n}, l_{x+2n}... by using.

Next step is to compute the ${}_nL_x$ column for the abridged life table. If we assume that central death rate ${}_nm_x$ in the life table stationary population then by definition, we get

$${}_nm_x = \frac{{}_nd_x}{{}_nL_x}$$

$$\Rightarrow \quad {}_nL_m = \frac{{}_nd_x}{{}_nm_x} \qquad \text{... (5)}$$

where ${}_nm_x$ are given values and ${}_nd_x$ are computed from.

Another approximation to ${}_nL_x$, based on numerical quadrature is given by the formula

$${}_nL_x = \int_0^n l_{n+t}\,dt$$

$$= \frac{n}{2}(l_x + l_{x+n}) + \frac{n}{24}({}_nd_{x+n} - {}_nd_{x-n})$$

and it provides more accurate results as compared with

If ω + n is the terminal age, i.e., the age at which l_m vanishes, which implies that $l_{w+n} = 0$, then we get from and

$${}_nL_w = \frac{{}_nd_w}{{}_nm_w} = \frac{l_w}{{}_nm_w}$$

The next column of the table, *viz*.,

$$T_x = {}_nL_x + {}_nL_{x+n} + ... + {}_nL_w$$

is obtained from the ${}_nL_x$ column by taking cumulative totals starting from the bottom of the table. The calculations may be simplified by using the relation.

$$T_x = {}_nL_x + T_{x+n}$$

Finally the last column giving the expectation of life is obtained from the relation.

$$e_x^0 = \frac{T_x}{l_x}$$

Remarks

1. It may be pointed out that before applying this method, it must be seen if the assumptions that ${}_nm_x$ follows Gompertz law is justified by he observed data.
2. An estimate of C may be obtained from an average of the values.

$$[{}_nm_{x+n}/{}_nm_x]^{1/n}$$

Short Question and Answers

1. Explain the uses of vital statistics.

Ans :

Uses

Vital statistics are being Extensively used in almost all the spheres of human activity. Given below are some of the important applications of vital statistics.

i) Study of Population Trend

The study of births and deaths gives us an idea of the population trend of any region. Community (or) country.

If ‘Birth rate > Death rate ; there is an increasing trend . If Birth rate < Death rate, there is decreasing trend.

The division of the population of different regions by birth and death rates enables is to form some idea about the population trend of the regions or countries and the general standard of living and virility of the races.

In under developed countries, the birth rate is fairly high but at the same time it is accompanied by high infant mortality rate showing thereby the lack of medical facilities poor hygenic conditions, malnutrition and low standard of living.

ii) Use of Public Administration

The study of population movement i.e., population estimation, population projections and other allied studies together with birth and death statistics according to age and sex distributions provides any administration with fundamen-tal tools which are indispensable for the overall planing and evaluation of economic and social development programes.

iii) Mortality

Mortality and natality statistics also provide guide sports for use by the researchers in Medical and pharmaceutical profession.

iv) Use to Operating Agencies

The facts relating to births, deaths and marringes are of Extreme importance to various official agencies for a variety of administrative purposes. Mortality statistics serve as a guide to the health authorities for sanitary improvements, improved medical facilities and public cleanliness.

2. Standardized Death Rates

Ans :

The crude death rates in terms of age specific death rates for two regions A and B are given respectively by

$$m^a = \frac{D^a}{P^a} \times 1000 = \frac{\sum_x m_x^a p_x^a}{\sum_x P_x^a} \quad \ldots (1)$$

$$\text{and } m^b = \frac{D^b}{P^b} \times 1{,}000 = \frac{\sum_x m_x^b P_x^b}{\sum_x P_x^b} \quad \ldots (2)$$

The Expressions in (1) & (2) are nothing but the weighted arithmetic means of the age S.D.R., the weights being the corresponding populations in the age groups we observe that even if A.S.D.R’S are some, i.e.,

$m_x^a = m_x^b \ \forall x,\ m^a \neq m^b$ since in general,

$$\frac{P_x^a}{\sum_x P_x^a} \neq \frac{P_x{}^b}{\sum_x P_x^b}$$

i.e., Since the age distributions of the populations in the two regions A and B are not identical. This drawback is removed if the same set of weights is used in $\rightarrow$(1) & (2) for computing the weighted average of the A.S.D.R’S. This is what is done in standardized death rates or adjusted death rates, used with a prefix to identify the basis of adjustment as, for Example, age-adjusted death rates and so on. We discuss below the two methods of age adjustments which are in common use.

(i) Direct Method of Standardization

This method consists in weighting the age specific death rates not by the corresponding population of the area to which they refer (as is done in C.D.R). But by the population contribution of another region chosen as a standard Thus if P_x^S is the number of persons in the age - group x to x +1 in the standard population, then the standardised death rates for the regions A and B are given respectively by

$$(\text{STDR})_A = \frac{\sum_x m_x^a P_x^s}{\sum_x P_x^s} \text{ and}$$

$$(\text{STDR})_B = \frac{\sum_x m_x^a P_x^s}{\sum_x P_x^s}$$

These age adjusted death rates for regions A and B respectively are nothing but the crude death rates that would be observed in the standard population if it were subject to the age S.D.R of the regions A and B.

(ii) Indirect Standardization

In compiling the standardised death rates by formula in direct method. It is necessary to know the number of persons and the age - specific death rates for different age -groups. Quite often we have a population classified by age but the age S.D.R's may not be known . However, the total number of deaths and hence C.D.R. may be known.

In such case, we use the indirect method of standardisation which consists in multiplying the crude death rate of the region A, say, by adjustment factor 'c' measuring the relative 'mortality'. Proneness of the population of the region such that the result is equal to the standardised death rate. Thus the problem is to find C such that.

ST. D.R = C.D.R × C

$$\frac{\sum_x m_x{}^s P_x{}^s}{\sum_x P_x{}^s} = C \times \frac{\sum_x m_x{}^s p_x{}^a}{\sum_x P_x{}^a}$$

3. Abridged life tables.

Ans :

A abridged life table is similar to the complete life table under which most of the important ages are mentioned for a period of more than 1 year that could be 5 years or 10 years.

The abridged life table can be available in two different forms or types. Such as,

1. It could be available in the form where some integral values of x which are at some distance apart, usually 5 or 10 years.
2. It could be available in the form where the values of the age group x are stated for 5 or 10 years.

4. Assumptions of Life table.

Ans :

The assumptions which are used in the construction of life tables are

- The hypothetical group is closed for migration i.e., there is no change in the group expect the losses due to deaths.
- The individuals die at each age according to some predetermined schedule which is fixed and doesn't change.
- The deaths are distributed uniformly over the period of age 'x' to 'x+1'.

5. Crude Birth Rate (C.B.R)

Ans :

This is the simplest of all the measures of fertility and consists in relating the number of live births to the total population. This provides an index of the relative speed at which additions are being made through child birth. The fertility pattern of the above mentioned measure is given by crude birth rate (C.B.R) defined as follows.

$$\text{Crude Birth Rate} = \frac{B^t}{P^t} \times k \qquad \text{... (1)}$$

Where

B^t = Total number of live births in the given region or locality drawing a given period, say t.

P^t = Total population of the given region during the period t.

K = A constant, usually 1000.

Merits

It is simple, easy to calculate and readily comprehensible. It is based only on the number of births (B^t) and the total size of the population (P^t) and does not necessitate the knowledge of these figures for different sections of the community or the population.

Demerits

- The crude birth rate, though simple, is only a crude measure of fertility and is unreliable since it completely ignores the age and sex distribution of the population.
- C.B.R is the not a probability ratio, since the whole population P^t cannot be regarded as Exposed to the risk of producing children . In fact, only the females and only those between the child bearing age group (usually 15 to 49 years) are Exposed to risk and as such whole of the male population and the female population outside the child-bearing age should be Excluded from P^t, the demoninator in ....(1). Moreover, even among the females who are Exposed to risk, the risk varies from one age group to another, a woman under 30 is certainly under greater risk as compared to a woman over 40.
- As a consequence of variation of climatic conditions in various countries, the child bearing age groups are not identical in all the countries. In tropical countries, the period starts at an apparent earlier date than in countries with cold weather. Accordingly, crude birth rate does not enable us to compare the fertility situation in different countries.

- Crude birth rate assumes that women in all the ages have the same fertility, an assumption which is not true since younger women have, in general higher fertility than elderly women. CBR thus gives us an estimate of a hetrogeneous figure and is unsuitable for comparative studies.

6. Specific Death Rate

Ans :

To formulate ideas mathematically, Let $_nD_x$ = Number of deaths in the age -group (x, x + n) i.e., the number occurring during a given period, t.

$_nP_x$ = Total population of the age -group x to (x+n) then the age -specific death rate for the age -group x to x+n, usually denoted by $_nm_x$ is given by

$$_nm_x = \frac{_nD_x}{_nP_x} \times 100$$

Taking n = 1, we get the annual A.S.D.R given by $m_x = \frac{D_x}{P_x} \times 1000$.

To be more specific, the A.S.D.R for males is given by

$$^m_nm_x = \frac{^m_nD_x}{^m_nP_x} \times 1000 \quad \text{... (1)}$$

Where $^m_nP_x$ is the number of males in the population in the age group x to x + n and $^m_nD_x$ is the number of deaths amongst this population.

Similarly, the A.S.D.R for females is given by the formula.

$$f_nm_x = \frac{f_nD_x}{^f_nP_x} \times 1000 \quad \text{... (2)}$$

Equation (1) and (2) give the death rates specific to both age and sex.

Merits

The death rates specific to age and sex overcome the drawback of C.D.R, Since they are computed by taking into consideration the age and sex composition of the population. By eliminating the variation in the death rates due to age-sex distribution of the population S.D.R's provide more appropriate measures of the relative mortality situation in the regions.

- For general analytical purpose, the death rate specific for age and sex is one of the most important and widely applicable type of death rates.

Demerits

More over, in addition to age and sex distribution of the population social, occupational and topographical factors come into operation causing what is called differential mortality. S.D.R'S completely ignore these factors.

7. Define vital statistics.

Ans :

Introduction

Vital statistics is defined as that a branch of Biometry which deals with data and the laws of human mortality, morbidity and demography. The term vital statistics refers to the numerical data or the techniques used in the analysis of the data pertaining to vital events occurring in the given section of the population. By vital events we mean such events of human life as fertility and mortality (births and deaths), marriage, divorce, separation, adoptions, etc.

Definition

The term vital statistics refers to the Numerical data (or) the technique use in the numerical data (or) the technique use in the analysis of data. Vital statistics forms perhaps the most important branch of statistic as it deals with mankind in the aggregate.

8. Crude Death Rate

Ans :

This is the simplest of all the indices of mortality and is defined as the number of deaths per k persons in the population of any given region or community during a given period. Thus, in particular, the annual crude death rate (C.D.R) denoted by m for any given region (or) community is given by.

$$\text{CDR} = \frac{\text{Annual Deaths}}{\text{Annual Mean Population}} \times k$$

where k = 1,000

The crude death rate for any period gives the rate at which the populated is depleted through deaths over the course of the period.

Merits

- It is simple to understand and calculate.
- C.D.R is perhaps the most widely used of any vital statistics rates. As an index of mortality, it is used in numerous demographic and public health problems.
- Since the entire population of the region is Exposed to the risk of mortatity.

Demerits

- The important drawback of crude death rate is "it completely ignores the age and sex distribution of the population".

9. Gross Reproductive Rate

Ans :

The population growth mainly depends on the birth of female children who are the future mothers. In order to have a better idea about the rate of population growth, In addition to the age and sex composits of the population. We must take into account the sex of the new born children because it is ultimately the female births who are the potential future mother's and result in an increase in the population. One of such measure is the G.R.R. and it is defined as the sum of the annual female age specific fertility rate.

The composition of Gross Reproductive rate requires the classification of births and the sex of the newly born babies. In such cases an approximate value of G.R.R. may be obtained under the assumption that sex ratio of birth remain more (or) less constants at all the age of the human in the reproductive period and is given by

$$\text{G.R.R} = \frac{\text{Total no. of female births}}{\text{No. of births}} \times \text{T.F.R}$$

10. Define life table

Ans :

It is a conventional method of expressing the most fundamental and essential facts about the age distribution of mortality in a tabular form. It is a powerful tool for measuring the probability of life death of various age stress.

The life table gives the life history of hypothetical group as it is gradually decreased by deaths.

11. Sources of Vital Statistics.

Ans :

The data or information for vital statistics can be collected through different sources or methods. They are as following.

1. Registration Method
2. Census Enumeration Method
3. Survey Method
4. Sample Registration System (SRS)
5. Analytical Methods.

12. Age Specific Rate

Ans :

The age specific fertility rate (ASFR) measure is used to calculate the birth rate i.e., how many number of child bom alive during a year per thousand women of particular age group. It allows to calculate the fertility or birth separately for various age groups of females who are in child bearing ages. Moreover, while calculating this rate number of live births to women of a specific age group are considered. For example, age between 15-20, 20-25 etc.

Choose the Correct Answer

1. ________ assumes that there will be no change in the census except the losses due to death. [d]
   (a) Index table (b) Pivot table
   (c) Abridged table (d) Life table

2. Registration method is one of the major source of ________ [b]
   (a) Business statistics (b) Vital statistics
   (c) Basic statistics (d) None of the above

3. There are ________ types of measures of population growth. [c]
   (a) Two (b) Three
   (c) Four (d) Five

4. Vital ratios are also known as ________ [c]
   (a) Vital statistics (b) Vital methods
   (c) Vital rates (d) Vital events

5. ________ are conventionally numerical records of marriages, births, sickness etc. [b]
   (a) Business statistics (b) Vital Statistics
   (c) Sources of statistics (d) Basic statistics

6. ________ is very useful in measuring population of country. [d]
   (a) Registration method (b) Survey method
   (c) Analytical method (d) Census enumeration method

7. ________ are the rates of vital events. [a]
   (a) Vital rates (b) Ratios
   (c) Index numbers (d) None of the above

8. ________ IMR means . [d]
   (a) Inflation measurement rate (b) Instant measurement report
   (c) Instant mortality rate (d) Infant mortality rate

9. ________ Mortality means . [a]
   (a) Death (b) Birth
   (c) Life (d) Insolvency

10. Life table is used for preparing a ________ based on age and sex. [c]
    (a) Financial report (b) Economy report
    (c) Population report (d) Inventory report

Fill in the Blanks

1. ________ is defined as that a branch of Biometry which deals with data and the laws of human mortality.
2. CDR stands for ________.
3. ________ rates are inadequate to give as any idea about the rate of population growth since they ignore the sex of the newly born children and their mortality.
4. Crude Rate of Natural Increase ________.
5. Pearl's vital index ________.
6. The ________ gives the life history of a hypothetical group or cohort as it is gradually diminished by deaths.
7. The ________ group is closed for migration i.e., there is no change in the group expect the losses due to deaths.
8. The data for constructing a life table are the ________ and ________ data.
9. Life tables are as well constructed on regional basis and other factors accounting differential ________.
10. Vital word is derived from ________.

Answers

1. Vital statistics
2. Crude Death Rate
3. Fertility
4. C.B.R – C.D.R
5. $\dfrac{\text{No. of births in the given period t}}{\text{No. of deaths in the given period t}}$
6. Life table
7. Hypothetical
8. Census data, Death registration
9. Mortality
10. Latin word

UNIT IV	**Indian Official Statistics**: Functions and organization of CSO and NSSO. Agricultural Statistics, area and yield statistics. National Income and its computation, utility and difficulties in estimation of national income. **Index Numbers :** Concept, construction, uses and limitations of simple and weighted index numbers. Laspeyer's, Paasche's and Fisher's index numbers, criterion of a good index numbers, problems involved in the construction of index numbers. Fisher's index as an ideal index number. Fixed and chain base index numbers. Cost of living index numbers and wholesale price index numbers. Base shifting, splicing and deflation of index numbers.

4.1 Indian Official Statistics

Q1. Define Official Statistics.

Ans :

Official Statistics is often referred to national statistics which is in the form of quantitative data.

Definition

According to Kerrison and Macfarlane, official statistics is defined as 'collected, commissioned or published by the central government departments and agencies, local government and NHS'.

Governments involves official statistics in registers where it records the details of specified events as when they occur such as births, deaths, marriages, divorces, crimes, certain contagious diseases and later notifiable diseases such as cancer, AIDS and so on.

4.1.1 Functions and organization of CSO and NSSO

Q2. Explain Functions and organization of CSO.

Ans : **(Dec.-21, Oct.-20, June-19)**

Functions

Central government established central statistical organization, under cabinet secretariat with the objective of creating coordination of large variety of statistical information, collected at the centre and state level. It performs many more functions as listed below.

1. Coordination of statistical activities at the centre and the state.
2. Advisory work concerning the statistical matter, particularly standardization of concepts and definitions to maintain uniformity, throughout the country.
3. Collection of statistical data related to planning.
4. Training of statistical personnel.
5. Compilation of national income estimates.
6. To provide statistical data of the nation to the united nations statistical offices and other international institutions.
7. To plan and coordinate the conduct of the annual survey of industries and publish the results.
8. To attend to the work of international statistical institutes (conferences) held in India and abroad.
9. To display of charts and graphs pertaining to the national data which are of administrative interest.
10. Circulation of regular publications.

This organisation does many other things concerning statistical matter, needed from time to time. Each of the above mentioned functions is looked after by a separate division with its head quarters in Delhi, except the industrial statistics wing at kolkata.

The CSO examines various schemes running at the state level and makes specific suggestions for their improvement. Now CSO has a network all over the country. CSO coordinates statistics collected by different ministries. It also coordinates with national sample survey organisation.

The central statistical organisation, supplies statistical information about India to united Nations statistical office and international agencies. It provides data for publication in

(i) The U.N monthly bulletin of statistics.

(ii) U.N quarterly bulletin on commodity trade statistics

(iii) U.N Demographic year book

(iv) The economic council for Asia and for East (ECAFE), Quarterly bulletin and annual surveys.

Besides these, CSO supplies compiled data to many other international agencies for publication. It also supplies information regarding the methodology adopted in the collection of data, the coverage and the scope.

Q3. Explain briefly about National Sample Survey Organisation (NSSO).

Ans : **(July-21)**

Census studies have been very expensive and time consuming. In many situations, we cannot afford to wait for long, to get the overall picture of various aspects of socio-economic study. Hence the emphasis was laid down to sampling studies which led to the creation of national sample survey organisation. As given earlier, it was established in January, 1950 in the department of Economic affairs, Ministry of Finance, to conduct country wide multipurpose sample surveys, covering all aspects of national economy required by national income committee (NIC), planning commission and other ministries of Government of India. The directorate of national sample survey, used to work under the statistics department of the cabinet secretaries its main functions are.

1. Collection of socio-economic data relating to demographic conditions for the whole country on regular basis.
2. To provide statistical data for national income and planning.
3. To conduct annual surveys in the organised industrial sector.
4. Training of personnel and providing guidance to the states in the conduct of surveys.

The work of NSSO was being done under dual control for two decades. The field work was carried out by the directorate of NSS, whereas designing of surveys, processing of data and preparation of reports was entrusted to the Indian statistical institute. In March 1970, the Government of India accepted the recommendation that all aspects of the NSS be brought under an unified control, with a governing council to govern its activities. The council was given the requisite independence and autonomy in decision making. Now NSSO is under the department of statistics, ministry of planning government of India. The objective of the NSSO, as laid down in the background of the aims of government statistical work, as an autonomous organisation, are

1. To provide statistical and other information needed for the efficient conduct of government business.
2. To evolve statistical technique to bear on the analysis of information, the solution of administrative problems and the estimation of future trends.
3. To collect and publish information which will be of use to those engaged in economic activities in the country.
4. To provide and analyse information which are useful to the research workers.
5. The assist in keeping the public informed of the new developments in the economic and the social fields.

In short, NSSO has unified control of the governing council with regard to survey designs, field operations, data processing, economic analysis and publication of NSS data. The present structure of NSSO consists of four functional divisions, with a chief executive officer. The four divisions are.

1. Survey design and research
2. Field operations
3. Data processing
4. Economic analysis

The executive officer, as the member secretary of the Governing council, obtains the approval of the council to the programme of work and directs the appropriate division to implement it. The national sample survey organisation works in close coordination with CSO in respect of certain projects and programmes. The tabulation and analysis of data are attended by the CSO and party, by the labour Bureau and National Building organisation (NBO). As far as the agriculture is concerned, the field operations division (FOD) is responsible for all activities with the increasing intake of new personnel, NSSO felt the need of reorientation of its staff.

Q4. Discuss in detail organization of NSSO.

Ans :

NSSO has conducted many rounds of surveys on various socioeconomic activities. They are shown in the below table,

Round	Year	Coverage Area (Area of Survey)	Survey Conducted
37th Round	Jan.-1982 - Dec.-1982	Whole of India, except Ladakh and Kargil districts of Jammu and Kashmir and the rural areas of Nagaland.	Survey on land holdings and live stocks, Debt and Investments.
38th Round	Jan.-1983 –Dec.-1983	Whole of India, except Ladakh and Kargil districts of Jammu and Kashmir and the rural areas of Nagaland.	Survey on Holdings consumer expenditure, employment and unemployment.
39th Round	Jan.-1984 -June-1984	Whole of India, except Ladakh and Kargil districts of Jammu and Kashmir and the rural areas of Nagaland.	Survey on population, the Births and the Deaths for the preceding two years.
46th Round	July-1990 - June-1991	Whole of India, except Ladakh and Kargil, Interior Villages of Nagaland and rural areas of Andaman and	Survey on wholesale and retail trade transacted by Non-Directory Trading Establishments (NDTE's).
47th Round	July-1991 - Dec.-1991	Whole of India, except Ladakh and Kargil, Interior Villages of Nagaland and rural areas of Andaman and Nicobar.	(i) Survey on disability, literacy and culture. (ii) Survey on consumer expenditure by taking two households per village or per urban block.
48th Round	Jan.-1992 - Dec.-1992	Whole of India, except Ladakh and Kargil, Interior Villages of Nagaland and rural areas of Andaman and Nicobar.	Survey on Landholding, Livestock holding, debt and investment.
49th Round	Jan.-1992 - July.-1992	Whole of India, except Ladakh and Kargil, Interior Villages of Nagaland and rural areas of Andaman and Nicobar.	Survey on Housing conditions, construction and migration.
61th Round	June-2004 July-2005	Whole of India, except Ladakh and Kargil, Interior Villages of Nagaland and rural areas of Andaman and Nicobar.	Survey on employment and unemployment.
65th Round	1st July-2008 30th June -2009	Whole of India, except Ladakh and Kargil, Interior Villages of Nagaland and rural areas of Andaman and Nicobar.	Survey on Domestic Tourism, Housing condition and Urban slums.

4.2 Agricultural Statistics

Q5. Explain the concept of Agricultural Statistics.

Ans : **(OCT.-20)**

India is an agricultural country and its economy largely depends on agriculture. In ancient days, agriculture had been the main source of income of all the rulers. They collected money in the form of land revenue. Therefore, most of the rulers maintained records of agricultural land and their categories on the basis of fertility and production. Records were also maintained for orchards. The main objective of keeping there records was the collection of revenue.

After independence, a lot of emphasis has been given to agriculture. Indian economy is based mostly on agriculture even today, hence a comprehensive and reliable statistics are necessary for good planning and development of agriculture and that of the country. As a matter of fact, all statistics which have an impact on agricultural economy may be regarded as agricultural statistics. The statistics pertaining to land utilization, production of crops, live stock and agricultural prices poultry, foresty, etc. will come under the category of agricultural statistics.

Agricultural statistics is mostly collected by the directorate of economics and statistics (DES) at the centre, as well as the state level. Each state has a Directorate of Economics and statistics. The statistics regarding area, crop production, poultry and forests are collected. Total area statistics are obtained from two sources,

1. The surveyor general of India
2. The village records maintained by the revenue department.

The statistics regarding area, forests are collected. The land is further divided into non-agricultural lands, non-cultivatable lands, grazing lands, irrigated lands and unirrigated lands. The area under crops is obtained by two sources.

1. Official series based on village records
2. NSS series based on sample surveys.

NSS collects data during the regular rounds of survey on area under different crops. The area figures provided by these two sources are far apart, because of the difference in method of the coverage of crops, difference in the field work, the classification of area under feed and the fodder crops, the allocation of area under mixed crops and due to sampling error in NSS estimates. The land utilisation statistics are published in India Agricultural statistics. The individual states publish these statistics in season and crop reports. Land utilisation statistics are also published in

1. Agricultural situation in India (monthly)
2. Abstract of agricultural statistics (DES Annual)
3. Statistical abstract of the Indian union (CSO Annual).

4.2.1 Area and Yield Statistics

Q6. How are the agricultural statistics of area and yield collected in India.

Ans : **(Imp.)**

Agricultural production includes the production of food and non-food crops but forests, livestock and fisheries are excluded at this place. We will give a brief account of these separately. Crop output statistics can not be collected at the time of harvesting, as millions of farmers harvest their crops at the same time, mostly they are uneducated and maintain no record. Hence the output is estimated by multiplying the area under a crop by the average expected yield per hectare, in the season. The yield statistics are collected by official machinery and NSSO. As official machinery, the work of the survey is undertaken by the directorate of economics and statistics. Two methods are adopted to collect yield statistics of various crops.

1. Traditional method
2. Random sampling method

1. Traditional Method

This method of estimation is known as annawari system. In this method, the condition of the crop is judged by patwari or any other official in relation to the normal crop. This is known as the condition factor or the annawari estimate. The yield estimate is obtained by the annawari system, by the formula,

$$\text{Crop yield} = \frac{\text{Annas judged}}{\text{Total annas}} \times \text{Normal yield} \times \text{Area}$$

2. Random Sampling Method

This method was recommended by the board of agriculture in early 1919, but was abandoned due to financial burden. Again it was introduced in 1942 by the Indian council of agriculture, this is used for cotton crops. In this method few villages are selected at randomly and in that village few crops are selected randomly. This experiments are conducted by the directorate of economics and statistics and statistical section at the board of revenue. The yield data collected by two different agencies. They differ in this estimates rounds, due to the size of the cuts and shape. The results are published mainly in

1. The agricultural statistics
2. The weekly bulletin of agriculture prices
3. The agricultural situation in India.

4.3 National Income and its Computation

Q7. Define National Income. Explain the uses of National Income.

Ans : **(Dec.-21)**

Meaning

National income estimates are of great importance to get a broad view of the entire economy of a country. They also provide the information about the changes occurring in economy from year to year. The economic policies and planning are mostly framed keeping in view, the national income of a nation. The National Income Committee (NIC) set up by the government of India in 1949, produced for the first time national income estimates for the entire Indian union. The estimates along with the methodology were published in the first and final report of the national income committee (ministry of finance) in 1951 and 1954. Efforts were made to improve the empirical data and overcome the gaps. The first result of these efforts were presented by the CSO in the National income statistics.

National income (or) national dividend is the sum of total goods and services produced in a country during the given period of time, generally in a year. It is the total output produced by the four factors for production in a country in a given year.

Use of National Income

1. The first and foremost use of national income estimate is that, it gives us a correct picture of the structure of the economy of the nation, as well as the distribution of income according to regions, industrial origin and income from functional services and persons during a specified period of time.
2. National income statistics provide a useful guideline in the formulation of the budget of a country.
3. National income figures have been found useful for studies of the problems of the economically underdeveloped countries.
4. National income estimates give us an Idea of the purchasing power of the people in the country inflationary and deflationary gaps of purchasing power are revealed by income and product figures. This is of great value in working out the details and timing of anti-inflationary and deflationary programmes.
5. National income statistics help in making the interregional and interstate comparisons of national income distribution. It is also useful in comparing the economic conditions of a country over a period of time.
6. National income estimates provide a basis for the future planning of a country. In the absence of such statistics, no comprehensive economic plan can be chalked out. It also helps in evaluating the success of an economic plan.
7. Estimate of national measure the size of the economy and level of performance of the economy.
8. They are useful to know the contributions of the different sectors of the economy to national income and variations in them.
9. They help the government in formulating suitable policies and development plans to increase growth rates.
10. They are also helpful in making international comparisons of peoples standards of living.

11. They help business firms in forecasting future demand for their products.
12. They are also helpful in making projections about future development trend of the economy.
13. They are also helpful in fixing future targets for the different sectors of economy based on their previous performance.

Q8. Explain various methods of national income.

(OR)

Describe briefly the methods of calculating national income.

Ans : **(June-19)**

The methods used to estimate the national income of a country necessarily depend upon the availability of statistics commonly the following three methods are used for the calculation of national income.

1. Output or production method
2. Income method
3. Expenditure method

1. Output or production method

Output (or) production method is also called inventory method. It consists of finding out the market value of all goods and services produced by the individuals and business enterprises during a specified period of time. The national income or net domestic product can be calculated by the following equation.

National income = (Value of goods and services + self consumption + increase in stock – depreciation of capital – income from abroad).

2. Income Method

The income method consists of adding together all incomes by the way of wages, interests rents and profits wages include the value of services (mental and physical), in terms of money rendered by the nationals. Interest is the amount of money, received by the entrepreneurs, by way of capital investment. Similarly, rent may be defined as the factor income generated by the letting and the use of land for agricultural and other purposes, buildings (residential and non residential), machinery, equipment and other fixed assets. It is treated as rental income from property. Also any amount of money saved by the individuals or institutions, after deducting the expenditure on production, may be termed as profit.

3. Expenditure Method

The national income is equal to the sum of consumption and saving. Symbolically we can write.

$$y = C + S$$

Where y = National income

C = Consumption

S = Saving

The method is rarely used because of the non-availability of data required for its calculation. Because of the predominance of unorganized sectors in underdeveloped and developing countries, there is a lack of proper maintenance of records of consumption, expenditure savings etc.

4.3.1 Utility and difficulties in estimation of national income

Q9. Explain the utility and difficulties in estimation of national income.

Ans : **(Imp)**

Utilities

National income estimates are of great importance partly as a measuring rod of economic welfare and growth and partly as a food of analysis changes and of comparison of economic performance of a country vis-a-vis other countries of the world.

There are six main utilities of national income estimates.

1. Indicator of Economic Welfare

Until recently, GNP per capital has served as an indicator of economic well being of people in a country. The higher the GNP per capita, the higher is the level of the economic welfare of people in a country.

2. Measure of Economic Growth and Development

Economic growth is measured as a percent increase in GNP over a period of one accounting year. For instance, if GNP increases from Rs. 20,00,000 crores to Rs. 21,00,000 crores over a period of one year, the growth rate over the year for the country would work out at 5%.

Economic growth leads to economic development A higher growth would lead to a higher level of economic development, other things remaining the same.

3. **Study and analysis of Structural changes in different sectors of an Economy**

National income estimates essentially reveal changes in incomes and outputs of different sectors performance of various sectors of the economy most importantly, it helps locating the surplus and the deficit sectors which is crucial for formulation of monetary, fiscal and foreign trade policies.

4. **Comparison of Economic performance of a country with that of other countries**

National income estimates facilitate comparison of economic performance of a country vis-a-vis other countries of the world. For instance, comparing rate of growth of national income of our country with that of the other countries tells us how our policies of investment and resource allocation have fared in comparison to those of the rest of the world. This helps in making decisions whether a policy change is called far or not.

5. **Significance to business policy formulation**

National income statistics provide data about the past performance of different industries and the likely trends of future demand for the products of these industries. This helps the firms in every industry to chalk out their future plans of production.

6. **Significance to trade unions**

Trade unions and labour organizations may link their wages and salaries to the contribution of the labour force in the GDP. This helps investigation whether the current salaries and wages are in line with the contribution of the labour force in the GDP.

Difficulties in Estimation of National income

Here six major difficulties faced by a country during computation of national income.

1. **Types of Goods and Services**

The kinds of goods and services which should be included in national income pose a problem.

Goods and services having money value are included in the national income but there are goods and services which may have no corresponding flow or money payments. Services which are performed for love, kindness and mercy and not for money have an economic value but have no money value.

The difficulty is whether these services should be included in national income and how to measure their money value. e.g a paid maid servant's services are included in the national income but later when she marries the master, she is not paid any more, though she continues to perform the services. There is, thus, a reduction in the national income.

Similarly, when a house wife cooks for the family, her activity is not included in the GNP. When she cooks in a restaurant and gets paid, her services are included in GNP. Much of the activity, goods and services which have money value and are considered economic in U.K and U.S.A on the basis of their marketability, are treated as non-economic in India because they are carried on in the household sector.

2. **Problems of Double counting**

Another difficulty is of double counting usually associated with the inventory method. Double counting implies the possibility of a commodity like raw material or labour being included in national income more than once, e.g., a farmer sells maize worth rupees two hundred to a mill - owner, the mill owner further sells the maize flour to a whole sale dealer, who further sells it to consumer, if we calculate it at every stage, its money value will increase to eight hundred rupees but actually the increase in national income has been to the extent of two hundred rupees only.

3. Excluded Market Transactions

Certain transactions that take place in the market are excluded from the computation of national income because they violate the general rule for the recognition of income. The good or service must be currently available scarce resources. Many transactions are such that represent merely the transfer of wealth or the exchange of commodities produced in some previous accounting period. These excluded transactions relate to

(i) Transfer payment

(ii) Capital gains

(iii) Illegal activities

(iv) second - hand sales

4. Problem of Inputted Values

There are certain goods and services which do not appear in or cannot be brought to the market. In such cases we have to input values to them. It means to give or to fix their values, in case they had been brought to the market. The procedure although very logical yet is beset with number of practical difficulties because the task of imputing or fixing values is not easy. But values, onces they are imputed or fixed are included in the national income accounts.

The true monthly income of the managing director will be the imputed money value of fringe benefits enjoyed by him plus the monthly salary. Thus, when some goods and services, representing current economic activity in the economy do not appear on the market, an imputed value equal to the market value of similar goods and services is assigned to them for purposes of including the value of these goods and services in the national product and income accounts.

5. Inventory Adjustments

Inventory adjustments i.e., changes in the stock of capital goods or final products are also to be taken into account while computing national income. If a jute mill adds to its inventory of jute products during the year, it represents an increase in output and must, therefore, be included in GNP. The difference between the difficially published figures and the figures obtained from the business accounting data calls for inventory valuation adjustment on account of the change in the physical volume of inventories and the change in the prices at which there inventories are valued by business units, inventory valuation adjustment becomes essential.

6. Depreciation

Depreciation implies a reduction in the value of capital stock or capital goods due to wear and tear, constant use etc. During the process of production the wear and tear or capital consumption occurs, resulting in, at the same time, a decline in the relative efficiency of the plant and equipment on account of obsolescence. However, the problem of correctly estimating depreciation is equally a difficult task e.g., a machine may be used more intensively in one year than the other. But the rate of depreciation remains the same, though it should differ between two years. Again the depreciation of similar equipment may differ between two business units.

4.4 Index Numbers

4.4.1 Concept

Q10. Define Index Numbers.

Ans :

Meaning

An Index, simply stated, is an indicator. It indicates the broad change in a given phenomenon. It is relative and indicates changes in the level of the given phenomenon in course of time, across geographical locations or in respect of any other characteristic. For example, the BSE Sensex (stands for Bombay Stock Exchange Sensitive Index) is an index of movement in stock prices. It tells us, very broadly, whether the prices of stocks of various companies have gone up or moved down on the Bombay Stock Exchange. Similarly, the Wholesale Price Index (WPI) gives us an indication of whether the general price level in the economy is going up or falling down. Similarly, Index numbers (or indices) can be constructed to measure any phenomenon, such as Industrial production, Number of road accidents, cost of Living or any other activity. Thus, Index Numbers are barometers measuring change in the level of a phenomenon.

Definitions

(i) **According to Croton and Cowden** Different Experts have defined Index Numbers in different words. Some of the definitions are stated as below.

(ii) **According to Speiges** "Index numbers are devices for measuring differences in the magnitude of a group of related variables."

(iii) **According to Weldon** "An Index Number is a statistical measure designed to show changes in a variable or a group of related variables with respect to time, geographical location or any the characteristic."

(iv) **According to Edgeworth** "An Index Number is a statistical device for indicating the relative movements of data where measurement of actual movement is difficult or incapable of being made."

(v) **According to Bowley** "Index Number shows by its variations the changes in magnitude which is not susceptible either of accurate measurement in itself or of direct valuation in practice."

"Index Numbers are used to measure the change in some quantity which cannot be observed directly, which we know to have a definite influence on many other quantities which we can so observe, tending to increase all or diminish all, while this influence is concealed by the action of many causes affecting the separate quantities in various ways."

On the basis of the above definitions, the following points can be made:

1. Index Numbers are a measurement device
2. They do not measure or state the actual level attained by the phenomenon being studied. They measure the change in the phenomenon being studied.
3. The situations for which Index Numbers are used for comparison are not restricted in any manner. It can be a comparison of two time periods, two geographical locations, two groups of people or any other phenomenon.
4. Index Numbers are the result of a numerical calculation. They do not have any units such as kgs or rupees.
5. Index Numbers are relative terms and hence, they are normally expressed in percentage terms.

Q11. What are the uses of index numbers ?

(OR)

Explain the importance of index numbers.

Ans : **(Imp.)**

1. Economic Barometers

Index Numbers can be constructed for any phenomenon for which quantitative information is available. They capture the various changes taking place in the general economy and business activities. They provide a fair view of the general trade, the economic development and business activity of the country. Thus, they are aptly termed as 'Economic barometers'.

2. Study of Trend

Index Numbers study the relative changes in the level of a phenomenon over different periods of time. They are especially useful for study of general trend of the general trend for a group phenomenon in a time series data.

3. Policy Formulation

Index Numbers are indispensable for any organization in efficient planning and formulation of executive decisions. For example, the Dearness Allowance payable to employees is determined on the basis of Cost of living index numbers. Similarly, Psychiatrists use Intelligence Quotients to assess a child's intelligence in relation to his/her age, which further can be used in framing the education policy.

4. Deflation

Index Numbers can be deflated to find out the real picture pertaining to the phenomenon. For example, we can know if the real income of employees has been increasing by deflating the nominal wages with the help of index numbers.

5. Forecasting

Index Numbers provide valuable information that aids forecasting. For example an Index of sales, along with related indices such as cost of living index, helps in forecasting future demand and future sales of business.

6. Measurement of Purchasing Power

The cost of Living Index helps us in finding out the intrinsic worth of money. It is one of the key indicators touching the life of the common man.

7. Simplicity

Index Numbers eliminate the clutter of large numbers and complicated calculations, providing the underlying information in a manner that is simple and easy to underlying information in a manner that is simple and easy to understand. For example, many people may not understand the dynamics of stock markets, but they can follow the movements of the BSE Sensex with ease.

Q12. What are the characteristics of Index Numbers ?

Ans :

The following are the main characteristics of index numbers :

i) Index Numbers are Specialized Averages.

ii) Index Numbers as Percentages and Measure of Relative Change.

iii) Basis of Comparison.

iv) Universal Applications.

(i) Index Numbers are Specialized Averages: According to L.R. Connor, "In its simplest form an index number represents a special case of an average, generally a weighted average compiled from a sample of items judged to be representative of the whole."

In simple average the data treated have same units of measurement but index number average variables having different units of measurement. We may find an index number for a group of items including food, clothing, rent, cooking gas etc. which are all measured in different units.

(ii) Index Numbers as Percentages and measure of Relative Change: Index numbers are expressed in percentages but the word or symbol for percentage is never used. They measure relative changes in price over years with reference to some period called Base Period or Base Year. If the index number in 2000 is 150 with reference to 100 in 1995. It will mean that there is 50% increase in price in 2000 as compared to 1995. Since the index numbers are quantitative expressions of relative changes, they are expressed in numbers.

(iii) Basis of Comparison : Index numbers are used to make comparisons over different time periods with reference to base period.

(iv) Universal Applications : They are applied universally to ascertain different types of changes in different sectors of the economy. These changes may be with regard to prices, standard of living of the people, industrial or agricultural production, national or per capita income, exports or imports etc.

Q13. Explain the various types of index numbers.

Ans :

An Index is an indicator. Index Numbers can be constructed for study of any phenomenon, if such phenomenon is capable of being quantified. In relation to data pertaining to business and economy, index numbers may be classified into the following.

1. Price Index Numbers

Price Index numbers measure the changes in prices. They are the most common indices. There are various methods by which the indices can be calculated. Most of the discussion in this chapter shall relate to price index numbers.

2. Quantity Index Numbers

Quantity Index Numbers study the changes in the volume of goods produced, consumed or transacted. The Index of Industrial Production is an example of a quantity index. However Quantity Index numbers are not very popular.

3. Value Index Numbers

Value Index numbers are rarely used. They study the change in total value, rather than mere change in prices of changes in quantities. They may have to be supplemented with price and quantity indices.

4.4.2 Construction

Q14. What are the various methods of Constructing Index Numbers?

Ans : **(Imp.)**

The various methods of constructing index numbers are shown in the following figure :

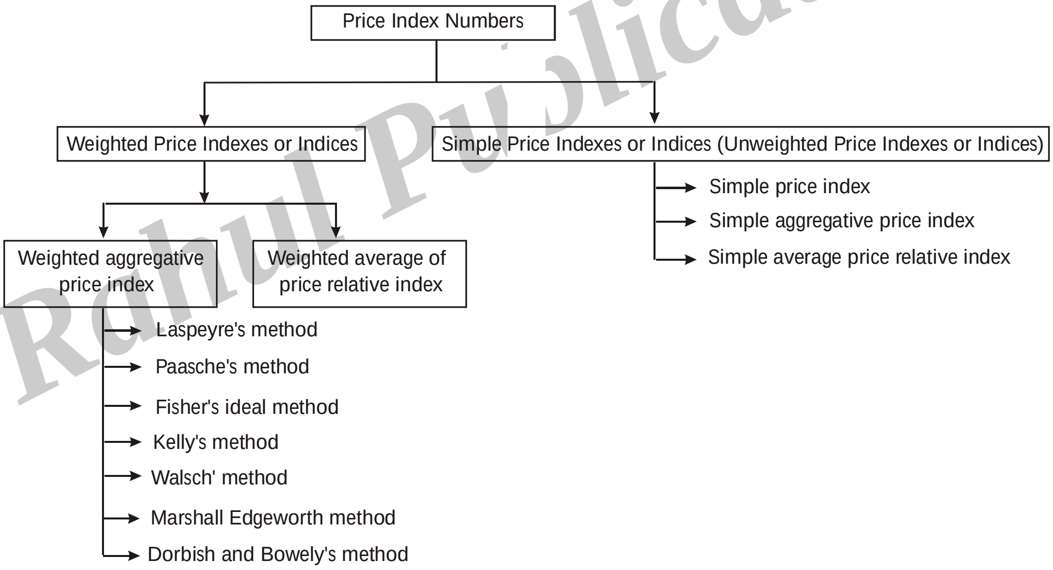


Fig: Types and Methods of Price Index Numbers

1. Weighted Price Indexes

At the time of constructing the weighted price indexes or indices, the rational weights are allocated in an explicit manner. These rational weights show the relative significance of items or commodities which are related with the computation of an index. Quantity weights and value weights are used in this weighted indexes or indices. Weighted price indexes or indices are further divided into two types as follows,

(a) Weighted aggregate price index

(b) Weighted average of price relative index.

(a) Weighted Aggregate Price Index

In a weighted aggregate price index, certain weight is assigned to each and every commodity or item of group in accordance with its significance. This helps in gathering more information and improving accuracy of the estimates.

The following methods are used in weighted aggregate price index:

(i) Laspeyre's method

(ii) Paasche's method

(iii) Fisher's ideal method

(iv) Kelly's method

(v) Walsch's method

(vi) Marshal Edgeworth's method

(vii) Dorbish and Bowley's method.

(b) Weighted Average of Relatives Method

The basis methodology of calculating the price relatives is same as in case of simple average of relatives method. However, instead of calculating simple average weights are assigned to price relatives and a weighted average is calculated. Such weighted average can be arithmetic mean or geometric mean of the weighted price relatives. The following steps have to be followed.

Case (i) If weighted Arithmetic Mean is used.

Step 1

Calculate Price Relatives P for each item, $p = (P_1 / P_0) \times 100$.

Step 2

Calculate weights with which price relatives are to be multiplied. Normally, the value of the item in the base year (i.e. P_0q_0) is taken as weight. How ever, weights can be P_0q, p_1q_0 or p_1q_1. Weights are denoted by V.

Step 3

Calculate weighted price relatives by multiplying the price relatives with their correspon-ding weights. In other words, Calculate PV.

Step 4

Add weighted price relatives obtained in step 3. Denote as $\sum PV$.

Step 5

Calculate the sum of weights. This is denoted by $\sum V$.

Step 6

Weighted Average of Price Relatives = $p_{0.1} = \frac{\Sigma PV}{\Sigma V}$.

Case (ii) If Geometric mean is to be used.

Step 1

Calculate Price relatives P for each item. $P = (P_1/p_0) \times 100$

Step 2

Calculate the logarithm value of P. This is denoted by log P

Step 3

Calculate weights for each them. This is denoted by V.

Step 4

Calculate V log P for each item by multiplying weight obtained in step 3 with logarithm value of price relatives obtained in step 2.

Step 5

Add the weighted logarithm values of Price relatives. Denote it as $\sum(V.\log P)$.

Step 6

Weighted Average to price relatives

$$= P_{0.1} = \text{Antilog}\left[\frac{\sum(V.\log P)}{\sum V}\right]$$

2. Unweighted Price Index

(a) Simple Aggregative Method: This method involves aggregation of prices in the current period and expressing the total as a percentage of aggregate of prices in the base period. The following steps are followed.

Step 1

Calculate $\sum P_1, \sum P_1$ is the sum total of prices of all items in the current year.

Step 2

Calculate $\sum P_0, \sum P_0$ is the sum total of prices of all items in the base year.

Step 3

Calculate $P_{0.1}$, $P_{0.1}$, is the price index number of the current year with respect to the base year. It is expressed in percentage terms and calculated as under:

$$P_{0.1} = \left(\frac{\sum P_1}{\sum P_0}\right) \times 100$$

(b) Simple Average of Price Relative Method

Under this method, the price of each item in the current year is expressed as a percentage of its price in the base year. The figure so obtained is called Price Relative. The price relatives are then averaged to calculate the index number for that year. Either Arithmetic mean or Geometric mean can be used for calculation of average of price relatives. The following steps are following.

(a) If Arithmetic mean is used for the purpose of averaging then

Step 1

Calculate Price relative for each item. Price Relative of an item. Is obtained by the formula $\left(\frac{p_1}{p_0}\right)$ × 100 Where P1 is price of the item in the current year and P0 is the price of the item in the base year.

Step 2

Calculate average of price relatives to obtain the Index Number P0.1

$$P_{0.1} = \sum \frac{\left(\frac{p_1}{p_0} \times 100\right)}{N}$$

Where N = Number of Items

(b) If Geometric Mean is used for averaging the price relatives, then

Step 1

Calculate Price Relative (p) of each item (p = P1/ p0) × 100

Step 2

Calculate Logarithm Value of each Price relative (log p)

Step 3

Calculate simple average of logarithm values obtained in step 2 $\left(\frac{\text{Logp}}{N}\right)$

Step 4

Calculate Antilog of value obtained in step 3

Thus

$$\text{Index Number } P_{0.1}, = \text{Antilog}\left[\frac{\sum \log p}{N}\right]$$

$$\text{Where } p = \left(\frac{p_1}{p_0} \times 100\right)$$

4.4.3 Uses and Limitations of Simple and Weighted Index Numbers

Q15. Explain the Uses and Limitations of Index Numbers.

Ans : **(Imp.)**

Uses

The main uses of index numbers can be summarised as follows :

1. Index Numbers as Economic Barometers

"Index numbers are today one of the most widely used statistical devices... They are used to take the pulse of the economy and they have come to be used as indicators of inflationary or deflationary tendencies". They are indispensable tools for the management personnel and in bussiness planning and formulation of executive decisions. The indices of prices (wholesale and retail), output (volume of trade, import and export, industrial and agricultural production) and bank deposits, foreign exchange and reserves to., throw light on the nature of variation in the general economic and business activity of the country and gives us a, fairly good appraisal of the general trade, economic development and business activity of the country.

2. Index Numbers help in Studying Trends and Tendencies

Since the index numbers study the relative changes in the level of a pheno-menon at different periods of time, they are specially useful for the study of the general trend for a group phenomenon in a time series data. The indices of output (industrial and agricultural production), volume of trade import and export etc., are extremely useful for studying the changes in the level of phenomenon due to the various components of a time series viz., secular trend, seasonal and cyclical variations and irregular components and reflect upon the general trend of production and business activity. As a measure of average change in extensive group, the index numbers can be used to forecast future events.

3. Index Numbers help in Formulating Decisions and Policies

Index numbers of the data relating to prices, production, profits, imports and exports, personnel and financial matters are indispensable for any organisation in efficient planning and formulation of executive decisions. For example, the cost of living index numbers are used by the government and the industrial and business concerns for the regulation of dearness allowance (D.A.) or grant of bonus to the workers so as to enable them to meet the increased cost of living from time to time. Although index numbers are now widely used to study the general economic and business conditítions of the society, they are also applied with advantage by sociologists (population indices), psychologists (I. Q's), health and educational authorities etc., for formulating and revising their policies from time to time.

Limitations of Index Numbers

Although index numbers are indispensable tools in economics, business management etc, they have their limitations and proper care should be taken in using and interpreting then.. Some of their limitations are enumerated below :

1. Since index numbers are computed from sample data, all the errors inherent in any sampling procedure creap in its construction. Hence the index numbers reflect only approximate changes in the relative level of a phenomenon.
2. At each stage of the construction of the index numbers, starting from selection of commodities to the choice of formula there is likelihood of the error being introduced. An attempt should be made to minimise these errors, as far as, possible.
3. Due to rapid advancements in science and technology these days, there is a rapid change in the tastes, customs and fashions and consequently in the pattern of consumption of various commodities among the people in a society. Hence index numbers, may not be able to keep pace with the changes in the nature and quality of the commodities consumed at the two periods being considered and hence may not be truly representative.
4. None of the formula for the construction of index numbers is exact and contains the so called 'Formula error'. For example, Laspeyer's dndex has an upward bias while Paasche's index has a downward bias.

4.4.4 Laspeyer's, Paasche's and Fisher's Index Numbers

Q16. What is Laspeyre's index method ?

Ans :

This method takes the quantities of the commodities in the base period as the weight of that commodity for the purpose of calculating the index numbers. The following steps may be following.

Step 1

Multiply the current year price (represented by p_0) with the quantities of the base year (q_0) for each commodity.

Step 2

Add the numbers obtained in step 1. The resultant sum is represented as $\sum p_1q_0$

Step 3

Multiply the prices of base year (represented by p_0) with the quantities of the base year for each commodity.

Step 4

Add the numbers obtained in step 3. The resultant sum is represented as $\sum p_1q_0$

Step 5

The index number as per Laspeyre's method

$$P_{0.1} = \frac{\sum P_1q_0}{\sum P_0q_0} \times 100$$

Q17. What is Paasche's index method ?

Ans :

Paasche's Method is similar to Laspeyre's method. The only difference is in assignment of weights. As per this method quantities consumed of the commodities in the current year is taken as basis. The following steps need to be followed.

Step 1

Multiply current year's prices (p_1) with current year's quantities (q_1)

Step 2

Add the numbers obtained in step (1). The resultant sum is $\sum p_1q_1$

Step 3

Multiply base year's Prices (p0) with current year's quantities (q_1)

Step 4

Add the numbers obtained in step (2). The resultant sum is $\sum p_0q_1$

Step 5

Index Number as per Paasche's method = $P_{0.1} = \frac{\sum p_1q_1}{\sum p_0q_1} \times 100.$

Q18. Explain the comparison of Laspeyre's and Paasche's method ?

Ans :

S.No.	Laspeyre's Index Number	S.No.	Paasche's Index Number
1.	Here, quantity of the base year is assumed to be the quantity of the current year.	1.	Here, quantity of the current year is assumed to be the quantity of the base year.
2.	It has an upward bias i.e. the numerator of the index number is increased due to the assignment of higher weights fixed on the basis of the base year's quantities even though there might have been a fall in the quantity consumed during the current year due to rise, or fall in price and change in in tastes, habits and customs etc. in the current year.	2.	It has a downward bias i.e. the numerator of the index number is decreased due to the assignment of lower weights fixed on the basis of the current year's quantities even though the quantities in the current year might have fallen due to rise or fall in price, or change in habits of consumption.
3.	As the quantity of the base year are used as weights, the influence of price changes on quantities demanded do not get reflected in the index number.	3.	As the quantities of the current year are used as weights, the influence of price changes on quantities demanded get reflected in the index number.
4.	It measures changes in a fixed marked basket of goods and services as the same quantities are used in each period.	4.	It continually updates the quantities to the level of current consumption.
5.	Here, weights remain constant.	5.	Here, weights are determined every time an index number is constructed.

Q19. What is Fisher's Ideal Index ?

Ans : **(July-22, Imp.)**

Fisher's Ideal Index

This is the most popular amongst all weighted aggregative index numbers. It is obtained by calculating the Geometric Mean (G.M) of Laspeyre's and Paasche's index numbers. The formula for calculating fisher's ideal index is an under.

$$P_{0.1} = \left[\sqrt{\frac{\sum p_1q_0 \times \sum p_1q_1}{\sum p_0q_0 \times \sum p_0q_1}}\right] \times 100$$

Reasons for Fisher's Index being called an Ideal Index

Reasons

(i) It gives weightage to both current consumption and base year consumption.

(ii) It is free from upward or downward bias.

(iii) It satisfies both time reversal and factor reversal tests.

(iv) It is a Geometric mean of Laspeyre's index and Paasche's index

Problems

1. From the following data construct an Index number for 1994 taking 1999 as base as per simple Aggregative method.

Commodities	Price in 1993 (Rs)	Price in 1994 (Rs)
A	40	60
B	60	90
C	85	125
D	25	35
E	30	40

Sol :

Construction of Price Index

Commodities	Price in 1993 in Rs. (P_o)	Price in 1994 in Rs. (P_1)
A	40	60
B	60	90
C	85	125
D	25	35
E	30	40
	$\Sigma P_o = 240$	$\Sigma P_1 = 350$

$$P_{01} = \left(\frac{\Sigma p1}{\Sigma p0}\right) \times 100 \ \frac{350}{240} \times 100 = 145.83$$

2. From the following data construct an Index number for 1994 taking 1999 as base as per simple Aggregative method.

Year	Commodity					
	A	B	C	D	E	F
1970	45	60	20	50	85	120
1975	55	70	30	75	90	130

Sol :

(a) Arithmetic Mean

Commodity	P_0 (Price in 1970)	P_1 (Price in 1975)	Price Relative P	Log P
A	45	55	55/45× 100 = 122.22	2.0871
B	60	70	116.67	2.0668
C	20	30	150.00	2.1761
D	50	75	150.00	2.1761
E	85	90	105.88	2.0245
F	120	130	108.33	2.0346
			ΣP=753.10	**12.565.2 = Σ Logp**

Step 2 : Average of Price Relatives = $\frac{753.10}{6} = 125.52$

(b) Geometric Mean

$$P_{0.1} = \text{Antilog}\left[\frac{\Sigma \text{Logp}}{N}\right] = \text{Antilog}\left(\frac{12.5652}{6}\right) = \text{Antilog (2.0942)} = 124.3$$

3. From the following data calculate a price Index based on price Relatives Method using Arithmetic Mean.

Commodity	A	B	C	D	E	F
Price 2015(Rs.)	45	60	20	50	85	120
Price 2016 (Rs.)	55	70	30	75	90	130

Sol :

Index Number using Arithmetic mean of price relations.

Commodity	Price in 2015 (P_o) (Rs)	Price in 2016 (P_1) (Rs)	Price Relatives $\frac{P_1}{P_o} \times 100$
A	45	55	$\frac{55}{45} \times 100 = 122.22$
B	60	70	116.666
C	20	30	150
D	50	75	150
E	85	90	105.88
F	120	130	108.33
N = 6		Total	$\Sigma \frac{P_1}{P_2} \times 100 = 753.09$

Arithmetic Mean $P_{01} = \frac{\Sigma \frac{P_1}{P_0} \times 100}{N} = \frac{753.09}{6}$

$= 125.515$

$\boxed{A.M = 125.515}$

4. From the following data construct on Index Number for 2017 taking 2016 as base as per simple aggregative Method.

Commodity	P	Q	R	S	T
Price 2015 (`)	40	60	85	25	30
Price 2017 (`)	60	90	125	30	40

Sol :

Price

Commodity	2015 (P_0)	2017 (P_1)
P	40	60
Q	60	90
R	85	125
S	25	35
T	30	40
Total	$\Sigma p_0 = 240$	$\Sigma p_1 = 350$

$$P_{0,1} = \frac{\Sigma p_1}{\Sigma p_0} \times 100$$

$$= \frac{350}{240} \times 100 = 1.4583 \times 100$$

$$= 145.83$$

5. Calculate Index number by average price relative method by using arithmetic mean.

Commodity	P	Q	R	S	T
Price 2017	2	6	10	5	12
Price 2018	4	8	15	5	8

Sol :

Commodity	P_0	P_1	$P = \frac{P_1}{P_0} \times 100$
P	2	4	200.00
Q	6	8	133.33
R	10	15	150.00
S	5	5	100.00
T	12	8	66.66
			649.99

Average of price relatives using A.M.

$$P_{0.1} = \frac{\Sigma P}{N} = \frac{649.99}{5} = 129.998 = 130$$

6. From the following data compute Laspeyre's Index number for 2012 :

	Price		Quantity	
Items	**2014**	**2017**	**2014**	**2017**
P	**20**	**25**	**10**	**12**
Q	**18**	**32**	**16**	**10**
R	**35**	**48**	**8**	**12**
S	**28**	**40**	**12**	**10**

Sol :

Computation of the Laspeyre's Index number

Items	q_0	p_0	q_1	p_1	p_1q_0	p_0q_0
P	10	20	12	25	250	200
Q	16	18	10	32	512	288
R	8	35	12	48	384	280
S	12	28	10	40	480	336
Total	–	–	–	–	1626	1104

We have, $P_{01}\,(L) = \frac{\sum p_1q_0}{\sum p_0q_0} \times 100 = \frac{1625}{1104} \times 100 = 147.28$

4.4.5 Criterion of a Good Index Numbers

Q20. Explain the various tests of Consistency of Index Number.

Ans : **(Imp.)**

1. Unit Test

This test states that the formula of index number should be independent of the units in which the prices or quantities of various commodities (or items) are quoted. All the formulae, except the index number based on simple aggregate of prices (quantities) satisfy this test.

2. Time Reversal Test

This test was proposed by Prof. Irwin Fisher. According to fisher 'the formula for calculating the index number should be such that it gives the same ratio between one point of comparison and the other, no matter which of the two is taken as the base, or putting it another way, the index number reckoned forward should be reciprocal of the one reckoned backward.'

In simple terms, given two time periods I and II, if an index number is calculated for period II taking period I as base, its value should be the reciprocal value of the index number for period I taking period II as base. The index numbers for the purpose of the test, should be in decimal form and not in percentage form. In other words, $P_{0.1}$ and $P_{1.0}$ should not be multiplied with 100. Symbolically.

$$\boxed{P_{01} \times P_{1.0} = 1}$$

Where

$P_{0.1}$, = Fishers index for period II taking period I as base and

$P_{0.1}$, = Fishers index for period I taking Period II as base

Time Reversal test is satisfied by Marshall-Edge worth, Fisher, Walsh. Kelly' s index numbers and also by simple Aggregative Index, simple geometric mean of price relatives and weighted average of price relatives. Laspeyre's and Paasche's index numbers do not satisfy the time reversal test.

3. Factor Reversal Test

This test was also proposed by proof Fisher. This test requires that the product of two index numbers, one measuring price taking quantities as base, and the other measuring quantities taking price as base, should be equal to the net increase in total value from one period to another. Let us illustrate the same with the help of an example.

If $P_{0,1}$, is price index number, $Q_{0,1}$, is quantity index number, the product of the two index numbers should be equal to the value index number V_{01}

$$\boxed{P_{0,1} \times Q_{0,1} = V_{0,1}}$$

Fishers index number satisfies the factor reversal test. No other method satisfies the factor reversal test.

4. Circular Test

The circular test was proposed by Weztergaard. It is an extension of the time-reversal test. It more than two time periods are considered, price index is calculated for each period with the previous year as base period. Lastly, the price index for the first year is calculated taking the last period as the base. The product of all the price index numbers should be equal to 1. Symbolically, if three time periods are considered, $\boxed{P_{0,1} \times P_{1,2} \times P_{2,1} = 1}$

Only simple geometric mean of price relatives method and Kelly's method satisfy the circular test.

PROBLEMS

7. Calculate Fisher's Ideal Index from the following data and prove that it satisfies both the time reversal and factor reversal tests :

Commodity	**2010**		**2011**	
	Price	**Qty**	**Price**	**Qty**
A	**8**	**80**	**10**	**120**
B	**10**	**120**	**12**	**96**
C	**5**	**40**	**5**	**50**
D	**4**	**56**	**3**	**60**
E	**20**	**100**	**25**	**150**

Sol :

Calculation of Fisher's Ideal Index

Commodity	P_0	q_0	P_1	q_1	$p_1 q_0$	$p_0 q_0$	$p_1 q_1$	$p_0 q_1$
A	8	10	10	12	100	80	120	96
B	10	12	12	8	144	120	96	80
C	5	8	5	10	40	40	50	50
D	4	14	3	20	42	56	60	80
E	20	5	25	6	125	100	150	120
					451	396	476	426

$$p_{01} = \sqrt{\frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1}} \times 100 = \sqrt{\frac{451}{396} \times \frac{476}{426}} \times 100$$

$$\sqrt{1.2726} \times 100 = 1.125 \times 100 = 112.8$$

Time Reversal Test

Time reversal test is satisfied when $P_{01} \times P_{10} = 1$

$$p_{01} = \sqrt{\frac{\Sigma p_0 q_1}{\Sigma p_1 q_1} \times \frac{\Sigma p_0 q_0}{\Sigma p_1 q_0}} = \sqrt{\frac{426}{476} \times \frac{396}{451}}$$

$$P_{01} \times P_{10} = \sqrt{\frac{451}{396} \times \frac{476}{426} \times \frac{426}{476} \times \frac{396}{451}} = \sqrt{1} = 1$$

Hence, time reversal test is satisfied.

Factor Reversal Test

Factor reversal test is satisfied when :

$$P_{01} \times P_{10} = \frac{\Sigma p_1 q_1}{\Sigma p_0 q_0}$$

$$Q_{01} = \sqrt{\frac{\Sigma q_1 p_0}{\Sigma q_0 p_0} \times \frac{\Sigma q_1 p_1}{\Sigma q_1 p_0}}$$

$$P_{01} \times P_{10} = \sqrt{\frac{451}{396} \times \frac{476}{426} \times \frac{426}{396} \times \frac{476}{451}} = \frac{476}{396}$$

This is also the value of $\frac{\Sigma p_1 q_1}{\Sigma p_0 q_0}$. Hence, the above data also satisfies the Factor Reversal Test

8. Construct a Fisher's Ideal Indeed from the following data and show that it satisfies time reversal and factor reversal test :

	1995		1996	
Commodity	P_0	q_0	P_1	q_1
A	10	40	12	45
B	11	50	11	52
C	14	30	17	30
D	8	28	10	29
E	12	15	13	20

Sol :

Construction of Fisher's Ideal Index

Items	P_0	q_0	P_1	q_1	$p_1 q_0$	$p_0 q_0$	$p_1 q_1$	$p_0 q_1$
A	10	40	12	45	480	400	540	450
B	11	50	11	52	550	550	572	572
C	14	30	17	30	510	420	510	420
D	8	28	10	29	280	224	290	232
E	12	15	13	20	195	180	260	240
					$\Sigma p_1q_0 = 2015$	$\Sigma p_0q_0 = 1774$	$\Sigma p_1q_1 = 2172$	$\Sigma p_0q_1 = 1914$

Fisher's Ideal Index

$$p_{01} = \sqrt{\frac{\Sigma\, p_1q_0}{\Sigma\, p_0q_0} \times \frac{\Sigma\, p_1q_1}{\Sigma\, p_0q_1}}$$

$$= \sqrt{\frac{2015}{1774} \times \frac{2172}{1914}} \times 100 = 1.135 \times 100 = 113.5$$

Time Reversal Test

Time Reversal test is satisfied when

$$P_{01} \times P_{10} = 1$$

$$P_{10} \sqrt{\frac{\Sigma p_0q_1}{\Sigma p_1q_1} \times \frac{\Sigma p_0q_0}{\Sigma p_1q_0}} = \sqrt{\frac{1914}{2172} \times \frac{1774}{2015}}$$

$$P_{0,1} \times P_{1,0} = \sqrt{\frac{2015}{1774} \times \frac{2172}{1914} \times \frac{1914}{2172} \times \frac{1774}{2015}} = 1$$

Hence, time reversal test is satisfied by the given data

Factor Reversal Test : Factor reversal test is satisfied when :

$$P_{0,1} \times q_{0,1} = \frac{\Sigma p_1q_1}{\Sigma p_0q_0}$$

$$p_{01} = \sqrt{\frac{\Sigma q_1p_0}{\Sigma q_0p_0} \times \frac{\Sigma q_1p_1}{\Sigma q_0p_1}} = \sqrt{\frac{1914}{1774} \times \frac{2172}{2015}}$$

$$p_{0,1} \times q_{0,1} = \sqrt{\frac{2015}{1774} \times \frac{2172}{1914} \times \frac{1914}{1774} \times \frac{2172}{2015}} = \frac{2172}{1774}$$

$\frac{\Sigma p_1q_1}{\Sigma p_0q_0}$, is also equal to $\frac{2172}{1774}$. Hence factor reversal test is satisfied by the given data.

4.5 Problems Involved in the Construction of Index Numbers

Q21. What are the problems involved in construction of index numbers ? Explain.

Ans : **(Imp.)**

The following problems are mainly faced in the construction of index numbers :

i) Definition of the Purpose,

ii) Selection of the Base Period,

iii) Selection of Items,

iv) Selection of Sources of Data and Collection of Data,

v) Selection of Average,

vi) System of Weighting.

(i) Definition of the Purpose

There are no all purpose index numbers. Therefore, before constructing an index number the specific purpose, i.e., objective for which it is designed must be clearly and rigorously defined. Haberler has rightly said "Different index numbers are constructed to fulfil different objectives and before setting to construct a particular number one must clearly define one's objective of study because it is on the objective of the study that the nature and format of the index number depends".

(ii) Selection of the Base Period

Selection of the proper Base Period is an important factor in the construction of index numbers. Base Period is a reference point with which changes in other periods are measured. About the selection of Base Period, following observations may be noted.

Morris Hamburg observes, "It is desirable that the Base Period be not too far away in times from the present. The further away we move from the Base Period, the dimmer are our recollections of Economic conditions prevailing at that time. Consequently comparisons with these remote periods tend to lose significance and to become rather tenuous in meaning."

According to George Simpson and Fritz Katka, "Since practical decisions are made in terms of index numbers, and economic practices so often are a matter of the short run, we wish to make comparisons between a base which lies in the same general economic framework as the years of immediate interest. Therefore, we choose a base relatively close to the years being studied." In the fast changing world of today the base year should not be more than a decade old. There is a psychological reason also for taking a recent period as base.

(iii) Selection of Items

Selection of items or 'Regimen' or 'Basket'. In any index number neither it is possible nor necessary to include all the items or commodities. Each index number tries to measure changes pertaining to a particular group.

Selection of items also depends on the purpose of index number. Moreover, items selected should be such as are widely consumed. The items selected for an index number should be relevant, representative, reliable and comparable. In general the larger is the number of items, the lesser will be the chances of error in the average. Since we cannot include a very large number of items. A compromise is always required between the number of items and the reasonable standard of accuracy. We must, however, have manageable number of items and should also aim at reasonable standard of accuracy.

(iv) Selection of Sources of Data and Collection of Data

For sources of data and collection of data we are mainly concerned with she prices. Use of wholesale prices or retail prices depends on the objective of study. Price quotations should be obtained from important markets. In order to ensure better results, it is advisable to take a standard price which implies representative price of a commodity for whole interval under consideration.

(v) Selection of Average

Since index numbers measure the relative changes, Geometric mean should be the best average but due to certain difficulties in calculations with G.M., for all practical purposes Arithmetic mean is used.

(vi) System of Weighting

According to John Giffin, "In simple terms, weighting is designed to give component series an importance in proper relation to their real significance." In order to allow each commodity to have a reasonable influence on the index it is advisable to use a suitable weighting system.

In case of an unweighted index number of prices, all commodities are given equal importance. But in actual practice different commodities need a different degree of importance.

The weights may be according to :

- The value or quantity produced.
- The value or quantity consumed.
- The value or quality sold.

Selection of an Appropriate Formula

A large number of formulae have been devised for constructing the index. The problem very often is that of selecting the most appropriate formula. The choice of the formula would depend not only on the purpose of the index but also on the data available.

Prof. Irving Fisher has suggested that an appropriate index is that which satisfies time reversal test and factor reversal test. Theoretically. Fisher's method is considered as "ideal" for constructing index number. However, from a practical point of view there are certain limitations of this index which shall be discussed later. As such, no one particular formula can be regarded as the best under all circumstances. On the basis of this knowledge of the characteristics of different formulae, a discriminating investigator will choose technical methods adapted to his data and appropriate to his purposes.

4.6 Fixed and Chain base Index Numbers

Q22. Define Fixed and Chain base Index Numbers.

Ans :

Fixed base is the method of calculating index numbers. In this method, a specific year is chosen as base for calculating index numbers. The year which is chosen as base year should not face any abnormal situations like earthquakes, wars, etc., and also should not be too far from the current year. In this method, index numbers are calculated by price relative of current year.

$$\text{Price Relative} = \frac{\text{Current Years Link Relative} \times \text{Previous Years Price Relative}}{100}$$

Index number in fixed base method is calculated as,

$$P_{0n} = \frac{P_n}{P_0} \times 100$$

Where, P_n = price of current year

P = Price of base year

Chain Base Index Numbers

In this method of calculating Index numbers, no specific year is chosen as base year. Instead, immediately preceding year is taken as base year for calculating index numbers of a particular year. The process involved in calculation of chain base index numbers is as follows,

Step 1

Calculate the link relatives. The formula used for calculating link relatives is,

$$\text{Link Relatives} = \frac{\text{Price Relative for the Current Period}}{\text{Price Relative for the Previous Period}} \times 100$$

Step 2

Calculate the chain inx dex of current year. The formula used for calculating chain index of current year is as follows,

$$\text{Chain Index to Current Year} = \frac{\text{Link Relative of Current Year} \times \text{Chain Index of Previous Year}}{100}$$

Q23. Distinguish between fixed base and chain base index numbers.

Ans : **(July-22)**

The difference between fixed base method and chain base method are as follows:

Sl.No.	Fixed Base Method	Sl.No.	Chain Base Method
1.	The year with normal events is taken as the base year.	1.	The year preceding the current year for which the index is to be obtained,
2.	The base year remains constant for computing index number for the given time period.	2.	The base year keeps on changing for the given time period.
3.	Since the base year is fixed, uniformity is maintained in comparing changes in the values of a variable quantity.	3.	Since the base year is changing, uniformity is not maintained in comparing changes in the values of variable quantity.
4.	In this method, new items in demand cannot be included and old items out of use or having no preferences cannot	4.	This method permits inclusion of items in demand and exclusion of items out of use or having no preferences.
5.	The work of selection of a base year is difficult.	5.	The question of selecting a base year does not arise as it is automatically selected.
6.	The base year is to be changed with lapse of time.	6.	The question of selecting a base year does not
7.	This method is quite useful for comparing long-term changes in the value of a variable quantity.	7.	This method is useful only for comparing short-term changes in the value of a variable quantity.
8.	It is easy to understand and easy in computation	8.	In this method if there is any mistake in the calculations of Index number of one year then that mistake is carried on in all the subsequent years.

PROBLEM

9. Distinguish between fixed base and chain base index numbers. From the fixed base index numbers given below, construct chain base index numbers.

Year	2003	2004	2005	2006	2007	2008
Fixed Base index	94	98	102	95	98	100

Sol : **(July-22)**

Conversion of FBI to CBI

Year	F.B.I	Link Relatives	Chain Base Index Numbers
2003	94	94	94
2004	98	$\frac{98}{94} \times 100 = 104.25$	$\frac{94 \times 104.25}{100} = 97.99$
2005	102	$\frac{102}{98} \times 100 = 104.08$	$\frac{98 \times 104.08}{100} = 101.99$
2006	95	$\frac{95}{102} \times 100 = 93.13$	$\frac{95 \times 93.13}{100} = 88.47$
2007	98	$\frac{98}{95} \times 100 = 103.15$	$\frac{98 \times 103.15}{100} = 101.09$
2008	100	$\frac{100}{98} \times 100 = 102.04$	$\frac{100 \times 102.04}{100} = 102.04$

4.7 COST OF LIVING INDEX NUMBERS AND WHOLESALE PRICE INDEX NUMBERS

Q24. Define Cost of Living Index Numbers. Explain various methods of construction of Cost of Living Index Numbers.

(OR)

Define cost of living index numbers. Describe various methods of its computation.

Ans : **(July-22)**

Cost of Living Index or Consumer Price Index is an index number measuring change in retail prices. While WPI (Wholesale Price Index) measures changes in general level of prices in the economy, they do not reflect changes in cost of living standard of any chosen group of people. The importance of various commodities is different for different types of people. Hence, separate indices are constructed for different groups. Therefore, each index tells us about the variations in cost of living of only a particular group. Moreover, the construction of Consumer Price Index takes into account retail prices and hence, reflects the cost of living of consumers with greater accuracy.

Construction of Cost of Living Index Numbers

Cost of Living Index Numbers are weighted index numbers. The commodities that constitute the index are given weights according to their importance. Normally, the weights are in the ratio of amounts spent on each item. There are two methods of constructing the cost of Living Index Numbers.

1. Aggregate Expenditure Method (or) Weighted Aggregative Method

Thus method is similar to Laspeyre's method. The quantities consumed in the base year are taken as weights. The formula is :

$$\text{Consumer Price Index} = \text{Cost of Living Index} = \frac{\Sigma p1q0}{\Sigma p0q1} \times 100$$

Since P represents price and q represents quantity, pq is the amount spent of given commodity. Thus Σpq represents total amount spent on all items. In other words it represents total expenditure. Σp_0q_0 total expenditure incurred in the base period. Σp_1q_0 is total expenditure incurred in the current year at base period. Thus

$$\text{Consumer Price Index} = \frac{\text{Total Expenditure in Current Year at Base Year Price}}{\text{Total Expenditure in Base Year}}$$

2. Family Budget Method (or) Method of Weighted Relatives

The cost of living index is obtained by taking a weighted average of price relatives. The quantities consumed in the base year are taken as weights.

The formula is:

$$\text{Cost of Living Index} = \frac{\Sigma PV}{\Sigma v} \text{ where}$$

$P = (p_1/p_0) \times 100$ for each item and V = Value Weights = 100

$$\text{Thus, Cost of Living Index} = \sum \frac{\left(\frac{p_1}{p_0}\times 100\right)\times p_0q_0}{\Sigma p_0q_0} = \frac{\Sigma p_1q_0}{\Sigma p_0q_0} \times 100$$

Thus, the cost of living index figure is one and the same, irrespective of method of construction.

PROBLEMS

10. Calculate the index nuber using both the aggregate. Expenditure method and Family budget method for the year 1973 with 1960 as base year from the following data.

Commodity	Quantity in Units in 1960	Price per units in 1960 (Rs)	Price per unit in 1973 (Rs)
A	100	8.00	12.00
B	25	6.00	7.50
C	10	5.00	5.25
D	20	48.00	52.00
E	25	15.00	16.50
F	30	9.00	27.00

Sol :

Calculation of Consumer Price Index

Commodity	Quantity (q_0)	Price (1960) P_0	Price (1975) P_1	p_0q_0 = V	p_1q_0 × 100	P = (P_1/P_0)	PV
A	100	8.00	12.00	800	1200	150	120000
B	25	6.00	7.50	150	187.50	125	18750
C	10	5.00	5.25	50	52.50	105	5250
D	20	48.00	52.00	960	1040.00	108.33	1040000
E	25	15.00	16.50	375	412.50	110	41250
F	30	9.00	27.00	270	810.00	300	81000
				2605	**3702.50**		**370205**

(1) Aggregate Expenditure Method

$$CP_1 = \frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times 100 = \frac{3702.05}{2605} \times 100 = 142.13$$

(2) Family budget Method

$$CP_1 = \frac{\Sigma PV}{\Sigma V} = \frac{3702.05}{2605} = 142.13.$$

11. In the construction of a certain cost of living number, the following group index numbers are found. Calculate the Cost of Living Index Number by using (i) Weighted Arithmetic Mean and (ii) Weighted Geometric Mean.

Group	Index Numbers	Weights
Food	**350**	**5**
Fuel and Lighting	**200**	**1**
Clothing	**240**	**1**
House Rent	**160**	**1**
Miscellaneous	**250**	**2**

Sol :

Computation of Consumer Price Index

Group	Index No. (I)	Weights (W)	Weighted (WI)	Log I	W.log I
Food	350	5	1750	2.5441	12.7205
Fuel Lighting	200	1	200	2.3010	2.3010
Clothing	240	1	240	2.3802	2.3802
House Rent	160	1	160	2.2041	2.2041
Miscellaneous	250	2	500	2.3979	4.7958
		10	**2850**		**24.4016**

Consumer Price Index

(i) Using Arithmetic Mean $= \frac{\Sigma IW}{\Sigma W} = \frac{2850}{10} = 285$

(ii) Using Geometric Mean $= \text{Antilog}\left[\frac{\Sigma w \log 1}{\Sigma w}\right] = \text{antilog}\left[\frac{24.4016}{10}\right]$

$= \text{Antilog } 2.44016 = 27.55$

Q25. Explain briefly about wholesale price index numbers.

Ans :

The general change occurred in the wholesale prices of the commodities which are represented in index numbers is termed as wholesale price index number. Wholesale price indices are used to study the changes that take place in the general price level of a country.

The Indian Ministry of Commerce and Industry constructed the first wholesale price index number in January 1947 with 1939 as base year.

Various Wholesale Price Index Numbers

On the recommendation of Wholesale Industrial Price Review Committee, the series of construction of index number was started. They are discussed as below,

(i) Revised index number of wholesale prices

(ii) New series of index numbers of wholesale prices

(iii) Series of index numbers of wholesale prices

(iv) Wholesale price index number series

(v) Current series of wholesale price indices.

4.8 Base Shifting, Splicing and Deflation of Index Numbers

Q26. What is Base Shifting ?

Ans :

For a variety of reasons, it frequently becomes necessary to change the reference base of an index number series from one time period to another without returning to the original raw data and recomputing the entire series. This change of reference base period is usually referred to as *"shifting the base".* There are two important reasons for shifting the base :

(i) The previous base has become too old and is almost useless for purposes of comparison. By shifting the base it is possible to state the series in terms of a more recent time period.

(ii) It may be desired to compare several index number series which have been computed on different base periods; particularly if the several series are to be shown on the same graph, it may be desirable for them to have the same base period. This may necessitate a shift in the base period.

When base period is to be changed, one possibility is to recompute all index numbers using the new base period. A simpler approximate method is to divide all index numbers for the various years corresponding to the old base period by the index number corresponding to the new base period, expressing the results as percentages. These results represent the new index numbers, the index number for the new base period being 100 per cent.

Mathematically speaking, this method is strictly applicable only if the index numbers satisfy the circular test.

$$\text{Index number with new base} = \frac{\text{Index of current year}}{\text{Index of new base year}} \times 100$$

PROBLEMS

12. Reconstruct the series of index numbers given below by shifting the base to 2010.

Year	2004	2005	2006	2007	2008	2009	2010	2011	2012
Index No	100	120	132	140	150	164	180	208	220

Sol :

The given series of index numbers along with another series with 2010 = 100 is shown in Table. To illustrate, for 2004, we have (100/180) × 100 = 55.56

Shifting the Base Year

Year	Index No. 2004 = 100	Index No. 2010 = 100
2004	100	55.56
2005	120	66.67
2006	132	73.33
2007	140	77.78
2008	150	83.33
2009	164	91.11
2010	180	100.00
2011	208	115.56
2012	220	122.22

13. The index of 2010 is 100. It rises by 5% in 2011, falls 2% in 2012, rises 4% in 2013, rises 10% in 2014, falls 3% in 2015, and rises 8% in 2016. Find out the indices for the 7 years assuming that all the increases and decreases are the percentages of the respective proceeding years - Also recast the indices shifting base to 2014.

Sol : **(Jan.-21)**

Calculation of Index number for base year and change of base year 2014

Year	Old Index Number (Base Year 2010 = 100) $= \frac{100+\%}{100} \times$ Previous Year Index Number	New Index Number (New Base Year 2014 = 116.59) $= \frac{100}{\text{Value of New Base Year 2014}} \times$ old index number of year
2010	100 (Given)	$\frac{100}{116.59} \times 100 = 85.77$
2011	$\frac{100+2\%}{100} \times 100 = 104$	$\frac{100}{116.59} \times 104 = 89.20$
2012	$\frac{100-2\%}{100} \times 104 = 101.92$	$\frac{100}{116.59} \times 101.92 = 87.42$

2013	$\frac{100+4\%}{100}\times 101.92 = 105.99$	$\frac{100}{116.59}\times 105.99 = 90.91$
2014	$\frac{100+10\%}{100}\times 105.99 = 116.59$	$\frac{100}{116.59}\times 116.59 = 99.99$
2015	$\frac{100-3\%}{100}\times 116.59 = 113.09$	$\frac{100}{116.59}\times 113.09 = 96.99$
2016	$\frac{100+8\%}{100}\times 113.09 = 122.14$	$\frac{100}{116.59}\times 122.14 = 104.76$

14. The following are the indices (2007. Base)

Year	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017
Indices	100	120	122	116	120	120	137	136	149	156	137

Shift the base to 2012 and recast the index numbers.

Sol.

$$\text{Current year F.B.I} = \frac{\text{Current tear C.B.I}\times\text{Previous year F.B.I}}{100}$$

The first year F.B.I being same as first year C.B.I; we obtain the F.B.I. numbers as given in table.

Conversion of C.B.I Numbers to F.B.I Numbers

Calculation of Index number for base year

Year	Indics	Fixed Base Index Number (Base 2012)
2007	100	$\frac{100}{120}\times 100 = 83.33$
2008	120	$\frac{120}{120}\times 100 = 100$
2009	122	$\frac{122}{120}\times 100 = 101.66$
2010	116	$\frac{116}{120}\times 100 = 96.66$
2011	120	$\frac{120}{120}\times 100 = 100$
2012	120	$\frac{100}{120}\times 100 = 100$
2013	137	$\frac{137}{120}\times 100 = 114.16$

2014	136	$\frac{136}{120} \times 100 = 113.33$
2015	149	$\frac{149}{120} \times 100 = 124.16$
2016	156	$\frac{156}{120} \times 100 = 130$
2017	157	$\frac{157}{120} \times 100 = 114.16$

Q27. What is Splicing?

Ans : **(July-22)**

Combining two or more series of overlapping index numbers to obtain a single index number on a common base is called splicing of index numbers. Splicing of index numbers can be done only if the index numbers are constructed with the same items and have an overlapping year.

Splicing is generally done when an old index number with an old base is being discontinued and a new index with a new base is being started.

The process of splicing is very simple and is akin to that used in shifting the base. It is expressed in the form of a formula as follows :

$$\text{Spliced Index No.} \ \frac{\text{Index No. of current year} \times \text{Old Index of New Base Year}}{100}$$

PROBLEMS

15. The index A given was started in 1996 and continued up to 2006 in which year another indexed B was started. Splice the index B to index A so that a continuous series of index.

Year	Index A	Index B
1996	100	
1997	110	
1998	112	
2005	138	
2006	150	100
2007		120
2008		140
2009		130
2010		150

Sol :

Index B Spliced to Index A

Year	Index A	Index B	Index B spliced to index A 1982 as base
1996	100		
1997	110		
1998	112		
–			
–			
2005	138		
2006	150	100	$\frac{150}{100} \times 100 = 150$
2007		120	$\frac{150}{100} \times 120 = 180$
2008		140	$\frac{150}{100} \times 140 = 210$
2009		130	$\frac{150}{100} \times 130 = 195$
2010		150	$\frac{150}{100} \times 150 = 225$

The spliced index now refers to 1996 as base and we can make a continuous comparison of index numbers from 1996 onwards

In the above, it is also possible to splice the new index in such as manner that a comparison could be made with 2006 as base. This would be done by multiplying the old index by the ratio $\frac{100}{150}$. Thus the spliced index for 1996 would be $\frac{100}{150} \times 100 = 66.7$ for 1997, $\frac{100}{150} \times 110 = 73.3$ for 1998, $\frac{100}{150} \times 112 = 74.6$ etc. This process appears to be more useful because a recent year can be kept as a base. However, much would depend upon the object.

The continuation of series A or series B may be done by the equivalence of the index values in the overlapping year. Splicing of B to A is also called forward splicing while splicing of series A to B is called backward splicing.

16. Consider the following series of index numbers :

Year	2004	2005	2006	2007	2008	2009	2010	2011	2012
Series A: (2004=100)	100	120	150	180	220				
Series B : (2008=100)					100	110	150	160	175

a) Splice series A to series B

b) Splice series B to series A.

Sol :

Hence series A has base year as 2004 while series B has base year as 2008. The year overlapping in the two series is 2008 for which respective index value are 220 and 100. To splice A to B, we need to find the index values for the years 2004, 2005, etc. These can be obtained by multiplying the ratio 100/220 to the given index value of the year. For the year 2004, for example, we have

$$\text{Index value} = \frac{100 \times 100}{220} = 45.45$$

Splicing Index Number Series

Year	Series A	Series B	Splicing A to B	Splicing B to A
2004	100		45.45	100
2005	120		54.55	120
2006	150		68.18	150
2007	180		81.82	180
2008	220	100	100.00	220
2009		110	110.00	242
2010		150	150.00	330
2011		160	160.00	352
2012		175	175.00	385

Q28. What is Deflating ?

Ans :

The index numbers can be used to eliminate from a given series the effect of inflation over the long term. The income of a worker may be observed to increase over years on account of promotions, salary increments, etc. This is because the prices also tend to rise over time. Now, if the rise in the prices is higher than the rise in his money income, then the purchasing power of his income, or the real income as it is called, would in fact decrease. If, however, the price increases are slower than the income increases, then the real income would rise. Thus, we calculate real income by adjusting the money income by appropriate price index. A similar calculation can be done to obtain the Gross Domestic Product (GDP) of a country in real terms for which GDP at market prices is adjusted for the price level changes. This process of eliminating the price effect from a given set of monetary values is termed deflating. The real values are obtained by dividing the monetary values by the price index value and multiplying the result by 100.

Q29. What are the various methods of Deflating?

Ans :

1. Purchasing power of money

This can be deflated by the following formula

$$\text{Purchasing power of money} = \frac{100}{\text{Price index}}$$

Thus, if the price rises by 25%, the price index becomes 125 and in that case, the purchasing power of every rupee would be $\frac{100}{125}$ = 0.80 ps. Or 80 ps. This means that a rupee in the current year is equal to 80 ps. in the base year.

2. Real wage, or income

This can be deflated by the following formula :

Real Wage $= \frac{100}{\text{Price Index}} \times$ Money Wage or $\boxed{\text{Real wage } = \frac{\text{Mone wage}}{\text{Price index}} \times 100}$

Here, price index should preferably be the consumer price index rather than the wholesale price index as the former reflects very well the change in the purchasing power of a wage earner.

Thus, if a worker earns ` 1500 during a year in which the price index stands at 150, the real value of his wage in comparison to the wage of the base year would be $\frac{1500}{150} \times 100$ = ` 1,000. This means that his present wage of ` 1500 is equal to the wage of ` 1000 earned in the base year.

3. Real wage index, or Real income index

This can be deflated by the following formula :

Real Wage Index $= \frac{100}{\text{Price Index}} \times$ index of Money Wage

or Real wage index no. $= \frac{\text{Index of money wage}}{\text{Price index number}} \times 100$

or Again, $\boxed{\text{RWI } = \frac{\text{Real wage of the year}}{\text{Real wage of the base year}} \times 100}$

PROBLEMS

17. The table below shows the average wages in rupees per day of a group of industrial workers during the years 2005-2012. The consumer price indices for these years are also shown

a) Determine the real average wages of the workers furring the years

b) Calculate the index of real wages taking 2005=100

c) Represent the average actual and real wages graphically

Year	2005	2006	2007	2008	2009	2010	2011	2012
Average Wage	119	133	144	157	175	184	189	194
CPI No	100.0	107.6	106.6	107.6	116.2	118.9	119.8	120.2

Sol :

The real wages may be calculated as follows :

Real wage $= \frac{\text{Average wage}}{\text{CPI No.}} = 100$

The illustrate, for the year 2006, we have

$$\text{Real wage} = \frac{133}{107.6} \times 100 = 123.61$$

Further, to calculate the real wage indices, we take the index for the year 2005 as 100. For each year, we have

$$\text{Real wage index} = \frac{\text{Average real wage for the year}}{\text{Average real wage for 2005}} \times 100$$

For example, for the year 2006, we have

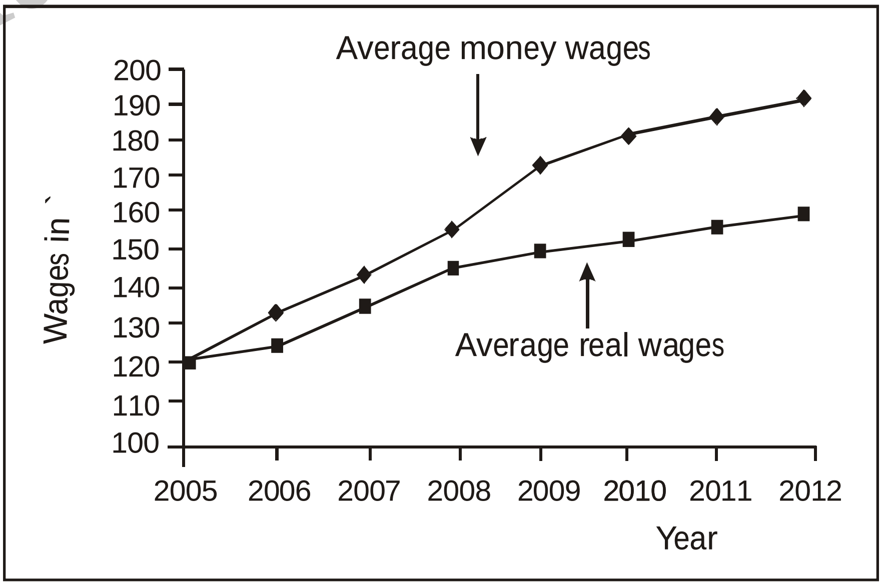
$$\text{Real wage index} = \frac{123.61}{119.00} \times 100 = 103.87$$

The results are given in Table

Calculation of Real Wages and Real Wages Indices

Year	Average Wage (`)	CPI No.	Average Real Wage(`)	Real Wage Index
2005	119	100.0	119.00	100.00
2006	133	107.6	123.61	103.87
2007	144	106.6	135.08	113.52
2008	157	107.6	145.91	122.61
2009	175	116.2	150.60	126.56
2010	184	118.9	154.75	130.04
2011	189	119.8	157.76	132.57
2012	194	120.2	161.40	135.63

The actual and real wages are shown plotted in figure. It is evident that while both of these are rising with time, the increase in real wages has been lower than the increase in money wages



18. The following table gives the annual income of worker and the general index numbers of price during 1988-1996. Prepare Index Number to show the changes in the real income of the teacher and comment on price increase.

Year	2001	2002	2003	2004	2005	2006	2007	2008	2009
Income Price	3600	4200	5000	5500	6000	6400	6800	7200	7500
Index No	100	120	145	160	250	320	450	530	600

Sol :

Index Number showing Changes in the Real Income of the Worker

Year	Income (Rs.)	Price Index No.	Real Income (Rs.)	Real Income Index No.
2001	3600	100	$\frac{3600}{100}\times 100 = 3600.00$	100.00
2002	4200	120	$\frac{4200}{120}\times 100 = 3500.00$	97.00
2003	5000	145	$\frac{5000}{145}\times 100 = 3448.27$	95.78
2004	5500	160	$\frac{5500}{160}\times 100 = 3437.50$	95.48
2005	6000	250	$\frac{6000}{250}\times 100 = 2400.00$	66.60
2006	6400	320	$\frac{6400}{320}\times 100 = 2000.00$	55.55
2007	6800	450	$\frac{6800}{450}\times 100 = 1511.11$	41.97
2008	7200	530	$\frac{7200}{530}\times 100 = 1358.49$	37.73
2009	7500	600	$\frac{7500}{600}\times 100 = 1250.00$	34.72

Short Question and Answers

1. Explain various methods of national income.

Ans :

i) Output or production method

Output (or) production method is also called inventory method. It consists of finding out the market value of all goods and services produced by the individuals and business enterprises during a specified period of time. The national income or net domestic product can be calculated by the following equation.

National income = (Value of goods and services + self consumption + increase in stock – depreciation of capital – income from abroad).

ii) Income Method

The income method consists of adding together all incomes by the way of wages, interests rents and profits wages include the value of services (mental and physical), in terms of money rendered by the nationals. Interest is the amount of money, received by the entrepreneurs, by way of capital investment. Similarly, rent may be defined as the factor income generated by the letting and the use of land for agricultural and other purposes, buildings (residential and non residential), machinery, equipment and other fixed assets. It is treated as rental income from property. Also any amount of money saved by the individuals or institutions, after deducting the expenditure on production, may be termed as profit.

iii) Expenditure Method

The national income is equal to the sum of consumption and saving. Symbolically we can write.

$$y = C + S$$

Where y = National income

C = Consumption

S = Saving

The method is rarely used because of the non-availability of data required for its calculation. Because of the predominance of unorganized sectors in underdeveloped and developing countries, there is a lack of proper maintenance of records of consumption, expenditure savings etc.

2. Agricultural Statistics.

Ans :

India is an agricultural country and its economy largely depends on agriculture. In ancient days, agriculture had been the main source of income of all the rulers. They collected money in the form of land revenue. Therefore, most of the rulers maintained records of agricultural land and their categories on the basis of fertility and production. Records were also maintained for orchards. The main objective of keeping there records was the collection of revenue.

After independence, a lot of emphasis has been given to agriculture. Indian economy is based mostly on agriculture even today, hence a comprehensive and reliable statistics are necessary for good planning and development of agriculture and that of the country. As a matter of fact, all statistics which have an impact on agricultural economy may be regarded as agricultural statistics. The statistics pertaining to land utilization, production of crops, live stock and agricultural prices poultry, foresty, etc. will come under the category of agricultural statistics.

3. Functions of CSO

Ans :

i) Coordination of statistical activities at the centre and the state.

ii) Advisory work concerning the statistical matter, particularly standardization of concepts and definitions to maintain uniformity, throughout the country.

iii) Collection of statistical data related to planning.

iv) Training of statistical personnel.

v) Compilation of national income estimates.

vi) To provide statistical data of the nation to the united nations statistical offices and other international institutions.

vii) To plan and coordinate the conduct of the annual survey of industries and publish the results.

viii) To attend to the work of international statistical institutes (conferences) held in India and abroad.

ix) To display of charts and graphs pertaining to the national data which are of administrative interest.

x) Circulation of regular publications.

4. Define National Income.

Ans :

National income estimates are of great importance to get a broad view of the entire economy of a country. They also provide the information about the changes occurring in economy from year to year. The economic policies and planning are mostly framed keeping in view, the national income of a nation. The National Income Committee (NIC) set up by the government of India in 1949, produced for the first time national income estimates for the entire Indian union. The estimates along with the methodology were published in the first and final report of the national income committee (ministry of finance) in 1951 and 1954. Efforts were made to improve the empirical data and overcome the gaps. The first result of these efforts were presented by the CSO in the National income statistics.

National income (or) national dividend is the sum of total goods and services produced in a country during the given period of time, generally in a year. It is the total output produced by the four factors for production in a country in a given year.

5. Use of National Income

Ans :

1. The first and foremost use of national income estimate is that, it gives us a correct picture of the structure of the economy of the nation, as well as the distribution of income according to regions, industrial origin and income from functional services and persons during a specified period of time.
2. National income statistics provide a useful guideline in the formulation of the budget of a country.
3. National income figures have been found useful for studies of the problems of the economically underdeveloped countries.
4. National income estimates give us an Idea of the purchasing power of the people in the country inflationary and deflationary gaps of purchasing power are revealed by income and product figures. This is of great value in working out the details and timing of anti-inflationary and deflationary programmes.
5. National income statistics help in making the interregional and interstate comparisons of national income distribution. It is also useful in comparing the economic conditions of a country over a period of time.
6. National income estimates provide a basis for the future planning of a country. In the absence of such statistics, no comprehensive economic plan can be chalked out. It also helps in evaluating the success of an economic plan.

6. Functions of NSSO

Ans :

1. Collection of socio-economic data relating to demographic conditions for the whole country on regular basis.
2. To provide statistical data for national income and planning.
3. To conduct annual surveys in the organised industrial sector.
4. Training of personnel and providing guidance to the states in the conduct of surveys.

7. What is Fisher's Ideal Index ?

Ans :

Fisher's Ideal Index

This is the most popular amongst all weighted aggregative index numbers. It is obtained by calculating the Geometric Mean (G.M) of Laspeyre's and Paasche's index numbers. The formula for calculating fisher's ideal index is an under.

$$P_{0.1} = \left[\sqrt{\frac{\sum p_1q_0 \times \sum p_1q_1}{\sum p_0q_0 \times \sum p_0q_1}}\right] \times 100$$

Reasons for Fisher's Index being called an Ideal Index

Reasons

(i) It gives weightage to both current consumption and base year consumption.

(ii) It is free from upward or downward bias.

(iii) It satisfies both time reversal and factor reversal tests.

(iv) It is a Geometric mean of Laspeyre's index and Paasche's index

8. What is Splicing?

Ans :

Combining two or more series of overlapping index numbers to obtain a single index number on a common base is called splicing of index numbers. Splicing of index numbers can be done only if the index numbers are constructed with the same items and have an overlapping year.

Splicing is generally done when an old index number with an old base is being discontinued and a new index with a new base is being started.

The process of splicing is very simple and is akin to that used in shifting the base. It is expressed in the form of a formula as follows :

$$\text{Spliced Index No.} \; \frac{\text{Index No. of current year} \times \text{Old Index of New Base Year}}{100}$$

9. Define cost of living index numbers.

Ans :

Cost of Living Index or Consumer Price Index is an index number measuring change in retail prices. While WPI (Wholesale Price Index) measures changes in general level of prices in the economy, they do not reflect changes in cost of living standard of any chosen group of people. The importance of various commodities is different for different types of people. Hence, separate indices are constructed for different groups. Therefore, each index tells us about the variations in cost of living of only a particular group. Moreover, the construction of Consumer Price Index takes into account retail prices and hence, reflects the cost of living of consumers with greater accuracy.

10. What is Base Shifting ?

Ans :

For a variety of reasons, it frequently becomes necessary to change the reference base of an index number series from one time period to another without returning to the original raw data and recomputing the entire series. This change of reference base period is usually referred to as *"shifting the base".* There are two important reasons for shifting the base :

(i) The previous base has become too old and is almost useless for purposes of comparison. By shifting the base it is possible to state the series in terms of a more recent time period.

(ii) It may be desired to compare several index number series which have been computed on different base periods; particularly if the several series are to be shown on the same graph, it may be desirable for them to have the same base period. This may necessitate a shift in the base period.

When base period is to be changed, one possibility is to recompute all index numbers using the new base period. A simpler approximate method is to divide all index numbers for the various years corresponding to the old base period by the index number corresponding to the new base period, expressing the results as percentages. These results represent the new index numbers, the index number for the new base period being 100 per cent.

Mathematically speaking, this method is strictly applicable only if the index numbers satisfy the circular test.

$$\text{Index number with new base} = \frac{\text{Index of current year}}{\text{Index of new base year}} \times 100$$

Choose the Correct Answer

1. Which of the following are the major characteristics of index numbers? [d]
   (a) It is expressed in percentages
   (b) It measures the net or relative changes in variables
   (c) It measures changes over a period of time
   (d) All of the above
2. The index number for base year is always __________. [c]
   (a) 1000 (b) 200
   (c) 100 (d) None of the above
3. The time period for which an index number is determined is known as _______. [c]
   (a) Base period (b) Normal period
   (c) Current period (d) None of the above
4. Index number is a type of ___________. [c]
   (a) Dispersion (b) Correlation
   (c) Average (d) None of the above
5. Which of the following statements is/are correct about average prices if the price index is 110? [a]
   (a) The prices have increased by 10 per cent
   (b) The prices have increased by 110 per cent
   (c) The prices have decreased by 10 per cent
   (d) None of the above
6. Which of the following are limitations of using index numbers? [d]
   (a) The use of each index number is restricted to a specific object
   (b) It ignores the quality of commodities
   (c) It is useful only for short term comparison
   (d) All of the above
7. Which of the following is the variation within two or more variable studies by the index? [b]
   (a) Price index (b) Composite index
   (c) Simple index (d) None of the above
8. The weights used in a quantity index are _______. [c]
   (a) Quantity (b) Values
   (c) Price (d) None of the above
9. Which of the following methods is used to calculate the Consumer Price Index? [a]
   (a) Laspeyres's formula (b) Fisher's formula
   (c) Palgrave's formula (d) None of the above
10. An index number that can serve many purposes is known as a ______. [a]
   (a) General purpose index (b) Special purpose index
   (c) Both a and b are incorrect (d) Both a and b are correct

Fill in the blanks

1. The ________ examines various schemes running at the state level and makes specific suggestions for their improvement.
2. NSSO stands for ________.
3. Output (or) production method is also called ________ .
4. ________ are devices for measuring differences in the magnitude of a group of related variables.
5. Time Reversal Test was proposed by ________ .
6. The circular test was proposed by ________.
7. ________ is an index number measuring change in retail prices.
8. The cost of living index is obtained by taking a weighted average of ________ .
9. NIC stands ________.
10. ________ Method is similar to Laspeyre's method.

ANSWERS

1. CSO
2. National Sample Survey Ogranisation
3. Inventory method
4. Index numbers
5. Prof. Irwin Fisher
6. Weztergaard
7. Consumer Price Index
8. Price Relatives
9. National Income Cm
10. Paasche's

FACULTY OF SCIENCE

B.Sc. VI-Semester(CBCS) Examination

MODEL PAPER - I

Subject : STATISTICS

Paper - VI - A : Applied Statistics - II

Time : 3 Hours] **[Max. Marks : 80**

PART - A (8 × 4 = 32 Marks)

Note : Answer any Eight questions. All questions carry equal marks.

ANSWERS

1. Replication. **(Unit-I, SQA-3)**
2. Applications of Design of Experiments. **(Unit-I, SQA-9)**
3. Find the expectation of error sum of squares in two-way classification. **(Unit-I, SQA-2)**
4. Randomized Block Design (R.B.D). **(Unit-II, SQA-10)**
5. Advantages of LSD over CRD. **(Unit-II, SQA-4)**
6. Discuss in detail about statistical analysis of CRD. **(Unit-II, SQA-1)**
7. Specific Death Rate. **(Unit-III, SQA-6)**
8. Assumptions of Life table. **(Unit-III, SQA-4)**
9. Define vital statistics. **(Unit-III, SQA-7)**
10. What is Base Shifting ? **(Unit-IV, SQA-10)**
11. What is Splicing ? **(Unit-IV, SQA-8)**
12. Agricultural Statistics. **(Unit-IV, SQA-2)**

PART - B (4 × 12 = 48 Marks)

Note : Answer all the questions. All questions carry equal marks.

13. (a) Discuss the concept of Gauss-Markoff Linear Model with Examples. **(Unit-I, Q.No.1)**

OR

(b) Explain the analysis of variance of two-way classification. **(Unit-I, Q.No.6)**

14. (a) Explain in detail about Completely randomized Design (C.R.D). **(Unit-II, Q.No.1)**

OR

(b) How do you estimate the missing observations in LSD. **(Unit-II, Q.No.9)**

15. (a) Explain various measurement of mortality. **(Unit-III, Q.No.4)**

OR

(b) Define Life table. What are the assumptions of Life table. **(Unit-III, Q.No.7)**

16. (a) Explain Functions and organization of CSO. **(Unit-IV, Q.No.2)**

OR

(b) Construct a Fisher's Ideal Indeed from the following data and show that it satisfies time reversal and factor reversal test :

	1995		1996	
Commodity	P_0	q_0	P_1	q_1
A	10	40	12	45
B	11	50	11	52
C	14	30	17	30
D	8	28	10	29
E	12	15	13	20

(Unit-IV, Prob.8)

FACULTY OF SCIENCE

B.Sc. VI-Semester(CBCS) Examination

MODEL PAPER - II

Subject : STATISTICS

Paper - VI - A : Applied Statistics - II

Time : 3 Hours] **[Max. Marks : 80**

PART - A (8 × 4 = 32 Marks)

Note : Answer any Eight questions. All questions carry equal marks.

ANSWERS

1. Assumptions **(Unit-I, SQA-7)**
2. Randomisation **(Unit-I, SQA-5)**
3. Local Control **(Unit-I, SQA-4)**
4. How do you estimate the missing observations in LSD? **(Unit-II, SQA-6)**
5. ANOVA TABLE FOR $m \times m$ L.S.D. **(Unit-II, SQA-2)**
6. Advantages of LSD over CRD. **(Unit-II, SQA-4)**
7. Explain the uses of vital statistics. **(Unit-III, SQA-1)**
8. Abridged life tables. **(Unit-III, SQA-3)**
9. Crude Birth Rate (C.B.R) **(Unit-III, SQA-5)**
10. Explain various methods of national income. **(Unit-IV, SQA-1)**
11. Functions of CSO **(Unit-IV, SQA-3)**
12. Use of National Income **(Unit-IV, SQA-5)**

PART - B (4 × 12 = 48 Marks)

Note : Answer all the questions. All questions carry equal marks.

13. (a) Suppose that we are interested in establishing the yield producing ability of four types of soya beans A, B, C and D. We have three blocks of land X,Y and Z which may be different in fertility. Each block of land is divided into four plots and the different types of soya beans are assigned to the plots in each block by a random procedure. The following results are obtained:

	Soya Bean			
Block	**Type A**	**Type B**	**Type C**	**Type D**
X	5	9	11	10
Y	4	7	8	10
Z	3	5	8	9

Test whether A,B,C and D are significantly different. **(Unit-I, Prob.6)**

OR

(b) Find the expectation of error sum of squares in two-way classification. **(Unit-I, Q.No.8)**

14. (a) Explain the Statistical Analysis of LSD. **(Unit-II, Q.No.8)**

OR

(b) In the below table are the yields of 6 varieties in a 4 replicate experiment for which one value is missing. Estimate, the missing value and analyze the data.

Block	**1**	**2**	**3**	**4**	**5**	**6**	**Block Totals (B_1)**
1	18.5	15 7	16.2	14.1	13.0	13.6	91.1
2	11.7	–	12.9	14.4	16.9	12.5	68.4
3	15.4	16.6	15.5	20.3	18.4	21.5	107.8
4	16.5	18.6	12.7	15.7	16.5	18.0	98.0
Treatment Totals (T_1)	62.1	50.9	57.3	64.5	64.8	65.7	365.3

(Unit-II, Prob.3)

15. (a) Explain different types of fertility rates. **(Unit-III, Q.No.6)**

OR

(b) Define vital statistics. Explain the uses of vital statistics. **(Unit-III, Q.No.1)**

16. (a) Define Cost of Living Index Numbers. Explain various methods of construction of Cost of Living Index Numbers. **(Unit-IV, Q.No.24)**

OR

(b) Explain briefly about National Sample Survey Organisation (NSSO). **(Unit-IV, Q.No.3)**

FACULTY OF SCIENCE

B.Sc. VI-Semester(CBCS) Examination

MODEL PAPER - III

Subject : STATISTICS

Paper - VI - A : Applied Statistics - II

Time : 3 Hours] **[Max. Marks : 80**

PART - A (8 × 4 = 32 Marks)

Note : Answer any Eight questions. All questions carry equal marks.

ANSWERS

1. Efficiency of a Design. **(Unit-I, SQA-10)**
2. Importance of Design of Experiments. **(Unit-I, SQA-8)**
3. Analysis of variance of two-way classification. **(Unit-I, SQA-6)**
4. Merits and Demerits. **(Unit-II, SQA-9)**
5. Efficiency of a design. **(Unit-II, SQA-3)**
6. Advantages of RBD. **(Unit-II, SQA-5)**
7. Define life table **(Unit-III, SQA-10)**
8. Crude Death Rate **(Unit-III, SQA-8)**
9. Gross Reproductive Rate. **(Unit-III, SQA-9)**
10. Define National Income. **(Unit-IV, SQA-4)**
11. What is Fisher's Ideal Index ? **(Unit-IV, SQA-7)**
12. Define cost of living index numbers. **(Unit-IV, SQA-9)**

PART - B (4 × 12 = 48 Marks)

Note : Answer all the questions. All questions carry equal marks.

13. (a) What is an experimental design. Explain the principles of experimental design. **(Unit-I, Q.No.10)**

OR

(b) Give complete statistical analysis of ANOVA One-way classification. **(Unit-I, Q.No.5)**

14. (a) Compare the efficiency of RBD relative to CRD. **(Unit-II, Q.No.12)**

OR

(b) Discuss about Randomized Block Design (R.B.D). **(Unit-II, Q.No.3)**

15. (a) Write in detail about Abridged life tables. **(Unit-III, Q.No.9)**

OR

(b) Explain different types of fertility rates. **(Unit-III, Q.No.5)**

16. (a) Explain the utility and difficulties in estimation of national income. **(Unit-IV, Q.No.9)**

OR

(b) Calculate Fisher's Ideal Index from the following data and prove that it satisfies both the time reversal and factor reversal tests :

Commodity	2010		2011	
	Price	Qty	Price	Qty
A	8	80	10	120
B	10	120	12	96
C	5	40	5	50
D	4	56	3	60
E	20	100	25	150

(Unit-IV, Prob.7)

FACULTY OF SCIENCE

B.Sc. VI-Semester(CBCS) Examination

May / June - 2019

STATISTICS

(Design of Experiments, Vital Statistics, Official Statistics and Business Forecasting)

PAPER - VII DSC

Time : 3 Hours] **[Max. Marks : 60**

PART - A (5 × 3 = 15 Marks)

Note : Answer any five of the following questions.

ANSWERS

1. ANOVA Table for a two-way classified data **(Unit-I, SQA-6)**
2. Efficiency of a Design **(Unit-II, SQA-3)**
3. Advantages of LSD over CRD **(Unit-II, SQA-4)**
4. Describe briefly the methods for calculating national income. **(Unit-IV, SQA-8)**
5. Role of forecasting in Business **(Out of Syllabus)**
6. Principle of Local control **(Unit-I, SQA-4)**
7. Abridged Life Tables **(Unit-III, SQA-3)**
8. Specific Death Rate **(Unit-III, SQA-6)**

PART - B (4 × 12 = 48 Marks)

Note : Answer all the questions. All questions carry equal marks.

9. (a) Give complete statistical analysis of ANOVA one-way classification. **(Unit-I, Q.No.5)**

OR

(b) Give the layout of a Randomized Block Design and estimate the missing observation in this Design. Also give its statistical analysis. **(Unit-II, Q.No.3,4,6)**

10. (a) What is the Design of Latin Square? Obtain the expectations of various sum of squares in Latin Square Design. **(Unit-II, Q.No.7,8)**

OR

(b) Explain the methods to collect yield statistics of various crops and / explain functions of CSO. **(Unit-IV, Q.No.2,6)**

11. (a) Define and compare various measures of fertility. **(Unit-III, Q.No.5)**

OR

(b) Define a Life Table. Describe various components of a Life Table. **(Unit-III, Q.No.7,8)**

FACULTY OF SCIENCE

B.Sc. VI-Semester(CBCS) Examination

September / October - 2020

STATISTICS

(Design of Experiments, Vital Statistics, Official Statistics and Business Forecasting)

PAPER - VII DSC

Time : 2 Hours] **[Max. Marks : 60**

PART - A (4 × 5 = 20 Marks)

Note : Answer any four questions.

ANSWERS

1. Define Birth rate and Age specific birth rate. **(Unit-III, SQA-5,12)**
2. Explain the advantages of Randomised Blocked Design (RBD). **(Unit-II, SQA-5)**
3. State Cochran's theorem. Give its applications. **(Unit-I, SQA-2)**
4. Write the formulae for one missing observation in Latin Square Design (LSD) and explain them. **(Unit-II, SQA-6)**
5. Explain the assumptions in Life Table. **(Unit-III, SQA-4)**
6. Explain the steps in Forecasting. **(Out of Syllabus)**
7. Explain about Agricultural statistics in brief. **(Unit-IV, SQA-2)**
8. Define standardized death rates. Why do we need them? **(Unit-III, SQA-2)**

PART - B (2 × 20 = 40 Marks)

Note : Answer any two questions

9. Explain analysis of ANOVA for two-way classification and stating the assumptions. **(Unit-I, Q.No.6)**
10. Write in detail about principles of Experimentation. **(Unit-I, Q.No.10)**
11. Write a detailed notes on Central Statistical Organization (CSO). **(Unit-IV, Q.No.2)**
12. Estimate the missing value in a Randomized Blocked Design (RBD) and state the differences in its analysis when compared to complete randomized design. **(Unit-II, Q.No.6,12)**
13. Write a detailed notes on population growth and how it can be measured. Explain. **(Unit-III, Q.No.6)**
14. Define various fertility rates and give suitable examples. **(Unit-III, Q.No.5)**

FACULTY OF SCIENCE

B.Sc. VI-Semester(CBCS) Examination

July - 2021

STATISTICS

(Design of Experiments, Vital Statistics, Official Statistics and Business Forecasting)

PAPER - VII DSC

Time : 2 Hours] [Max. Marks : 60

PART - A (4 × 5 = 20 Marks)

Note : Answer any four questions.

		ANSWERS
1.	State Gauss-Markov of theorem.	**(Unit-I, SQA-1)**
2.	Define Completely Randomized Design. Write its merits and demerits.	**(Unit-II, SQA-7)**
3.	Define Latin Square Design (LSD). Write 4x4 Latin square.	**(Unit-II, SQA-8)**
4.	State the functions of CSO.	**(Unit-IV, SQA-3)**
5.	Discuss the role of forecasting in Business.	**(Out of Syllabus)**
6.	Wrife the assumptions and uses of life table.	**(Unit-III, SQA-4)**
7.	Define Gross Reproduction Rate (GRR) and Net Reproduction Rate (NRR).	**(Unit-III, SQA-9)**
8.	Define Crude Birth Rate. Write its merits and demerits.	**(Unit-III, SQA-5)**

PART - B (2 × 20 = 40 Marks)

Note : Answer any two questions

9.	State the assumptions, explain the analysis of variance for One-way classified data.	**(Unit-I, Q.No.3,4)**
10.	Write the complete analysis of Randomized Block Design (RBD).	**(Unit-II, Q.No.3)**
11.	How can you estimate the missing observation (one) in LSD, then write is ANOVA?	**(Unit-II, Q.No.9)**
12.	What is NS SO? What are its functions and explain how NS SO helps the government of India in evaluating economic status from time to time?	**(Unit-IV, Q.No.2)**
13.	Define and explain various measures of Mortality's Rates.	**(Unit-III, Q.No.4)**
14.	What is population growth measure? Explain different population growth measures.	**(Unit-III, Q.No.6)**

FACULTY OF SCIENCE

B.Sc. VI-Semester(CBCS) Examination

November / December - 2021

STATISTICS

(Design of Experiments, Vital Statistics, Official Statistics and Business Forecasting)

PAPER - VII DSC

Time : 3 Hours] **[Max. Marks : 60**

PART - A (4 × 5 = 20 Marks)

Note : Answer any four questions.

ANSWERS

1. State Cochran's theorem and write its need of study. **(Unit-I, SQA-2)**
2. What is Completely Randomized Design (CRD)? **(Unit-II, SQA-7)**
3. Define Replication and Randomization. **(Unit-I, SQA-3,5)**
4. Write the merits and demerits of Latin Square Design (LSD). **(Unit-II, SQA-9)**
5. Describe National Income and its uses. **(Unit-IV, SQA-4,5)**
6. Explain the steps involved in Business forecasting. **(Out of Syllabus)**
7. Define Abridged Life table. **(Unit-III, SQA-3)**
8. Write on the sources of vital statistics. **(Unit-III, SQA-11)**

PART - B (2 × 20 = 40 Marks)

Note : Answer any two questions

9. By stating the assumptions explain the analysis of variance for two-way classification data. **(Unit-I, Q.No.6)**
10. Give the Layout of Randomized Block Design (RBD) with four treatments A, B, C, D each replicated 5 times. Derive the efficiency of RBD over CRD. **(Unit-II, Q.No.3,12)**
11. How can you estimate the missing observations in a Randomized Block Design (when two observations are missing)? Then write its ANOVA. **(Unit-II, Q.No.6)**
12. Explain functions and organizations of CSO. **(Unit-IV, Q.No.2)**
13. Define and explain various measures of fertility rate. Discuss their use. **(Unit-III, Q.No.5)**
14. Stating the assumptions, explain the construction of complete Life table. **(Unit-III, Q.No.7,8)**

FACULTY OF SCIENCE

B.Sc. VI-Semester(CBCS) Examination

July - 2022

Subject : STATISTICS

Paper - VI - A : Applied Statistics - II

Time : 3 Hours] **[Max. Marks : 80**

PART - A (8 × 4 = 32 Marks)

Note : Answer any Eight questions.

ANSWERS

1. Explain the concept of gauss markov linear model. **(Unit-I, SQA-1)**
2. Find the expectation of error sum of squares in two way classification. **(Unit-I, SQA-2)**
3. Write the statistical analysis of CRD. **(Unit-II, SQA-1)**
4. Explain replication and local control. **(Unit-I, SQA-3,4)**
5. Write the ANOVA table of LSD. **(Unit-II, SQA-2)**
6. Explain the concept of critical difference. **(Unit-II, SQA-11)**
7. Explain the uses of vital statistics. **(Unit-III, SQA-1)**
8. Explain Standardised death rates. **(Unit-III, SQA-2)**
9. Explain the abridged Life table. **(Unit-III, SQA-3)**
10. Explain the functions of NSSO. **(Unit-IV, SQA-6)**
11. Why Fisher's index is called Ideal? Explain. **(Unit-IV, SQA-7)**
12. Explain Splicing. **(Unit-IV, SQA-8)**

PART - A (4 × 12 = 48 Marks)

Note : Answer all the questions.

13. (a) Explain the Analysis of Variance of Two Way classification. **(Unit-I, Q.No.6)**

OR

(b) Find the Expectation of Treatment and Error sum of squares in one-way Classification. **(Unit-I, Q.No.5)**

14. (a) Find the Efficiency of RBD over CRD. **(Unit-II, Q.No.12)**

OR

(b) How do you estimate the missing observation in LSD? Give its statistical Analysis. **(Unit-II, Q.No.9)**

15. (a) Explain different types of Fertility rates. **(Unit-III, Q.No.5)**

OR

(b) What is Complete Life Table? Describe Various Components of a Life Table. **(Unit-III, Q.No.7,8)**

16. (a) Distinguish between fixed base and chain base index numbers. From the fixed base index numbers given below, construct chain base index numbers.

Year	2003	2004	2005	2006	2007	2008
Fixed Base index	94	98	102	95	98	100

(Unit-IV, Q.No.23, Prob.9)

OR

(b) Define Cost of Living Index Numbers. Describe various methods of its computation also give its uses. **(Unit-IV, Q.No.24)**